On Differentially Private Stochastic Convex Optimization
with Heavy-tailed Data

Di Wang [* 1 2] **Hanshen Xiao** [* 3] **Srini Devadas** [3] **Jinhui Xu** [1]

Abstract

In this paper, we consider the problem of designing Differentially Private (DP) algorithms for Stochastic Convex Optimization (SCO) on heavy-tailed data. The irregularity of such data violates some key assumptions used in almost all existing DP-SCO and DP-ERM methods, resulting in failure to provide the DP guarantees. To better understand this type of challenges, we provide in this paper a comprehensive study of DP-SCO under various settings. First, we consider the case where the loss function is strongly convex and smooth. For this case, we propose a method based on the sample-and-aggregate framework, which has an excess population risk of $\tilde{O}(\frac{d^3}{n\epsilon^4})$ (after omitting other factors), where n is the sample size and d is the dimensionality of the data. Then, we show that with some additional assumptions on the loss functions, it is possible to reduce the *expected* excess population risk to $\tilde{O}(\frac{d^2}{n\epsilon^2})$. To lift these additional conditions, we also provide a gradient smoothing and trimming based scheme to achieve excess population risks of $\tilde{O}(\frac{d^2}{n\epsilon^2})$ and $\tilde{O}(\frac{d^{\frac{2}{3}}}{(n\epsilon^2)^{\frac{1}{3}}})$ for strongly convex and general convex loss functions, respectively, *with high probability*. Experiments suggest that our algorithms can effectively deal with the challenges caused by data irregularity.

1. Introduction

Stochastic Convex Optimization (SCO) (Vapnik, 2013) and its empirical form, Empirical Risk Minimization (ERM), are the most fundamental problems in supervised learning and statistics. They find numerous applications in many areas such as medicine, finance, genomics and social science. One often encountered challenge in such models is how to handle sensitive data, such as those in biomedical datasets. As a commonly-accepted approach for preserving privacy, differential privacy (Dwork et al., 2006) provides provable protection against identification and is resilient to arbitrary auxiliary information that might be available to attackers. Methods to guarantee differential privacy have been widely studied, and recently adopted in industry (Tang et al., 2017; Ding et al., 2017).

Differentially Private Stochastic Convex Optimization and Empirical Risk Minimization (*i.e.,* DP-SCO and DP-ERM) have been extensively studied in the past decade, starting from (Chaudhuri & Monteleoni, 2009; Chaudhuri et al., 2011). Later on, a long list of works have attacked the problems from different perspectives: (Bassily et al., 2014; Wang et al., 2017; 2019a; Wu et al., 2017; Bassily et al., 2019) studied the problems in the low dimensional case and the central model, (Kasiviswanathan & Jin, 2016; Kifer et al., 2012; Talwar et al., 2015) considered the problems in the high dimensional sparse case and the central model, (Smith et al., 2017; Wang et al., 2018; 2019b; Duchi et al., 2013) focused on the problems in the local model.

It is worth noting that all previous results need to assume that either the loss function is $O(1)$-Lipschitz or each data sample has bounded ℓ_2 or ℓ_∞ norm. This is particularly true for those output perturbation based (Chaudhuri et al., 2011) and objective or gradient perturbation based (Bassily et al., 2014) DP methods. However, such assumptions may not always hold when dealing with real-world datasets, especially those from biomedicine and finance, implying that existing algorithms may fail. The main reason is that in such applications, the datasets are often unbounded or even heavy-tailed (Woolson & Clarke, 2011; Biswas et al., 2007; Ibragimov et al., 2015). As pointed out by Mandelbrot and Fama in their influential finance papers (Mandelbrot, 1997; Fama, 1963), asset prices in the early 1960s exhibit some power-law behavior. The heavy-tailed data could lead to unbounded gradient and thus violate the Lipschitz condition. For example, consider the linear squared

---
[*]Equal contribution [1]Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY [2]King Abdullah University of Science and Technology, Thuwal, Saudi Arabia [3]CSAIL, MIT, Cambridge, MA. Correspondence to: Di Wang <dwang45@buffalo.edu>.

loss $\ell(w, x, y) = (w^T x - y)^2$. When x is heavy-tailed, the gradient of $\ell(w, x, y)$ becomes unbounded.

With the above understanding, our questions now are: **What is the behavior of DP-SCO on heavy-tailed data and is there any effective method for the problem?**

To answer these questions, we will conduct, in this paper, a comprehensive study of the DP-SCO problem. Our contributions can be summarized as follows.

1. We first consider the case where the loss function is strongly convex and smooth. For this case, we propose an (ϵ, δ)-DP method based on the sample-and-aggregate framework by (Nissim et al., 2007) and show that under some assumptions, with high probability, the excess population risk of the output is $\tilde{O}(\frac{d^3}{n\epsilon^4} L_{\mathcal{D}}(w^*))$, where n is the sample size, d is the dimensionality and $L_{\mathcal{D}}(w^*)$ is the minimal value of the population risk.

2. Then, we study the case with the additional assumptions: each coordinate of the gradient of the loss function is sub-exponential and Lipschitz. For this case, we introduce an (ϵ, δ)-DP algorithm based on the gradient descent method and a recent algorithm on private 1-dimensional mean estimation (Bun & Steinke, 2019) (*i.e.*, Algorithm 3). We show that the expected excess population risk for this case can be improved to $\tilde{O}(\frac{d^2 \log \frac{1}{\delta}}{n\epsilon^2})$.

3. We also consider the general case, where the loss function does not need the above additional assumptions and can be general convex, instead of strongly convex. For this case, we present a gradient descent method based on the strategy of trimming the unbounded gradient (Algorithm 4). We show that if each coordinate of the gradient of the loss function has bounded second-order moment, then with high probability, the output of our algorithm achieves excess population risks of $\tilde{O}(\frac{d^2 \log \frac{1}{\delta}}{n\epsilon^2})$ and $\tilde{O}(\frac{\log \frac{1}{\delta} d^{\frac{2}{3}}}{(n\epsilon^2)^{\frac{1}{3}}})$ for strongly convex and general convex loss functions, respectively. It is notable that compared with Algorithm 4, Algorithm 3 uses stronger assumptions and yields weaker results.

4. Finally, we test our proposed aglorithms on both synthetic and real-world datasets. Experimental results are consistent with our theoretical claims and reveal the effectiveness of our algorithms in handling heavy-tailed datasets.

Due to the space limit, some definitions, all the proofs are relegated to the appendix in the Supplementary Material, which also includes the codes of experiments.

2. Related Work

As mentioned earlier, there is a long list of works on DP-SCO or DP-ERM. However, none of them considers the case with heavy-tailed data. Recently, a number of works have studied the SCO and ERM problems with heavy-tailed data (Brownlees et al., 2015; Minsker et al., 2015; Hsu & Sabato, 2016; Lecué et al., 2018). However, all of them focus on the non-private version of the problem. It is not clear whether they can be adapted to private versions. To our best knowledge, the work presented in this paper is the first one on general DP-SCO with heavy-tailed data.

The works that are most related to ours are perhaps those dealing with unbounded sensitivity. (Dwork & Lei, 2009) proposed a general framework called propose-test-release and applied it to mean estimation. They obtained asymptotic results which are incomparable with ours. Also, it is not clear whether such a framework can be applied to our problem. In our second result, we adopt the private mean estimation procedure in (Bun & Steinke, 2019). However, their results are in expectation form, which is not preferred in robust estimation (Brownlees et al., 2015). For this reason, we propose a new algorithm which yields theoretically guaranteed bounds with high probability. (Karwa & Vadhan, 2017) considered the confidence interval estimation problem for Gaussian distributions which was later extended to general distributions (Feldman & Steinke, 2018). However, it was unknown how to extend them to the DP-SCO problem. (Abadi et al., 2016) proposed a DP-SGD method based on truncating the gradient, which could deal with the infinity sensitivity issue. However, there is no theoretical guarantees on the excess population risk.

3. Preliminaries

Definition 1 (Differential Privacy (Dwork et al., 2006))**.** Given a data universe $\mathcal{X}$, we say that two datasets $D, D' \subseteq \mathcal{X}$ are neighbors if they differ by only one entry, which is denoted as $D \sim D'$. A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private (DP) if for all neighboring datasets D, D' and for all events S in the output space of $\mathcal{A}$, the following holds

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

Definition 2 (DP-SCO (Bassily et al., 2014))**.** Given a dataset $D = \{x_1, \cdots, x_n\}$ from a data universe $\mathcal{X}$ where x_i are i.i.d. samples from some unknown distribution $\mathcal{D}$, a convex loss function $\ell(\cdot, \cdot)$, and a convex constraint set $\mathcal{W} \subseteq \mathbb{R}^d$, 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 the algorithm is measured by the *(expected) excess population risk*, that is $\mathbb{E}_{\mathcal{A}}[L_{\mathcal{D}}(w^{\text{priv}})] - \min_{w \in \mathcal{W}} L_{\mathcal{D}}(w)$, where

the expectation of $\mathcal{A}$ is taken over all the randomness of the algorithm. Besides the population risk, we can also measure the *empirical risk* of dataset D: $\hat{L}(w, D) = \frac{1}{n}\sum_{i=1}^{n}\ell(w, x_i)$.

Definition 3. A random variable X with mean μ is called τ-sub-exponential if $\mathbb{E}[\exp(\lambda(X - \mu))] \leq \exp(\frac{1}{2}\tau^2\lambda^2), \forall|\lambda| \leq \frac{1}{\tau}$.

Definition 4. A function f is L-Lipschitz if for all $w, w' \in \mathcal{W}$, $|f(w) - f(w')| \leq L\|w - w'\|_2$.

Definition 5. A function f is α-strongly convex on $\mathcal{W}$ if for all $w, w' \in \mathcal{W}$, $f(w') \geq f(w) + \langle\nabla f(w), w' - w\rangle + \frac{\alpha}{2}\|w' - w\|_2^2$.

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

Assumption 1. For the loss function and the population risk, we assume the following.

1. The loss function $\ell(w, x)$ is non-negative, differentiable and convex for all $w \in \mathcal{W}$ and $x \in \mathcal{X}$.

2. The population risk $L_\mathcal{D}(w)$ is β-smooth.

3. The convex constraint set $\mathcal{W}$ is bounded with diameter $\Delta = \max_{w,w'\in\mathcal{W}}\|w - w'\|_2 < \infty$.

4. The optimal solution $w^* = \arg\min_{w\in\mathcal{W}} L_\mathcal{D}(w)$ satisfies $\nabla L_\mathcal{D}(w^*) = 0$.

Assumption 2. There exists a number n_α such that when the sample size $|D| \geq n_\alpha$, the empirical risk $\hat{L}(\cdot, D)$ is α-strongly convex with probability at least $\frac{5}{6}$ over the choice of i.i.d. samples in D.

We note that Assumptions 1 and 2 are commonly used in the studies on the problem of Stochastic Strongly Convex Optimization with heavy-tailed data, such as (Hsu & Sabato, 2016; Holland, 2019). Also the probability of $\frac{5}{6}$ in Assumption 2 is only for convenience.

Assumption 3. We assume the following for the loss functions.

1. For any $w \in \mathcal{W}$ and each coordinate $j \in [d]$, we assume that the random variable $\nabla_j\ell(w, x)$ is τ-sub-exponential and β_j-Lipschitz (that is $\ell_j(w, x)$ is β_j-smooth), where ∇_j represents the j-th coordinate of the gradient.

2. There are known constants $a, b = O(1)$ such that $a \leq \mathbb{E}[\nabla_j\ell(w, x)] \leq b$ for all $w \in \mathcal{W}$.

Assumption 4. For any $w \in \mathcal{W}$ and each coordinate $j \in [d]$, we have $\mathbb{E}[(\nabla_j\ell(w, x))^2] \leq v = O(1)$, where v is some known constant.

We can see that, compared with Assumption 3, Assumption 4 needs fewer assumptions on the loss functions, because we only need to assume the gradient of the loss function has bounded second-order moment. We also note that Assumption 4 is more suitable to the problem of Stochastic Convex Optimization with heavy-tailed data and has been used in some previous works such as (Holland & Ikeda, 2017; Brownlees et al., 2015).

4. Sample-aggregation based method

In this section we first summarize the sample-aggregate framework introduced in (Nissim et al., 2007).

Most of the existing privacy-preserving frameworks are based on the notion of *global sensitivity*, which is defined as the maximum output perturbation $\|f(D) - f(D')\|_\xi$, where the maximum is over all neighboring datasets D, D' and $\xi = 1, 2$. However, in some problems such as clustering (Nissim et al., 2007; Wang et al., 2015) the sensitivity could be very high and thus ruin the utility of the algorithm.

To circumvent this issue, (Nissim et al., 2007) introduced the sample-aggregate framework based on a smooth version of *local sensitivity*. Unlike the global sensitivity, local sensitivity measures the maximum perturbation $\|f(D) - f(D')\|_\xi$ over all databases D' neighboring the input database D. The proposed sample-aggregate framework (Algorithm 1) enjoys local sensitivity and comes with the following guarantee:

Theorem 1 (Theorem 4.2 in (Nissim et al., 2007))**.** Let $f : \mathcal{D} \mapsto \mathbb{R}^d$ be a function where $\mathcal{D}$ is the collection of all databases and d is the dimensionality of the output space. Let $d_\mathcal{M}(\cdot, \cdot)$ be a semi-metric on the output space of f. Set $\epsilon > \frac{2d}{\sqrt{m}}$ and $m = \omega(\log^2 n)$. The sample-aggregate algorithm $\mathcal{A}$ in Algorithm 1 is an efficient (ϵ, δ)-DP algorithm.[1] Furthermore, if f and m are chosen such that the ℓ_1 norm of the output of f is bounded by Λ and

$$\Pr_{D_S\subseteq D}[d_\mathcal{M}(f(D_S), c) \leq r] \geq \frac{3}{4} \tag{1}$$

for some $c \in \mathbb{R}^d$ and $r > 0$, then the standard deviation of Gaussian noise added is upper bounded by $O(\frac{r}{\epsilon} + \frac{\Lambda}{\epsilon}e^{-\Omega(\frac{\epsilon\sqrt{m}}{d})})$. In addition, when $m = \omega(\frac{d^2\log^2(r/\Lambda)}{\epsilon^2})$, with high probability each coordinate of $\mathcal{A}(D) - \bar{c}$ is upper bounded by $O(\frac{r}{\epsilon})$, where $\bar{c}$ depending on $\mathcal{A}(D)$ satisfies $d_\mathcal{M}(c, \bar{c}) = O(r)$.

We have the following Lemma 1, which shows that the minimum of the empirical risk satisfies (1).

Lemma 1. Let $w_D = f(D) = \arg\min_{w\in\mathcal{W}}\hat{L}(w, D)$ where $|D| = n$. Then, under Assumptions 1 and 2, if

---

[1] Here the efficiency means that the time complexity is polynomial in all terms.

Algorithm 1 Sample-aggregate Framework (Nissim et al., 2007)

Input: $D = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$, number of subsets m, privacy parameters ϵ, δ; $f, d_{\mathcal{M}}$.

1: **Initialize:** $s = \sqrt{m}, \gamma = \frac{\epsilon}{5\sqrt{2\log(2/\delta)}}$ and $\beta = \frac{\epsilon}{4(d+\log(2/\delta))}$.

2: **Subsampling:** Select m random subsets of size $\frac{n}{m}$ of D independently and uniformly at random without replacement. Repeat this step until no single data point appears in more than $\sqrt{m}$ of the sets. Mark the subsampled subsets $D_{S_1}, D_{S_2}, \cdots, D_{S_m}$.

3: Compute $\mathcal{S} = \{s_i\}_{i=1}^m$, where $s_i = f(D_{S_i})$.

4: Compute $g(\mathcal{S}) = s_{i^*}$, where $i^* = \arg\min_{i=1}^m r_i(t_0)$ with $t_0 = \frac{m+s}{2} + 1$. Here $r_i(t_0)$ denotes the distance $d_{\mathcal{M}}(\cdot, \cdot)$ between s_i and the t_0-th nearest neighbor to s_i in $\mathcal{S}$.

5: **Noise Calibration:** Compute $S(\mathcal{S}) = 2\max_k(\rho(t_0 + (k+1)s) \cdot e^{-\beta k})$, where $\rho(t)$ is the mean of the top $\lceil \frac{s}{\beta} \rceil$ values in $\{r_1(t), \cdots, r_m(t)\}$.

6: Return $\mathcal{A}(D) = g(\mathcal{S}) + \frac{S(\mathcal{S})}{\gamma} u$, where u is a standard Gaussian random vector.

$n \geq n_\alpha$, the following holds

$$\Pr[\|w_D - w^*\|_2 \leq \eta] \geq \frac{3}{4}, \quad (2)$$

where $\eta = O(\sqrt{\frac{\mathbb{E}\|\nabla\ell(w^*, x)\|_2^2}{n\alpha^2}})$.

Combining Lemma 1 and Theorem 1, we get the following upper bound for DP-SCO with heavy-tailed data and strongly convex loss functions.

Theorem 2. Under Assumptions 1 and 2, for any $\epsilon, \delta > 0$, if $n \geq \tilde{\Omega}(\frac{n_\alpha d^2}{\epsilon^2})$, $m \geq \tilde{\omega}(\frac{d^2}{\epsilon^2})$, $f(D) = \arg\min_{w \in \mathcal{W}} \hat{L}(w, D)$ and $d_{\mathcal{M}}(x, y) = \|x - y\|_2$, then Algorithm 1 is (ϵ, δ)-DP. Moreover, with high probability the output of $\mathcal{A}(D)$ ensures that

$$L_{\mathcal{D}}(\mathcal{A}(D)) - L_{\mathcal{D}}(w^*) \leq \tilde{O}((\frac{\beta}{\alpha})^2 \frac{d^3}{n\epsilon^4} L_{\mathcal{D}}(w^*)), \quad (3)$$

where the Big-$\tilde{O}, \Omega$ and small-ω notations omit the logarithmic terms.

Remark 1. For DP-SCO with Lipschitz and strongly-convex loss function and bounded data, (Bassily et al., 2014; Wang et al., 2017; Bassily et al., 2019) showed that the upper bound of the excess population risk is $O(\frac{\sqrt{d}}{n\epsilon})$, and the lower bound is $\Omega(\frac{d}{n^2\epsilon^2})$ [2]. This suggests that the

---

[2] (Bassily et al., 2014) only shows the lower bound of the excess empirical risk. We can obtain the lower bound of the excess population risk by using the reduction from private ERM to private SCO (Bassily et al., 2019).

bound in Theorem 2 has some additional factors related to d and $\frac{1}{\epsilon}$. We note that the upper bound in Theorem 2 has a multiplicative term of $L_{\mathcal{D}}(w^*)$. This means that when $L_{\mathcal{D}}(w^*)$ is small, our bound is better. For example, when $L_{\mathcal{D}}(w^*) = 0$, our algorithm can recover w^* exactly and results in an excess risk of 0. Notice that there is no previous work on DP-ERM or DP-SCO that has a multiplicative error with respect to $L_{\mathcal{D}}(w^*)$.

5. Gradient descent based methods

There are several issues in the sample-aggregation based method presented in last section. Firstly, function $f(D)$ in Theorem 2 needs to solve the optimization problem exactly, which could be quite inefficient in practice. Second, previous empirical evidence suggests that sample-aggregation based methods often suffer from poor utility in practice (Su et al., 2016; Wang et al., 2015). Thirdly, Theorem 2 needs to assume strong convexity for the empirical risk and it is unclear whether it can be extended to the general convex case. Finally, from Eq.(3) we can see that when $L_{\mathcal{D}}(w^*) = \Theta(1)$, the excess population risk is quite large as compared to the ones in (Bassily et al., 2014). Thus, an immediate question is whether we can further lower the upper bound. To answer this question and resolve the above issues, we propose in this section two DP algorithms based on the Gradient Descent method under different assumptions.

Recently, (Bun & Steinke, 2019) studied the problem of estimating the mean of a 1-dimensional heavy-tailed distribution and proposed algorithms based on the idea of truncating the empirical mean and the local sensitivity. Motivated by this DP algorithm that has the capability of handling heavy-tailed data, we plan to develop a new method by borrowing some ideas from the work (Bun & Steinke, 2019) and robust gradient descent. Our method is inspired by their theorem that follows and uses the Arsinh-Normal mechanism (see Algorithm 2 and Prop. 5 in (Bun & Steinke, 2019)).

Theorem 3 (Theorem 7 in (Bun & Steinke, 2019)). Let $0 < \epsilon, \delta \leq 1$ be two constants and n be some integer $\geq O(\log(\frac{n(b-a)/\sigma}{\epsilon})$. Then, there exists a $\frac{1}{2}\epsilon^2$-zero concentrated Differentially Private (zCDP) (see Appendix for the definition of zCDP) algorithm (Algorithm 2) $M : \mathbb{R}^n \mapsto \mathbb{R}$ such that the following holds: Let $\mathcal{D}$ be a distribution with mean $\mu \in [a, b]$, where a, b are given constants and unknown variance σ^2. Then,

$$\mathbb{E}_{X \sim \mathcal{D}^n, Z}[(M(X) - \mu)^2] \leq O(\frac{\sigma^2 \log n}{n\epsilon^2}).$$

The key idea of our algorithm is that, in each iteration, after getting w^{t-1}, we use the mechanism in Theorem 3 on each coordinate of $\nabla\ell(w, x_i)$. See Algorithm 3 for details.

By the composition theorem and the relationship between

Algorithm 2 Mechanism $\mathcal{M}$ in (Bun & Steinke, 2019)

Input: $D = \{x_i\}_{i=1}^n \subset \mathbb{R}, \epsilon, a, b$.

1: Let $t = \frac{\epsilon^2}{16}$ and $s = \frac{\epsilon}{4}$. Sort $\{x_i\}_{i=1}^n$ in the ascending order as $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Calculate the upper bound of the smooth sensitivity for the trimming and truncating step:

$$S^t_{[\text{trim}_m(\cdot)]_{[a,b]}}(D) = \max\{\frac{x_{(n)} - x_{(1)}}{n - 2m}, e^{-mt}(b-a)\},$$

where $m = O(1) \leq \frac{n}{2}$ is a constant.

2: Do the average trimming and truncating step:

$$[\text{Trim}_m(D)]_{[a,b]} = [\frac{x_{(m+1)} + \cdots + x_{(n-m)}}{n - 2m}]_{[a,b]},$$

where $[x]_{[a,b]} = x$ if $a \leq x \leq b$, equals to a if $x < a$ and otherwise equals to b.

3: Output $[\text{Trim}_m(D)]_{[a,b]} + \frac{1}{s} S^t_{[\text{trim}_m(\cdot)]_{[a,b]}}(D) \cdot Z$, where $Z = \sinh(Y) = \frac{e^Y - e^{-Y}}{2}$ and Y is the Standard Gaussian.

Algorithm 3 Heavy-tailed DP-SCO with known mean

Input: $D = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$, privacy parameters ϵ, δ; loss function $\ell(\cdot, \cdot)$, initial parameter w^0, a, b which satisfy Assumption 3, and the number of iterations T (to be specified later).

1: Let $\tilde{\epsilon} = \sqrt{2\log\frac{1}{\delta} + 2\epsilon} - \sqrt{2\log\frac{1}{\delta}}$.

2: **for** $t = 1, 2, \cdots, T$ **do**

3:     For each $j \in [d]$, calculate $D_{t-1,j}(w^{t-1}) = \{\nabla_j \ell(w^{t-1}, x_i)\}_{i=1}^n$.

4:     Run Algorithm 2 for each $D_{t-1,j}$ and denote the output $\tilde{\nabla}_{t-1,j}(w^{t-1}) = (\mathcal{M}(D_{t-1,j}(w^{t-1})), \frac{\tilde{\epsilon}}{\sqrt{dT}}, a, b)$. Denote

$$\nabla\tilde{L}(w^{t-1}, D) = (\tilde{\nabla}_{t-1,1}(w^{t-1}) \cdots, \tilde{\nabla}_{t-1,d}(w^{t-1})).$$

5:     Updating $w^t = \mathcal{P}_{\mathcal{W}}(w^{t-1} - \eta_{t-1}\nabla\tilde{L}(w^{t-1}, D))$, where η_{t-1} is some step size and $\mathcal{P}_{\mathcal{W}}$ is the projection operator.

6: **end for**

$zCDP$ and (ϵ, δ)-DP (Bun & Steinke, 2016), we have the DP guarantee.

Theorem 4. For any $0 < \epsilon, \delta \leq 1$, Algorithm 3 is (ϵ, δ)-differentially private.

To show the *expected* excess population risk of Algorithm 3, we cannot use the upper bound in Theorem 3 directly for the following reasons. First, since the upper bound is for the expectation w.r.t. X and Z while the *expected* excess population risk depends only on the randomness of the algorithm instead of the data. Thus, we need to obtain an upper bound for $\mathbb{E}_Z[(M(X) - \mu)^2]$ (with high probability w.r.t. X). Secondly, to get an upper bound, it is sufficient to analyze the term $\|\nabla\tilde{L}(w^{t-1}, D) - \nabla L_{\mathcal{D}}(w^{t-1})\|_2$ in each iteration. However, since the parameter w^{t-1} at any step depends on the random draw of the dataset $\{x_i\}_{i=1}^n$, upper bounds on the estimation error need to be uniform in $w \in \mathcal{W}$ in order to capture all contingencies. To resolve these two issues, we use the same technique as in (Chen et al., 2017; Vershynin, 2010) (under Assumption 3) to obtain the following lemma.

Lemma 2. Under Assumption 3, with probability at least $1 - \frac{2dn}{(1+n\hat{\beta}\Delta)^d}$ the following holds for all $w \in \mathcal{W}$,

$$\mathbb{E}_Z \|\nabla\tilde{L}(w, D) - \nabla L_{\mathcal{D}}(w)\|_2 \leq O(\frac{\tau d\sqrt{T\log n}}{\sqrt{n}\tilde{\epsilon}}), \quad (4)$$

where $\hat{\beta} = \sqrt{\beta_1^2 + \cdots + \beta_d^2}$, the expectation is w.r.t. the random variables $\{Z_i\}_{i=1}^d$ and the Big-O notation omits other factors.

Next, we show the expected excess population risk for strongly convex loss functions.

Theorem 5 (Strongly-convex case). Under Assumptions 1 and 3, if the population risk is α-strongly convex and T and η are set to be $T = O(\frac{\beta}{\alpha}\log n)$ and $\eta = \frac{1}{\beta}$, respectively, in Algorithm 3, then with probability at least $1 - \Omega(\frac{\beta}{\alpha}\frac{2dn\log n}{(1+n\hat{\beta}\Delta)^d})$ the output satisfies the following for all $D \sim \mathcal{D}^n$,

$$\mathbb{E}[L_{\mathcal{D}}(w^T)] - L_{\mathcal{D}}(w^*) \leq O(\frac{\Delta^2\beta^2\tau^2 d^2 \log^2 n \log\frac{1}{\delta}}{\alpha^3 n\epsilon^2}).$$

Compared with the bound in Theorem 2, we can see that the bound in Theorem 5 improves a factor of $\tilde{O}(\frac{d}{\epsilon^2})$ (if we omit other terms). However, there are more assumptions on the distribution and the loss functions. Specifically, in Assumption 3 we need to assume the sub-exponential property, *i.e.,* the moment of $\nabla_j\ell(w, x)$ exists for every order. Also, we need to assume that $\nabla_j\ell(w, x)$ is Lipschitz and the range of its mean is known. These assumptions are quite strong, compared to those used in the literature of learning with heavy-tailed data, such as (Holland & Ikeda, 2017; Brownlees et al., 2015; Hsu & Sabato, 2016; Minsker et al., 2015).

To improve the above result, we consider the following. First, we would like to relax those assumptions in the theorem. Second, in the problem of ERM with heavy-tailed data, it is expected to have an excess population risk bound that is in the form of *with high probability* instead of its *expectation* (Brownlees et al., 2015). However, it is unclear whether Algorithm 3 can achieve a high probability bound.

This is due to the fact that the noise added in each iteration is a combination of log-normal distributions, which is non-sub-exponential and thus is hard to get tail bounds. Third, Algorithm 3 depends on the local sensitivity and thus cannot be extended to the distributed settings or local differential privacy model. Finally, the practical performance of Algorithm 3 has poor utility and is unstable due to the noise added in each iteration (see Section 6 for details), which means that Algorithm 3 is still impractical. To resolve all these issues and still keeping (approximately) the same upper bound, we propose a new algorithm that is simply based on the Gaussian mechanism.

In the following we will study the problem under Assumptions 1 and 4. Note that compared with Assumption 3, we only need to assume that the second-order moment of $\nabla_j \ell(w, x)$ exists for all $w \in \mathcal{W}$ and $j \in [d]$ and its upper bound is known.

Our method is motivated by the robust mean estimator given in (Holland, 2019). To be self-contained, we first review their estimator. Now, we consider 1-dimensional random variable x and assume that $x_1, x_2, \cdots, x_n$ are i.i.d. sampled from x. The estimator consists of the following steps:

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

$$\frac{s}{n} \sum_{i=1}^{n} \phi(\frac{x_i}{s}) \approx \mathbb{E}X. \quad (5)$$

Here, we use the function given in (Catoni & Giulini, 2017),

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

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

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,

$$\tilde{x}(\eta) = \frac{s}{n} \sum_{i=1}^{n} \phi(\frac{x_i + \eta_i x_i}{s}). \quad (7)$$

Noise Smoothing In this final step, we smooth the multiplicative noise by taking the expectation w.r.t. the distribu-

tions. That is,

$$\hat{x} = \mathbb{E}\tilde{x}(\eta) = \frac{s}{n} \sum_{i=1}^{n} \int \phi(\frac{x_i + \eta_i x_i}{s}) d\chi(\eta_i). \quad (8)$$

Computing the explicit form of each integral in (8) depends on the function $\phi(\cdot)$ and the distribution χ. Fortunately, (Catoni & Giulini, 2017) showed that when ϕ is in (6) and $\chi \sim \mathcal{N}(0, \frac{1}{\beta})$ (where β will be specified later), we have for any a, b

$$\mathbb{E}_\eta \phi(a + b\sqrt{\beta}\eta) = a(1 - \frac{b^2}{2}) - \frac{a^3}{6} + C(a, b), \quad (9)$$

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

(Holland, 2019) showed the following estimation error for the mean estimator $\hat{x}$ after these three steps.

Lemma 3 (Lemma 5 in (Holland, 2019))**.** Let $x_1, x_2, \cdots, x_n$ be i.i.d. samples from distribution $x \sim \mu$. Assume that there is some known upper bound on the second-order moment, *i.e.*, $\mathbb{E}_\mu x^2 \leq v$. For a given failure probability δ', if set $\beta = 2 \log \frac{1}{\delta'}$ and $s = \sqrt{\frac{nv}{2 \log \frac{1}{\delta'}}}$, then with probability at least $1 - \delta'$ the following holds

$$|\hat{x} - \mathbb{E}x| \leq O(\sqrt{\frac{v \log \frac{1}{\delta'}}{n}}). \quad (11)$$

To obtain an (ϵ, δ)-DP estimator, the key observation is that the bounded function ϕ in (6) also makes the integral form of (9) bounded by $\frac{2\sqrt{2}}{3}$. Thus, we know that the ℓ_2-norm sensitivity is $\frac{s}{n}\frac{4\sqrt{2}}{3}$. Hence, the query

$$\mathcal{A}(D) = \hat{x} + Z, Z \sim \mathcal{N}(0, \sigma^2), \sigma^2 = O(\frac{s^2 \log \frac{1}{\delta}}{\epsilon^2 n^2}) \quad (12)$$

will be (ϵ, δ)-DP, which leads to the following theorem.

Theorem 6. Under the assumptions in Lemma 3, with probability at least $1 - \delta'$ the following holds

$$|\mathcal{A}(D) - \mathbb{E}(x)| \leq O(\sqrt{\frac{v \log \frac{1}{\delta} \log \frac{1}{\delta'}}{n\epsilon^2}}). \quad (13)$$

Comparing with Theorem 3, we can see that the upper bound in Theorem 6 is in the form of 'with high probability' (after transferring zCDP to (ϵ, δ)-DP (Bun & Steinke, 2016)). Moreover, we improve by a factor of $O(\log n)$ in the error bound.

Inspired by Theorem 6 and Algorithm 3, we propose a new method (Algorithm 4), which uses our private mean estimator (12) on each coordinate of the gradient in each iteration. The following theorem shows the error bound when the loss function is strongly convex.

Algorithm 4 Heavy-tailed DP-SCO with known variance

Input: $D = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$, privacy parameters ϵ, δ, loss function $\ell(\cdot, \cdot)$, initial parameter w^0, v which satisfies Assumption 4, the number of iterations T (to be specified later), and failure probability δ'.

1: Let $\tilde{\epsilon} = (\sqrt{\log \frac{1}{\delta} + \epsilon} - \sqrt{\log \frac{1}{\delta}})^2$, $s = \sqrt{\frac{nv}{2\log \frac{1}{\delta'}}}$, $\beta = \log \frac{1}{\delta'}$.

2: **for** $t = 1, 2, \cdots, T$ **do**

3:    For each $j \in [d]$, calculate the robust gradient by (7)-(9), that is

$$g_j^{t-1}(w^{t-1}) = \frac{1}{n} \sum_{i=1}^n \left(\nabla_j \ell(w^{t-1}, x_i)\left(1 - \frac{\nabla_j^2 \ell(w^{t-1}, x_i)}{2s^2\beta}\right) - \frac{\nabla_j^3 \ell(w^{t-1}, x_i)}{6s^2} \right)$$
$$+ \frac{s}{n} \sum_{i=1}^n C\left(\frac{\nabla_j \ell(w^{t-1}, x_i)}{s}, \frac{|\nabla_j \ell(w^{t-1}, x_i)|}{s\sqrt{\beta}} \right) + Z_j^{t-1}, \quad (10)$$

where $Z_j^{t-1} \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = \frac{8vdT}{9\log \frac{1}{\delta'} n\tilde{\epsilon}}$.

4:    Let vector $g^{t-1}(w^{t-1}) \in \mathbb{R}^d$ to denote $g^{t-1}(w^{t-1}) = (g_1^{t-1}(w^{t-1}), g_2^{t-1}(w^{t-1}), \cdots, g_d^{t-1}(w^{t-1}))$.

5:    Update $w^t = \mathcal{P}_\mathcal{W}(w^{t-1} - \eta_{t-1} g^{t-1})$.

6: **end for**

Theorem 7. For any $0 < \epsilon, \delta < 1$, Algorithm 4 is (ϵ, δ)-DP. Under Assumptions 1 and 4, if the population risk is α-strongly convex and η_t and T in Algorithm 4 are set to be $\eta_t = \frac{1}{\beta}$ and $T = O(\frac{\beta}{\alpha} \log n)$, respectively, then for any $\delta' > 0$, with probability at least $1 - 2\delta'T$ the output w^T satisfies

$$L_\mathcal{D}(w^T) - L_\mathcal{D}(w^*) \leq O(\frac{v\Delta^2 \beta^4 d^2 \log^2 n \log \frac{1}{\delta} \log \frac{1}{\delta'}}{\alpha^3 n\epsilon^2}).$$

Comparing with Theorem 7 and 5, we can see that if we omit other terms, the bounds are asymptotically the same and Theorem 7 needs fewer assumptions.

With the high probability guarantee on the error in Theorem 6, we can actually get an upper bound for general convex loss functions. For this general convex case, we need the following mild technical assumption on the constraint set $\mathcal{W}$.

Assumption 5. The constraint set $\mathcal{W}$ contains the following ℓ_2-ball centered at w^*: $\{w : \|w - w^*\|_2 \leq 2\|w^0 - w^*\|_2\}$.

Theorem 8 (Convex case). Under Assumptions 1, 4 and 5, if we take $\eta = \frac{1}{\beta}$ and $T = \tilde{O}\left(\left(\frac{\|w^0 - w^*\|_2 \sqrt{n}\sqrt{\epsilon}}{d}\right)^{\frac{2}{3}} \right)$ in Algorithm 4, then for any given failure probability δ', with probability at least $1 - T\delta'$ the following holds

$$L_\mathcal{D}(w^T) - L_\mathcal{D}(w^*) \leq \tilde{O}(\frac{\log^{\frac{1}{3}} \frac{1}{\delta} \sqrt{\log \frac{1}{\delta'}} d^{\frac{2}{3}}}{(n\epsilon^2)^{\frac{1}{3}}}) \quad (14)$$

when $n \geq \tilde{\Omega}(\frac{d^2}{\epsilon^2})$, where the Big-$\tilde{O}$ notation omits other logarithmic factors and the term of v, β.

6. Experiments

Baseline Methods As mentioned earlier, sample-aggregation based methods often have poor practical performance. Thus, we will not conduct experiments on Algorithm 1. Moreover, as this is the first paper studying DP-SCO with heavy-tailed data and almost all previous methods on DP-SCO that have theoretical guarantees fail to provide DP guarantees, we do not compare our methods with them, and instead focus on comparing the performance of Algorithm 3 and Algorithm 4. To show the effectiveness of our methods, we use the non-private heavy-tailed SCO method in (Holland, 2019), denoted by (stochastic) RGD in the following, as our baseline method.

Experimental Settings For synthetic data, we consider the linear and binary logistic models. Specifically, we generate the synthetic datasets in the following way. Each dataset has a size of 1×10^5 and each data point (x_i, y_i) is generated by the model of $y_i = \langle \omega^*, x_i \rangle + e_i$ and $y_i = \text{sign}[\frac{1}{1 + e^{\langle \omega^*, x_i \rangle + e_i}} - \frac{1}{2}]$, respectively, where $x_i \in \mathbb{R}^{10}$ and $y_i \in \mathbb{R}$. In the first model, the zero mean noise e_i is generated as follows. We first generate a noise Δ_i from the (μ, σ) log-normal distribution, i.e., $\mathbb{P}(\Delta_i = x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$, and then let $e_i = \Delta_i - \mathbb{E}[\Delta_i]$. For the second model, we first generate a noise Δ_i from the (μ, σ) log-logistic distribution, i.e., $\mathbb{P}(\Delta_i = x) = \frac{e^z}{\sigma x(1 + e^z)^2}$, where $x > 0$ and $z = \frac{\log(x) - \mu}{\sigma}$. Then, we let $e_i = \Delta_i - \mathbb{E}[\Delta_i]$. Accordingly, we implement Algorithm 3 and Algorithm 4, together with RGD, on the ridge and logistic regressions.

For real-world data, we use the Adult dataset from the

UCI Repository (Dua & Graff, 2017). We aim to predict whether the annual income of an individual is above 50,000. We select 30,000 samples, 28,000 amongst which are used as the training set and the rest are used for test.

For the privacy parameters, we will choose $\epsilon = \{0.1, 0.5, 1\}$ and $\delta = O(\frac{1}{n})$. See Appendix for the selections of other parameters. For Algorithm 3, the strength of prior knowledge is modeled by $\kappa = b - a$.

Experimental Results Figure 1 and 2 show the results of ridge and logistic regressions on synthetic and real datasets w.r.t iteration, respectively. Since there is no ground truth in the real dataset, we use the empirical risk on test data as the measurement. To test scalability of Algorithm 4 dealing with large-scaling data, experiments on stochastic versions of Algorithm 4 and RGD with minibatch size 1000 are also conducted. We can see that the performance of Algorithm 3 bears a larger variation compared to Algorithm 4, since we have to apply a heavy-tailed noise to fit the smooth sensitivity. Moreover, the performance of Algorithm 3 is sensitive to the parameter κ. Thus, these results show that Algorithm 3 has poor performance and the results of Algorithm 4 are comparable to the non-private ones. In Figure 3 and 4 we test the estimation error w.r.t different dimensionality d and sample size n, respectively. From these results we can see that when n increases or d decreases, the estimation error will decrease. Also, with fixed n and d, we can see that the estimation error will decrease as ϵ becomes larger. Thus, all these results confirm our previous theoretical analysis.

7. Discussion

In this paper, we provide the first comprehensive study on DP-SCO with heavy-tailed data. To the best of our knowledge, this is the first work on this problem. Specifically, we give a systematic analysis on the problem and design the first efficient algorithms to solve it. In various settings, we bound the (expected) excess generalization risk in both addictive and multiplicative manners. However, the problem is far from being closed. First, it is unclear whether the upper bounds of the excess population risk for strongly convex and general convex loss functions can be further improved. The second open problem is that we do not know what the lower bound for the excess population risk for these two cases is. Finally, it is an open problem to determine whether we can further relax the assumptions in our previous theorems. We leave these open problems for future research.

Acknowledgements

References

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

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

Bassily, R., Feldman, V., Talwar, K., and Thakurta, A. Private stochastic convex optimization with optimal rates. In *NeurIPS*, 2019.

Biswas, A., Datta, S., Fine, J. P., and Segal, M. R. *Statistical advances in the biomedical science*. Wiley Online Library, 2007.

Brownlees, C., Joly, E., Lugosi, G., et al. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536, 2015.

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

Bun, M. and Steinke, T. Average-case averages: Private algorithms for smooth sensitivity and mean estimation. *arXiv preprint arXiv:1906.02830*, 2019.

Catoni, O. and Giulini, I. Dimension-free pac-bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*, 2017.

Chaudhuri, K. and Monteleoni, C. Privacy-preserving logistic regression. In *Advances in neural information processing systems*, pp. 289–296, 2009.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

Chen, Y., Su, L., and Xu, J. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):44, 2017.

Ding, B., Kulkarni, J., and Yekhanin, S. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*, pp. 3571–3580, 2017.
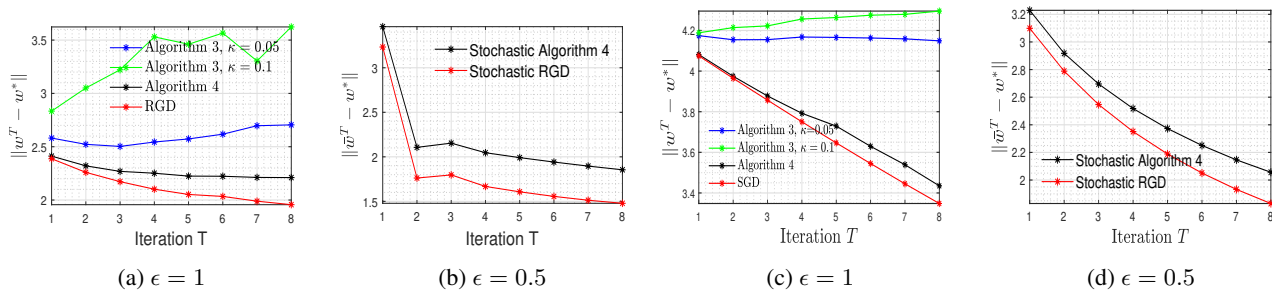
Figure 1: Experiments on synthetic datasets. Figures 1a and 1b are for ridge regressions over synthetic data with Lognormal noises. Figures 1c and 1d are for logistic regressions over synthetic data with Loglogistic noises.
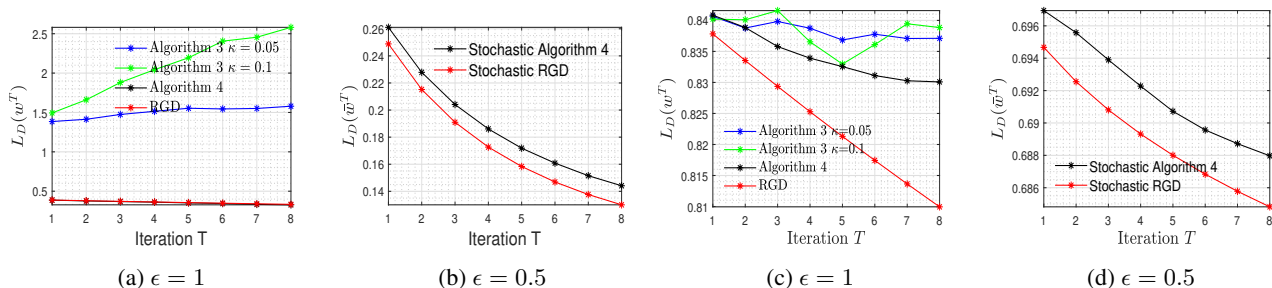

Figure 2: Experiments on UCI Adult dataset. Figures 2a and 2b are for ridge regressions. Figures 2c and 2d are for logistic regressions.
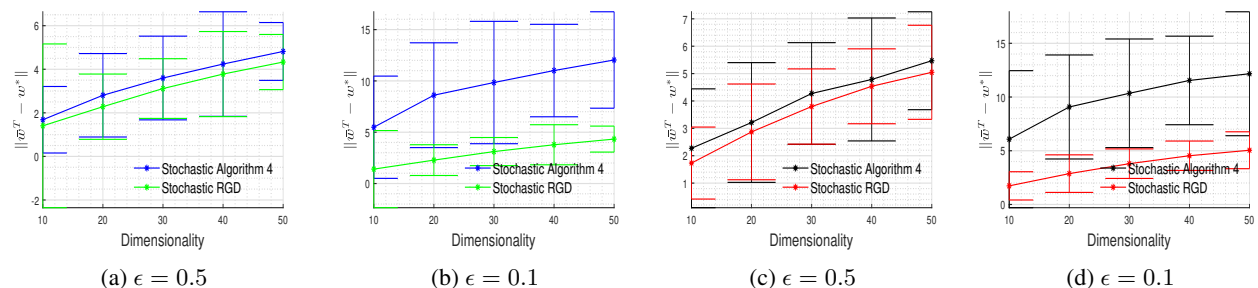

Figure 3: Experiments for the impact of dimensionality. Figure 3a and 3b are for ridge regressions. Figure 3c and 3d are for logistic regressions.
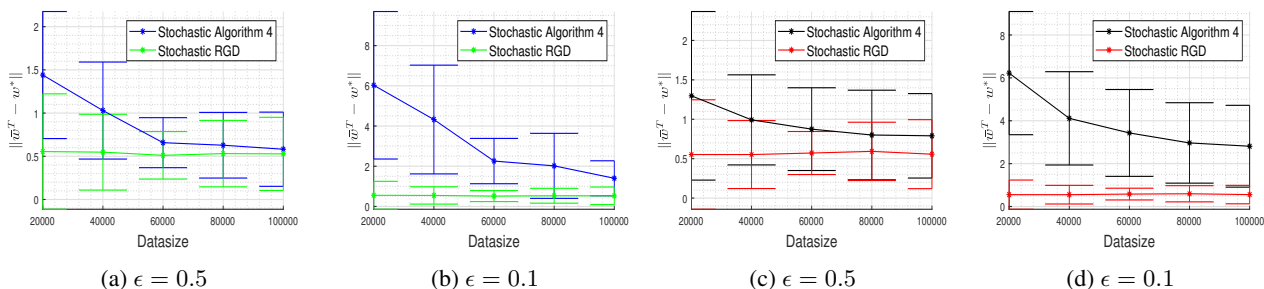

Figure 4: Experiments for the impact of the size of the dataset. Figure 4a and 4b are for ridge regressions. Figure 4c and 4d are for logistic regressions.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

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

Dwork, C. and Lei, J. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 371–380. ACM, 2009.

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

Fama, E. F. Mandelbrot and the stable paretian hypothesis. *The journal of business*, 36(4):420–429, 1963.

Feldman, V. and Steinke, T. Calibrating noise to variance in adaptive data analysis. In *Conference On Learning Theory*, pp. 535–544, 2018.

Holland, M. J. Robust descent using smoothed multiplicative noise. In *22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pp. 703–711, 2019.

Holland, M. J. and Ikeda, K. Efficient learning with robust gradient descent. *Machine Learning*, pp. 1–38, 2017.

Hsu, D. and Sabato, S. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016.

Ibragimov, M., Ibragimov, R., and Walden, J. *Heavy-tailed distributions and robustness in economics and finance*, volume 214. Springer, 2015.

Karwa, V. and Vadhan, S. Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908*, 2017.

Kasiviswanathan, S. P. and Jin, H. Efficient private empirical risk minimization for high-dimensional learning. In *International Conference on Machine Learning*, pp. 488–497, 2016.

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

Lecué, G., Lerasle, M., and Mathieu, T. Robust classification via mom minimization. *arXiv preprint arXiv:1808.03106*, 2018.

Mandelbrot, B. B. The variation of certain speculative prices. In *Fractals and scaling in finance*, pp. 371–418. Springer, 1997.

Minsker, S. et al. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

Nissim, K., Raskhodnikova, S., and Smith, A. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 75–84. ACM, 2007.

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

Su, D., Cao, J., Li, N., Bertino, E., and Jin, H. Differentially private k-means clustering. In *Proceedings of the sixth ACM conference on data and application security and privacy*, pp. 26–37. ACM, 2016.

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

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

Vapnik, V. *The nature of statistical learning theory*. Springer science & business media, 2013.

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

Wang, D., Ye, M., and Xu, J. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems*, pp. 2722–2731, 2017.

Wang, D., Gaboardi, M., and Xu, J. Empirical risk minimization in non-interactive local differential privacy revisited. In *Advances in Neural Information Processing Systems*, pp. 965–974, 2018.

Wang, D., Chen, C., and Xu, J. Differentially private empirical risk minimization with non-convex loss functions. In *International Conference on Machine Learning*, pp. 6526–6535, 2019a.

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

Wang, Y., Wang, Y.-X., and Singh, A. Differentially private subspace clustering. In *Advances in Neural Information Processing Systems*, pp. 1000–1008, 2015.

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

Wu, X., Li, F., Kumar, A., Chaudhuri, K., Jha, S., and Naughton, J. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pp. 1307–1322. ACM, 2017.