High Dimensional Differentially Private Stochastic Optimization with Heavy-tailed Data

High Dimensional Differentially Private Stochastic Optimization with Heavy-tailed Data

Lijie Hu
King Abdullah University of Science and Technology
Thuwal, Makkah, Saudi Arabia
lijie.hu@kaust.edu.sa

Shuo Ni
University of Southern California
Los Angeles, California, United States
shuoni@usc.edu

Hanshen Xiao
Massachusetts Institute of Technology
Cambridge, Massachusetts, United States
hsxiao@mit.edu

Di Wang
King Abdullah University of Science and Technology
Thuwal, Makkah, Saudi Arabia
di.wang@kaust.edu.sa

ABSTRACT

As one of the most fundamental problems in machine learning, statistics and differential privacy, Differentially Private Stochastic Convex Optimization (DP-SCO) has been extensively studied in recent years. However, most of the previous work can only handle either regular data distributions or irregular data in the low dimensional space case. To better understand the challenges arising from irregular data distributions, in this paper we provide the first study on the problem of DP-SCO with heavy-tailed data in the high dimensional space. In the first part we focus on the problem over some polytope constraint (such as the ℓ_1-norm ball). We show that if the loss function is smooth and its gradient has bounded second order moment, it is possible to get a (high probability) error bound (excess population risk) of $\tilde{O}(\frac{\log d}{(n\epsilon)^{\frac{1}{3}}})$ in the ϵ-DP model, where n is the sample size and d is the dimension of the underlying space. Next, for LASSO, if the data distribution has bounded fourth-order moments, we improve the bound to $\tilde{O}(\frac{\log d}{(n\epsilon)^{\frac{2}{5}}})$ in the (ϵ, δ)-DP model. In the second part of the paper, we study sparse learning with heavy-tailed data. We first revisit the sparse linear model and propose a truncated DP-IHT method whose output could achieve an error of $\tilde{O}(\frac{s^{*2}\log^2 d}{n\epsilon})$, where s^* is the sparsity of the underlying parameter. Then we study a more general problem over the sparsity (*i.e.,* ℓ_0-norm) constraint, and show that it is possible to achieve an error of $\tilde{O}(\frac{s^{*\frac{3}{2}}\log d}{n\epsilon})$, which is also near optimal up to a factor of $\tilde{O}(\sqrt{s^*})$, if the loss function is smooth and strongly convex.

CCS CONCEPTS

• **Security and privacy → Data anonymization and sanitization**; • **Computing methodologies → Machine learning algorithms**.

PODS '22, June 12–17, 2022, Philadelphia, PA, USA

ACM ISBN 978-1-4503-9260-0/22/06…$15.00
https://doi.org/10.1145/3517804.3524144

KEYWORDS

differential privacy; stochastic convex optimization; high dimensional statistics; robust statistics

ACM Reference Format:
Lijie Hu, Shuo Ni, Hanshen Xiao, and Di Wang. 2022. High Dimensional Differentially Private Stochastic Optimization with Heavy-tailed Data. In *Proceedings of the 41st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS '22), June 12–17, 2022, Philadelphia, PA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3517804.3524144

1 INTRODUCTION

Privacy-preservation has become an important consideration and now is a challenging task for machine learning algorithms with sensitive data. To address the privacy issue, Differential Privacy (DP) has received a great deal of attentions and now has established itself as a de facto notation of privacy for data analysis. Methods to guarantee differential privacy have been widely studied, and recently adopted in industry [19, 46].

Stochastic Convex Optimization (SCO) [48] and its empirical form, Empirical Risk Minimization (ERM), are the most fundamental problems in machine learning and statistics, which include several basic models, such as linear regression and logistic regression. They find numerous applications in many areas such as medicine, finance, genomics and social science. Due to their importance, the problem of designing DP algorithms for SCO or ERM (*i.e.,* DP-SCO and DP-ERM) have been extensively studied for nearly a decade starting from [17, 18]. Later on, a long list of works have attacked the problems from different perspectives: [5–7, 26, 31, 44, 58] studied the problems in the low dimensional case and the central model, [14, 34, 35, 45, 55] considered the problems in the high dimensional sparse case and the central model, [20, 21, 43, 50] focused on the problems in the local model.

However, most of those previous work can only handle regular data, *i.e.,* they need to assume either the underlying data distribution is bounded or sub-Gaussian, or the loss function is $O(1)$-Lipschitz for all the data. This is particularly true for those output perturbation based [18] and objective or gradient perturbation based [7] DP methods. However, such assumptions may not always hold when dealing with real-world datasets, especially those from biomedicine and finance, which are often heavy-tailed [8, 30, 56], implying that existing algorithms may fail to guarantee the DP property.

Compared with bounded data, heavy-tailed data could lead to unbounded gradient and thus violate the Lipschitz condition. For example, consider the linear squared loss $\ell(w,(x,y)) = (w^T x - y)^2$. When x is heavy-tailed, the gradient of $\ell(w,(x,y))$ becomes unbounded. To address the issue, one potential approach is to truncating or trimming the gradient, such as in [1]. However, there is no existing convergence result based on their algorithm. Thus, new private and robust estimation methods for heavy-tailed data are needed.

Recently, there are several work studied private mean estimation or DP-SCO with heavy-tailed data [4, 33, 37, 51] (see Section 2 for details). However, the estimation errors of these results all are dependent on polynomial in the dimension of the underlying space, which impedes them to be implemented to the high dimensional setting, where the dimension is far greater than the sample size. In contrast, as we mentioned earlier, high dimensional DP-SCO with regular data has been studied quite well. Thus, our question is, what are the theoretical behaviors of DP-SCO with heavy-tailed data in the high dimensional space? In this paper, we provide a comprehensive and the first study on the problem under different settings by providing several new methods. Our contributions can be summarized as the following,

(1) We first study DP-SCO over some polytope constraint, which has been studied in [2, 45] for regular data. We first show that if the loss function is smooth and its gradient has bounded second order moment, it is possible to get an excess population risk (error bound) of $\tilde{O}(\frac{\log d}{(n\epsilon)^{\frac{1}{3}}})$ with high probability in the ϵ-DP model, where n is the sample size and d is the dimensionality of the underlying space. Next, for LASSO, if the data distribution has bounded fourth-order moments, we improve the bound to $\tilde{O}(\frac{\log d}{(n\epsilon)^{\frac{2}{5}}})$ in the (ϵ,δ)-DP model.

(2) We then study DP-SCO for sparse learning with heavy-tailed data in the (ϵ,δ)-DP model, which has been studied in [13, 52, 54] in the regular data case. We first revisit the sparse linear regression problem and propose a new method whose output could achieve an error bound of $\tilde{O}(\frac{s^{*2}\log^2 d}{n\epsilon})$, where s^* is the sparsity of the underlying parameter. Then we study a general DP-SCO problem under the sparsity constraint, and show that it is possible to achieve an error of $\tilde{O}(\frac{s^{*\frac{3}{2}}\log d}{n\epsilon})$, if the loss function is smooth and strongly convex. We also show this bound is near optimal up to a factor of $O(\sqrt{s^*}\log^2 n)$. To get these results, we provide several new methods and hard instances which may be used to in other machine learning problems.

Due to space limit, all the proofs and lemmas, and experiments on synthetic and real-world data are included in the full version of the paper [29].

2 RELATED WORK

As mentioned earlier, there is a long list of results on DP-SCO and DP-ERM. However, most of them consider the case where the underlying data distribution is sub-Gaussian and cannot be extended to heavy-tailed case. On the other side, in the non-private case, recently a number of works have studied the SCO and ERM problems with heavy-tailed data, such as [9, 27, 28, 36, 39, 40, 42]. It is not clear whether they can be adapted to private versions and in the high dimensional setting.

For DP-SCO or private estimation for heavy-tailed distribution, [4] provides the first study on private mean estimation for distributions with bounded moment and proposes the minimax private rates. Their methods are based on truncating the data to make each data record has a bounded ℓ_2-norm. However, as [33] mentioned, they need a stronger assumption on the bounded moment, *e.g., for the mean estimation problem they need to assume* $\mathbb{E}[\|x\|_2^2] \le 1$ *while we only assume* $\mathbb{E}[x_j^2] \le 1$ *for each coordinate* $j \in [d]$. Moreover, their method cannot be extended to the high dimensional sparse setting directly, and their error bound is in the expectation form, while in the robust statistics it is preferable to get high probability results (see Definition 3 for details). Later, [33] also studies the heavy-tailed mean estimation, which is also studied by [37] recently. However, their results for general d dimensional space are still not the high probability form (they can only show their results hold with probability at least 0.7). Thus, their methods cannot be used to DP-SCO directly. Moreover, it is unknown whether their methods could be extended to the high dimensional or the sparse setting. [10] recently also studies the same problem and proposes a method based on the PTR mechanism [22]. However, their method can be only used in the 1-dimensional space and and needs stronger assumptions.

Meanwhile, instead of the mean estimation, [51] provides the first study on DP-SCO with heavy-tailed data and proposes three methods based on different assumptions. Their first method is based on the Sample-and-Aggregate framework [41]. However, this method needs enormous assumptions and its error bound is quite large. Their second method is still based on the smooth sensitivity [12]. However, [51] needs to assume the distribution is sub-exponential. It also provides a new private estimator motivated by the previous work in robust statistics. While some our estimators are quite similar as theirs, they are quite a lot differences (see Remark 1 for details). Based on the mean estimator in [33], [32] recently studies DP-SCO and improves the (expected) excess population risk to $\tilde{O}((\frac{d}{\epsilon n})^{\frac{1}{2}})$ and $\tilde{O}(\frac{d}{\epsilon n})$ for convex and strongly convex loss functions respectively under the assumption that the gradient of the loss has bounded second order moment. These results match the best known result of the heavy-tailed mean estimation problem. However, all of these results are in the expectation form instead of the high probability form. Moreover, their method cannot be extended to the linear model, where the bounded second order moment of loss assumption is quite strong (see Assumption 3 for details). We note that all these methods cannot be directly extended to the high dimensional case or the sparse learning problem. [1]

3 PRELIMINARIES

Notations: For vectors $v, v_i \in \mathbb{R}^d$, we denote v_j and $v_{i,j}$ as their corresponding the j-th coordinate. Given a set of indices $S \subseteq [d]$, we denote the vector $v_S \in \mathbb{R}^d$ as the projection of v onto S, *i.e.,* $v_{S,j} = v_j$ if $j \in S$, and $v_{S,j} = 0$ otherwise. We also denote $|S|$ as the number of elements in S and $\mathrm{supp}(w) = \{j \in [d] : w_j \neq 0\} \subseteq [d]$

[1] We refer readers the reference [33, 51] to see more related work on DP methods for unbounded sensitivity.

for w. For a constraint set $\mathcal{W}$, we denote $\|\mathcal{W}\|_1$ as it is ℓ_1-norm diameter, *i.e.,* $\|\mathcal{W}\|_1 = \max_{u,v\in\mathcal{W}} \|u - v\|_1$.

Definition 1 (Differential Privacy [23])**.** Given a data universe $\mathcal{X}$, we say that two datasets $D, D' \subseteq \mathcal{X}$ are neighbors if they differ by only one data sample, which is denoted as $D \sim D'$. A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private (DP) if for all neighboring datasets D, D' and for all events S in the output space of $\mathcal{A}$, we have

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

In this paper, we will mainly use the Laplacian and the Exponential mechanism, and the Advanced Composition Theorem to guarantee DP property.

Definition 2 (Laplacian Mechanism)**.** Given a function $q : \mathcal{X}^n \to \mathbb{R}^d$, the Laplacian Mechanism is defined as: $\mathcal{M}_L(D, q, \epsilon) = q(D) + (Y_1, Y_2, \cdots, Y_d)$, where Y_i is i.i.d. drawn from a Laplacian Distribution $\mathrm{Lap}(\frac{\Delta_1(q)}{\epsilon})$, where $\Delta_1(q)$ is the ℓ_1-sensitivity of the function q, *i.e.,* $\Delta_1(q) = \sup_{D\sim D'} ||q(D) - q(D')||_1$. For a parameter λ, the Laplacian distribution has the density function $\mathrm{Lap}(\lambda)(x) = \frac{1}{2\lambda}\exp(-\frac{x}{\lambda})$. Laplacian Mechanism preserves ϵ-DP.

Definition 3 (Exponential Mechanism)**.** The Exponential Mechanism allows differentially private computation over arbitrary domains and range $\mathcal{R}$, parametrized by a score function $u(D, r)$ which maps a pair of input data set D and candidate result $r \in \mathcal{R}$ to a real valued score. With the score function u and privacy budget ϵ, the mechanism yields an output with exponential bias in favor of high scoring outputs. Let $\mathcal{M}(D, u, \mathcal{R})$ denote the exponential mechanism, and Δ be the sensitivity of u in the range $\mathcal{R}$, *i.e.,* $\Delta = \max_{r\in\mathcal{R}} \max_{D\sim D'} |u(D, r) - u(D', r)|$. Then if $\mathcal{M}(D, u, R)$ selects and outputs an element $r \in \mathcal{R}$ with probability proportional to $\exp(\frac{\epsilon u(D,r)}{2\Delta u})$, it preserves ϵ-DP.

The output of exponential mechanism has the following utility.

Lemma 1 ([24])**.** For the exponential mechanism $\mathcal{M}(D, u, \mathcal{R})$, we have

$$\Pr\{u(\mathcal{M}(D, u, \mathcal{R})) \leq \mathrm{OPT}_u(x) - \frac{2\Delta u}{\epsilon}(\ln|\mathcal{R}| + t)\} \leq e^{-t}.$$

where $\mathrm{OPT}_u(x)$ is the highest score in the range $\mathcal{R}$, *i.e.* $\max_{r\in\mathcal{R}} u(D, r)$.

Lemma 2 (Advanced Composition Theorem)**.** Given target privacy parameters $0 < \epsilon < 1$ and $0 < \delta < 1$, to ensure $(\epsilon, T\delta' + \delta)$-DP over T mechanisms, it suffices that each mechanism is (ϵ', δ')-DP, where $\epsilon' = \frac{\epsilon}{2\sqrt{2T\ln(2/\delta)}}$ and $\delta' = \frac{\delta}{T}$.

Definition 4 (DP-SCO [7])**.** Given a dataset $D = \{z_1, \cdots, z_n\}$ from a data universe $\mathcal{Z}$ where $z_i = (x_i, y_i)$ with a feature vector x_i and a label/response y_i are i.i.d. samples from some unknown distribution $\mathcal{D}$, a convex constraint set $\mathcal{W} \subseteq \mathbb{R}^d$, and a convex loss function $\ell : \mathcal{W}\times\mathcal{Z} \mapsto \mathbb{R}$. Differentially Private Stochastic Convex Optimization (DP-SCO) is to find w^{priv} so as to minimize the population risk, *i.e.,* $L_{\mathcal{D}}(w) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(w, z)]$ with the guarantee of being differentially private.[2] The utility of the algorithm is measured by the excess population risk, that is

$$L_{\mathcal{D}}(w^{\mathrm{priv}}) - \min_{w\in\mathcal{W}} L_{\mathcal{D}}(w).$$

[2]Note that in this paper we consider the improper learning case, that is w^{priv} may not in $\mathcal{W}$.

Besides the population risk, we can also measure the *empirical risk* of dataset D: $\hat{L}(w, D) = \frac{1}{n}\sum_{i=1}^n \ell(w, z_i)$. It is notable that in the **high probability setting**, we need to get a high probability excess population risk. That is given a failure probability $0 < \zeta < 1$, we want get a (polynomial) function $f(d, \log\frac{1}{\delta}, \log\frac{1}{\zeta}, \frac{1}{n}, \frac{1}{\epsilon})$ such that with probability at least $1-\zeta$ (over the randomness of the algorithm and the data distribution),

$$L_{\mathcal{D}}(w^{\mathrm{priv}}) - \min_{w\in\mathcal{W}} L_{\mathcal{D}}(w) \leq O(f(d, \log\frac{1}{\delta}, \log\frac{1}{\zeta}, \frac{1}{n}, \frac{1}{\epsilon})).$$

Compared with the high probability setting, there is another setting namely the expectation setting where our goal is to get a (polynomial) function $f(d, \log\frac{1}{\delta}, \frac{1}{n}, \frac{1}{\epsilon})$ such that

$$\mathbb{E}L_{\mathcal{D}}(w^{\mathrm{priv}}) - \min_{w\in\mathcal{W}} L_{\mathcal{D}}(w) \leq O(f(d, \log\frac{1}{\delta}, \frac{1}{n}, \frac{1}{\epsilon})),$$

where the expectation takes over the randomness of the data records and the algorithm.

It is notable that, in the regular data case where the data distribution $\mathcal{D}$ or the gradient of the loss is bounded or sub-Gaussian, it is easy to transform an expected excess population risk to an excess population risk with high probability. However, this is not true for the heavy-tailed case. [3] Thus, all of the recent studies on robust statistics such as [9, 27, 28, 36, 39, 40, 42] focused on the high probability setting. In the paper, we will study the problem in the high probability setting. Moreover, throughout the paper we focus on the high dimensional case where d could be far greater than n. Thus we wish the error bounds (excess population risk) be logarithmic of d.

The following two definitions on loss functions are commonly used in machine learning, optimization and statistics.

Definition 5. A function f is L-Lipschitz w.r.t the norm $\|\cdot\|$ if for all $w, w' \in \mathcal{W}$, $|f(w) - f(w')| \leq L\|w - w'\|$.

Definition 6. A function f is α-smooth on $\mathcal{W}$ if for all $w, w' \in \mathcal{W}$, $f(w') \leq f(w) + \langle\nabla f(w), w' - w\rangle + \frac{\alpha}{2}\|w' - w\|_2^2$.

4 HIGH DIMENSIONAL DP-SCO OVER POLYTOPE DOMAIN

In this section we will study DP-SCO over polytope domain, *i.e.,* the underlying constraint set $\mathcal{W}$ is some polytope and thus could be written as the convex hull of a finite set V. This contains numerous of learning models that address high dimensional data, such as LASSO and minimization over probability simplex.

[45] first studied the problem of DP-ERM over polytope domain in the regular data setting (*i.e.,* the gradient of loss function has bounded norm). Specifically, they showed that when the loss function is Lipschitz w.r.t ℓ_1-norm, there is an (ϵ, δ)-DP algorithm (DP Frank-Wolfe) whose output could achieve an error of $O(\frac{\log(|V|n)}{(n\epsilon)^{\frac{2}{3}}})$. However, to generalize to the heavy-tailed data setting, the main difficulty is that the assumption of ℓ_1-norm Lipschitz does not hold anymore. To address the problem, one possible approach may be truncating the gradient to make it has bounded ℓ_∞-norm (since ℓ_1-norm Lipschitz is equivalent to its gradient has bounded ℓ_∞-norm). However, as mentioned in [51], it could introduce enormous

[3]See [15] for the necessity to consider the high probability setting.

amount of error and it is difficult to select the best threshold parameter. In the following we will propose a new method to overcome this challenge. We will focus on the case where the gradient of the loss is heavy-tailed. Specifically, following from the previous work on robust statistics such as [27, 42], here we propose the following assumption on the gradient of the loss function.

Assumption 1. We assume $L_D(\cdot)$ is α-smooth, and there exists a $\tau > 0$ such that for any $w \in \mathcal{W}$ and each coordinate $j \in [d]$, we have $\mathbb{E}[(\nabla_j \ell(w,x))^2] \le \tau$.

First, it is notable that the smoothness condition in Assumption 1 is necessary for the high dimensional setting. As shown by [2], when the loss function is non-smooth and ℓ_1-norm Lipschitz, even in the regular data setting the excess population risk is lower bounded by $\Omega(\sqrt{\frac{\log d}{n}} + \frac{\sqrt{d}}{n\epsilon})$, which depends on $\Omega(\sqrt{d})$. Secondly, in some other work on studying private estimation for distributions with bounded second-order moment (such as [33]), they assume that for each unit vector $u \in \mathbb{R}^d$, $\mathbb{E}[\langle u, \nabla \ell(w,x)\rangle^2] \le \tau = O(1)$. Thus, our assumption on the moment is reasonable. Thirdly, we note that τ may be not a constant, it could depend on the structure of the loss function, data distribution and the underlying structure of $\mathcal{W}$ [49]. Throughout the whole paper we assume τ is known, which is commonly used in other related work in robust statistics such as [11, 33].

Our approach, namely Heavy-tailed DP-FW, could be seen as a generalization of the DP Frank-Wolfe method in [45]. The approach is motivated by a robust mean estimator for heavy-tailed distribution given by [16] which was extended by [27]. For simplicity, we first consider a 1-dimensional random variable x and assume that $x_1, x_2, \cdots, x_n$ are i.i.d. sampled from x. The robust mean estimator consists of three steps:

Scaling and Truncation For each sample x_i, we first re-scale it by dividing s (which will be specified later). Then, the re-scaled one was passed through a soft truncation function ϕ. Finally, we put the truncated mean back to the original scale. That is,

$$\frac{s}{n}\sum_{i=1}^{n}\phi(\frac{x_i}{s}) \approx \mathbb{E}x. \tag{1}$$

Here, we use the function given in [16],

$$\phi(x) = \begin{cases} x - \frac{x^3}{6}, & -\sqrt{2} \le x \le \sqrt{2} \\ \frac{2\sqrt{2}}{3}, & x > \sqrt{2} \\ -\frac{2\sqrt{2}}{3}, & x < -\sqrt{2}. \end{cases} \tag{2}$$

A key property for ϕ is that ϕ is bounded, that is, $|\phi(x)| \le \frac{2\sqrt{2}}{3}$.

Noise Multiplication Let $\eta_1, \eta_2, \cdots, \eta_n$ be random noise generated from a common distribution $\eta \sim \chi$ with $\mathbb{E}\eta = 0$. We multiply each data x_i by a factor of $1 + \eta_i$, and then perform the scaling and truncation step on the term $x_i(1+\eta_i)$. That is,

$$\tilde{x}(\eta) = \frac{s}{n}\sum_{i=1}^{n}\phi(\frac{x_i + \eta_i x_i}{s}). \tag{3}$$

Noise Smoothing In this final step, we smooth the multiplicative noise by taking the expectation w.r.t. the distributions. In total the robust mean estimator $\hat{x}(s,\beta)$ could be written as,

$$\hat{x}(s,\beta) = \mathbb{E}\tilde{x}(\eta, s, \beta) = \frac{s}{n}\sum_{i=1}^{n}\int \phi(\frac{x_i + \eta_i x_i}{s}) d\chi(\eta_i). \tag{4}$$

Computing the explicit form of each integral in (4) depends on the function $\phi(\cdot)$ and the distribution χ. Fortunately, [16] showed that when ϕ is in (2) and $\chi \sim \mathcal{N}(0, \frac{1}{\beta})$ (where β will be specified later), we have for any a and $b > 0$

$$\mathbb{E}_\eta \phi(a + b\sqrt{\beta}\eta) = a(1 - \frac{b^2}{2}) - \frac{a^3}{6} + \hat{C}(a,b), \tag{5}$$

where $\hat{C}(a,b)$ is a correction form which is easy to implement and its explicit form will be given in Appendix.

The key idea of our method is that, by the definition of $\hat{x}(s,\beta)$ in (4) and the function ϕ is in (2), we can see that the value of $\hat{x}(s,\beta)$ will be changed at most $\frac{4\sqrt{2}s}{3n}$ if we change one sample in the data, *i.e.,* the sensitivity of $\hat{x}(s,\beta)$ is bounded by $\frac{4\sqrt{2}s}{3n}$. That is, given a fixed vector w and n gradients $\{\nabla\ell(w,z_i)\}_{i=1}^n$, we can use the above estimator to the entrywise of these gradients to get an estimator (we denote it as $\tilde{g}(w,D)$) of $\mathbb{E}[\ell(w,z)]$. Moreover, we can see the ℓ_∞-norm sensitivity of $\tilde{g}(w,D)$ is bounded $\frac{4\sqrt{2}s}{3n}$, *i.e.,* $\|\tilde{g}(w,D) - \tilde{g}(w,D')\|_\infty \le \frac{4\sqrt{2}s}{3n}$, where D and D' are neighboring datasets. Combing this result with DP Frank Wolfe method, we propose our algorithm. See Algorithm 1 for details.

Algorithm 1 Heavy-tailed DP-FW

1: **Input:** n-size dataset D, loss function $\ell(\cdot,\cdot)$, initial parameter w^0, parameters $s, T, \beta, \{\eta_t\}_t$ (will be specified later), privacy parameter ϵ, failure probability ζ. $\mathcal{W}$ is the convex hull of a finite set V.
2: Split the data D into T parts $\{D_t\}_{t=1}^T$ with $|D_t| = m = \frac{n}{T}$.
3: **for** $t = 1, \cdots, T$ **do**
4: For each $j \in [d]$, calculate the robust gradient by (2)-(5), that is

$$g_j^{t-1}(w^{t-1}, D_t) = \frac{1}{m}\sum_{x\in D_t}\left(\nabla_j\ell(w^{t-1},x)\big(1 - \frac{\nabla_j^2\ell(w^{t-1},x)}{2s^2\beta}\big) - \frac{\nabla_j^3\ell(w^{t-1},x)}{6s^2}\right) + \frac{s}{m}\sum_{x\in D_t}\hat{C}\left(\frac{\nabla_j\ell(w^{t-1},x)}{s}, \frac{|\nabla_j\ell(w^{t-1},x)|}{s\sqrt{\beta}}\right).$$

5: Let vector $\tilde{g}(w^{t-1}, D_t) \in \mathbb{R}^d$ as $\tilde{g}(w^{t-1}, D_t) = (g_1^{t-1}(w^{t-1}, D_t), g_2^{t-1}(w^{t-1}, D_t), \cdots, g_d^{t-1}(w^{t-1}, D_t))$.
6: Denote the score function $u(D_t, \cdot) : V \mapsto \mathbb{R}$ such that for each $v \in V$ let $u(D_t, v) = -\langle v, \tilde{g}(w^{t-1}, D_t)\rangle$. Run the exponential mechanism with the range $R = V$, sensitivity $\Delta = \frac{4\|\mathcal{W}\|_1\sqrt{2}s}{3m}$ and the privacy budget ϵ. Denote the output as $\tilde{w}^{t-1} \in \mathcal{W}$.
7: Let $w^t = (1 - \eta_{t-1})w^{t-1} + \eta_{t-1}\tilde{w}^{t-1}$.
8: **end for**
9: **return** w^T.

Theorem 1. For any $\epsilon > 0$, Algorithm 1 is ϵ-DP.

Theorem 2. Under Assumption 1 and if $\mathcal{W}$ is a convex hull of a finite compact set V. Then for any given probability of failure $0 < \zeta < 1$, if we set $T = \tilde{O}\big((\frac{n\epsilon\alpha^2}{\tau \log \frac{|V|d}{\zeta}})^{\frac{1}{3}}\big)$, $\beta = O(1), s = O(\sqrt{\frac{n\epsilon\tau}{T \log \frac{|V|dT}{\zeta}}})$ and $\eta_{t-1} = \frac{2}{t+2}$ in Algorithm 1, with probability at least $1-\zeta$,

$$L_{\mathcal{D}}(w^T) - \min_{w\in\mathcal{W}} L_{\mathcal{D}}(w) \leq O\big(\frac{\|\mathcal{W}\|_1(\alpha\tau \log \frac{n|V|d}{\zeta})^{\frac{1}{3}}}{(n\epsilon)^{\frac{1}{3}}}\big). \quad (6)$$

Remark 1. From Theorem 2 we can see that when $|V| = \text{poly}(d)$ and $\tau = O(1)$, the excess population risk will be upper bounded by $\tilde{O}(\frac{1}{(n\epsilon)^{\frac{1}{3}}})$. Compared with the previous results in private heavy-tailed estimation [10, 33, 51], we improve the error bound from $O(d)$ to $O(\log d)$. It is also notable that [51] also used a similar robust estimator as ours. However, there are several differences: First, [51] first performs the robust estimator to each coordinate of the gradients and then add Gaussian noise to the whole vector to ensure DP. Thus, all the errors in [51] depend on $\text{poly}(d)$ and their method cannot be extended to high dimensional space directly. Secondly, [51] sets $s = O(\sqrt{n})$ while our s depends on both n, ϵ and T. We provide a finer analysis on the trade-off between the bias and variance of the robust estimator, and the noise we added in each iteration (see the proof of Theorem 2 for details). Thus, our error is much lower than theirs and our method could be used in [51] and improve their bounds.

Corollary 1. Consider the LASSO problem where $L_D(w) = \mathbb{E}(\langle x, w\rangle - y)^2$ and $\mathcal{W} = \{w \in \mathbb{R}^d : \|w\|_1 \leq 1\}$. We know that the population risk function is $\lambda_{\max}(\mathbb{E}(xx^T))$-smooth, where $\lambda_{\max}(M)$ is the maximal eigenvalue of the matrix M. If we further assume each coordinate of the gradient has bound second moment *i.e.,* for each $w \in \mathcal{W}$ and $j \in [d]$, $\mathbb{E}[(x_j(\langle x, w\rangle - y))^2] \leq O(1)$ (for example x_j and y are $O(1)$-sub-Gaussian). Then the output of Algorithm 1 satisfies the following with probability at least $1-\zeta$:

$$L_{\mathcal{D}}(w^T) - \min_{w\in\mathcal{W}} L_{\mathcal{D}}(w) \leq O\big(\frac{(\lambda_{\max}(\mathbb{E}(xx^T) \log \frac{d}{\zeta} \log n)^{\frac{1}{3}}}{(n\epsilon)^{\frac{1}{3}}}\big). \quad (7)$$

In the previous theorem, we need to assume the loss function is convex. However, we can also show that Algorithm 1 could be used to some specific non-convex loss functions. Below we will study the Robust Regression and provide an upper bound of $\tilde{O}(\frac{1}{(n\epsilon)^{\frac{1}{4}}})$. For $\mathcal{W} = \{w \in \mathbb{R}^d | \|w\|_1 \leq 1\}$, and a non-convex positive loss function ψ, the loss of robust regression is defined as $\ell(w, (x, y)) = \psi(\langle x, w\rangle - y)$. We make the following assumptions on ψ, which includes the biweight loss function [4] [38].

Assumption 2. We assume that

(1) There is a constant $C_\psi \geq 1$, s.t. $\max\{\psi'(s), \psi''(s)\} \leq C_\psi = O(1)$, for for all s.
(2) $\psi'(\cdot)$ is odd with $\psi'(s) > 0$, for $\forall s > 0$; and $h(s) := \mathbb{E}_\xi[\psi'(s+\xi)]$ satisfies $h'(0) > c_\psi$, where $c_\psi = O(1) > 0$.
(3) There is $w^* \in \mathcal{W}$ such that $y = \langle w^*, x\rangle + \xi$, where ξ is symmetric noise with a zero-mean given x. Also we assume that for each coordinate $j \in [d]$, x_j has bounded second order moment, that is $\mathbb{E}x_j^2 \leq O(1)$.

Theorem 3. Under Assumption 2, for any given probability of failure $0 < \zeta < 1$, if we set $\beta = O(1)$, $s = O(\frac{\sqrt{n\epsilon}}{\sqrt{T \log \frac{dT}{\zeta}}})$, $\eta = \frac{1}{\sqrt{T}}$, and $T = \tilde{O}(\sqrt{\frac{n\epsilon}{\log \frac{d}{\zeta}}})$ in Algorithm 1. Then with probability at least $1-\zeta$ (we omit the C_ψ and c_ψ term),

$$L_{\mathcal{D}}(w^T) - \min_{w\in\mathcal{W}} L_{\mathcal{D}}(w) \leq O\big(\frac{\lambda_{\max}(\mathbb{E}(xx^T)) \log^{\frac{1}{4}} \frac{dn}{\zeta}}{(n\epsilon)^{\frac{1}{4}}}.\big). \quad (8)$$

For LASSO, there are enormous differences between our results and the results in [45]. First, [45] needs to assume that each $|x_{ij}| \leq O(1)$ and $|y_i| \leq O(1)$ to guarantee the loss function be ℓ_1-norm Lipschitz, while here we just need a bounded second order moment condition. Secondly, [45] only considers the empirical risk function while here we consider the population risk. It is notable that their method cannot be extended to population risk directly based on their theoretical analysis. Thus, our result of $\tilde{O}(\frac{1}{(n\epsilon)^{\frac{1}{3}}})$ cannot be compared with theirs directly. Recently [2] considers DP-SCO with ℓ_1-norm Lipschitz loss functions and it provides an upper bound of $\tilde{O}(\frac{1}{\sqrt{\epsilon n}})$ and $\tilde{O}(\sqrt{\frac{1}{n}} + \frac{1}{(n\epsilon)^{\frac{2}{3}}})$ for ϵ and (ϵ, δ)-DP model respectively. Compared with this, we can see, due to the heavy-tailed distribution, the upper bound now decreases to $\tilde{O}(\frac{1}{(n\epsilon)^{\frac{1}{3}}})$ for ϵ-DP. Thirdly, the DP Frank Wolfe algorithm given by [45] could guarantee both ϵ and (ϵ, δ)-DP with error upper bounds of $\tilde{O}(\frac{1}{\sqrt{n\epsilon}})$ and $\tilde{O}(\frac{1}{(n\epsilon)^{\frac{2}{3}}})$ respectively.[5] However, our method can only guarantee ϵ-DP and cannot get improved bounds in the (ϵ, δ)-DP model. The mainly reason is that, [45] performs the exponential mechanism on the whole data to achieve $O(\frac{\epsilon}{\sqrt{T \log \frac{1}{\delta}}})$-DP in each iteration, then the whole algorithm will be (ϵ, δ)-DP due to the advanced composition theorem. However here we cannot adopt this technique directly. The major difficulty is that if we use whole dataset in each iteration then w^{t-1} will depend on the whole dataset. And this cause us in the proof to analyze an upper bound of $\sup_{v\in V} \sup_{w\in\mathcal{W}} \langle v, \tilde{g}(w, D) - \mathbb{E}[\nabla\ell(w; z)]\rangle$, which is difficult to analyze due to the complex form of our estimator $\tilde{g}(w, D)$ in step 5. Thus, we need to get avoid of the dependency. Our strategy is splitting the whole dataset into several parts and in each iteration we use the exponential mechanism on one subset. That is why here we only consider the ϵ-DP model. It is an open problem that whether we can get an improved (ϵ, δ)-DP method in general. Below we will show that for LASSO it is possible to improve the upper bound from $\tilde{O}(\frac{1}{(n\epsilon)^{\frac{1}{3}}})$ in Corollary 1 to $\tilde{O}(\frac{1}{(n\epsilon)^{\frac{2}{5}}})$ in (ϵ, δ)-DP model if the data distribution has bounded fourth-order moments.

The algorithm consists of two parts. In the first part, motivated by [25], we shrunk each entry of each sample by a threshold K, which will be determined later. That is, for each $i \in [n]$ and $j \in [d]$, we

[4] For a fixed parameter $c > 0$, the biweight loss is defined as $\psi(s) = \frac{c^2}{6} \cdot \begin{cases} 1 - (1 - (\frac{s}{c})^2)^3, |t| \leq c \\ 1, |t| \geq c. \end{cases}$

[5] We can adopt the idea in [45] and get the result for ϵ-DP

let $\tilde{x}_{i,j} = \text{sign}(x_{i,j})\min\{|x_{i,j}|, K\}$ and $\tilde{y}_i = \text{sign}(y_i)\min\{|y_i|, K\}$. Note that since now each entry is bounded, the loss function will be ℓ_1-norm Lipschitz with $O(K^2)$. Thus, in the second part, we perform the DP-FW in [45] on the shrunken data. See Algorithm 2 for details.

Algorithm 2 Heavy-tailed Private LASSO

Input: n-size dataset $D = \{(x_i, y_i)\}_{i=1}^n$, loss function $\ell(w,(x,y)) = (\langle w, x\rangle - y)^2$, initial parameter w^0, parameters $K, T, \{\eta_t\}$ (will be specified later), privacy parameter ϵ, δ, failure probability ζ. $\mathcal{W}$ is the ℓ_1-norm ball with set of vertices V.
For each $i \in [n]$, we denote a truncated sample $\tilde{x}_i \in \mathbb{R}^d$ where for $j \in [d]$ $\tilde{x}_{i,j} = \text{sign}(x_{i,j})\min\{|x_{i,j}|, K\}$, and $\tilde{y}_i = \text{sign}(y_i)\min\{|y_i|, K\}$. Denote the truncated dataset as $\tilde{D} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^n$.
for $t = 1, \cdots, T$ **do**.
    Denote the score function $u(\tilde{D}, \cdot) : V \mapsto \mathbb{R}$ such that for each $v \in V$ let $u(\tilde{D}, v) = -\langle v, \tilde{g}(w^{t-1}, \tilde{D})\rangle$, where $\tilde{g}(w^{t-1}, \tilde{D}) = \frac{2}{n}\sum_{i=1}^n \tilde{x}_i(\langle \tilde{x}_i, w^{t-1}\rangle - \tilde{y}_i)$. Run the exponential mechanism with the range $R = V$, sensitivity $\Delta = \frac{8\|\mathcal{W}\|_1 K^2}{n}$ and the privacy budget $\frac{\epsilon}{2\sqrt{2T\log\frac{1}{\delta}}}$. Denote the output as $\tilde{w}^{t-1} \in V$.
    Let $w^t = (1-\eta_{t-1})w^{t-1} + \eta_{t-1}\tilde{w}^{t-1}$.
end for
return w^T.

Theorem 4. For any $0 < \epsilon, \delta < 1$, Algorithm 1 is (ϵ, δ)-DP.

Assumption 3. We assume that x and y have bounded forth order moment, *i.e.*, for each $j_1, j_2 \in [d]$, $\mathbb{E}(x_{j_1}x_{j_2})^2 \le M$, and $\mathbb{E}[y^4] \le M$, where $M = O(1)$ is a constant.

Remark 2. We note that Assumption 1 implies $\mathbb{E}(x_j x_k)^2 \le O(\tau)$ for any $j, k \in [d]$. Since in Assumption 1 we can get $\mathbb{E}[(x_j y)^2] \le \tau$ if we take $w = 0$. And we have $\mathbb{E}[(x_j(x_k - y))^2] \le \tau$ when we take $w = e_k$ (the k-th basis vector), thus $\mathbb{E}[(x_j x_k)^2] \le O(\tau)$. From this view, Assumption 3 is weaker than Assumption 1. Moreover, in Assumption 1 we need to assume that the term $\mathbb{E}[x_i^2(\langle w, x\rangle - y)^2]$ is bounded for each $\|w\|_1 \le 1$, which is hard to be verified and is unnatural for the linear model compared with the previous work on linear regression with heavy-tailed data [28, 57].

Theorem 5. Under Assumption 3, for any given probability of failure $0 < \zeta < 1$, if we set $K = \frac{(n\epsilon)^{\frac{1}{4}}}{T^{\frac{1}{8}}}$, $T = \tilde{O}((\frac{\sqrt{n\epsilon}\lambda_{\max}(\mathbb{E}(xx^T))}{\sqrt{\log\frac{1}{\delta}\log\frac{dT}{\zeta}}})^{\frac{4}{5}})$ and $\eta_{t-1} = \frac{2}{t+2}$ in Algorithm 2, then with probability at least $1-\zeta$,

$$L_{\mathcal{D}}(w^T) - \min_{w\in\mathcal{W}} L_{\mathcal{D}}(w) \le O(\frac{\lambda_{\max}^{\frac{1}{5}}(\mathbb{E}(xx^T))(\sqrt{\log\frac{1}{\delta}}\log\frac{dn}{\zeta})^{\frac{4}{5}}}{(n\epsilon)^{\frac{2}{5}}}). \tag{9}$$

Truncating or shrunking the data to let them has bounded norm (or bounded sensitivity) is a commonly used technique in previous study on DP machine learning such as [4, 13, 14]. However, all of these methods need to assume the data distribution is sub-Gaussian so that truncation may not lose too much information about the original record. Here we generalized to a heavy-tailed case, which may could be used to other problems. Moreover, the thresholds in the truncation step for sub-Gaussian and heavy-tailed cases are also quite different. In the sub-Gaussian case, the threshold always depends on the sub-Gaussian parameter and $\log n, \log d$, while in Algorithm 2 we set the threshold as a function of n, ϵ and T.

5 HEAVY-TAILED DP-SCO FOR SPARSE LEARNING

5.1 Private Heavy-tailed Sparse Linear Regression

In the previous section, we studied DP-SCO over polytope constraint. However, in the high dimensional statistics we always assume the underlying parameter has additional structure of sparsity. Directly solving DP-SCO over ℓ_1-norm ball constraint may not provide efficient estimation to the sparse parameters. In this section, we will focus on sparse learning with heavy-tailed data. Specifically, we will consider two canonical models, one is the sparse linear model, the other one is the DP-SCO over sparsity constraint, which includes sparse regularized logistic regression and spase mean estimation. First we consider the sparse linear regression, where for each pair (x, y) we have a linear model,

$$y = \langle w^*, x\rangle + \iota,$$

here ι is some randomized noise and $\|w^*\|_2 \le C$ (for simplicity we assume $C = 1$) and w^* is s^*-sparse.

Similar to the previous section, here we assume Assumption 3 holds. Instead of using DP variants of the Frank-Wolfe method, here we will adopt a private variant of the iterative hard thresholding (IHT) method. Specifically, first we will shrunk the original heavy-tailed data, which is similar to Algorithm 2. After that we will perform the DP-IHT procedure. That is, in each iteration, we fist calculate the gradient on the shrunken data, and update our vector via the gradient descent. Next, we perform a DP-thresholding step, provided by [13] (Algorithm 4). That is, we will privately select the indices with largest s magnitude of the vector, keep the entries of vectors among these indices and let the remain entries be 0. See Algorithm 3 and 4 for details.

Theorem 6. For any $0 < \epsilon, \delta < 1$, Algorithm 3 is (ϵ, δ)-DP.

Theorem 7. Under Assumption 3, if $\|w^*\|_2 \le \frac{1}{2}$, the initial vector w^1 satisfies $\|w^1 - w^*\| \le O(\frac{\gamma}{\mu})$ and n is sufficiently large such that $n \ge \tilde{O}(\frac{s^2 M\log^2\frac{d}{\zeta}\log\frac{1}{\delta}}{\gamma\mu^4\epsilon})$. Then if we set $T = \tilde{O}(\frac{\gamma}{\mu}\log n)$, $K = \frac{(n\epsilon)^{\frac{1}{4}}}{(sT)^{\frac{1}{4}}}$, $s \ge 72(\frac{\gamma}{\mu})^2 s^*$ and $\eta = \frac{2}{3\gamma}$ in Algorithm 3, then with probability at least $1-\zeta$

$$L_{\mathcal{D}}(w^{T+1}) - L_{\mathcal{D}}(w^*) \le O(\frac{M\gamma^4 s^{*2}\log n\log^2\frac{d}{\zeta}\log\frac{1}{\delta}}{\mu^7 n\epsilon}),$$

where $\gamma = \lambda_{\max}(\mathbb{E}(xx^T))$ and $\mu = \lambda_{\min}(\mathbb{E}(xx^T))$ and the Big-O notation omits other log terms.

Remark 3. In Theorem 7, we need to assume that $\|w^*\|_2 \le \frac{1}{2}$ and the initial vector be close to w^*. These two conditions guarantee $\|w^{t+0.75}\|_2 \le 1$ in each iteration, which simplify our theoretical

Algorithm 3 Heavy-tailed Private Sparse Linear Regression

Input: n-size dataset $D = \{(x_i, y_i)\}_{i=1}^n$, loss function $\ell(w,(x,y)) = (\langle w,x\rangle - y)^2$, initial vector w^1 satisfies $\|w^1\|_2 \le 1$ and is s-sparse, parameters K, T, η_0, s (will be specified later), privacy parameter ϵ, δ, failure probability ζ. $\mathcal{W}$ is the unit ℓ_2-norm ball.
For each $i \in [n]$, we denote a truncated sample $\tilde{x}_i \in \mathbb{R}^d$ where for $j \in [d]$ $\tilde{x}_{i,j} = \text{sign}(x_{i,j})\min\{|x_{i,j}|, K\}$, and $\tilde{y}_i = \text{sign}(y_i)\min\{|y_i|, K\}$. Denote the truncated dataset as $\tilde{D} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^n$.
Split the data $\tilde{D}$ into T parts $\{\tilde{D}_t\}_{t=1}^T$, each with $m = \frac{n}{T}$ samples.
for $t = 1, \cdots, T$ **do**.
  Denote $w^{t+0.5} = w^t - \frac{\eta_0}{m}\sum_{x\in\tilde{D}_t}\tilde{x}(\langle\tilde{x}, w^t\rangle - \tilde{y})$
  Let $w^{t+0.75} = \text{Peeling}(w^{t+0.5}, D_t, s, \epsilon, \delta, \frac{2K^2\eta_0(\sqrt{s}+1)}{m})$.
  Let $w^{t+1} = \Pi_{\mathcal{W}}(w^{t+0.75})$
end for
return w^{T+1}.

Algorithm 4 Peeling [13]

Input: Vector $v = v(D) \in \mathbb{R}^d$ which depends on the data D, sparsity s, privacy parameter ϵ, δ, and noise scale λ.
Initialize $S = \emptyset$.
for $i = 1 \cdots s$ **do**
  Generate $w_i \in \mathbb{R}^d$ with $w_{i,1}, \cdots, w_{i,d} \sim \text{Lap}(\frac{2\lambda\sqrt{3s\log\frac{1}{\delta}}}{\epsilon})$.
  Append $j^* = \arg\max_{j\in[d]\setminus S}|v_j| + w_{i,j}$ to S.
end for
Generate $\tilde{w} \in \mathbb{R}^d$ with $\tilde{w}_1, \cdots, \tilde{w}_d \sim \text{Lap}(\frac{2\lambda\sqrt{3s\log\frac{1}{\delta}}}{\epsilon})$.
return $v_S + \tilde{w}_S$.

analysis. For sub-Gaussian data, with some other additional assumptions, [13, 54] showed that the optimal rate is $\tilde{O}(\frac{s^*\log d}{n} + \frac{(s^*\log d)^2}{(n\epsilon)^2})$. Thus, due to the data irregularity, the error now increases to $\tilde{O}(\frac{s^2\log^2 d}{n\epsilon})$. Moreover, we can see although both Algorithm 3 and 2 shrunk the data in the first step, the threshold value K are quite different, where $K = \frac{(n\epsilon)^{\frac{1}{4}}}{T^{\frac{1}{8}}}$ in LASSO and $K = \frac{(n\epsilon)^{\frac{1}{4}}}{(sT)^{\frac{1}{4}}}$ in the sparse linear model. This is due to different trade-offs between the bias, variance in the estimation error and the noises we added.

5.2 Extending to Sparse Learning

In this section, we extend our previous ideas and methods to the problem of DP-SCO over sparsity constraints. That is, $\mathcal{W}$ is defined as $\mathcal{W} = \{w : \|w\|_0 \le s^*\}$. We note that such a formulation encapsulates several important problems such as the ℓ_0-constrained linear/logistic regression [3]. DP-SCO over sparsity constraints has been studied previously [52–55]. However, all of the previous methods need either the loss function is Lipschitz, or the data follows some sub-Gaussian distribution [13, 14]. In the following we extend to the heavy-tailed case. We first introduce some assumptions to the loss functions, which are commonly used in previous research on sparse learning.

Definition 7 (Restricted Strong Convexity, RSC)**.** A differentiable function $f(x)$ is restricted ρ_r-strongly convex with parameter r if there exists a constant $\mu_r > 0$ such that for any x, x' with $\|x - x'\|_0 \le r$, we have $f(x) - f(x') - \langle\nabla f(x'), x - x'\rangle \ge \frac{\mu_r}{2}\|x - x'\|_2^2$.

Definition 8 (Restricted Strong Smoothness, RSS)**.** A differentiable function $f(x)$ is restricted μ_s-strong smooth with parameter r if there exists a constant $\gamma_r > 0$ such that for any x, x' with $\|x - x'\|_0 \le r$, we have $f(x) - f(x') - \langle\nabla f(x'), x - x'\rangle \le \frac{\gamma_r}{2}\|x - x'\|_2^2$.

Assumption 4. We assume that the objective function $L_{\mathcal{D}}(\cdot)$ is μ_r-RSC and $\ell(w,z)$ is γ_r-RSS with parameter $r = 2s + s^*$, where $s = O((\frac{\gamma_r}{\mu_r})^2 s^*)$. We also assume for any $w \in \mathcal{W}'$ and each coordinate $j \in [d]$, we have $\mathbb{E}[(\nabla_j\ell(w,x))^2] \le \tau = O(1)$, where τ is some known constant and $\mathcal{W}' = \{w | \|w\|_0 \le s\}$.

Many problems satisfy Assumption 4, e.g., mean estimation and ℓ_2-norm regularized generalized linear loss where $L_{\mathcal{D}}(w) = \mathbb{E}[\ell(y\langle w,x\rangle)] + \frac{\lambda}{2}\|w\|_2^2$. If $|\ell'(\cdot)| \le O(1)$, $|\ell''(\cdot)| \le O(1)$ (such as the logistic loss) and x_j has bounded second-order moment, then we can see it satisfies Assumption 4.

Since now the loss function becomes non-linear, the approach of shrunking the data in Algorithm 3 may introduce tremendous error. However, since in Assumption 4 we have stronger assumptions on the loss function, we may use the private estimator in Algorithm 1. Thus, our idea is that, we first perform the robust one-dimensional mean estimator in (2)-(5) to each coordinate of the gradient, then we use the private selection algorithm to select top s indices, which is the same as in Algorithm 3. Note that [51] also provides a similar method in the low dimensional space. However, the main difference is that here we do not add noise directly to the vector $\tilde{g}(w^{t-1}, D_t)$. Instead, we first privately select the top s indices and then add noises to the corresponding sub-vector. See Algorithm 5 for details.

Algorithm 5 Heavy-tailed Private Sparse Optimization

Input: n-size dataset $D = \{(x_i, y_i)\}_{i=1}^n$, initial parameter w^1 is s-sparse, parameters s, β, k, T, η (will be specified later), privacy parameter ϵ, δ, failure probability ζ.
Split the data D into T parts $\{D_t\}_{t=1}^T$, each with $m = \frac{n}{T}$ samples.
for $t = 1, \cdots, T$ **do**.
  For each $j \in [d]$, calculate the robust gradient by (2)-(5), that is

$$g_j^{t-1}(w^{t-1}, D_t) = \frac{1}{m}\sum_{x\in D_t}\left(\nabla_j\ell(w^{t-1},x)\big(1 - \frac{\nabla_j^2\ell(w^{t-1},x)}{2k^2\beta}\big) - \frac{\nabla_j^3\ell(w^{t-1},x)}{6k^2}\right) + \frac{k}{m}\sum_{x\in D_t}\hat{C}\left(\frac{\nabla_j\ell(w^{t-1},x)}{k}, \frac{|\nabla_j\ell(w^{t-1},x)|}{k\sqrt{\beta}}\right).$$

  Let vector $\tilde{g}(w^{t-1}, D_t) \in \mathbb{R}^d$ as $\tilde{g}(w^{t-1}, D_t) = (g_1^{t-1}(w^{t-1}, D_t), g_2^{t-1}(w^{t-1}, D_t), \cdots, g_d^{t-1}(w^{t-1}, D_t))$.
  Denote $w^{t+0.5} = w^t - \eta\tilde{g}(w^{t-1}, D_t)$
  Let $w^{t+1} = \text{Peeling}(w^{t+0.5}, D_t, s, \epsilon, \delta, \frac{4k\sqrt{2}\eta}{m})$.
end for
return w^{T+1}.

Theorem 8. For any $0 < \epsilon, \delta < 1$, Algorithm 5 is (ϵ, δ)-DP. Moreover, under Assumption 4, if we set $T = \tilde{O}(\frac{\gamma_r}{\mu_r}\log n)$, $s = O((\frac{\gamma_r}{\mu_r})^2 s^*)$, $\beta = O(1)$, $\eta = \frac{2}{3\gamma_r}$ and $k = \tilde{O}(\sqrt{n\epsilon\tau})$, then with probability at least $1-\zeta$,

$$L_{\mathcal{D}}(w^{T+1}) - L_{\mathcal{D}}(w^*) \leq O(\frac{\tau\gamma_r^4 s^{*\frac{3}{2}} \log n \log\frac{d}{\zeta}\sqrt{\log\frac{1}{\delta}}}{\mu_r^5 n\epsilon}),$$

where the Big-O notation omits other log terms.

Remark 4. Compared with Theorem 7, we can see here we do not need the assumptions on $\|w^*\|_2$ and the initial vector. This is due to that we have stronger assumptions on the loss function. Compared with the bound $\tilde{O}(\frac{s^{*2}}{n\epsilon})$ in Theorem 7, it seems like here our bound is lower. However, we note that they are incomparable due to different assumptions. For example, there is a τ in the bound of Theorem 8, which could also depend on the sparsity s^* [49]. [52] also studies DP-SCO over sparsity constraint, it provides an upper bound of $\tilde{O}(\frac{s^*}{n^2\epsilon^2})$ under the assumption that the loss function is Lipschitz. Moreover, for high dimensional sparse mean estimation and Generalized Linear Model (GLM) with the Lipschitz loss and sub-Gaussian data, [13, 14] provided optimal rates of $\tilde{O}(\frac{s^*\log d}{n} + \frac{(s^*\log d)^2}{(n\epsilon)^2})$. We can see that compared with these results, the error bound now becomes to $\tilde{O}(\frac{\tau s^{*\frac{3}{2}}}{n\epsilon})$ due to data irregularity. Moreover, we can see that in the regular data case, the optimal rates of linear regression and GLM are the same, while in the heavy-tailed data case, there is a gap of $\tilde{O}(\sqrt{s^*})$ in the upper bounds. We conjecture this gap is necessary and will leave it as future research.

In the following we will focus on the lower bound of the loss functions in Theorem 8. Since our lower bound will be in the form of private minimax risk, we first introduce the classical statistical minimax risk before discussing its (ϵ, δ)-private version. More details can be found in [4].

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

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

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

In the (ϵ, δ)-DP model, the estimator $\hat{\theta}$ is obtained via some (ϵ, δ)-DP mechanism Q. Thus, we can also define the (ϵ, δ)-private minimax risk:

$$\mathcal{M}_n(\theta(\mathcal{P}), Q, \Phi\circ\rho) := \inf_{Q\in\mathcal{Q}} \inf_{\hat{\theta}} \sup_{p\in\mathcal{P}} \mathbb{E}_{p,Q}[\Phi(\rho(\hat{\theta}(D), \theta(p))],$$

where $\mathcal{Q}$ is the set of all the (ϵ, δ)-DP mechanisms.

To proof the lower bound, we consider the sparse mean estimation problem, *i.e.*, $L_{\mathcal{D}}(w) = \mathbb{E}_{x\sim\mathcal{D}}[\|x - w\|_2^2]$, where the mean of x, $\mu(\mathcal{D})$, is s^*-sparse. Thus, we can see that the population risk function satisfies Assumption 4 if we assume $\mathbb{E}x_j^2 \leq \tau$ for each $j \in [d]$. Moreover, we have $\min_{w\in\mathcal{W}} L_{\mathcal{D}}(w) = 0$ which indicates that the excess population risk of w is equal to $\mathbb{E}\|w - \mu(\mathcal{D})\|_2^2$. That is, the lower bound of Theorem 8 reduced to the sparse mean estimation problem. Therefore, it is sufficient for us to consider the (ϵ, δ)-private minimax rate for the sparse mean estimation problem with $\mathbb{E}x_j^2 \leq \tau$ for each $j \in [d]$.

In the non-private case, a standard approach to prove the lower bound of the minimax risk is reducing the original problem to a testing problem. Specifically, our goal is to identify a parameter $\theta \in \Theta$ from a finite collection of well-separated points. Given an index set $\mathcal{V}$ with finite cardinality, the indexed family of distributions $\{P_v, v \in \mathcal{V}\} \subset \mathcal{P}$ is said to be a 2γ-packing if $\rho(\theta(P_v), \theta(P_{v'})) \geq 2\gamma$ for all $v \neq v' \in \mathcal{V}$. In the standard hypothesis testing problem, nature chooses $V \in \mathcal{V}$ uniformly at random, then draws samples $X_1, \cdots X_n$ i.i.d. from the distribution P_V. The problem is to identify the index V. It has been shown that given a 2γ-packing $\{P_v, v \in \mathcal{V}\} \subset \mathcal{P}$,

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi\circ\rho) \geq \Phi(\gamma)\inf_{\psi}\mathbb{P}(\psi(D) \neq V),$$

where $\mathbb{P}$ denotes the probability under the joint distribution of both V and the samples D.

Similar to the non-private case, for the private minimax risk we have

$$\mathcal{M}_n(\theta(\mathcal{P}), Q, \Phi\circ\rho) \geq \inf_{Q\in\mathcal{Q}}\Phi(\gamma)\inf_{\psi}\mathbb{P}_Q(\psi(\hat{\theta}(D)) \neq V),$$

where $\hat{\theta}(D)$ is the private estimator via some (ϵ, δ)-DP algorithm Q, where $\mathbb{P}_Q$ denotes the probability under the joint distribution of both V, the samples D and $\hat{\theta}(D)$.

In the following we will consider a special indexed family of distributions $\{P_v\}_{v\in\mathcal{V}} \subset \mathcal{P}$, which will be used in our main proof. We assume there exists a distribution P_0 such that for some fixed $p \in [0, 1]$ we have $(1-p)P_0 + pP_v \in \mathcal{P}$ for all $v \in \mathcal{V}$. For simplicity for each $v \in \mathcal{V}$ we define the following parameter

$$\theta_v := \theta((1-p)P_0 + pP_v).$$

We then define the separation of the set $\{\theta\}_v$ by

$$\rho^*(\mathcal{V}) := \min\{\rho(\theta_v, \theta_{v'})|v, v' \in \mathcal{V}, v \neq v'\}.$$

We have the following lower bound of (ϵ, δ)-private minimax risk based on the the family of distributions $\{(1-p)P_0 + pP_v\}_{v\in\mathcal{V}}$.

Lemma 3 (Theorem 3 in [4])**.** Fix $p \in [0, 1]$ and define $P_{\theta_v} = (1-p)P_0 + pP_v \in \mathcal{P}$. Let $\hat{\theta}$ be an (ϵ, δ)-DP estimator. Then

$$\begin{aligned}
&\mathcal{M}_n(\theta(\mathcal{P}), Q, \Phi\circ\rho)\\
&\geq \Phi(\rho^*(\mathcal{V}))\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} P_{\theta_v}(\rho(\hat{\theta}, \theta_v) \geq \rho^*(\mathcal{V}))\\
&\geq \Phi(\rho^*(\mathcal{V}))\frac{(|\mathcal{V}|-1)(\frac{1}{2}e^{-\epsilon\lceil np\rceil} - \delta\frac{1-e^{-\epsilon\lceil np\rceil}}{1-e^{-\epsilon}})}{1+(|\mathcal{V}|-1)e^{-\epsilon\lceil np\rceil}}. \qquad (10)
\end{aligned}$$

By Lemma 3 and a set of hard distributions, we have the following result.

Theorem 9. Consider the class of distributions $\mathcal{P}_d^{s^*}(\tau)$ as distributions P in the d dimensional space satisfying that $\mathbb{E}_{X\sim P}X_j^2 \leq \tau$ for all $x_j \in [d]$ and the mean of P, $\mu(P)$ is s^*-sparse. Then the (ϵ, δ)-private minimax risk with $\Phi(x) = x^2$ and $\rho(x_1, x_2) = \|x_1 - x_2\|_2$ satisfies that

$$\mathcal{M}_n(\theta(\mathcal{P}_d^{s^*}(\tau)), Q, \Phi \circ \rho) \geq \Omega(\frac{\tau \min\{s^* \log d, \log \frac{1}{\delta}\}}{n\epsilon}). \quad (11)$$

Thus, for DP-SCO problem under Assumption 4. The information-theoretical lower bound of the expected population risk in the (ϵ, δ)-DP model is $\Omega(\frac{\tau \min\{s \log d, \log \frac{1}{\delta}\}}{n\epsilon})$.

Compared with the upper bound in Theorem 8 and the lower bound in Theorem 9, we can see there is still a gap of $\tilde{O}(\sqrt{s^*})$. It is an open problem that whether we can further improve the upper bound. For the low dimensional case, [4, 32] showed that the optimal rate of the mean estimation is $O(\frac{\tau d}{n\epsilon})$ in both ϵ and (ϵ, δ)-DP models under the assumption that the gradient of loss has bounded second order moment. Compared with this here we extend to the high dimensional sparse case.

6 CONCLUSION

In this paper, we studied the problem of Differentially Private Stochastic Convex Optimization (DP-SCO) in the high dimensional (sparse) setting, where the sample size n is far less than the dimension of the space d and the underlying data distribution may be heavy-tailed. We first considered the problem of DP-SCO where the constraint set is some polytope. We showed that if the gradient of loss function has bounded second order moment, then it is possible to achieve an excess population risk of $\tilde{O}(\frac{\log d}{(n\epsilon)^{\frac{1}{3}}})$ (with high probability) in the ϵ-DP model, if we omit other terms. Moreover, for the LASSO problem, we showed that it is possible to achieve an error of $\tilde{O}(\frac{\log d}{(n\epsilon)^{\frac{2}{5}}})$ in the (ϵ, δ)-DP model. Next we studied DP-SCO for sparse learning with heavy-tailed data. We first investigated the sparse linear model and proposed a method whose output could achieve an estimation error of $\tilde{O}(\frac{s^{*2}\log^2 d}{n\epsilon})$, where s^* is the sparsity of the underlying parameter. Then we studied a more general problem over the sparsity (*i.e.,* ℓ_0-norm) constraint, and show that it is possible to achieve an error of $\tilde{O}(\frac{s^{*\frac{3}{2}}\log d}{n\epsilon})$ if the loss function is smooth and strongly convex. Finally, we showed a lower bound of $\tilde{O}(\frac{s^* \log d}{n\epsilon})$ for the high dimensional heavy-tailed sparse mean estimation in the (ϵ, δ)-DP model.

Besides the open problems we mentioned in the previous sections, there are still many other future work. First, in this paper, we studied the problem under various settings and assumptions and provided some bounds of the excess population risk. While we showed a lower bound for the high dimensional heavy-tailed sparse mean problem, we still do not know the lower bounds of other problems. Previous results on the lower bounds need to assume the data is regular, thus we need new techniques or hard instances to get those lower bounds in the heavy-tailed setting. Secondly, in the heavy-tailed and low dimensional case, we know that the bounds of excess population risk may be different in the high probability form and expectation form [32, 51]. Thus, our question is, in the high dimensional case, if we relax to the expectation form, can we further improve these upper bounds? Thirdly, we need to assume the gradient of the loss has bounded second order moment throughout the paper. However, sometimes this will not be held and the data may only has the $1+v$-th moment with some $v \in (0, 1)$ [47]. Due to this weaker assumption, all the previous methods are failed. Thus, how to extend to this case in both low dimensional and high dimensional cases?

ACKNOWLEDGMENTS

Di Wang and Lijie Hu were support in part by the baseline funding BAS/1/1689-01-01 and funding from the AI Initiative REI/1/4811-10-01 of King Abdullah University of Science and Technology (KAUST).

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 308–318.

[2] Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. 2021. Private Stochastic Convex Optimization: Optimal Rates in ℓ_1 Geometry. *arXiv preprint arXiv:2103.01516* (2021).

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

[4] Rina Foygel Barber and John C Duchi. 2014. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv preprint arXiv:1412.4451* (2014).

[5] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. 2020. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems* 33 (2020).

[6] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradee Thakurta. 2019. Private Stochastic Convex Optimization with Optimal Rates. In *NeurIPS*.

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

[8] Atanu Biswas, Sujay Datta, Jason P Fine, and Mark R Segal. 2007. *Statistical advances in the biomedical science*. Wiley Online Library.

[9] Christian Brownlees, Emilien Joly, Gábor Lugosi, et al. 2015. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics* 43, 6 (2015), 2507–2536.

[10] Victor-Emmanuel Brunel and Marco Avella-Medina. 2020. Propose, Test, Release: Differentially private estimation with high probability. *arXiv preprint arXiv:2002.08774* (2020).

[11] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. 2013. Bandits with heavy tail. *IEEE Transactions on Information Theory* 59, 11 (2013), 7711–7717.

[12] Mark Bun and Thomas Steinke. 2019. Average-Case Averages: Private Algorithms for Smooth Sensitivity and Mean Estimation. *arXiv preprint arXiv:1906.02830* (2019).

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

[14] T Tony Cai, Yichen Wang, and Linjun Zhang. 2020. The Cost of Privacy in Generalized Linear Models: Algorithms and Minimax Lower Bounds. *arXiv preprint arXiv:2011.03900* (2020).

[15] Olivier Catoni. 2012. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, Vol. 48. 1148–1185.

[16] Olivier Catoni and Ilaria Giulini. 2017. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747* (2017).

[17] Kamalika Chaudhuri and Claire Monteleoni. 2009. Privacy-preserving logistic regression. In *Advances in neural information processing systems*. 289–296.

[18] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. 2011. Differentially private empirical risk minimization. *Journal of Machine Learning Research* 12, Mar (2011), 1069–1109.

[19] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. 2017. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*. 3571–3580.

[20] John C Duchi, Michael I Jordan, and Martin J Wainwright. 2013. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE, 429–438.

[21] John C Duchi, Michael I Jordan, and Martin J Wainwright. 2018. Minimax optimal procedures for locally private estimation. *J. Amer. Statist. Assoc.* 113, 521 (2018), 182–201.

[22] Cynthia Dwork and Jing Lei. 2009. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*. ACM, 371–380.

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

[24] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9, 3-4 (2014), 211–407.

[25] Jianqing Fan, Weichen Wang, and Ziwei Zhu. 2016. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *arXiv preprint arXiv:1603.08315* (2016).

[26] Vitaly Feldman, Tomer Koren, and Kunal Talwar. 2020. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. 439–449.

[27] Matthew Holland and Kazushi Ikeda. 2019. Better generalization with less data using robust gradient descent. In *International Conference on Machine Learning*. 2761–2770.

[28] Daniel Hsu and Sivan Sabato. 2016. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research* 17, 1 (2016), 543–582.

[29] Lijie Hu, Shuo Ni, Hanshen Xiao, and Di Wang. 2021. High Dimensional Differentially Private Stochastic Optimization with Heavy-tailed Data. *arXiv preprint arXiv:2107.11136* (2021).

[30] Marat Ibragimov, Rustam Ibragimov, and Johan Walden. 2015. *Heavy-tailed distributions and robustness in economics and finance*. Vol. 214. Springer.

[31] Roger Iyengar, Joseph P Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. 2019. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 299–316.

[32] Gautam Kamath, Xingtu Liu, and Huanyu Zhang. 2021. Improved Rates for Differentially Private Stochastic Convex Optimization with Heavy-Tailed Data. *arXiv preprint arXiv:2106.01336* (2021).

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

[34] Shiva Prasad Kasiviswanathan and Hongxia Jin. 2016. Efficient private empirical risk minimization for high-dimensional learning. In *International Conference on Machine Learning*. 488–497.

[35] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. 2012. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*. 25–1.

[36] Guillaume Lecué, Matthieu Lerasle, and Timothée Mathieu. 2018. Robust classification via MOM minimization. *arXiv preprint arXiv:1808.03106* (2018).

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

[38] Po-Ling Loh and Martin J Wainwright. 2013. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*. 476–484.

[39] Gábor Lugosi and Shahar Mendelson. 2019. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society* (2019).

[40] Stanislav Minsker et al. 2015. Geometric median and robust estimation in Banach spaces. *Bernoulli* 21, 4 (2015), 2308–2335.

[41] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2007. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. ACM, 75–84.

[42] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. 2018. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485* (2018).

[43] Adam Smith, Abhradeep Thakurta, and Jalaj Upadhyay. 2017. Is interaction necessary for distributed private learning?. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 58–77.

[44] Shuang Song, Om Thakkar, and Abhradeep Thakurta. 2020. Characterizing private clipped gradient descent on convex generalized linear problems. *arXiv preprint arXiv:2006.06783* (2020).

[45] Kunal Talwar, Abhradeep Thakurta, and Li Zhang. 2015. Nearly-optimal private LASSO. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*. 3025–3033.

[46] Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and XiaoFeng Wang. 2017. Privacy Loss in Apple's Implementation of Differential Privacy on MacOS 10.12. *CoRR* abs/1709.02753 (2017). arXiv:1709.02753

[47] Youming Tao, Yulian Wu, Peng Zhao, and Di Wang. 2021. Optimal Rates of (Locally) Differentially Private Heavy-tailed Multi-Armed Bandits. *arXiv preprint arXiv:2106.02575* (2021).

[48] Vladimir Vapnik. 2013. *The nature of statistical learning theory*. Springer science & business media.

[49] Di Wang, Jiahao Ding, Zejun Xie, Miao Pan, and Jinhui Xu. 2020. Differentially Private (Gradient) Expectation Maximization Algorithm with Statistical Guarantees. *CoRR* abs/2010.13520 (2020).

[50] Di Wang, Marco Gaboardi, Adam Smith, and Jinhui Xu. 2020. Empirical Risk Minimization in the Non-interactive Local Model of Differential Privacy. *Journal of Machine Learning Research* 21, 200 (2020), 1–39.

[51] Di Wang, Hanshen Xiao, Srini Devadas, and Jinhui Xu. 2020. On Differentially Private Stochastic Convex Optimization with Heavy-tailed Data. *arXiv preprint arXiv:2010.11082* (2020).

[52] Di Wang and Jinhui Xu. 2019. On Sparse Linear Regression in the Local Differential Privacy Model. In *ICML (Proceedings of Machine Learning Research, Vol. 97)*. PMLR, 6628–6637.

[53] Di Wang and Jinhui Xu. 2021. On Sparse Linear Regression in the Local Differential Privacy Model. *IEEE Trans. Inf. Theory* 67, 2 (2021), 1182–1200.

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

[55] Lingxiao Wang and Quanquan Gu. 2020. A Knowledge Transfer Framework for Differentially Private Sparse Learning.. In *AAAI*. 6235–6242.

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

[57] Lijun Zhang and Zhi-Hua Zhou. 2018. ℓ_1-regression with Heavy-tailed Distributions. In *Advances in Neural Information Processing Systems*. 1076–1086.

[58] Yingxue Zhou, Zhiwei Steven Wu, and Arindam Banerjee. 2020. Bypassing the ambient dimension: Private sgd with gradient subspace identification. *arXiv preprint arXiv:2007.03813* (2020).