Differential Private Stochastic Optimization with Heavy-tailed Data: Towards Optimal Rates

Puning Zhao[1] Jiafei Wu[1*], Zhe Liu[1], Chong Wang[2], Rongfei Fan[3], Qingming Li[4]

[1]School of Cyber Science and Technology, Sun Yat-sen University, Shenzhen, China
[2] Ningbo University, Ningbo, China
[3] Beijing Institute of Technology, Beijing, China
[4] Zhejiang University, Hangzhou, China
{pnzhao,wujiafei,zhe.liu}@zhejianglab.org, wangchong@nbu.edu.cn, fanrongfei@bit.edu.cn, liqm@zju.edu.cn

Abstract

We study convex optimization problems under differential privacy (DP). With heavy-tailed gradients, existing works achieve suboptimal rates. The main obstacle is that existing gradient estimators have suboptimal tail properties, resulting in a superfluous factor of d in the union bound. In this paper, we explore algorithms achieving optimal rates of DP optimization with heavy-tailed gradients. Our first method is a simple clipping approach. Under bounded p-th order moments of gradients, with n samples, it achieves $\tilde{O}(\sqrt{d/n} + \sqrt{d}(\sqrt{d}/n\epsilon)^{1-1/p})$ population risk with $\epsilon \leq 1/\sqrt{d}$. We then propose an iterative updating method, which is more complex but achieves this rate for all $\epsilon \leq 1$. The results significantly improve over existing methods. Such improvement relies on a careful treatment of the tail behavior of gradient estimators. Our results match the minimax lower bound, indicating that the theoretical limit of stochastic convex optimization under DP is achievable.

Introduction

Differential privacy (DP) (Dwork et al. 2006) is a prevailing framework for privacy protection. In recent years, significant progress has been made on deep learning under DP (Abadi et al. 2016; Tramer and Boneh 2021; Wei et al. 2022; De et al. 2022). While the practical performance continues to improve, the theoretical analysis lags behind. Existing analyses focus primarily on Lipschitz loss functions, such that the gradients are all bounded (Bassily, Smith, and Thakurta 2014; Bassily et al. 2019; Iyengar et al. 2019; Bassily and Sun 2023). However, many empirical studies have shown that in deep learning, gradient noise usually follows heavy-tailed distributions (Simsekli, Sagun, and Gurbuzbalaban 2019; Şimşekli et al. 2019; Zhang et al. 2020; Gurbuzbalaban, Simsekli, and Zhu 2021; Sha et al. 2024). To bridge the gap between theory and practice, it is worth investigating the DP stochastic optimization problem with heavy tails.

It has been shown in (Kamath, Liu, and Zhang 2022) that if the stochastic gradients have bounded p-th order moments for some $p \geq 2$, then the minimax lower bound of optimization risk is $\Omega\left(\sqrt{d/n} + \sqrt{d}\left(\sqrt{d}/\epsilon n\right)^{1-1/p}\right)$ under (ϵ, δ)-DP, which can be viewed as the theoretical limit of DP optimization. However, existing methods fail to achieve this rate. Compared with Lipschitz loss functions, a crucial challenge in analyzing heavy-tailed gradients is the design of an efficient mean estimator under DP. Various methods for DP mean estimation have been proposed (Huang, Liang, and Yi 2021; Hopkins, Kamath, and Majid 2022; Kamath, Singhal, and Ullman 2020; Liu et al. 2021). These methods have achieved optimal mean squared error, but the high probability bounds are not optimal. To bound the risk of optimization, we need a union bound of the bias and variance of gradient estimate over the whole hypothesis space. Therefore, a suboptimal high probability bound of mean estimation results in a suboptimal risk of optimization. To be best of our knowledge, currently, it is unknown whether the minimax lower bound shown in (Kamath, Liu, and Zhang 2022) is achievable.

In this paper, we answer this question affirmatively. We propose two methods, called the *simple clipping* method and the *iterative updating* method, respectively. For both methods, we derive the high probability bounds of mean estimation first, and then analyze the risk of optimization.

1) Simple clipping. This method just clips all gradients to a given radius R and calculates sample averages. R can be tuned based on the privacy requirement ϵ, δ and the number of samples n. Our analysis shows that the population risk is $\tilde{O}\left(\sqrt{d/n} + \sqrt{d}\left(\sqrt{d}/\epsilon n\right)^{1-1/p} + d^{\frac{3}{2}-\frac{1}{p}}/n^{1-\frac{1}{p}}\right)$, which improves over existing methods. This rate matches the minimax lower bound if $\epsilon \leq 1/\sqrt{d}$, under which the third term $d^{\frac{3}{2}-\frac{1}{p}}/n^{1-\frac{1}{p}}$ does not dominate. The key of such improvement is that we treat the tail behavior of mean estimation more carefully. In particular, we show that the mean estimation has a subexponential tail in all directions, which refines the union bounds and eventually leads to the risk bound mentioned above. The remaining drawback is that this method has an additional term $d^{\frac{3}{2}-\frac{1}{p}}/n^{1-\frac{1}{p}}$. Therefore, this method is suboptimal if $\epsilon > 1/\sqrt{d}$.

2) Iterative updating. This method is proposed to remove the additional term of the simple clipping method. It divides

*Corresponding author.

Source	Bound of risk
(Wang et al. 2020)	$\tilde{O}\left(\left(\frac{d^3}{\epsilon^2 n}\right)^{\frac{1}{3}}\right)$ [1]
(Kamath et al. 2022)	$\tilde{O}\left(\frac{d}{\sqrt{n}}+\sqrt{d}\left(\frac{d^{3/2}}{\epsilon n}\right)^{1-\frac{1}{p}}\right)$
(Kamath et al. 2022)	$\tilde{O}\left(\min_{0.5\le q\le 2}\left(\frac{d^{\frac{3-q}{2}}}{\sqrt{n}}+\frac{d^{\frac{1+q}{2}}}{\epsilon^{\frac{1}{2}}\sqrt{n}}\right)\right)$ [2]
Simple clipping	$\tilde{O}\left(\sqrt{\frac{d}{n}}+\sqrt{d}\left(\frac{\sqrt{d}}{\epsilon n}\right)^{1-\frac{1}{p}}+\frac{d^{\frac{3}{2}-\frac{1}{p}}}{n^{1-\frac{1}{p}}}\right)$
Iterative updating	$\tilde{O}\left(\sqrt{\frac{d}{n}}+\sqrt{d}\left(\frac{\sqrt{d}}{\epsilon n}\right)^{1-\frac{1}{p}}\right)$
Lower bound	$\Omega\left(\sqrt{\frac{d}{n}}+\sqrt{d}\left(\frac{\sqrt{d}}{\epsilon n}\right)^{1-\frac{1}{p}}\right)$

Table 1: Comparison of risk bounds of stochastic optimization under (ϵ, δ)-DP with p-th order bounded moments on gradients. Logarithmic factors are omitted here.

the data into k groups. For each group, this method calculates the group-wise mean and adds noise to meet DP requirements. After that, the mean estimate is iteratively updated based on the estimation of distances and directions to the ground truth $\nabla F(\mathbf{w}_t)$. Such design is inspired by several existing methods for non-private mean estimation with heavy-tailed data (Lugosi and Mendelson 2019b; Cherapanamjeri, Flammarion, and Bartlett 2019; Lei et al. 2020; Depersin and Lecué 2022). Compared with the simple clipping approach, this method improves the tail behavior of the mean estimator from subexponential to subgaussian. Moreover, this method is invariant to permutations of groups. As a result, the overall privacy of the final estimate is amplified compared with the privacy of each group (Erlingsson et al. 2019; Feldman, McMillan, and Talwar 2022). With this new algorithm and refined theoretical analysis, we achieve a risk bound $\tilde{O}\left(\sqrt{d/n}+\sqrt{d}\left(\sqrt{d}/\epsilon n\right)^{1-1/p}\right)$, matching the minimax lower bound.

Our results and comparison with existing works are summarized in Table 1.

Related Work

DP optimization. Early works focus on empirical risk minimization (ERM) under DP, which is a relatively simpler problem compared with stochastic optimization, such as (Chaudhuri and Monteleoni 2008; Kifer, Smith, and Thakurta 2012; Thakurta and Smith 2013; Bassily, Smith, and Thakurta 2014; Wang, Ye, and Xu 2017; Zhang, Mironov, and Hejazinia 2021). For stochastic optimization problem, (Bassily et al. 2019) shows that for Lipschitz loss functions, DP-SGD is minimax optimal with proper parameter selection. The analysis is then improved in several later works on time complexity (Feldman, Koren, and Talwar 2020; Kulkarni, Lee, and Liu 2021) and extended to different geometries (Asi et al. 2021; Bassily, Guzmán, and Nandi 2021). For heavy-tailed gradients, the non-private optimization has been widely studied (Pascanu, Mikolov, and Bengio 2012; Zhang et al. 2020; Gorbunov, Danilova, and Gasnikov 2020; Parletta et al. 2022; Liu et al. 2023; Nguyen et al. 2023; Eldowa and Paudice 2024; Liu and Zhou 2023; Armacki et al. 2023; Koloskova, Hendrikx, and Stich 2023). Under DP, (Wang et al. 2020) provides the upper bound of DP-SGD under the assumption that gradients have bounded second moments. (Hu et al. 2022) analyzes the sparse setting. (Kamath, Liu, and Zhang 2022) improves on the risk bound. Moreover, (Das et al. 2023) weakens the uniform Lipschitz assumption to a sample-wise one. (Lowy and Razaviyayn 2023) further discusses the case with large worst-case Lipschitz parameters.

Mean estimation with subgaussian rates. Non-private mean estimation for heavy-tailed distributions has received widespread attention (Lugosi and Mendelson 2019a). We hope to minimize the $1-\beta$ high probability bound of the estimation error. (Minsker 2015) shows that the median-of-means method achieves an error bound of $O(\sqrt{d\ln(1/\beta)/n})$ with probability $1-\beta$. (Lugosi and Mendelson 2019b) improves the bound to $O(\sqrt{(d+\ln(1/\beta))/n})$ for the first time, but the method in (Lugosi and Mendelson 2019b) is computationally expensive. After that, improved algorithms with the same high probability bounds and faster computation are proposed in (Cherapanamjeri, Flammarion, and Bartlett 2019; Lei et al. 2020; Depersin and Lecué 2022). Note that this rate is minimax optimal. (Catoni 2012) shows that the lower bound of estimation error with $1-\beta$ probability is $\Omega(\sqrt{(d+\ln(1/\beta))/n})$ for n samples.

Concurrent work. After the initial submission of this paper, we notice an independent work (Asi, Liu, and Tian 2024), which proposes a reduction-based approach. Moreover, (Wang and Xu 2025) extends the analysis to non-smooth loss functions.

Compared with non-private mean estimation, we need to randomize samples carefully to achieve a tradeoff between accuracy and privacy. This involves a refined analysis of tail behaviors, as well as privacy amplification by shuffling. As a result, we finally achieve optimal rates of DP optimization with heavy-tailed gradients.

Preliminaries

Denote $\mathcal{Z}$ as the space of samples, and $\mathcal{Y}$ as the output space. We state the standard definition of DP first.

Definition 1. *(Differential Privacy (DP) (Dwork et al. 2006)) A randomized algorithm $\mathcal{A}: \mathcal{Z}^n \to \mathcal{Y}$ satisfies (ϵ, δ)-DP if for any $S \subseteq \mathcal{Y}$ and any pairs of datasets $D, D' \in \mathcal{Z}^n$ such that D and D' differ in one element,*

$$P(\mathcal{A}(D) \in S) \le e^{\epsilon} P(\mathcal{A}(D') \in S) + \delta. \tag{1}$$

Moreover, $\mathcal{A}$ is ϵ-DP if (1) *holds with $\delta = 0$.*

[1] The analysis in (Wang et al. 2020) underestimates the dependence on d. We refer to (Kamath, Liu, and Zhang 2022) for a detailed discussion. The bounds listed in Table 1 are the corrected results from (Kamath, Liu, and Zhang 2022).

[2] (Kamath, Liu, and Zhang 2022)proposed two methods, whose risk bounds are shown in the second and third rows in Table 1, respectively.

For the convenience of analysis, we also introduce another definition of DP, called concentrated differential privacy, which was first proposed in (Dwork and Rothblum 2016). (Bun and Steinke 2016) gives a refinement called zero-concentrated differential privacy. Throughout this paper, we use the definition in (Bun and Steinke 2016).

Definition 2. *(Concentrated differential privacy (CDP) (Bun and Steinke 2016)) A randomized algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{Y}$ satisfies ρ-CDP if for any pairs of datasets $D, D' \in \mathcal{Z}^n$ such that D and D' differ in one element, any $S \subseteq \mathcal{Y}$, and any $\alpha \in (1, \infty)$, $D_\alpha(\mathcal{A}(D)||\mathcal{A}(D')) \leq \rho\alpha$, in which D_α is the α-Rényi divergence between two random variables.*[3]

Our analysis in this paper will use some basic rules about the composition of DP and CDP, as well as the conversion between them. These rules are summarized in Lemma 1.

Lemma 1. *There are several facts about DP and CDP:*

(1) (Advanced composition, (Dwork, Rothblum, and Vadhan 2010; Dwork, Roth et al. 2014)) If $\mathcal{A}_1, \ldots, \mathcal{A}_k$ are (ϵ, δ)-DP, then the composition $(\mathcal{A}_1, \ldots, \mathcal{A}_k)$ is $(\sqrt{2k\ln(1/\delta')}\epsilon + k\epsilon(e^\epsilon - 1), k\delta + \delta')$-DP for any $\delta' \in (0, 1)$;

(2) (Composition of CDP, (Bun and Steinke 2016)) If $\mathcal{A}_1, \ldots, \mathcal{A}_k$ are ρ-CDP, then the composition $(\mathcal{A}_1, \ldots, \mathcal{A}_k)$ is $k\rho$-CDP;

(3) (From DP to CDP, (Bun and Steinke 2016)) If a randomized algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{Y}$ is ϵ-DP, then $\mathcal{A}$ is $(\epsilon^2/2)$-CDP;

(4) (From CDP to DP, (Bun and Steinke 2016)) If $\mathcal{A}$ is $(\epsilon^2/2)$-CDP, then $\mathcal{A}$ is $(\epsilon^2/2 + \epsilon\sqrt{2\ln(1/\delta)}, \delta)$-DP.

Moreover, we need the following lemma on the noise mechanism.

Lemma 2. *(Additive noise mechanism, (Bun and Steinke 2016)) Let $\mathcal{A}_0$ be a non-private algorithm. Define*

$$\Delta_2(\mathcal{A}_0) = \max_{d_H(D,D')=1} \|\mathcal{A}_0(D) - \mathcal{A}_0(D')\|_2$$

as the ℓ_2 sensitivity of $\mathcal{A}_0$, in which d_H denotes the Hamming distance. Then $\mathcal{A}(D) = \mathcal{A}_0(D) + \mathbf{W}$ with $\mathbf{W} \sim \mathcal{N}(0, (\Delta_2^2(\mathcal{A}_0)/2\rho)\mathbf{I})$ satisfies ρ-CDP.

We then state the problem of stochastic optimization. Suppose there are n i.i.d samples $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ following a common distribution. Given a convex constraint $\mathcal{W} \subseteq \mathbb{R}^d$ and loss function $l : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ which is convex in $\mathcal{W}$, the goal is to find an estimated minimizer $\hat{\mathbf{w}}$ of the population risk $F(\mathbf{w}) := \mathbb{E}[l(\mathbf{w}, \mathbf{Z})]$. Denote $\mathbf{w}^* = \arg\min_{\mathbf{w}} F(\mathbf{w})$ as the minimizer of the population risk. The performance of a learning algorithm is evaluated by the expected excess risk $\mathbb{E}[F(\hat{\mathbf{w}})] - F(\mathbf{w}^*)$. Our analysis is based on the following assumptions, which are similar to (Kamath, Liu, and Zhang 2022), with simplified statements.

Assumption 1. *There exists constants L, λ, M such that*

(a) The diameter of parameter space $\mathcal{W}$ is bounded by L;

(b) F is λ-smooth, i.e. for any $\mathbf{w}, \mathbf{w}'$,

$$F(\mathbf{w}') \leq F(\mathbf{w}) + \langle \nabla F(\mathbf{w}), \mathbf{w}' - \mathbf{w} \rangle + \frac{\lambda}{2}\|\mathbf{w} - \mathbf{w}'\|^2; \quad (2)$$

(c) The gradients of loss function has p-th order bounded moment for some $p \geq 2$. To be more precise, for any $\mathbf{w} \in \mathcal{W}$ and any vector $\mathbf{u}$ with $\|\mathbf{u}\| = 1$,

$$\mathbb{E}\left[|\langle \mathbf{u}, \nabla l(\mathbf{w}, \mathbf{Z})\rangle|^p\right] \leq M^p. \quad (3)$$

In (b) and (c), $\|\cdot\|$ denotes ℓ_2 norm.

In Assumption 1, (a) and (b) are common in literatures about convex optimization. (c) controls the tail behavior of gradient vectors. Lower p indicates a heavier tail, and vice versa. The case with the Lipschitz loss function (i.e. bounded gradients) corresponds to the limit of $p \to \infty$. Our assumption (3) is slightly different from the assumptions in (Kamath, Liu, and Zhang 2022). We refer to the full paper (Zhao et al. 2024c) for detailed discussion.

Under Assumption 1, the minimax lower bound of optimization under DP has been established in (Kamath, Liu, and Zhang 2022). For consistency of notations, we restate it in the following theorem.

Theorem 1. *(Rephrased from (Kamath, Liu, and Zhang 2022), Theorem 6.4) Let $\mathcal{F}$ be the set of all λ-smooth functions on $\mathcal{W}$. Let $\hat{\mathbf{w}} = \mathcal{A}(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$, in which $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$ is an arbitrary learning algorithm satisfying (ϵ, δ)-DP. Then*

$$\inf_{\mathcal{A}} \sup_{F \in \mathcal{F}} \mathcal{R}(\hat{\mathbf{w}}) \gtrsim \sqrt{\frac{d}{n}} + \sqrt{d}\left(\frac{\sqrt{d}}{n\epsilon}\right)^{\frac{p-1}{p}} \ln\frac{1}{\delta}. \quad (4)$$

Theorem 1 describes the theoretical limit of optimization risk under the differential privacy requirements. Under the light tail limit, i.e. $p \to \infty$, the right hand side of (4) becomes $\Omega(\sqrt{d/n} + d/(n\epsilon))$. Recall that for bounded gradients, the bound of excess risk is $O(1/\sqrt{n}) + \sqrt{d}/(n\epsilon))$ (Bassily et al. 2019). At first glance, it appears that the result in Theorem 1 is larger by a factor of $\sqrt{d}$. However, this discrepancy comes from the difference of assumptions. Under our tail assumption (3), the expectation of the ℓ_2 norm of the gradient vector is only bounded by $O(\sqrt{d})$, while (Bassily et al. 2019) requires the gradients to be bounded by $O(1)$. After adjustments of assumptions, Theorem 1 matches (Bassily et al. 2019) under the limit $p \to \infty$. Similar discussions can also be found in (Kamath, Liu, and Zhang 2022).

Finally, we discuss the convergence property of stochastic optimization. The framework is shown as follows. At each step t, let $g(\mathbf{w}_t)$ be the gradient estimate by $\nabla l(\mathbf{w}_t, \mathbf{Z}_1), \ldots, \nabla l(\mathbf{w}_t, \mathbf{Z}_n)$ using either Algorithm 2 or 3 with some appropriate privacy constraints. The model weights are then updated with

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}(\mathbf{w}_t - \eta g(\mathbf{w}_t)), \quad (5)$$

in which $\Pi_{\mathcal{W}}$ is the projection operator on $\mathcal{W}$. Finally, the algorithm returns

$$\hat{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t. \quad (6)$$

The whole procedures are shown in Algorithm 1. The risk can be bounded using the bias and variance of gradient estimates.

[3] The α-Rényi divergence between two distributions P and Q is defined as $D_\alpha(P||Q) = \frac{1}{\alpha - 1}\ln \mathbb{E}_{X \sim Q}\left[\left(\frac{P(X)}{Q(X)}\right)^\alpha\right]$.

Algorithm 1: Stochastic optimization

Input: dataset $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$, privacy requirement (ϵ, δ)
Output: Final iterate $\hat{\mathbf{w}}$
Parameter: Initial point $\mathbf{w}_0$, learning rate η, number of steps T
for $t = 1, \ldots, T$ **do**
  Calculate $g(\mathbf{w}_t)$, which estimates $\nabla F(\mathbf{w}_t)$ using $\nabla l(\mathbf{w}_t, \mathbf{Z}_1), \ldots, \nabla l(\mathbf{w}_t, \mathbf{Z}_n)$;
  $\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}(\mathbf{w}_t - \eta g(\mathbf{w}_t))$;
end for
$\hat{\mathbf{w}} = (1/T)\sum_{t=1}^{T} \mathbf{w}_t$;
return $\hat{\mathbf{w}}$

Lemma 3. *((Kamath, Liu, and Zhang 2022), Theorem 3.1) Define*

$$B := \max_t \|\mathbb{E}[g(\mathbf{w}_t)] - \nabla F(\mathbf{w}_t)\|, \tag{7}$$

$$G^2 := \max_t \mathbb{E}[\|g(\mathbf{w}_t) - \nabla F(\mathbf{w}_t)\|^2]. \tag{8}$$

Then the risk of optimization with updating rule (5) *and output* (6) *is bounded by*

$$\mathbb{E}[F(\hat{\mathbf{w}})] - F(\mathbf{w}^*) \le \frac{L^2}{2\eta T} + LB + \eta(\lambda^2 L^2 + G^2). \tag{9}$$

For completeness, we show the proof of Lemma 3 in the full paper (Zhao et al. 2024c). Based on Lemma 3, to bound the excess risk, we need to give bounds of B and G^2. It is relatively simple to bound $\|\mathbb{E}[g(\mathbf{w})] - \nabla F(\mathbf{w})\|$ and $\mathbb{E}[\|g(\mathbf{w}) - \nabla F(\mathbf{w})\|^2]$ for any fixed $\mathbf{w}$. The challenging part is that $\mathbf{w}_t$ depends on the data, therefore the bounds with respect to fixed $\mathbf{w}$ do not imply the bounds of B and G^2. In the following two sections, we propose two methods and provide bounds of B and G^2 for each method.

Simple Clipping Method

The simple clipping method is stated as follows. In each round, the algorithm just clips the gradient to some radius R and then adds noise to protect the privacy. Such a simple clipping method is convenient to implement and is close to the popular DP-SGD algorithm in (Abadi et al. 2016). Therefore, an in-depth analysis of this method will be helpful to bridge the gap between theory and practice in deep learning with DP.

Mean Estimation

Suppose there are n i.i.d samples $\mathbf{X}_1, \ldots, \mathbf{X}_n$ following a common distribution with mean μ. Here we assume that for any unit vector $\mathbf{u}$ with $\|\mathbf{u}\| = 1$, $\mathbb{E}[|\langle \mathbf{u}, \mathbf{X}\rangle|^p] \le M^p$, which matches Assumption 1(c).

Since samples follow a heavy-tailed distribution, some of them may be far away from μ. A simple averaging of these samples has infinite sensitivity. To ensure that the overall sensitivity is bounded, we clip them with a radius R. To be more precise, for each $i = 1, \ldots, n$, let $\mathbf{Y}_i = \mathrm{Clip}(\mathbf{X}_i, R)$,

Algorithm 2: Simple clipping method for mean estimation

Input: dataset $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and privacy requirement ρ (under CDP)
Output: Estimate $\hat{\mu}$
Parameter: R
1: **for** $i = 1, \ldots, n$ **do**
2: $\quad \mathbf{Y}_i = \mathrm{Clip}(\mathbf{X}_i, R)$;
3: **end for**
4: $\hat{\mu} = (1/n)\sum_{i=1}^{n} \mathbf{Y}_i + \mathbf{W}$, in which $\mathbf{W} \sim \mathcal{N}(0, 2R^2/(\rho n^2))$;
5: **return** $\hat{\mu}$

in which $\mathrm{Clip}(\mathbf{x}, R) = \min\{1, R/\|\mathbf{x}\|\}\mathbf{x}$. Then the final estimate is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Y}_i + \mathbf{W}, \tag{10}$$

in which $\mathbf{W} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is the noise added to meet the privacy requirement. We then show the following theorem, which determines the strength of $\mathbf{W}$,

Lemma 4. *Let $\sigma^2 = 2R^2/(\rho n^2)$, then $\hat{\mu}$ is ρ-CDP.*

Proof. Since $\|\mathbf{Y}_i\| \le R$, the sensitivity of $(1/n)\sum_{i=1}^{n} \mathbf{Y}_i$ is $\Delta_2(\hat{\mu}) = 2R/n$. According to Lemma 2, the estimator (10) is ρ-CDP. □

The procedure is summarized in Algorithm 2.

The following theorem provides a high probability bound of the estimation error.

Theorem 2. *Under the condition $\mathbb{E}[|\langle \mathbf{u}, \mathbf{X}\rangle|^p] \le M^p$ for some $p \ge 2$, under ρ-CDP, with probability $1-\beta$, the simple clipping method achieves*

$$\|\hat{\mu} - \mu\| \le \max\left\{\sqrt{\frac{12M^2}{n}\ln\frac{2\times 6^d}{\beta}}, \frac{8R}{n}\ln\frac{2\times 6^d}{\beta}\right\} + \frac{d^{\frac{p}{2}}M^p}{p-1}R^{1-p} + \frac{4R}{n\sqrt{\rho}}\sqrt{\ln\frac{2\times 6^d}{\beta}}. \tag{11}$$

Now we provide an intuitive interpretation of the result. The second term $d^{p/2}M^pR^{1-p}/(p-1)$ in (11) is the clipping bias $\|\mu_Y - \mu\|$, in which $\mu_Y = \mathbb{E}[\mathbf{Y}_i]$ is the expectation after clipping. The third term in (11) is caused by the noise $\mathbf{W}$. The first term is the bound of $\left\|\bar{\mathbf{Y}} - \mu_Y\right\|$, which is caused by the randomness of samples. Here $\bar{\mathbf{Y}} = (1/n)\sum_{i=1}^{n} \mathbf{Y}_i$ is the sample average of $\mathbf{Y}_i$. It can be written as $O\left(\sqrt{\frac{d+\ln(1/\beta)}{n}} + \frac{R}{n}(d + \ln\frac{1}{\beta})\right)$, indicating that $\bar{\mathbf{Y}}$ is subgaussian around its mean μ_Y, followed by a subexponential tail.

In (11), the factor $\ln(2 \times 6^d/\beta)$ is an important improvement over (Kamath, Liu, and Zhang 2022). The corresponding factor in (Kamath, Liu, and Zhang 2022) is $O(d\ln(1/\beta))$, while we achieve $O(d + \ln(1/\beta))$ here. Such difference does not lead to improvement in the bias and

variance of mean estimation. However, the high probability bound is improved significantly. In optimization problems, we need to take union bounds over all possible model weights $\mathbf{w}$, which requires β to be very small. In this case, $d+\ln(1/\beta) \ll d\ln(1/\beta)$. As a result, our method improves over (Kamath, Liu, and Zhang 2022) in the dependence of d. Despite such improvement, (11) has a drawback of exponential tail. As will be shown later, due to the subexponential tail $\frac{8R}{n}\ln\frac{2\times 6^d}{\beta}$, the optimization risk is not completely optimal.

Optimization

Based on the simple clipping approach, we then analyze the performance of stochastic optimization. We first discuss the DP property of the optimization problem.

Theorem 3. *If $\epsilon \le 1$, let the gradient estimator be the simple clipping method under ρ/T-CDP, in which*

$$\rho = \frac{\epsilon^2}{\left(1+2\sqrt{\ln\frac{1}{\delta}}\right)^2}, \tag{12}$$

then the whole optimization process is (ϵ,δ)-DP.

Proof. By Lemma 1(2), since each step is ρ/T-CDP, the whole process is ρ-CDP. By Lemma 1(4), ρ-CDP implies $(\rho + 2\sqrt{\ln(1/\delta)},\delta)$-DP. Since $\epsilon \le 1$, from (12), $\rho \le 1$. Therefore

$$\rho + 2\sqrt{\rho\ln\frac{1}{\delta}} \le \sqrt{\rho}\left(1+2\sqrt{\ln\frac{1}{\delta}}\right) \le \epsilon. \tag{13}$$

Therefore the optimization process is (ϵ,δ)-DP. □

From Theorem 3, we let each step satisfy ρ/T-CDP, in which ρ takes value according to (12). By Lemma 4, this requires the noise variance be $\sigma^2 = 2R^2T/(\rho n^2)$. As discussed earlier, $\mathbf{w}_t$ depends on previous steps, which depend on the data. Therefore, we need to get union bounds of estimation error to calculate B and G defined in (7) and (8). The results are shown in the following lemma.

Lemma 5. *B and G^2 defined in* (7) *and* (8) *are bounded by*

$$B \lesssim \sqrt{\frac{d\ln n}{n}} + \frac{Rd\ln n}{n}\ln n + d^{\frac{p}{2}}R^{1-p}, \tag{14}$$

$$G^2 \lesssim \frac{d\ln n}{n} + \frac{R^2d^2}{n^2}\ln^2 n + d^pR^{2(1-p)} + \frac{R^2Td}{\rho n^2}. \tag{15}$$

With Lemma 3 and 5, we then show the following theorem, which bounds the overall excess risk.

Theorem 4. *Let $T = \rho n^2/(dR^2)$, $\eta = 1/\sqrt{2T\lambda^2}$, and*

$$R = \sqrt{d}\left(\frac{n\sqrt{\rho}}{\sqrt{d}}\right)^{\frac{1}{p}} \wedge \sqrt{d}\left(\frac{n}{d}\right)^{\frac{1}{p}}, \tag{16}$$

in which ρ is determined with (12)*. Then under Assumption 1, the excess risk of Algorithm 1 is bounded by*

$$\mathbb{E}[F(\hat{\mathbf{w}})] - F(\mathbf{w}^*) \lesssim \sqrt{\frac{d\ln n}{n}}$$

$$+\sqrt{d}\left(\frac{\sqrt{d}}{n\epsilon}\sqrt{\ln\frac{1}{\delta}}\right)^{1-\frac{1}{p}} + \frac{d^{\frac{3}{2}-\frac{1}{p}}}{n^{1-\frac{1}{p}}}\ln n. \tag{17}$$

Compared with the lower bound in Theorem 1, the first two terms in (17) match (4) up to logarithmic factors. However, there is an additional term $d^{3/2-1/p}\ln n/n^{1-1/p}$. If $\epsilon \le 1/\sqrt{d}$, then this term does not dominate. Therefore, the simple clipping method is minimax optimal (up to logarithmic factors) for $\epsilon \le 1/\sqrt{d}$.

Iterative Updating Method

The previous section shows that the simple clipping method is not always optimal due to an additional term $d^{3/2-1/p}\ln n/n^{1-1/p}$. In this section, we show an improved method to avoid this term, which is inspired by some existing methods for non-private mean estimation (Cherapanamjeri, Flammarion, and Bartlett 2019; Lugosi and Mendelson 2019b; Lei et al. 2020). To begin with, to illustrate the idea of design, we provide some basic intuition. The mean estimation algorithm is then described in detail. Finally, we analyze the risk of optimization with the new mean estimator.

Intuition

The suboptimality of the simple clipping approach comes from the subexponential tails. Ideally, under ρ-CDP, we would like a $\|\hat{\mu}-\mu\| \lesssim \sqrt{(d+\ln(1/\beta))/n} + R\sqrt{d+\ln(1/\beta)}/(n\sqrt{\rho}) + d^{p/2}R^{1-p}$ error bound that holds with probability $1-\beta$. However, from (11), the simple clipping method has an additional term $O(R(d+\ln(1/\beta))/n)$, which indicates a subexponential tail behavior. To remove the subexponential tail, a classical approach is median-of-means, which divides data into multiple groups, calculates the mean of each group, and then finds the median of all group-wise means. However, (Minsker 2015) shows that even for non-private estimation, the geometric median-of-mean method achieves a suboptimal bound of $O(\sqrt{d\ln(1/\beta)/n})$ with probability $1-\beta$. While this bound has optimal dependence on d and n, the dependence on β is not optimal. The calculation of the union bound of estimation error usually encounters very small β. As a result, the suboptimal dependence of the error bound on β leads to a larger union bound.

To handle this issue, we design a new estimator, which is inspired by several later works (Lugosi and Mendelson 2019b; Cherapanamjeri, Flammarion, and Bartlett 2019; Lei et al. 2020) that improve the non-private bound to $O(\sqrt{(d+\ln(1/\beta))/n})$. The basic idea of the mean estimator is to iteratively update the current estimate $\mathbf{c}_t$ based on the estimation of distance and direction to the truth. To make the estimator satisfy the DP requirement, we add appropriate noise. The estimator is permutation invariant with respect to the group-wise means, thus equivalently, we can view these group-wise means as being shuffled. The shuffling operation makes an amplification to DP (Erlingsson et al. 2019; Feldman, McMillan, and Talwar 2022). Therefore, each group only needs to satisfy a weaker privacy requirement than (ϵ,δ)-DP.

Algorithm 3: Iterative updating method for mean estimation

Input: dataset $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and privacy requirement ϵ, δ
Output: Estimate $\hat{\mu}$
Parameter: R, k, initial point $\mathbf{c}_1$

Divide samples into k groups randomly;
for $j = 1, \ldots, k$ **do**
  $\mathbf{Q}_j = (1/m)\sum_{i \in B_j} \mathbf{Y}_i + \mathbf{W}_j$, in which $\mathbf{W}_j \sim \mathcal{N}(0, \sigma^2)$, with $\sigma^2 = 2R^2/(\rho m^2)$;
end for
for $l = 1, \ldots, t_c$ **do**
  $d_l, \mathbf{g}_l = \text{Estimate}(\mathbf{Q}_1, \ldots, \mathbf{Q}_k, \mathbf{c}_l)$;
  $\mathbf{c}_{l+1} = \mathbf{c}_l + \eta d_l \mathbf{g}_l$;
end for
$l^* = \arg\min_l d_l$;
$\hat{\mu} = \mathbf{c}_{l^*}$;
return $\hat{\mu}$

The Mean Estimation Algorithm

Here we state the result first and then show the construction of the mean estimator.

Theorem 5. *There exists a constant c, if $\epsilon \le c\sqrt{(1/k)\ln(1/\delta)}$ and $\delta \in (0,1)$, there exists an estimator satisfying (ϵ, δ)-DP, such that with probability $1-\beta$,*

$$\|\hat{\mu} - \mu\| \lesssim \sqrt{\frac{d + \ln\frac{1}{\beta}}{n}} + d^{\frac{p}{2}} R^{1-p} + \frac{R}{n\epsilon}\sqrt{d + \ln\frac{1}{\beta}}\left(\ln\frac{1}{\delta} + \sqrt{\ln\frac{1}{\delta}\ln\frac{1}{\beta}}\right). \quad (18)$$

With $\epsilon \to \infty$ or $\delta \to 1$, one can just let R to be sufficiently large, then $\|\hat{\mu} - \mu\| \lesssim \sqrt{(d + \ln(1/\beta))/n}$, which matches existing results on non-private mean estimation (Lugosi and Mendelson 2019b; Cherapanamjeri, Flammarion, and Bartlett 2019; Lei et al. 2020). Note that the factor $d + \ln(1/\beta)$ is important. If we use the median-of-means method instead, then this factor will become $d\ln(1/\beta)$, which will yield a suboptimal union bound.

The remainder of this section explains the algorithm and proves Theorem 5. The whole process of mean estimation is shown in Algorithm 3. The idea uses (Cherapanamjeri, Flammarion, and Bartlett 2019). The difference from (Cherapanamjeri, Flammarion, and Bartlett 2019) is that we need to let the result satisfy (ϵ, δ)-DP, thus the truncation radius needs to be carefully tuned. Moreover, the distance estimation is with respect to the truncated mean $\mu_Y = \mathbb{E}[\mathbf{Y}_i]$ instead of ground truth μ. Compared with (Cherapanamjeri, Flammarion, and Bartlett 2019), we make a simplified algorithm statement here.

Now we explain key steps in Algorithm 3.

1) Group-wise averages (step 3). The algorithm begins with dividing samples into into k bins $B_1, \ldots, B_k$. The size of each bin is m, then $n = mk$. Let

$$\mathbf{Q}_j = \frac{1}{m}\sum_{i \in B_j} \mathbf{Y}_i + \mathbf{W}_j, \quad (19)$$

with $\mathbf{W}_j \sim \mathcal{N}(0, \sigma^2)$. Here we let $\sigma^2 = 2R^2/(\rho m^2)$.

The final estimate is based on $\mathbf{Q}_1, \ldots, \mathbf{Q}_k$. Note that the sensitivities of $\mathbf{Q}_j$, $j = 1, \ldots, k$ are all $2R/m$ over $\mathbf{X}_i$, $i \in B_j$. By Lemma 2, $\mathbf{Q}_1, \ldots, \mathbf{Q}_k$ are all ρ-CDP. Before constructing the final estimator $\hat{\mu}$, we first show that $\hat{\mu}$ is (ϵ, δ)-DP under some necessary conditions.

Theorem 6. *Suppose $\epsilon \le 8e^2\sqrt{(1/k)\ln(4/\delta)}$. Let the noise variance be $\sigma^2 = 2R^2/(\rho m^2)$. If an estimator $\hat{\mu}$ only depends on $\mathbf{Q}_1, \ldots, \mathbf{Q}_k$, and is permutation invariant with respect to $\mathbf{Q}_1, \ldots, \mathbf{Q}_k$. If $\mathbf{Q}_1, \ldots, \mathbf{Q}_k$ are all ρ-CDP with*

$$\rho = \frac{1}{64e^4}\frac{\epsilon^2 k}{\ln\frac{8}{\delta}\left(1 + 2\sqrt{\ln\frac{12k}{\delta}}\right)^2}, \quad (20)$$

then $\hat{\mu}$ is (ϵ, δ)-DP.

Since $\mathbf{Q}_j$ is ρ-CDP with respect to $\{\mathbf{X}_i | i \in B_j\}$, it is straightforward to see that any estimator $\hat{\mu}$ that depends only on $\mathbf{Q}_1, \ldots, \mathbf{Q}_k$ is also ρ-CDP. However, if $\hat{\mu}$ is permutation invariant with respect to $\mathbf{Q}_1, \ldots, \mathbf{Q}_k$, then the privacy guarantee becomes stronger, since $\hat{\mu}$ does not change if we shuffle $\mathbf{Q}_1, \ldots, \mathbf{Q}_k$ randomly. According to (Erlingsson et al. 2019; Feldman, McMillan, and Talwar 2022), the privacy guarantee can be amplified. As a result, to ensure that the whole algorithm satisfies (ϵ, δ)-DP, the privacy requirement for each group is only ρ-CDP with (20), which is weaker than (12) for sufficiently large k.

Under such settings, we show the bound of the estimation error. Similar to existing research on non-private mean estimation, we show that most elements in $\{\mathbf{Q}_1, \ldots, \mathbf{Q}_k\}$ are not far away from the truncated mean μ_Y.

Lemma 6. *There exists a constant C_0. For any $\beta > 0$, let $k = 800\ln(1/\beta)$. Define*

$$r_0 = C_0\left(\sqrt{\frac{d + \ln(1/\beta)}{n}} + \frac{R}{n}\sqrt{\frac{k}{\rho}}\sqrt{d + \ln(1/\beta)}\right). \quad (21)$$

Then with probability at least $1-\beta$,

$$\sup_{\mathbf{u}:\|\mathbf{u}\|=1} \sum_{j=1}^{k} \mathbf{1}\left(\langle \mathbf{u}, \mathbf{Q}_j - \mu_Y\rangle > r_0\right) \le \frac{1}{10}k. \quad (22)$$

(Lugosi and Mendelson 2019b) shows that for the non-private case, for arbitrary unit vector $\mathbf{u}$, most of elements in $\{\mathbf{Q}_1, \ldots, \mathbf{Q}_k\}$ satisfy $\langle \mathbf{u}, \mathbf{Q}_j - \mu\rangle \lesssim \sqrt{(d + \ln(1/\beta))/n}$, which matches the first term in (22). Compared with (Lugosi and Mendelson 2019b), we extend the analysis to the case with the clipping operation and random noise.

2) Distance and gradient estimation (step 6). We introduce the following optimization problem:

$$\begin{aligned} \max \quad & s \\ \text{subject to} \quad & b_j\langle \mathbf{Q}_j - \mathbf{c}, \mathbf{u}\rangle \ge b_j s, j = 1, \ldots, k \\ & \sum_{j=1}^{k} b_j \ge 0.9k \\ & b_j \in \{0,1\}, j = 1, \ldots, k, \\ & \|\mathbf{u}\| = 1. \end{aligned} \quad (23)$$

Now we explain (23). This optimization problem attempts to find maximum s, such that there exists a unit vector $\mathbf{u}$, at least 90% of the projection of $\{\mathbf{Q}_1, \ldots, \mathbf{Q}_k\}$ on $\mathbf{u}$ are at least s far away from the current iterate $\mathbf{c}$. Denote the function of estimation as $d, \mathbf{g} = \mathrm{Estimate}(\mathbf{Q}_1, \ldots, \mathbf{Q}_k, \mathbf{c})$. The program Estimate solves the optimization problem (23), and returns d and $\mathbf{g}$, in which d is the optimal value of s, which estimates the distance $\|\mathbf{c} - \mu_Y\|$, and $\mathbf{g}$ is the corresponding value of $\mathbf{u}$, which estimates the direction from $\mathbf{c}$ to μ_Y, i.e. $(\mathbf{c} - \mu_Y)/\|\mathbf{c} - \mu_Y\|$. Such estimation is analyzed in the following lemma, in which we follow the analysis in (Cherapanamjeri, Flammarion, and Bartlett 2019).

Lemma 7. *Let $d, \mathbf{g} = \mathrm{Estimate}(\mathbf{Q}_1, \ldots, \mathbf{Q}_k, \mathbf{c})$, which is calculated by solving the optimization problem* (23)*. If* (22) *is satisfied, then*

$$|d - \|\mathbf{c} - \mu_Y\|\,| \le r_0. \tag{24}$$

Moreover, if $\|\mu_Y - \mathbf{c}\| \ge 4r_0$, then

$$\left\langle \mathbf{g}, \frac{\mu_Y - \mathbf{c}}{\|\mu_Y - \mathbf{c}\|} \right\rangle \ge \frac{1}{2}. \tag{25}$$

(24) shows that under (22), which holds with probability at least $1 - \beta$, the error of distance estimate is at least r_0. Moreover, (25) shows that if the current iterate $\mathbf{c}$ is sufficiently far away from the truncated mean μ_Y, then the angle between gradient estimate and the ideal update direction along $\mu_Y - \mathbf{c}$ is no more than $\pi/3$. These analysis validates that the update rule $\mathbf{c}_{l+1} = \mathbf{c}_l + \eta d_l g_l$ makes the iterate point $\mathbf{c}_l$ close to μ_Y with large l. To be more precise, we show the following lemma:

Lemma 8. *Let $\eta = 1/4$. If $\|\mathbf{c}_1 - \mu_Y\| \le 4r_0$, or $t_c \ge 2\ln\frac{\|\mathbf{c}_1 - \mu_Y\|}{4r_0} / \ln\frac{256}{233}$, then the estimate $\hat{\mu}$ with Algorithm 3 satisfies $\|\hat{\mu} - \mu_Y\| \le 6r_0$, in which $\mu_Y = \mathbb{E}[\mathbf{Y}_i]$.*

Lemma 8 bounds the estimation error with respect to the truncated mean μ_Y. Recall the definition of r_0 in (21). Considering the clipping bias, we have

$$\|\hat{\mu} - \mu\| \lesssim \sqrt{\frac{d + \ln(1/\beta)}{n}} + \frac{R}{n}\sqrt{\frac{k}{\rho}}\sqrt{d + \ln(1/\beta)} + d^{\frac{p}{2}} R^{1-p}. \tag{26}$$

(18) can then be obtained using (26) and (20). The construction of the mean estimator and the proof of Theorem 5 is complete.

Application in DP Optimization

Now we have constructed a mean estimator under (ϵ, δ)-DP, whose estimation error is bounded with Theorem 5. For the stochastic optimization problem, we need to estimate the gradient for T steps, and the (ϵ, δ)-DP requirement is imposed on the whole process. Therefore, each step needs to satisfy stronger privacy requirements. According to advanced composition theorem (Lemma 1(1)), here we ensure that each step satisfies (ϵ_0, δ_0)-DP, with

$$\epsilon_0 = \frac{\epsilon}{2\sqrt{2T\ln\frac{2}{\delta}}}, \delta_0 = \frac{\delta}{2T}. \tag{27}$$

Then the optimization process with T steps is (ϵ, δ)-DP. For any fixed $\mathbf{w}$, let $g(\mathbf{w})$ be the mean estimate using $\nabla l(\mathbf{w}, \mathbf{Z}_1), \ldots, \nabla l(\mathbf{w}, \mathbf{Z}_n)$ under (ϵ_0, δ_0)-DP. According to Theorem 5, for any fixed $\mathbf{w}$, the gradient estimate at each step satisfies

$$\|g(\mathbf{w}) - \nabla F(\mathbf{w})\| \lesssim \sqrt{\frac{d + \ln\frac{1}{\beta}}{n}} + d^{\frac{p}{2}} R^{1-p} + \frac{R}{n\epsilon_0}\sqrt{d + \ln\frac{1}{\beta}}\left(\ln\frac{1}{\delta_0} + \sqrt{\ln\frac{1}{\delta_0}\ln\ln\frac{1}{\beta}}\right). \tag{28}$$

As discussed earlier, since $\mathbf{w}_t$ depends on the data, the bias and variance of gradient estimation at time t, i.e. $B = \max_t \|\mathbb{E}[g(\mathbf{w}_t)] - \nabla F(\mathbf{w}_t)\|$ and $G^2 = \max_t \mathbb{E}[\|g(\mathbf{w}_t) - \nabla F(\mathbf{w}_t)\|^2]$ can not be bounded simply using the bias and variance with fixed $\mathbf{w}$. Instead, similar to Lemma 5, we need to derive a union bound of all $\mathbf{w} \in \mathcal{W}$. The results are shown in Lemma 9.

Lemma 9. *For the mean estimation algorithm, B and G^2 are bounded by $B \lesssim \sqrt{\frac{d}{n}} + d^{\frac{p}{2}} R^{1-p}$ and*

$$G^2 \lesssim \frac{d}{n} + \frac{R^2 d}{n^2\epsilon_0^2}\left(\ln\frac{1}{\delta_0} + \sqrt{\ln\frac{1}{\delta_0}\ln d}\right) + d^p R^{2(1-p)}. \tag{29}$$

Based on Lemma 3 and 9, we can then derive the following bound on the excess risk.

Theorem 7. *Let $T = n^2\epsilon^2/(dR^2)$, $\eta = 1/\sqrt{2T\lambda^2}$, and $R = \sqrt{d}\left(n\epsilon/\sqrt{d}\right)^{\frac{1}{p}}$. If $\epsilon \le 1$, then*

$$\mathbb{E}[F(\hat{\mathbf{w}})] - F(\mathbf{w}^*) \lesssim \sqrt{\frac{d}{n}} + \sqrt{d}\left(\frac{\sqrt{d}}{n\epsilon}\right)^{1-\frac{1}{p}} \ln\frac{1}{\delta}\ln(nd).$$

The proof of Theorem 7 is shown in Appendix H in the full paper (Zhao et al. 2024c). The bound shown in Theorem 7 matches the minimax lower bound in Theorem 1, indicating that the new method is minimax rate optimal.

Conclusion

In this paper, we have improved the convergence of population risk of stochastic optimization under DP. We have proposed two methods. The simple clipping method is relatively convenient to implement. It achieves $\tilde{O}\left(\sqrt{\frac{d}{n}} + \sqrt{d}\left(\frac{\sqrt{d}}{\epsilon n}\right)^{1-\frac{1}{p}} + \frac{d^{\frac{3}{2}-\frac{1}{p}}}{n^{1-\frac{1}{p}}}\right)$ risk bound. The iterative updating method further improves the risk bound to $\tilde{O}\left(\sqrt{\frac{d}{n}} + \sqrt{d}\left(\frac{\sqrt{d}}{\epsilon n}\right)^{1-\frac{1}{p}}\right)$, which matches the minimax lower bound, indicating that this method is optimal.

Acknowledgements

The work of Z.Liu is supported by the National Natural Science Foundation of China (No.62132008 and U22B2030), Natural Science Foundation of Jiangsu Province(BK20220075). The work of R. Fan is supported by National Natural Science Foundation of China under

Grant No. 62171034. The work of C. Wang is supported by Ningbo Municipal Natural Science Foundation of China (No. 2022J114), Ningbo S&T Project (No.2024Z004) and Ningbo Major Research and Development Plan Project (No.2023Z225). The work of Q. Li is supported by Priority-Funded Postdoctoral Research Project, Zhejiang Province (No.ZJ2024001).

We thank Prof.Andrew Lowy for his fruitful discussion.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 308–318.

Armacki, A.; Sharma, P.; Joshi, G.; Bajovic, D.; Jakovetic, D.; and Kar, S. 2023. High-probability Convergence Bounds for Nonlinear Stochastic Gradient Descent Under Heavy-tailed Noise. *arXiv preprint arXiv:2310.18784*.

Asi, H.; Feldman, V.; Koren, T.; and Talwar, K. 2021. Private stochastic convex optimization: Optimal rates in l1 geometry. In *ICML*, 393–403.

Asi, H.; Liu, D.; and Tian, K. 2024. Private Stochastic Convex Optimization with Heavy Tails: Near-Optimality from Simple Reductions. *arXiv preprint arXiv:2406.02789*.

Bassily, R.; Feldman, V.; Talwar, K.; and Guha Thakurta, A. 2019. Private stochastic convex optimization with optimal rates. *NeurIPS*, 32.

Bassily, R.; Guzmán, C.; and Nandi, A. 2021. Non-euclidean differentially private stochastic convex optimization. In *Conference on Learning Theory*, 474–499.

Bassily, R.; Smith, A.; and Thakurta, A. 2014. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, 464–473. IEEE.

Bassily, R.; and Sun, Z. 2023. User-level private stochastic convex optimization with optimal rates. In *ICML*, 1838–1851.

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

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

Chaudhuri, K.; and Monteleoni, C. 2008. Privacy-preserving logistic regression. *NeurIPS*, 21.

Cherapanamjeri, Y.; Flammarion, N.; and Bartlett, P. L. 2019. Fast mean estimation with sub-gaussian rates. In *Conference on Learning Theory*, 786–806.

Das, R.; Kale, S.; Xu, Z.; Zhang, T.; and Sanghavi, S. 2023. Beyond uniform lipschitz condition in differentially private optimization. In *ICML*, 7066–7101.

De, S.; Berrada, L.; Hayes, J.; Smith, S. L.; and Balle, B. 2022. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*.

Depersin, J.; and Lecué, G. 2022. Robust sub-Gaussian estimation of a mean vector in nearly linear time. *The Annals of Statistics*, 50(1): 511–536.

Dong, Y.; Yao, J.; Wang, J.; Liang, Y.; Liao, S.; and Xiao, M. 2024. Dynamic fraud detection: Integrating reinforcement learning into graph neural networks. In *2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS)*, 818–823. IEEE.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, 265–284. Springer.

Dwork, C.; Roth, A.; et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4): 211–407.

Dwork, C.; and Rothblum, G. N. 2016. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*.

Dwork, C.; Rothblum, G. N.; and Vadhan, S. 2010. Boosting and differential privacy. In *2010 IEEE 51st annual symposium on foundations of computer science*, 51–60. IEEE.

Eldowa, K.; and Paudice, A. 2024. General Tail Bounds for Non-Smooth Stochastic Mirror Descent. In *International Conference on Artificial Intelligence and Statistics*, 3205–3213.

Erlingsson, Ú.; Feldman, V.; Mironov, I.; Raghunathan, A.; Talwar, K.; and Thakurta, A. 2019. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2468–2479. SIAM.

Feldman, V.; Koren, T.; and Talwar, K. 2020. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, 439–449.

Feldman, V.; McMillan, A.; and Talwar, K. 2022. 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)*, 954–964. IEEE.

Feng, J.; Wu, Y.; Sun, H.; Zhang, S.; and Liu, D. 2025, DOI: 10.1109/TIFS.2025.3526063. Panther: Practical Secure Two-Party Neural Network Inference. *IEEE Transactions on Information Forensics and Security*.

Gorbunov, E.; Danilova, M.; and Gasnikov, A. 2020. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *NeurIPS*, 33: 15042–15053.

Gurbuzbalaban, M.; Simsekli, U.; and Zhu, L. 2021. The heavy-tail phenomenon in SGD. In *ICML*, 3964–3975.

Hopkins, S. B.; Kamath, G.; and Majid, M. 2022. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, 1406–1417.

Hsieh, W.; Bi, Z.; Jiang, C.; Liu, J.; Peng, B.; Zhang, S.; Pan, X.; Xu, J.; Wang, J.; Chen, K.; et al. 2024. A Comprehensive Guide to Explainable AI: From Classical Models to LLMs. *arXiv preprint arXiv:2412.00800*.

Hu, L.; Ni, S.; Xiao, H.; and Wang, D. 2022. High dimensional differentially private stochastic optimization with heavy-tailed data. In *Proceedings of the 41st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 227–236.

Huang, Z.; Liang, Y.; and Yi, K. 2021. Instance-optimal mean estimation under differential privacy. *NeurIPS*, 34: 25993–26004.

Huo, M.; Lu, K.; Li, Y.; and Zhu, Q. 2025. CT-PatchTST: Channel-Time Patch Time-Series Transformer for Long-Term Renewable Energy Forecasting. arXiv:2501.08620.

Iyengar, R.; Near, J. P.; Song, D.; Thakkar, O.; Thakurta, A.; and Wang, L. 2019. Towards practical differentially private convex optimization. In *2019 IEEE symposium on security and privacy (SP)*, 299–316. IEEE.

Kamath, G.; Liu, X.; and Zhang, H. 2022. Improved rates for differentially private stochastic convex optimization with heavy-tailed data. In *ICML*, 10633–10660.

Kamath, G.; Singhal, V.; and Ullman, J. 2020. Private mean estimation of heavy-tailed distributions. In *Conference on Learning Theory*, 2204–2235.

Ke, Z.; Xu, J.; Zhang, Z.; Cheng, Y.; and Wu, W. 2024. A Consolidated Volatility Prediction with Back Propagation Neural Network and Genetic Algorithm. *arXiv preprint arXiv:2412.07223*.

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

Koloskova, A.; Hendrikx, H.; and Stich, S. U. 2023. Revisiting Gradient Clipping: Stochastic bias and tight convergence guarantees. In *ICML*, 17343–17363.

Kulkarni, J.; Lee, Y. T.; and Liu, D. 2021. Private non-smooth erm and sco in subquadratic steps. *NeurIPS*, 34: 4053–4064.

Lei, Z.; Luh, K.; Venkat, P.; and Zhang, F. 2020. A fast spectral algorithm for mean estimation with sub-gaussian rates. In *Conference on Learning Theory*, 2598–2612.

Li, Z.; Cui, J.; Chen, H.; Lu, H.; Zhou, F.; Rocha, P. R. F.; and Yang, C. 2025. Research Progress of All-Fiber Optic Current Transformers in Novel Power Systems: A Review. *Microwave and Optical Technology Letters*, 67(1): e70061.

Liu, X.; Kong, W.; Kakade, S.; and Oh, S. 2021. Robust and differentially private mean estimation. *NeurIPS*, 34: 3887–3901.

Liu, Z.; Nguyen, T. D.; Nguyen, T. H.; Ene, A.; and Nguyen, H. 2023. High probability convergence of stochastic gradient methods. In *ICML*, 21884–21914.

Liu, Z.; and Zhou, Z. 2023. Stochastic Nonsmooth Convex Optimization with Heavy-Tailed Noises: High-Probability Bound, In-Expectation Rate and Initial Distance Adaptation. *arXiv preprint arXiv:2303.12277*.

Lowy, A.; and Razaviyayn, M. 2023. Private stochastic optimization with large worst-case lipschitz parameter: Optimal rates for (non-smooth) convex losses and extension to non-convex losses. In *International Conference on Algorithmic Learning Theory*, 986–1054. PMLR.

Lugosi, G.; and Mendelson, S. 2019a. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5): 1145–1190.

Lugosi, g.; and Mendelson, S. 2019b. Sub-gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2): 783–794.

Lyu, W.; Zheng, S.; Pang, L.; Ling, H.; and Chen, C. 2023. Attention-Enhancing Backdoor Attacks Against BERT-based Models. In *EMNLP*, 10672–10690.

Minsker, S. 2015. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 2308–2335.

Nguyen, T. D.; Nguyen, T. H.; Ene, A.; and Nguyen, H. 2023. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. *NeurIPS*, 36: 24191–24222.

Parletta, D. A.; Paudice, A.; Pontil, M.; and Salzo, S. 2022. High probability bounds for stochastic subgradient schemes with heavy tailed noise. *arXiv preprint arXiv:2208.08567*.

Pascanu, R.; Mikolov, T.; and Bengio, Y. 2012. Understanding the exploding gradient problem. *CoRR, abs/1211.5063*, 2(417): 1.

Peng, B.; Bi, Z.; Niu, Q.; Liu, M.; Feng, P.; Wang, T.; Yan, L. K.; Wen, Y.; Zhang, Y.; and Yin, C. H. 2024. Jailbreaking and mitigation of vulnerabilities in large language models. *arXiv preprint arXiv:2410.15236*.

Sha, H.; Cao, Y.; Liu, Y.; Wu, Y.; Liu, R.; and Chen, H. 2024. Clip Body and Tail Separately: High Probability Guarantees for DPSGD with Heavy Tails. *arXiv preprint arXiv:2405.17529*.

Şimşekli, U.; Gürbüzbalaban, M.; Nguyen, T. H.; Richard, G.; and Sagun, L. 2019. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv preprint arXiv:1912.00018*.

Simsekli, U.; Sagun, L.; and Gurbuzbalaban, M. 2019. A tail-index analysis of stochastic gradient noise in deep neural networks. In *ICML*, 5827–5837.

Thakurta, A. G.; and Smith, A. 2013. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory*, 819–850.

Tramer, F.; and Boneh, D. 2021. Differentially Private Learning Needs Better Features (or Much More Data). In *ICLR*.

Wang, D.; Xiao, H.; Devadas, S.; and Xu, J. 2020. On differentially private stochastic convex optimization with heavy-tailed data. In *ICML*, 10081–10091.

Wang, D.; and Xu, J. 2025. Private Least Absolute Deviations with Heavy-tailed Data. *Theoretical Computer Science*, 115071.

Wang, D.; Ye, M.; and Xu, J. 2017. Differentially private empirical risk minimization revisited: Faster and more general. *NeurIPS*, 30.

Wang, S.; Jiang, R.; Wang, Z.; and Zhou, Y. 2024. Deep learning-based anomaly detection and log analysis for computer networks. *arXiv preprint arXiv:2407.05639*.

Wei, J.; Bao, E.; Xiao, X.; and Yang, Y. 2022. Dpis: An enhanced mechanism for differentially private sgd with importance sampling. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2885–2899.

Zhang, H.; Mironov, I.; and Hejazinia, M. 2021. Wide network learning with differential privacy. *arXiv preprint arXiv:2103.01294*.

Zhang, J.; Karimireddy, S. P.; Veit, A.; Kim, S.; Reddi, S.; Kumar, S.; and Sra, S. 2020. Why are adaptive methods good for attention models? *NeurIPS*, 33: 15383–15393.

Zhao, P.; Fan, R.; Wu, H.; Li, Q.; Wu, J.; and Liu, Z. 2024a. Enhancing Learning with Label Differential Privacy by Vector Approximation. *arXiv preprint arXiv:2405.15150*.

Zhao, P.; and Lai, L. 2022. Analysis of knn density estimation. *IEEE Transactions on Information Theory*, 68(12): 7971–7995.

Zhao, P.; Lai, L.; Shen, L.; Li, Q.; Wu, J.; and Liu, Z. 2024b. A Huber Loss Minimization Approach to Mean Estimation under User-level Differential Privacy. In *NeurIPS*.

Zhao, P.; Wu, J.; Liu, Z.; Wang, C.; Fan, R.; and Li, Q. 2024c. Differential Private Stochastic Optimization with Heavy-tailed Data: Towards Optimal Rates. *arXiv preprint arXiv:2408.09891*.

Zhao, P.; Yu, F.; and Wan, Z. 2024. A huber loss minimization approach to byzantine robust federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 21806–21814.