User-level Private Stochastic Convex Optimization with Optimal Rates

Raef Bassily [1] **Ziteng Sun** [2]

Abstract

We study the problem of differentially private (DP) stochastic convex optimization (SCO) under the notion of user-level differential privacy. In this problem, there are n users, each contributing $m > 1$ samples to the input dataset of the private SCO algorithm, and the notion of indistinguishability embedded in DP is w.r.t. replacing the entire local dataset of any given user.

Under smoothness conditions of the loss, we establish the optimal rates for user-level DP-SCO in both the central and local models of DP. In particular, we show, roughly, that the optimal rate is $\frac{1}{\sqrt{nm}} + \frac{\sqrt{d}}{\varepsilon n\sqrt{m}}$ in the central setting and is $\frac{\sqrt{d}}{\varepsilon\sqrt{nm}}$ in the local setting, where d is the dimensionality of the problem and ε is the privacy parameter. Our algorithms combine new user-level DP mean estimation techniques with carefully designed first-order stochastic optimization methods. For the central DP setting, our optimal rate improves over the rate attained for the same setting in Levy et al. (2021) by $\sqrt{d}$ factor. One of the main ingredients that enabled such an improvement is a novel application of the generalization properties of DP in the context of multi-pass stochastic gradient methods.

1. Introduction

Differential privacy (DP) (Dwork et al., 2006) has become the gold standard for rigorous privacy protection in machine learning. Given the fundamental importance of stochastic convex optimization (SCO) in machine learning, many works studied stochastic optimization algorithms under the constraint of differential privacy, a problem referred to as differentially private SCO (DP-SCO) (Bassily et al., 2019; Feldman et al., 2020; Bassily et al., 2021b; Asi et al., 2021; Kulkarni et al., 2021; Bassily et al., 2021a; Song et al., 2021; Arora et al., 2022).

However, most of the existing works on DP-SCO study a basic setting, where there are n individuals (users) and each user contributes a single data point to the input dataset of the algorithm, and hence the privacy guarantee of the proposed algorithms is *item-level DP*. A more general setting of substantial practical importance is when each user contributes a local dataset of m data points to the input dataset, where $m > 1$. In such scenarios, item-level DP does not provide sufficient privacy protection for each user. Instead, *user-level DP* would offer a more meaningful and stronger privacy protection in these scenarios. In user-level DP, the output of the differentially private (DP) algorithm needs to be insensitive to replacing the entire local dataset of any user. Under item-level DP, optimal rates for DP-SCO (optimal excess population risk bounds) are known in the central model of DP (Bassily et al., 2014) and the local model (Duchi et al., 2013) (see the first row of Table 1).

A fundamental question is: *what are the optimal rates for DP-SCO under user-level DP?* In particular, can we attain better rates for user-level DP than those rates implied directly from the optimal algorithms for item-level DP? That is, can we devise new algorithms that exploit the fact that each user contributes multiple samples to the input dataset? Note that a naive baseline would be using the existing algorithms designed for item-level DP while only modifying the notion of privacy to user-level DP. However, this yields rates that are essentially the same as in the case where $m = 1$ (Levy et al., 2021). Related to this question, the work of Levy et al. (2021) gives an upper bound on the DP-SCO rate in the central model of DP when the loss function is sufficiently smooth. Under the Lipschitz condition, their rate scales as $\approx \frac{1}{\sqrt{nm}} + \frac{d}{\varepsilon n\sqrt{m}}$ in the worst case. Although their bound shows an improvement in terms of the dependence on m compared to the naive baseline mentioned above, it is worse than such naive baseline in terms of its dependence on d, particularly by a factor of $\sqrt{d}$. The picture in the local model is even less clear as there are no known DP-SCO rates better than what can be implied via the naive approach.

*Equal contribution [1]Department of Computer Science Engineering and the Translational Data Analytics Institute (TDAI), The Ohio State University [2]Google Research, New York. Correspondence to: Ziteng Sun <zitengsun@google.com>, Raef Bassily <bassily.1@osu.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Model	Central DP	Local DP
Item-level	$\Theta\left(\frac{1}{\sqrt{mn}} + \frac{\sqrt{d\log(1/\delta)}}{nm\varepsilon}\right)$ (Bassily et al., 2019)	$\Theta\left(\sqrt{\frac{d}{nm\varepsilon^2}}\right)$ (Duchi et al., 2013)
User-level	$\tilde{\Theta}\left(\frac{1}{\sqrt{mn}} + \frac{\sqrt{d\log(1/\delta)}}{n\sqrt{m}\varepsilon}\right)\dagger$ (Theorem 3.1)	$\tilde{\Theta}\left(\sqrt{\frac{d}{nm\varepsilon^2}}\right)\dagger$ (Theorem 3.2)

Table 1. Comparison of the excess population risks for SCO with $R = L = 1$ in different privacy models. n : number of users; m : number of samples per user; d :dimension of the problem. Results marked by † require additional conditions on the smoothness or parameter ranges.

Our contributions: In this work, we give an answer to the fundamental question above in the smooth setting of DP-SCO. In particular, *under reasonable smoothness conditions on the loss function, we prove optimal rates for DP-SCO with user-level DP in both the central and local models of DP.* Our rates are stated in the second row of Table 1.

Our overarching approach is based on combining new techniques for user-level DP mean estimation with carefully chosen first-order (gradient-based) optimization methods. In particular, our approach entails devising two DP versions of a stochastic gradient oracle: one for the central model of DP and another for the local model. Those instantiations are based on techniques of DP mean estimation that take advantage of the multiple samples at each user to provide a more accurate DP estimate of the gradient of the population loss, whose variance scales roughly as $O\left(\frac{1}{m}\right)$, hence, effectively reducing the Lipschitz constant of the loss by a factor of $\approx \frac{1}{\sqrt{m}}$ (after appropriate recentering of the gradient estimates).

One of the main contributions of our work lies in how we attain the optimal rate in the central model that does not suffer from the extra $\sqrt{d}$ factor in the rate of Levy et al. (2021). We provide a novel application of the generalization and concentration properties of DP (Dwork et al., 2015; Bassily et al., 2016; Feldman & Steinke, 2018) in the context of multi-pass stochastic gradient methods. When each user has multiple i.i.d. samples, the average local gradient at each user will be concentrated around the true population gradient and hence recent advances on private estimation of *concentrated random variables* can be used. However, for private multi-pass algorithms, each iterate is a function of previous estimates based on the users' data, which breaks the independence structure of local gradients. The work of Levy et al. (2021) uses uniform concentration of the gradients to get around this problem, but this leads to an extra factor of $\sqrt{d}$ in the attained rate. In this work, we show that the generalization properties of DP can be used effectively to ensure concentration of the local gradient estimates across all users and all iterations of the algorithm without essentially any extra cost in the rate. See Section 3.1 for a detailed discussion of the technique.

We also deviate from Levy et al. (2021) in terms of the gradient-based algorithm. Apart from our construction of the underlying DP stochastic gradient oracle described earlier, our algorithm is quite simple and has a similar outline to the noisy mini-batch stochastic gradient descent (SGD) algorithm of Bassily et al. (2019).

In the local DP setting, we propose new variance-reduced local DP gradient estimator in the high-dimensional case for concentrated random variables, which is crucial for obtaining the optimal rate. Moreover, apart from this, our private optimization algorithm for the local model is still different from the optimal local DP-SCO algorithm with item-level DP (Duchi et al., 2013). It turns out that using the standard one-pass noisy SGD algorithm requires a relatively strong assumption on the smoothness of the loss to yield the optimal rate for user-level LDP, even after we replace the noisy gradients with our variance-reduced local DP gradient estimator. To attain the optimal rate in this case with a milder smoothness condition, we give a new private algorithm based on accelerated mini-batch SGD (Cotter et al., 2011).

Finally, our results entail a condition on the total number of users n. We prove a lower bound showing that this condition is necessary.

2. Preliminaries

Stochastic convex optimization (SCO). Let $\ell(\cdot, z)$ be a loss function which is convex in its first argument. Let P be a distribution over $\mathcal{Z}$. For all $\theta \in \Theta$, define $F(\theta) = \mathbb{E}_{Z\sim P}\left[\ell(\theta, Z)\right]$. Given i.i.d samples from P, the goal is to find $\widehat{\theta}$ with small excess risk, $F(\widehat{\theta}) - \min_{\theta\in\Theta} F(\theta) \leq \alpha$. In this paper, we use $\nabla\ell(\cdot, z)$ to refer to the gradient with respect the first argument. We put additional assumptions on the loss function and parameter space:

- **Lipschitzness:** We assume $\forall z \in \mathcal{Z}$, $\ell(\theta, z)$ is L-lipschitz in its first argument, *i.e.,* $\forall z \in \mathcal{Z}, \theta \in \Theta$,

$$\|\nabla\ell(\theta, z)\|_2 \leq L.$$

- **Bounded parameter range:** We assume $\Theta = \{\theta \in$

$\mathbb{R}^d \mid \|\theta\|_2 \le R\}$.

- **Smoothness:** $\ell(\cdot,\cdot)$ is said to be β-smooth if $\forall z \in \mathcal{Z}$, $\theta_1, \theta_2 \in \Theta$

$$\|\nabla \ell(\theta_1, z) - \nabla \ell(\theta_2, z)\| \le \beta \|\theta_1 - \theta_2\|.$$

Differential privacy at user-level. We consider the setting where the samples are contributed by multiple users and each user contributes more than one samples. More specifically, there are n users and each user observes m i.i.d samples from P. We denote the ith user's samples as $Z_i = (Z_{i,1}, Z_{i,2}, \ldots, Z_{i,m})$. When $m = 1$, we use Z_i to denote the ith user's single sample. The dataset consisting of all user's samples are denoted as $S = Z^n := (Z_1, Z_2, \ldots, Z_n)$.

We will consider differential privacy (DP) both in the central and local model, based on the following indistinguishability notion.

Definition 2.1 (Indistinguishability)**.** For $\varepsilon > 0$ and $\delta \in (0,1)$, two distributions P and Q supported on $\mathcal{O}$ are called (ε, δ)-indistinguishable (denoted as $P \sim_{(\varepsilon,\delta)} Q$) if for all event O in the probability space,

$$e^{-\varepsilon}(P(O) - \delta) \le Q(O) \le e^{\varepsilon} P(O) + \delta.$$

Note that when $\varepsilon = 0$, the notion is equivalent to $d_{\mathrm{TV}}(P, Q) \le \delta$, where $d_{\mathrm{TV}}(P, Q) = \sup_{O \in \mathcal{O}} |P(O) - Q(O)|$ is the total variantion distance between P and Q.

Next we give the definitions of central differential privacy (DP) and local differential privacy (LDP).

Definition 2.2 (Differential privacy)**.** An algorithm $\mathcal{A}$ is said to be (ε, δ)-differentially private (DP) if for any two datasets S, S' differing on at most one user's contribution, i.e., $\sum_{i=1}^n \mathbb{1}\{Z_i \neq Z_i'\} \le 1$, we have

$$\mathcal{A}(S) \sim_{(\varepsilon,\delta)} \mathcal{A}(S').$$

When $m > 1$, the definition is referred to as *user-level* DP. When $m = 1$, the definition is the same as the canonical *item-level* DP.

Definition 2.3 (Local differential privacy)**.** An randomizer $\mathcal{R}$ is said to be (ε, δ)-LDP if for all $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}^m$, we have

$$\mathcal{R}(\mathbf{z}) \sim_{(\varepsilon,\delta)} \mathcal{R}(\mathbf{z}').$$

An algorithm $\mathcal{A}$ is said to be (ε, δ)-LDP if for all i, $\mathcal{A}$ can only access Z_i through an (ε, δ)-LDP randomizer.

The following properties of DP will be useful in the analysis.

Theorem 2.4 (Advanced composition (Dwork et al., 2014))**.** *If $\forall i \in [k], \mathcal{A}_i$ is (ε, δ)-DP, $\forall \delta' \in (0,1)$, their (adaptive) composition $(\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k)$ is $(\varepsilon\sqrt{2k\log(1/\delta')} + k\varepsilon(e^{\varepsilon} - 1), \delta' + k\delta)$-DP.*

Theorem 2.5 (Amplification by subsampling (Balle et al., 2018))**.** *For $\varepsilon < 1$ and $\delta \in (0,1)$, let $\mathcal{A} : \mathcal{Z}^k \to \Theta$ be a (ε, δ)-private algorithm. For $n > k$ and a dataset $S \subset \mathcal{Z}^n$, let S^{wor} be a size k dataset obtained from randomly sample without replacement from S. Then $\mathcal{A}'$ obtained from $\mathcal{A}'(S) = \mathcal{A}(S^{wor})$ is $((e-1)\frac{k}{n}\varepsilon, \frac{k}{n}\delta)$-DP.*

Throughout the paper, we often need to deal with concentrated random variables, defined below.

Definition 2.6. A sequence of n random vectors $X^n = (X_1, \ldots, X_n)$, where $X_i \in \mathbb{R}^d$, $\forall i \in [n]$, is said to be (τ, γ)-*concentrated* if with probability $1 - \gamma$, there exists $x \in \mathbb{R}^d$ such that

$$\forall i \in [n], \qquad \|X_i - x\|_2 \le \tau.$$

One example of such concentrated random variables is the *subgaussian* random variables. A d-dimensional random variable $X \sim P$ is said to be σ-subgaussian if for any $v \in \mathbb{R}^d$ with $\|v\|_2 = 1$, we have $\forall t > 0$,

$$\Pr\left(|v \cdot X| \ge t\right) \le 2e^{-\frac{t^2}{2\sigma^2}}.$$

If can be verified that when $X^n \sim_{i.i.d} P$ for a σ-subgaussian distribution P, we have $\forall \gamma > 0$, X^n is $(\sigma\sqrt{2\log(2n/\gamma)}, \gamma)$-concentrated.

Additional notations. We use $\mathcal{B}_2^d(x, R)$ to denote the d-dimensional ℓ_2 ball of radius R centered around x. When $x = \vec{0}$, we drop x and simply use $\mathcal{B}_2^d(R)$. For a convex set Ω, Π_Ω is used to denote the ℓ_2-projection onto Ω, *i.e.,*

$$\Pi_\Omega(x) := \min_{x' \in \Omega} \|x - x'\|_2.$$

For $X^n \in \mathbb{R}^{d \times n}$, we use

$$\mu(X^n) := \frac{1}{n}\sum_{i=1}^n X_i$$

to denote its mean.

3. Our results and techniques.

SCO under central DP. There has been a rich literature on private SCO in under *item-level* ($m = 1$) central DP recently (Bassily et al., 2019; Feldman et al., 2020; Bassily et al., 2021b; Asi et al., 2021; Kulkarni et al., 2021; Bassily et al., 2021a; Song et al., 2021; Arora et al., 2022). It has been shown that for $\varepsilon = O(1)$, there exists an (ε, δ)-DP algorithm $\mathcal{A}$ with expected excess risk of

$$\tilde{O}\left(\frac{RL}{\sqrt{n}} + \frac{RL\sqrt{d\log(1/\delta)}}{n\varepsilon}\right). \tag{1}$$

Moreover, the rate is shown to be tight. Levy et al. (2021) study SCO under *user-level* privacy and it is shown that under certain smoothness conditions, the following excess risk can be obtained

$$\tilde{O}\left(\frac{RL}{\sqrt{nm}} + \frac{Rd\min\{L,\sigma\}}{n\sqrt{m}\varepsilon}\right),$$

where σ is the subgaussian parameter of $\nabla\ell(\theta, Z)$ when $Z \sim P$. The result shows that each user contributing more samples can indeed help in certain cases. Although the dependence on σ is shown to be tight in certain cases, σ can be as large as L in the worst case. In this case, when $m < d$, the privacy rate is worse than the baseline of (1), which can be achieved by each user ignoring $m-1$ additional samples. Hence it is not clear whether more samples can help when m is small.

In this paper, we prove the following theorem, which shows that under certain smoothness conditions, collecting more samples from each user can provably improve the excess risk under Lipschitz assumption.

Theorem 3.1. *For $\varepsilon \in (0,1), \delta \in (0, \frac{d}{4n^{5/2}\sqrt{m}\varepsilon^2})$, when $n > \sqrt{d}/\varepsilon, m < \max\{\sqrt{d}, n\varepsilon^2/\sqrt{d}\}$ and $\beta = \tilde{O}(\frac{2L}{R}\min\{\frac{n^{3/2}\varepsilon^2}{d\sqrt{m}}, \frac{n\varepsilon}{\sqrt{md}}\})$, there exists an (ε,δ) user-level private algorithm $\mathcal{A}$ with*

$$\mathbb{E}\left[F(\mathcal{A}(Z^n)) - \min_{\theta\in\Theta} F(\theta)\right] = \tilde{O}\left(\frac{RL}{\sqrt{nm}} + \frac{RL\sqrt{d}}{\sqrt{m}n\varepsilon}\right).$$

Moreover, given the lower bound in Levy et al. (2021) the rate is tight up to logarithmic factors (see (2) in Section 5 for log factors).

Parameter requirements. As shown in Levy et al. (2021, Theorem 9), under fixed n and ε, the optimality gap won't approach zero even when $m = \infty$. This implies that the rate in Theorem 3.1 won't hold for arbitrarily large m. Whether the requirement on m and β can be relaxed is an interesting direction to explore.

SCO under LDP. Under local DP constraint, due to the more stringent privacy notion, at *item-level* ($m = 1$), the optimal rate for the excess risk is shown to be (Duchi et al., 2013)

$$\Theta\left(RL\sqrt{\frac{d}{n\varepsilon^2}}\right).$$

The result under *user-level* LDP is less explored in previous work to the best of our knowledge. Our next result shows that under the local setting, collecting more samples from each user can provably improve the performance under the same privacy level.

Theorem 3.2. *For $\varepsilon = O(1)$ and $\delta < \varepsilon, m < d/\varepsilon^2$, when $\beta < \frac{L\varepsilon^3}{R}\sqrt{\frac{n^3}{md^3}}$ and $n \geq d/\varepsilon^2$ (or $n > md/\varepsilon^2$ when β is unbounded), there exists an (ε,δ) user-level LDP algorithm $\mathcal{A}$ with*

$$\mathbb{E}\left[F(\mathcal{A}(Z^n)) - \min_{\theta\in\Theta} F(\theta)\right] = \tilde{O}\left(RL\sqrt{\frac{d}{nm\varepsilon^2}}\right),$$

Moreover, the rate is tight up to logarithmic factors.

Interestingly, in contrast to the result in the central case, the risk decreases at the same rate for m and n. This shows that to achieve similar excess risk, less users are needed if we are allowed to collect multiple samples from each user. Moreover, this doesn't come at the cost of leaking more information about each user since the user-level privacy parameter is fixed.

When the function is not smooth, Theorem 3.2 has an additional requirement that $n > md/\varepsilon^2$ in the nonsmooth case and $n > d/\varepsilon^2$ in the smooth case. In Theorem 3.3, we show that this requirement is mild when m is small in the sense that $n > d/\varepsilon^2$ is required to get any non-trivial optimization guarantee for any m. Whether the requirement can be removed when $d/\varepsilon^2 < n < md/\varepsilon^2$ in the nonsmooth case is an interesting future direction to explore.

Theorem 3.3. *$\forall m > 0$ and ε user-level LDP algorithm $\mathcal{A}$, when $n = o(d/\varepsilon^2)$, there exists a distribution P such that*

$$\mathbb{E}\left[F(\mathcal{A}(Z^n)) - \min_{\theta\in\Theta} F(\theta)\right] = \Omega(RL).$$

Private mean estimation. Our LDP optimization algorithm relies on private mean estimation of concentrated random variables in the high-dimensional setting. The problem has been well-studied in the central setting (Smith, 2011; Karwa & Vadhan, 2017; Cai et al., 2019; Kamath et al., 2019; Biswas et al., 2020; Levy et al., 2021). For the local setting, we propose an computationally efficient extension of the algorithm in (Gaboardi et al., 2019) to the high-dimensional setting. The result is stated in Theorem 4.3.

The independent work of Girgis et al. (2022) also provides a similar result for private mean estimation, which is used to solve the task of user-level LDP empirical risk minization (ERM). While related, ERM and SCO are fundamentally different problems that have been investigated separately in the private optimization literature (*e.g.,* Bassily et al. (2014; 2019)). The optimal rates for DP-ERM don't imply optimal rates for DP-SCO in general.

3.1. Our technique: concentration via DP

Similar to Levy et al. (2021), our algorithm is based on the gradient-base optimization algorithms and the challenge

comes from privately estimating the gradient at each iterate. When each user has multiple i.i.d samples, the averaged local gradient at each user will be concentrated around the true population gradient and hence recent advances on private estimation of *concentrated random variables* can be used. However, for private optimization algorithms, the queried parameter at each iteration is a function of previous estimates based on the users' data, which breaks the independence structure of the local gradients. Levy et al. (2021) resorts to uniform concentration of the gradients, and this leads to an extra factor of $\sqrt{d}$ in the private risk. In this work, we resolve this issue by carefully exploiting the generalization property of differential privacy for concentrated random variables in the adaptive estimation setting, as shown below.

Theorem 3.4. *Let $Z^n = (Z_1, \ldots, Z_n) \in \bar{\mathcal{Z}}^n$ be a sequence of samples drawn from a distribution P over a user's data universe $\bar{\mathcal{Z}}$. Let $f : \Theta \times \bar{\mathcal{Z}} \to \mathbb{R}^d$ and P be such that $\forall \theta \in \Theta$,*

$$\Pr\left(\|f(\theta, Z) - f(\theta, P)]\| > \tau\right) \leq \gamma,$$

where $f(\theta, P) := \mathbb{E}_{Z\sim P}[f(\theta, Z)]$. Let $\mathcal{A} : \bar{\mathcal{Z}}^n \to \Theta$ be an (ε, δ)-differentially private algorithm and $\boldsymbol{\theta}_{Z^n} := \mathcal{A}(Z^n)$. Then, the sequence

$$\vec{f}(\boldsymbol{\theta}_{Z^n}, Z^n) := (f(\boldsymbol{\theta}_{Z^n}, Z_1), f(\boldsymbol{\theta}_{Z^n}, Z_2), \ldots, f(\boldsymbol{\theta}_{Z^n}, Z_n))$$

is (τ, γ')-concentrated, where $\gamma' = n\big(e^{2\varepsilon}\gamma + \delta\big)$.

When θ is independent of Z^n, union bound would imply that $\vec{f}(\boldsymbol{\theta}, Z^n)$ is $(\tau, n\gamma)$-concentrated. The theorem shows that when $\boldsymbol{\theta}_{Z^n}$ is a function of the dataset Z^n, the concentration of function queries to the dataset still holds with a slightly worse property as long as $\boldsymbol{\theta}_{Z^n}$ is differentially private. Note that here the vector $\vec{f}(\boldsymbol{\theta}_{Z^n}, Z^n)$ involves n quantities and a naive application of group privacy would incur a multiplicative term of $e^{n\varepsilon}$. We get around this by using the fact that $\forall i \in [n]$, $\boldsymbol{\theta}_{Z^n}$ can be viewed as a private randomization of Z_i and $\forall \theta$, $f(\theta, Z_i)$ is concentrated itself. Hence we can apply the generalization property of DP on each entry and use union bound to argue about the concentration of the sequence. We present its proof in Section 3.2.

Our algorithm for the central DP case builds on this and use algorithms for private estimation of concentrated random variables obtained from adaptive but private queries. This is important to achieve the desired rate for iterative optimization methods. We then carefully choose the number of rounds and learning rate to balance optimization loss, privacy loss, and generalization error, which we detail in Section 5.

3.2. Proof of Theorem 3.4

By the union bound and the definition of (τ, γ)-concentration, it would be enough if we prove that $\forall i \in [n]$,

$$\Pr\left(\|f(\boldsymbol{\theta}_{Z^n}, Z_i) - f(\boldsymbol{\theta}_{Z^n}, P)\| \geq \tau\right) \leq e^{2\varepsilon}\gamma + \delta.$$

Our first observation is that $\forall i \in [n], \boldsymbol{\theta}_{Z^n}$ is an (ε, δ)-DP randomization of Z_i. By the following lemma from Feldman et al. (2022), we know it is close to a $(2\varepsilon, 0)$-randomization of Z_i.

Lemma 3.5. *Let $\mathcal{A}$ be an (ε, δ) randomization of Z, then there exists an $(2\varepsilon, 0)$-DP algorithm $\mathcal{A}'$ such that*

$$d_{\mathrm{TV}}(\mathcal{A}(Z), \mathcal{A}'(Z)) \leq \delta.$$

Let $\mathcal{A}'$ be the algorithm defined in Lemma 3.5 and $\boldsymbol{\theta}'_{Z^n} := \mathcal{A}'(Z^n)$. We know that there exists a coupling between $\boldsymbol{\theta}'_{Z^n}$ and $\boldsymbol{\theta}_{Z^n}$ such that

$$\Pr\left(\boldsymbol{\theta}'_{Z^n} \neq \boldsymbol{\theta}_{Z^n}\right) \leq \delta.$$

Hence it would be enough to prove that

$$\Pr\left(\|f(\boldsymbol{\theta}'_{Z^n}, Z_i) - f(\boldsymbol{\theta}'_{Z^n}, P)\| \geq \tau\right) \leq e^{2\varepsilon}\gamma.$$

Note that

$$\begin{aligned}
&\Pr\left(\|f(\boldsymbol{\theta}'_{Z^n}, Z_i) - f(\boldsymbol{\theta}'_{Z^n}, P)\|_2 \geq \tau\right) \\
&= \sum\nolimits_{\theta} \Pr\left(\boldsymbol{\theta}'_{Z^n} = \theta\right) \cdot \\
&\qquad \Pr\left(\|f(\theta, Z_i \mid \boldsymbol{\theta}'_{Z^n} = \theta) - f(\theta, P)\|_2 \geq \tau\right) \\
&\leq \max_{\theta} \Pr\left(\|f(\theta, Z_i \mid \boldsymbol{\theta}'_{Z^n} = \theta) - f(\theta, P)\|_2 \geq \tau\right).
\end{aligned}$$

Since $\boldsymbol{\theta}'_{Z^n}$ is 2ε-DP, we have $\forall z$ and θ,

$$\frac{\Pr\left(Z_i = z \mid \boldsymbol{\theta}'_{Z^n} = \theta\right)}{\Pr\left(Z_i = z\right)} = \frac{\Pr\left(\boldsymbol{\theta}'_{Z^n} = \theta \mid Z_i = z\right)}{\Pr\left(\boldsymbol{\theta}'_{Z^n} = \theta\right)} \leq e^{2\varepsilon}.$$

Hence we have $\forall \theta \in \Theta$,

$$\begin{aligned}
&\Pr\left(\|f(\theta, Z_i \mid \boldsymbol{\theta}'_{Z^n} = \theta) - f(\theta, P)\|_2 \geq \tau\right) \\
&= \sum\nolimits_{z} \Pr\left(Z_i = z \mid \boldsymbol{\theta}'_{Z^n} = \theta\right) \cdot \\
&\qquad \Pr\left(\|f(\theta, z) - f(\theta, P)\|_2 \geq \tau\right) \\
&\leq e^{2\varepsilon} \sum\nolimits_{z} \Pr\left(Z_i = z\right) \Pr\left(\|f(\theta, z) - f(\theta, P)\|_2 \geq \tau\right) \\
&\leq e^{2\varepsilon} \Pr\left(\|f(\theta, Z_i) = \theta) - f(\theta, P)\|_2 \geq \tau\right) \\
&\leq e^{2\varepsilon}\gamma,
\end{aligned}$$

where the last inequality follows from the concentration assumption of $f(\theta, Z_i)$. This completes the proof.

4. User-level private mean estimation

In this section, we will describe the mean estimation primitives for *concentrated random variables*. We start by describing a meta-algorithm that can be instantialized to give mean estimation algorithms under central DP and local DP, respectively. At a high level, the algorithm starts by privately computing a crude estimate of the mean which is close to most of the data points (within radius C). Then it truncates each sample and computes the noisy mean of the truncated samples. The details of the algorithm are described in Algorithm 1.

Algorithm 1 Truncated mean estimation

Input: $X^k = (X_1, X_2, \dots X_k) \in \mathcal{B}_2^d(R)$; σ : noise level; C : truncation radius. **CrudeMean**: a private crude mean estimator.

1: Compute a crude estimate of the mean using **CrudeMean** up to radius C.

$$\tilde{\mu} = \textbf{CrudeMean}(X^k, C).$$

2: Compute the noisy truncated mean

$$\widehat{\mu} = \frac{1}{k}\sum_{i=1}^{k}\left(\Pi_{\mathcal{B}_2^d(\tilde{\mu},C)}(X_i) + \mathcal{N}(0, \sigma^2 \mathbb{I}_d)\right)$$

3: **Return:** $\widehat{\mu}$.

4.1. Mean estimation under central DP

Under central DP, Levy et al. (2021) study mean estimation of concentrated random variables. We will use the following result from Levy et al. (2021).

Theorem 4.1 (Theorem 2 (Levy et al., 2021))**.** *For $\varepsilon \in (0,1)$ and $\delta \in (0, 1/n)$, there exists a private mean estimator $\mathcal{A}$, which is an instantiation of Algorithm 1 such that if X^k is (τ,γ)-concentrated and $k = \Omega\Big(\sqrt{d\log(1/\delta)}\log(dRn/(\tau/\gamma))/\varepsilon\Big)$, we have*

$$\mathcal{A}(X^k) \sim_{(0,2\gamma)} \frac{1}{k}\sum_{i=1}^{k} X_i + \mathcal{N}(0, \sigma^2 \mathbb{I}_d),$$

with $\sigma^2 = O\Big(\frac{\tau^2 \log(dn/\gamma)\log(1/\delta)}{k^2\varepsilon^2}\Big)$.

The statement shows that if the data is concentrated, the private estimator can be close to a Gaussian perturbation of the empirical mean with noise level scaling quadratically with the concentration radius τ (up to log factors) instead of quadratically with the worst-case bound of R. For completeness, we give a description of algorithm in Appendix B.

4.2. Mean estimation under local DP

Next we consider mean estimation of concentrated random variables under local DP. The algorithm will also be an instantiation of Algorithm 1. First, we describe the crude mean estimator we will use for the first step. The algorithm can be viewed as a high-dimensional extension of Gaboardi et al. (2019), which focuses on the one-dimensional case.

Algorithm 2 LDP Range - scalar

Input: $X^k = (X_1, X_2, \dots X_k) \in [-R, R]$; ε : privacy level; concentration radius τ.

1: Divide $[-R, R]$ into $t = R/\tau$ nonoverlapping intervals of width 2τ, denoted as $I_1, I_2, \dots, I_t$.

2: $\forall i \in [k]$, let Y_i be a t-dimensional vector with $\forall j \in [t]$,

$$Y_i(j) = \begin{cases} \mathbb{1}\{X_i \in I_j\} & \text{with prob } \frac{e^{\varepsilon/2}}{e^{\varepsilon/2}+1}, \\ 1 - \mathbb{1}\{X_i \in I_j\} & \text{with prob } \frac{1}{e^{\varepsilon/2}+1}. \end{cases}$$

3: Let $\bar{Y} = \sum_{i=1}^{k} Y_i$.

4: **Output:** the middle point of I_{j^*} where

$$j^* = \arg\max_{j\in[t]} \bar{Y}(j).$$

Algorithm 3 LDP Range - High Dim

Input: $X^k = (X_1, X_2, \dots X_k) \in \mathcal{B}_2^d(R)$; ε : privacy level.

1: Apply a random rotation matrix $R = H_d D$ on each X_i to get

$$X_i' = RX_i.$$

where H_d is a d-dimensional Hadamard matrix and D is a diagonal matrix with Rademacher entries ($+1$ or -1 with equal probability).

2: Divide k users into d non-overlapping groups $G_1, \dots, G_d$ with equal size.

3: For $j \in [d]$, compute

$$\tilde{\mu}'(j) = \textbf{LDPRange1D}(\{X_t'\}_{t\in G_j}, \varepsilon).$$

4: Let $\tilde{\mu}' = (\tilde{\mu}'(1), \tilde{\mu}'(2), \dots, \tilde{\mu}'(d))$. **Return**

$$\tilde{\mu} = R^{-1}\tilde{\mu}'.$$

The algorithm is described in Algorithm 3 and the performance is stated in Lemma 4.2. The one-dimensional version of the algorithm is stated in Algorithm 2.

In the high dimensional case, we apply a random rotation on the data and estimate the range on each dimension separately. The algorithm is stated in Algorithm 3 and the guarantee is stated in Lemma 4.2.

Lemma 4.2. *There exists an (ε, δ)-LDP algorithm $\tilde{\mu}$ such that when $k \geq 4d\log(\sqrt{d^3}R/(\tau\gamma))/\varepsilon^2$ and X^k is (τ, γ)-concentrated, we have with probability at least $1 - 2\gamma$,*

$$\|\tilde{\mu}(X^k) - \mu(X^k)\| = O\Big(\tau\sqrt{\log(dk/\gamma)}\Big).$$

With the guarantee of Algorithm 3, we are ready to state the guarantee for LDP mean estimation.

Theorem 4.3. *For $\varepsilon \in (0, 1)$ and $\delta \in (0, \frac{1}{k})$, let $\mathcal{A}$ be the Algorithm 1 with following instantiation: i) Use Algorithm 3 with $\varepsilon' = \varepsilon/2$ as **CrudeMean**; ii) $C = \tau\sqrt{\log(dn/\gamma)}$; iii) $\sigma = \frac{C\sqrt{8\log(1.25/\delta)}}{\varepsilon}$. Then $\mathcal{A}$ is (ε, δ)-LDP. When X^k is (τ, γ)-concentrated and $n \geq 4d\log(\sqrt{d^3}R/(\tau\gamma))/\varepsilon^2$, we have*

$$\mathcal{A}(X^k) \sim_{(0,2\gamma)} \frac{1}{k}\sum_{i=1}^{k} X_i + \mathcal{N}(0, \frac{\sigma^2}{k}\mathbb{I}_d).$$

We leave the proof of Lemma 4.2 and Theorem 4.3 to Appendix B.

5. User-level DP-SCO with Central DP

Here we describe our central DP algorithm, detailed in Algorithm 4, and prove the guarantee stated in Theorem 3.1.

Proof of Theorem 3.1: The privacy guarantee of the algorithm holds since by Theorem 2.5, the algorithm in each round satisfies $(\frac{\varepsilon}{\sqrt{2T\log(2/\delta)}}, \frac{\delta}{2T})$-DP, and choosing $\delta' = \delta/2, k = T$ in Theorem 2.4 leads to final privacy guarantee.

To prove the utility guarantee, we show that with high probability, in each round, the gradient estimate $\tilde{\nabla}F(\theta_t)$ is the same as a stochastic gradient oracle, stated below.

Lemma 5.1. *Let $(\theta_0, \ldots, \theta_T)$ be the parameter trajectory of Algorithm 4, denoted by $\mathcal{A}$. Let $(\theta'_0, \ldots, \theta'_T)$ be the parameter trajectory of $\mathcal{A}'$ where $\forall t \in [T]$, $\theta'_{t+1} \sim \Pi_\Theta\Big(\theta'_t - \eta\tilde{\nabla}F'(\theta'_t)\Big)$, where*

$$\tilde{\nabla}F'(\theta'_t) \sim \frac{1}{B}\sum_{i\in[B]} g_i(\theta'_t) + \mathcal{N}(0, \sigma^2\mathbb{I}_d),$$

with $\sigma^2 = O\Big(\frac{L^2\log^2(n/\gamma)\log(n/\delta)}{mB^2}\Big)$. When $\varepsilon \in (0, 1), \delta \in (0, \frac{d}{4n^{5/2}\sqrt{m}\varepsilon^2})$, the trajectories satisfy

$$(\theta_0, \ldots, \theta_T) \sim_{(0,\gamma)} (\theta'_0, \ldots, \theta'_T)$$

with $\gamma = 1/\sqrt{mn}$.

Then the proof follows similar as other SCO algorithms based on stochastic gradient oracles (*e.g.,* (Bassily et al., 2019)). We first prove Theorem 3.1 based on Lemma 5.1, and then give the proof of Lemma 5.1.

Algorithm 4 User-level private noisy SGD

Input: n users, each with m *i.i.d.*samples from P. Privacy parameter ε, δ. Lipschitz parameter L, parameter set Θ with radius R.
$\mathcal{M}_{\mathrm{DP}}$: private mean estimation algorithm in Theorem 4.1.

1: Initialize $\theta_0 = \vec{0}$.
2: Take $T = \frac{n^2\varepsilon^2}{c^2_{n,\delta}d}$ with $c_{n,\delta} = \Theta(\log n\log(1/\delta))$
$\eta = \frac{R}{L}\Big(\frac{d\sqrt{m}}{n^{3/2}\varepsilon^2} + \frac{c_{n,\delta}\sqrt{md}}{n\varepsilon}\Big)$, $\varepsilon_0 = 1$,$\delta_0 = \frac{n\delta}{2TB}$, $B = \frac{n\varepsilon}{(e-1)\sqrt{2T\log(2/\delta)}}$.
3: **for** $t = 0, 1, 2, \ldots, T-1$ **do**
4: Choose a random subset S_t of users with size B without replacement.
5: Compute the average gradient at each user at θ_t, $\forall i \in S_t$,
$$g_i(\theta_t) = \frac{1}{m}\sum_{j=1}^{m}\nabla\ell(\theta_t, Z_{i,j}).$$
6: Compute a noisy version of the average gradients using $\mathcal{M}$ and get
$$\tilde{\nabla}F(\theta_t) = \mathcal{M}_{\mathrm{DP}}(\{g_i(\theta_t)\}_{i\in S_t}, \varepsilon_0, \delta_0).$$
7: Update the parameter with
$$\theta_{t+1} = \Pi_\Theta\Big(\theta_t - \eta\tilde{\nabla}F(\theta_t)\Big).$$
8: **end for**
9: **Return:** $\bar{\theta}_T = \frac{1}{T}\sum_{t=1}^{T}\theta_t$.

By Lemma 5.1, we have $\mathcal{A}(Z^n) \sim_{(0,\gamma)} \mathcal{A}'(Z^n)$, and hence

$$\begin{aligned}
&\mathbb{E}\left[F(\mathcal{A}(Z^n)) - \min_{\theta\in\Theta}F(\theta)\right] \\
&\leq \mathbb{E}\left[F(\mathcal{A}'(Z^n)) - \min_{\theta\in\Theta}F(\theta)\right] + \gamma RL \\
&\leq \mathbb{E}\left[F(\mathcal{A}'(Z^n)) - \min_{\theta\in\Theta}F(\theta)\right] + \frac{RL}{\sqrt{nm}}.
\end{aligned}$$

Hence it would be enough to prove

$$\mathbb{E}\left[F(\mathcal{A}'(Z^n)) - \min_{\theta\in\Theta}F(\theta)\right] = \tilde{O}\left(\frac{RL}{\sqrt{nm}} + \frac{RL\sqrt{d}}{\sqrt{mn}\varepsilon}\right).$$

Let $\hat{F}(\theta) = \sum_{i\in[n]}\sum_{j\in[m]}\ell(\theta, Z_{i,j})$. Let $\tilde{\nabla}F'(\theta_t)$ be the gradient estimate of $\mathcal{A}'$ as defined in Lemma 5.1, we have

$\mathbb{E}\left[\tilde{\nabla}F'(\theta_t)\right] = \nabla\hat{F}(\theta)$, and

$$\mathbb{E}\left[\|\tilde{\nabla}F'(\theta_t) - \nabla\hat{F}(\theta)\|_2^2\right] \le \frac{L^2}{B} + d\sigma^2.$$

Hence by standard analysis of stochastic gradient descent for smooth functions (*e.g.,* (Bubeck, 2014)), we have

$$\mathbb{E}\left[\hat{F}(\mathcal{A}'(Z^n)) - \min_{\theta\in\Theta}\hat{F}(\theta)\right] \le \frac{\beta R^2}{T} + \frac{R^2}{\eta T} + \frac{\eta}{2}\left(\frac{L^2}{B} + d\sigma^2\right).$$

Next we bound the generalization error. The generalization analysis follows similarly as the stability-based analysis in Bassily et al. (2019, Lemma 2.2 and 3.4). When $\eta \le 2/\beta$ (this holds for the parameter range stated in Theorem 4.1), the generalization error can be shown to be upper bounded by $L^2\frac{T\eta}{nm}$. Hence we have

$$\mathbb{E}\left[F(\mathcal{A}'(Z^n)) - \hat{F}(\mathcal{A}'(Z^n))\right] \le L^2\frac{T\eta}{nm}.$$

Combining the above two inequalities, we have

$$\begin{aligned}
&\mathbb{E}\left[F(\mathcal{A}'(Z^n)) - \min_{\theta\in\Theta}F(\theta)\right] \\
&\le \mathbb{E}\left[F(\mathcal{A}'(Z^n)) - \min_{\theta\in\Theta}\hat{F}(\theta)\right] \\
&\le \mathbb{E}\left[\hat{F}(\mathcal{A}'(Z^n)) - \min_{\theta\in\Theta}\hat{F}(\theta)\right] + \\
&\qquad \mathbb{E}\left[F(\mathcal{A}'(Z^n)) - \hat{F}(\mathcal{A}'(Z^n))\right] \\
&\le \frac{\beta R^2}{T} + \frac{R^2}{\eta T} + \frac{\eta}{2}\left(\frac{L^2}{B} + d\sigma^2\right) + L^2\frac{T\eta}{nm}.
\end{aligned}$$

Plugging in the values of the parameters in Algorithm 4 and Lemma 5.1, we get when $\beta \le \frac{2L}{R}\min\{\frac{n^{3/2}\varepsilon^2}{d\sqrt{m}}, \frac{n\varepsilon}{c_{n,\delta}\sqrt{md}}\}$

$$\mathbb{E}\left[F(\mathcal{A}'(Z^n)) - \min_{\theta\in\Theta}F(\theta)\right] = O\left(\frac{RL}{\sqrt{nm}} + c_{n,\delta}\cdot\frac{RL\sqrt{d}}{\sqrt{m}n\varepsilon}\right), \tag{2}$$

where $c_{n,\delta}$ is as defined in Algorithm 4.

□

Proof of Lemma 5.1: By union bound, it would be enough to show that for all $t \in [T]$, we have

$$\tilde{\nabla}F(\theta_t) \sim_{(0,\gamma/T)} \tilde{\nabla}F'(\theta_t).$$

Note that $\forall i, g_i(\theta)$ is $L/\sqrt{m}$-subgaussian. And hence $\forall S \subset [n]$ and $|S| = B$, $\{g_i(\theta)\}_{i\in S_t}$ is $(L\sqrt{2\log(8eTn/\gamma)/m}, \gamma/(4eT))$-concentrated. Since θ_t is (ε,δ)-private with respect to S_t, by Theorem 3.4, we have $\{g_i(\theta_t)\}_{i\in S_t}$ is $(L\sqrt{2\log(8eTn/\gamma)/m}, \gamma')$ concentrated with $\gamma' = e^{\varepsilon}\gamma/(4eT) + \delta \le \gamma/2T$, where we use the fact that in the required parameter range, $\delta \le \gamma/4T$.

Moreover, $B = \frac{n\varepsilon}{(e-1)\sqrt{2T\log(2/\delta)}} = \tilde{\Omega}(\sqrt{d}/\varepsilon_0)$. Hence by Theorem 4.1, we obtain the desired bound in Lemma 5.1.
□

6. User-level DP-SCO with Local DP

Here we describe the details of our LDP algorithm in Algorithm 5 and prove the guarantee stated in Theorem 3.2. The proof of Theorem 3.3 will be in Appendix A.

Algorithm 5 also relies on private mean estimation primitives (under the more strigent LDP setting) to obtain the a gradient estimate at each round. However, compared to Algorithm 4, there are two main differences: (1) Non-overlapping batches of users are used in each round and the number of gradient queries is nm, linear in the total number of samples; (2) The gradient update rule is based on accelerated gradient methods instead of SGD (Cotter et al., 2011). This leads to a faster convergence rate and a smaller smoothness parameter is required.

Algorithm 5 User-level LDP SCO

Input: n users, each with m *i.i.d.*samples from P. Privacy parameter ε. Lipschitz parameter L, parameter set Θ with radius R.
$\mathcal{M}_{\mathrm{LDP}}$: private mean estimation algorithm in Theorem 4.3.

1: Initialize $\theta_0 = \vec{0}$. and $\theta^{ag} = \theta_0$.
2: Take $T = n\varepsilon^2/d$, and $\{\eta_t, \gamma_t\}_{t\in[T]}$ as in Lemma 6.1.
3: **for** $t = 0, 1, 2, \ldots, T-1$ **do**
4: Compute $\theta_t^{md} = \gamma_t^{-1}\theta_t + (1-\gamma_t^{-1})\theta_t^{ag}$.
5: Choose a fresh batch S_t of $n_0 = \lfloor n/T \rfloor$ users.
6: Compute the average gradient at each user at θ_t, $\forall i \in S_t$,

$$g_i(\theta_t^{md}) = \frac{1}{m}\sum_{j=1}^{m}\nabla\ell(\theta_t^{md}, Z_{i,j}).$$

7: Compute a noisy version of the average gradients using $\mathcal{M}_{\mathrm{LDP}}$ and get

$$\tilde{\nabla}F(\theta_t^{md}) = \mathcal{M}_{\mathrm{LDP}}\left(\{g_i(\theta_t^{md})\}_{i\in S_t}, \varepsilon, \delta\right).$$

8: Update $\theta_{t+1} = \theta_t^{md} - \eta_t\tilde{\nabla}F(\theta_t^{md})$.
9: Compute $\theta_{t+1}^{ag} = \gamma_t^{-1}\theta_{t+1} + (1-\gamma_t^{-1})\theta_t^{ag}$.
10: **end for**
11: **Return:** θ_T^{ag}.

Proof of Theorem 3.2: The privacy guarantee follows from the privacy guarantee of $\mathcal{M}_{\mathrm{LDP}}$ and the fact that the batches are not overlapping.

To prove the utility guarantee, similar to the proof of Theorem 3.1, we first show that the parameter trajectory is close to the parameter trajectory where the gradient estimate at each round is replaced by an unbiased stochastic

gradient oracle. More precisely, let $(\theta_1^{ag}, \theta_2^{ag}, \ldots, \theta_T^{ag})$ be the parameter trajector of Algorithm 5 (denoted by $\mathcal{A}$), and $(\theta_1^{'ag}, \theta_2^{'ag}, \ldots, \theta_T^{'ag})$ be the parameter trajectory of an algorithm $\mathcal{A}'$ which replaces the gradient estimate $\tilde{\nabla} F(\theta_t^{md})$ by

$$\tilde{\nabla} F'(\theta_t^{md}) \sim \frac{1}{n_0} \sum_{i \in S_t} g_i(\theta_t^{md}) + \mathcal{N}(0, \sigma^2 \mathbb{I}_d),$$

with $\sigma^2 = \tilde{O}(\frac{L^2}{n_0 m \varepsilon^2})$. By union bound and Theorem 4.3, we have

$$(\theta_1^{'ag}, \ldots, \theta_T^{'ag}) \sim_{(0,\gamma)} (\theta_1^{ag}, \ldots, \theta_T^{ag}),$$

with $\gamma = \sqrt{\frac{d}{mn\varepsilon^2}}$. Hence we have

$$\mathbb{E}\left[F(\theta_T^{ag})\right] \leq \mathbb{E}\left[F(\theta_T^{'ag})\right] + RL\sqrt{\frac{d}{mn\varepsilon^2}}.$$

Next we bound $\mathbb{E}\left[F(\theta_T^{'ag})\right]$. Note that since S_t's are disjoint. We have $\mathbb{E}\left[\tilde{\nabla} F'(\theta_t^{md})\right] = \nabla F(\theta_t^{md})$, and

$$\begin{aligned}\mathbb{E}\left[\|\tilde{\nabla} F'(\theta_t^{md}) - \nabla F(\theta_t^{md})\|_2^2\right] &\leq \frac{L^2}{n_0} + d\sigma^2 \\ &= \tilde{O}\left(\frac{dL^2}{n_0 m \varepsilon^2}\right),\end{aligned}$$

where we use $m < d/\varepsilon^2$. To move forward, we need the following guarantee for accelerated gradient method.

Lemma 6.1 ((Cotter et al., 2011; Lan, 2012)). *Suppose each $\tilde{\nabla} F(\theta)$ is an unbiased stochastic oracle to $\nabla F(\theta)$ with variance ν^2. If $F(\theta)$ is β-smooth, there exists settings of $\{\eta_t, \gamma_t\}_{t \in [T]}$ such that*

$$F(\theta_T^{'ag}) - \min_{\theta \in \Theta} F(\theta) = \tilde{O}\left(\beta \frac{R^2}{T^2} + \frac{R\nu}{\sqrt{T}}\right).$$

Plugging in the value of ν^2, $T = n\varepsilon^2/d$ and $n_0 = \lfloor n/T \rfloor$, we get

$$F(\theta_T^{ag}) - F(\theta^*) = O\left(\beta \frac{d^2 R^2}{n^2 \varepsilon^4} + RL\sqrt{\frac{d}{mn\varepsilon^2}}\right).$$

When $\beta = \tilde{O}\left(\frac{L\varepsilon^3}{R}\sqrt{\frac{n^3}{md^3}}\right)$, we get the desired rate in Theorem 3.2.

Discussion on the smoothness condition. When the smoothness assumption doesn't hold, using Moreau envelope smoothing method (Nesterov, 2005; Bassily et al., 2019), there exists a smoothed version of f, denoted by f_β, which is β-smooth and $2L$-Lipschitz for all $\theta \in \Theta$ and $\forall z, f(\theta, z) \leq f_\beta(\theta, z) \leq f(\theta, z) + \frac{L^2}{2\beta}$. Moreover, the gradient of f_β can be computed from f. Hence we can instead optimize f_{β_m} with

$$\beta_m = \frac{L^2}{RL\sqrt{\frac{d}{mn\varepsilon^2}}} = \frac{L}{R}\sqrt{\frac{mn\varepsilon^2}{d}}$$

and this won't affect the optimality result up to constants. When $n > \frac{dm}{\varepsilon^2}$, we have $\beta_m \leq \frac{L\varepsilon^3}{R}\sqrt{\frac{n^3}{md^3}}$, and hence the guarantee of Algorithm 5 discussed above can be used. □

Acknowledgements

Raef Bassily's research is supported by NSF CAREER Award 2144532, NSF Award AF-1908281, and NSF Award 2112471.

References

Arora, R., Bassily, R., Guzmán, C., Menart, M., and Ullah, E. Differentially private generalized linear models revisited. *arXiv preprint arXiv:2205.03014*, 2022.

Asi, H., Feldman, V., Koren, T., and Talwar, K. Private stochastic convex optimization: Optimal rates in l1 geometry. In *International Conference on Machine Learning*, pp. 393–403. PMLR, 2021.

Balle, B., Barthe, G., and Gaboardi, M. Privacy amplification by subsampling: Tight analyses via couplings and divergences, 2018.

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

Bassily, R., Nissim, K., Smith, A., Steinke, T., Stemmer, U., and Ullman, J. Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 1046–1059, 2016.

Bassily, R., Feldman, V., Talwar, K., and Thakurta, A. G. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, pp. 11279–11288, 2019.

Bassily, R., Guzmán, C., and Menart, M. Differentially private stochastic optimization: New results in convex and non-convex settings. *Advances in Neural Information Processing Systems*, 34:9317–9329, 2021a.

Bassily, R., Guzmán, C., and Nandi, A. Non-euclidean differentially private stochastic convex optimization. In

Conference on Learning Theory, pp. 474–499. PMLR, 2021b.

Biswas, S., Dong, Y., Kamath, G., and Ullman, J. Coinpress: Practical private mean and covariance estimation. *Advances in Neural Information Processing Systems*, 33: 14475–14485, 2020.

Bubeck, S. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

Cai, T. T., Wang, Y., and Zhang, L. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *arXiv preprint arXiv:1902.04495*, 2019.

Cotter, A., Shamir, O., Srebro, N., and Sridharan, K. Better mini-batch algorithms via accelerated gradient methods. *Advances in neural information processing systems*, 24, 2011.

Duchi, J. C. and Wainwright, M. J. Distance-based and continuum fano inequalities with applications to statistical estimation. *arXiv preprint arXiv:1311.2669*, 2013.

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

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

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. 2014.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. Generalization in adaptive data analysis and holdout reuse. *Advances in Neural Information Processing Systems*, 28, 2015.

Feldman, V. and Steinke, T. Generalization for adaptively-chosen estimators via stable median. In Kale, S. and Shamir, O. (eds.), *ICML*, volume 65 of *Proceedings of Machine Learning Research*, pp. 728–757, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.

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

Feldman, V., Koren, T., and Talwar, K. Private stochastic convex optimization: Optimal rates in linear time, 2020.

Feldman, V., McMillan, A., and Talwar, K. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 954–964, 2022. doi: 10.1109/FOCS52979.2021.00096.

Gaboardi, M., Rogers, R., and Sheffet, O. Locally private mean estimation: z-test and tight confidence intervals. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2545–2554. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/gaboardi19a.html.

Girgis, A. M., Data, D., and Diggavi, S. Distributed user-level private mean estimation. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pp. 2196–2201, 2022. doi: 10.1109/ISIT50566.2022.9834713.

Kairouz, P., Bonawitz, K., and Ramage, D. Discrete distribution estimation under local privacy. *arXiv preprint arXiv:1602.07387*, 2016.

Kamath, G., Li, J., Singhal, V., and Ullman, J. Privately learning high-dimensional distributions. In *Conference on Learning Theory*, pp. 1853–1902. PMLR, 2019.

Karwa, V. and Vadhan, S. Finite sample differentially private confidence intervals, 2017.

Kulkarni, J., Lee, Y. T., and Liu, D. Private non-smooth erm and sco in subquadratic steps. *Advances in Neural Information Processing Systems*, 34:4053–4064, 2021.

Lan, G. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

Levy, D. A. N., Sun, Z., Amin, K., Kale, S., Kulesza, A., Mohri, M., and Suresh, A. T. Learning with user-level privacy. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=G1jmxFOtY_.

Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

Smith, A. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 813–822, 2011.

Song, S., Steinke, T., Thakkar, O., and Thakurta, A. Evading the curse of dimensionality in unconstrained private glms. In *International Conference on Artificial Intelligence and Statistics*, pp. 2638–2646. PMLR, 2021.

Warner, S. L. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

A. Proof of Theorem 3.3

We will focus on the case when $m = \infty$, i.e., each user can draw as many local samples as possible. A lower bound for $m = \infty$ would naturally imply the same lower bound for any finite m as well.

Similar to (Bassily et al., 2014; Levy et al., 2021), consider the following class of linear function where

$$\ell(\theta, Z) = -\theta^T Z.$$

Without loss of generality, we assume $R = L = 1$ and prove a lower bound of $\Omega(1)$. Otherwise we can scale the domain of Z and θ by L and R respectively. Let P be drawn from the following class of point mass distributions. Let $v \in \{\pm 1\}^d$ and P_v is such that

$$P_v(z) = \begin{cases} 1 & \text{if } z = \frac{v}{\sqrt{d}} \\ 0 & \text{o.w.} \end{cases}$$

Note that this distribution is deterministic. Hence observing $m = \infty$ samples is equivalent to observing one sample. We will assum each user observes one sample from P_v in the rest of the proof. Then we have

$$F_v(\theta) := \mathbb{E}_{Z \sim P_v}\left[\ell(\theta, Z)\right] = -\frac{\theta^T v}{\sqrt{d}}.$$

Then it can be verified that under P_v, the loss is minimized when $\theta_v = \frac{v}{\sqrt{d}}$. Moreover, for any $\theta \in \Theta$,

$$F(\theta) - F(\theta_v) = \frac{(\theta_v - \theta)^T v}{\sqrt{d}} = 1 - \frac{\theta^T v}{\sqrt{d}} \geq \frac{1}{2}\|\theta - \theta_v\|_2^2.$$

Hence it would be enough to show that for any user-level private algorithm $\mathcal{A}$, there exists a v such that

$$\mathbb{E}\left[\|\mathcal{A}(Z^n) - \theta_v\|_2^2\right] = \Omega(1).$$

We proceed by using proof by contradiction. Suppose there exists an algorithm $\mathcal{A}$ such that $\forall v \in \{\pm 1\}^d$,

$$\mathbb{E}\left[\|\mathcal{A}(Z^n) - \theta_v\|_2^2\right] \leq \frac{1}{200}. \tag{3}$$

We allow the LDP protocol to be sequentially interactive in our proof. More specifically, consider the following process

- Draw V uniformly at random from $\{+1, -1\}^d$.
- n users each observe one sample from P_V, denoted by $(Z_1, \dots Z_n)$.
- n users come in sequence and the ith user observes $Y^{i-1} := (Y_1, \dots, Y_{i-1})$ and sends message Y_i, which is a private randomization of
$$Y_i = \mathcal{R}^{Y^{i-1}}(Z_i).$$
- The server observed Y^n and makes an inference.

The next lemma shows that any algorithm with small error for the mean estimation, we must be able to extract enough information about Z from the messages Y^n.

Lemma A.1. *Suppose there exists $\mathcal{A}$ such that Equation (3) holds, we must have*

$$I(V; Y^n) = \Omega(d).$$

Proof. Let $\hat{V} = \arg\max_{v'}\{\|\mathcal{A}(Z^n) - \theta'_v\|_2^2\}$. If (3) holds, we have

$$\mathbb{E}\left[d_{\text{ham}}(V, \hat{V})\right] = d\mathbb{E}\left[\|\theta_V - \theta_{V'}\|_2^2\right] \leq 2d\mathbb{E}\left[\|\theta_V - \mathcal{A}(Z^n)\|_2^2\right] \leq \frac{d}{100}.$$

Hence we have

$$\Pr\left(d_{\text{ham}}(V,\hat{V}) \geq \frac{d}{50}\right) \leq \frac{1}{2}.$$

By Fano's inequality (*e.g.,*, the distance-based variant in (Duchi & Wainwright, 2013) (Corollary 1)), we have

$$\Pr\left(d_{\text{ham}}(V,\hat{V}) \geq \frac{d}{100}\right) > 1 - \frac{\log 2 + I(V;Y^n)}{d/200}.$$

Combining the above equations completes the proof. □

Next we prove a contradiction when $n = o(d/\varepsilon^2)$. By chain rule of mutual information, we have

$$\begin{aligned}
I(V;Y^n) &= \sum_{i=1}^{n} I(V;Y_i \mid Y^{i-1}) \\
&= \sum_{i=1}^{n} \mathbb{E}_{Y^{i-1}}\left[I(V;\mathcal{R}^{Y^{i-1}} \circ Z_i \mid Y^{i-1})\right] \\
&\leq \sum_{i=1}^{n} \mathbb{E}_{Y^{i-1}}\left[\max_{\mathcal{R}:\varepsilon-LDP} I(V;\mathcal{R} \circ Z_i \mid Y^{i-1})\right] \\
&= \sum_{i=1}^{n} \mathbb{E}_{Y^{i-1}}\left[\max_{\mathcal{R}:\varepsilon-LDP} I(V;\mathcal{R} \circ Z_i \mid Y^{i-1})\right] \\
&= \sum_{i=1}^{n} \mathbb{E}_{Y^{i-1}}\left[\max_{\mathcal{R}:\varepsilon-LDP} \mathbb{E}_{V|Y^{i-1}}\left[\text{KL}(\mathcal{R} \circ P(Z_i \mid V)||\mathcal{R} \circ P(Z_i \mid Y^{i-1}))\right]\right]
\end{aligned}$$

It has been shown in (Duchi et al., 2013) (Theorem 1) that for any ε-LDP $\mathcal{R}$ and distributions P_1, P_2, we have

$$\text{KL}(\mathcal{R} \circ P_1||\mathcal{R} \circ P_2) = O\big(\varepsilon^2\big).$$

Combining the above, we get:

$$I(Z;Y^n) = O\big(n\varepsilon^2\big).$$

With Lemma A.1, we get for any algorithm such that Equation (3) holds, we must have

$$n = \Omega\left(\frac{d}{\varepsilon^2}\right).$$

B. Details of the mean estimation algorithms in Section 4

B.1. User-level private mean estimation algorithm in Levy et al. (2021).

The algorithm follows a similar procedure as Algorithm 1 and the crude mean estimator is also based on the combination of random rotation and one-dimensional estimation as stated in Algorithm 3 except for that it works in the central model. Hence here we only state the one-dimensional range estimation algorithm in Algorithm 6.

B.2. User-level LDP mean estimation

The proof of Lemma 4.2 and Theorem 4.3 relies on the guarantees of Algorithm 2, which is the one-dimensional version of Algorithm 3. The guarantees of Algorithm 2 are stated in the following lemma:

Lemma B.1. *Algorithm 2 is an $(\varepsilon, 0)$-LDP algorithm. Let $\tilde{\mu}(X^k)$ denote its output. When X^k is $(\tau, \gamma/2)$-concentrated and $k > 4\log(R/(\tau\gamma))/\varepsilon^2$, we have with probability at least $1-\gamma$,*

$$|\tilde{\mu}(X^k) - \mu(X^k)| \leq 4\tau.$$

Algorithm 6 PrivateRange($X^n, \varepsilon, \tau, B$): Private Range Estimation (Feldman & Steinke, 2017)

Input: $X^n := (X_1, X_2, ..., X_n) \in [-B, B]^n$, τ : concentration radius, privacy parameter $\varepsilon > 0$.

1: Divide the interval $[-B, B]$ into $l = B/\tau$ disjoint bins, each with width 2τ[1]. Let T be the set of middle points of intervals.

2: $\forall i \in [n]$, let $X_i' = \min_{x \in T} |X_i - x|$ be the point in T closest to X_i.

3: $\forall x \in T$, define cost function

$$c(x) = \max\{|\{i \in [n] \mid X_i' < x\}|, |\{i \in [n] \mid X_i' > x\}|\}.$$

4: Sample $x \in T$ based on the following distribution:

$$\Pr(\hat{\mu} = x) = \frac{e^{-\varepsilon c(x)/2}}{\sum_{x' \in T} e^{-\varepsilon c(x')/2}}.$$

5: Return $R = [\hat{\mu} - 2\tau, \hat{\mu} + 2\tau]$.

Proof. First, the privacy guarantee is straightforward and follows from the privacy guarantee of the randomized response mechanism of LDP (Warner, 1965; Kairouz et al., 2016). Particularly, fix any $i \in [k]$ and consider any pair $X_i, X_i' \in [-R, R]$. Let Y_i and Y_i' be the corresponding randomized versions of X_i and X_i', respectively, as generated by step 2. Note that for any $\mathbf{b} \in \{0,1\}^t$,

$$\frac{\mathbb{P}[Y_i = \mathbf{b}]}{\mathbb{P}[Y_i' = \mathbf{b}]} \in [e^{-\varepsilon}, e^{\varepsilon}].$$

Next, we prove the accuracy guarantee. Suppose X^k is $(\tau, \gamma/2)$-concentrated. For each $i \in [k]$ and each $j \in [t]$, let $b_i(j) \triangleq \mathbb{1}\{X_i \in I_j\}$. Note that by the concentration property of X^k, there exists $j' \in [t-1]$ such that $\sum_{i=1}^k (b_i(j') + b_i(j'+1)) = k$ (and hence, $\sum_{i=1}^k b_i(j) = 0$ for all $j \in [t] \setminus \{j', j'+1\}$). Fix any $j \in [t]$. Note that $(Y_1(j), \ldots, Y_k(j))$ is a sequence of independent Bernoulli random variables with means $c_\varepsilon (b_1(j), \ldots, b_k(j))$, where $c_\varepsilon = \frac{e^{\varepsilon/2}-1}{e^{\varepsilon/2}+1} \approx \varepsilon$. Thus, by Chernoff's bound together with the union bound over $j \in [t]$, with probability at least $1 - \gamma$, the following conditions are simultaneously satisfied:

$$\sum_{i=1}^k (Y_i(j') + Y_i(j'+1)) \geq c_\varepsilon k - \sqrt{k \log(t/\gamma)},$$

$$\sum_{i=1}^k Y_i(j) < \sqrt{k \log(t/\gamma)}, \ \forall j \in [t] \setminus \{j', j'+1\}.$$

Now, since $k > 4\frac{\log(t/\gamma)}{\varepsilon^2}$ then the lower bound in the first event above $k - \sqrt{k \log(t/\gamma)}$ is greater than the upper bound in the second event $\sqrt{k \log(t/\gamma)}$. Thus, with probability at least $1 - \gamma$, we must have $j^* \in \{j', j'+1\}$, where $j^* = \arg\max_{j \in [t]} \sum_{i=1}^k Y_i(j)$ is the index obtained in the final step of the algorithm. Letting $\tilde{\mu}(X^k)$ be the mid-point of I_{j^*} (which is the output of the algorithm), we must then have $|\tilde{\mu}(X^k) - \mu(X^k)| \leq |I_{j'}| + |I_{j'+1}| = 4\tau$.

□

Given Lemma B.1, we now give a proof for Lemma 4.2.

Proof of Lemma 4.2: Since X^k is $(\tau, \gamma/2)$-concentrated, by definition, there exists x_0 such that with probability at least $1 - \gamma/2$, we have

$$\max_{i \in [k]} \|X_i - x_0\|_2 \leq \tau.$$

Under this event, the random rotation step (Step 1) in Algorithm 3 guarantees that X_i' are concentrated along each direction. More specifically, in Levy et al. (2021, Lemma 2), we have that with probability at least $1 - \gamma/4$

$$\max_{i \in [k]} \|X_i' - R x_0\|_\infty \leq \frac{10 \max_{i \in [k]} \|X_i - x_0\|_2 \sqrt{\log(4kd/\gamma)}}{\sqrt{d}}.$$

Hence by union bound, with probability at least $1 - 3/4\gamma$, we have

$$\max_{i\in[k]} \|X_i' - Rx_0\|_\infty \le \frac{10\tau\sqrt{\log(4kd/\gamma)}}{\sqrt{d}},$$

We denote the right hand side bound as τ'. Applying the guarantee of Lemma B.1 with $\tau = \tau'$ and $\gamma = \gamma/(4d)$ on each dimension, we get by union bound, when $k \ge 4d\log(\sqrt{d^3}R/(\tau\gamma))/\varepsilon^2$, we have with probability $1 - \gamma/4$,

$$\|\tilde{\mu} - \mu(X^k)\|_\infty \le 4\max_{i\in[k]} \|X_i' - Rx_0\|_\infty.$$

Hence, by union bound, we have with probability at least $1 - \gamma$,

$$\|\tilde{\mu} - \mu(X^k)\|_\infty = O\Big(\tau\sqrt{\log(kd/\gamma)}\Big)$$

Proof of Theorem 4.3 The privacy guarantee follows from the privacy guarantee of Gaussian mechanism. And the utility guarantee follows from that by union bound, with probability at least $1 - 2\gamma$,

$$\max_{i\in[k]} \|X_i - \tilde{\mu}(X^k)\| \le \max_{i\in[k]} \|X_i - \mu(X^k)\| + \|\tilde{\mu}(X^k) - \mu(X^k)\| \le O\Big(\tau\sqrt{\log(dn/\gamma)}\Big).$$

Note that when the above is true, the averaged clipped mean is the same as the actual mean.