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

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

Abstract

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

1 Introduction

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

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

Even though there are numerous works on DP-SCO, a critical issue in most existing results is that loss function has to be assumed to satisfy the $O(1)$-Lipschitz property, or the underlying data distribution is assumed to be sub-Gaussian or even bounded. Despite simplifying the procedure of designing DP algorithms, such assumptions are unrealistic and may not always hold when dealing with real-world datasets, especially those from biomedicine and finance, as it has been observed that they are often heavy-tailed [Woolson and Clarke, 2011]. The heavy-tailed data could lead to unbounded gradients and thus break the Lipschitz assumption, which implies that previous algorithms may fail to provide DP guarantee. Recently, to tackle this issue, there have been several works studying DP-SCO with heavy-tailed data [Wang *et al.*, 2020b; Kamath *et al.*, 2022; Hu *et al.*, 2022] or private mean estimation for heavy-tailed distributions [Barber and Duchi, 2014; Kamath *et al.*, 2020; Liu *et al.*, 2021]. However, all these results still need to assume that the distribution of each coordinate of the gradient of the loss function has bounded second moment, which implies the data is still well-behaved to some extent. Thus, a natural question is,

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

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

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

[*] Contact Author

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

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

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

2 Related Work

As mentioned in Section 1, there are a vast number of works on DP-SCO and DP-ERM. Due to space limit, here we just discuss the results that are the most related to ours. For DP-SCO with heavy-tailed data, [Wang *et al.*, 2020b] first studies the problem by proposing three methods based on different assumptions. However, all the three methods need to assume the distribution of gradient of the loss is sub-exponential or has at least bounded second moment. [Kamath *et al.*, 2022] recently revisits the problem under the same assumption as in [Wang *et al.*, 2020b] and improves the (expected) excess population risk for both convex and strongly convex loss functions. It also provides the lower bounds in both (ϵ, δ)-DP and ϵ-DP models. Note that although some ideas of our algorithms and proofs are the same as theirs, their methods cannot be used in our relaxed setting and there are several differences. See Remark 1 and 2 for details.

DP sparse learning has been studied previously [Wang and Gu, 2019; Wang and Xu, 2019; Wang and Xu, 2021]. However, all of the previous methods need either the loss function be Lipschitz, or the data distribution be sub-Gaussian or even bounded [Cai *et al.*, 2021]. [Hu *et al.*, 2022] recently extends to the heavy-tailed case where each coordinate of the gradient has bounded second moment. However, due to their private estimator, it is impossible to use their methods to the bounded $(1 + v)$-th moment case. See Remark 3 for details.

3 Preliminaries

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Remark 1. When $v = 1$, the error becomes $O\left(\frac{1}{\sqrt{n\epsilon}}\right)$ (if we omit other terms), which matches the bound in [Wang *et al.*, 2020a] and is optimal [Kamath *et al.*, 2020]. However, previous results are for the case where the distribution has bounded second moment. Here we relax it to the case where the distribution only has its $(1 + v)$-th moment bounded. Thus, our method can be thought as a generalization of the previous results. Recently [Tao *et al.*, 2022] also considers a similar problem. However, there are several differences: First, [Tao *et al.*, 2022] focuses on the online setting in the ϵ-DP model while we consider the offline setting and the (ϵ, δ)-DP model. Secondly, the methods in [Tao *et al.*, 2022] are based on the tree mechanism and Laplacian mechanism, while we mainly use the Gaussian mechanism. Thus, our error bound is lower. Thirdly, besides the upper bound, we also show that the bound of $O((u^{\frac{1}{v+1}}(\frac{1}{n\epsilon})^{\frac{v}{1+v}}))$ in Theorem 1 is optimal by showing its lower bound in the Theorem 2 below, which has not been studied in [Tao *et al.*, 2022].

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

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

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

5 DP-SCO with Heavy-Tailed Data

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

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

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

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

Assumption 1. We make the following assumptions:

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

There are several notes on the terms in Assumption 1. Firstly, the first three terms in Assumption 1 are commonly used in the previous works on DP-SCO with heavy-tailed data [Kamath *et al.*, 2022; Wang *et al.*, 2020b]. The fourth condition assumes that the gradient of the loss is heavy-tailed, which is a commonly used assumption in the study of robust learning with heavy-tailed data. However, as mentioned previously, most of those works only assume that the gradient has at least bounded second moment while here we relax to the $(1+v)$-th moment. Moreover, we can see it is a relaxation of the Lipschitz condition that $\|\nabla L_{\mathcal{D}}(w)\|_2 \le L$ for all w.

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

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

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

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

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

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

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

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

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

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

Remark 2. To privately estimate the mean of heavy-tailed distributions with bounded second moments, in general there are three approaches: The first one is directly using one dimensional private estimator to each coordinate [Wang *et al.*, 2020b; Wang *et al.*, 2020a]. However, the bound of this approach is only sub-optimal. [Kamath *et al.*, 2020] proposes a method which aggressively truncate the distribution around a point, and compute the noisy empirical mean. However, their theoretical guarantee only holds with constant probability. Instead of these two approaches, here we adopt the idea of the third one, which is a private version of the MoM method and was recently proposed by [Kamath *et al.*, 2022]. However, there are two crucial differences: First, the truncation step is quite different, [Kamath *et al.*, 2022] truncates each data into an interval $[a, b]$, *i.e.,* for the sample x, if $x > b$ then we will let $x' = b$ and if $x < a$ then we will let $x' = a$. However, our approach could be seen as thresholding, *i.e.,* when $|x| > \tau$ then $x' = 0$. Second, to get the theoretical guarantee, [Kamath *et al.*, 2022] needs Lemma 4.4 in [Kamath *et al.*, 2020] (or Lemma A.2 in [Kamath *et al.*, 2022]), which only holds for data distributions that have at least bounded second moment. Thus, we cannot adopt their approach in our setting. To overcome the challenge we provide the following lemma on the concentration of heavy-tailed distributions, which is the key to prove Theorem 3.

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

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

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

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

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

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

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

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

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

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

When $v = 1$, the rate becomes $\widetilde{O}\left(\frac{d^{5/4}}{(n\epsilon)^{\frac{1}{2}}} + \frac{d^{15/4}}{(n\epsilon)^{\frac{3}{2}}}\right)$ and $\widetilde{O}\left(\frac{d^{3/2}}{n\epsilon}\right)$ for convex and strongly convex case respectively. These bounds are consistent with the best known result in [Kamath *et al.*, 2022]. However, the methods in [Kamath *et al.*, 2022] cannot be extended to the case where $v \in (0, 1)$. In the following we show that the term of $O\left(\frac{1}{(\epsilon n)^{\frac{v}{1+v}}}\right)$ and $O\left(\frac{1}{(\epsilon n)^{\frac{2v}{1+v}}}\right)$ is optimal for convex and strongly convex loss in (ϵ, δ)-DP respectively. Moreover, we provide lower bounds in the ϵ-DP model.

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

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

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

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

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

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

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

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

When $v = 1$, all the rates in Theorem 8 and 9 match the lower bounds in [Kamath *et al.*, 2022]. Thus, our results can be seen as extensions of the previous results. In Theorem 8, the gap between the rates in ϵ-DP and (ϵ, δ)-DP is $O(d^{\frac{v}{1+v}})$, while it is $O\left(d^{\frac{v}{2(1+v)}}\right)$ in Theorem 9. This is quite different with the case when the loss is Lipschitz [Bassily *et al.*, 2014]. To prove the lower bounds, we first reduce the problem to mean estimation, and then we use the private version of Fano's lemma in [Acharya *et al.*, 2021; Kamath *et al.*, 2022], based on the packing of distributions in [Barber and Duchi, 2014].

6 Differentially Private Sparse Learning with Heavy-tailed Data

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

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

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

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

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

Algorithm 4 Peeling$(v, s, \epsilon, \delta, \lambda)$[Cai *et al.*, 2021]

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

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

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

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

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

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

where the Big-O notation omits γ_r and μ_r.

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

Algorithm 5 Heavy-Tailed Private Sparse Optimization

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

Remark 3. For DP sparse learning with Lipschitz loss or regular data, [Wang and Xu, 2019] provided an upper bound of $\widetilde{O}(\frac{s^*}{n^2\epsilon^2})$. Moreover, for high dimensional sparse mean estimation and Generalized Linear Model (GLM) with the Lipschitz loss and sub-Gaussian data, [Cai *et al.*, 2020; Cai *et al.*, 2021] provided optimal rates of $\widetilde{O}\big(\frac{s^*\log d}{n} + \frac{(s^*\log d)^2}{(n\epsilon)^2}\big)$. We can see that compared with these results, the error bound now becomes $\widetilde{O}\big(\frac{(s^*)^{\frac{1+2v}{1+v}} u^{\frac{2}{1+v}}}{(n\epsilon)^{\frac{2v}{1+v}}}\big)$ due to data irregularity. When $v = 1$, the error bound now becomes to $\widetilde{O}\Big(\frac{(s^*)^{\frac{3}{2}} u}{(n\epsilon)}\Big)$, which matches the result in [Hu *et al.*, 2022]. Thus, our result can be seen as a generalization of the previous ones.

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

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

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

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

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

Acknowledgments

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

References

[Acharya *et al.*, 2018] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private testing of identity and closeness of discrete distributions. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*, pages 6879–6891, 2018.

[Acharya *et al.*, 2021] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private assouad, fano, and le cam. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory (ALT)*, 2021.

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

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

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

[Bassily *et al.*, 2019] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradee Thakurta. Private stochastic convex optimization with optimal rates. In *Proceedings of the 33rd Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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

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

[Cai *et al.*, 2021] T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021.

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

[Dwork *et al.*, 2006] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Theory of Cryptography Conference (TCC)*. Springer, 2006.

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

[Hu *et al.*, 2022] Lijie Hu, Shuo Ni, Hanshen Xiao, and Di Wang. High dimensional differentially private stochastic optimization with heavy-tailed data. In *Proceedings of the 41st ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS)*, 2022.

[Jain *et al.*, 2014] Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Proceedings of the 28th Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[Kamath *et al.*, 2020] Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. In *Proceedings of 33rd Conference on Learning Theory (COLT)*, pages 2204–2235. PMLR, 2020.

[Kamath *et al.*, 2022] Gautam Kamath, Xingtu Liu, and Huanyu Zhang. Improved rates for differentially private stochastic convex optimization with heavy-tailed data. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

[Liu *et al.*, 2021] Xiyang Liu, Weihao Kong, Sham Kakade, and Sewoong Oh. Robust and differentially private mean estimation. In *Proceedings of the 35th Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[Raskutti *et al.*, 2011] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57(10):6976–6994, 2011.

[Tao *et al.*, 2022] Youming Tao, Yulian Wu, Peng Zhao, and Di Wang. Optimal rates of (locally) differentially private heavy-tailed multi-armed bandits. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

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

[Vershynin, 2018] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[von Bahr and Esseen, 1965] Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the rth Absolute Moment of a Sum of Random Variables, $1 \leqq r \leqq 2$. *The Annals of Mathematical Statistics*, 36(1):299 – 303, 1965.

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

[Wang and Xu, 2019] Di Wang and Jinhui Xu. On sparse linear regression in the local differential privacy model. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

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

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

[Wang *et al.*, 2020b] Di Wang, Hanshen Xiao, Srinivas Devadas, and Jinhui Xu. On differentially private stochastic convex optimization with heavy-tailed data. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 10081–10091. PMLR, 2020.

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

A Some Useful Results

Lemma 4 (Tail Bound of Laplacian Vairable [Dwork *et al.*, 2006])**.** If $X \sim \text{Lap}(b)$, then

$$\mathbb{P}(|X| \geq t \cdot b) = \exp(-t).$$

Lemma 5 (Tail Bound of Gaussian Variable)**.** If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\mathbb{P}(X \geq \mu + t) \leq \exp(-\frac{t^2}{2\sigma^2}).$$

Lemma 6 (Bernstein's Inequality [Vershynin, 2018])**.** Let $X_1, \cdots X_n$ be n independent zero-mean random variables. Suppose $|X_i| \leq M$ and $\mathbb{E}[X_i^2] \leq s$ for all i. Then for any $t > 0$, we have

$$\mathbb{P}\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq t\} \leq \exp(-\frac{\frac{1}{2}t^2 n}{s + \frac{1}{3}Mt})$$

Lemma 7 (Chebyshev's Inequality)**.** Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with bounded $(1+v)$-th moment for some $v \in (0, 1]$. Then the following holds for any t,

$$\mathbb{P}_{x \sim \mathcal{D}}(\|x\|_2 > t) \leq \frac{\mathbb{E}[\|x\|_2^{1+v}]}{t^{1+v}}$$

Lemma 8 (Jensen's Inequality)**.** Let X be an integrable, real-valued random variable, and ψ be a convex function. Then

$$\psi(\mathbb{E}[X]) \leq \mathbb{E}[\psi(X)].$$

Lemma 9 (Holder's Inequality)**.** Let X, Y be random variables over $\mathbb{R}$, and let $k > 1$. Then,

$$\mathbb{E}[|XY|] \leq \left(\mathbb{E}\left[|X|^k\right]\right)^{\frac{1}{k}} \left(\mathbb{E}\left[|Y|^{\frac{k}{k-1}}\right]\right)^{\frac{k-1}{k}}.$$

Lemma 10. Given a random variable X with $\mathbb{E}[|X|^{1+v}] \leq u$ for some $v \in (0, 1]$, for any $B > 0$ we have

$$\mathbb{E}[X \mathbb{1}_{|X|>B}] \leq \frac{u}{B^v}.$$

Proof of Lemma 10. The the definition of expectation we have

$$\begin{aligned}
u \geq \mathbb{E}|X|^{1+v} &= \int_0^\infty (1+v)t^{1+v-1}\mathbb{P}(|X| > t)dt \\
&\geq \int_B^\infty t^v \mathbb{P}(|X| > t)dt \\
&\geq B^v \int_B^\infty \mathbb{P}(|X| > t)dt \\
&= B^v \int_0^\infty \mathbb{P}(X\mathbb{1}_{|X|>B} > t)dt \\
&= B^v \mathbb{E}[X\mathbb{1}_{|X|>B}]
\end{aligned}$$

□

Lemma 11. ([von Bahr and Esseen, 1965, Theorem 2]). Let $X_1, \cdots, X_n$ be *independent* random variables over $\mathbb{R}$ such that for $\forall i \in [n]$, $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[|X_i|^{1+v}] < \infty$, where $v \in (0, 1]$. Then

$$\mathbb{E}\left[\left|\sum\nolimits_{i=1}^{n} X_i\right|^{1+v}\right] \leq 2\sum\nolimits_{i=1}^{n} \mathbb{E}\left[|X_i|^{1+v}\right].$$

The lemma imples the following result:

Lemma 12. Let $x_1, \ldots, x_n \in \mathbb{R}$ be independent random variables such that for $\forall i \in [n]$, we have $\mathbb{E}[x_i] = 0$ and $\mathbb{E}[|x_i|^{1+v}] \leq M$ for some $v \in (0, 1]$. Then

$$\mathbb{E}\left[\left|\frac{\sum_{i=1}^{n} x_i}{n}\right|^{1+v}\right] \leq \frac{2M}{n^v}.$$

Proof. Note that $\mathbb{E}[x_i] = 0$ and x_i's are independent, according to Lemma 11, we have $\mathbb{E}[|\sum_{i=1}^n x_i|^{1+v}] \leq 2 \sum\limits_{i=1}^{n} \mathbb{E}[|x_i|^{1+v}] = 2nM$ for $v \in (0, 1]$. Thus we have

$$\mathbb{E}\left[\left|\frac{\sum_{i=1}^n x_i}{n}\right|^{1+v}\right] \leq \frac{2nM}{n^{1+v}} = \frac{2M}{n^v}$$

□

Lemma 13 (Relation between Raw Moment and Central Moment)**.** Let x be a random variable over $\mathbb{R}^d$ such that with mean μ and $(1+v)$-th raw moment bounded by a constant M, i.e., $\mathbb{E}[x] = \mu$ and $\mathbb{E}[|\langle x, \mathcal{X}\rangle|^{1+v}] < \infty$ where $\mathcal{X} \in \mathcal{S}^{d-1}$ is an arbitrary unit vector and $v \in (0, 1]$. Then we have

$$\mathbb{E}[|\langle x - \mu, \mathcal{X}\rangle|^{1+v}] \leq 4\mathbb{E}[|\langle x, \mathcal{X}\rangle|^{1+v}].$$

Moreover, when $\mathbb{E}[|\langle x - \mu, \mathcal{X}\rangle|^{1+v}] < \infty$ for an arbitrary unit vector $\mathcal{X}$ then we have

$$\mathbb{E}[|\langle x, \mathcal{X}\rangle|^{1+v}] \leq 2\mathbb{E}[(|\langle x - \mu, \mathcal{X}\rangle|^{1+v})] + 2\|\mu\|_2^{1+v}$$

Proof. When $\mathbb{E}[|\langle x, \mathcal{X}\rangle|^{1+v}] < \infty$ then

$$\begin{aligned}
\mathbb{E}[|\langle x - \mu, \mathcal{X}\rangle|^{1+v}] &= \mathbb{E}[|\langle x, \mathcal{X}\rangle - \langle \mu, \mathcal{X}\rangle|^{1+v}] \\
&\leq \mathbb{E}[2(|\langle x, \mathcal{X}\rangle|^{1+v} + |\langle \mu, \mathcal{X}\rangle|^{1+v})] \\
&= 2\mathbb{E}[|\langle x, \mathcal{X}\rangle|^{1+v}] + 2\mathbb{E}[|\langle \mu, \mathcal{X}\rangle|^{1+v}] \\
&= 2\mathbb{E}[|\langle x, \mathcal{X}\rangle|^{1+v}] + 2|\langle \mu, \mathcal{X}\rangle|^{1+v} \\
&= 2\mathbb{E}[|\langle x, \mathcal{X}\rangle|^{1+v}] + 2|\langle \mathbb{E}[x], \mathcal{X}\rangle|^{1+v} \\
&= 2\mathbb{E}[|\langle x, \mathcal{X}\rangle|^{1+v}] + 2|\mathbb{E}[\langle x, \mathcal{X}\rangle]|^{1+v} \\
&\leq 2\mathbb{E}[|\langle x, \mathcal{X}\rangle|^{1+v}] + 2\mathbb{E}[|\langle x, \mathcal{X}\rangle|^{1+v}] \\
&= 4\mathbb{E}[|\langle x, \mathcal{X}\rangle|^{1+v}],
\end{aligned}$$

where the first inequality is due to the inequality (6) of [von Bahr and Esseen, 1965], and the last inequality is due to Jensen's inequality (Lemma 8).

When $\mathbb{E}[|\langle x - \mu, \mathcal{X}\rangle|^{1+v}] < \infty$ then we have

$$\mathbb{E}[|\langle x, \mathcal{X}\rangle|^{1+v}] \leq 2\mathbb{E}[|\langle x - \mu, \mathcal{X}\rangle|^{1+v}] + 2\|\mu\|_2^{1+v}$$

□

Since the proofs will involves the private minimax risk, we first introduce the classical statistical minimax risk before discussing its private version. More details can be found in [Barber and Duchi, 2014].

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

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

where the supremum is taken over distributions $p \in \mathcal{P}$ and the infimum over all estimators $Q(X)$.

In the $(\epsilon, \delta)/\epsilon$-DP model, the estimator $Q(X)$ is obtained via some $(\epsilon, \delta)/\epsilon$-DP mechanism Q. Thus, we can also define the $(\epsilon, \delta)/\epsilon$-private minimax risk:

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

where $\mathcal{Q}$ is the set of all the $(\epsilon, \delta)/\epsilon$-DP mechanisms. When $\delta = 0$, we denote it as $\mathcal{M}_n(\theta(\mathcal{P}), Q, \Phi \circ \rho, \epsilon)$

Next, we recall two private Fano's Lemmas given in [Acharya *et al.*, 2021; Kamath *et al.*, 2022].

Lemma 14 (Theorem 2 in [Acharya *et al.*, 2021])**.** Consider a set of distributions $\mathcal{V} = \{p_1, p_2, \cdots, p_M\} \subseteq \mathcal{P}$ such that for all $i \neq j$,

- $\Phi(\rho(\theta(p_i), \theta_(p_j))) \geq \alpha$,

- $D_{KL}(p_i, p_j) \leq \beta$, where D_{KL} is the KL-divergence,
- $D_{TV}(p_i, p_j) \leq \gamma$,

then we have

$$\mathcal{M}_n(\theta(\mathcal{P}), Q, \Phi \circ \rho, \epsilon) \geq \frac{1}{M} \sum_{i \in [M]} \mathbb{E}_{X \sim p_i^n, Q}[\Phi(\rho(Q(X), \theta(p_i)))] \geq \max\{\frac{\alpha}{2}(1 - \frac{n\beta + \log 2}{\log M}), 0.4\alpha \min\{1, \frac{M}{e^{10\epsilon n \gamma}}\}\}. \quad (3)$$

Lemma 15. [Theorem 1.4 in [Kamath *et al.*, 2022]] In the case where $\epsilon \ll \log \frac{1}{\delta}$, consider a set of distributions $\mathcal{V} = \{p_1, p_2, \cdots, p_M\} \subseteq \mathcal{P}$ such that for all $i \neq j$,

- $\Phi(\rho(\theta(p_i), \theta_(p_j)) \geq \alpha$,
- $D_{KL}(p_i, p_j) \leq \beta$, where D_{KL} is the KL-divergence,
- $D_{TV}(p_i, p_j) \leq \gamma$,

then we have

$$\begin{aligned}\mathcal{M}_n(\theta(\mathcal{P}), Q, \Phi \circ \rho, \epsilon, \delta) &\geq \frac{1}{M} \sum_{i \in [M]} \mathbb{E}_{X \sim p_i^n, Q}[\Phi(\rho(Q(X), \theta(p_i)))] \\ &\geq \frac{\alpha}{2} \max\{1 - \frac{n\beta + \log 2}{\log M}, 1 - \frac{\frac{\epsilon^2}{4 \log \frac{1}{\delta}}(n^2\gamma^2 + n\gamma(1-\gamma)) + \log 2}{\log M}\}.\end{aligned}$$

Proof. The proof is directly followed by Theorem 1.4 in [Kamath *et al.*, 2022] where we know that each $\rho = (\sqrt{\log \frac{1}{\delta} + \epsilon} - \sqrt{\log \frac{1}{\delta}})^2$-zCDP is (ϵ, δ)-DP. Since $(\sqrt{\log \frac{1}{\delta} + \epsilon} - \sqrt{\log \frac{1}{\delta}})^2 \approx \frac{\epsilon^2}{4 \log \frac{1}{\delta}}$ when $\epsilon^2 \ll \log \frac{1}{\delta}$, we can get the proof. □

B Omitted Proofs in Section 4

Proof of Theorem 1. For the proof of DP, note that the sensitivity of $\widehat{\mu}$ is $\frac{2B}{n}$. Then the guarantee of DP follows directly from the Gaussian Mechanism. Next we focus on the utility. With probability at least $1 - 3\xi$, we have

$$\begin{aligned}|\widetilde{\mu} - \mu| &= \left| \frac{1}{n} \sum_{i=1}^n x_i \cdot \mathbf{1}_{|x_i| \leq B} + \nu - \mathbb{E}x \right| \\ &\leq \left| \frac{1}{n} \sum_{i=1}^n x_i \cdot \mathbf{1}_{|x_i| \leq B} - \mathbb{E}x \right| + |\nu| \\ &= \left| \frac{1}{n} \sum_{i=1}^n x_i \cdot \mathbf{1}_{|x_i| \leq B} - \mathbb{E}[x \cdot \mathbf{1}_{|x| \leq B}] + \mathbb{E}[x \cdot \mathbf{1}_{|x| \leq B}] - \mathbb{E}x \right| + |\nu| \\ &\leq \left| \frac{1}{n} \sum_{i=1}^n x_i \cdot \mathbf{1}_{|x_i| \leq B} - \mathbb{E}[x \cdot \mathbf{1}_{|x| \leq B}] \right| + \left| \mathbb{E}[x \cdot \mathbf{1}_{|x| \leq B}] - \mathbb{E}x \right| + |\nu| \\ &= \left| \frac{1}{n} \sum_{i=1}^n x_i \cdot \mathbf{1}_{|x_i| \leq B} - \mathbb{E}[x \cdot \mathbf{1}_{|x| \leq B}] \right| + \left| \mathbb{E}[x \cdot \mathbf{1}_{|x| > B}] \right| + |\nu| \\ &\leq \sqrt{\frac{2B^{1-v} u \log \frac{1}{\xi}}{n}} + \frac{B \log \frac{1}{\xi}}{3n} + \frac{u}{B^v} + \frac{4\sqrt{2}B}{n\epsilon} \sqrt{\log \frac{1.25}{\delta}} \sqrt{\log \frac{1}{\xi}}, \end{aligned} \quad (4)$$

where the last inequality is due to lemma 10, the lemma 5 that with probability at least $1 - 2\xi$,

$$|\nu| \leq \frac{4\sqrt{2}B}{n\epsilon} \sqrt{\log \frac{1.25}{\delta}} \sqrt{\log \frac{1}{\xi}},$$

and the lemma 6 that with probability at least $1 - \xi$,

$$\left| \frac{1}{n} \sum_{i=1}^n x_i \cdot \mathbf{1}_{|x_i| \leq B} - \mathbb{E}[X \cdot \mathbf{1}_{|X| \leq B}] \right| \leq \sqrt{\frac{2B^{1-v} u \log \frac{1}{\xi}}{n}} + \frac{B \log \frac{1}{\xi}}{3n}. \quad (5)$$

Since $B = \left(\frac{un\epsilon}{\log\frac{1}{\xi}\sqrt{\log\frac{1.25}{\delta}}}\right)^{\frac{1}{1+v}}$ and $\epsilon \leq 1$, we can bound each term in (4) as follows,

$$\begin{aligned}
\sqrt{\frac{2B^{1-v}u\log\frac{1}{\xi}}{n}} &= \sqrt{\frac{2u^{\frac{2}{1+v}}\epsilon^{\frac{1-v}{1+v}}\left(\log\frac{1}{\xi}\right)^{\frac{2v}{1+v}}}{n^{\frac{2v}{1+v}}\left(\sqrt{\log\frac{1.25}{\delta}}\right)^{\frac{1-v}{1+v}}}} \leq \sqrt{\frac{2u^{\frac{2}{1+v}}\left(\log\frac{1}{\xi}\right)^{\frac{2v}{1+v}}\left(\sqrt{\log\frac{1.25}{\delta}}\right)^{\frac{v}{1+v}}}{n^{\frac{2v}{1+v}}\epsilon^{\frac{v}{1+v}}}} \\
&\leq \sqrt{\frac{2u^{\frac{2}{1+v}}\left(\log\frac{1}{\xi}\right)^{\frac{2v}{1+v}}\left(\sqrt{\log\frac{1.25}{\delta}}\right)^{\frac{2v}{1+v}}}{n^{\frac{2v}{1+v}}\epsilon^{\frac{2v}{1+v}}}} \\
&= \sqrt{2}u^{\frac{1}{1+v}}\left(\frac{\sqrt{\log\frac{1.25}{\delta}}\log\frac{1}{\xi}}{n\epsilon}\right)^{\frac{v}{1+v}},
\end{aligned} \tag{6}$$

$$\frac{B\log\frac{1}{\xi}}{3n} = \frac{1}{3}u^{\frac{1}{1+v}}\frac{\left(\log\frac{1}{\xi}\right)^{\frac{v}{1+v}}\epsilon^{\frac{1}{1+v}}}{n^{\frac{v}{1+v}}\left(\sqrt{\log\frac{1.25}{\delta}}\right)^{\frac{1}{1+v}}} \leq \frac{1}{3}u^{\frac{1}{1+v}}\left(\frac{\sqrt{\log\frac{1.25}{\delta}}\log\frac{1}{\xi}}{n\epsilon}\right)^{\frac{v}{1+v}}, \tag{7}$$

$$\frac{u}{B^v} = u^{\frac{1}{1+v}}\left(\frac{\sqrt{\log\frac{1.25}{\delta}}\log\frac{2}{\xi}}{n\epsilon}\right)^{\frac{v}{1+v}}, \tag{8}$$

$$\begin{aligned}
\frac{4\sqrt{2}B}{n\epsilon}\sqrt{\log\frac{1.25}{\delta}}\sqrt{\log\frac{1}{\xi}} &= 4\sqrt{2}u^{\frac{1}{1+v}}\left(\frac{\sqrt{\log\frac{1.25}{\delta}}}{n\epsilon}\right)^{\frac{v}{1+v}}\left(\log\frac{1}{\xi}\right)^{\frac{v-1}{2(1+v)}} \\
&\leq 4\sqrt{2}u^{\frac{1}{1+v}}\left(\frac{\sqrt{\log\frac{1.25}{\delta}}\log\frac{1}{\xi}}{n\epsilon}\right)^{\frac{v}{1+v}}.
\end{aligned} \tag{9}$$

Based on (6), (7), (8), (9), we can conclude that, with probability at least $1 - 3\xi$,

$$|\widetilde{\mu} - \mu| \leq 9u^{\frac{1}{1+v}}\left(\frac{\sqrt{\log\frac{1.25}{\delta}}\log\frac{1}{\xi}}{n\epsilon}\right)^{\frac{v}{1+v}} \tag{10}$$

□

Proof of Theorem 2. First, we construct two distributions that are close and show that any (ϵ, δ)-DP algorithm that distinguishes between them requires a large number of samples. The result is shown in the following lemma.

Lemma 16. Let $\epsilon, \delta, \alpha > 0$. Suppose $\mathcal{D}_1$ is an one-point distribution over $\{0\}$ and $\mathcal{D}_2$ is an binomial distribution defined as follows:

$$\mathcal{D}_2 \triangleq \begin{cases} \tau, & \text{w. p.} \quad p \\ 0, & \text{o. w.} \end{cases}$$

where τ is a positive such that $p\tau = \alpha$ and $\alpha^{\frac{1+v}{v}} = pu^{\frac{1}{v}}$ (that is, $\tau = \left(\frac{u}{p}\right)^{\frac{1}{1+v}}$). Then the following holds.

1. The $((1+v))$-th raw moment of $\mathcal{D}_2$ is bounded by u, i.e., $\mathbb{E}_{X\sim\mathcal{D}_2}[|x|^{1+v}] \leq u$.
2. Any (ϵ, δ)-DP algorithm that distinguishes between $\mathcal{D}_1$ and $\mathcal{D}_2$ with a constant probability requires at least $\frac{1}{\epsilon\alpha^{\frac{1+v}{v}}}$ samples.

Proof. The first part follows from direct calculation.

$$\mathbb{E}_{X\sim\mathcal{D}_2}[|X|^{1+v}] = p\tau^{v+1} = u \tag{11}$$

Next we focus on the second part. Note that $|\mathbb{E}_{X\sim\mathcal{D}_1}[X] - \mathbb{E}_{X\sim\mathcal{D}_2}[X]| = \alpha$. Suppose we take n samples each from $\mathcal{D}_1$ and $\mathcal{D}_2$, then by Theorem 11 of [Acharya *et al.*, 2018], we know that any ϵ-DP algorithm that distinguishes between $\mathcal{D}_1$ and $\mathcal{D}_2$ with a constant error probability must satisfy

$$pn = \frac{\alpha^{\frac{1+v}{v}}}{u^{\frac{1}{v}}} n \in \Omega\left(\frac{1}{\epsilon}\right),$$

which gives

$$n \in \Omega\left(u^{\frac{1}{v}} \frac{1}{\epsilon\alpha^{\frac{1+v}{v}}}\right).$$

Thus, by using the equivalence of pure DP and approximate DP for testing problems (e.g., Lemma 2 and Lemma 3 of [Acharya *et al.*, 2018]), we conclude the proof. □

Next we back to the proof of Theorem 2. It can be concluded from Lemma 16 that, any (ϵ,δ)-DP algorithm takes at least $n \in \Omega\left(u^{\frac{1}{v}} \frac{1}{\epsilon\alpha^{\frac{1+v}{v}}}\right)$ samples to get an estimation $\widehat{\mu}$ for μ such that $|\widehat{\mu}-\mu| \leq \frac{\alpha}{2}$ with a constant probability. Finally, we show that, even if the algorithm takes enough samples, we must have $|\widehat{\mu}-\mu| \in \Omega\left(u^{\frac{1}{v+1}} (\frac{1}{n\epsilon})^{\frac{v}{1+v}}\right)$. If not, we have $|\widehat{\mu}-\mu| \in o\left(u^{\frac{1}{v+1}} (\frac{1}{n\epsilon})^{\frac{v}{1+v}}\right)$, i.e., $\alpha \in o\left(u^{\frac{1}{v+1}} (\frac{1}{n\epsilon})^{\frac{v}{1+v}}\right)$. Then according to $n \in \Omega\left(u^{\frac{1}{v+1}} \frac{1}{\epsilon\alpha^{\frac{1+v}{v}}}\right)$, we have $n \in \omega\left(\frac{1}{\epsilon((\frac{1}{n\epsilon})^{\frac{v}{1+v}})^{\frac{1+v}{v}}}\right) = \omega(n)$, which is impossible. Thus the lower bound concludes. □

C Omitted Proofs in Section 5

Proof of Theorem 3. For the proof of DP, we first bound the sensitivity of the non-private $\widehat{\mu}$. Fixing a dimension $j \in [d]$, for two neighbouring dataset X and X', we have $|\widehat{\mu}_j(X) - \widehat{\mu}_j(X')| \leq \frac{2\tau m}{n}$. Therefore, the ℓ_2 sensitivity of $\widehat{\mu}$ is upper bounded by $\frac{2\tau m\sqrt{d}}{n}$. Thus, by the Gaussian mechanism, Theorem 3 is (ϵ,δ)-DP.

For the upper bound, we first show the accuracy guarantees of the non-private estimator $\widehat{\mu}$. We analyze the algorithm coordinatewisely. For a fixed dimension $j \in [d]$ and a fixed sample batch k,

$$\begin{aligned}
\left|\widehat{\mu}_j^k - \mu_j\right| &= \left|\frac{m}{n}\sum_{i=(k-1)\cdot\frac{n}{m}+1}^{k\cdot\frac{n}{m}} [x_i']_j - \mu_j\right| \\
&= \left|\frac{m}{n}\sum_{i=(k-1)\cdot\frac{n}{m}+1}^{k\cdot\frac{n}{m}} [x_i]_j \cdot \mathbf{1}_{|[x_i]_j|\leq\tau} - \mathbb{E}[[x]_j]\right| \\
&= \left|\frac{m}{n}\sum_{i=(k-1)\cdot\frac{n}{m}+1}^{k\cdot\frac{n}{m}} [x_i]_j \cdot \mathbf{1}_{|[x_i]_j|\leq\tau} - \mathbb{E}[[x]_j \cdot \mathbf{1}_{|[x]_j|\leq\tau}] + \mathbb{E}[[x]_j \cdot \mathbf{1}_{|[x]_j|\leq\tau}] - \mathbb{E}[[x]_j]\right| \\
&\leq \left|\frac{m}{n}\sum_{i=(k-1)\cdot\frac{n}{m}+1}^{k\cdot\frac{n}{m}} [x_i]_j \cdot \mathbf{1}_{|[x_i]_j|\leq\tau} - \mathbb{E}[[x]_j \cdot \mathbf{1}_{|[x]_j|\leq\tau}]\right| + \left|\mathbb{E}[[x]_j \cdot \mathbf{1}_{|[x]_j|\leq\tau}] - \mathbb{E}x\right| \\
&= \left|\frac{m}{n}\sum_{i=(k-1)\cdot\frac{n}{m}+1}^{k\cdot\frac{n}{m}} [x_i]_j \cdot \mathbf{1}_{|[x_i]_j|\leq B} - \mathbb{E}[[x]_j \cdot \mathbf{1}_{|[x]_j|\leq\tau}]\right| + \left|\mathbb{E}[[x]_j \cdot \mathbf{1}_{|[x]_j|>\tau}]\right|.
\end{aligned} \tag{12}$$

The first term in (12) can be bounded According to Theorem 4 as

$$\mathbb{P}\left(\left|\frac{m}{n}\sum_{i=(k-1)\cdot\frac{n}{m}+1}^{k\cdot\frac{n}{m}} [x_i]_j \cdot \mathbf{1}_{|[x_i]_j|\leq B} - \mathbb{E}[[x]_j \cdot \mathbf{1}_{|[x]_j|\leq B}]\right| \leq (80u)^{\frac{1}{1+v}}\left(\frac{m}{n}\right)^{\frac{v}{1+v}}\right) \geq 0.9 \tag{13}$$

The second term in (12) is bounded in Lemma 10 such that

$$\mathbb{E}[[x]_j \mathbb{1}_{|[x]_j|>\tau}] \leq \frac{u}{\tau^v}. \tag{14}$$

For batch k, denote $\mathcal{E}_k$ as the event that

$$\left|\frac{m}{n}\sum_{i=(k-1)\cdot\frac{n}{m}+1}^{k\cdot\frac{n}{m}}[x_i]_j\cdot\mathbf{1}_{|[x_i]_j|\le B}-\mathbb{E}[[x]_j\cdot\mathbf{1}_{|[x]_j|\le B}]\right|\le(80u)^{\frac{1}{1+v}}\left(\frac{m}{n}\right)^{\frac{v}{1+v}}$$

Note that $\widehat{\mu}_j=\operatorname{median}(\widehat{\mu}_j^1,\cdots,\widehat{\mu}_j^m)$. Suppose $\widehat{\mu}_j=\widehat{\mu}_j^{k_0}$, $k_0\in[m]$, event $\mathcal{E}_{k_0}$ happens if and only if at least a half events in $\{\mathcal{E}_k\}_{k=1}^m$ happens. By Hoeffding's inequality,

$$\mathbb{P}\left(\left|\frac{m}{n}\sum_{i=(k_0-1)\cdot\frac{n}{m}+1}^{k_0\cdot\frac{n}{m}}[x_i]_j\cdot\mathbf{1}_{|[x_i]_j|\le B}-\mathbb{E}[[x]_j\cdot\mathbf{1}_{|[x]_j|\le B}]\right|\le(80u)^{\frac{1}{1+v}}\left(\frac{m}{n}\right)^{\frac{v}{1+v}}\right)\le e^{-\frac{m}{4}}. \tag{15}$$

Apply the union bound to all the dimensions, and combine the result with (14), we get

$$\mathbb{P}\left(\|\widehat{\mu}-\mu\|_2\ge\sqrt{d}\left((80u)^{\frac{1}{1+v}}\left(\frac{m}{n}\right)^{\frac{v}{1+v}}+\frac{u}{\tau^v}\right)\right)\le d\cdot e^{-\frac{m}{4}}\le\frac{\xi}{2}. \tag{16}$$

Next, we consider the private estimator $\widetilde{\mu}$. Since the noise $\nu\sim\mathcal{N}\left(0,\sigma^2\right)$, where $\sigma^2=\frac{8\tau^2m^2d}{\epsilon^2n^2}\ln\frac{1.25}{\delta}\cdot\mathbb{I}_{d\times d}$, by the tail property of chi-squared distribution,

$$\mathbb{P}\left(\|\nu\|_2\ge2\sigma\left(\sqrt{d}+\sqrt{\log\left(\frac{1}{\xi}\right)}\right)\right)\le\frac{\xi}{2}. \tag{17}$$

Note that $\|\widetilde{\mu}-\mu\|_2\le\|\widehat{\mu}-\mu\|+\|\nu\|_2$, we conclude the proof by the union bound. □

Proof of Theorem 4. By Lemma 13, we have $\mathbb{E}[|x_i-\mu|^{1+v}]\le4M$. According to Lemma 12, we have

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n x_i-\mu\right|^{1+v}\right]\le\frac{8M}{n^v}. \tag{18}$$

Follow Chebyshev's inequality, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n x_i-\mu\right|\ge t\right)\le\frac{\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n x_i-\mu\right|^{1+v}\right]}{t^{1+v}} \tag{19}$$

$$\le\frac{8M}{n^v}\frac{1}{t^{1+v}}. \tag{20}$$

Set $t=\left(\frac{80M}{n^v}\right)^{\frac{1}{1+v}}$, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n x_i-\mu\right|\ge\left(\frac{80M}{n^v}\right)^{\frac{1}{1+v}}\right)\le\frac{1}{10} \tag{21}$$

□

Proof of Theorem 5. By the composition property of DP (Lemma 1), Algorithm 2 preserves (ϵ,δ)-DP. □

C.1 Proof Upper Bounds

Algorithm 6 SCO Framework $\mathrm{SCOF}_{\eta,T,\mathrm{MeanEstimator}(X)}$

```
Input: Data samples $X=\{x_i\}_{i=1}^n, x_i\in\mathbb{R}$, algorithm MeanEstimator, parameters $\eta, T$.
Output: Iteratives $\{w_1,w_2,\cdots,w_T\}$.
for $t\leftarrow 1,\cdots,T$ do
    $\nabla\widehat{L}_{\mathcal{D}}(w_{t-1})=\mathrm{MeanEstimator}(\{\nabla\ell(w_{t-1},x)\}_{x\in X_t})$
    $w_t=\mathrm{Proj}_{\mathcal{W}}(w_{t-1}-\eta\nabla\widehat{L}_{\mathcal{D}}(w_{t-1}))$
end for
return $\{w_1,w_2,\cdots,w_T\}$
```

Our idea of proof follows [Kamath *et al.*, 2022]. To obtain the accuracy guarantee of Algorithm 2, we first recall the the relationship between the excess population risk of general SCO framework (shown in Algorithm 6) and the accuracy of MeanEstimator it uses. We summarize the results in Lemma 17 and 18 for general convex population risk functions and the population risk functions that are both strongly convex and smooth respectively.

Lemma 17. (Convex[Kamath *et al.*, 2022, Theorem 3.1]) Suppose that MeanEstimator guarantees that, for any $w \in \mathcal{W}$, $\|\mathbb{E}[\nabla \widetilde{L}_{\mathcal{D}}(w)] - \nabla L_{\mathcal{D}}(w)\|_2 \leq \mathcal{B}$ and $\mathbb{E}[\|\nabla \widetilde{L}_{\mathcal{D}}(w) - \nabla L_{\mathcal{D}}(w)\|_2^2] \leq \mathcal{V}^2$. Under the Assumption 1, for any $\eta > 0$ the output $w^{\text{priv}} = \frac{1}{T}\sum_{t\in[T]} w_t$ satisfies

$$\mathbb{E}_{X\sim\mathcal{D}^n,\mathcal{A}}[L_{\mathcal{D}}(w^{\text{priv}}) - L_{\mathcal{D}}(w^*)] \leq \frac{\Delta^2}{2\eta T} + \frac{\eta R^2}{2} + \frac{\eta \mathcal{V}^2}{2} + \Delta\mathcal{B}.$$

where $w^* = \arg\min_w L_{\mathcal{D}}(w)$.

Lemma 18. (Strongly Convex[Kamath *et al.*, 2022, Theorem 3.2]) Suppose that MeanEstimator guarantees that, for any $w \in \mathcal{W}$, $\mathbb{E}[\|\nabla \widetilde{L}_{\mathcal{D}}(w) - \nabla L_{\mathcal{D}}(w)\|_2] \leq \mathcal{V}$. Under Assumption 1, and the further assumption that the population risk function $L_{\mathcal{D}}(\cdot)$ is α-strongly convex and β-smooth, if $\eta = \frac{1}{\alpha+\beta}$, the output $w^{\text{priv}} = w_T$ satisfies

$$\mathbb{E}_{X\sim\mathcal{D}^n,\mathcal{A}}[L_{\mathcal{D}}(w^{\text{priv}}) - L_{\mathcal{D}}(w^*)] \leq (1 - \frac{\alpha+\beta}{(\alpha+\beta)^2})^T\Delta + \frac{(\alpha+\beta)\mathcal{V}}{\alpha\beta}$$

Specifically, if we set $T = \log(\frac{(\alpha+\beta)\mathcal{V}}{\alpha\beta})/\log(\frac{\alpha^2+\beta^2+\alpha\beta}{(\alpha^2+\beta^2)})$, the output w^{priv} satisfies

$$\mathbb{E}_{X\sim\mathcal{D}^n,\mathcal{A}}[L_{\mathcal{D}}(w^{\text{priv}}) - L_{\mathcal{D}}(w^*)] \leq \frac{(\alpha+\beta)^2(\Delta+1)^2\mathcal{V}^2}{2\alpha^2\beta^2}$$

where $w^* = \arg\min_w L_{\mathcal{D}}(w)$.

To provide the performance guarantee on our DP-SCO algorithm, we need to obtain bounds for the gradient estimator that hold uniformly over the choice of $w \in \mathcal{W}$. Specifically, we have the following results.

Lemma 19. Consider our Algorithm 2, the following holds for all $w \in \mathcal{W}$ simultaneously:

1. $\left\|\mathbb{E}\left[\nabla \widetilde{L}_{\mathcal{D}}(w)\right] - \nabla L_{\mathcal{D}}(w)\right\|_2 \leq \widetilde{O}\left(u^{\frac{1}{1+v}}\frac{d^{\frac{1+3v}{2(1+v)}}}{n^{\frac{v}{1+v}}} + u\frac{\sqrt{d}}{\tau^v}\right)$
2. $\mathbb{E}\left[\left\|\nabla \widetilde{L}_{\mathcal{D}}(w) - \nabla L_{\mathcal{D}}(w)\right\|_2^2\right] \leq \widetilde{O}\left(\frac{\tau^2 d^4 T}{\epsilon^2 n^2} + u^{\frac{2}{1+v}}\frac{d^{\frac{1+3v}{1+v}}}{n^{\frac{2v}{1+v}}} + u^2\frac{d}{\tau^{2v}}\right)$

Proof. We start with the proving the first part. By the law of total expectation, we have $\mathbb{E}\left[\nabla \widetilde{L}_{\mathcal{D}}(w)\right] = \mathbb{E}\left[\mathbb{E}\left[\nabla \widetilde{L}_{\mathcal{D}}(w)\middle|\nabla \widehat{L}_{\mathcal{D}}(w)\right]\right] = \mathbb{E}\left[\nabla \widehat{L}_{\mathcal{D}}(w)\right]$. Thus, we only need to focus on the non-private $\nabla \widehat{L}_{\mathcal{D}}(w)$. In order to obtain the bounds that hold uniformly over the parameter space $\mathcal{W}$, we follow a standard covering net argument. Suppose the parameter space $\mathcal{W}$ is covered by a set of balls with radius γ. Then number of balls, denoted by N_γ, is upper bounded by $\left(\frac{\Delta}{\gamma}\right)^d$, where Δ is the diameter of $\mathcal{W}$. Let $\mathcal{W}_\gamma = \{\widetilde{w}_1, \cdots, \widetilde{w}_{N_\gamma}\}$ denotes the centers of this covering. For an arbitrary $w \in \mathcal{W}$, there exists some $\widetilde{w} \in \mathcal{W}_\gamma$ such that $\|w - \widetilde{w}\|_2 \leq \gamma$. Then,

$$\begin{aligned}\left\|\mathbb{E}\left[\nabla \widetilde{L}_{\mathcal{D}}(w)\right] - \nabla L_{\mathcal{D}}(w)\right\|_2 &= \left\|\mathbb{E}\left[\nabla \widehat{L}_{\mathcal{D}}(w)\right] - \nabla L_{\mathcal{D}}(w)\right\|_2 \leq \mathbb{E}\left[\left\|\nabla \widehat{L}_{\mathcal{D}}(w) - \nabla L_{\mathcal{D}}(w)\right\|_2\right] \\ &\leq \mathbb{E}\left[\left\|\nabla \widehat{L}_{\mathcal{D}}(w) - \nabla \widehat{L}_{\mathcal{D}}(\widetilde{w})\right\|_2 + \left\|\nabla \widehat{L}_{\mathcal{D}}(\widetilde{w}) - \nabla L_{\mathcal{D}}(\widetilde{w})\right\|_2 + \|\nabla L_{\mathcal{D}}(\widetilde{w}) - \nabla L_{\mathcal{D}}(w)\|_2\right].\end{aligned}$$

We bound each term in the last inequality above respectively. For the first term, we need to analyze how much the output of the non-private estimator $\nabla \widehat{L}_{\mathcal{D}}(w)$ changes when the input switches from w to $\widetilde{w}$. According to the β-smooth assumption, for each dimension $j \in [d]$ and batch $k \in [m]$, the average differs no more than $\beta\gamma$. For each dimension j, the median differ no more than $\beta\gamma$. Summing over all the dimensions,

$$\left\|\nabla \widehat{L}_{\mathcal{D}}(w) - \nabla \widehat{L}_{\mathcal{D}}(\widetilde{w})\right\|_2 \leq \beta\gamma \cdot \sqrt{d}.$$

For the second term, according to Theorem 3, with probability at least $1 - \xi$,

$$\left\|\nabla \widehat{L}_{\mathcal{D}}(\widetilde{w}) - \nabla L_{\mathcal{D}}(\widetilde{w})\right\|_2 \leq O\left(\sqrt{d}\left(u^{\frac{1}{1+v}}\left(\frac{\log\frac{d}{\xi}}{n}\right)^{\frac{v}{1+v}} + \frac{u}{\tau^v}\right)\right).$$

Let $\xi = \left(\frac{\gamma}{\Delta}\right)^{2d}$, by union bound, with probability at least $1 - \xi \cdot N_\gamma \geq 1 - \left(\frac{\gamma}{\Delta}\right)^d$, for all $\widetilde{w} \in \mathcal{W}_\gamma$,

$$\|\nabla \widehat{L}_{\mathcal{D}}(\widetilde{w}) - \nabla L_{\mathcal{D}}(\widetilde{w})\|_2 \leq O\left(\sqrt{d}\left(u^{\frac{1}{1+v}}\left(\frac{\log\frac{d}{\xi}}{n}\right)^{\frac{v}{1+v}} + \frac{u}{\tau^v}\right)\right).$$

Taking expectation, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla \widehat{L}_{\mathcal{D}}(\widetilde{w}) - \nabla L_{\mathcal{D}}(\widetilde{w})\right\|_2\right] &\leq O\left(\sqrt{d}\left(u^{\frac{1}{1+v}}\left(\frac{\log \frac{d}{\xi}}{n}\right)^{\frac{v}{1+v}} + \frac{u}{\tau^v}\right)\right) + R \cdot \left(\frac{\gamma}{\Delta}\right)^d \\ &\leq O\left(\sqrt{d}\left(u^{\frac{1}{1+v}}\left(\frac{\log \frac{d}{\xi}}{n}\right)^{\frac{v}{1+v}} + \frac{u}{\tau^v}\right) + \gamma^d\right),\end{aligned}$$

where we assume R and Δ are constants. For the third term, by the smoothness assumption,

$$\|\nabla L_{\mathcal{D}}(\widehat{w}) - \nabla L_{\mathcal{D}}(w)\|_2 \leq \beta\gamma.$$

Summing up all three terms and taking $\gamma = \frac{1}{n^{\frac{v}{1+v}}}$, we have

$$\left\|\mathbb{E}\left[\nabla \widetilde{L}_{\mathcal{D}}(w)\right] - \nabla L_{\mathcal{D}}(w)\right\|_2 \leq \mathbb{E}\left[\left\|\nabla \widehat{L}_{\mathcal{D}}(w) - \nabla L_{\mathcal{D}}(w)\right\|_2\right] \leq \widetilde{O}\left(u^{\frac{1}{1+v}}\frac{d^{\frac{1+3v}{2(1+v)}}}{n^{\frac{v}{1+v}}} + u\frac{\sqrt{d}}{\tau^v}\right). \tag{22}$$

Next, we prove the second part. By the Cauchy-Schwartz inequality, we have

$$\mathbb{E}\left[\left\|\nabla \widetilde{L}_{\mathcal{D}}(w) - \nabla L_{\mathcal{D}}(w)\right\|_2^2\right] \leq 2\mathbb{E}\left[\left\|\nabla \widetilde{L}_{\mathcal{D}}(w) - \nabla \widehat{L}_{\mathcal{D}}(w)\right\|_2^2\right] + 2\mathbb{E}\left[\left\|\nabla \widehat{L}_{\mathcal{D}}(w) - \nabla L_{\mathcal{D}}(w)\right\|_2^2\right].$$

According to the Gaussian mechanism, we know that $\mathbb{E}\left[\left\|\nabla \widetilde{L}_{\mathcal{D}}(w) - \nabla \widehat{L}_{\mathcal{D}}(w)\right\|_2^2\right] \leq \frac{8\tau^2 m^2 d^2}{\epsilon'^2 n^2} \ln \frac{1.25}{\delta'} = \frac{1024\tau^2\left[\log(2d) + 2d\log\left(\Delta n^{\frac{v}{1+v}}\right)\right]^2 d^2 T \log\frac{2.5T+2}{\delta}}{\epsilon^2 n^2}$, and by (22), we have

$$\mathbb{E}\left[\left\|\nabla \widetilde{L}_{\mathcal{D}}(w) - \nabla L_{\mathcal{D}}(w)\right\|_2^2\right] \leq \widetilde{O}\left(\frac{\tau^2 d^4 T}{\epsilon^2 n^2} + u^{\frac{2}{1+v}}\frac{d^{\frac{1+3v}{1+v}}}{n^{\frac{2v}{1+v}}} + u^2\frac{d}{\tau^{2v}}\right).$$

□

Based on the results above, we are able to show the upper bounds for Algorithm 2 by the appropriate choice of τ, η and T.

Proof of Theorem 6.

$$\begin{aligned}&\frac{\Delta^2}{2\eta T} = O\left(\Delta\frac{d^{\frac{1+4v}{2+2v}}}{(\epsilon n)^{\frac{v}{1+v}}}\right), \\ &\frac{\eta}{2}R^2 = O\left(\Delta\frac{d^{\frac{1+4v}{2+2v}}}{(\epsilon n)^{\frac{v}{1+v}}}\right), \\ &\Delta B \leq \widetilde{O}\left(\Delta u^{\frac{1}{1+v}}\frac{d^{\frac{1+3v}{2+2v}}}{n^{\frac{v}{1+v}}} + \Delta u\frac{d^{\frac{1+4v}{2+2v}}}{(\epsilon n)^{\frac{v}{1+v}}}\right) \leq \widetilde{O}\left(\Delta \max\{u, u^{\frac{1}{1+v}}\}\frac{d^{\frac{1+4v}{2+2v}}}{(\epsilon n)^{\frac{v}{1+v}}}\right), \\ &\frac{\eta\mathcal{V}^2}{2} \leq \widetilde{O}\left(\Delta\frac{d^{\frac{1+4v}{2+2v}}}{(\epsilon n)^{\frac{v}{1+v}}} + \frac{\Delta u^{\frac{2}{1+v}}}{R^2}\frac{d^{\frac{3+10v}{2+2v}}}{\epsilon^{\frac{v}{1+v}} n^{\frac{3v}{1+v}}} + \frac{\Delta u^2}{R^2}\frac{d^{\frac{3+12v}{2+2v}}}{(\epsilon n)^{\frac{3v}{1+v}}}\right) \leq \widetilde{O}\left(\Delta \max\{u^2, u^{\frac{2}{1+v}}\}(\frac{d^{\frac{1+4v}{2+2v}}}{(\epsilon n)^{\frac{v}{1+v}}} + \frac{d^{\frac{3+12v}{2+2v}}}{(\epsilon n)^{\frac{3v}{1+v}}})\right).\end{aligned}$$

Combining all the terms together, we have

$$\mathop{\mathbb{E}}_{X \sim \mathcal{D}^n, \mathcal{A}}\left[L_{\mathcal{D}}\left(w^{priv}\right) - L_{\mathcal{D}}\left(w^*\right)\right] \leq \widetilde{O}\left(\Delta\frac{d^{\frac{3+12v}{2+2v}}}{(\epsilon n)^{\frac{v}{1+v}}}\right).$$

The proof then concludes by applying the results of Lemma 17. □

Proof of Theorem 7. Setting $\xi = 0.1$ and $\tau = \left(\frac{u\epsilon n}{\sqrt{d}}\right)^{\frac{1}{1+v}}$ in Theorem 3 gives the following corollary.

Corollary 1. Taking $\tau = \left(\frac{u\epsilon n}{\sqrt{d}}\right)^{\frac{1}{1+v}}$ in Algorithm 3, then with probability at least 0.9, the output $\widetilde{\mu}$ of Algorithm 3 satisfies

$$\|\widetilde{\mu} - \mu\|_2 = \widetilde{O}\Bigg(u^{\frac{1}{1+v}}\sqrt{d}\Big(\frac{1}{n}\Big)^{\frac{v}{1+v}} + u^{\frac{1}{1+v}}\frac{d^{\frac{1+2v}{2+2v}}}{(\epsilon n)^{\frac{v}{1+v}}}\Bigg). \tag{23}$$

For the DP-SCO in the strongly convex setting, for each iteration, the input of the mean estimator DPHDME is disjoint and independent, with size of $\frac{n}{T}$. Due to the choice of τ in Algorithm 2, the Corollary 1 immediately guarantees the following accuracy for Algorithm 2.

Lemma 20. Consider the Algorithm 2 with $\tau = \left(\frac{u\epsilon n}{\sqrt{d}T}\right)^{\frac{1}{1+v}}$. Under Assumption 1, the following holds for all w_t, $t \in [T]$:

$$\mathbb{E}[\nabla \widetilde{L}_{\mathcal{D}}(w_t) - \nabla L_{\mathcal{D}}(w_t)] \le \widetilde{O}\left(u^{\frac{1}{1+v}}\sqrt{d}\left(\frac{T}{n}\right)^{\frac{v}{1+v}} + u^{\frac{1}{1+v}} d^{\frac{1+2v}{2+2v}}\left(\frac{T}{\epsilon n}\right)^{\frac{v}{1+v}}\right). \tag{24}$$

Note that T is poly-logarithmic on n and d. The proof of the theorem follows directly from the results of Lemma 18 □

C.2 Proof of Lower Bounds

Proof of Theorem 8. We first prove the theorem by studying the lower bound for DP mean estimation, which is shown in Lemma 21. Then we show a reduction from DP mean estimation to DP-SCO.

Lemma 21. For any given $v \in (0, 1]$ and ϵ-DP algorithm $\mathcal{A}$, there exists a distribution $\mathcal{D}$ over $\mathbb{R}^d$ with $\|\mathbb{E}_{\mathcal{D}}[x] = \mu\|_2 \le 1$ and $\sup_{j\in[d]} \mathbb{E}_{x\sim\mathcal{D}}[|x_j|^{1+v}] \le u$, such that when $n \ge \Omega(u^{\frac{1}{v}} d^{\frac{1+3v}{2v}}/\epsilon)$,

$$\mathbb{E}_{x\sim\mathcal{D},\mathcal{A}}[\|\mathcal{A}(x) - \mu\|_2] \ge \Omega\left(\sqrt{d}u^{\frac{1}{1+v}}\left(\frac{d}{\epsilon n}\right)^{\frac{v}{1+v}}\right).$$

By Jensen's inequality, the above yields

$$\mathbb{E}_{x\sim\mathcal{D},\mathcal{A}}[\|\mathcal{A}(x) - \mu\|_2^2] \ge \Omega\left(u^{\frac{2}{1+v}} d\left(\frac{d}{\epsilon n}\right)^{\frac{2v}{1+v}}\right).$$

For any (ϵ, δ)-DP algorithm $\mathcal{A}$ with $\epsilon \ll \log\frac{1}{\delta}$, there exists a distribution $\mathcal{D}$ with $\|\mathbb{E}_{\mathcal{D}}[x_j] = \mu\|_2 \le u$ and $\sup_{j\in[d]} \mathbb{E}_{x\sim\mathcal{D}}[|x_j|^{1+v}] \le u$, such that when $n \ge \Omega(u^{\frac{1}{v}}\sqrt{\log\frac{1}{\delta}}d^{\frac{1+2v}{2v}}/\epsilon)$

$$\mathbb{E}_{x\sim\mathcal{D},\mathcal{A}}[\|\mathcal{A}(x) - \mu\|_2^2] \ge \Omega\left(u^{\frac{2}{1+v}} d\left(\frac{\sqrt{d\log\frac{1}{\delta}}}{\epsilon n}\right)^{\frac{2v}{1+v}}\right).$$

Proof. We adopt the packing argument from [Barber and Duchi, 2014]. A binary *code* of length d is a set $\Theta \subseteq \{0, 1\}^d$ and each $\theta \in \Theta$ is a *codeword*. The weight of a binary codeword θ is the number of 1's in θ, i.e., $|\{i : \theta_i = 1\}|$. We say a binary code is a constant weight code if each $\theta \in \Theta$ has the same weight. Given $\theta \in \Theta$, let $Q_\theta = (1-p)P_0 + pP_\theta$ for some $p \in (0, 1]$, where P_0 is a point mass on $\{D = 0\}$ and P_θ is a point mass on $\left\{D = (u/p)^{\frac{1}{1+v}}\theta\right\}$. Given Q_θ, let $\mu_\theta \in \mathbb{R}^d$ be the mean of Q_θ, i.e., $\mu_\theta = \mathbb{E}_{x\sim Q_\theta}[x]$. According to the Gilbert-Varshamov bound for constant-weight codes (see, e.g., [Acharya *et al.*, 2021]), there exists a code Θ such that,

- The cardinality of Θ satisfies $|\Theta| \ge 2^{\frac{7}{128}d}$.
- For all $\theta \in \Theta$, $\theta \in \{0, 1\}^d$ with $\|\theta\|_1 = \frac{d}{2}$.
- For all $\theta_1, \theta_2 \in \Theta$, $d_{\mathrm{Ham}}(\theta_1, \theta_2) = \sum_{j=1}^{d} \mathbb{I}[(\theta_1)_j \ne (\theta_2)_j] \ge \frac{d}{8}$.

We first compute the norm of μ_θ. Note that for all $\theta \in \Theta$, $\|\mu_\theta\|_2$ is the same, which is denoted by ι.

$$\|\mu_\theta\|_2 = \|\mathbb{E}_{x\sim Q_\theta}[x]\|_2 = u^{\frac{1}{1+v}} p^{\frac{v}{1+v}}\sqrt{\frac{d}{2}} \triangleq \iota.$$

Let $x \sim Q_\theta$ we have

$$\sup_{j\in[d]} \mathbb{E}_{x\sim Q_\theta}[|x_j|^{1+v}] \le p \cdot \left((u/p)^{\frac{1}{1+v}}\right)^{1+v} = u.$$

Now, we are able to bound the error. Define the family $\mathcal{D}_v(1)$ of heavy-tailed distributions supported on $\mathbb{R}^d$ by

$$\mathcal{D}_v(1) \triangleq \{\mathcal{D}|\operatorname{supp}\mathcal{D} \subseteq \mathbb{R}^d \quad \text{and} \quad \mathbb{E}_{x\sim\mathcal{D}}[|x_j|^{1+v}] \le 1\forall j \in [d]\}.$$

Note that $|\Theta| \geq 2^{\frac{7}{128}d}$. Furthermore, $\forall \theta \neq \theta'$, $\|\mu_\theta - \mu_{\theta'}\|_2 \geq 2\iota$ and $d_{\mathrm{TV}}(Q_\mu, Q_{\mu'}) = p$. Thus, by Lemma 14 with $\Phi(x) = x$ and ρ as the ℓ_2-norm difference we have

$$\mathcal{M}_n(\theta(\mathcal{D}_v(1)), Q, \|\cdot\|_2, \epsilon) \geq \frac{1}{|\Theta|}\sum_{\theta\in\Theta}\mathbb{E}_{X\sim Q_\theta^n, Q}[\|(Q(X) - \mu_\theta\|_2] \geq \Omega(\iota\min\{1, \frac{|\Theta|}{e^{\epsilon np}}) = \Omega(u^{\frac{1}{1+v}}\sqrt{d}p^{\frac{v}{1+v}}\min\{1, \frac{|\Theta|}{e^{\epsilon np}}\}). \tag{25}$$

Take $p = \Omega(\min\{1, \frac{d}{n\epsilon}\})$ we have the result. Next we just enforce $\|\mu_\theta\|_2 = \iota \leq 1$. This holds when $n \geq \Omega(u^{\frac{1}{v}}d^{\frac{1+3v}{2v}}/\epsilon)$.

For (ϵ, δ)-DP, by Lemma 15 we have

$$\mathcal{M}_n(\theta(\mathcal{D}_v(1)), Q, \|\cdot\|_2, \epsilon, \delta) \geq \frac{1}{|\Theta|}\sum_{\theta\in\Theta}\mathbb{E}_{X\sim Q_\theta^n, Q}[\|(Q(X) - \mu_\theta\|_2] \geq \Omega(u^{\frac{1}{1+v}}\sqrt{d}p^{\frac{v}{1+v}}(1 - \frac{\frac{\epsilon^2}{4\log\frac{1}{\delta}}(n^2p^2 + np(1-p)) + \log 2}{\log|\Theta|}). \tag{26}$$

Take $p = \Omega(\min\{1, \frac{\sqrt{d\log\frac{1}{\delta}}}{n\epsilon}\})$ we have the result. Next we just enforce $\|\mu_\theta\|_2 = \iota \leq 1$. This holds when $n \geq \Omega(u^{\frac{1}{v}}\sqrt{\log\frac{1}{\delta}}d^{\frac{1+2v}{2v}}/\epsilon)$. □

Now we back our proof. We focus on the (ϵ, δ)-DP (it is the same for $\delta = 0$). By Lemma 21 we know that there exists a distribution $\mathcal{D}$ with $\mathbb{E}_{\mathcal{D}}[x] = \mu$, $\|\mu\|_2 \leq 1$ and $\sup_{j\in[d]}\mathbb{E}_{x\sim\mathcal{D}}[|x|^{1+v}] \leq u$, such that

$$\mathbb{E}_{x\sim\mathcal{D},\mathcal{A}}[\|\mathcal{A}(x) - \mu\|_2^2] \geq \Omega\left(u^{\frac{2}{1+v}}d\left(\frac{\sqrt{d\log\frac{1}{\delta}}}{\epsilon n}\right)^{\frac{2v}{1+v}}\right).$$

For this distribution $\mathcal{D}$, we consider the following SCO $L_{\mathcal{D}}(w) = \mathbb{E}_{x\sim\mathcal{D}}\frac{1}{2}\|x - w\|_2^2$. We can see that $\nabla L_{\mathcal{D}}(w) = \mathbb{E}_{x\sim\mathcal{D}}[x - w] = \mu - w$. Since $\|\mu\|_2 \leq 1$ we can se that $w^* = \arg\min_{w\in\mathcal{W}} L_{\mathcal{D}}(w) = \mu$ and

$$\begin{aligned}
L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) &= \frac{1}{2}\mathbb{E}_{x\sim\mathcal{D}}[\|w - x\|_2^2 - \|w^* - x\|_2^2] \\
&= \frac{1}{2}\mathbb{E}_{x\sim\mathcal{D}}[\|w\|_2^2 - 2\langle w, x\rangle + \|x\|_2^2 - \|w^*\|_2^2 + 2\langle w^*, x\rangle - \|x\|_2^2] \\
&= \frac{1}{2}(\|w\|_2^2 - 2\langle w, w^*\rangle - \|w^*\|_2^2 + 2\langle w^*, w^*\rangle) \\
&= \frac{1}{2}(\|w\|_2^2 - 2\langle w, w^*\rangle + \|w^*\|_2^2) \\
&= \frac{1}{2}\|w - w^*\|_2^2
\end{aligned}$$

Thus we have

$$\mathbb{E}_{D,\mathcal{A}}L_{\mathcal{D}}(w^{\mathrm{priv}}) - L_{\mathcal{D}}(w^*) = \frac{1}{2}\mathbb{E}\|w^{\mathrm{priv}} - w^*\|_2^2 \geq \Omega\left(u^{\frac{2}{1+v}}d\left(\frac{\sqrt{d\log\frac{1}{\delta}}}{\epsilon n}\right)^{\frac{2v}{1+v}}\right).$$

□

Proof of Theorem 9. We first prove the private term (the second term). Again, we adopt the packing argument. Given $\theta \in \Theta$ with $\|\theta\|_1 = \frac{d}{2}$ and $\theta \in \{0,1\}^d$, let $Q_\theta = (1-p)P_0 + pP_\theta$ for some $p \in [0,1]$, where P_0 is a point mass on $\{D = 0\}$ and P_θ is a point mass on $\left\{D = (u/p)^{\frac{1}{1+v}}\theta\right\}$. Let μ_θ be the mean of Q_θ, i.e., $\mu_\theta = \mathbb{E}_{x\sim Q_\theta}[x]$. Additionally, we define $\bar{\mu}_\theta$ be the normalization of μ_θ, i.e., $\bar{\mu}_\theta = \frac{\mu_\theta}{\|\mu_\theta\|_2}$. Note that $\bar{\mu}_\theta$ is in the same direction as μ_θ, with $\|\bar{\mu}_\theta\|_2 = 1$. By the Gilbert-Varshamov bound for constant-weight codes, there exists a set Θ such that

- The cardinality of Θ satisfies $|\Theta| \geq 2^{\frac{7}{128}d}$.
- For all $\theta \in \Theta$, $\theta \in \{0,1\}^d$ with $\|\theta\|_1 = \frac{d}{2}$.
- For all $\theta_1, \theta_2 \in \Theta$, $d_{\mathrm{Ham}}(\theta_1, \theta_2) \geq \frac{d}{8}$.

For $\forall\theta \in \Theta$, $\|\mu_\theta\|_2$ is the same, which is denoted by ι.

$$\|\mu_\theta\|_2 = \|\mathbb{E}_{x\sim Q_\theta}[x]\|_2 = u^{\frac{1}{1+v}} p^{\frac{v}{1+v}} \cdot \sqrt{\frac{d}{2}} \triangleq \iota.$$

Without loss of generality, we assume the parameter space $\|\mathcal{W}\|_2 = 1$, which is a unit ball. Then we define the loss function $\ell(w, x)$. Given $\theta \in \Theta$ and $x \sim Q_\theta$, we let

$$\ell(w, x) = -\langle w, x\rangle,$$

and

$$L_{Q_\theta}(w) = \mathbb{E}_{x\sim Q_\theta}[\ell(w, x)] = -\langle w, \mu_\theta\rangle.$$

Note that ℓ is both convex and smooth. Let $x \sim Q_\theta$. Note that $\nabla\ell(w, x) = -x$ and $\mathbb{E}[\nabla\ell(w, x)] = -\mu_\theta$,

$$\sup_{j\in[d]} \mathbb{E}_{x\sim Q_\theta}[|\nabla_j\ell(w, x)|^{1+v}] = \sup_{j\in[d]} \mathbb{E}_{x\sim Q_\theta}[| - x_j|^{1+v}] \le p \cdot [(u/p)^{\frac{1}{1+v}}]^{1+v} = u.$$

Define the family $\mathcal{D}_v(u)$ of heavy-tailed distributions of x supported on $\mathbb{R}^d$ by

$$\mathcal{D}_v(u) \triangleq \{\mathcal{D} | \operatorname{supp}\mathcal{D} \subseteq \mathbb{R}^d \quad \text{and} \quad \mathbb{E}_{x\sim\mathcal{D}}[|\langle X, e_j\rangle|^{1+v}] \le u, \quad \forall j \in [d]\}.$$

Next we bound the error of SCO.

$$\begin{aligned}
\sup_{\mathcal{D}\in\mathcal{D}_v(u)} \mathbb{E}\left[L_{\mathcal{D}}(w^{\text{priv}}) - \min_{w\in\mathcal{W}} L_{\mathcal{D}}(w)\right] &\ge \frac{1}{|\Theta|}\sum_{\theta\in\Theta} \mathbb{E}\left[L_{Q_\theta}(w^{\text{priv}}) - \min_{w\in\mathcal{W}} L_{Q_\theta}(w)\right] \\
&\ge \frac{1}{|\Theta|}\sum_{\theta\in\Theta} \mathbb{E}\left[\langle \frac{\mu_\theta}{\|\mu\|_2}, \mu_\theta\rangle - \langle w^{\text{priv}}, \mu_\theta\rangle\right] \\
&= \frac{1}{|\Theta|}\sum_{\theta\in\Theta} \mathbb{E}\left[\|\mu\|_2 - \langle w^{\text{priv}} - \mu_\theta\rangle\right] \\
&= \frac{1}{|\Theta|}\sum_{\theta\in\Theta} \mathbb{E}\left[\|\mu\|_2 \cdot (1 - \langle w^{\text{priv}}, \bar{\mu}_\theta\rangle)\right] \\
&\ge \frac{1}{|\Theta|}\sum_{\theta\in\Theta} \mathbb{E}\left[\|\mu\|_2 \cdot \frac{1}{2}(\|w^{\text{priv}}\|_2^2 + \|\bar{\mu}_\theta\|_2^2 - 2\langle w^{\text{priv}}, \bar{\mu}_\theta\rangle)\right] \\
&= \frac{1}{|\Theta|}\sum_{\theta\in\Theta} \mathbb{E}\left[\frac{1}{2} \cdot \|\mu\|_2 \cdot \|w^{\text{priv}} - \bar{\mu}_\theta\|_2^2\right] = \frac{\iota}{2}\frac{1}{|\Theta|}\sum_{\theta\in\Theta} \mathbb{E}\left[\|w^{\text{priv}} - \bar{\mu}_\theta\|_2^2\right],
\end{aligned}$$

where the first inequality comes from the fact that the worst case loss is no smaller than the average loss, the second inequality is due to the fact that $\bar{\mu}_\theta = \arg\min_{w\in\mathcal{W}} L_{Q_\theta}(w) = -\langle w, \mu_\theta\rangle$, and the third inequality comes from the fact that $\|w^{\text{priv}}\|_2 \le 1$ and $\|\bar{\mu}_\theta\|_2 \le 1$. Note that $|\Theta| \ge 2^{\frac{7}{128}d}$, and for $\forall\theta, \theta'$, $\|\bar{\mu}_\theta - \bar{\mu}_{\theta'}\|_2 = \Omega(1)$, $D_{\text{TV}}(Q_\theta, Q_{\theta'}) = p$. Thus, by Lemma 14 with $\Phi(x) = x$ and ρ as the ℓ_2-norm difference we have

$$\frac{1}{|\Theta|}\sum_{\theta\in\Theta} \mathbb{E}\left[\|w^{\text{priv}} - \bar{\mu}_\theta\|_2^2\right] \ge \Omega(\min\{1, \frac{|\Theta|}{e^{10\epsilon np}}\})$$

Take $p = \Omega(\min(1, \frac{d}{n\epsilon}))$, we have

$$\frac{1}{|\Theta|}\sum_{\theta\in\Theta} \mathbb{E}[\|w^{\text{priv}} - \bar{\mu}_\theta\|_2^2] = \Omega(1).$$

Then,

$$\mathbb{E}[L_{\mathcal{D}}(w^{\text{priv}}) - \min_{w\in\mathcal{W}} L_{\mathcal{D}}(w)] \ge \Omega(1) \cdot \iota = \Omega\left(\sqrt{d}u^{\frac{1}{1+v}} \cdot \min\left\{1, \left(\frac{d}{\epsilon n}\right)^{\frac{v}{1+v}}\right\}\right).$$

For (ϵ, δ)-DP, by Lemma 15 we have

$$\frac{1}{|\Theta|}\sum_{\theta\in\Theta} \mathbb{E}\left[\|w^{\text{priv}} - \bar{\mu}_\theta\|_2^2\right] \ge \Omega(1 - \frac{\frac{\epsilon^2}{4\log\frac{1}{\delta}}(n^2\gamma^2 + n\gamma(1-\gamma)) + \log 2}{\log|\Theta|}).$$

Take $p = \Omega(\min(1, \frac{\sqrt{d \log \frac{1}{\delta}}}{n\epsilon}))$, we have

$$\mathbb{E}[L_{\mathcal{D}}(w^{\text{priv}}) - \min_{w \in \mathcal{W}} L_{\mathcal{D}}(w)] \geq \Omega(1) \cdot \iota = \Omega\left(\sqrt{d}u^{\frac{1}{1+v}} \cdot \min\left\{1, \left(\frac{\sqrt{d \log \frac{1}{\delta}}}{\epsilon n}\right)^{\frac{v}{1+v}}\right\}\right).$$

□

D Proofs in Section 6

Proof of Theorem 10. For the guarantee of DP. First recall that for any $0 < \epsilon, \delta < 1$, the Peeling mechanism Algorithm 4 is (ϵ, δ)-DP if the input vector $v(D)$ satisfies $\|v(D)\|_\infty \leq \lambda$.

Lemma 22 (Lemma 3.3 in [Cai *et al.*, 2021]). If for every pair of neighboring datasets D, D' we have $\|v(D) - v(D')\|_\infty \leq \lambda$, then Algorithm 4 is (ϵ, δ)-DP.

Since in each iteration of Algorithm 5 we use a new data. Thus, it is sufficient to show it is (ϵ, δ)-DP in each iteration. Since we have for any neighboring data $X \sim X'$

$$\|w^{t+1} - w'^{t+1}\|_\infty = \|\eta \widetilde{g}(w^{t-1}, X_t) - \eta \widetilde{g}(w^{t-1}, X'_t)\|_\infty \leq \frac{2B\eta s}{3m}.$$

Thus, by Lemma 22 we can see it is (ϵ, δ)-DP.

In the following we will proof the utility. For simplicity we will omit the subscript r in u_r, λ_r. Before the proof, let us first recall two lemmas related to the output of Algorithm 5.

Lemma 23 (Lemma 3.4 in [Cai *et al.*, 2021]). Let S and $\{w_i\}_{i=1}^s$ be defined is Algorithm 5. For every $R_1 \subseteq S$ and $R_2 \subseteq S^c$ such that $|R_1| = |R_2|$ and every $c > 0$, we have

$$\|v_{R_2}\|_2^2 \leq (1+c)\|v_{R_1}\|_2^2 + 4(1 + \frac{1}{c}) \sum_{i \in [s]} \|w_i\|_\infty^2,$$

where v is the input vector of Algorithm 5.

Lemma 24 (Lemma A.3 in [Cai *et al.*, 2021]). Consider in Algorithm 5 with input vector $\widetilde{v}$ and the index set S. For any index set I, any $v \in \mathbb{R}^{|I|}$ which is a subvector of $\widetilde{v}$ and $\hat{v}$ such that $\|\hat{v}\|_0 \leq \hat{s} \leq s$, we have that for every $c > 0$,

$$\|v_S - v\|_2^2 \leq (1 + \frac{1}{c})\frac{|I| - s}{|I| - \hat{s}}\|\hat{v} - v\|_2^2 + 4(1+c)\sum_{i \in [s]} \|w_i\|_\infty^2.$$

We denote $\widetilde{g}^t = \widetilde{g}(w^t, X_t)$ and $g^t = \nabla L_{\mathcal{D}}(w^t)$, $S^t = \text{supp}(w^t)$, $S^{t+1} = \text{supp}(w^{t+1})$, $S^* = \text{supp}(w^*)$ and $I^t = S^{t+1} \bigcup S^t \bigcup S^*$. We can see that $|S^t| \leq s$, $|S^{t+1}| \leq 2$ and $|I^t| \leq 2s + s^*$. We also denote $W^t = 4\sum_{i \in [s]} \|w_i\|_\infty^2$, where $\{w_i\}$ are the vectors in Algorithm 5 in the t-th iteration. We let $\eta_0 = \frac{\eta}{\gamma}$ for some η.

Then the smooth Lipschitz property we have

$$\begin{aligned} & L_{\mathcal{D}}(w^{t+1}) - L_{\mathcal{D}}(w^t) \\ & \leq \langle w^{t+1} - w^t, g^t \rangle + \frac{\gamma}{2}\|w^{t+1} - w^t\|_2^2 \\ & = \langle w_{I^t}^{t+1} - w_{I^t}^t, g_{I^t}^t \rangle + \frac{\gamma}{2}\|w_{I^t}^{t+1} - w_{I^t}^t\|_2^2 \\ & \leq \frac{\gamma}{2}\|w_{I^t}^{t+1} - w_{I^t}^t + \frac{\eta}{\gamma}g_{I^t}^t\|_2^2 - \frac{\eta^2}{2\gamma}\|g_{I^t}^t\|_2^2 + (1-\eta)\langle w^{t+1} - w^t, g^t \rangle \end{aligned} \tag{27}$$

First, let us focus on the third term of (27)

$$\begin{aligned} \langle w^{t+1} - w^t, g^t \rangle & = \langle w_{S^{t+1} \bigcup S^t}^{t+1} - w_{S^{t+1} \bigcup S^t}^t, g_{S^{t+1} \bigcup S^t}^t \rangle \\ & = \langle w_{S^{t+1}}^{t+1} - w_{S^{t+1}}^t, g_{S^{t+1}}^t \rangle + \langle w_{S^t \backslash S^{t+1}}^{t+1} - w_{S^t \backslash S^{t+1}}^t, g_{S^t \backslash S^{t+1}}^t \rangle \\ & = \langle w_{S^{t+1}}^{t+1} - w_{S^{t+1}}^t, g_{S^{t+1}}^t \rangle - \langle w_{S^t \backslash S^{t+1}}^t, g_{S^t \backslash S^{t+1}}^t \rangle. \end{aligned} \tag{28}$$

From the definition we know that $w^{t+1} = \hat{w}^{t+1} + \widetilde{w}_{S^{t+1}}$, where $\hat{w}^{t+1} = (w^t - \eta_0 \widetilde{g}^t)_{S^{t+1}}$. Thus,

$$\langle w^{t+1} - w^t, g^t \rangle = \langle \hat{w}_{S^{t+1}}^{t+1} - w_{S^{t+1}}^t, g_{S^{t+1}}^t \rangle + \langle \widetilde{w}_{S^{t+1}}, g_{S^{t+1}}^t \rangle - \langle w_{S^t \backslash S^{t+1}}^t - g_{S^t \backslash S^{t+1}}^t \rangle. \tag{29}$$

For the first term in (29) we have

$$
\begin{aligned}
\langle \hat{w}^{t+1}_{S^{t+1}} - w^t_{S^{t+1}}, g^t_{S^{t+1}} \rangle &= \langle -\eta_0 \widetilde{g}^t_{S^{t+1}}, g^t_{S^{t+1}} \rangle = -\frac{\eta}{\gamma} \langle \widetilde{g}^t_{S^{t+1}}, g^t_{S^{t+1}} \rangle \\
&= -\frac{\eta}{\gamma} \|g^t_{S^{t+1}}\|_2^2 - \frac{\eta}{\gamma} \langle \widetilde{g}^t_{S^{t+1}} - g^t_{S^{t+1}}, g^t_{S^{t+1}} \rangle \\
&\leq -\frac{\eta}{\gamma} \|g^t_{S^{t+1}}\|_2^2 + \frac{\eta}{2\gamma} \|g^t_{S^{t+1}}\|_2^2 + \frac{\eta}{2\gamma} \|\widetilde{g}^t_{S^{t+1}} - g^t_{S^{t+1}}\|_2^2 \\
&= -\frac{\eta}{2\gamma} \|g^t_{S^{t+1}}\|_2^2 + \frac{\eta}{2\gamma} \|\widetilde{g}^t_{S^{t+1}} - g^t_{S^{t+1}}\|_2^2. \qquad (30)
\end{aligned}
$$

Take (30) into (29) we have for $c > 1$

$$
\langle w^{t+1} - w^t, g^t \rangle \;\leq\; -\frac{\eta}{2\gamma} \|g^t_{S^{t+1}}\|_2^2 + \frac{\eta}{2\gamma} \|\widetilde{g}^t_{S^{t+1}} - g^t_{S^{t+1}}\|_2^2 + c \|\widetilde{w}_{S^{t+1}}\|_2^2 + \frac{1}{4c} \|g^t_{S^{t+1}}\|_2^2 - \langle w^t_{S^t \backslash S^{t+1}} - g^t_{S^t \backslash S^{t+1}} \rangle. \qquad (31)
$$

For the last term of (31) we have

$$
\begin{aligned}
-\langle w^t_{S^t \backslash S^{t+1}} - g^t_{S^t \backslash S^{t+1}} \rangle &\leq \frac{\gamma}{2\eta} (\|w^t_{S^t \backslash S^{t+1}} - \frac{\eta}{\gamma} g^t_{S^t \backslash S^{t+1}}\|_2^2 - (\frac{\eta}{\gamma})^2 \|g^t_{S^t \backslash S^{t+1}}\|_2^2) \\
&= \frac{\gamma}{2\eta} \|w^t_{S^t \backslash S^{t+1}} - \frac{\eta}{\gamma} g^t_{S^t \backslash S^{t+1}}\|_2^2 - \frac{\eta}{2\gamma} \|g^t_{S^t \backslash S^{t+1}}\|_2^2. \qquad (32)
\end{aligned}
$$

In Lemma 23, let $v = w^t - \frac{\eta}{\gamma} \widetilde{g}^t$, $R_2 = S^t \backslash S^{t+1}$ and $R_1 = S^{t+1} \backslash S^t$. We have for $c > 1$

$$
\|w^t_{S^t \backslash S^{t+1}} - \frac{\eta}{\gamma} \widetilde{g}^t_{S^t \backslash S^{t+1}}\|_2^2 \leq (1 + \frac{1}{c}) \|w^t_{S^{t+1} \backslash S^t} - \frac{\eta}{\gamma} \widetilde{g}^t_{S^{t+1} \backslash S^t}\|_2^2 + (1 + c) W^t.
$$

Since for every $c > 1$, $(1 - \frac{1}{c}) \|a\|^2 - (c - 1) \|b\|_2^2 \leq \|a + b\|^2 \leq (1 + \frac{1}{c}) \|a\|_2^2 + (1 + c) \|b\|_2^2$ we have

$$
\begin{aligned}
&(1 - \frac{1}{c}) \|w^t_{S^t \backslash S^{t+1}} - \frac{\eta}{\gamma} g^t_{S^t \backslash S^{t+1}}\|_2^2 - (c - 1) \frac{\eta^2}{\gamma^2} \|g^t_{S^t \backslash S^{t+1}} - \widetilde{g}^t_{S^t \backslash S^{t+1}}\|_2^2 \\
&\leq \|w^t_{S^t \backslash S^{t+1}} - \frac{\eta}{\gamma} \widetilde{g}^t_{S^t \backslash S^{t+1}}\|_2^2 \leq (1 + \frac{1}{c}) \|w^t_{S^{t+1} \backslash S^t} - \frac{\eta}{\gamma} \widetilde{g}^t_{S^{t+1} \backslash S^t}\|_2^2 + (1 + c) W^t \\
&\leq (1 + \frac{1}{c}) [(1 + 1/c) \|w^t_{S^{t+1} \backslash S^t} - \frac{\eta}{\gamma} g^t_{S^{t+1} \backslash S^t}\|_2^2 + (1 + c) \frac{\eta^2}{\gamma^2} \|g^t_{S^t \backslash S^{t+1}} - \widetilde{g}^t_{S^t \backslash S^{t+1}}\|_2^2] + 2(1 + c) W^t. \qquad (33)
\end{aligned}
$$

That is

$$
\begin{aligned}
\|w^t_{S^t \backslash S^{t+1}} - \frac{\eta}{\gamma} g^t_{S^t \backslash S^{t+1}}\|_2^2 \leq{}& \frac{(c + 1)^2}{c(c - 1)} \|w^t_{S^{t+1} \backslash S^t} - \frac{\eta}{\gamma} g^t_{S^{t+1} \backslash S^t}\|_2^2 \\
&+ (c + \frac{(c + 1)^2}{c}) \frac{\eta}{2\gamma} (\|g^t_{S^t \backslash S^{t+1}} - \widetilde{g}^t_{S^t \backslash S^{t+1}}\|_2^2 + \|g^t_{S^{t+1} \backslash S^t} - \widetilde{g}^t_{S^{t+1} \backslash S^t}\|_2^2) + \frac{c(1 + c)}{c - 1} W^t.
\end{aligned}
$$

Thus

$$
\begin{aligned}
-\langle w^t_{S^t \backslash S^{t+1}} - g^t_{S^t \backslash S^{t+1}} \rangle \leq{}& \frac{(c + 1)^2}{2c(c - 1)} \frac{\eta}{\gamma} \|g^t_{S^{t+1} \backslash S^t}\|_2^2 \\
&+ \frac{\gamma c(1 + c)}{2\eta(c - 1)} W^t - \frac{\eta}{2\gamma} \|g^t_{S^t \backslash S^{t+1}}\|_2^2 + \frac{(2c + 3)\eta}{2\gamma} (\|g^t_{S^t \backslash S^{t+1}} - \widetilde{g}^t_{S^t \backslash S^{t+1}}\|_2^2 + \|g^t_{S^{t+1} \backslash S^t} - \widetilde{g}^t_{S^{t+1} \backslash S^t}\|_2^2).
\end{aligned}
$$

Thus in (31) we have

$$\begin{aligned}
&\langle w^{t+1} - w^t, g^t\rangle \leq -\frac{\eta}{2\gamma}\|g^t_{S^{t+1}}\|_2^2 + \frac{\eta}{2\gamma}\|\widetilde{g}^t_{S^{t+1}} - g^t_{S^{t+1}}\|_2^2 + c\|\widetilde{w}_{S^{t+1}}\|_2^2 \\
&\qquad + \frac{1}{4c}\|g^t_{S^{t+1}}\|_2^2 - \langle w^t_{S^t\backslash S^{t+1}} - g^t_{S^t\backslash S^{t+1}}\rangle \\
&\leq -\frac{\eta}{2\gamma}\|g^t_{S^{t+1}}\|_2^2 + \frac{\eta}{2\gamma}\|\widetilde{g}^t_{S^{t+1}} - g^t_{S^{t+1}}\|_2^2 + c\|\widetilde{w}_{S^{t+1}}\|_2^2 + \frac{1}{4c}\|g^t_{S^{t+1}}\|_2^2 - \frac{\eta}{2\gamma}\|g^t_{S^{t+1}}\|_2^2 \\
&+ \frac{(c+1)^2}{2c(c-1)}\frac{\eta}{\gamma}\|g^t_{S^{t+1}\backslash S^t}\|_2^2 + \frac{(2c+3)\eta}{2\gamma}(\|g^t_{S^t\backslash S^{t+1}} - \widetilde{g}^t_{S^t\backslash S^{t+1}}\|_2^2 + \|g^t_{S^{t+1}\backslash S^t} - \widetilde{g}^t_{S^{t+1}\backslash S^t}\|_2^2) \\
&= \frac{\eta}{2\gamma}\|g^t_{S^{t+1}\backslash S^t}\|_2^2 + \frac{\eta}{2\gamma}\frac{3c+1}{c(c-1)}\|g^t_{S^{t+1}\backslash S^t}\|_2^2 - \frac{\eta}{2\gamma}\|g^t_{S^t\backslash S^{t+1}}\|_2^2 - \frac{\eta}{2\gamma}\|g^t_{S^{t+1}}\|_2^2 \\
&\qquad + \frac{1}{4c}\|g^t_{S^{t+1}}\|_2^2 + \frac{\gamma c(1+c)}{2\eta(c-1)}W^t + \frac{\eta}{2\gamma}\|\widetilde{g}^t_{S^{t+1}} - g^t_{S^{t+1}}\|_2^2 + c\|\widetilde{w}_{S^{t+1}}\|_2^2 \\
&\qquad + \frac{(2c+3)\eta}{\gamma}(\|g^t_{S^t\backslash S^{t+1}} - \widetilde{g}^t_{S^t\backslash S^{t+1}}\|_2^2 + \|g^t_{S^{t+1}\backslash S^t} - \widetilde{g}^t_{S^{t+1}\backslash S^t}\|_2^2) \\
&= \frac{\eta}{2\gamma}\|g^t_{S^{t+1}\backslash S^t}\|_2^2 - \frac{\eta}{2\gamma}\|g^t_{S^t\backslash S^{t+1}}\|_2^2 - \frac{\eta}{2\gamma}\|g^t_{S^{t+1}}\|_2^2 + \frac{1}{c}(\frac{1}{4} + \frac{\eta}{2\gamma} + \frac{\eta}{2\gamma}\frac{3c+1}{(c-1)})\|g^t_{S^{t+1}}\|_2^2 \\
&\underbrace{+ \frac{\gamma c(1+c)}{2\eta(c-1)}W^t + \frac{\eta}{2\gamma}\|\widetilde{g}^t_{S^{t+1}} - g^t_{S^{t+1}}\|_2^2 + c\|\widetilde{w}_{S^{t+1}}\|_2^2 + \frac{(2c+3)\eta}{\gamma}(\|g^t_{S^t\backslash S^{t+1}} - \widetilde{g}^t_{S^t\backslash S^{t+1}}\|_2^2 + \|g^t_{S^{t+1}\backslash S^t} - \widetilde{g}^t_{S^{t+1}\backslash S^t}\|_2^2)}_{N^t} \\
&\leq \frac{\eta}{2\gamma}\|g^t_{S^{t+1}\backslash S^t}\|_2^2 - \frac{\eta}{2\gamma}\|g^t_{S^t\backslash S^{t+1}}\|_2^2 - \frac{\eta}{2\gamma}\|g^t_{S^{t+1}}\|_2^2 + C_1\frac{\eta}{\gamma c}\|g^t_{S^{t+1}}\|_2^2 + N^t, \qquad (34)
\end{aligned}$$

where $C_1 > 0$ is some constant. We can easily see that

$$\begin{aligned}
\frac{\eta}{2\gamma}\|g^t_{S^{t+1}\backslash S^t}\|_2^2 - \frac{\eta}{2\gamma}\|g^t_{S^t\backslash S^{t+1}}\|_2^2 - \frac{\eta}{2\gamma}\|g^t_{S^{t+1}}\|_2^2 &= -\frac{\eta}{2\gamma}\|g^t_{S^t\backslash S^{t+1}}\|_2^2 - \frac{\eta}{2\gamma}\|g^t_{S^{t+1}\bigcap S^t}\|_2^2 \\
&= -\frac{\eta}{2\gamma}\|g^t_{S^{t+1}\bigcup S^t}\|_2^2
\end{aligned}$$

In total

$$\langle w^{t+1} - w^t, g^t\rangle \leq -\frac{\eta}{2\gamma}\|g^t_{S^{t+1}\bigcup S^t}\|_2^2 + C_1\frac{\eta}{\gamma c}\|g^t_{S^{t+1}}\|_2^2 + N_1. \qquad (35)$$

Take (35) into (27) we have

$$\begin{aligned}
&L_{\mathcal{D}}(w^{t+1}) - L_{\mathcal{D}}(w^t) \leq \frac{\gamma}{2}\|w^{t+1}_{I^t} - w^t_{I^t} + \frac{\eta}{\gamma}g^t_{I^t}\|_2^2 - \frac{\eta^2}{2\gamma}\|g^t_{I^t}\|_2^2 + (1-\eta)\langle w^{t+1} - w^t, g^t\rangle \\
&\leq \frac{\gamma}{2}\|w^{t+1}_{I^t} - w^t_{I^t} + \frac{\eta}{\gamma}g^t_{I^t}\|_2^2 - \frac{\eta^2}{2\gamma}\|g^t_{I^t}\|_2^2 - \frac{(1-\eta)\eta}{2\gamma}\|g^t_{S^{t+1}\bigcup S^t}\|_2^2 + C_1\frac{(1-\eta)}{c}\frac{\eta}{\gamma}\|g^t_{S^{t+1}}\|_2^2 \\
&\qquad + (1-\eta)N^t \\
&\leq \frac{\gamma}{2}\|w^{t+1}_{I^t} - w^t_{I^t} + \frac{\eta}{\gamma}g^t_{I^t}\|_2^2 - \frac{\eta^2}{2\gamma}\|g^t_{I^t\backslash(S^t\bigcup S^*)}\|_2^2 - \frac{\eta^2}{2\gamma}\|g^t_{(S^t\bigcup S^*)}\|_2^2 - \frac{(1-\eta)\eta}{2\gamma}\|g^t_{S^{t+1}\bigcup S^t}\|_2^2 \\
&+ C_1\frac{(1-\eta)}{c}\frac{\eta}{\gamma}\|g^t_{S^{t+1}}\|_2^2 + (1-\eta)N^t \\
&\leq \frac{\gamma}{2}\|w^{t+1}_{I^t} - w^t_{I^t} + \frac{\eta}{\gamma}g^t_{I^t}\|_2^2 - \frac{\eta^2}{2\gamma}\|g^t_{I^t\backslash(S^t\bigcup S^*)}\|_2^2 - \frac{\eta^2}{2\gamma}\|g^t_{(S^t\bigcup S^*)}\|_2^2 \\
&- \frac{(1-\eta)\eta}{2\gamma}\|g^t_{S^{t+1}\backslash(S^*\bigcup S^t)}\|_2^2 + C_1\frac{(1-\eta)}{c}\frac{\eta}{\gamma}\|g^t_{S^{t+1}}\|_2^2 + (1-\eta)N^t, \qquad (36)
\end{aligned}$$

where the last inequality is due to $S^{t+1}\backslash(S^*\bigcup S^t) \subseteq S^{t+1}\bigcup S^t$. Next we will analyze the term $\frac{\gamma}{2}\|w^{t+1}_{I^t} - w^t_{I^t} + \frac{\eta}{\gamma}g^t_{I^t}\|_2^2 - \frac{\eta^2}{2\gamma}\|g^t_{I^t\backslash(S^t\bigcup S^*)}\|_2^2$ in (36). Let R be a subset of $S^t\backslash S^{t+1}$ such that $|R| = |I^t\backslash(S^*\bigcup S^t)| = |S^{t+1}\backslash(S^t\bigcup S^*)|$. In Lemma 23, we take $v = w^t - \frac{\eta}{\gamma}\widetilde{g}^t$, $R_2 = R$ and $R_1 = I^t\backslash(S^*\bigcup S^t)$ we have for $c > 1$,

$$\|w^t_R - \frac{\eta}{\gamma}\widetilde{g}^t_R\|_2^2 \leq (1+c)\|(w^t - \frac{\eta}{\gamma}\widetilde{g}^t)_{I^t\backslash(S^*\bigcup S^t)}\|_2^2 + (1+\frac{1}{c})W^t. \qquad (37)$$

Just as in (33) we have is for $c > 1$,

$$(\frac{\eta}{\gamma})^2\|g^t_{I^t\backslash(S^*\bigcup S^t)}\|_2^2 \geq (1-\frac{1}{c})\|w^t_R - \frac{\eta}{\gamma}g^t_R\|_2^2 - cW^t - c\frac{\eta^2}{\gamma^2}(\|\widetilde{g}^t_R - g^t_R\|_2^2 + \|g^t_{I^t\backslash(S^*\bigcup S^t)} - \widetilde{g}^t_{I^t\backslash(S^*\bigcup S^t)}\|_2^2). \tag{38}$$

Then we have

$$\begin{aligned}
&\frac{\gamma}{2}\|w^{t+1}_{I^t} - w^t_{I^t} + \frac{\eta}{\gamma}g^t_{I^t}\|_2^2 - \frac{\eta^2}{2\gamma}\|g^t_{I^t\backslash(S^t\bigcup S^*)}\|_2^2\\
&\leq \frac{\gamma}{2}\|\widetilde{w}_{S^{t+1}}\|_2^2 + \frac{\gamma}{2}\|\hat{w}^{t+1}_{I^t} - w^t_{I^t} + \frac{\eta}{\gamma}g^t_{I^t}\|_2^2 - \frac{\gamma}{2}(1-\frac{1}{c})\|w^t_R - \frac{\eta}{\gamma}g^t_R\|_2^2 + \frac{\gamma c}{2}W^t\\
&+ c\frac{\eta^2}{2\gamma}(\|\widetilde{g}^t_R - g^t_R\|_2^2 + \|g^t_{I^t\backslash(S^*\bigcup S^t)} - \widetilde{g}^t_{I^t\backslash(S^*\bigcup S^t)}\|_2^2) + c\frac{\eta^2}{\gamma^2}(\|\widetilde{g}^t_R - g^t_R\|_2^2 \qquad (39)\\
&= \frac{\gamma}{2}\|\hat{w}^{t+1}_{I^t} - w^t_{I^t} + \frac{\eta}{\gamma}g^t_{I^t}\|_2^2 - \frac{\gamma}{2}\|\hat{w}^{t+1}_R - w^t_R + \frac{\eta}{\gamma}g^t_R\|_2^2 + \frac{\gamma}{2}\|\widetilde{w}_{S^{t+1}}\|_2^2 + \frac{\gamma c}{2}W^t\\
&+ \frac{\gamma}{2c}\|w^t_R - \frac{\eta}{\gamma}g^t_R\|_2^2 + c\frac{\eta^2}{2\gamma}(\|\widetilde{g}^t_R - g^t_R\|_2^2 + \|g^t_{I^t\backslash(S^*\bigcup S^t)} - \widetilde{g}^t_{I^t\backslash(S^*\bigcup S^t)}\|_2^2) \qquad (40)\\
&\leq \frac{\gamma}{2}\|\hat{w}^{t+1}_{I^t\backslash R} - w^t_{I^t\backslash R} + \frac{\eta}{\gamma}g^t_{I^t\backslash R}\|_2^2 + \frac{\gamma}{2c}(1+\frac{1}{c})\|\frac{\eta}{\gamma}g^t_{I^t\backslash(S^*\bigcup S^t)}\|_2^2 + \frac{\gamma}{2}\|\widetilde{w}_{S^{t+1}}\|_2^2\\
&+ \underbrace{\frac{\gamma c}{2}W^t + \frac{\gamma}{2c}(1+c)W^t + C_2c\frac{\eta^2}{2\gamma}(\|\widetilde{g}^t_R - g^t_R\|_2^2 + \|g^t_{I^t\backslash(S^*\bigcup S^t)} - \widetilde{g}^t_{I^t\backslash(S^*\bigcup S^t)}\|_2^2)}_{N_1^t}. \qquad (41)
\end{aligned}$$

(39) is due to that $[\hat{w}^{t+1}_{I^t} - (w^t_{I^t} - \frac{\eta}{\gamma}g^t_{I^t})]_{S^{t+1}} = 0$, thus $\langle\widetilde{w}_{S^{t+1}}, \hat{w}^{t+1}_{I^t} - (w^t_{I^t} - \frac{\eta}{\gamma}g^t_{I^t})\rangle = 0$ and (38). (40) is due to $\hat{w}^{t+1}_R = 0$, (41) is due to (37) by the same technique as in (33) and $w^t_{I^t\backslash(S^*\bigcup S^t)} = 0$. In the following we will consider the first term in (41).

In Lemma 24, take $v = w^t_{I^t\backslash R} - \frac{\eta}{\gamma}\widetilde{g}^t_{I^t\backslash R}$, $\hat{v} = w^*$, $S = S^{t+1}$ we have for all $c > 1$

$$\|\hat{w}^{t+1}_{I^t\backslash R} - w^t_{I^t\backslash R} + \frac{\eta}{\gamma}\widetilde{g}^t_{I^t\backslash R}\|_2^2 \leq (1+\frac{1}{c})\frac{|I^t\backslash R| - s}{|I^t\backslash R| - s^*}\|w^* - w^t_{I^t\backslash R} + \frac{\eta}{\gamma}\widetilde{g}^t_{I^t\backslash R}\|_2^2 + (1+c)W^t.$$

Then we have

$$\begin{aligned}
&(1-\frac{1}{c})\|\hat{w}^{t+1}_{I^t\backslash R} - w^t_{I^t\backslash R} + \frac{\eta}{\gamma}g^t_{I^t\backslash R}\|_2^2 - (c-1)\frac{\eta^2}{\gamma^2}\|g^t_{I^t\backslash R} - \widetilde{g}^t_{I^t\backslash R}\|_2^2\\
&\leq \|\hat{w}^{t+1}_{I^t\backslash R} - w^t_{I^t\backslash R} + \frac{\eta}{\gamma}\widetilde{g}^t_{I^t\backslash R}\|_2^2\\
&\leq (1+\frac{1}{c})\frac{|I^t\backslash R| - s}{|I^t\backslash R| - s^*}\|w^* - w^t_{I^t\backslash R} + \frac{\eta}{\gamma}\widetilde{g}^t_{I^t\backslash R}\|_2^2 + (1+c)W^t\\
&\leq (1+\frac{1}{c})\frac{|I^t\backslash R| - s}{|I^t\backslash R| - s^*}[(1+\frac{1}{c})\|w^* - w^t_{I^t\backslash R} + \frac{\eta}{\gamma}g^t_{I^t\backslash R}\|_2^2 + (1+c)\frac{\eta^2}{\gamma^2}\|g^t_{I^t\backslash R} - \widetilde{g}^t_{I^t\backslash R}\|_2^2] + (1+c)W^t
\end{aligned}$$

That is

$$\begin{aligned}
\|\hat{w}^{t+1}_{I^t\backslash R} - w^t_{I^t\backslash R} + \frac{\eta}{\gamma}g^t_{I^t\backslash R}\|_2^2 \leq{}& \frac{(c+1)^2}{c(c-1)}\frac{2s^*}{s+s^*}\|w^* - w^t_{I^t\backslash R} + \frac{\eta}{\gamma}g^t_{I^t\backslash R}\|_2^2\\
&+ \frac{(c+1)^2}{c-1}\frac{\eta^2}{\gamma^2}\|g^t_{I^t\backslash R} - \widetilde{g}^t_{I^t\backslash R}\|_2^2 + c\frac{\eta^2}{\gamma^2}\|g^t_{I^t\backslash R} - \widetilde{g}^t_{I^t\backslash R}\|_2^2 + \frac{(1+c)c}{c-1}W^t
\end{aligned}$$

Take $c \geq \frac{\sqrt{3}}{\sqrt{3}-\sqrt{2}}$, and since $|I^t\backslash R| \leq 2s^* + s$, we have

$$\begin{aligned}
&\|\hat{w}^{t+1}_{I^t\backslash R} - w^t_{I^t\backslash R} + \frac{\eta}{\gamma}g^t_{I^t\backslash R}\|_2^2 \leq \frac{3}{2}\frac{2s^*}{s+s^*}\|w^* - w^t_{I^t\backslash R} + \frac{\eta}{\gamma}g^t_{I^t\backslash R}\|_2^2\\
&+ \underbrace{C_3c(\|g^t_{I^t\backslash R} - \widetilde{g}^t_{I^t\backslash R}\|_2^2 + \|g^t_{I^t\backslash R} - \widetilde{g}^t_{I^t\backslash R}\|_2^2 + W^t)}_{N_3^t}. \qquad (42)
\end{aligned}$$

Take (42) into (41) we have

$$
\begin{aligned}
&\frac{\gamma}{2}\|w^{t+1}_{I^t} - w^t_{I^t} + \frac{\eta}{\gamma} g^t_{I^t}\|_2^2 - \frac{\eta^2}{2\gamma}\|g^t_{I^t\setminus(S^t\bigcup S^*)}\|_2^2 \\
&\leq \frac{3\gamma s^*}{2(s+s^*)}\|w^* - w^t_{I^t\setminus R} + \frac{\eta}{\gamma} g^t_{I^t\setminus R}\|_2^2 + \frac{\gamma}{2c}(1+\frac{1}{c})\|\frac{\eta}{\gamma} g^t_{I^t\setminus(S^*\bigcup S^t)}\|_2^2 + \frac{\gamma}{2}\|\widetilde{w}_{S^{t+1}}\|_2^2 \\
&+ N_1^t + N_3^t \qquad (43) \\
&= \frac{3\gamma s^*}{2(s+s^*)}\|w^* - w^t_{I^t\setminus R} + \frac{\eta}{\gamma} g^t_{I^t\setminus R}\|_2^2 + \frac{\gamma}{2c}(1+\frac{1}{c})\|\frac{\eta}{\gamma} g^t_{S^{t+1}}\|_2^2 \\
&\quad + \frac{\gamma}{2}\|\widetilde{w}_{S^{t+1}}\|_2^2 + N_1^t + N_3^t \qquad (44) \\
&= \frac{3s^*}{s+s^*}(\eta\langle w^* - w^t, g^t\rangle + \frac{\gamma}{2}\|w^* - w^t\|_2^2 + \frac{\eta^2}{2c\gamma}\|g^t_{I^t}\|_2^2) + \frac{\eta^2}{2c\gamma}(1+\frac{1}{c})\|g^t_{S^{t+1}}\|_2^2 \\
&\quad + \frac{\gamma}{2}\|\widetilde{w}_{S^{t+1}}\|_2^2 + N_1^t + N_3^t \qquad (45) \\
&\leq \frac{3s^*}{s+s^*}(\eta(L_{\mathcal{D}}(w^*) - L_{\mathcal{D}}(w^t)) + \frac{\gamma-\eta\mu}{2}\|w^* - w^t\|_2^2 + \frac{\eta^2}{2c\gamma}\|g^t_{I^t}\|_2^2) + \frac{\eta^2}{2c\gamma}(1+\frac{1}{c})\|g^t_{S^{t+1}}\|_2^2 \\
&\quad + \underbrace{\frac{\gamma}{2}\|\widetilde{w}_{S^{t+1}}\|_2^2 + N_1^t + N_3^t}_{N_2^t}. \qquad (46)
\end{aligned}
$$

Take (46) into (36) we have

$$
\begin{aligned}
&L_{\mathcal{D}}(w^{t+1}) - L_{\mathcal{D}}(w^t) \leq \frac{\gamma}{2}\|w^{t+1}_{I^t} - w^t_{I^t} + \frac{\eta}{\gamma} g^t_{I^t}\|_2^2 - \frac{\eta^2}{2\gamma}\|g^t_{I^t\setminus(S^t\bigcup S^*)}\|_2^2 - \frac{\eta^2}{2\gamma}\|g^t_{(S^t\bigcup S^*)}\|_2^2 \\
&- \frac{(1-\eta)\eta}{2\gamma}\|g^t_{S^{t+1}\setminus(S^*\bigcup S^t)}\|_2^2 + C_1\frac{(1-\eta)}{c}\frac{\eta}{\gamma}\|g^t_{S^{t+1}}\|_2^2 + (1-\eta)N^t \\
&\leq \frac{3s^*}{s+s^*}(\eta(L_{\mathcal{D}}(w^*) - L_{\mathcal{D}}(w^t)) + \frac{\gamma-\eta\mu}{2}\|w^* - w^t\|_2^2 + \frac{\eta^2}{2c\gamma}\|g^t_{I^t}\|_2^2) + \frac{\eta^2}{2c\gamma}(1+\frac{1}{c})\|g^t_{S^{t+1}}\|_2^2 \\
&- \frac{\eta^2}{2\gamma}\|g^t_{(S^t\bigcup S^*)}\|_2^2 - \frac{(1-\eta)\eta}{2\gamma}\|g^t_{S^{t+1}\setminus(S^*\bigcup S^t)}\|_2^2 + C_1\frac{(1-\eta)}{c}\frac{\eta}{\gamma}\|g^t_{S^{t+1}}\|_2^2\|g^t_{S^{t+1}}\|_2^2 + (1-\eta)N^t + N_2^t. \qquad (47)
\end{aligned}
$$

We have when $c \to \infty$

$$
\frac{\eta^2}{2c\gamma}(1+\frac{1}{c})\|g^t_{S^{t+1}}\|_2^2 + C_1\frac{(1-\eta)}{c}\frac{\eta}{\gamma}\|g^t_{S^{t+1}}\|_2^2 \to 0.
$$

Thus, if $\eta \geq \frac{1}{2}$ there must exits a sufficient large c such that

$$
\begin{aligned}
&\frac{\eta^2}{2c\gamma}(1+\frac{1}{c})\|g^t_{S^{t+1}}\|_2^2 + \frac{(1-\eta)}{c}(\frac{1}{4}+\frac{\eta}{2\gamma})\|g^t_{S^{t+1}}\|_2^2 \leq \frac{\eta(1-\eta)}{4\gamma}\|g^t_{S^{t+1}}\|_2^2 \\
&\leq \frac{\eta^2}{4\gamma}\|g^t_{(S^t\bigcup S^*)}\|_2^2 + \frac{(1-\eta)\eta}{4\gamma}\|g^t_{S^{t+1}\setminus(S^*\bigcup S^t)}\|_2^2 \qquad (48)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&L_{\mathcal{D}}(w^{t+1}) - L_{\mathcal{D}}(w^t) \leq \frac{\gamma}{2}\|w^{t+1}_{I^t} - w^t_{I^t} + \frac{\eta}{\gamma} g^t_{I^t}\|_2^2 - \frac{\eta^2}{2\gamma}\|g^t_{I^t\setminus(S^t\bigcup S^*)}\|_2^2 - \frac{\eta^2}{2\gamma}\|g^t_{(S^t\bigcup S^*)}\|_2^2 \\
&- \frac{(1-\eta)\eta}{2\gamma}\|g^t_{S^{t+1}\setminus(S^*\bigcup S^t)}\|_2^2 + C_1\frac{(1-\eta)}{c}\frac{\eta}{\gamma}\|g^t_{S^{t+1}}\|_2^2 + (1-\eta)N^t \\
&\leq \frac{3s^*}{s+s^*}(\eta(L_{\mathcal{D}}(w^*) - L_{\mathcal{D}}(w^t)) + \frac{\gamma-\eta\mu}{2}\|w^* - w^t\|_2^2 + \frac{\eta^2}{2c\gamma}\|g^t_{I^t}\|_2^2) \\
&- \frac{\eta^2}{4\gamma}\|g^t_{(S^t\bigcup S^*)}\|_2^2 - \frac{(1-\eta)\eta}{4\gamma}\|g^t_{S^{t+1}\setminus(S^*\bigcup S^t)}\|_2^2 + (1-\eta)N^t + N_2^t. \qquad (49)
\end{aligned}
$$

Take $\eta = \frac{2}{3}$, $s = 72\frac{\gamma^2}{\mu^2}s^*$ so that $\frac{3s^*}{s+s^*} \leq \frac{\mu^2}{24\gamma(\gamma-\eta\mu)} \leq \frac{1}{8}$. We have

$$\begin{aligned}
L_{\mathcal{D}}(w^{t+1}) - L_{\mathcal{D}}(w^t) &\leq \frac{3s^*}{s+s^*}(\eta(L_{\mathcal{D}}(w^*) - L_{\mathcal{D}}(w^t)) + \frac{\gamma - \eta\mu}{2}\|w^* - w^t\|_2^2 + \frac{\eta^2}{2c\gamma}\|g^t_{I^t}\|_2^2) \\
&- \frac{\eta^2}{4\gamma}\|g^t_{(S^t \bigcup S^*)}\|_2^2 - \frac{(1-\eta)\eta}{4\gamma}\|g^t_{S^{t+1}\backslash(S^* \bigcup S^t)}\|_2^2 + (1-\eta)N^t + N_2^t \\
&\leq \frac{2s^*}{s+s^*}(L_{\mathcal{D}}(w^*) - L_{\mathcal{D}}(w^t)) + \frac{\mu^2}{48\gamma}\|w^* - w^t\|_2^2 + \frac{1}{36\gamma}\|g^t_{I^t}\|_2^2 \\
&\quad - \frac{1}{9\gamma}\|g^t_{(S^t \bigcup S^*)}\|_2^2 - \frac{1}{18\gamma}\|g^t_{S^{t+1}\backslash(S^* \bigcup S^t)}\|_2^2 + O(N^t + N_2^t) \\
&\leq \frac{2s^*}{s+s^*}(L_{\mathcal{D}}(w^*) - L_{\mathcal{D}}(w^t)) - \frac{3}{36\gamma}(\|g^t_{(S^t \bigcup S^*)}\|_2^2 - \frac{\mu^2}{4}\|w^* - w^t\|_2^2) + O(N^t + N_2^t) \quad (50) \\
&\leq (\frac{2s^*}{s+s^*} + \frac{\mu}{24\gamma})(L_{\mathcal{D}}(w^*) - L_{\mathcal{D}}(w^t)) + O(N^t + N_2^t). \quad (51)
\end{aligned}$$

Where (50) is due to the following lemma:

Lemma 25. [Lemma 6 in [Jain *et al.*, 2014]]

$$|g^t_{(S^t \bigcup S^*)}\|_2^2 - \frac{\mu^2}{4}\|w^* - w^t\|_2^2 \geq \frac{\mu}{2}(L_{\mathcal{D}}(w^t) - L_{\mathcal{D}}(w^*)). \quad (52)$$

Thus

$$L_{\mathcal{D}}(w^{t+1}) - L_{\mathcal{D}}(w^*) \leq (1 - \frac{5}{72}\frac{\mu}{\gamma})(L_{\mathcal{D}}(w^t) - L_{\mathcal{D}}(w^*)) + O(N^t + N_2^t).$$

Where

$$N^t + N_2^t \leq O(\sum_{i\in[s]}\|w_i\|_\infty^2 + (2s+s^*)\|\widetilde{g}^t - g^t\|_\infty^2 + s\|\widetilde{w}\|_\infty^2), \quad (53)$$

where each coordinate of $w_i, \widetilde{w}$ sampled from $\sim \text{Lap}(O(\frac{\eta_0 TB\sqrt{s\log\frac{1}{\delta}}}{n\epsilon}))$. We first bound the term of $\sum_{i\in[s]}\|w_i\|_\infty^2 + s\|\widetilde{w}\|_\infty^2$. We recall the following lemma:

Lemma 26 (Lemma A.1 in [Cai *et al.*, 2021])**.** For a random vector $w \in \mathbb{R}^d$, where $w_i \sim \text{Lap}(\lambda)$, then for any $\xi > 0$,

$$\Pr(\|w\|_\infty^2 \geq 4\lambda^2\log^2\frac{1}{\xi}\log^2 d) \leq \xi.$$

Take the union we have with probability at least $1-\xi$,

$$\sum_{i\in[s]}\|w_i\|_\infty^2 + s\|\widetilde{w}\|_\infty^2 \leq O(\frac{\eta_0^2 T^2 B^2 s^2 \log\frac{1}{\delta}\log^2\frac{Ts}{\xi}}{n^2\epsilon^2}).$$

For $s\|\widetilde{g}^t - g^t\|_\infty^2$, due to the assumption on the moment and (5) we have with probability at least $1-\xi$

$$(2s+s^*)\|\widetilde{g}^t - g^t\|_\infty^2 \leq (2s+s^*)(\sqrt{\frac{2B^{1-v}u\log\frac{d}{\xi}}{m}} + \frac{B\log\frac{1}{\xi}}{3m} + \frac{u}{B^v})^2 = O(s(\frac{TB^{1-v}u\log\frac{d}{\xi}}{n} + \frac{B^2T^2\log^2\frac{1}{\xi}}{n^2} + \frac{u^2}{B^{2v}})). \quad (54)$$

Thus, we have with probability at least $1-2\xi$,

$$N^t + N_2^t = O(s(\frac{TB^{1-v}u\log\frac{d}{\xi}}{n} + \frac{u^2}{B^{2v}} + \frac{T^2B^2s\log\frac{1}{\delta}\log^2\frac{Ts}{\xi}}{\gamma^2 n^2\epsilon^2})). \quad (55)$$

Take $B = O(\left(\frac{\gamma u n\epsilon}{T\log\frac{d}{\xi}\sqrt{s\log\frac{1}{\delta}}}\right)^{\frac{1}{1+v}})$ we have

$$N^t + N_2^t = O\left(s^{\frac{1+3v}{2+2v}}u^{\frac{2}{1+v}}(\frac{T\log\frac{d}{\xi}}{n})^{\frac{2v}{1+v}} + su^{\frac{2}{1+v}}\left(\frac{T\log\frac{d}{\xi}\sqrt{s\log\frac{1}{\delta}}}{\gamma n\epsilon}\right)^{\frac{2v}{1+v}}\right) = O\left(su^{\frac{2}{1+v}}\left(\frac{T\log\frac{d}{\xi}\sqrt{s\log\frac{1}{\delta}}}{\gamma n\epsilon}\right)^{\frac{2v}{1+v}}\right). \quad (56)$$

In total we have for all $t \in [T]$ with probability at least $1 - \xi$,

$$L_{\mathcal{D}}(w^{t+1}) - L_{\mathcal{D}}(w^*) \leq (1 - \frac{\mu}{12\gamma})(L_{\mathcal{D}}(w^t) - L_{\mathcal{D}}(w^*)) + O\left(su^{\frac{2}{1+v}} \left(\frac{T \log \frac{d}{\xi} \sqrt{s \log \frac{1}{\delta}}}{\gamma n \epsilon} \right)^{\frac{2v}{1+v}} \right)$$

$$L_{\mathcal{D}}(w^{T+1}) - L_{\mathcal{D}}(w^*) \leq (1 - \frac{\mu}{12\gamma})^{T+1}(L_{\mathcal{D}}(w^1) - L_{\mathcal{D}}(w^*)) + O\left(\frac{\gamma}{\mu} \cdot su^{\frac{2}{1+v}} \left(\frac{T \log \frac{d}{\xi} \sqrt{s \log \frac{1}{\delta}}}{\gamma n \epsilon} \right)^{\frac{2v}{1+v}} \right).$$

Take $T = \widetilde{O}(\frac{\gamma}{\mu} \log n)$ and $s = O((\frac{\gamma}{\mu})^2 s^*)$ we have the result. □

Proof of Theorem 11. We first show the following result:

Lemma 27. For any (ϵ, δ)-DP algorithm $\mathcal{A}$, there exists a distribution $\mathcal{D}$ over $\mathbb{E}_{x \sim \mathcal{D}}[x] = \mu$ with $\|\mu\|_0 \leq s^*$ and $\sup_{j \in [d]} \mathbb{E}_{x \sim \mathcal{D}}[|x_j|^{1+v}] \leq u$, such that

$$\mathbb{E}_{x \sim \mathcal{D}, \mathcal{A}}[\|\mathcal{A}(x) - \mu\|_2] \geq \Omega\left(u^{\frac{1}{1+v}} \min\left(1, \left\{ \frac{s^* \log d}{\epsilon n} \right)^{\frac{v}{1+v}} \right\} \right).$$

By Jensen's inequality, the above yields

$$\mathbb{E}_{x \sim \mathcal{D}, \mathcal{A}}[\|\mathcal{A}(x) - \mu\|_2^2] \geq \Omega\left(u^{\frac{2}{1+v}} \min\left\{1, \left(\frac{s^* \log d}{\epsilon n} \right)^{\frac{2v}{1+v}} \right\} \right).$$

For any (ϵ, δ)-DP algorithm $\mathcal{A}$, there exists a distribution $\mathcal{D}$ over $\mathbb{E}_{x \sim \mathcal{D}}[x] = \mu$ with $\|\mu\|_0 \leq s^*$ and $\sup_{j \in [d]} \mathbb{E}_{x \sim \mathcal{D}}[|x_j|^{1+v}] \leq u$, such that

$$\mathbb{E}_{x \sim \mathcal{D}, \mathcal{A}}[\|\mathcal{A}(x) - \mu\|_2^2] \geq \Omega\left(u^{\frac{2}{1+v}} \min\left\{1, \left(\frac{\sqrt{s^* \log d \log \frac{1}{\delta}}}{\epsilon n} \right)^{\frac{2v}{1+v}} \right\} \right).$$

Proof. We first recall the following lemma:

Lemma 28. [[Raskutti *et al.*, 2011]] For any $s^* \in [d]$, define the set

$$\mathcal{H}(s^*) := \{z \in \{-1, 0, +1\}^d \mid \|z\|_0 = s^*\}$$

with Hamming distance $\rho_H(z, z') = \sum_{i=1}^d 1[z_j \neq z'_j]$ between the vectors z and z'. Then, there exists a subset $\widetilde{\mathcal{H}} \subset \mathcal{H}$ with cardinality $|\widetilde{\mathcal{H}}| \geq \exp(\frac{s}{2} \log \frac{d - s^*}{s^*/2})$ such that $\rho_H(z, z') \geq \frac{s^*}{2}$ for all $z, z' \in \widetilde{\mathcal{H}}$.

We denote the index set $\Theta = \frac{1}{\sqrt{2s^*}} \widetilde{H}$ where $\widetilde{H}$ is in Lemma 28. We can see that for any $\theta, \theta' \in \Theta$ we have $\|\theta - \theta'\|_2 \geq \sqrt{2}$ and each $\|\theta\|_2 \leq 1$. For each θ we construct the following distribution: $P_\theta := (1-p)P_0 + pP_\theta \in \mathcal{P}$ with some $p \in [0, 1]$, where Let P_0 be a point mass distribution supported on $X = 0$, let P_θ be a point mass supported on $X = (\frac{u}{p})^{\frac{1}{1+v}} \theta$. We can see that $\mathbb{E}_{x \sim P_\theta}[\|x\|_2^{1+v}] \leq u$ and $\|\mu_\theta\|_0 = \|\mathbb{E}_{x \sim P_\theta}[x]\|_0 \leq s^*$. Thus, $P_\theta \in \mathcal{D}_{v,s^*}(u)$, where the family $\mathcal{D}_{v,s^*}(u)$ of heavy-tailed distributions of x supported on $\mathbb{R}^d$ by

$$\mathcal{D}_{v,s^*}(u) \triangleq \{\mathcal{D} | \operatorname{supp} \mathcal{D} \subseteq \mathbb{R}^d \quad \text{and} \quad \mathbb{E}_{x \sim \mathcal{D}}[|\langle X, e_j \rangle|^{1+v}] \leq u, \forall j \in [d] \quad \text{and} \quad \|\mathbb{E}_{x \sim \mathcal{D}}[x]\|_0 \leq s^*\}.$$

Moreover we have $\|\mu_\theta - \mu_{\theta'}\|_2 \geq \sqrt{2} p^{\frac{v}{1+v}} u^{\frac{1}{1+v}}$ for all $\theta, \theta' \in \Theta$, and $D_{TV}(P_\theta, P_{\theta'}) \leq p$. Thus, by Lemma 14 we have

$$\mathcal{M}_n(\theta(\mathcal{D}_{v,s^*}(u)), Q, \|\cdot\|_2, \epsilon) \geq \frac{1}{|\Theta|} \sum_{\theta \in \Theta} \mathbb{E}_{X \sim P_\theta^n, Q}[\|Q(X) - \mu_\theta\|] \geq \Omega(p^{\frac{v}{1+v}} u^{\frac{1}{1+v}} \min\{1, \frac{|\Theta|}{e^{10\epsilon n p}}\}). \tag{57}$$

Let $p = \Omega(\{1, \frac{s^* \log d}{n\epsilon}\})$ we have the result for all ϵ-DP algorithms.

For (ϵ, δ)-DP, by Lemma 15 we have

$$\mathcal{M}_n(\theta(\mathcal{D}_{v,s^*}(1)), Q, \|\cdot\|_2, \epsilon, \delta) \geq \frac{1}{|\Theta|} \sum_{\theta \in \Theta} \mathbb{E}_{X \sim P_\theta^n, Q}[\|Q(X) - \mu_\theta\|] \geq \Omega(p^{\frac{v}{1+v}} u^{\frac{1}{1+v}} (1 - \frac{\frac{\epsilon^2}{4 \log \frac{1}{\delta}}(n^2 p^2 + np(1-p)) + \log 2}{\log |\Theta|}). \tag{58}$$

Take $p = \Omega(\{1, \frac{\sqrt{s^* \log d \log \frac{1}{\delta}}}{n\epsilon}\})$ we have the result. □

Now we back to the proof, we can reduce the problem to mean estimation for each $\mathcal{D} \in \mathcal{D}_{v,s^*}(u)$, where $L_{\mathcal{D}}(w) = \mathbb{E}_{x\sim\mathcal{D}}[\frac{1}{2}\|x - w\|_2^2]$. Note that $\nabla L_{\mathcal{D}}(w) = \mathbb{E}_{x\sim\mathcal{D}}[x] - w$. Thus $w^* = \mathbb{E}_{x\sim\mathcal{D}}[x] \in \mathcal{W}$ and $L_{\mathcal{D}}(w^{\text{priv}}) - L(w^*) = \frac{1}{2}\mathbb{E}\|w^{\text{priv}} - \mathbb{E}_{x\sim\mathcal{D}}[x]\|_2$. By Lemma 27 we complete the proof. □