Journal of Machine Learning Research 25 (2024) 1-41 Submitted 1/22; Revised 9/23; Published 4/24

Faster Rates of Differentially Private Stochastic Convex Optimization

Jinyan Su* JS3673@CORNELL.EDU
Cornell University

Lijie Hu LIJIE.HU@KAUST.EDU.SA
Provable Responsible AI and Data Analytics Lab
Division of CEMSE
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

Di Wang DI.WANG@KAUST.EDU.SA
Provable Responsible AI and Data Analytics Lab
Division of CEMSE
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

Editor: Moritz Hardt

Abstract

In this paper, we revisit the problem of Differentially Private Stochastic Convex Optimization (DP-SCO) and provide excess population risks for some special classes of functions that are faster than the previous results of general convex and strongly convex functions. In the first part of the paper, we study the case where the population risk function satisfies the Tysbakov Noise Condition (TNC) with some parameter $\theta > 1$. Specifically, we first show that under some mild assumptions on the loss functions, there is an algorithm whose output could achieve an upper bound of $\widetilde{O}((\frac{1}{\sqrt{n}} + \frac{d}{n\epsilon})^{\frac{\theta}{\theta-1}})$ and $\widetilde{O}((\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{\theta}{\theta-1}})$ for ϵ-DP and (ϵ, δ)-DP, respectively when $\theta \geq 2$, where n is the sample size and d is the dimension of the space. Then we address the inefficiency issue, improve the upper bounds by $\mathrm{Poly}(\log n)$ factors and extend to the case where $\theta \geq \bar{\theta} > 1$ for some known $\bar{\theta}$. Next, we show that the excess population risk of population functions satisfying TNC with parameter $\theta \geq 2$ is always lower bounded by $\Omega((\frac{d}{n\epsilon})^{\frac{\theta}{\theta-1}})$ and $\Omega((\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{\theta}{\theta-1}})$ for ϵ-DP and (ϵ, δ)-DP, respectively, which matches our upper bounds. In the second part, we focus on a special case where the population risk function is strongly convex. Unlike the previous studies, here we assume the loss function is *non-negative* and *the optimal value of population risk is sufficiently small*. With these additional assumptions, we propose a new method whose output could achieve an upper bound of $O(\frac{d\log(1/\delta)}{n^2\epsilon^2} + \frac{1}{n^\tau})$ and $O(\frac{d^2}{n^2\epsilon^2} + \frac{1}{n^\tau})$ for any $\tau > 1$ in (ϵ, δ)-DP and ϵ-DP model respectively if the sample size n is sufficiently large. These results circumvent their corresponding lower bounds in (Feldman et al., 2020) for general strongly convex functions. Finally, we conduct experiments of our new methods on real-world data. Experimental results also provide new insights into established theories.

∗. Part of the work was done when Jinyan Su was a research intern at KAUST. An abstract version of this paper will be presented at The 33rd International Conference on Algorithmic Learning Theory (ALT 2022) (Su et al., 2022).

Keywords: Differential Privacy, Stochastic Convex Optimization

1. Introduction

Preserving the privacy of training data has become an important consideration and now is a challenging task for machine learning algorithms. To address the privacy issue, Differential Privacy (DP) (Dwork et al., 2006), which roots in cryptography, is a strong mathematical scheme for privacy preservation. It allows for rich statistical and machine learning analysis and is now becoming a de facto notation for private data analysis. Methods to guarantee differential privacy have been widely studied, and recently adopted in industry (Tang et al., 2017; Ding et al., 2017).

As one of the most important problems in Machine Learning and Differential Privacy community, the Empirical Risk Minimization problem in the DP model, *i.e.,* DP-ERM, has been studied quite well in the last decade, starting from (Chaudhuri et al., 2011), such as (Bassily et al., 2014; Wang et al., 2017, 2019a; Wu et al., 2017; Kasiviswanathan and Jin, 2016; Kifer et al., 2012; Smith et al., 2017; Wang et al., 2018a, 2019b; Asi et al., 2021a). Besides DP-ERM, its population (or expected) version, namely Differentially Private Stochastic Convex Optimization (DP-SCO), has received much attention in recent years, starting from (Bassily et al., 2014). Specifically, (Bassily et al., 2019) first provides the optimal rate of DP-SCO with general convex loss functions in (ϵ, δ)-DP, which is quite different from the optimal rate in DP-ERM. Later, (Feldman et al., 2020) extends this problem to strongly convex and (or) non-smooth cases by providing a general localization technique. Moreover, their methods have linear time complexity if the loss functions are smooth. For non-smooth loss functions, (Kulkarni et al., 2021) recently proposes a new method that only needs subquadratic gradient complexity. While there are already a large number of studies on DP-SCO, the problem is still far from well understood. A key observation is that all of the previous works only focus on the case where the loss functions are either general convex or strongly convex. However, there are also many problems that are even stronger than strongly convex functions, or fall between convex and strongly convex functions. In the non-private counterpart, various studies have attempted to get faster rates by imposing additional assumptions on the loss functions. And it has been shown that it is indeed possible to achieve rates that are faster than the rates of general convex loss functions (Yang et al., 2018; Koren and Levy, 2015; van Erven et al., 2015), or it could even achieve the same rate as in the strongly convex case even if the function is not strongly convex (Karimi et al., 2016; Liu et al., 2018; Xu et al., 2017). Motivated by this, our question is,

For the problem of DP-SCO with special classes of population risk functions, is it possible to achieve faster rates of excess population risk than the optimal ones of general convex and (or) strongly convex cases?

In this paper, we provide an affirmative answer by studying some classes of population risk functions. Particularly, we will mainly focus on the case where the population risk function satisfies the Tysbakov Noise Condition (TNC) [1], which includes strongly convex functions, SVM, ℓ_1-regularized stochastic optimization and linear regression as special cases

1. In some related work, it is also called the Error Bound Condition or the Growth Condition (Liu et al., 2018; Xu et al., 2017).

Method	Assumptions	Upper Bound	Lower Bound
Algorithm 2♣	$F(\cdot)$ satisfies (θ,λ)-TNC, $\theta \geq 2$ is unknown, the loss function is convex, smooth and Lipschitz	$O\big((\frac{1}{\sqrt{n}}+\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{\theta}{\theta-1}}\cdot\text{Poly}(\log n)\big)$	[Theorem 19] $\Omega\left((\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{\theta}{\theta-1}}\right)$
Algorithm 4⋆	$F(\cdot)$ satisfies (θ,λ)-TNC, $\theta > 1$ is known, the loss function is convex, smooth and Lipschitz	$O\big((\frac{1}{\sqrt{n}}+\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{\theta}{\theta-1}}\cdot\text{Poly}(\log n)\big)$	
Algorithm 5♠	$F(\cdot)$ satisfies (θ,λ)-TNC, θ unknown but is lower bounded by some known $\bar{\theta} > 1 (\theta \geq \bar{\theta} > 1)$, the loss function is convex, smooth and Lipschitz	$O\big((\frac{1}{\sqrt{n}}+\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{\theta}{\theta-1}}\big)$	
Algorithm 8♯	$F(\cdot)$ is λ-strongly convex, β-smooth and Lipschitz, the loss function is nonnegative, $n \geq \kappa^\tau$ for some constant $\tau > 1$ with $\kappa = \frac{\beta}{\lambda}$	$O\big(\frac{d\log(1/\delta)}{n^2\epsilon^2}+\frac{4^\tau\cdot F(w^*)}{n}$ $+\frac{2^{2\tau^2+4\tau}}{n^\tau}+\frac{2^{4\tau^2+10\tau}\cdot d\log(1/\delta)}{n^{2\tau}\cdot\epsilon^2}\big)$	(Bassily et al., 2019) $\Omega\big(\frac{1}{n}+\frac{d\log(1/\delta)}{n^2\epsilon^2}\big)$

Table 1: Summary of the main results (in terms of the excess population risk) of the (ϵ,δ)-DP algorithms proposed in our paper. All the methods can be extended to the ϵ-DP case and the term of $O(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})$ will be replaced by $O(\frac{d}{n\epsilon})$ in the above upper bounds and lower bound. The Big-O and Big-Ω notations omit other terms of λ, smoothness, and Lipschitz constant. For Algorithm 2, 4 and 5, the smoothness assumption can be further removed with the same upper bounds (see Section 4.2 for details). ♣: The algorithm needs to efficiently implement the projection of a given vector onto the intersection of the underlying constraint set $\mathcal{W}$ and any given ℓ_2-norm ball, which is difficult to implement in practice. ⋆: The algorithms needs the prior knowledge of θ, i.e., θ should be an input of the algorithm. ♠: Here θ could be unknown in advance but we assume that $\theta \geq \bar{\theta} > 1$ for some known $\bar{\theta}$, i.e., $\bar{\theta}$ will be an input of the algorithm. Unlike other upper bounds, here the upper bound is not for the exact excess population risk (see Theorem 12). ♯: $F(w^*)$ is the optimal minimal function value of $F(\cdot)$.

(see Fact 1-4 for details). Our contributions can be summarized as follows (see Table 1 for details).

- In the first part of the paper, we study the problem where the population risk satisfying TNC with parameter θ and propose three methods. When $\theta \geq 2$, we first propose a method that could achieve an excess population risk of $\widetilde{O}((\frac{1}{\sqrt{n}}+\frac{d}{n\epsilon})^{\frac{\theta}{\theta-1}})$ and $\widetilde{O}((\frac{1}{\sqrt{n}}+\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{\theta}{\theta-1}})$ in ϵ-DP and (ϵ,δ)-DP model respectively under the assumption that the loss function is smooth and Lipschitz, where n is the sample size of the data and d is the dimension of the space. We then propose another method to resolve the inefficiency issue under the assumption that θ is known. Moreover, we propose an improved

method. Compared with the previous two methods, it improves the upper bounds of error by Poly($\log n$) factors. And it only needs a relaxed assumption of $\theta \geq \bar{\theta} > 1$ for some known $\bar{\theta}$ instead of θ being known or $\theta \geq 2$. Moreover, it outperforms the previous methods practically. Next, we focus on the lower bounds of the excess population risk. Specifically, for any $\theta \geq 2$, we show that there is a population risk function satisfying TNC with parameter θ such that for any ϵ-DP ((ϵ, δ)-DP) algorithm, its output achieves an excess risk of $\Omega((\frac{d}{n\epsilon})^{\frac{\theta}{\theta-1}})$ ($\Omega((\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{\theta}{\theta-1}})$) with high probability.

- In the second part of the paper, we will focus on the problem where the population risk function is strongly convex, which is a special case of TNC functions with $\theta = 2$. Unlike the previous studies, here we assume the loss function is non-negative and the optimal value of the population is sufficiently small. With these additional assumptions, we propose a new method whose output could achieve an upper bound of $O(\frac{d\log(1/\delta)}{n^2\epsilon^2} + \frac{1}{n^\tau})$ and $O(\frac{d^2}{n^2\epsilon^2} + \frac{1}{n^\tau})$ for any $\tau > 1$ in (ϵ, δ)-DP and ϵ-DP model respectively if the sample size n is sufficiently large. These rates circumvent their corresponding lower bounds for general strong convex functions in (Feldman et al., 2020), *i.e.,* $\Theta(\frac{d^2}{n^2\epsilon^2} + \frac{1}{n})$ for ϵ-DP and $\Theta(\frac{d\log(1/\delta)}{n^2\epsilon^2} + \frac{1}{n})$ for (ϵ, δ)-DP.

2. Related Work

Starting from (Chaudhuri et al., 2011), a long list of works have attacked the problems of DP-ERM from different perspectives: (Bassily et al., 2014; Iyengar et al., 2019; Zhou et al., 2020; Song et al., 2020; Wang et al., 2017; Zhang et al., 2017) studied the problems in the low dimensional case and the central model, (Kasiviswanathan and Jin, 2016; Kifer et al., 2012; Talwar et al., 2015; Wang and Gu, 2020; Cai et al., 2020) considered the problems in the high dimensional sparse case and the central model, (Smith et al., 2017; Duchi et al., 2013; Wang et al., 2020a; Duchi et al., 2018) focused on the problems in the local model. However, almost all of these works only focus the case where the empirical risk function is either general convex or strongly convex. For a special class of functions, (Wang et al., 2017) studies the empirical risk functions satisfying Polyak-Lojasiewicz (PL) condition, which is weaker than strongly convexity and show that it is possible to achieve an excess empirical risk of $O(\frac{d\log(1/\delta)}{n^2\epsilon^2})$, which is the same as the strongly convex loss. As we will mention in Remark 16, the PL condition is equivalent to TNC with parameter $\theta = 2$. Thus, in this paper, we extend the result from the empirical risk to the population risk function.

For the problem of DP-SCO, besides the related work we mentioned in the previous section, there is another direction that studies some special cases of DP-SCO. For example, (Bassily et al., 2021) and (Asi et al., 2021a) consider the case where the underlying constraint set $\mathcal{W}$ has specific geometric structures, such as polyhedron. (Guzmán et al., 2021) studies the (non)smooth and (non)convex generalized linear loss. (Wang et al., 2020b) and (Kamath et al., 2021) focus on the case where the distribution of the data or the gradient of the loss function is heavy-tailed. However, none of these works study the case where the population risk satisfies TNC. (Liu et al., 2021) recently studies the theoretical guarantees of the PATE model (Papernot et al., 2016) under the assumption that the population risk function satisfies TNC and shows that it is possible to achieve faster rates than in the convex case

(Bassily et al., 2018). However, since here we focus on a different problem, their results cannot be used for DP-SCO.

2.1 Comparison with Concurrent Work

We notice that (Asi et al., 2021b) also studies DP-SCO with TNC population risk functions concurrently. However, compared with its results there are several critical differences. In the following, we will give the details of these differences.

1) For the upper bound, the idea of Algorithm 2 in (Asi et al., 2021b) is similar to Algorithm 2 in our paper. However, the idea of proof and the choice of parameters are quite different. Moreover, we then propose new methods and get improved upper bounds (Theorem 14) as compared to Theorem 2 and 3 in (Asi et al., 2021b).

2) There are some efficiency issues on implementing Algorithm 2 in (Asi et al., 2021b). Firstly, the algorithm in (Asi et al., 2021b) heavily depends on a localization algorithm (Algorithm 1 in (Asi et al., 2021b)) that needs to get the exact optimal solution of some ERM problem, which is inefficient in general. Compare to it, our Algorithm 2 does not need to exactly solve ERM problems. Moreover, even if we are allowed to use some optimization methods to solve the ERM problem in (Asi et al., 2021b) at each phase, we still need a subroutine of projecting a vector onto the ball $\mathcal{W} \cap \mathbb{B}(\hat{w}_{k-1}, R_{k-1})$ at each iteration of the optimization method. However, such a projection step could be costly and even inefficient. Compared to (Asi et al., 2021b), we further propose two other more efficient and practical algorithms (Algorithm 3 and 5) with nearly the same utility upper bounds.

3) For lower bounds, (Asi et al., 2021b) also considers the (ϵ, δ)-DP model (for excess empirical risk). Specifically, they also shows a worst-case lower bound of $\Omega((\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{\theta}{\theta-1}})$ for $\theta \geq 2$ (Theorem 6 and Proposition 3 in (Asi et al., 2021b)), which is the same as our Theorem 19. Although the hard instance in (Asi et al., 2021b) is similar to ours, the proofs of lower bounds are different. Specifically, (Asi et al., 2021b) shows a reduction in solving general convex ERM in the DP model while we construct hard instances under the TNC condition. For ϵ-DP, (Asi et al., 2021b) considers the minimax rate and shows an **information-theoretic lower bound** of $\widetilde{\Omega}((\frac{1}{n} + \frac{d^2}{n^2\epsilon^2})^{\frac{\theta}{2(\theta-1)}})$ for $\kappa \geq 2$ (Theorem 4 in (Asi et al., 2021b)) while we show a **worst-case lower bound** of $\Omega((\frac{d^2}{n^2\epsilon^2})^{\frac{\theta}{2(\theta-1)}})$. It is also notable that for the case where $d = 1$, (Asi et al., 2021b) also shows an **information-theoretic lower bound** of $\widetilde{\Omega}((\frac{1}{n}+\frac{1}{n^2\epsilon^2})^{\frac{\theta}{\theta-1}})$ for $\kappa \in (1, 2]$ (Theorem 5 in (Asi et al., 2021b)). Thus, due to different notations of the lower bound, our results on ϵ-DP are incomparable to the results in (Asi et al., 2021b).

4) In this paper, we also provide experimental results on the problem which has not been studied in (Asi et al., 2021b).

5) Besides TNC population risk functions, in this paper we also provide faster rates of DP-SCO with strongly convex loss function with additional assumptions which also has not been studied in (Asi et al., 2021b).

3. Preliminaries

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

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

If $\delta = 0$, we call algorithm $\mathcal{A}$ is ϵ-DP.

In this paper, we will focus on both ϵ and (ϵ, δ)-DP and we will mainly use the Gaussian mechanism and Laplacian mechanism to guarantee the DP property.

Definition 2 (Gaussian Mechanism) *Given any function $q : \mathcal{X}^n \to \mathbb{R}^d$, the Gaussian mechanism is defined as $q(S) + \xi$ where $\xi \sim \mathcal{N}(0, \frac{16\Delta_2^2(q)\log(1/\delta)}{\epsilon^2}\mathbb{I}_d)$, [2] where where $\Delta_2(q)$ is the ℓ_2-sensitivity of the function q,* i.e., *$\Delta_2(q) = \sup_{S\sim S'} ||q(S) - q(S')||_2$. Gaussian mechanism preserves (ϵ, δ)-DP for $0 < \epsilon, \delta \leq 1$.*

Definition 3 (Laplacian Mechanism) *Given any function $q : \mathcal{X}^n \to \mathbb{R}^d$, the Laplacian mechanism is defined as $\mathcal{M}_G(S, q, \epsilon) = q(S) + (Y_1, Y_2, \cdots, Y_d)$, where each Y_i is i.i.d. drawn from a Laplacian Distribution $Lap(\frac{\Delta_1(q)}{\epsilon})$, where $\Delta_1(q)$ is the ℓ_1-sensitivity of the function q,* i.e., *$\Delta_1(q) = \sup_{S\sim S'} ||q(S) - q(S')||_1$. For a parameter λ, the Laplacian distribution has the density function: $Lap(x|\lambda) = \frac{1}{2\lambda}\exp(-\frac{x}{\lambda})$. Laplacian Mechanism preserves ϵ-DP.*

Definition 4 (DP-SCO (Bassily et al., 2014)) *Given a dataset $S = \{x_1, \cdots, x_n\}$ from a data universe $\mathcal{X}$ where x_i are i.i.d. samples from some unknown distribution $\mathcal{D}$, a convex loss function $f(\cdot, \cdot)$, and a convex constraint set $\mathcal{W} \subseteq \mathbb{R}^d$, Differentially Private Stochastic Convex Optimization (DP-SCO) is to find w^{priv} so as to minimize the population risk,* i.e., *$F(w) = \mathbb{E}_{x\sim\mathcal{D}}[f(w, x)]$ with the guarantee of being differentially private. The utility of the algorithm is measured by the (expected) excess population risk, that is*

$$\mathbb{E}_{\mathcal{A}}[F(w^{priv})] - \min_{w\in\mathcal{W}} F(w),$$

where the expectation of $\mathcal{A}$ is taken over all the randomness of the algorithm. Besides the population risk, we may also measure the empirical risk of dataset S: $\bar{F}(w, S) = \frac{1}{n}\sum_{i=1}^n f(w, x_i)$.

Definition 5 *A function $f : \mathcal{W} \mapsto \mathbb{R}$ is L-Lipschitz over the domain $\mathcal{W}$ if for all $w, w' \in \mathcal{W}$, $|f(w) - f(w')| \leq L||w - w'||_2$.*

Definition 6 *A function $f : \mathcal{W} \mapsto \mathbb{R}$ is β-smooth over the domain $\mathcal{W}$ if for all $w, w' \in \mathcal{W}$,*

$$f(w) \leq f(w') + \langle \nabla f(w'), w - w' \rangle + \frac{\beta}{2}||w - w'||_2^2.$$

2. For simplicity to theoretical analysis, throughout the paper we use constant 16 for Gaussian mechanism. In practice we can use smaller constants.

Definition 7 *A function $F : \mathcal{W} \mapsto \mathbb{R}$ is λ-strongly convex over the domain $\mathcal{W}$ if, for all $w, w' \in \mathcal{W}$,*

$$F(w) + \langle \nabla F(w), w' - w \rangle + \frac{\lambda}{2}||w' - w||_2^2 \leq F(w').$$

Let $w^* = \arg\min_{w \in \mathcal{W}} F(w)$ be the minimizer, strongly convexity implies (Hazan and Kale, 2011):

$$F(w) - F(w^*) \geq \frac{\lambda}{2}||w - w^*||_2^2, \forall w \in \mathcal{W}. \tag{1}$$

Previous work on DP-SCO only focused on cases where the loss function is either convex or strongly convex (Bassily et al., 2019; Feldman et al., 2020). In this paper, we will mainly study the case where the population risk satisfies the Tysbakov Noise Condition (TNC) (Ramdas and Singh, 2012; Liu et al., 2018; Ramdas and Singh, 2013), which has been studied quite well and has been shown that it could achieve faster rates than the optimal one of general convex loss functions in the non-private case. Below we provide the definition of TNC.

Definition 8 *For a convex function $F(\cdot)$, let $\mathcal{W}_* = \arg\min_{w \in \mathcal{W}} F(w)$ denote the optimal set and for any $w \in \mathcal{W}$, let $w^* = \arg\min_{u \in \mathcal{W}_*} \|u - w\|_2$ denote the projection of w onto the optimal set $\mathcal{W}_*$. Function F satisfies (θ, λ)-TNC for some $\theta > 1$ and $\lambda > 0$ if for any $w \in \mathcal{W}$ the following inequality holds*

$$F(w) - F(w^*) \geq \lambda ||w - w^*||_2^\theta. \tag{2}$$

From the definition of TNC and (1) we can see that for a λ-strong convex function it is $(2, \frac{\lambda}{2})$-TNC. Moreover, if a function is (θ, λ)-TNC, then it is also (θ', λ)-TNC for any $\theta < \theta'$. Throughout the whole paper we will assume that θ is a constant and thus we will omit the term of c^θ in the Big-O notation if c is a constant.

Lemma 9 (Lemma 2 in (Ramdas and Singh, 2012)) *If the function $F(\cdot)$ is (θ, λ)-TNC and L-Lipschitz, then we have $||w - w^*||_2 \leq (L\lambda^{-1})^{\frac{1}{\theta - 1}}$ and $F(w) - F(w^*) \leq (L^\theta \lambda^{-1})^{\frac{1}{\theta - 1}}$ for all $w \in \mathcal{W}$, where w^* is defined as in Definition 8.*

4. Optimal Rates of Excess Population Risk

4.1 Upper Bounds of Excess Population Risk

In this section, we will concentrate on the case where the population risk function is (θ, λ)-TNC and provide some upper bounds of its excess population risk. To provide a clear intuition of our methods, we will first assume that the loss functions are smooth. Later we will extend to the non-smooth case based on the same ideas.

We first consider an easier case, where the TNC parameter θ satisfies $\theta \geq 2$. Our algorithm is based on the localization technique proposed by (Feldman et al., 2020), which provides an algorithm, namely Phased-SGD (Algorithm 1) for DP-SCO with general convex loss functions and shows that the algorithm could achieve the optimal rate of expected excess population risk.

Lemma 10 (Theorem 4.4 in (Feldman et al., 2020)) *Let $\mathcal{W} \subseteq \mathbb{R}^d$ be a closed convex set and $f(\cdot, x)$ be β-smooth, convex and L-Lipschitz function over $\mathcal{W}$ for each x. If we set $\eta = \frac{D}{L}\min\{\frac{4}{\sqrt{n}}, \frac{\epsilon}{2\sqrt{d\log(1/\delta)}}\}$ and if $\eta \leq \frac{1}{\beta}$ (*i.e., *n is sufficiently large), then Algorithm 1 will be (ϵ, δ)-DP for all $\epsilon \leq 2\log(1/\delta)$. The output satisfies*

$$\mathbb{E}[F(w_k)] - \min_{w\in\mathcal{W}} F(w) \leq 10LD\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon n}\right).$$

Set $\eta = \frac{D}{L}\min\{\frac{4}{\sqrt{n}}, \frac{\epsilon}{d}\}$ and if $\eta \leq \frac{1}{\beta}$, then Algorithm 1 will be ϵ-DP. Moreover, the output satisfies

$$\mathbb{E}[F(w_k)] - \min_{w\in\mathcal{W}} F(w) \leq 10LD\left(\frac{1}{\sqrt{n}} + \frac{d}{n\epsilon}\right),$$

where $D > 0$ satisfies that $||w_0 - w^*||_2 \leq D$.

We propose our adaptive stochastic approximation algorithm, which is presented in Algorithm 2. The updates are divided into m stages. At each stage, the Phased-SGD algorithm is applied with n_0 samples. Each employment of the Phased-SGD algorithm is warm-started by the initial point that is returned from the last stage.

Algorithm 1 Phased-SGD$(w_0, \eta, n, \mathcal{W})$ algorithm (Feldman et al., 2020)

Input: Dataset $S = \{x_1, \cdots, x_n\}$, convex function $f : \mathcal{W} \times \mathcal{X} \mapsto \mathbb{R}$, initial point $w_0 \in \mathcal{W}$, step size η (will be specified later), privacy parameter ϵ and (or) δ.
Set $k = \lceil \log_2 n \rceil$. partition the whole dataset S into k subsets $\{S_1, \cdots, S_k\}$. Denote n_i as the number of samples in S_i, *i.e.,* $|S_i| = n_i$, where $n_i = \lfloor 2^{-i}n \rfloor$.
for $i = 1, \cdots, k$ **do**
    Let $\eta_i = 4^{-i}\eta$, $w_i^1 = w_{i-1}$.
    for $t = 1, \cdots, n_i$ **do**
        Update $w_i^{t+1} = \prod_{\mathcal{W}}(w_i^t - \eta_i \nabla_w f(w_i^t, x_i^t))$, where x_i^t is the t-th sample of the set S_i.
    end for
    Set $\overline{w}_i = \frac{1}{n_i+1}\sum_{t=1}^{n_i+1} w_i^t$.
    For (ϵ, δ)-DP, $w_i = \overline{w}_i + \xi_i$, where $\xi_i \sim \mathcal{N}(0, \sigma_i^2 \mathbb{I}_d)$ with $\sigma_i = \frac{4L\eta_i\sqrt{\log(1/\delta)}}{\epsilon}$.
    For ϵ-DP, $w_i = \overline{w}_i + \xi_i$, where $\xi_i = (\zeta_1, \cdots, \zeta_d)$ with each $\zeta_j \sim \text{Lap}(\lambda)$ and $\lambda = \frac{4L\eta_i\sqrt{d}}{\epsilon}$.
end for
return w_k

The following theorem states that the output of Algorithm 2 achieves an excess population risk of $\widetilde{O}((\frac{1}{\sqrt{n}} + \frac{d}{n\epsilon})^{\frac{\theta}{\theta-1}})$ and $\widetilde{O}((\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{\theta}{\theta-1}})$ for ϵ-DP and (ϵ, δ)-DP, respectively, if the population risk function satisfies TNC with $\theta \geq 2$.

Theorem 11 *Assume that $F(\cdot)$ satisfies (θ, λ)-TNC and $f(\cdot, x)$ is convex, β-smooth and L-Lipschitz for each x. Then Algorithm 2 is ϵ-DP or (ϵ, δ)-DP based on different stepsizes*

Algorithm 2 Private Stochastic Approximation(w_1, n, R_0)

1: **Input:** Dataset $S = \{x_1, \cdots, x_n\}$, convex function $f : \mathcal{W} \times \mathcal{X} \mapsto \mathbb{R}$, initial point $w_0 \in \mathcal{W}$, privacy parameter ϵ and (or) δ, R_0 satisfies $R_0 \geq \|w_0 - w^*\|_2$.
2: Set $\hat{w}_0 = w_0, m = \lfloor \frac{1}{2}\log_2 \frac{2n}{\log_2 n} \rfloor - 1, n_0 = \lfloor \frac{n}{m} \rfloor$.
3: partition the data S into m disjoint subsets $\{S_1, \cdots, S_m\}$ with each S_i containing n_0 samples.
4: **for** $k = 1, \cdots, m$ **do**
5: For (ϵ, δ)-DP, set $\gamma_k = \frac{R_{k-1}}{L} \cdot \min\left\{\frac{4}{\sqrt{n_0}}, \frac{\epsilon}{2\sqrt{d\log(1/\delta)}}\right\}$ and $R_k = \frac{R_{k-1}}{2}$.
6: For ϵ-DP, set $\gamma_k = \frac{R_{k-1}}{L} \cdot \min\left\{\frac{4}{\sqrt{n_0}}, \frac{\epsilon}{d}\right\}$ and $R_k = \frac{R_{k-1}}{2}$.
7: Denote $\hat{w}_k = \text{Phased-SGD}(\hat{w}_{k-1}, \gamma_k, n_0, \mathcal{W} \cap \mathbb{B}(\hat{w}_{k-1}, R_{k-1}))$, where $\mathbb{B}(\hat{w}_{k-1}, R_{k-1})$ is a ball with center $\hat{w}_{k-1}$ and radius R_{k-1}. The Phased-SGD runs on the k-th subset S_i.
8: **end for**
9: **return** $\hat{w}_m$

$\{\gamma_k\}_{k=1}^m$ and noises if $\gamma_k \leq \frac{1}{\beta}$. Moreover, if $\theta \geq 2$ and n is sufficiently large such that $n \geq 256$ and $\|w_0 - w^*\|_2 \leq R_0$, for (ϵ, δ)-DP we have

$$\mathbb{E}[F(\hat{w}_m)] - \min_{w\in\mathcal{W}} F(w) = O\left(\left(\frac{L^\theta}{\lambda}\right)^{\frac{1}{\theta-1}} \cdot \left(\frac{\sqrt{\log n}}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}\log n}{n\epsilon}\right)^{\frac{\theta}{\theta-1}}\right).$$

And for ϵ-DP we have $\mathbb{E}[F(\hat{w}_m)] - \min_{w\in\mathcal{W}} F(w) = O\left(\left(\frac{L^\theta}{\lambda}\right)^{\frac{1}{\theta-1}} \cdot \left(\frac{\sqrt{\log n}}{\sqrt{n}} + \frac{d\log n}{n\epsilon}\right)^{\frac{\theta}{\theta-1}}\right)$.

Algorithm 3 Phased-SGD-SC($w_0, \gamma, \epsilon, \delta, \mathcal{W}$)

1: **Input:** Dataset $S = \{x_1, \cdots, x_n\}$, convex function $f : \mathcal{W} \times \mathcal{X} \mapsto \mathbb{R}$, initial point $w_0 \in \mathcal{W}$, privacy parameter ϵ and (or) δ. D is a constant satisfying $D \geq \|w_0 - w^*\|_2$.
2: partition the data S into k disjoint subsets $\{S_1, \cdots, S_k\}$, where $k = \lceil \log\log n \rceil$ and for each $i \in [k]$, $|S_i| = n_i = \lfloor \frac{2^{i-2}n}{\log n} \rfloor$.
3: **for** $t = 1, \cdots, s$ **do**
4: Let $w_t = \text{Phased-SGD}(w_{t-1}, \eta_t, n_t, \mathcal{W})$, where the Phased-SGD runs on the t-th subset S_i with loss function $f(w, x) + \frac{1}{2\gamma}\|w - w_0\|_2^2$. For (ϵ, δ)-DP, $\eta_t = \frac{D}{L}\min\{\frac{4}{\sqrt{n_t}}, \frac{\epsilon}{2\sqrt{d\log(1/\delta)}}\}$. For ϵ-DP, $\eta_t = \frac{D}{L}\min\{\frac{4}{\sqrt{n_t}}, \frac{\epsilon}{d}\}$.
5: **end for**
6: **return** w_s

In practice, the main difficulty in implementing Algorithm 2 is the projection onto the ball $\mathcal{W} \cap \mathbb{B}(\hat{w}_{k-1}, R_{k-1})$ in each iteration of the Phased-SGD in each phase. In practice, this could be solved by using the Dykstra's algorithm (Dykstra, 1983; Boyle and Dykstra, 1986), which studied the *best approximation problem:* given m closed and convex sets $W_1, \cdots, \mathcal{W}_m \subseteq \mathbb{R}^d$ and a point $y \in \mathbb{R}^d$, we seek the point in $\mathcal{W}_1 \bigcap \cdots \bigcap \mathcal{W}_m$ (assumed

Algorithm 4 Private Stochastic Approximation-II($w_0, n, \mathcal{W}$)

Input: Dataset $S = \{x_1, \cdots, x_n\}$, convex function $f : \mathcal{W} \times \mathcal{X} \mapsto \mathbb{R}$, initial point $w_0 \in \mathcal{W}$, privacy parameter ϵ and (or) δ, χ_0 is a constant such that $\chi_0 \geq F(w_0) - \min_{w\in\mathcal{W}} F(w)$, TNC parameter θ.
For (ϵ, δ)-DP, set $m = \lfloor -\frac{\theta}{2(\theta-1)} \log_2(\frac{L^2}{\lambda^{\frac{2}{\theta}}}(\frac{1}{n} + \frac{d\log(1/\delta)}{n^2\epsilon^2}))\rfloor, n_0 = \lfloor \frac{n}{m} \rfloor$. For ϵ-DP, set $m = \lfloor -\frac{\theta}{2(\theta-1)} \log_2(\frac{L^2}{\lambda^{\frac{2}{\theta}}}(\frac{1}{n} + \frac{d^2}{n^2\epsilon^2}))\rfloor, n_0 = \lfloor \frac{n}{m} \rfloor$. partition the dataset S into m disjoint subsets $\{S_1, \cdots, S_m\}$ with each S_i containing n_0 samples.
Set $\gamma_0 = \chi_0/(6400L^2(\frac{1}{n_0} + \frac{d\log(1/\delta)}{n_0^2\epsilon^2}))$ for (ϵ, δ)-DP and $\gamma_0 = \chi_0/(6400L^2(\frac{1}{n_0} + \frac{d^2}{n_0^2\epsilon^2}))$ for ϵ-DP.
for $k = 1, \cdots, m$ **do**
    Set $\gamma_k = \frac{\gamma_{k-1}}{2}$.
    Denote $w_k = $ Phased-SGD-SC$(w_{k-1}, \gamma_k, \epsilon, \delta, \mathcal{W})$.
end for
return w_m

nonempty) closest to y, and solve $\min_{u\in\mathcal{W}_1\bigcap\cdots\bigcap\mathcal{W}_m} \|u - y\|_2$. However, in theory, the theoretical guarantee of Theorem 11 may not be held if we use Dykstra's algorithm under the privacy constraint. The main reason is that Dykstra's algorithm can only provide an approximate solution for the projection step. However, the approximate solution may not have the same ℓ_2 (or ℓ_1)-norm sensitivity as the exact solution. Thus, from this view, Algorithm 2 lacks efficiency.

Instead of using Dykstra's algorithm, motivated by (Xu et al., 2017), in the following, we present a new algorithm that only needs the projection onto $\mathcal{W}$. Briefly speaking, instead of considering the original stochastic function, we focus on the problem with an additional strongly convex regularization, *i.e.,* $\min_{w\in\mathcal{W}} F(w) + \frac{1}{2\gamma}\|w - w_1\|_2^2$, where $w_1 \in \mathcal{W}$ is some reference point and γ is some parameter.

Specifically, the same as in Algorithm 2, we first divide the whole algorithm into m stages. In each stage we hope to find a private estimator w^k such that $w^k \approx \arg\min_{w\in\mathcal{W}} F(w) + \frac{1}{2\gamma_k}\|w - w_{k-1}\|_2^2$ with γ_k changing with k. Specifically, we use Algorithm 3 to get such a private estimator. Note that due to the additional ℓ_2 regularization, now the function is strongly convex. Thus, instead of using the original Phased-SGD (Algorithm 1) for general convex loss, here we use a strongly convex version of Phased-SGD, which is adapted from (Feldman et al., 2020). Moreover, since now we have an additional ℓ_2-norm regularization, here we do not need the projection onto the balls $\mathcal{W} \cap \mathbb{B}(\hat{w}_{k-1}, R_{k-1})$ during updates compared with Algorithm 2.

Theorem 12 *Assume that $F(\cdot)$ satisfies (θ, λ)-TNC and $f(\cdot, x)$ is convex, β-smooth and L-Lipschitz for each x. Then Algorithm 4 is ϵ-DP or (ϵ, δ)-DP based on different stepsizes, noises and $\{\gamma_k\}_{k=1}^m$, under the assumption that n is sufficiently large such that $\gamma_0 \geq \frac{\|\mathcal{W}\|_2}{L}$, where $\|\mathcal{W}\|_2$ is the diameter of the set $\mathcal{W}$,* i.e., $\|\mathcal{W}\|_2 = \max_{w,w'\in\mathcal{W}} \|w - w'\|_2$. *Moreover, if θ is known in advance and n is sufficiently large such that $\theta \geq 2\frac{\log\log n}{\log n}$, for (ϵ, δ)-DP, we*

have

$$\min_{k=1,\cdots,m}\mathbb{E}[F(w_k)]-\min_{w\in\mathcal{W}}F(w)=O\left(\left(\frac{L^2}{\lambda^{\frac{2}{\theta}}}\left(\frac{\log n}{n}+\frac{d\log n^2\log(1/\delta)}{n^2\epsilon^2}\right)\right)^{\frac{\theta}{2(\theta-1)}}\right).$$

For ϵ-DP we have $\min_{k=1,\cdots,m}\mathbb{E}[F(w_k)]-\min_{w\in\mathcal{W}}F(w)=O\left(\left(\frac{L^2}{\lambda^{\frac{2}{\theta}}}\left(\frac{\log n}{n}+\frac{d^2\log n^2}{n^2\epsilon^2}\right)\right)^{\frac{\theta}{2(\theta-1)}}\right)$.

So far we have proposed two algorithms. However, there are still several issues: First, both of the previous methods need strong assumptions on θ. To achieve those utility upper bounds, Algorithm 4 needs θ to be known in advance while Algorithm 2 needs to assume $\theta\geq 2$. Thus, can we develop a method that only needs a weaker assumption on θ to get similar utility upper bounds? Secondly, both of the previous two algorithms could achieve rates of $\widetilde{O}((\frac{1}{\sqrt{n}}+\frac{d}{n\epsilon})^{\frac{\theta}{\theta-1}})$ and $\widetilde{O}((\frac{1}{\sqrt{n}}+\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{\theta}{\theta-1}})$ for ϵ-DP and (ϵ,δ)-DP, respectively. Can we further improve these bounds? Thirdly, the two methods are either impractical or inefficient. Specifically, for Algorithm 4, as we can see from our theoretical analysis, we need to **exactly** set γ_0 as $\chi_0/(6400L^2(\frac{1}{n_0}+\frac{d\log(1/\delta)}{n_0^2\epsilon^2}))$ with $\chi_0\geq F(w_0)-\min_{w\in\mathcal{W}}F(w)$ in the (ϵ,δ)-DP model (similar to ϵ-DP). However, getting such an upper bound of $F(w_0)-\min_{w\in\mathcal{W}}F(w)$ in general is difficult to get. And we can see that in Theorem 12 we can only guarantee there exists a w_k that achieves the upper bound of error, it is still unknown how to find such w_k privately with the same theoretical guarantees. For Algorithm 2, it runs several times longer than other algorithms due to the approximate projection steps. Thus, how to design improved methods both theoretically and practically? In the following, we will focus on these three issues by developing a new method.

The idea of our algorithm is as follows: assuming that the value of θ is unknown, but θ is lower bounded by some known constant $\bar{\theta}>1$, namely $\theta\geq\bar{\theta}>1$. We first divide the whole dataset into $k=\lfloor(\log_{\bar{\theta}}2)\cdot\log\log n\rfloor$ disjoint subsets, where the i-th subset has $n_i=\frac{2^{i-1}n}{(\log n)^{\log_{\bar{\theta}}2}}$ samples; then we repeat the Algorithm 1 for k times where each phase runs on the i-th subset and is initialized at the output of the previous phase. See Algorithm 5 for details.

Algorithm 5 Iterated Phased-SGD$(w_1,n,\mathcal{W},\bar{\theta})$

Input: Dataset $S=\{x_1,\cdots,x_n\}$, convex function $f:\mathcal{W}\times\mathcal{X}\mapsto\mathbb{R}$, initial point $w_0\in\mathcal{W}$, privacy parameter ϵ and (or) δ, D is a constant satisfying $D\geq\|w_0-w^*\|_2$.
partition the data S into k disjoint subsets $\{S_1,\cdots,S_k\}$, where $k=\lfloor(\log_{\bar{\theta}}2)\cdot\log\log n\rfloor$ and for each $i\in[k]$, $|S_i|=n_i=\lfloor\frac{2^{i-1}n}{(\log n)^{\log_{\bar{\theta}}2}}\rfloor$.
for $t=1,\cdots,k$ **do**
    Let $w_t=$ Phased-SGD$(w_{t-1},\eta_t,n_t,\mathcal{W})$, where the Phased-SGD runs on the t-th subset S_i. For (ϵ,δ)-DP, $\eta_t=\frac{D}{L}\min\{\frac{4}{\sqrt{n_i}},\frac{\epsilon}{2\sqrt{d\log(1/\delta)}}\}$. For ϵ-DP, $\eta_t=\frac{D}{L}\min\{\frac{4}{\sqrt{n_i}},\frac{\epsilon}{d}\}$.
end for
return w_k

Remark 13 *Although both Algorithm 5 and 2 partition the data into several parts and perform the Phased-SGD several times. There are several differences: First, the sizes of*

subsets in Algorithm 2 are equal, while we partition the data aggressively in Algorithm 5. Secondly, in each phase of Algorithm 5, the convex set to be projected is invariant while in Algorithm 2 we constantly replace it to $\mathcal{W} \cap \mathbb{B}(\hat{w}_{k-1}, R_{k-1})$, which is necessary based on our theoretical analysis.

Theorem 14 *Assume that $F(\cdot)$ is (θ, λ)-TNC with $\theta \geq \bar{\theta} > 1$ for some known constant $\bar{\theta}$, and $f(\cdot, x)$ is convex, β-smooth and L-Lipschitz for each x. If the sample size n is sufficiently large such that $\bar{\theta} \geq 2^{\frac{\log\log n}{(\log n)-1}}$, then Algorithm 5 is either ϵ-DP or (ϵ, δ)-DP for any $\epsilon \leq 2\log(1/\delta)$, based on different step sizes and noises under the assumption that $\eta_t \leq \frac{1}{\beta}$. Moreover, for (ϵ, δ)-DP, the output satisfies*

$$\mathbb{E}[F(w_k)] - \min_{w \in \mathcal{W}} F(w) = O\left(\left(\frac{L^\theta}{\lambda}\right)^{\frac{1}{\theta-1}} \cdot \left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon n}\right)^{\frac{\theta}{\theta-1}}\right).$$

For ϵ-DP, we have

$$\mathbb{E}[F(w_k)] - \min_{w \in \mathcal{W}} F(w) = O\left(\left(\frac{L^\theta}{\lambda}\right)^{\frac{1}{\theta-1}} \cdot \left(\frac{1}{\sqrt{n}} + \frac{d}{\epsilon n}\right)^{\frac{\theta}{\theta-1}}\right).$$

Remark 15 *Compared with Theorem 11 and 12, we can see the upper bounds in Theorem 14 improve factors of Poly$(\log n)$ in both ϵ-DP and (ϵ, δ)-DP models. Moreover, instead of $\theta \geq 2$, we only need the assumption of $\theta \geq \bar{\theta}$ for some known $\bar{\theta} > 1$ in Theorem 14. And as we will see in the experimental part, Algorithm 5 achieves good performance.*

Remark 16 *We can see that it is possible to get faster rates than the rates of strongly convex loss if $\theta < 2$. For example, when $\theta = \frac{3}{2}$, the upper bound of error becomes $O((\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon n})^3)$ in the (ϵ, δ)-DP model. Moreover, when $\theta > 1$, then the bounds will be always higher than the optimal rate for general convex loss as $\frac{\theta}{\theta-1} > 1$. When $\theta = 2$, we have an excess population risk of $O((\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^2)$ and $O((\frac{1}{\sqrt{n}} + \frac{d}{\epsilon n})^2)$ for ϵ-DP and (ϵ, δ)-DP respectively, which matches the optimal rate of DP-SCO with strongly convex function (Feldman et al., 2020). Besides strongly convex functions, there are other problems that satisfy $(2, \lambda)$-TNC, such as the functions satisfying Weak Strong Convexity, Restricted Secant Inequality (RSI), Error Bound (EB) and Polyak-Lojasiewicz (PL) conditions (see Section 2.1 in (Karimi et al., 2016) for details). Thus, Theorem 14 with $\theta = 2$ could be seen as a generalization of the strongly convex case. For Polyak- Lojasiewicz (PL) functions, (Wang et al., 2018b) shows an upper bound of $O(\frac{d\log(1/\delta)}{n^2\epsilon^2})$ for the empirical risk. However, their method cannot be extended to the population risk. In the following we provide some examples that satisfy TNC with $\theta = 2$.*

Fact 1 (Quadratic Problem (Liu et al., 2018)) *Consider the quadratic problem $F(w) = w^T\mathbb{E}_x[A(x)]w + w^T\mathbb{E}_x[b(x)] + c$, where c is a constant. If $\mathbb{E}[A(x)]$ is a positive* **semi-definite** *matrix, the loss function $f(w, x) = w^TA(x)w + w^Tb(x) + c$ is Lipschitz (e.g., $\max\{\|A(x)\|_2, \|b(x)\|_2\} \leq O(1)$) and $\mathcal{W}$ is a bounded polyhedron (e.g., ℓ_1-norm or ℓ_∞-norm*

ball), then the population risk function will be TNC with $\theta = 2$ and the problem will satisfy the assumptions in Theorem 14.

Moreover, for the ℓ_p regularized quadratic problem $F(w) = w^T\mathbb{E}_x[A(x)]w + w^T\mathbb{E}_x[b(x)] + c + \lambda\|w\|_p^p$ for any $p \geq 2$. Under the same assumptions as above, the population risk function will be TNC with $\theta = p$ and the problem will satisfy the assumptions in Theorem 14.

By Fact 1 we can see that for the linear regression problem where $F(w) = \mathbb{E}(x^T w - y)^2$ over a bounded polyhedron $\mathcal{W}$. It is possible to achieve an upper bound of $O(\frac{1}{n} + \frac{d\log(1/\delta)}{n^2\epsilon^2})$ and $O(\frac{1}{n} + \frac{d^2}{n^2\epsilon^2})$ for the excess population risk in the (ϵ, δ)-DP and ϵ-DP model, respectively.

Fact 2 (SCO over ℓ_2-norm ball (Liu et al., 2018)) *Consider the problem of SCO over ℓ_2-norm ball $\min_{\|w\|_2 \leq B} F(w) = \mathbb{E}[f(w, x)]$. If $f(\cdot, x)$ is convex, smooth and Lipschitz, and $\min_{w\in\mathbb{R}^d} F(w) < \min_{\|w\|_2\leq B} F(w)$. Then the population risk is TNC with $\theta = 2$ and satisfies the assumptions in Theorem 14.*

4.2 Extension to Non-smooth Loss

In the previous section, we provided several methods for TNC population risk functions under the assumption that the loss function is smooth. However, we constantly meet the case where the loss is non-smooth. In this section, we will extend the previous methods to the non-smooth case. The observation is that, in both Algorithm 5 and Algorithm 2, we use the Phased-SGD (Algorithm 1) as a sub-routine for several phases. And we need the smoothness condition in Phased-SGD to get the upper bounds in Lemma 10. Thus, to extend to the non-smooth case, the most direct way is to change Phased-SGD to a non-smooth version in both Algorithm 5 and Algorithm 2. (Feldman et al., 2020) provided non-smooth version of Phased-SGD based on proximal mapping for (ϵ, δ)-DP model, namely Phased-ERM, which is shown in Algorithm 6.

Algorithm 6 Phased-ERM$(w_0, \eta, n, \mathcal{W})$ algorithm (Feldman et al., 2020)

1: **Input:** Dataset $S = \{x_1, \cdots, x_n\}$, convex function $f : \mathcal{W} \times \mathcal{X} \mapsto \mathbb{R}$, initial point $w_0 \in \mathcal{W}$, step size η (will be specified later), privacy parameters ϵ, δ.
2: Set $k = \lceil \log_2 n \rceil$. partition the whole dataset into k subsets $\{S_1, \cdots, S_k\}$ where $|S_i| = \lfloor 2^{-i}n \rfloor$.
3: **for** $t = 1, \cdots, k$ **do**
4: Let $n_i = 2^{-i}n$, $\eta_i = 4^{-i}\eta$.
5: Compute $\widetilde{w}_i \in \mathcal{W}$ such that $F_i(\widetilde{w}_i) - \min_{w\in\mathcal{W}} F_i(w) \leq \frac{L^2\eta_i}{n_i}$ with probability at least $1 - \delta$ for

$$F_i(w) = \frac{1}{n_i}\sum_{x\in S_i} f(w, x) + \frac{1}{\eta_i n_i}\|w - w_{i-1}\|_2^2.$$

6: Set $w_i = \widetilde{w}_i + \xi_i$, where $\xi_i \sim \mathcal{N}(0, \sigma_i\mathbb{I}_d)$ with $\sigma_i = \frac{4L\eta_i\sqrt{\log(1/\delta)}}{\epsilon}$.
7: **end for**
8: **return** w_k

Lemma 17 (Theorem 4.8 in (Feldman et al., 2020)) *Set* $\eta = \frac{D}{L}\min\{\frac{4}{\sqrt{n}}, \frac{\epsilon}{2\sqrt{d\log(1/\delta)}}\}$. *Then for the output of Algorithm 6 we have*

$$\mathbb{E}[F(\hat{w})] - \min_{w\in\mathcal{W}} F(w) \leq 10LD\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right).$$

By using Algorithm 6 as subroutine in Algorithm 5 and 2 we have the following result, which is similar to Theorem 14 and 11.

Theorem 18 *Assume that* $F(\cdot)$ *is* (θ,λ)*-TNC and* $f(\cdot,x)$ *is convex and L-Lipschitz for each* x. *For any* $0<\epsilon,\delta<1$, *if we replace the Phased-SGD with Phased-ERM in Algorithm 5 and 2 (we also change the stepsizes), then the two algorithms are* (ϵ,δ)*-DP. Moreover, in Algorithm 5, the output satisfies*

$$\mathbb{E}[F(w_k)] - \min_{w\in\mathcal{W}} F(w) = O\left(\left(\frac{L^\theta}{\lambda}\right)^{\frac{1}{\theta-1}} \cdot \left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right)^{\frac{\theta}{\theta-1}}\right).$$

If n *is sufficiently large such that* $n \geq 256$, *in Algorithm 2, the output satisfies*

$$\mathbb{E}[F(\hat{w}_m)] - F(w^*) = O\left(\left(\frac{L^\theta}{\lambda}\right)^{\frac{1}{\theta-1}} \cdot \left(\frac{\sqrt{\log n}}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}\log n}{n\epsilon}\right)^{\frac{\theta}{\theta-1}}\right).$$

In the following, we provide some examples that satisfy TNC with $\theta = 2$ and with non-smooth loss.

Fact 3 (Hinge Loss (Xu et al., 2017)) *Consider the SVM problem with hinge loss*

$$\min_{w\in\mathcal{W}} F(w) = \mathbb{E}[(1 - y\langle w,x\rangle)_+],$$

where $\mathcal{W}$ is an ℓ_1-norm or ℓ_∞-norm ball and $|\langle w,x\rangle| \leq 1$ for all x and $w\in\mathcal{W}$. Then $F(\cdot)$ satisfies TNC with $\theta = 2$.

Fact 4 (ℓ_1-regularized Problems (Xu et al., 2017)) *Consider the following* ℓ_1*-regularized problem*

$$\min_{\|w\|_1\leq B} F(w) = \mathbb{E}[f(w,x)] + \lambda\|w\|_1,$$

where $\mathbb{E}[f(w,x)]$ *is convex quadratic or piecewise linear, then* $F(w)$ *satisfies TNC with* $\theta = 2$.

4.3 Lower Bounds of Excess Population Risk

In the previous section, we provide an algorithm whose output could achieve an excess population risk of $O((\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{\theta}{\theta-1}})$ and $O((\frac{1}{\sqrt{n}} + \frac{d}{\epsilon n})^{\frac{\theta}{\theta-1}})$ for ϵ-DP and (ϵ,δ)-DP respectively. The question is, can we further improve these bounds? In this section, we show that for all $\theta \geq 2$, the term of $O((\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{\theta}{\theta-1}})$ and $O((\frac{d}{\epsilon n})^{\frac{\theta}{\theta-1}})$ cannot be further improved. We consider the following loss function. Define

$$f(w,x) = -\langle w,x\rangle + \frac{1}{\theta}\|w\|_2^\theta, \|w\|_2 \leq 1, x\in\{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}^d. \tag{3}$$

Theorem 19 (Lower bound of (ϵ, δ)-DP) *Let $n, d \in \mathbb{N}$, $\theta \geq 2$, $\epsilon > 0$ and $\delta = o(\frac{1}{n})$ such that $n \geq \Omega(\frac{\sqrt{d\log(1/\delta)}}{\epsilon})$. For every (ϵ, δ)-Differentially Private algorithm, there is a dataset $S = \{x_1, \cdots, x_n\}$ where each $x_i \in \{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}^d$ such that with probability at least $\frac{1}{3}$ over the randomness of the algorithm, its output w_{priv} satisfies*

$$F(w_{priv}) - \min_{\|w\|_2 \leq 1} F(w) = \Omega\left((\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{\theta}{\theta-1}} \right),$$

where the loss function is given by (3) which is $O(1)$-Lipschitz, and the population risk satisfies $(\theta, O(1))$-TNC.

Theorem 20 (Lower bound of ϵ-DP) *Let $n, d \in \mathbb{N}$, $\theta \geq 2$ and $\epsilon > 0$ such that $n \geq \Omega(\frac{\sqrt{d}}{\epsilon})$. For every ϵ-Differentially Private algorithm, there is a dataset $S = \{x_1, \cdots, x_n\}$ where each $x_i \in \{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}^d$ such that with probability at least $\frac{1}{3}$ over the randomness of the algorithm, its output w_{priv} satisfies*

$$F(w_{priv}) - \min_{\|w\|_2 \leq 1} F(w) = \Omega\left(\left(\frac{d}{n\epsilon}\right)^{\frac{\theta}{\theta-1}} \right),$$

where loss function is given by (3) which is $O(1)$-Lipschitz, and the population risk satisfies $(\theta, O(1))$-TNC.

Remark 21 *From the above theorems, we can see that for the case where $\theta = 2$, the loss function in (3) is reduced to the squared loss, which was used to the lower bound proof of strongly convex loss in (Bassily et al., 2014).*

5. Improved Rates for Strongly Convex Loss

In the previous section, we showed upper and lower bounds of the excess population risk for general TNC population risk functions. Moreover, from Theorem 18 we can see that we get asymptotically the same bound for smooth and non-smooth loss functions in the (ϵ, δ)-DP model. However, in the non-private case, it has been shown that for the strongly convex loss functions, it is possible to get an improved rate compared with the non-smooth ones (Zhang and Zhou, 2019). Thus, our question is, can we get improved rates if the loss functions have additional properties? In the following we will study the strongly convex loss case. Specifically, we will show that when the loss function $f(\cdot, x)$ has additional assumptions on non-negativity and if the optimal value $F(w^*)$ is sufficiently small, it is possible to achieve an upper bound of $O(\frac{d\log(1/\delta)}{n^2\epsilon^2} + \frac{1}{n^\tau})$ for any $\tau > 1$ if the sample size n is sufficiently large.

There are two parts in the algorithm. In the first part, we perform the original Iterated Phased-SGD (Algorithm 5) on the first half of the data to get a good solution to the optimal parameter w^*. After that we perform a new method, namely Epoch-DP-SGD (Algorithm 7) on the second half of the data, which may also be used in other problems. We note that although Algorithm 7 and Algorithm 1 both perform the original DP-SGD algorithm in

Algorithm 7 Epoch-DP-SGD(η_1, n_1, n, w_0)

Input: Parameter λ, dataset $S = \{x_1, \cdots, x_n\}$, convex function $f : \mathcal{W} \times \mathcal{X} \mapsto \mathbb{R}$, the first partition n_1, initial point $w_0 \in \mathcal{W}$, privacy parameter ϵ and (or) δ.
Set $k = \lceil \log \frac{n}{2n_1} + 1 \rceil$ and partition the whole dataset into $\{S_1, S_2, \cdots, S_k\}$. Denote $n_i = |S_i|$, which satisfies $n_{i+1} = 2n_i$ (if there are left samples, we will add them to the last subset).
for $t = 1, \cdots, k$ **do**
    Set $w_i^1 = w_{i-1}$.
    for $t = 1, \cdots, n_i$ **do**
        Update

$$w_i^{t+1} = \prod_{\mathcal{W}} (w_i^t - \eta_i \nabla_w f(w_i^t, x_i^t)), \tag{4}$$

    where x_i^t is the t-th sample in the set S_i.
    end for
    Update $\overline{w}_i = \frac{1}{n_i+1} \sum_{t=1}^{n_i+1} w_i^t$.
    Let $w_i = \overline{w}_i + \xi_i$, where $\xi_i \sim \mathcal{N}(0, \sigma_i^2 \mathbb{I}_d)$ with $\sigma_i = \frac{4L^2\sqrt{\log(1/\delta)}}{n_i \epsilon \lambda}$ for (ϵ, δ)-DP and $\xi_i = (\zeta_1, \cdots, \zeta_d)$ with each $\zeta_j \sim \text{Lap}(\lambda)$ and $\lambda = \frac{4L^2\sqrt{d}}{\lambda n_i \epsilon}$ for ϵ-DP.
    Set $\eta_{i+1} = \eta_i/2$.
end for
return w_k

Algorithm 8 Faster-DPSGD-SC

Input: Parameter β, λ, $\kappa = \frac{\beta}{\lambda}$ and τ. Dataset $S = \{x_1, \cdots, x_n\}$, convex function $f : \mathcal{W} \times \mathcal{X} \mapsto \mathbb{R}$, initial point $w_0 \in \mathcal{W}$, privacy parameter ϵ and (or) δ.
Split the dataset S into S_1, S_2 where $|S_1| = |S_2| = \frac{n}{2}$.
Perform Iterated Phased-SGD($w_0, \frac{n}{2}, \mathcal{W}$) with $\theta = 2$ on S_1. Denote the returned solution as $\hat{w}$.
Perform Epoch-DP-SGD($\frac{1}{4\beta}, 2^{2\tau+3} \cdot \kappa, \frac{n}{2}, \hat{w}$) on S_2. Denote the returned solution by $\widetilde{w}$.
return $\widetilde{w}$

(Bassily et al., 2014) for several phases or epochs. They are quite different: First, as the phase/epoch increases, we decrease the size of the subset (or the number of iterations) in Algorithm 1. While in Algorithm 7 we will increase the size of the subset (or the number of iterations). As we will see in the proof, this increase is necessary. Specifically, we can show that, for strongly convex loss functions, by using our strategy on the size of the subset and stepsize, for any epoch e in Algorithm 7, we have

$$\mathbb{E}[F(w_e)] - F(w^*) \leq \left(\frac{32dL^4\beta \log(1/\delta)}{n_e^2 \epsilon^2 \lambda^2} + \frac{2^{2\tau+3} \cdot \kappa \cdot F(w^*)}{n_e} \right) \cdot \sum_{i=1}^{e} \frac{1}{2^{2(i-1)(\tau-1)}} + \frac{c_1^2 L^2}{\lambda} \left(\frac{2^{2\tau^2+\tau}}{n_e^\tau} + \frac{2^{4\tau^2+4\tau} \cdot d \log(1/\delta)}{n_e^{2\tau} \cdot \epsilon^2} \right).$$

Thus, to get a low utility upper bound, we need a large n_e at the last iteration, and decreasing the size of the subset in Algorithm 1 cannot get such a bound. The second difference is that the initial size of the subset in Algorithm 1 is $\frac{n}{2}$ while it is $2^{2\tau+3}\kappa$ in Algorithm 7, where κ is the condition number $\kappa = \beta/\lambda$ of the population risk functions.

Theorem 22 *Given ϵ and δ, if $f(\cdot, x)$ is convex, L-Lipschitz and β-smooth for all x, Algorithm 8 is either ϵ-DP or (ϵ, δ)-DP, based on different choices on the stepsizes and noises, under the assumption that $\eta_k \leq \frac{2}{\beta}$ in Algorithm 5.*

Theorem 23 *Denote $\min_{w\in\mathcal{W}} F(w) = F(w^*)$ and suppose $n \geq \kappa^\tau$ for some constant $\tau > 1$, and $F(w)$ is L-Lipschitz, λ-strongly convex and β-smooth. For (ϵ, δ)-DP, the output returned by algorithm 8 satisfies*

$$\mathbb{E}[F(\widetilde{w})] - F(w^*) = O\left(\frac{L^4\beta d\log(1/\delta)}{\lambda^2 n^2\epsilon^2} + \frac{4^\tau\cdot\kappa F(w^*)}{n} + \frac{L^2}{\lambda}\left(\frac{2^{2\tau^2+4\tau}}{n^\tau} + \frac{2^{4\tau^2+10\tau}\cdot d\log(1/\delta)}{n^{2\tau}\cdot\epsilon^2}\right)\right).$$

Specifically, when $\tau = \log_\kappa n$, we have for any n,

$$\mathbb{E}[F(\widetilde{w})] - F(w^*) = O\Big(\frac{L^4\beta d\log(1/\delta)}{\lambda^2 n^2\epsilon^2} + \frac{\kappa F(w^*)}{n^{1-2\log_\kappa 2}} + \frac{L^2}{\lambda}\left(\frac{1}{n^{(1-4\log_\kappa 2)\log_\kappa n - 4\log_\kappa 2}} + \frac{2^{4\tau^2+10\tau}\cdot d\log(1/\delta)}{n^{(2-4\log_\kappa 2-10\log_\kappa 2)\log_\kappa n}\cdot\epsilon^2}\right)\Big).$$

For ϵ-DP, the output returned by algorithm 8 satisfies

$$\mathbb{E}[F(\widetilde{w})] - F(w^*) = O\left(\frac{L^4\beta d^2}{\lambda^2 n^2\epsilon^2} + \frac{4^\tau\cdot\kappa F(w^*)}{n} + \frac{L^2}{\lambda}\left(\frac{2^{2\tau^2+4\tau}}{n^\tau} + \frac{2^{4\tau^2+10\tau}\cdot d^2}{n^{2\tau}\cdot\epsilon^2}\right)\right).$$

We note that recently (Wang et al., 2020b) also showed that when the loss function is non-negative and the optimal value of the population risk is small, it is possible to get a non-trivial upper bound for DP-SCO. However, there are some differences: Firstly, (Wang et al., 2020b) only studied the case of DP-SCO with heavy-tailed data while here we study DP-SCO with strongly convex functions. Thus, the problems are different. Moreover, their method is based on the sample-and-aggregate framework, which is impractical, and their result is $O(\frac{d^3F(w^*)}{n\epsilon^4})$ under the assumption that $\nabla F(w^*) = 0$, which may not hold in the case where $\mathcal{W}$ is a close set. Compared with their work, we do not need such a strong assumption and in general, our bound is much smaller than theirs for $F(w^*) = O(1)$.

Remark 24 *Theorem 23 implies that when $n = \Omega(\kappa^\tau)$, the output of Algorithm 8 achieves excess population risks of $O(\frac{d\log(1/\delta)}{n^2\epsilon^2} + \frac{F(w^*)}{n} + \frac{1}{n^\tau})$ and $O(\frac{d^2}{n^2\ \epsilon^2} + \frac{F(w^*)}{n} + \frac{1}{n^\tau})$ for (ϵ, δ)-DP and ϵ-DP, respectively, which are faster than the optimal rates of $O(\frac{1}{n} + \frac{d\log(1/\delta)}{n^2\epsilon^2})$ and $O(\frac{d^2}{n^2\ \epsilon^2} + \frac{1}{n})$ for general strongly convex loss functions, under the assumption that the optimal risk $F(w^*)$ is relatively small. It is also notable that the bounds in Theorem 23 have exponential dependence on the parameter τ, which means τ also cannot be very large. Moreover, due to the large (hidden) constant in the upper bound, the practical performance of Theorem 23 is poor. We leave the problem of designing more practical algorithms for future research.*

Remark 25 *It is notable that recently (Asi et al., 2022) studies the problem of DP-SCO under TNC with $\theta = 2$ and in the interpolation regime, which is similar to our problem in this section. An instance of SCO is an interpolation problem if there exists $w^* \in \mathcal{W}_*$ such that $0 \in \partial f(w^*, x)$ for all $x \sim \mathcal{D}$, where $\mathcal{W}_* = \arg\min_{w\in\mathcal{W}} F(w)$ denote the optimal set. In other words, an interpolation SCO problem indicates that there exists a solution that simultaneously minimizes all the sample losses. Specifically, (Asi et al., 2022) shows that it is possible to achieve an excess population risk of $O\left(\frac{1}{n^\tau} + \exp(-\widetilde{\Theta}(n)) + \exp(-\widetilde{\Theta}(\frac{n\epsilon}{d}))\right)$ and $O\left(\frac{1}{n^\tau} + \exp(-\widetilde{\Theta}(n)) + \exp(-\widetilde{\Theta}(\frac{n\epsilon}{\sqrt{d\log(1/\delta)}}))\right)$ for ϵ-DP and (ϵ, δ)-DP, respectively, where $\tau > 0$ is any constant. It seems like their results are better than ours. However, we claim that due to different assumptions, our results are incomparable to theirs. Note that here we assume the loss is non-negative and the minimal value of $F(w^*)$ is small, which may not satisfy the interpolation problem condition that needs all sample losses to achieve the minimal value simultaneously (unless $F(w^*) = 0$). On the other hand, for an interpolation problem, we also cannot say its minimal value $F(w^*)$ is small.*

6. Experiments

In this section, we provide experimental studies to compare the effectiveness of the proposed methods for several problems satisfying TNC.

Experimental Settings

For the instances satisfying TNC, here we study three examples that have been studied in the previous related work such as (Liu et al., 2018; Xu et al., 2017). The first one is linear regression and the constrained set is the unit ℓ_1-norm ball. As we mentioned in Fact 1, it satisfies TNC with parameter $\theta = 2$. Specifically, we have

$$\min_{||w||_1\leq 1} F(w) \triangleq \mathbb{E}[(\langle w, x\rangle - y)^2]. \tag{5}$$

We also study the ℓ_2-norm regularized logistic regression (with the regularization parameter λ) under the unit ℓ_2-norm ball constraint, which is λ-strongly convex and thus satisfies $(2, \lambda)$-TNC. Specifically, let $h_w(x) = \frac{1}{1+e^{-\langle x,w\rangle}}$ and $y \in \{0, 1\}$, the problem can be written as

$$\min_{||w||_2\leq 1} F(w) \triangleq \mathbb{E}[-y \log h_w(x) - (1 - y)\log(1 - h_w(x))] + \frac{\lambda}{2}||w||_2^2. \tag{6}$$

Here we will set the parameter $\lambda = 10^{-3}$.

Finally, we consider the ℓ_1 constrained ℓ_4-norm linear regression, which has been studied in (Xu et al., 2017) and satisfies TNC with $\theta = 4$ (Liu et al., 2018). Specifically, it can be written as the following.

$$\min_{||w||_1\leq 1} F(w) \triangleq \mathbb{E}[(\langle w, x\rangle - y)^4]. \tag{7}$$

Methods

Although we studied both (ϵ, δ)-DP and ϵ-DP, in practice we are preferable to (ϵ, δ)-DP. Consequently, this section exclusively concentrates on the discussion of (ϵ, δ)-DP. In the context of the three aforementioned instances, we will consider the following baseline methods.

- **DP-SGD (Abadi et al., 2016)**. Notably, the initial version of DP-SGD was introduced in (Bassily et al., 2014). However, its practical performance in its original form was found to be unsatisfactory, as highlighted in (Wang et al., 2017). To address this, we adopt the batched and clipped variant as proposed by (Abadi et al., 2016), which demonstrates improved practical performance. It's important to mention, though, that the algorithm presented in (Abadi et al., 2016) with a general clipping threshold lacks a theoretical guarantee on the excess population risk. Our approach involves conducting hyperparameter tuning to yield optimal outcomes, and we will present the results based on the selected hyperparameters.

- **Phased-SGD (Algorithm 1)**. Theoretically, Phased-SGD in (Feldman et al., 2020) could be considered as the state-of-the-art method for DP-SCO with smooth convex loss functions. Here we adopt the parameter settings delineated in the theoretical results given by (Feldman et al., 2020).

- **Phased-SGD-SC (Algorithm 3)**. Theoretically, Phased-SGD-SC in (Feldman et al., 2020) could be considered as the state-of-the-art method for DP-SCO with smooth and strongly convex loss functions. Here we will follow the parameter setting in the theoretical results given by (Feldman et al., 2020).

- **SC-psgd (Wu et al., 2017)**. The Private Perturbation-based SGD (SC-psgd) for strongly convex loss method proposed by (Wu et al., 2017) is a practically feasible variant of the output perturbation method. As suggested by Iyengar et al. (2019), here we set constant learning rates as this scheme produces the most accurate models.

Regarding our methodologies, it's important to highlight that we will exclude the investigation of Algorithm 4 and Algorithm 8. A closer examination reveals that these algorithms, from a theoretical perspective, incorporate notably large constants, thereby diminishing their practical feasibility. As a result, our focus will be on assessing PSA (Algorithm 2) and Iterated Phased-SGD (Algorithm 5, we denote it as Iterated SGD) with parameter values $\bar{\theta} = 1.5$ and 2 using comparative analysis. As for the initial point in these algorithms, it will be randomly sampled from the constrained set $\mathcal{W}$.

Note that all the algorithms presented in the experimental results are conducted for 20 random runs and we take their averaged testing error over the 20 runs.

Dataset and Parameter Settings

We will implement all the above methods on four real-world datasets from the libsvm website[3], namely a8a ($n = 22,696, d = 123$ for training, and $n = 9,865$ for testing), a9a

3. https://www.csie.ntu.edu.tw/~cjlin/libsvm/

($n = 32,561, d = 123$ for training, and $n = 16,281$ for testing), ijcnn1 ($n = 49,990, d = 22$ for training, and $n = 91,701$ for testing), and w7a ($n = 24,692, d = 300$ for training, and $n = 25,057$ for testing). For each sample in each dataset, we preprocess it to make its feature vector satisfy $\|x\|_1 \leq 1$ so that the loss function will be Lipschitz for some constant.

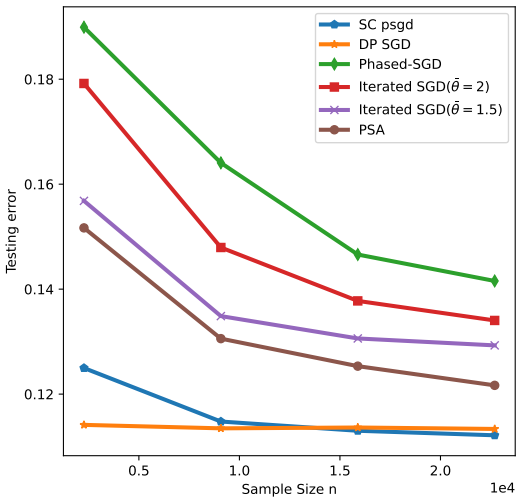
Since it is difficult to get the exact value of the population risk function, here we will use the testing error to approximate it, which is the value of the empirical risk on test data. In the experimental part, we study the above-mentioned three TNC problems and their corresponding testing errors with various sample sizes and privacy budgets ϵ. When performing the results for different sample sizes, we will fix $\epsilon = 0.5$ and consider different sample sizes n that are at most 3.5×10^4. When performing the results for different privacy budgets ϵ, we will use $n = 10^4$ samples and choose $\epsilon = \{0.5, 1, 1.5, 2\}$ respectively. We will set $\delta = \frac{1}{n^{1.1}}$ for all experiments.

Experimental Results

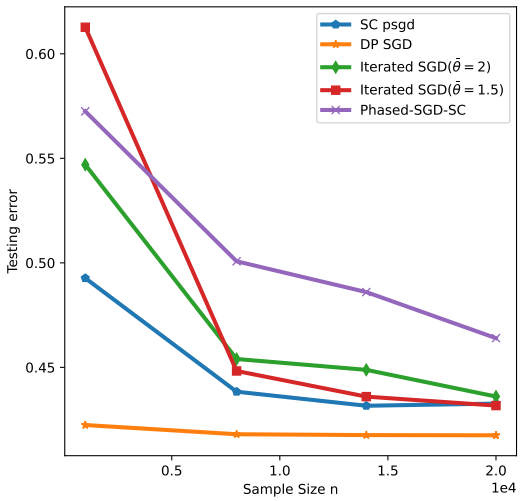
In Figure 1, we show the performance of Iterated SGD with different $\bar{\theta}$ comparing with three baseline methods for ℓ_2-norm regularized logistic regression. First, we can see that for all four datasets, DP-SGD and SC-psgd perform better than Phased-SGD-SC and Iterated SGD, indicating that the latter two methods are less efficient, although they have better upper bounds theoretically. Secondly, compared with Phased-SGD-SC, our methods are better, which is consistent with the observation that the previous linear-time optimal DP-SCO algorithms in (Feldman et al., 2020) do not perform well in practice. Finally, from the results of Iterated SGD with $\bar{\theta} = 2$ and $\bar{\theta} = 1.5$, we can see that our method is quite flexible as the performance difference between these methods is slight. This is due to that we showed that Theorem 14 will hold as long as $\theta \geq \bar{\theta} > 1$. However, we note that the performance could still be different for $\bar{\theta} = 1.5$ and $\bar{\theta} = 2$, and we find that $\bar{\theta} = 1.5$ is better than $\bar{\theta} = 2$. We conjecture it is because the hidden constant in the upper bound of Theorem 14 in the case of $\bar{\theta} = 1.5$ is relatively smaller than the case of $\bar{\theta} = 2$.

Figure 2 shows the results of linear regression for our three methods (Iterated SGD($\bar{\theta} = 2$), Iterated SGD($\bar{\theta} = 1.5$), and PSA) and three baseline methods. We can observe that our methods perform better than Phased-SGD in most cases, except the case when n is large in ijcann1 where Phased-SGD is better than Iterated SGD. However, we think it is acceptable as we can see such a gap is small. Moreover, we can see that compared with Iterated SGD, PSA performs better for all datasets. However, since PSA requires projection to the intersection of two convex sets, it is still inefficient and the performance depends heavily on how accurately we can do the projection while its efficiency depends on the efficiency of the projection. We can also see that DP-SGD and SC-psgd are still the best two methods.

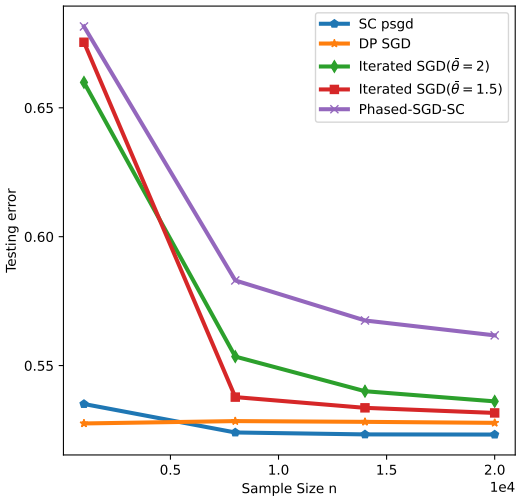
We study the behaviors of our methods and baselines with different privacy budget ϵ in Figure 3-5 for linear regression, ℓ_2-norm regularized logistic regression, and ℓ_4-norm linear regression. We can first see that for almost all the cases, DP-SGD and SC-psgd are the two best methods. However, for ℓ_4-norm linear regression and when ϵ is small, we can see PSA may be the best, e.g., Figure 5 (b). We think this is because PSA could leverage the TNC with $\theta = 4$ for ℓ_4-norm linear regression. Secondly, we can see unlike linear regression or ℓ_4-norm linear regression, PSA has worse performance than Iterated SGD on IJCNN data.



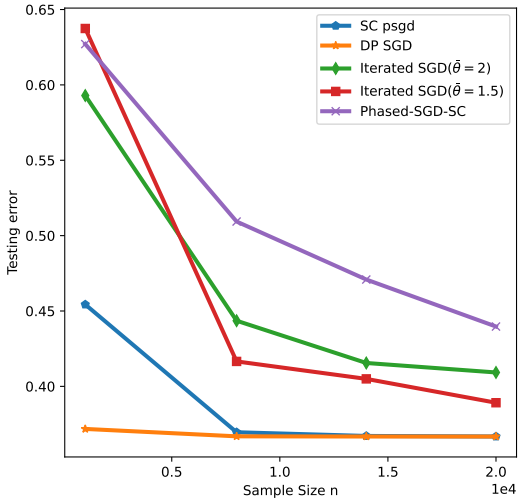
(a) a8a



(b) a9a

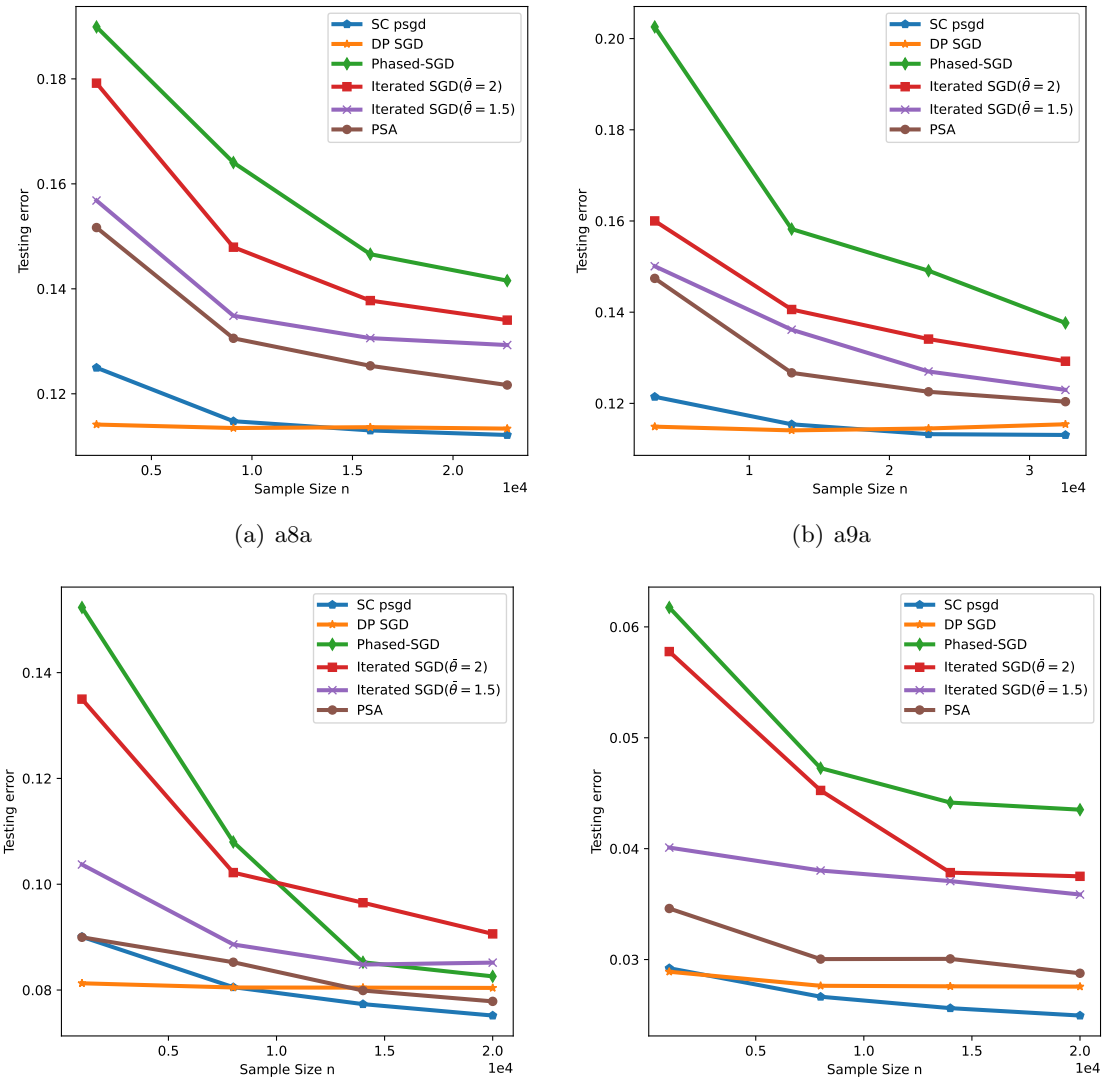


(c) ijcnn1



(d) w7a

Figure 1: Results of l_2-norm regularized logistic regression with different training sample sizes.



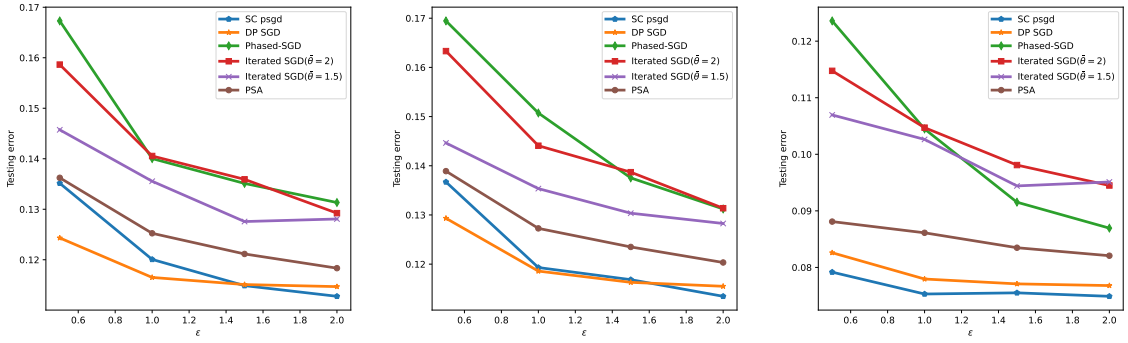
(a) a8a

(b) a9a

(c) ijcnn1

(d) w7a

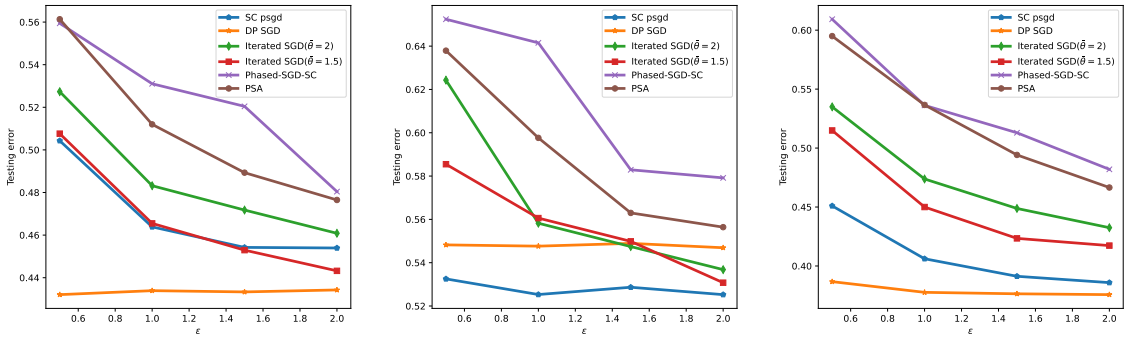
Figure 2: Results of linear regression with different training sample sizes.



(a) a8a (b) a9a (c) ijcnn1

Figure 3: Results of linear regression with different privacy budget ϵ.

Thus, we conjecture the performance of PSA and Iterated SGD heavily depends on the dataset.

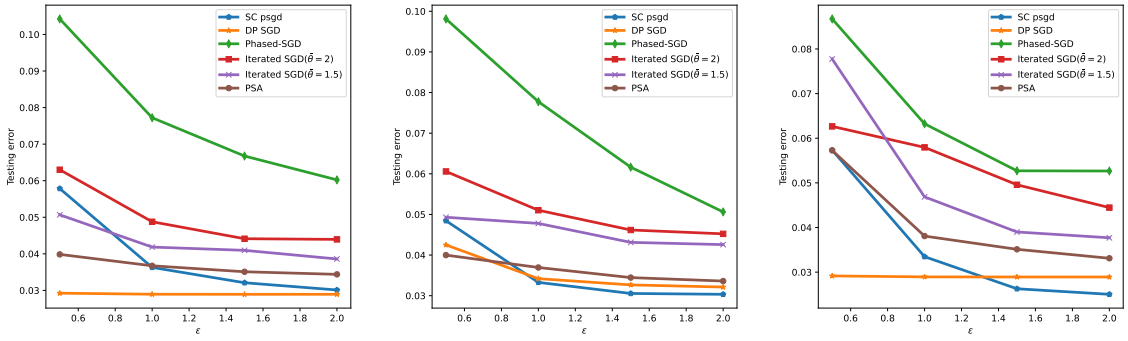


(a) a9a (b) ijcnn1 (c) w7a

Figure 4: Results of l_2-norm regularized logistic regression with different privacy budget ϵ.

7. Conclusion

In this paper, we studied DP-SCO with special classes of population functions. In the first part of the paper, we study the case where the population function satisfies TNC with the parameter $\theta > 1$. Specifically, we first provided several methods which could achieve upper bounds of $\widetilde{O}((\frac{1}{\sqrt{n}}+\frac{d}{n\epsilon})^{\frac{\theta}{\theta-1}})$ and $\widetilde{O}((\frac{1}{\sqrt{n}}+\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{\theta}{\theta-1}})$ for ϵ-DP and (ϵ,δ)-DP, respectively. Then we showed that for any $\theta > 1$, there is a population risk function satisfies TNC with θ such that for any ϵ-DP ((ϵ,δ)-DP) algorithm, the excess population risk of its output is lower bounded by $\Omega((\frac{d}{n\epsilon})^{\frac{\theta}{\theta-1}})$ and $\Omega((\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{\theta}{\theta-1}})$ for ϵ-DP and (ϵ,δ)-DP, respectively. In the second part of the paper, we revisited DP-SCO with strongly convex loss functions. We claimed that when the loss function is non-negative and the optimal value of the population



(a) a8a (b) a9a (c) ijcnn1

Figure 5: Results of ℓ_4-norm linear regression with different privacy budget ϵ.

function is small enough, it is possible to achieve an upper bound of $O(\frac{d\log(1/\delta)}{n^2\epsilon^2}+\frac{1}{n^\tau})$ and $O(\frac{d^2}{n^2\epsilon^2}+\frac{1}{n^\tau})$ for any $\tau>1$ in (ϵ,δ)-DP and ϵ-DP model respectively if the sample size n is sufficiently large.

Besides the open problems we mentioned in the previous parts, there are other unsolved problems: 1) From the theoretical results in this paper, we can see there is still a gap of $O(\frac{1}{n^{\frac{\theta}{2(\theta-1)}}})$ between upper bounds and lower bounds in both ϵ-DP and (ϵ,δ)-DP models. Thus, the optimal rates of excess population risk is still unknown. 2) In this paper we provide faster rates of DP-SCO with special class of functions, especially for TNC population functions. However, besides TNC, there are other special classes of functions which have faster rates in the non-private case, such as exponential concave loss (Koren and Levy, 2015). It is still unknown whether we can get faster rates under the differential privacy constraint. We will leave these problems for future research.

Acknowledgments

Di Wang and Lijie Hu are supported in part by the baseline funding BAS/1/1689-01-01, funding from the CRG grand URF/1/4663-01-01, FCC/1/1976-49-01 from the Computational Bioscience Research Center (CBRC) and funding from the AI Initiative REI/1/4811-10-01 of King Abdullah University of Science and Technology (KAUST). They are also supported by the funding of the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI).

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

Hilal Asi, John Duchi, Alireza Fallah, Omid Javidbakht, and Kunal Talwar. Private adaptive gradient methods for convex optimization. In *International Conference on Machine Learning*, pages 383–392. PMLR, 2021a.

Hilal Asi, Daniel Lévy, and John C Duchi. Adapting to function difficulty and growth conditions in private optimization. *Advances in Neural Information Processing Systems*, 34:19069–19081, 2021b.

Hilal Asi, Karan Chadha, Gary Cheng, and John Duchi. Private optimization in the interpolation regime: faster rates and hardness results. In *International Conference on Machine Learning*, pages 1025–1045. PMLR, 2022.

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

Raef Bassily, Abhradeep Guha Thakurta, and Om Dipakbhai Thakkar. Model-agnostic private learning. *Advances in Neural Information Processing Systems*, 2018.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Thakurta. Private stochastic convex optimization with optimal rates. *arXiv preprint arXiv:1908.09970*, 2019.

Raef Bassily, Cristóbal Guzmán, and Anupama Nandi. Non-euclidean differentially private stochastic convex optimization. *arXiv preprint arXiv:2103.01278*, 2021.

James P Boyle and Richard L Dykstra. A method for finding projections onto the intersection of convex sets in hilbert spaces. In *Advances in order restricted statistical inference*, pages 28–47. Springer, 1986.

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

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

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

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

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

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

Richard L Dykstra. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842, 1983.

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

Cristóbal Guzmán, Raef Bassily, and Michael Menart. Differentially private stochastic optimization: New results in convex and non-convex settings. *arXiv preprint arXiv:2107.05585*, 2021.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv e-prints*, pages arXiv–1509, 2015.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 421–436. JMLR Workshop and Conference Proceedings, 2011.

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

Gautam Kamath, Xingtu Liu, and Huanyu Zhang. Improved rates for differentially private stochastic convex optimization with heavy-tailed data. *arXiv preprint arXiv:2106.01336*, 2021.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

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

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

Tomer Koren and Kfir Y Levy. Fast rates for exp-concave empirical risk minimization. In *NIPS*, pages 1477–1485, 2015.

Janardhan Kulkarni, Yin Tat Lee, and Daogao Liu. Private non-smooth empirical risk minimization and stochastic convex optimization in subquadratic steps. *arXiv preprint arXiv:2103.15352*, 2021.

Chong Liu, Yuqing Zhu, Kamalika Chaudhuri, and Yu-Xiang Wang. Revisiting model-agnostic private learning: Faster rates and active learning. In *International Conference on Artificial Intelligence and Statistics*, pages 838–846. PMLR, 2021.

Mingrui Liu, Xiaoxuan Zhang, Lijun Zhang, Rong Jin, and Tianbao Yang. Fast rates of erm and stochastic approximation: Adaptive to error bound conditions. *arXiv preprint arXiv:1805.04577*, 2018.

Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

Aaditya Ramdas and Aarti Singh. Optimal rates for first-order stochastic convex optimization under tsybakov noise condition. *arXiv preprint arXiv:1207.3012*, 2012.

Aaditya Ramdas and Aarti Singh. Algorithmic connections between active learning and stochastic convex optimization. In *International Conference on Algorithmic Learning Theory*, pages 339–353. Springer, 2013.

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

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

Karthik Sridharan and Ambuj Tewari. Convex games in banach spaces. In *COLT*, pages 1–13. Citeseer, 2010.

Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *arXiv preprint arXiv:1501.06095*, 2015.

Jinyan Su, Lijie Hu, and Di Wang. Faster rates of private stochastic convex optimization. In *Algorithmic Learning Theory*, 2022.

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

Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and XiaoFeng Wang. Privacy loss in apple's implementation of differential privacy on macos 10.12. *CoRR*, abs/1709.02753, 2017.

Tim van Erven, Peter D Grünwald, Nishant A Mehta, Mark D Reid, and Robert C Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems*, pages 2722–2731, 2017.

Di Wang, Marco Gaboardi, and Jinhui Xu. Empirical risk minimization in non-interactive local differential privacy revisited. In *Advances in Neural Information Processing Systems*, pages 965–974, 2018a.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. *arXiv preprint arXiv:1802.05251*, 2018b.

Di Wang, Changyou Chen, and Jinhui Xu. Differentially private empirical risk minimization with non-convex loss functions. In *International Conference on Machine Learning*, pages 6526–6535, 2019a.

Di Wang, Adam Smith, and Jinhui Xu. Noninteractive locally private learning of linear models via polynomial approximations. In *Algorithmic Learning Theory*, pages 897–902, 2019b.

Di Wang, Marco Gaboardi, Adam Smith, and Jinhui Xu. Empirical risk minimization in the non-interactive local model of differential privacy. *Journal of Machine Learning Research*, 21(200):1–39, 2020a.

Di Wang, Hanshen Xiao, Srinivas Devadas, and Jinhui Xu. On differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*, pages 10081–10091. PMLR, 2020b.

Lingxiao Wang and Quanquan Gu. A knowledge transfer framework for differentially private sparse learning. In *AAAI*, pages 6235–6242, 2020.

Xi Wu, Fengan Li, Arun Kumar, Kamalika Chaudhuri, Somesh Jha, and Jeffrey Naughton. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1307–1322. ACM, 2017.

Yi Xu, Qihang Lin, and Tianbao Yang. Stochastic convex optimization: Faster local growth implies faster global convergence. In *International Conference on Machine Learning*, pages 3821–3830. PMLR, 2017.

Tianbao Yang and Qihang Lin. Rsg: Beating subgradient method without smoothness and strong convexity. *The Journal of Machine Learning Research*, 19(1):236–268, 2018.

Tianbao Yang, Zhe Li, and Lijun Zhang. A simple analysis for exp-concave empirical minimization with arbitrary convex regularizer. In *International Conference on Artificial Intelligence and Statistics*, pages 445–453. PMLR, 2018.

Jiaqi Zhang, Kai Zheng, Wenlong Mou, and Liwei Wang. Efficient private erm for smooth objectives. *arXiv preprint arXiv:1703.09947*, 2017.

Lijun Zhang and Zhi-Hua Zhou. Stochastic approximation of smooth and strongly convex functions: Beyond the $o(1/t)$ convergence rate. In *Conference on Learning Theory*, pages 3160–3179. PMLR, 2019.

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

Appendix A. Omitted Proofs

Proof [**Proof of Theorem 11**] For convenience here we only show the proof of (ϵ, δ)-DP. The proof of ϵ-DP is almost the same by replacing the term $(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon n})$ to $(\frac{1}{\sqrt{n}} + \frac{d}{n\epsilon})$ in the following proof.

The guarantee of (ϵ, δ)-DP is just followed by Lemma 10 and the parallel theorem of Differential Privacy. In the following we will focus on the utility.

For simplicity, we denote $a(n) = 10L\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon n}\right)$. We set $\mu_0 = 2R_0^{1-\theta}a(n_0)$, $\mu_k = 2^{(\theta-1)k}\mu_0$ and $R_k = \frac{R_0}{2^k}$, where $k = 1, \cdots, m$.

Then we have $\mu_k \cdot R_k^\theta = 2^{-k}\mu_0 R_0^\theta$. We can also assume that $\lambda \le \frac{L}{R_0^{\theta-1}}$, otherwise we can set $\lambda = \frac{L}{R_0^{\theta-1}}$, which makes TNC still hold.

Recall that $m = \lfloor \frac{1}{2}\log_2 \frac{2n}{\log_2 n} \rfloor - 1$, when $n \ge 256$, it follows that

$$0 < \frac{1}{2}\log_2 \frac{2n}{\log_2 n} - 2 \le m \le \frac{1}{2}\log_2 \frac{2n}{\log_2 n} - 1 \le \frac{1}{2}\log_2 n.$$

Thus, we have $2^m \ge \frac{1}{4}\sqrt{\frac{2n}{\log_2 n}}$.

Thus

$$\begin{aligned}
\mu_m = 2^{(\theta-1)m}\mu_0 &\ge 2^m \mu_0 \\
&\ge \frac{1}{4}\sqrt{\frac{2n}{\log_2 n}} \cdot 2 \cdot R_0^{1-\theta} a(n_0) \\
&= 5 \cdot L R_0^{1-\theta} \sqrt{\frac{2n}{\log_2 n}} \left(\frac{1}{\sqrt{\frac{n}{m}}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon \cdot \frac{n}{m}}\right) \\
&\ge 5 \cdot L R_0^{1-\theta} \sqrt{\frac{2n}{\log_2 n}} \left(\frac{1}{\sqrt{\frac{2n}{\log_2 2n - \log_2\log_2 n - 4}}}\right) \qquad (8) \\
&= 5 \cdot L R_0^{1-\theta} \sqrt{\frac{\log_2 2n - \log\log_2 n - 4}{\log_2 n}} \\
&\ge L R_0^{1-\theta} \quad \left(\text{Since } 5 \cdot \sqrt{\frac{\log_2 2n - \log\log_2 n - 4}{\log_2 n}} \ge 1 \text{ when } n \ge 256\right) \\
&\ge \lambda \text{ (By assumption).}
\end{aligned}$$

where the third inequality is given by throwing away the $\frac{\sqrt{d\log(1/\delta)}}{\epsilon \cdot \frac{n}{m}}$ term and substituting m in term $\frac{1}{\sqrt{\frac{n}{m}}}$ with $\frac{1}{2}\log_2 \frac{2n}{\log_2 n} - 2$.

Below, we consider the following two cases.

Case 1 *If $\lambda \ge \mu_0$, then $\mu_0 \le \lambda \le \mu_m$. We have the following lemma.*

Lemma 26 *Let k^* satisfies $\mu_{k^*} \leq \lambda \leq 2^{\theta-1}\mu_{k^*}$, then for any $1 \leq k \leq k^*$, the points $\{\hat{w}_k\}_{k=1}^m$ generated by Algorithm 2 satisfy*

$$\mathbb{E}[||\hat{w}_{k-1} - w^*||_2] \leq R_{k-1} = 2^{-(k-1)} \cdot R_0, \tag{9}$$

$$\mathbb{E}[F(\hat{w}_k)] - F(w^*) \leq \mu_k R_k^\theta = 2^{-k}\mu_0 R_0^\theta. \tag{10}$$

Moreover, for $k \geq k^*$, we have

$$\mathbb{E}[F(\hat{w}_k)] - \mathbb{E}[F(\hat{w}_{k^*})] \leq \mu_{k^*} R_{k^*}^\theta. \tag{11}$$

Proof *[Proof of Lemma 26] We prove (9), (10) by induction. Note that (9) holds for $k = 1$. Assume (9) is true for some $k > 1$, then we have*

$$\begin{aligned}
\mathbb{E}[F(\hat{w}_k)] - F(w^*) &\leq 10R_{k-1} \cdot L\left(\frac{1}{\sqrt{n_0}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon \cdot n_0}\right) \\
&= R_{k-1}a(n_0) \\
&= \frac{1}{2}\mu_k 2^{(1-\theta)k} R_0^{\theta-1} R_{k-1} \\
&= \mu_k R_k^\theta
\end{aligned} \tag{12}$$

Which is (10). By the definition of TNC, we have

$$\begin{aligned}
\mathbb{E}||\hat{w}_k - w^*||_2^\theta &\leq \frac{1}{\lambda}(\mathbb{E}[F(\hat{w}_k)] - F(w^*)) \\
&\leq \frac{\mathbb{E}[F(\hat{w}_k)] - F(w^*)}{\mu_{k^*}} \\
&\leq \frac{\mu_k R_k^\theta}{\mu_{k^*}} \leq R_k^\theta
\end{aligned} \tag{13}$$

Thus (9) is true for $k+1$.

Now we prove (11). Referring to Lemma 10, we know that

$$\begin{aligned}
\mathbb{E}[F(\hat{w}_k)] - \mathbb{E}[F(\hat{w}_{k-1})] &\leq R_{k-1} \cdot a(n_0) \\
&= 2^{k^*-k} R_{k^*-1} a(n_0) \\
&= 2^{k^*-k}\mu_{k^*} R_{k^*}^\theta \\
&= \mu_k R_k^\theta
\end{aligned}$$

Thus, for $k > k^*$,

$$\begin{aligned}
\mathbb{E}[F(\hat{w}_k)] - \mathbb{E}[F(\hat{w}_{k^*})] &= \sum_{j=k^*+1}^{k} (\mathbb{E}[F(\hat{w}_j)] - \mathbb{E}[F(\hat{w}_{j-1})]) \\
&\leq \sum_{j=k^*+1}^{k} 2^{k^*-j}\mu_{k^*} R_{k^*}^\theta \\
&= (1 - 2^{k^*-k})\mu_{k^*} R_{k^*}^\theta \\
&\leq \mu_{k^*} R_{k^*}^\theta
\end{aligned}$$

Here completes the proof of the lemma. ■

Now we proceed to prove theorem 11 in this case.

$$\begin{aligned}
\mathbb{E}[F(\hat{w}_m)] - F(w^*) &= (\mathbb{E}[F(\hat{w}_m)] - \mathbb{E}[F(\hat{w}_{k^*})]) + (\mathbb{E}[F(\hat{w}_{k^*})] - F(w^*)) \\
&\leq 2\mu_{k^*} R_{k^*}^{\theta} \\
&\leq 4\left(\frac{\mu_{k^*}}{\lambda}\right)^{\frac{1}{\theta-1}} \mu_{k^*} R_{k^*}^{\theta} \ (Since \left(\frac{\mu_{k^*}}{\lambda}\right)^{\frac{1}{\theta-1}} \geq \frac{1}{2}) \\
&= 4\left(\frac{2^{(\theta-1)k^*}\mu_0}{\lambda}\right)^{\frac{1}{\theta-1}} \mu_{k^*} R_{k^*}^{\theta} \\
&= 4(2^{k^*} \mu_{k^*} R_{k^*}^{\theta} \mu_0^{\frac{1}{\theta-1}} \left(\frac{1}{\lambda}\right)^{\frac{1}{\theta-1}}) \\
&= 4(\mu_0 R_0^{\theta} \mu_0^{\frac{1}{\theta-1}} \left(\frac{1}{\lambda}\right)^{\frac{1}{\theta-1}}) \\
&= 4(R_0^{\theta} \mu_0^{\frac{\theta}{\theta-1}} \left(\frac{1}{\lambda}\right)^{\frac{1}{\theta-1}}) \\
&= 4 \cdot ((2 \cdot a(n_0))^{\frac{\theta}{\theta-1}} \left(\frac{1}{\lambda}\right)^{\frac{1}{\theta-1}}) \\
&= 4 \cdot \left(\frac{1}{\lambda}\right)^{\frac{1}{\theta-1}} \cdot \left(20L\left(\frac{\sqrt{m}}{\sqrt{n}} + \frac{m}{n}\frac{\sqrt{d\log(1/\delta)}}{\epsilon}\right)\right)^{\frac{\theta}{\theta-1}}
\end{aligned} \tag{14}$$

where $m = O(\log_2 n)$.(Recall that $m \leq \frac{1}{2}\log_2 n$).

Case 2 *If $\lambda < \mu_0$, then*

$$\begin{aligned}
\mathbb{E}[F(\hat{w}_1)] - F(w^*) &\leq R_0 a(n_0) \\
&= \left(\frac{2}{\mu_0}\right)^{\frac{1}{\theta-1}} \cdot a(n_0)^{\frac{\theta}{\theta-1}} \\
&< \left(\frac{2}{\lambda}\right)^{\frac{1}{\theta-1}} \cdot a(n_0)^{\frac{\theta}{\theta-1}}
\end{aligned}$$

Also, we have

$$\begin{aligned}
\mathbb{E}[F(\hat{w}_m)] - \mathbb{E}[F(\hat{w}_1)] &= \sum_{j=2}^{m} (\mathbb{E}[F(\hat{w}_j)] - \mathbb{E}[F(\hat{w}_{j-1})]) \\
&\leq \sum_{j=2}^{m} R_{j-1} \cdot a(n_0) \\
&= \sum_{j=2}^{m} 2^{-(j-1)} R_0 \cdot a(n_0) \\
&= (1 - (1/2)^{m-1}) R_0 \cdot a(n_0) < R_0 \cdot a(n_0)
\end{aligned}$$

By a similar argument process as in ***Case 1****, we have*

$$\begin{aligned}\mathbb{E}[F(\hat{w}_m)] - F(w^*) &= (\mathbb{E}[F(\hat{w}_m)] - \mathbb{E}[F(\hat{w}_1)]) + (\mathbb{E}[F(\hat{w}_1)] - F(w^*)) \\ &\leq 2R_0 a(n_0) \leq 2\left(\frac{2}{\lambda}\right)^{\frac{1}{\theta-1}} \cdot a(n_0)^{\frac{\theta}{\theta-1}} \\ &= 2 \cdot \left(\frac{2}{\lambda}\right)^{\frac{1}{\theta-1}} \cdot \left(10L\left(\frac{\sqrt{m}}{\sqrt{n}} + \frac{m}{n}\frac{\sqrt{d\log(1/\delta)}}{\epsilon}\right)\right)^{\frac{\theta}{\theta-1}}\end{aligned} \tag{15}$$

Combining the two cases, we conclude that

$$\mathbb{E}[F(\hat{w}_m)] - F(w^*) \leq O\left(\left(\frac{L^\theta}{\lambda}\right)^{\frac{1}{\theta-1}} \cdot \left(\frac{\sqrt{\log n}}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)} \cdot \log n}{n\epsilon}\right)^{\frac{\theta}{\theta-1}}\right)$$

■

Proof [**Proof of Theorem 12**] Before our proof, we provide some notations. We denote $F^* = \min_{w\in\mathcal{W}} F(w)$. For a given error ρ, we denote $\mathcal{L}_\rho$ the ρ-level set of function $F(W)$ and $\mathcal{S}_\rho$ the ρ-sublevel set $F(w)$, respectively, *i.e.,* $\mathcal{L}_\rho = \{w \in \mathcal{W} : F(w) = F^* + \rho\}$, $\mathcal{S}_\rho = \{w \in \mathcal{W} : F(w) \leq F^* + \rho\}$. For any $w \in \mathcal{W}$, we denote w_ρ^+ as the closet point in the ρ-sublevel set to w, *i.e.,*

$$w_\rho^+ = \arg\min_{v\in\mathcal{S}_\rho} \|v - w\|_2^2.$$

Using the KKT condition, it is easy to check that when $w \notin \mathcal{S}_\rho$ then $w_\rho^+ \in \mathcal{L}_\rho$. We first recall the following lemma, given by (Yang and Lin, 2018).

Lemma 27 (Lemma 1 in (Yang and Lin, 2018)) *For any $w \in \mathcal{W}$ and $\rho > 0$ we have*

$$\|w - w_\rho^+\|_2 \leq \frac{dist(w_\rho^+, \mathcal{W}_*)}{\rho}(F(w) - F(w_\rho^+)),$$

where $\mathcal{W}_* = \{w : w \in \arg\min_{w\in\mathcal{W}} F(w)\}$ and $dist(w_\rho^+, \mathcal{W}_*)$ is the distance from the point w_ρ^+ to the set $\mathcal{W}_*$.

Lemma 28 *If $f(\cdot, x)$ is convex, β-smooth and L-Lipschitz for each x and $\gamma \geq \frac{\|\mathcal{W}\|_2}{L}$, where $\|\mathcal{W}\|_2$ is the diameter of the set $\mathcal{W}$,* i.e., *$\|\mathcal{W}\|_2 = \max_{w,w'\in\mathcal{W}} \|w - w'\|_2$. Based on different noises and stepsizes in Algorithm 3, Algorithm 3 is (ϵ, δ) or ϵ-DP if $\eta \leq \frac{1}{\beta}$. Given $w_0 \in \mathcal{W}$, for the output w_k in Algorithm 3. In the case of (ϵ, δ)-DP, we have*

$$\mathbb{E}[\hat{F}(w_s)] - \min_{w\in\mathcal{W}} \hat{F}(w) \leq 3200L^2\gamma(\frac{1}{n} + \frac{d\log(1/\delta)}{n^2\epsilon^2}).$$

In the case of ϵ-DP, we have

$$\mathbb{E}[\hat{F}(w_s)] - \min_{w\in\mathcal{W}} \hat{F}(w) \leq 3200L^2\gamma(\frac{1}{n} + \frac{d^2}{n^2\epsilon^2}),$$

where $\hat{F}(w) = F(w) + \frac{1}{2\gamma}||w - w_0||_2^2$ and w_0 is the initial point.

Proof [**Proof of Lemma 28**] We can see the regularized function of $\hat{F}(w)$ as a population risk with loss function $\widetilde{f}(w,x) = f(w,x) + \frac{1}{2\gamma}\|w - w_0\|_2^2$. Thus, by the assumption of $f(\cdot,x)$, we have $\widetilde{f}(\cdot,x)$ is $L + \frac{\|\mathcal{W}\|_2}{\gamma} \leq 2L$-Lipschitz, $\beta + \frac{1}{\gamma}$-smooth and $\frac{1}{\gamma}$-strongly convex. Thus, by Theorem 5.1 in (Feldman et al., 2020) we have the results.

■

Next we start our proof. For convenience here we only focus on (ϵ,δ)-DP, the proof of ϵ-DP is almost the same. The guarantee of (ϵ,δ)-DP is simply followed by Lemma 28. We also note that Lemma 28 implies that for any $w \in \mathcal{W}$,

$$\mathbb{E}[F(w_k)] - F(w) \leq \frac{1}{2\gamma}||w - w_0||_2^2 + 3200L^2\gamma(\frac{1}{n} + \frac{d\log(1/\delta)}{n^2\epsilon^2}). \tag{16}$$

We denote $\rho = (\frac{8\times 3200L^2}{\lambda^{\frac{2}{\theta}}}(\frac{1}{n_0} + \frac{d\log(1/\delta)}{n_0^2\epsilon^2}))^{\frac{\theta}{2(\theta-1)}}$, $\chi_k = \frac{\chi_0}{2^k}$ and $\gamma_k = \frac{\gamma_0}{2^k}$. Then we have

$$\frac{1}{\gamma_0} = \frac{\lambda^{\frac{2}{\theta}}}{4\chi_0}\rho^{\frac{2(\theta-1)}{\theta}} = \frac{2^{k-2}\lambda^{\frac{2}{\theta}}}{\chi_k}\rho^{\frac{2(\theta-1)}{\theta}}. \tag{17}$$

We assume that for all $i \in \{0, 1\cdots, m-1\}$, $\mathbb{E}[F(w_i)] - F^* > 2\rho$. Otherwise we have proved the theorem.

We will show by induction that

$$\mathbb{E}[F(w_k)] - F^* \leq \chi_k + \rho. \tag{18}$$

If this is true then when $w = m$ we have

$$\mathbb{E}[F(w_k)] - F^* \leq O\left((\frac{L^2}{\lambda^{\frac{2}{\theta}}}(\frac{1}{n_0} + \frac{d\log(1/\delta)}{n_0^2\epsilon^2}))^{\frac{\theta}{2(\theta-1)}}\right).$$

In the following we will show (18). For $k = 0$, by the definition of χ, it is true. Now, consider the k-th phase. By (16) we have

$$\mathbb{E}[F(w_k) - F(w^+_{k-1,\rho})] \leq \underbrace{\frac{1}{2\gamma_k}\mathbb{E}\|w^+_{k-1,\rho} - w_{k-1}\|_2^2}_{A} + \underbrace{3200L^2\gamma_k(\frac{1}{n_0} + \frac{d\log(1/\delta)}{n_0^2\epsilon^2})}_{B}.$$

Since $w_{k-1} \notin \mathcal{S}_\rho$, $w^+_{k-1,\rho} \in \mathcal{L}_\rho$. Moreover, since we have $\mathbb{E}[F(w_{k-1})] - F(w^*) \leq \chi_{k-1} + \rho$, we have $\mathbb{E}[F(w_{k-1})] - \mathbb{E}[F^+(w_{k-1,\rho})] \leq \chi_k$. For term A, by Lemma 27 we have

$$\mathbb{E}\|w^+_{k-1,\rho} - w_{k-1}\|_2 \leq \frac{1}{\lambda^{\frac{1}{\theta}}\rho^{1-\frac{1}{\theta}}}\chi_{k-1}.$$

Thus,

$$\frac{1}{2\gamma_k}\mathbb{E}\|w^+_{k-1,\rho} - w_{k-1}\|_2^2 \leq \frac{1}{2\gamma_k}(\frac{1}{\lambda^{\frac{2}{\theta}}\rho^{\frac{2(\theta-1)}{\theta}}}\chi_{k-1}^2) = \frac{\chi_{k-1}}{4},$$

where the last equality is due to (17).

For term B, we have

$$3200L^2\gamma_k(\frac{1}{n_0}+\frac{d\log(1/\delta)}{n_0^2\epsilon^2}) = 3200L^2\frac{4\chi_0}{2^k\lambda^{\frac{2}{\theta}}\rho^{\frac{2(\theta-1)}{\theta}}}(\frac{1}{n_0}+\frac{d\log(1/\delta)}{n_0^2\epsilon^2}) = \frac{\chi_0}{4\times 2^{k-1}} = \frac{\chi_{k-1}}{4},$$

where the first equality is due to (17). Thus, in total we have

$$\mathbb{E}[F(w_k) - F(w^+_{k-1,\rho})] \le \frac{\chi_{k-1}}{2} = \chi_k.$$

That is $\mathbb{E}[F(w_k)] - F^* \le \chi_k + \rho$. ■

Proof [**Proof of Theorem 14**] In the following we only consider the (ϵ,δ)-DP case. It is almost the same for ϵ-DP.

The guarantee of (ϵ,δ)-DP is just followed by Lemma 10 and the parallel theorem of Differential Privacy. In the following we will focus on the utility.

Since $k = \lfloor(\log_{\bar\theta} 2)\cdot\log\log n\rfloor$, then $k \le (\log_{\bar\theta} 2)\cdot\log\log n$, namely $2^k \le (\log n)^{\log_{\bar\theta} 2}$ and $\frac{2^k-1}{(\log n)^{\log_{\bar\theta} 2}} \le 1$. Observe that the total sample number used in the algorithm is $\sum_{i=1}^k n_i \le \sum_{i=1}^k \frac{2^{i-1}n}{(\log n)^{\log_{\bar\theta} 2}} = \frac{(2^k-1)n}{(\log n)^{\log_{\bar\theta} 2}} \le n$.

For the output of phase i, denote $\Delta_i = \mathbb{E}[F(w_i)] - F(w^*)$, and let $D_i^\theta = \mathbb{E}[||w_i - w^*||_2^\theta]$. The assumption of TNC implies that $F(w_i) - F(w^*) \ge \lambda||w_i - w^*||_2^\theta$, which will be $\mathbb{E}[F(w_i)] - F(w^*) \ge \lambda\mathbb{E}[||w_i - w^*||_2^\theta]$ when we take expectations at both sides, namely

$$\Delta_i \ge \lambda D_i^\theta. \tag{19}$$

Thus, we have

$$\Delta_i \le cLD_{i-1}(\frac{1}{\sqrt{n_i}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon n_i}) \overset{(19)}{\le} cL(\frac{\Delta_{i-1}}{\lambda})^{\frac{1}{\theta}}(\frac{1}{\sqrt{n_i}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon n_i}), \tag{20}$$

where the first inequality comes from Lemma 10 and the second inequality uses (19). Denote $E_i = \frac{c^\theta L^\theta}{\lambda}(\frac{1}{\sqrt{n_i}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon n_i})^\theta$. Then (20) can be simplified as

$$\Delta_i \le (\Delta_{i-1}E_i)^{\frac{1}{\theta}}. \tag{21}$$

Notice that $n_i/n_{i-1} = 2$, then $\frac{E_{i-1}}{E_i} \le (\frac{n_i}{n_{i-1}})^\theta = 2^\theta$, namely:

$$E_i \ge 2^{-\theta}E_{i-1}. \tag{22}$$

Then we can rearrange the above inequality as

$$\frac{\Delta_i}{E_i^{\frac{1}{\theta-1}}} \le \frac{(\Delta_{i-1}E_i)^{\frac{1}{\theta}}}{E_i^{\frac{1}{\theta-1}}} \le 2^{\frac{1}{\theta-1}}\left(\frac{\Delta_{i-1}}{E_{i-1}^{\frac{1}{\theta-1}}}\right)^{\frac{1}{\theta}}, \tag{23}$$

where the first inequality uses (21) and the second inequality applies (22).

It can be verified that (23) is equivalent to

$$\frac{\Delta_i}{2^{\frac{\theta}{(\theta-1)^2}} E_i^{\frac{1}{\theta-1}}} \leq \left(\frac{\Delta_{i-1}}{2^{\frac{\theta}{(\theta-1)^2}} E_{i-1}^{\frac{1}{\theta-1}}} \right)^{\frac{1}{\theta}} \leq \left(\frac{\Delta_1}{2^{\frac{\theta}{(\theta-1)^2}} E_1^{\frac{1}{\theta-1}}} \right)^{\frac{1}{\theta^{i-1}}}.$$

According to Lemma 9, $\Delta_1 \leq (L^\theta \lambda^{-1})^{\frac{1}{\theta-1}}$. Also observe that

$$E_1 = \frac{c^\theta L^\theta}{\lambda} (\frac{1}{\sqrt{n_1}} + \frac{\sqrt{d \log(1/\delta)}}{\epsilon n_1})^\theta \geq \frac{c^\theta L^\theta}{\lambda} \frac{1}{(\sqrt{n_1})^\theta} \geq c^\theta \frac{L^\theta}{\lambda} \frac{1}{n^\theta}.$$

Let $c_1 = c^{\frac{\theta}{\theta-1}} 2^{\frac{\theta}{(\theta-1)^2}}$, then $\frac{\Delta_1}{2^{\frac{\theta}{(\theta-1)^2}} E_1^{\frac{1}{\theta-1}}} \leq \frac{n^{\frac{\theta}{\theta-1}}}{c_1}$, which implies that for $k = \lfloor (\log_{\bar{\theta}} 2) \cdot \log\log n \rfloor$,

$$\frac{\Delta_k}{2^{\frac{\theta}{(\theta-1)^2}} E_k^{\frac{1}{\theta-1}}} \leq \left(\frac{n^{\frac{\theta}{\theta-1}}}{c_1} \right)^{\frac{1}{\theta^{k-1}}}.$$

Let $C_1 = 2^{\frac{\theta^3}{\theta-1} + \theta^2 |\log c_1|}$. In the following we will prove that

$$\left(\frac{n^{\frac{\theta}{\theta-1}}}{c_1} \right)^{\frac{1}{\theta^{k-1}}} \leq C_1.$$

Since $k+1 \geq (\log_{\bar{\theta}} 2) \log\log n \geq (\log_\theta 2) \log\log n$, it follows that

$$(k-1)\log\theta + \log\log C_1 \geq \log(\frac{\theta}{\theta-1} + |\log c_1|) + \log\log n,$$

which indicates

$$(\frac{\theta}{\theta-1} + |\log c_1|) \log n \leq \theta^{k-1} \log C_1.$$

Thus we have $\frac{\theta}{\theta-1} \log n - \log c_1 \leq \theta^{k-1} \log C_1$, which is equivalent to our object $\left(\frac{n^{\frac{\theta}{\theta-1}}}{c_1} \right)^{\frac{1}{\theta^{k-1}}} \leq C_1$.

Now we know

$$\frac{\Delta_k}{2^{\frac{\theta^2}{(\theta-1)^2}} E_k^{\frac{1}{\theta-1}}} \leq \left(\frac{n^{\frac{\theta}{\theta-1}}}{c_1} \right)^{\frac{1}{\theta^{k-1}}} \leq C_1,$$

which indicates that $\frac{\Delta_k}{E_k^{\frac{1}{\theta-1}}} \leq 2^{\frac{\theta}{(\theta-1)^2}} C_1 = 2^{\theta^2 (\frac{\theta^2-\theta+1}{(\theta-1)^2} + |\log c_1|)} := C$.

As a result, we hold a solution with error:

$$\begin{aligned}
\mathbb{E}[F(w_k)] - F(w^*) \leq C E_k^{\frac{1}{\theta-1}} &= C \left(\frac{c^\theta L^\theta}{\lambda} \right)^{\frac{1}{\theta-1}} \left(\frac{1}{\sqrt{n_k}} + \frac{\sqrt{d \log(1/\delta)}}{\epsilon n_k} \right)^{\frac{\theta}{\theta-1}} \\
&\leq 2^{\frac{3\theta}{2(\theta-1)}} \cdot C \left(\frac{c^\theta L^\theta}{\lambda} \right)^{\frac{1}{\theta-1}} \left(\frac{1}{n} + \frac{d \log(1/\delta)}{\epsilon^2 n^2} \right)^{\frac{\theta}{2(\theta-1)}}
\end{aligned}$$

where we use the fact that $n_k = \frac{2^{k-1}}{(\log n)^{\log_{\bar{\theta}} 2}} \geq \frac{1}{2}n$ and $(a+b)^2 \leq 2(a^2+b^2)$. ■

Remark 29 *To perform valid Phased SGD (Subroutine of Iterated Phased-SGD) for k times, it should satisfy* $n_i = \frac{2^{i-1}n}{(\log n)^{\log_{\bar{\theta}} 2}} \geq 2$ *for any* $i \in [k]$. *Otherwise, the Phased SGD cannot function properly to get the bound in Lemma 10. As a result, n should be sufficiently large such that* $\bar{\theta} \geq 2^{\frac{\log\log n}{(\log n)-1}}$.

Proof [**Proof of Theorem 19**]

Based on the fact that a lower bound on excess empirical risk implies nearly the same lower bound on the excess population risk (Bassily et al., 2019), here we consider the empirical risk, then we can use the boosting technique to the population loss. See (Bassily et al., 2019) for details.

Based on the definition of the loss function in (3), we can see that $f(w,x)$ is 2-Lipschiz in $||w||_2 \leq 1$, and it is (θ,λ)-TNC with some constant λ (Sridharan and Tewari, 2010).

For any dataset $S = \{x_1, \cdots, x_n\}$ with data point drawn from $x \in \{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}^d$, and any $w \in \mathcal{W}$, we define the empirical risk function as the following,

$$\hat{F}(w;S) = \sum_{i=1}^{n} \frac{1}{n} f(w,x_i) = -\langle w, \frac{1}{n}\sum_{i=1}^{n} x_i \rangle + \frac{1}{\theta}||w||_2^{\theta}.$$

In the following, we first show that there is a point w^* satisfying $||w^*||_2 \leq 1$, s.t. $\nabla \hat{F}(w^*;S) = 0$. To prove this, we first take the derivative of $\hat{F}(w;S)$ and let it be 0, so we get

$$\nabla \hat{F}(w^*;S) = 0 \Leftrightarrow ||w^*||_2^{\theta-2} \cdot w^* = \frac{\sum_{i=1}^{n} x_i}{n} \tag{24}$$

That is $||w^*||_2^{\theta-1} = ||\frac{\sum_{i=1}^{n} x_i}{n}||_2 \leq 1$, thus w^* must satisfies $\|w^*\|_2 \leq 1$ when $\theta > 1$.

In the following, we denote $\overline{\mathrm{Z}} = \frac{\sum_{i=1}^{n} x_i}{n}$, then $||w^*||_2 = ||\overline{\mathrm{Z}}||_2^{\frac{1}{\theta-1}}$. Thus from (24) we can get $w^* = \frac{\overline{\mathrm{Z}}}{||\overline{\mathrm{Z}}||_2^{\frac{\theta-2}{\theta-1}}}$. Let w_{priv} denote the output of the (ϵ,δ)-differentially private algorithm $\mathcal{A}$, we will show that with probability at least $\frac{1}{3}$,

$$||w_{priv} - w^*|| \geq \Omega\left(\left(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right)^{\frac{1}{\theta-1}}\right)$$

We prove it by showing that the following inequality leads to contradiction.

$$||w_{priv} - w^*|| \leq O\left(\left(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right)^{\frac{1}{\theta-1}}\right) \tag{25}$$

If (25) holds, then

$$||||\overline{\mathrm{Z}}||_2^{\frac{\theta-2}{\theta-1}} w_{priv} - \overline{\mathrm{Z}}|| \leq O\left(\left(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right)^{\frac{1}{\theta-1}} \cdot ||\overline{\mathrm{Z}}||_2^{\frac{\theta-2}{\theta-1}}\right) \tag{26}$$

Recall the following lemma.

Lemma 30 (Lemma 5.1 in (Steinke and Ullman, 2015; Bassily et al., 2014)) *Let $n, d \in \mathbb{N}$, $\epsilon > 0$ and $\delta = o(\frac{1}{n})$. There is a number $M = \Omega\left(\min\left(n, \frac{\sqrt{d\log(1/\delta)}}{\epsilon}\right)\right)$ such that for every (ϵ, δ)-differentially private algorithm $\mathcal{A}$, there is a dataset $S = \{x_1, \cdots, x_n\} \subseteq \{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}^d$ with $||\sum_{i=1}^n x_i||_2 \in [M-1, M+1]$ such that w.p. $\frac{1}{3}$, we have*

$$||\mathcal{A}(S) - \frac{1}{n}\sum_{i=1}^n x_i||_2 = \Omega\left(\min\left(1, \frac{\sqrt{d\log(1/\delta)}}{\epsilon n}\right)\right)$$

For the sake of contradiction, we consider such S described in the above lemma, with probability more than $\frac{2}{3}$, (25) holds. Let $\widetilde{\mathcal{A}}$ be an (ϵ, δ)-differentially private algorithm that first runs $\mathcal{A}$ on the data and then outputs $||\overline{\mathrm{Z}}||_2^{\frac{\theta-2}{\theta-1}} w_{priv}$, and let n be sufficiently large that $n \geq \frac{\sqrt{d\log(1/\delta)}}{\epsilon}$.

Then we have $||\overline{\mathrm{Z}}||_2^{\frac{\theta-2}{\theta-1}} = \Theta\left(\left(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right)^{\frac{\theta-2}{\theta-1}}\right)$, and (26) will become

$$||||\overline{\mathrm{Z}}||_2^{\frac{\theta-2}{\theta-1}} w_{priv} - \overline{\mathrm{Z}}|| = ||\widetilde{\mathcal{A}} - \overline{\mathrm{Z}}|| \leq O\left(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right)$$

which contradicts to Lemma 30. Thus

$$\hat{F}(w_{priv}, S) - \hat{F}(w^*, S) \geq \Omega\left(\left(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right)^{\frac{\theta}{\theta-1}}\right) \tag{27}$$

By the boosting technique in (Bassily et al., 2019), we have with probability at least $\frac{1}{3}$,

$$F(w_{priv}) - \min_{||w||_2 \leq 1} F(w) \geq \Omega\left(\left(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right)^{\frac{\theta}{\theta-1}}\right).$$

■

Proof [**Proof of Theorem 20**] The proof of Theorem 20 is almost the same as the proof of Theorem 19. Instead of using Lemma 30 we use the following lemma:

Lemma 31 (Lemma 5.1 in (Bassily et al., 2014)) *Let $n, d \in \mathbb{N}$, $\epsilon > 0$ such that $n \geq \Omega(\frac{d}{\epsilon})$. There is a number $M = \Omega\left(\min\left(n, \frac{d}{\epsilon}\right)\right)$ such that for every (ϵ, δ)-differentially private algorithm $\mathcal{A}$, there is a dataset $S = \{x_1, \cdots, x_n\} \subseteq \{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}^d$ with $||\sum_{i=1}^n x_i||_2 \in [M-1, M+1]$ such that w.p. $\frac{1}{3}$, we have*

$$||\mathcal{A}(S) - \frac{1}{n}\sum_{i=1}^n x_i||_2 = \Omega\left(\min\left(1, \frac{d}{\epsilon n}\right)\right). \tag{28}$$

■

Proof [**Proof of Theorem 22**] For simplicity, here we only focus on (ϵ, δ)-DP. It is almost the same for ϵ-DP.

In the first step we perform Algorithm 1, which is (ϵ, δ)-DP. Thus, it is sufficient to show that Algorithm 7 is also (ϵ, δ)-DP, *i.e.,* each epoch is (ϵ, δ)-DP. To prove this, we first revoke the stability of One -Pass Projected SGD for strongly convex loss functions, which is given by (Hardt et al., 2015).

Lemma 32 *[Theorem 3.9 in (Hardt et al., 2015)] Assume the loss function $f(\cdot, x)$ is λ-strongly convex and β-smooth with respect to $w \in \mathcal{W}$ for all x. Let S_i and $S_i^{'}$ be two samples of size n_i differing in only a single element. Denote w_i^t and w'^t_i as the outputs of the projected stochastic gradient method (4) on datasets S_i and $S_i^{'}$ respectively, then if $\eta \leq \frac{1}{\beta}$ we have*

$$||w_i^t - w'^t_i|| \leq \frac{2L^2}{\lambda n_i} \tag{29}$$

Recall that in each epoch we perform projected gradient descent for n_i steps using n_i samples, according to Lemma 32, we can bound the sensitivity of w_i^t for each t and we have $||w_i^t - w'^t_i|| \leq \frac{2L^2}{\lambda n_i}$ for all t, where w_i^t and w'^t_i correspond to the solution of two neighboring dataset S_i and $S_i^{'}$ that differs in one sample.

Thus, the sensitivity of $\overline{w}_i = \frac{1}{n_i}\sum_{t=1}^{n_i} w_i^t$ is also $\frac{2L^2}{\lambda n_i}$. By the Gaussian mechanism, adding Gaussian noise with $\sigma_i = \frac{8L^2\sqrt{\log(1/\delta)}}{n_i\lambda\epsilon}$ will preserve (ϵ, δ)-DP. ■

Proof [**Proof of Theorem 23**] For convenience here we only focus on (ϵ, δ)-DP, the proof is almost the same as for ϵ-DP.

Since $F(\cdot)$ is λ-strongly convex, it satisfies $(2, \frac{\lambda}{2})$-TNC. Thus, by Theorem 11 we have

$$\begin{aligned}\mathbb{E}[F(\hat{w})] - F(w^*) \leq \frac{c^2L^2}{\lambda/2}\left(\frac{1}{n/2} + \frac{d\log(1/\delta)}{\epsilon^2(n/2)^2}\right) &\leq \frac{c_1^2L^2}{\lambda}\left(\frac{1}{n} + \frac{d\log(1/\delta)}{\epsilon^2n^2}\right) \\ &\leq \frac{c_1^2L^2}{\lambda}\left(\frac{1}{\kappa^\tau} + \frac{d\log(1/\delta)}{\epsilon^2\kappa^{2\tau}}\right),\end{aligned} \tag{30}$$

where c and c_1 are universal constants and the last inequality is due to the condition of $n \geq \kappa^\tau$.

Now we proceed to analyze the solution returned by Epoch-DP-SGD (Algorithm 7). The following lemma shows how the excess population risk decreases in each epoch.

Lemma 33 (Lemma 1 in (Zhang and Zhou, 2019)) *Assume $f(\cdot, x)$ is non-negative and β-smooth for all x and $F(\cdot)$ is convex. Apply n_i iterations of (4),* i.e., $w_i^{t+1} = \prod_{\mathcal{W}}(w_i^t - \eta_i\nabla_w f(w_i^t, x_i^t))$ *with $\eta_i < 1/(2\beta)$. Then for any $w \in \mathcal{W}$, we have*

$$\mathbb{E}[F(\overline{w}_i)] - F(w) \leq \frac{1}{2\eta_i n_i(1-2\eta_i\beta)}\mathbb{E}[||w_i^1 - w||^2] + \frac{2\eta_i\beta}{(1-2\eta_i\beta)}F(w),$$

where $\overline{w}_i = \frac{1}{n_i}\sum_{t=1}^{n_i} w_i^t$.

Since $f(\cdot, x)$ is β-smooth for all x, we have

$$\begin{aligned} f(w_i) - f(\overline{w}_i) \leq & \langle \nabla f(\overline{w}_i), w_i - \overline{w}_i \rangle + \frac{\beta}{2} ||w_i - \overline{w}_i||_2^2 \\ = & \langle \nabla f(\overline{w}_i), \xi_i \rangle + \frac{\beta}{2} ||\xi_i||_2^2 \end{aligned}$$

Take expectations on both sides w.r.t the data and ξ_i we get

$$\mathbb{E}[F(w_i)] - F(\overline{w}_i) \leq \frac{\beta}{2} \mathbb{E}[||\xi_i||_2^2] = \frac{d\beta\sigma_i^2}{2} = \frac{32dL^4\beta \log(1/\delta)}{n_i^2 \epsilon^2 \lambda^2}.$$

Combining with Lemma 33, we have

$$\begin{aligned} & \mathbb{E}[F(w_i)] - F(w^*) \\ = & \mathbb{E}[F(w_i)] - F(\overline{w}_i) + F(\overline{w}_i) - F(w^*) \\ \leq & \frac{32dL^4\beta \log(1/\delta)}{n_i^2 \epsilon^2 \lambda^2} + \frac{1}{2\eta_i n_i (1 - 2\eta_i \beta)} \mathbb{E}[||w_i^1 - w^*||^2] + \frac{2\eta_i \beta}{(1 - 2\eta_i \beta)} F(w^*) \end{aligned} \tag{31}$$

Based on the above result, we establish the following result of excess population risk of each epoch in Epoch-DP-SGD (Algorithm 7).

Lemma 34 *For any epoch e in Epoch-DP-SGD (Algorithm 7), we have*

$$\begin{aligned} \mathbb{E}[F(w_e)] - F(w^*) \leq & \left(\frac{32dL^4\beta \log(1/\delta)}{n_e^2 \epsilon^2 \lambda^2} + \frac{2^{2\tau+3} \cdot \kappa \cdot F(w^*)}{n_e} \right) \cdot \sum_{i=1}^{e} \frac{1}{2^{2(i-1)(\tau-1)}} \\ & + \frac{c_1^2 L^2}{\lambda} \left(\frac{2^{2\tau^2+\tau}}{n_e^\tau} + \frac{2^{4\tau^2+4\tau} \cdot d \log(1/\delta)}{n_e^{2\tau} \cdot \epsilon^2} \right) \end{aligned}$$

Proof [**Proof of Lemma 34**] We will prove the lemma by induction on e.

Note that by iteration rules in our algorithm, $w_1^1 = \hat{w}$, $w_{e+1}^1 = w_e$, also, by the algorithm setting, we have for any epoch e,

$$\eta_e \beta \leq \eta_1 \beta = \frac{1}{4}. \tag{32}$$

$$\eta_e n_e = \eta_1 n_1 = 2^{2\tau+3} \kappa \cdot \frac{1}{4\beta}. \tag{33}$$

When $e = 1$, from (31), we have

$$\begin{aligned}
\mathbb{E}[F(w_1)] - F(w^*) \leq & \frac{32dL^4\beta\log(1/\delta)}{n_1^2\epsilon^2\lambda^2} + \frac{1}{2\eta_1 n_1(1-2\eta_1\beta)}\mathbb{E}[||w_1^1 - w^*||^2] + \frac{2\eta_1\beta}{(1-2\eta_1\beta)}F(w^*) \\
\overset{(33)}{\leq} & \frac{32dL^4\beta\log(1/\delta)}{n_1^2\epsilon^2\lambda^2} + \frac{\lambda}{2^{2\tau+1}}\mathbb{E}[||w_1^1 - w^*||^2] + 4\eta_1\beta F(w^*) \\
\leq & \frac{32dL^4\beta\log(1/\delta)}{n_1^2\epsilon^2\lambda^2} + \frac{\lambda}{2^{2\tau+1}}\cdot\frac{2}{\lambda}\mathbb{E}[F(w_1^1) - F(w^*)] + 4\eta_1\beta F(w^*) \\
\overset{(30)}{\leq} & \frac{32dL^4\beta\log(1/\delta)}{n_1^2\epsilon^2\lambda^2} + \frac{1}{2^{2\tau}}\frac{c_1^2L^2}{\lambda}\left(\frac{1}{\kappa^\tau} + \frac{d\log(1/\delta)}{\epsilon^2\kappa^{2\tau}}\right) + 4\eta_1\beta F(w^*) \\
\overset{(33)}{\leq} & \frac{32dL^4\beta\log(1/\delta)}{n_1^2\epsilon^2\lambda^2} + \frac{1}{2^{2\tau}}\frac{c_1^2L^2}{\lambda}\left(\frac{2^{2\tau^2+3\tau}}{n_1^\tau} + \frac{2^{4\tau^2+6\tau}d\log(1/\delta)}{\epsilon^2 n_1^{2\tau}}\right) + \frac{2^{2\tau+3}\cdot\kappa F(w^*)}{n_1} \\
\leq & \frac{32dL^4\beta\log(1/\delta)}{n_1^2\epsilon^2\lambda^2} + \frac{c_1^2L^2}{\lambda}\left(\frac{2^{2\tau^2+\tau}}{n_1^\tau} + \frac{2^{4\tau^2+4\tau}d\log(1/\delta)}{\epsilon^2 n_1^{2\tau}}\right) + \frac{2^{2\tau+3}\cdot\kappa F(w^*)}{n_1}.
\end{aligned}$$

Thus the lemma holds for $e = 1$. Now we assume the lemma is true for some $e \geq 1$, then for $e + 1$,

$$\begin{aligned}
& \mathbb{E}[F(w_{e+1})] - F(w^*) \\
\overset{(31)}{\leq} & \frac{32dL^4\beta\log(1/\delta)}{n_{e+1}^2\epsilon^2\lambda^2} + \frac{1}{2\eta_{e+1}n_{e+1}(1-2\eta_{e+1}\beta)}\mathbb{E}[||w_{e+1}^1 - w^*||^2] + \frac{2\eta_{e+1}\beta}{(1-2\eta_{e+1}\beta)}F(w^*) \\
\overset{(32)}{\leq} & \frac{32dL^4\beta\log(1/\delta)}{n_{e+1}^2\epsilon^2\lambda^2} + \frac{1}{\eta_{e+1}n_{e+1}}\mathbb{E}[||w_{e+1}^1 - w^*||^2] + 4\eta_{e+1}\beta F(w^*) \\
\overset{(33)}{\leq} & \frac{32dL^4\beta\log(1/\delta)}{n_{e+1}^2\epsilon^2\lambda^2} + \frac{\lambda}{2^{2\tau+1}}\cdot\frac{2}{\lambda}\mathbb{E}[F(w_e) - F(w^*)] + \frac{\kappa\cdot 2^{2\tau+3}}{n_{e+1}}F(w^*) \\
\leq & \frac{32dL^4\beta\log(1/\delta)}{n_{e+1}^2\epsilon^2\lambda^2} + \frac{1}{2^{2\tau}}\left(\frac{32dL^4\beta\log(1/\delta)}{n_e^2\epsilon^2\lambda^2} + \frac{2^{2\tau+3}\cdot\kappa\cdot F(w^*)}{n_e}\right)\cdot\sum_{i=1}^{e}\frac{1}{2^{2(i-1)(\tau-1)}} \\
& + \frac{1}{2^{2\tau}}\frac{c_1^2L^2}{\lambda}\left(\frac{2^{2\tau^2+\tau}}{n_e^\tau} + \frac{2^{4\tau^2+4\tau}\cdot d\log(1/\delta)}{n_e^{2\tau}\cdot\epsilon^2}\right) + \frac{\kappa\cdot 2^{2\tau+3}}{n_{e+1}}F(w^*) \\
< & \frac{32dL^4\beta\log(1/\delta)}{n_{e+1}^2\epsilon^2\lambda^2}\left(1 + \frac{1}{2^{2\tau-2}}\cdot\sum_{i=1}^{e}\frac{1}{2^{2(i-1)(\tau-1)}}\right) + \frac{c_1^2L^2}{\lambda}\left(\frac{2^{2\tau^2+\tau}}{n_{e+1}^\tau} + \frac{2^{4\tau^2+4\tau}\cdot d\log(1/\delta)}{n_{e+1}^{2\tau}\cdot\epsilon^2}\right) \\
& + \frac{\kappa\cdot 2^{2\tau+3}}{n_{e+1}}F(w^*)\left(1 + \frac{1}{2^{2\tau-1}}\sum_{i=1}^{e}\frac{1}{2^{2(i-1)(\tau-1)}}\right) \\
< & \left(\frac{32dL^4\beta\log(1/\delta)}{n_{e+1}^2\epsilon^2\lambda^2} + \frac{2^{2\tau+3}\cdot\kappa\cdot F(w^*)}{n_{e+1}}\right)\cdot\sum_{i=1}^{e+1}\frac{1}{2^{2(i-1)(\tau-1)}} \\
& + \frac{c_1^2L^2}{\lambda}\left(\frac{2^{2\tau^2+\tau}}{n_{e+1}^\tau} + \frac{2^{4\tau^2+4\tau}\cdot d\log(1/\delta)}{n_{e+1}^{2\tau}\cdot\epsilon^2}\right).
\end{aligned}$$

Thus the lemma holds for $e + 1$ which completes the proof. ■

Now we go back to our proof. The number of epochs made is given by the largest e which satisfies $\sum_{i=1}^{e} n_i \leq \frac{n}{2}$, i.e.,

$$\sum_{i=1}^{e} n_i = n_1(1 + 2 + \cdots + 2^{e-1}) = n_1(2^e - 1) \leq \frac{n}{2}$$

which means the largest value is $E = \lfloor \log_2(\frac{n}{2n_1} + 1) \rfloor$ and the final solution is $\widetilde{w} = w_E$.

From Lemma 34, we have

$$\begin{aligned}
\mathbb{E}[F(w_E)] - F(w^*) \leq & \left(\frac{32L^4 \beta d \log(1/\delta)}{\lambda^2 n_E^2 \cdot \epsilon^2} + \frac{2^{2\tau+3} \cdot \kappa F(w^*)}{n_E} \right) \cdot \sum_{i=1}^{E} \frac{1}{2^{2(i-1)(\tau-1)}} \\
& + \frac{c_1^2 L^2}{\lambda} \left(\frac{2^{2\tau^2+\tau}}{n_E^\tau} + \frac{2^{4\tau^2+4\tau} \cdot d \log(1/\delta)}{n_E^{2\tau} \cdot \epsilon^2} \right) \\
\leq & \left(\frac{32L^4 \beta d \log(1/\delta)}{\lambda^2 n_E^2 \epsilon^2} + \frac{2^{2\tau+3} \cdot \kappa F(w^*)}{n_E} \right) \cdot \frac{2^{2\tau-2}}{2^{2\tau-2} - 1} \\
& + \frac{c_1^2 L^2}{\lambda} \left(\frac{2^{2\tau^2+\tau}}{n_E^\tau} + \frac{2^{4\tau^2+4\tau} \cdot d \log(1/\delta)}{n_E^{2\tau} \cdot \epsilon^2} \right) \\
\leq & \left(\frac{2^{2\tau+9} L^4 \beta d \log(1/\delta)}{\lambda^2 n^2 \epsilon^2} + \frac{2^{4\tau+4} \cdot \kappa F(w^*)}{n} \right) \cdot \frac{1}{2^{2\tau-2} - 1} \\
& + \frac{c_1^2 L^2}{\lambda} \left(\frac{2^{2\tau^2+4\tau}}{n^\tau} + \frac{2^{4\tau^2+10\tau} \cdot d \log(1/\delta)}{n^{2\tau} \cdot \epsilon^2} \right) \\
= & O\left(\frac{L^4 \beta d \log(1/\delta)}{\lambda^2 n^2 \epsilon^2} + \frac{4^\tau \cdot \kappa F(w^*)}{n} + \frac{c_1^2 L^2}{\lambda} \left(\frac{2^{2\tau^2+4\tau}}{n^\tau} + \frac{2^{4\tau^2+10\tau} \cdot d \log(1/\delta)}{n^{2\tau} \cdot \epsilon^2} \right) \right)
\end{aligned}$$

where the last step is due to the fact that $n_E = n_1 2^{E-1} \geq \frac{n_1}{4}(\frac{n}{2n_1} + 1) \geq \frac{n}{8}$. ■