Differentially Private Empirical Risk Minimization Revisited: Faster and More General*

Di Wang
Dept. of Computer Science and Engineering
State University of New York at Buffalo
Buffalo, NY 14260
dwang45@buffalo.edu

Minwei Ye
Dept. of Computer Science and Engineering
State University of New York at Buffalo
Buffalo, NY 14260
minweiye@buffalo.edu

Jinhui Xu
Dept. of Computer Science and Engineering
State University of New York at Buffalo
Buffalo, NY 14260
jinhui@buffalo.edu

Abstract

In this paper we study the differentially private Empirical Risk Minimization (ERM) problem in different settings. For smooth (strongly) convex loss function with or without (non)-smooth regularization, we give algorithms that achieve either optimal or near optimal utility bounds with less gradient complexity compared with previous work. For ERM with smooth convex loss function in high-dimensional ($p \gg n$) setting, we give an algorithm which achieves the upper bound with less gradient complexity than previous ones. At last, we generalize the expected excess empirical risk from convex loss functions to non-convex ones satisfying the Polyak-Lojasiewicz condition and give a tighter upper bound on the utility than the one in [34].

1 Introduction

Privacy preserving is an important issue in learning. Nowadays, learning algorithms are often required to deal with sensitive data. This means that the algorithm needs to not only learn effectively from the data but also provide a certain level of guarantee on privacy preserving. Differential privacy [11] is a rigorous privacy definition for data analysis which provides meaningful guarantees regardless of what an adversary knows ahead of time about individual's data. As a commonly used supervised learning method, Empirical Risk Minimization (ERM) also faces the challenge of achieving simultaneously privacy preserving and learning. Differentially Private (DP) ERM with convex loss function has been extensively studied in the last decade, starting from [8]. In this paper, we revisit this problem and present several improved results.

Problem Setting Given a dataset $D = \{z_1, z_2 \cdots, z_n\}$ from a data universe $\mathcal{X}$, and a closed convex set $\mathcal{C} \subseteq \mathbb{R}^p$, DP-ERM is to find

$$x_* \in \arg\min_{x \in \mathcal{C}} F^r(x, D) = F(x, D) + r(x) = \frac{1}{n} \sum_{i=1}^{n} f(x, z_i) + r(x)$$

with the guarantee of being differentially private. We refer to f as loss function. $r(\cdot)$ is some simple (non)-smooth convex function called regularizer. If the loss function is convex, the utility of the

*This research was supported in part by NSF through grants IIS-1422591, CCF-1422324, and CCF-1716400.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

	Method	Utility Upper Bd	Gradient Complexity	Non smooth Regularizer?
[9][8]	Objective Perturbation	$O(\frac{p}{n^2\epsilon^2})$	N/A	No
[21]	Objective Perturbation	$O(\frac{p}{n^2\epsilon^2} + \frac{\lambda\|\|x_*\|\|^2}{n\epsilon})$	N/A	Yes
[6]	Gradient Perturbation	$O(\frac{p\log^2(n)}{n^2\epsilon^2})$	$O(n^2)$	Yes
[34]	Output Perturbation	$O(\frac{p}{n^2\epsilon^2})$	$O(n\kappa\log(\frac{n\epsilon}{\kappa}))$	No
This Paper	Gradient Perturbation	$O(\frac{p\log(n)}{n^2\epsilon^2})$	$O((n+\kappa)\log(\frac{n\epsilon\mu}{p}))$	Yes

Table 1: Comparison with previous (ϵ, δ)-DP algorithms. We assume that the loss function f is convex, 1-smooth, differentiable (twice differentiable for objective perturbation), and 1-Lipschitz. F^r is μ-strongly convex. Bound and complexity ignore multiplicative dependence on $\log(1/\delta)$. $\kappa = \frac{L}{\mu}$ is the condition number. The lower bound is $\Omega(\min\{1, \frac{p}{n^2\epsilon^2}\})$[6].

algorithm is measured by the expected excess empirical risk, *i.e.* $\mathbb{E}[F^r(x^{\text{private}}, D)] - F^r(x_*, D)$. The expectation is over the coins of the algorithm.

A number of approaches exist for this problem with convex loss function, which can be roughly classified into three categories. The first type of approaches is to perturb the output of a non-DP algorithm. [8] first proposed output perturbation approach which is extended by [34]. The second type of approaches is to perturb the objective function [8]. We referred to it as objective perturbation approach. The third type of approaches is to perturb gradients in first order optimization algorithms. [6] proposed gradient perturbation approach and gave the lower bound of the utility for both general convex and strongly convex loss functions. Later, [28] showed that this bound can actually be broken by adding more restrictions on the convex domain $\mathcal{C}$ of the problem.

As shown in the following tables[2] , the output perturbation approach can achieve the optimal bound of utility for strongly convex case. But it cannot be generalized to the case with non-smooth regularizer. The objective perturbation approach needs to obtain the optimal solution to ensure both differential privacy and utility, which is often intractable in practice, and cannot achieve the optimal bound. The gradient perturbation approach can overcome all the issues and thus is preferred in practice. However, its existing results are all based on Gradient Descent (GD) or Stochastic Gradient Descent (SGD). For large datasets, they are slow in general. In the first part of this paper, we present algorithms with tighter utility upper bound and less running time. Almost all the aforementioned results did not consider the case where the loss function is non-convex. Recently, [34] studied this case and measured the utility by gradient norm. In the second part of this paper, we generalize the expected excess empirical risk from convex to Polyak-Lojasiewicz condition, and give a tighter upper bound of the utility given in [34]. Due to space limit, we leave many details, proofs, and experimental studies in the supplement.

2 Related Work

There is a long list of works on differentially private ERM in the last decade which attack the problem from different perspectives. [17][30] and [2] investigated regret bound in online settings. [20] studied regression in incremental settings. [32] and [31] explored the problem from the perspective of learnability and stability. We will compare to the works that are most related to ours from the utility and gradient complexity (*i.e.,* the number (complexity) of first order oracle $(f(x, z_i), \nabla f(x, z_i))$ being called) points of view. **Table 1** is the comparison for the case that loss function is strongly convex and 1-smooth. Our algorithm achieves near optimal bound with less gradient complexity compared with previous ones. It is also robust to non-smooth regularizers.

Tables 2 and 3 show that for non-strongly convex and high-dimension cases, our algorithms outperform other peer methods. Particularly, we improve the gradient complexity from $O(n^2)$ to $O(n\log n)$ while preserving the optimal bound for non-strongly convex case. For high-dimension case, gradient complexity is reduced from $O(n^3)$ to $O(n^{1.5})$. Note that [19] also considered high-dimension case

[2] Bound and complexity ignore multiplicative dependence on $\log(1/\delta)$.

	Method	Utility Upper Bd	Gradient Complexity	Non smooth Regularizer?
[21]	Objective Perturbation	$O(\frac{\sqrt{p}}{n\epsilon})$	N/A	Yes
[6]	Gradient Perturbation	$O(\frac{\sqrt{p}\log^{3/2}(n)}{n\epsilon})$	$O(n^2)$	Yes
[34]	Output Perturbation	$O([\frac{\sqrt{p}}{n\epsilon}]^{\frac{2}{3}})$	$O(n[\frac{n\epsilon}{d}]^{\frac{2}{3}})$	No
This paper	Gradient Perturbation	$O(\frac{\sqrt{p}}{n\epsilon})$	$O(\frac{n\epsilon}{\sqrt{p}} + n\log(\frac{n\epsilon}{p}))$	Yes

Table 2: Comparison with previous (ϵ, δ)-DP algorithms, where F^r is not necessarily strongly convex. We assume that the loss function f is convex, 1-smooth, differentiable(twice differentiable for objective perturbation), and 1-Lipschitz. Bound and complexity ignore multiplicative dependence on $\log(1/\delta)$. The lower bound in this case is $\Omega(\min\{1, \frac{\sqrt{p}}{n\epsilon}\})$[6].

via dimension reduction. But their method requires the optimal value in the dimension-reduced space, in addition they considered loss functions under the condition rather than ℓ_2- norm Lipschitz.

For non-convex problem under differential privacy, [15][10][13] studied private SVD. [14] investigated k-median clustering. [34] studied ERM with non-convex smooth loss functions. In [34], the authors defined the utility using gradient norm as $\mathbb{E}[||\nabla F(x^{\text{private}})||^2]$. They achieved a qualified utility in $O(n^2)$ gradient complexity via DP-SGD. In this paper, we use DP-GD and show that it has a tighter utility upper bound.

	Method	Utility Upper Bd	Gradient Complexity	Non smooth Regularizer?
[28]	Gradient Perturbation	$O(\frac{\sqrt{G_{\mathcal{C}}^2+\|\|\mathcal{C}\|\|^2}\log(n)}{n\epsilon})$	$O(\frac{n^3\epsilon^2}{(G_{\mathcal{C}}^2+\|\|\mathcal{C}\|\|^2)\log^2(n)})$	Yes
[28]	Objective Perturbation	$O(\frac{G_{\mathcal{C}}+\lambda\|\|\mathcal{C}\|\|^2}{n\epsilon})$	N/A	No
[29]	Gradient Perturbation	$O(\frac{(G_{\mathcal{C}}^{\frac{2}{3}}\log^2(n))}{(n\epsilon)^{\frac{2}{3}}})$	$O(\frac{(n\epsilon)^{\frac{2}{3}}}{G_{\mathcal{C}}^{\frac{2}{3}}})$	Yes
This paper	Gradient Perturbation	$O(\frac{\sqrt{G_{\mathcal{C}}^2+\|\|\mathcal{C}\|\|^2}}{n\epsilon})$	$O\left(\frac{n^{1.5}\sqrt{\epsilon}}{(G_{\mathcal{C}}^2+\|\|\mathcal{C}\|\|^2)^{\frac{1}{4}}}\right)$	No

Table 3: Comparison with previous (ϵ, δ)-DP algorithms. We assume that the loss function f is convex, 1-smooth, differentiable(twice differentiable for objective perturbation), and 1-Lipschitz. The utility bound depends on $G_{\mathcal{C}}$, which is the Gaussian width of $\mathcal{C}$. Bound and complexity ignore multiplicative dependence on $\log(1/\delta)$.

3 Preliminaries

Notations: We let $[n]$ denote $\{1, 2, \ldots, n\}$. Vectors are in column form. For a vector v, we use $||v||_2$ to denote its ℓ_2-norm. For the gradient complexity notation, G, δ, ϵ are omitted unless specified. $D = \{z_1, \cdots, z_n\}$ is a dataset of n individuals.

Definition 3.1 (Lipschitz Function over θ)**.** A loss function $f : \mathcal{C} \times \mathcal{X} \to \mathbb{R}$ is G-Lipschitz (under ℓ_2-norm) over θ, if for any $z \in \mathcal{X}$ and $\theta_1, \theta_2 \in \mathcal{C}$, we have $|f(\theta_1, z) - f(\theta_2, z)| \leq G||\theta_1 - \theta_2||_2$.

Definition 3.2 (L-smooth Function over θ)**.** A loss function $f : \mathcal{C} \times \mathcal{X} \to \mathbb{R}$ is L-smooth over θ with respect to the norm $||\cdot||$ if for any $z \in \mathcal{X}$ and $\theta_1, \theta_2 \in \mathcal{C}$, we have

$$||\nabla f(\theta_1, z) - \nabla f(\theta_2, z)||_* \leq L||\theta_1 - \theta_2||,$$

where $||\cdot||_*$ is the dual norm of $||\cdot||$. If f is differentiable, this yields

$$f(\theta_1, z) \leq f(\theta_2, z) + \langle \nabla f(\theta_2, z), \theta_1 - \theta_2 \rangle + \frac{L}{2}||\theta_1 - \theta_2||^2.$$

We say that two datasets D, D' are neighbors if they differ by only one entry, denoted as $D \sim D'$.

Definition 3.3 (Differentially Private[11])**.** A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private if for all neighboring datasets D, D' and for all events S in the output space of $\mathcal{A}$, we have

$$Pr(\mathcal{A}(D) \in S) \leq e^{\epsilon} Pr(\mathcal{A}(D') \in S) + \delta,$$

when $\delta = 0$ and $\mathcal{A}$ is ϵ-differentially private.

We will use Gaussian Mechanism [11] and moments accountant [1] to guarantee (ϵ, δ)-DP.

Definition 3.4 (Gaussian Mechanism)**.** Given any function $q : \mathcal{X}^n \to \mathbb{R}^p$, the Gaussian Mechanism is defined as:

$$\mathcal{M}_G(D, q, \epsilon) = q(D) + Y,$$

where Y is drawn from Gaussian Distribution $\mathcal{N}(0, \sigma^2 I_p)$ with $\sigma \geq \frac{\sqrt{2\ln(1.25/\delta)}\Delta_2(q)}{\epsilon}$. Here $\Delta_2(q)$ is the ℓ_2-sensitivity of the function q, i.e. $\Delta_2(q) = \sup_{D \sim D'} ||q(D) - q(D')||_2$. Gaussian Mechanism preservers (ϵ, δ)-differentially private.

The moments accountant proposed in [1] is a method to accumulate the privacy cost which has tighter bound for ϵ and δ. Roughly speaking, when we use the Gaussian Mechanism on the (stochastic) gradient descent, we can save a factor of $\sqrt{\ln(T/\delta)}$ in the asymptotic bound of standard deviation of noise compared with the advanced composition theorem in [12].

Theorem 3.1 ([1])**.** For G-Lipschitz loss function, there exist constants c_1 and c_2 so that given the sampling probability $q = l/n$ and the number of steps T, for any $\epsilon < c_1 q^2 T$, a DP stochastic gradient algorithm with batch size l that injects Gaussian Noise with standard deviation $G\sigma$ to the gradients (Algorithm 1 in [1]), is (ϵ, δ)-differentially private for any $\delta > 0$ if

$$\sigma \geq c_2 \frac{q\sqrt{T\ln(1/\delta)}}{\epsilon}.$$

4 Differentially Private ERM with Convex Loss Function

In this section we will consider ERM with (non)-smooth regularizer[3], i.e.

$$\min_{x \in \mathbb{R}^p} F^r(x, D) = F(x, D) + r(x) = \frac{1}{n}\sum_{i=1}^{n} f(x, z_i) + r(x). \tag{1}$$

The loss function f is convex for every z. We define the proximal operator as

$$\text{prox}_r(y) = \arg\min_{x \in \mathbb{R}^p} \{\frac{1}{2}||x - y||_2^2 + r(x)\},$$

and denote $x_* = \arg\min_{x \in \mathbb{R}^p} F^r(x, D)$.

Algorithm 1 DP-SVRG($F^r, \tilde{x}_0, T, m, \eta, \sigma$)

Input: $f(x, z)$ is G-Lipschitz and L-smooth. $F^r(x, D)$ is μ-strongly convex w.r.t ℓ_2-norm. $\tilde{x}_0$ is the initial point, η is the step size, T, m are the iteration numbers.

for $s = 1, 2, \cdots, T$ **do**
  $\tilde{x} = \tilde{x}_{s-1}$
  $\tilde{v} = \nabla F(\tilde{x})$
  $x_0^s = \tilde{x}$
  for $t = 1, 2, \cdots, m$ **do**
    Pick $i_t^s \in [n]$
    $v_t^s = \nabla f(x_{t-1}^s, z_{i_t^s}) - \nabla f(\tilde{x}, z_{i_t^s}) + \tilde{v} + u_t^s$, where $u_t^s \sim \mathcal{N}(0, \sigma^2 I_p)$
    $x_t^s = \text{prox}_{\eta r}(x_{t-1}^s - \eta v_t^s)$
  end for
  $\tilde{x}_s = \frac{1}{m}\sum_{k=1}^{m} x_k^s$
end for
return $\tilde{x}_T$

[3] All of the algorithms and theorems in this section are applicable to closed convex set $\mathcal{C}$ rather than $\mathbb{R}^p$.

4.1 Strongly convex case

We first consider the case that $F^r(x, D)$ is μ-strongly convex, **Algorithm 1** is based on the Prox-SVRG [33], which is much faster than SGD or GD. We will show that DP-SVRG is also faster than DP-SGD or DP-GD in terms of the time needed to achieve the near optimal excess empirical risk bound.

Definition 4.1 (Strongly Convex)**.** The function $f(x)$ is μ-strongly convex with respect to norm $||\cdot||$ if for any $x, y \in \text{dom}(f)$, there exist $\mu > 0$ such that

$$f(y) \geq f(x) + \langle \partial f, y - x \rangle + \frac{\mu}{2}||y - x||^2, \tag{2}$$

where ∂f is any subgradient on x of f.

Theorem 4.1. In **DP-SVRG**(Algorithm 1), for $\epsilon \leq c_1 \frac{Tm}{n^2}$ with some constant c_1 and $\delta > 0$, it is (ϵ, δ)-differentially private if

$$\sigma^2 = c\frac{G^2 T m \ln(\frac{1}{\delta})}{n^2\epsilon^2} \tag{3}$$

for some constant c.

Remark 4.1. The constraint on ϵ in Theorems 4.1 and 4.3 comes from Theorem 3.1. This constraint can be removed if the noise σ is amplified by a factor of $O(\ln(T/\delta))$ in (3) and (6). But accordingly there will be a factor of $\tilde{O}(\log(Tm/\delta))$ in the utility bound in (5) and (7). In this case the guarantee of differential privacy is by advanced composition theorem and privacy amplification via sampling[6].

Theorem 4.2 (Utility guarantee)**.** Suppose that the loss function $f(x, z)$ is convex, G-Lipschitz and L-smooth over x. $F^r(x, D)$ is μ-strongly convex w.r.t ℓ_2-norm. In **DP-SVRG**(Algorithm 1), let σ be as in (3). If one chooses $\eta = \Theta(\frac{1}{L}) \leq \frac{1}{12L}$ and sufficiently large $m = \Theta(\frac{L}{\mu})$ so that they satisfy inequality

$$\frac{1}{\eta(1 - 8\eta L)\mu m} + \frac{8L\eta(m + 1)}{m(1 - 8L\eta)} < \frac{1}{2}, \tag{4}$$

then the following holds for $T = O\left(\log(\frac{n^2\epsilon^2\mu}{pG^2\ln(1/\delta)})\right)$,

$$\mathbb{E}[F^r(\tilde{x}_T, D)] - F^r(x_*, D) \leq \tilde{O}\left(\frac{p\log(n)G^2\log(1/\delta)}{n^2\epsilon^2\mu}\right), \tag{5}$$

where some insignificant logarithm terms are hiding in the $\tilde{O}$-notation. The total gradient complexity is $O\left((n + \frac{L}{\mu})\log\frac{n\epsilon\mu}{p}\right)$.

Remark 4.2. We can further use some acceleration methods to reduce the gradient complexity, see [25][3].

4.2 Non-strongly convex case

In some cases, $F^r(x, D)$ may not be strongly convex. For such cases, [5] has recently showed that SVRG++ has less gradient complexity than Accelerated Gradient Descent. Following the idea of DP-SVRG, we present the algorithm DP-SVRG++ for the non-strongly convex case. Unlike the previous one, this algorithm can achieve the optimal utility bound.

Theorem 4.3. In **DP-SVRG++**(Algorithm 2), for $\epsilon \leq c_1 \frac{2^T m}{n^2}$ with some constant c_1 and $\delta > 0$, it is (ϵ, δ)-differentially private if

$$\sigma^2 = c\frac{G^2 2^T m \ln(\frac{2}{\delta})}{n^2\epsilon^2} \tag{6}$$

for some constant c.

Theorem 4.4 (Utility guarantee)**.** Suppose that the loss function $f(x, z)$ is convex, G-Lipschitz and L-smooth. In **DP-SVRG++**(Algorithm 2), if σ is chosen as in (6), $\eta = \frac{1}{13L}$, and $m = \Theta(L)$ is sufficiently large, then the following holds for $T = O\left(\log(\frac{n\epsilon}{G\sqrt{p}\sqrt{\log(1/\delta)}})\right)$,

$$\mathbb{E}[F^r(\tilde{x}_T, D)] - F^r(x_*, D) \leq O\left(\frac{G\sqrt{p\ln(1/\delta))}}{n\epsilon}\right). \tag{7}$$

The gradient complexity is $O\left(\frac{nL\epsilon}{\sqrt{p}} + n\log(\frac{n\epsilon}{p})\right)$.

Algorithm 2 DP-SVRG++($F^r, \tilde{x}_0, T, m, \eta, \sigma$)

Input: $f(x, z)$ is G-Lipschitz, and L-smooth over $x \in \mathcal{C}$. $\tilde{x}_0$ is the initial point, η is the step size, and T, m are the iteration numbers.

$x_0^1 = \tilde{x}_0$
for $s = 1, 2, \cdots, T$ **do**
  $\tilde{v} = \nabla F(\tilde{x}_{s-1})$
  $m_s = 2^s m$
  for $t = 1, 2, \cdots, m_s$ **do**
    Pick $i_t^s \in [n]$
    $v_t^s = \nabla f(x_{t-1}^s, z_{i_t^s}) - \nabla f(\tilde{x}_{s-1}, z_{i_t^s}) + \tilde{v} + u_s^t$, where $u_s^t \sim \mathcal{N}(0, \sigma^2 I_p)$
    $x_t^s = \text{prox}_{\eta r}(x_{t-1}^s - \eta v_t^s)$
  end for
  $\tilde{x}_s = \frac{1}{m_s} \sum_{k=1}^{m_s} x_k^s$
  $x_0^{s+1} = x_{m_s}^s$
end for
return $\tilde{x}_T$

5 Differentially Private ERM for Convex Loss Function in High Dimensions

The utility bounds and gradient complexities in Section 4 depend on dimensionality p. In high-dimensional (i.e., $p \gg n$) case, such a dependence is not very desirable. To alleviate this issue, we can usually get rid of the dependence on dimensionality by reformulating the problem so that the goal is to find the parameter in some closed centrally symmetric convex set $\mathcal{C} \subseteq \mathbb{R}^p$ (such as l_1-norm ball), *i.e.,*

$$\min_{x \in \mathcal{C}} F(x, D) = \frac{1}{n} \sum_{i=1}^{n} f(x, z_i), \tag{8}$$

where the loss function is convex.

[28],[29] showed that the $\sqrt{p}$ term in (5),(7) can be replaced by the Gaussian Width of $\mathcal{C}$, which is no larger than $O(\sqrt{p})$ and can be significantly smaller in practice (for more detail and examples one may refer to [28]). In this section, we propose a faster algorithm to achieve the upper utility bound. We first give some definitions.

Algorithm 3 DP-AccMD(F, x_0, T, σ, w)

Input: $f(x, z)$ is G-Lipschitz , and L-smooth over $x \in \mathcal{C}$. $||\mathcal{C}||_2$ is the ℓ_2 norm diameter of the convex set $\mathcal{C}$. w is a function that is 1-strongly convex w.r.t $||\cdot||_{\mathcal{C}}$. x_0 is the initial point, and T is the iteration number.

Define $\mathcal{B}_w(y, x) = w(y) - \langle \nabla w(x), y - x \rangle - w(x)$
$y_0, z_0 = x_0$
for $k = 0, \cdots, T-1$ **do**
  $\alpha_{k+1} = \frac{k+2}{4L}$ and $r_k = \frac{1}{2\alpha_{k+1} L}$
  $x_{k+1} = r_k z_k + (1 - r_k) y_k$
  $y_{k+1} = \arg\min_{y \in \mathcal{C}} \{ \frac{L||\mathcal{C}||_2^2}{2} ||y - x_{k+1}||_{\mathcal{C}}^2 + \langle \nabla F(x_{k+1}), y - x_{k+1} \rangle \}$
  $z_{k+1} = \arg\min_{z \in \mathcal{C}} \{ \mathcal{B}_w(z, z_k) + \alpha_{k+1} \langle \nabla F(x_{k+1}) + b_{k+1}, z - z_k \rangle \}$, where $b_{k+1} \sim \mathcal{N}(0, \sigma^2 I_p)$
end for
return y_T

Definition 5.1 (Minkowski Norm)**.** The Minkowski norm (denoted by $||\cdot||_{\mathcal{C}}$) with respect to a centrally symmetric convex set $\mathcal{C} \subseteq \mathbb{R}^p$ is defined as follows. For any vector $v \in \mathbb{R}^p$,

$$||\cdot||_{\mathcal{C}} = \min\{r \in \mathbb{R}^+ : v \in r\mathcal{C}\}.$$

The dual norm of $||\cdot||_{\mathcal{C}}$ is denoted as $||\cdot||_{\mathcal{C}^*}$, for any vector $v \in \mathbb{R}^p$, $||v||_{\mathcal{C}^*} = \max_{w \in \mathcal{C}} |\langle w, v \rangle|$.

The following lemma implies that for every smooth convex function $f(x, z)$ which is L-smooth with respect to ℓ_2 norm, it is $L||\mathcal{C}||_2^2$-smooth with respect to $||\cdot||_{\mathcal{C}}$ norm.

Lemma 5.1. For any vector v, we have $||v||_2 \leq ||\mathcal{C}||_2||v||_{\mathcal{C}}$, where $||\mathcal{C}||_2$ is the ℓ_2-diameter and $||\mathcal{C}||_2 = \sup_{x,y\in\mathcal{C}} ||x - y||_2$.

Definition 5.2 (Gaussian Width)**.** Let $b \sim \mathcal{N}(0, I_p)$ be a Gaussian random vector in $\mathbb{R}^p$. The Gaussian width for a set $\mathcal{C}$ is defined as $G_{\mathcal{C}} = \mathbb{E}_b[\sup_{w\in\mathcal{C}}\langle b, w\rangle]$.

Lemma 5.2 ([28])**.** For $W = (\max_{w\in\mathcal{C}}\langle w, v\rangle)^2$ where $v \sim \mathcal{N}(0, I_p)$, we have $\mathbb{E}_v[W] = O(G_{\mathcal{C}}^2 + ||\mathcal{C}||_2^2)$.

Our algorithm **DP-AccMD** is based on the Accelerated Mirror Descent method, which was studied in [4],[23].

Theorem 5.3. In **DP-AccMD**(Algorithm 3), for $\epsilon, \delta > 0$, it is (ϵ, δ)-differentially private if

$$\sigma^2 = c\frac{G^2T\ln(1/\delta)}{n^2\epsilon^2} \tag{9}$$

for some constant c.

Theorem 5.4 (Utility Guarantee)**.** Suppose the loss function $f(x, z)$ is G-Lipschitz , and L-smooth over $x \in \mathcal{C}$. In DP-AccMD, let σ be as in (9) and w be a function that is 1-strongly convex with respect to $||\cdot||_{\mathcal{C}}$. Then if

$$T^2 = O\left(\frac{L||\mathcal{C}||_2^2\sqrt{\mathcal{B}_w(x_*, x_0)}n\epsilon}{G\sqrt{\ln(1/\delta)}\sqrt{G_{\mathcal{C}}^2 + ||\mathcal{C}||_2^2}}\right),$$

we have

$$\mathbb{E}[F(y_T, D)] - F(x_*, D) \leq O\left(\frac{\sqrt{\mathcal{B}_w(x_*, x_0)}\sqrt{G_{\mathcal{C}}^2 + ||\mathcal{C}||_2^2}G\sqrt{\ln(1/\delta)}}{n\epsilon}\right).$$

The total gradient complexity is $O\left(\frac{n^{1.5}\sqrt{\epsilon L}}{(G_{\mathcal{C}}^2+||\mathcal{C}||_2^2)^{\frac{1}{4}}}\right)$.

6 ERM for General Functions

In this section, we consider non-convex functions with similar objective function as before,

$$\min_{x\in\mathbb{R}^p} F(x, D) = \frac{1}{n}\sum_{i=1}^{n} f(x, z_i). \tag{10}$$

Algorithm 4 DP-GD($x_0, F, \eta, T, \sigma, D$)

Input: $f(x, z)$ is G-Lipschitz , and L-smooth over $x \in \mathcal{C}$. F is under the assumptions. $0 < \eta \leq \frac{1}{L}$ is the step size. T is the iteration number.

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

$x_t = x_{t-1} - \eta\left(\nabla F(x_{t-1}, D) + z_{t-1}\right)$, where $z_{t-1} \sim \mathcal{N}(0, \sigma^2 I_p)$

end for

return x_T(For section 6.1)

return x_m where m is uniform sampled from $\{0, 1, \cdots, m-1\}$(For section 6.2)

Theorem 6.1. In **DP-GD**(Algorithm 4), for $\epsilon, \delta > 0$, it is (ϵ, δ)-differentially private if

$$\sigma^2 = c\frac{G^2T\ln(1/\delta)}{n^2\epsilon^2} \tag{11}$$

for some constant c.

6.1 Excess empirical risk for functions under Polyak-Lojasiewicz condition

In this section, we consider excess empirical risk in the case where the objective function $F(x, D)$ satisfies Polyak-Lojasiewicz condition. This topic has been studied in [18][27][26][24][22].

Definition 6.1 (Polyak-Lojasiewicz condition)**.** For function $F(\cdot)$, denote $\mathcal{X}^* = \arg\min_{x\in\mathbb{R}^p} F(x)$ and $F^* = \min_{x\in\mathbb{R}^p} F(x)$. Then there exists $\mu > 0$ and for every x,

$$||\nabla F(x)||^2 \geq 2\mu(F(x) - F^*). \tag{12}$$

(12) guarantees that every critical point (*i.e.,* the point where the gradient vanish) is the global minimum. [18] shows that if F is differentiable and L-smooth w.r.t ℓ_2 norm, then we have the following chain of implications:

Strong Convex $\Rightarrow$ Essential Strong Convexity$\Rightarrow$ Weak Strongly Convexity $\Rightarrow$ Restricted Secant Inequality $\Rightarrow$ Polyak-Lojasiewicz Inequality $\Leftrightarrow$ Error Bound

Theorem 6.2. Suppose that $f(x, z)$ is G-Lipschitz, and L-smooth over $x\mathcal{C}$, and $F(x, D)$ satisfies the Polyak-Lojasiewicz condition. In **DP-GD**(Algorithm 4), let σ be as in (11) with $\eta = \frac{1}{L}$. Then if $T = \tilde{O}\left(\log(\frac{n^2\epsilon^2}{pG^2\log(1/\delta)})\right)$, the following holds

$$\mathbb{E}[F(x_T, D)] - F(x_*, D) \leq O(\frac{G^2 p\log^2(n)\log(1/\delta)}{n^2\epsilon^2}), \tag{13}$$

where $\tilde{O}$ hides other $\log, L, \mu$ terms.

DP-GD achieves near optimal bound since strongly convex functions can be seen as a special case in the class of functions satisfying Polyak-Lojasiewicz condition. The lower bound for strongly convex functions is $\Omega(\min\{1, \frac{p}{n^2\epsilon^2}\})$[6]. Our result has only a logarithmic multiplicative term comparing to that. Thus we achieve near optimal bound in this sense.

6.2 Tight upper bound for (non)-convex case

In [34], the authors considered (non)-convex smooth loss functions and measured the utility as $||F(x^{\text{private}}, D)||^2$. They proposed an algorithm with gradient complexity $O(n^2)$. For this algorithm, they showed that $\mathbb{E}[||F(x^{\text{private}}, D)||^2] \leq O(\frac{\log(n)\sqrt{p\log(1/\delta)}}{n\epsilon})$. By using DP-GD(Algorithm 4), we can eliminate the $\log(n)$ term.

Theorem 6.3. Suppose that $f(x, z)$ is G-Lipschitz, and L-smooth. In **DP-GD**(Algorithm 4), let σ be as in (11) with $\eta = \frac{1}{L}$. Then when $T = O(\frac{\sqrt{L}n\epsilon}{\sqrt{p\log(1/\delta)}G})$, we have

$$\mathbb{E}[||\nabla F(x_m, D)||^2] \leq O(\frac{\sqrt{L}G\sqrt{p\log(1/\delta)}}{n\epsilon}). \tag{14}$$

Remark 6.1. Although we can obtain the optimal bound by Theorem 3.1 using DP-SGD, there will be a constraint on ϵ. Also, we still do not know the lower bound of the utility using this measure. We leave it as an open problem.

7 Discussions

From the discussion in previous sections, we know that when gradient perturbation is combined with linearly converge first order methods, near optimal bound with less gradient complexity can be achieved. The remaining issue is whether the optimal bound can be obtained in this way. In Section 6.1, we considered functions satisfying the Polyak-Lojasiewicz condition, and achieved near optimal bound on the utility. It will be interesting to know the bound for functions satisfying other conditions (such as general Gradient-dominated functions [24], quasi-convex and locally-Lipschitz in [16]) under the differential privacy model. For general non-smooth convex loss function (such as SVM), we do not know whether the optimal bound is achievable with less time complexity. Finally, for non-convex loss function, proposing an easier interpretable measure for the utility is another direction for future work.

References

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

[2] N. Agarwal and K. Singh. The price of differential privacy for online learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 32–40, 2017.

[3] Z. Allen-Zhu. Katyusha: the first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1200–1205. ACM, 2017.

[4] Z. Allen-Zhu and L. Orecchia. Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent. In *Proceedings of the 8th Innovations in Theoretical Computer Science*, ITCS '17, 2017.

[5] Z. Allen-Zhu and Y. Yuan. Improved SVRG for Non-Strongly-Convex or Sum-of-Non-Convex Objectives. In *Proceedings of the 33rd International Conference on Machine Learning*, ICML '16, 2016.

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

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

[8] K. Chaudhuri and C. Monteleoni. Privacy-preserving logistic regression. In *Advances in Neural Information Processing Systems*, pages 289–296, 2009.

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

[10] K. Chaudhuri, A. Sarwate, and K. Sinha. Near-optimal differentially private principal components. In *Advances in Neural Information Processing Systems*, pages 989–997, 2012.

[11] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

[12] C. Dwork, G. N. Rothblum, and S. Vadhan. Boosting and differential privacy. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 51–60. IEEE, 2010.

[13] C. Dwork, K. Talwar, A. Thakurta, and L. Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 11–20. ACM, 2014.

[14] D. Feldman, A. Fiat, H. Kaplan, and K. Nissim. Private coresets. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 361–370. ACM, 2009.

[15] M. Hardt and A. Roth. Beyond worst-case analysis in private singular vector computation. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 331–340. ACM, 2013.

[16] E. Hazan, K. Levy, and S. Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. In *Advances in Neural Information Processing Systems*, pages 1594–1602, 2015.

[17] P. Jain, P. Kothari, and A. Thakurta. Differentially private online learning. In *COLT*, volume 23, pages 24–1, 2012.

[18] H. Karimi, J. Nutini, and M. 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.

[19] S. P. Kasiviswanathan and H. Jin. Efficient private empirical risk minimization for high-dimensional learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 488–497, 2016.

[20] S. P. Kasiviswanathan, K. Nissim, and H. Jin. Private incremental regression. In *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2017, Chicago, IL, USA, May 14-19, 2017*, pages 167–182, 2017.

[21] D. Kifer, A. Smith, and A. Thakurta. Private convex empirical risk minimization and high-dimensional regression. *Journal of Machine Learning Research*, 1(41):3–1, 2012.

[22] G. Li and T. K. Pong. Calculus of the exponent of kurdyka-{\ L} ojasiewicz inequality and its applications to linear convergence of first-order methods. *arXiv preprint arXiv:1602.02915*, 2016.

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

[24] Y. Nesterov and B. T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[25] A. Nitanda. Stochastic proximal gradient descent with acceleration techniques. In *Advances in Neural Information Processing Systems*, pages 1574–1582, 2014.

[26] B. T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

[27] S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323, 2016.

[28] K. Talwar, A. Thakurta, and L. Zhang. Private empirical risk minimization beyond the worst case: The effect of the constraint set geometry. *arXiv preprint arXiv:1411.5417*, 2014.

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

[30] A. G. Thakurta and A. Smith. (nearly) optimal algorithms for private online learning in full-information and bandit settings. In *Advances in Neural Information Processing Systems*, pages 2733–2741, 2013.

[31] Y.-X. Wang, J. Lei, and S. E. Fienberg. Learning with differential privacy: Stability, learnability and the sufficiency and necessity of erm principle. *Journal of Machine Learning Research*, 17(183):1–40, 2016.

[32] X. Wu, M. Fredrikson, W. Wu, S. Jha, and J. F. Naughton. Revisiting differentially private regression: Lessons from learning theory and their consequences. *arXiv preprint arXiv:1512.06388*, 2015.

[33] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

[34] J. Zhang, K. Zheng, W. Mou, and L. Wang. Efficient private ERM for smooth objectives. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 3922–3928, 2017.

8 Experiments

In this section, we validate our methods using Covertype dataset[4] and logistic regression. This dataset contains 581012 samples with 54 features. We use 200000 samples for training. We compare our **DP-SVRG** algorithm with the **DP-GD** method in [34] for logistic regression with L_2-norm regularization.

$$F^r(w, D) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(1 + y_i w^T x_i)) + \frac{\lambda}{2}||w||^2,$$

where λ is set to be 10^{-2}.

We also compare our **DP-SVRG++** algorithm with the **DP-GD** method in [34] for logistic regression,

$$F^r(w, D) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(1 + y_i w^T x_i))$$

We evaluate the optimality gap $\mathbb{E}[F^r(w^{\text{private}}, D)] - F^r(w^*, D)$ and the running time for $\epsilon = \{0.2, 0.5, 1\}$ and $\delta = 0.001$.

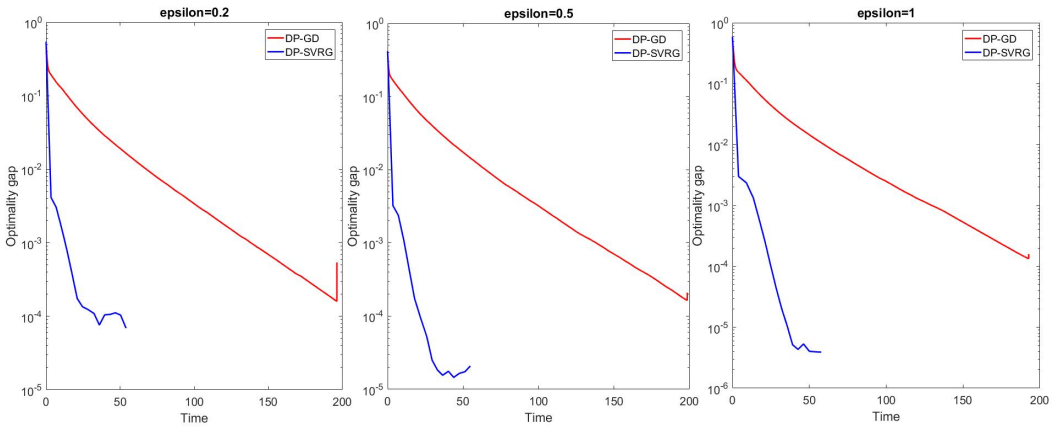

Figure 1: Comparison of DP-SVRG and DP-GD for Logistic regression with different ϵ and L_2-regularization. We set $T = 15, m = 5000$ and use SVRG-BB for step size update in DP-SVRG, $T = 1500$ in DP-GD.

From the figure, it is clear that our method outperform the previous results in both cases.

9 Details and proofs

9.1 Using Advance Composition Theorem to Guarantee (ϵ, δ)-differential private

As we can see that there are constrains on ϵ in Theorem 4.1 and Theorem 4.3. The constrains come from Theorem 3.1 (see the proof below). For general ϵ, we can just amplify a factor of $O(\ln(T/\delta))$ on the σ. However, in this case, we will amplify a factor of $O(\log(Tm/\delta))$ (neglecting other terms) in (5) and (7) in Theorem 4.2 and 4.4; the guarantee of DP is by advanced composition theorem and privacy amplification via sampling [6]. Below we will show this. Consider the i-th query:

$$M_i = \nabla f(x^s_{t-1}, z_{i^s_t}) - \nabla f(\tilde{x}, z_{i^s_t}) + \frac{1}{n}\sum_{i=1}^{n} \nabla f(\tilde{x}, z_i) + \mathcal{N}(0, \sigma^2 I_p),$$

[4] https://archive.ics.uci.edu/ml/datasets/covertype

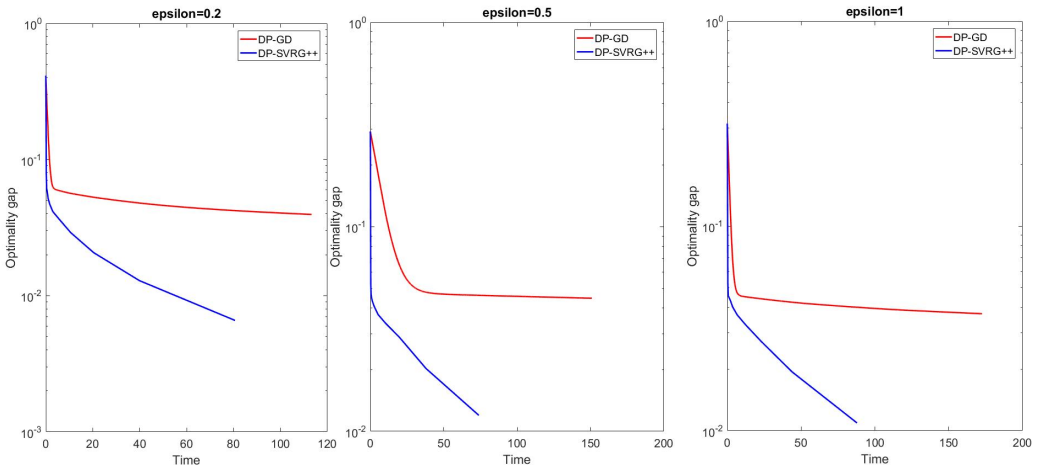

Figure 2: Comparison of DP-SVRG++ and DP-GD for Logistic regression with different ϵ. We set $T = 15, m = 10, \eta = 0.01$ in DP-SVRG++ and $T = 1000, \eta = 0.1$ in DP-GD.

where i_t^s is the uniform sampling. There are T-compositions of these queries. By advanced composition theorem, we know that in order to guarantee the (ϵ, δ)-differential private, we need $(c\frac{\epsilon}{\sqrt{T\log(1/\delta)}}, T/2\delta)$-differential private in each M_i for some constant c. Now consider M_i on the whole dataset (*i.e.,* with no random sample).

$$\tilde{M}_i = \sum_{i=1}^{n} \nabla f(x_{t-1}^s, z_i) - \sum_{i=1}^{n} \nabla f(\tilde{x}, z_i) + \frac{1}{n}\sum_{i=1}^{n} \nabla f(\tilde{x}, z_i) + \mathcal{N}(0, \sigma^2 I_p).$$

From the above, we can see that the L_2-sensitive of $\tilde{M}_i$ is $\Delta \leq 2G + \frac{G}{n} \leq 3G$. Thus if $\sigma^2 \geq c_1 \frac{G^2 \log(1/\delta')}{\epsilon'^2}$ for some c_1, $\tilde{M}_i$ will be $(\epsilon', \delta'))$-differential private. This implies that the query M_i will be $(2\frac{1}{n}\epsilon', \delta')$-differential private, which comes from the following lemma (see Theorem 2.1 and Lemma 2.2 in [6]).

Lemma 9.1. If an algorithm $\mathcal{A}$ is ϵ'-differentially private, then for any n-element dataset D, executing $\mathcal{A}$ on uniformly random γn entries ensures $2\gamma\epsilon'$-differential private.

Let $2\frac{1}{n}\epsilon' = c\frac{\epsilon}{\sqrt{T\log(1/\delta)}}$ and $\delta' = T/2\delta$, that is $\epsilon' = c'\frac{n\epsilon}{\sqrt{T\log(1/\delta)}}$ and

$$\sigma^2 \geq c_2 \frac{GT\log(T/\delta)\log(1/\delta)}{\epsilon^2 n^2}.$$

We can guarantee that T composition of M_i queries is (ϵ, δ)-differential private.

9.2 Proof of Theorem 4.1 and 4.3

Proof. W.l.o.g, we assume $G = 1$, *i.e.,* $||\nabla f|| \leq 1$ (otherwise we can rescale f).The Proof of Theorem 4.1 and Theorem 4.3 are the same instead of the iteration number (or number of queries). Let the difference data of D, D' be the n-th data. Now, consider the i-th query:

$$M_i = \nabla f(x_{t-1}^s, z_{i_t^s}) - \nabla f(\tilde{x}, z_{i_t^s}) + \frac{1}{n}\sum_{i=1}^{n} \nabla f(\tilde{x}, z_i) + u_t^s, u_t^s \sim \mathcal{N}(0, \sigma^2 I_p),$$

where $i_t^s \in [n]$ is a uniform sample. This query can be thought as the composition of two queries:

$$M_{i,1} = \nabla f(x_{t-1}^s, z_{i_t^s}) - \nabla f(\tilde{x}, z_{i_t^s}) + \mathcal{N}(0, \sigma_1^2 I_p) \tag{15}$$

and

$$M_{i,2} = \nabla F(\tilde{x}, D) + \mathcal{N}(0, \sigma_2^2 I_p) = \frac{1}{n}\sum_{i=1}^{n} \nabla f(\tilde{x}, z_i) + \mathcal{N}(0, \sigma_2^2 I_p) \tag{16}$$

for some σ_1, σ_2. By **Theorem 2.1** in [1] we have $\alpha_{M_i}(\lambda) \leq \alpha_{M_{i,1}}(\lambda) + \alpha_{M_{i,2}}(\lambda)$. Now we bound $\alpha_{M_{i,1}}(\lambda)$ and $\alpha_{M_{i,2}}(\lambda)$.

For $\alpha_{M_{i,1}}$, we can use **Lemma 3** in [1] directly, where $q = \frac{1}{n}$, $f(\cdot) = \nabla f(x_{t-1}^s, \cdot) - \nabla f(\tilde{x}, \cdot)$. For some constant c_1 and any integer $\lambda \leq \sigma_1^2 \ln(n/\sigma_1)$, we have

$$\alpha_{M_{i,1}}(\lambda) \leq c_1 \frac{\lambda^2}{n^2\sigma_1^2} + O(\frac{\lambda^3}{n^3\sigma_1^3}). \tag{17}$$

For $\alpha_{M_{i,2}}(\lambda)$, we use the relationship between moment account and Rényi divergence. By Definition 2.1 in [7] we have:

$$\alpha_{M_{i,2}}(\lambda) = \lambda D_{\lambda+1}(P||Q), \tag{18}$$

where $P = \nabla F(\tilde{x}, D) + \mathcal{N}(0, \sigma_2^2 I_p) = \mathcal{N}(\nabla F(\tilde{x}, D), \sigma_2^2)$ and $Q = \nabla F(\tilde{x}, D') + \mathcal{N}(0, \sigma_2^2 I_p) = \mathcal{N}(\nabla F(\tilde{x}, D'), \sigma_2^2)$. By Lemma 2.5 in [7], we have for some c_2:

$$\lambda D_{\lambda+1}(P||Q) = \frac{\lambda(\lambda+1)||\nabla F(\tilde{x}, D) - \nabla F(\tilde{x}, D')||^2}{2\sigma^2} \leq \frac{2\lambda(\lambda+1)}{n^2\sigma_2^2} \leq \frac{c_1\lambda^2}{n^2\sigma_2^2}. \tag{19}$$

Combining (17), (18) and (19), we have

$$\alpha_{M_i}(\lambda) \leq c_1 \frac{\lambda^2}{n^2\sigma_2^2} + c_2 \frac{\lambda^2}{n^2\sigma_1^2} + O(\frac{\lambda^3}{n^3\sigma_1^3}). \tag{20}$$

The rest is similar to the proof of Theorem 3.1.
After T iterations, we have for some c_1, c_2,

$$\alpha_M \leq \sum_{i=1}^{T} \alpha_{M_i} \leq c_1 \frac{\lambda^2}{n^2\sigma_2^2} + c_2 \frac{\lambda^2}{n^2\sigma_1^2}. \tag{21}$$

To be (ϵ, δ)-differential private, by Theorem 2.2 in [1], it suffices that

$$c_1 \frac{T\lambda^2}{n^2\sigma_2^2} + c_2 \frac{T\lambda^2}{n^2\sigma_1^2} \leq \frac{\lambda\epsilon}{2}$$

and

$$\exp(\frac{-\lambda\epsilon}{2}) \leq \delta.$$

In addition we need

$$\lambda \leq \sigma_1^2 \ln(n/\sigma_1). \tag{22}$$

It can be verified that when $\epsilon \leq c_3 \frac{T}{n^2}$ for some constant c_3, we have

$$\sigma_1 = c_4 \frac{\sqrt{T\log(1/\delta)}}{n\epsilon} \tag{23}$$

and

$$\sigma_2 = c_5 \frac{\sqrt{T\log(1/\delta)}}{n\epsilon}. \tag{24}$$

For some constant c_4, c_5, all the conditions can be satisfied. Since the sum of two Gaussian distributions is still a Gaussian distribution, and $M_i = M_{i,1} + M_{i,2}$, we have $\sigma = c\frac{\sqrt{T\log(1/\delta)}}{n\epsilon}$ for some c. Thus, T-fold of the queries.

$$M_i = \nabla f(x_{t-1}^s, z_{i_t^s}) - \nabla f(\tilde{x}, z_{i_t^s}) + \frac{1}{n}\sum_{i=1}^{n} \nabla f(\tilde{x}, z_i) + \mathcal{N}(0, \sigma^2 I_p)$$

will guarantee (ϵ, δ)-differential private when $\epsilon \leq c_3 \frac{T}{n^2}$.
For Theorem 4.1 $T = Tm$ while for Theorem 4.3 $T = 2^{T+1}m$. □

9.3 Proof of Theorem 5.3 and Theorem 6.1

Proof. The proof is similar to the above.

$$M_i = \nabla F(\tilde{x}, D) + \mathcal{N}(0, \sigma^2 I_p) = \frac{1}{n}\sum_{i=1}^{n} \nabla f(\tilde{x}, z_i) + \mathcal{N}(0, \sigma^2 I_p). \tag{25}$$

By (17) and (18), we have

$$\alpha_{M_i}(\lambda) \leq \frac{2\lambda(\lambda+1)}{n^2\sigma^2}. \tag{26}$$

Thus, after T-iterations, we have for some c

$$\alpha_M \leq \sum_{i=1}^{T} \alpha_{M_i} \leq c\frac{T\lambda^2}{n^2\sigma^2}. \tag{27}$$

Taking $\sigma = c_1 \frac{\sqrt{T\log(1/\delta)}}{n\epsilon}$ for some constant c_1, we can guarantee that

$$c\frac{T\lambda^2}{n^2\sigma^2} \leq \frac{\lambda\epsilon}{2}$$

and

$$\exp(\frac{-\lambda\epsilon}{2}) \leq \delta,$$

which means (ϵ, δ)-differential privacy due to Theorem 2.2 in [1]. □

9.4 Proof of Theorem 4.2

Proof. Let $g_t^s = \frac{1}{\eta}(x_{t-1}^s - \text{prox}_{\eta r}(x_{t-1}^s - \eta v_t^s))$. Then we have $x_t^s = x_{k-1}^s - \eta g_t^s$. Thus

$$||x_t^s - x_*||^2 = ||x_{t-1}^s - \eta g_t^s - x_*||^2 = ||x_{t-1}^s - x_*||^2 - 2\eta\langle g_t^s, x_{t-1}^s - x_*\rangle + \eta^2||g_t^s||^2. \tag{28}$$

By Lemma 3 in [33], we have the following inequality

$$-\langle g_t^s, x_{t-1}^s - x_*\rangle + \frac{\eta}{2}||g_t^s||^2 \leq F^r(x_*) - F^r(x_t^s) - \frac{\mu_F}{2}||x_{t-1}^s - x_*||^2 - \frac{\mu_r}{2}||x_t^s - x_*||^2 - \langle v_t^s - \nabla F(x_{t-1}^s), x_t^s - x^*\rangle. \tag{29}$$

Plugging (29) into (28), we have

$$||x_t^s - x_*||^2 \leq ||x_{t-1}^s - x_*||^2 - 2\eta[F^r(x_t^s) - F^r(x_*)] - 2\eta\langle v_t^s - \nabla F(x_{t-1}^s), x_t^s - x^*\rangle. \tag{30}$$

Next we bound $-2\eta\langle v_t^s - \nabla F(x_{t-1}^s), x_t^s - x^*\rangle$. Denote $\hat{x_t}^s = \text{prox}_{\eta r}(x_{t-1}^s - \eta\nabla F(x_{t-1}^s))$.

$$\begin{aligned}
&-2\eta\langle v_t^s - \nabla F(x_{t-1}^s), x_t^s - x^*\rangle = \\
&-2\eta\langle v_t^s - \nabla F(x_{t-1}^s), x_t^s - \hat{x_t^s}\rangle - 2\eta\langle v_t^s - \nabla F(x_{t-1}^s), \hat{x_t}^s - x_*\rangle && (31)\\
&\leq 2\eta||v_t^s - \nabla F(x_{t-1}^s)||||x_t^s - \hat{x_t^s}|| - 2\eta\langle v_t^s - \nabla F(x_{t-1}^s), \hat{x_t}^s - x_*\rangle && (32)\\
&\leq 2\eta||v_t^s - \nabla F(x_{t-1}^s)||||x_{t-1}^s - \eta v_t^s - (x_{t-1}^s - \nabla F(x_{t-1}^s)|| - 2\eta\langle v_t^s - \nabla F(x_{t-1}^s), \hat{x_t}^s - x_*\rangle && (33)\\
&\leq 2\eta^2||v_t^s - \nabla F(x_{t-1}^s)||^2 - 2\eta\langle v_t^s - \nabla F(x_{t-1}^s), \hat{x_t}^s - x_*\rangle && (34)
\end{aligned}$$

The first inequality is due to the following lemma,

Lemma 9.2. Let r be a closed convex function on $\mathbb{R}^p$. Then for any $x, y \in \text{dom}(R)$

$$||\text{prox}_r(x) - \text{prox}_r(y)|| \leq ||x - y||.$$

We can easily get $\mathbb{E}_{u_t^s, i_t^s}(v_t^s - \nabla F(x_{t-1}^s) = 0$ since u_t^s is independent with v_{t-1}^s. Also by Lemma 1 in [33] and $\mathbb{E}[||a+b||^2] \leq 2\mathbb{E}||a||^2 + 2\mathbb{E}||b||^2$, we have

$$\mathbb{E}_{i_t^s, u_t^s}||v_t^s - \nabla F(x_{t-1}^s)||^2 \leq 8L[F^r(x_{t-1}^s) - F^r(x_*) + F^r(\tilde{x}) - F^r(x_*)] + 2\sigma^2 p. \tag{35}$$

Plugging (20) into (30) and taking the expectation with i_t^s, u_t^s, we have

$$\mathbb{E}||x_t^s - x_*||^2 \leq ||x_{t-1}^s - x_*||^2 - 2\eta[\mathbb{E}(F^r(x_t^s) - F^r(x_*)] + 16\eta^2 L[F^r(x_{t-1}^s) - F^r(x_*) + F^r(\tilde{x}) - F^r(x_*)] + 4\eta^2\sigma^2 p. \quad (36)$$

Summing over $t = 1, 2, \cdots, m$ and taking the expectation, we have

$$\mathbb{E}[||x_m^s - x_*||^2] + 2\eta(1 - 8\eta L)\sum_{t=1}^{m}[\mathbb{E}(F^r(x_t^s)) - F^r(x_*)] \quad (37)$$

$$\leq ||\tilde{x} - x_*||^2 + 16L\eta^2(m+1)[F^r(\tilde{x}) - F^r(x_*)] + 4m\eta^2\sigma^2 p. \quad (38)$$

Since F^r is μ strongly convex, we have $||\tilde{x} - x_*||^2 \leq \frac{2}{\mu}(F^r(\tilde{x}) - F^r(x_*))$. Dividing $2m\eta(1 - 8L\eta)$ from both sides, we get

$$\mathbb{E}[F^r(\tilde{x}^s)] - F^r(x_*) \leq (\frac{1}{\eta(1-8\eta L)\mu m} + \frac{8L\eta(m+1)}{m(1-8L\eta)})(\mathbb{E}[F^r(\tilde{x}_{s-1})] - F^r(x_*)) + \frac{2\eta}{1-8L\eta}\sigma^2 p. \quad (39)$$

Thus we can choose $\eta = \Theta(\frac{1}{L}) < \frac{1}{12L}$ and $m = \Theta(\frac{L}{\mu})$ to make

$$A = \frac{1}{\eta(1-8\eta L)\mu m} + \frac{8L\eta(m+1)}{m(1-8L\eta)} < \frac{1}{2}$$

and $\frac{2\eta}{1-8L\eta} < \frac{1}{2L}$. By (39) and summing over $s = 1, 2 \cdots, T$ we can get

$$\mathbb{E}[F^r(\tilde{x}^T)] - F^r(x_*) \quad (40)$$

$$\leq A^T[F^r(x_0) - F^r(x_*)] + \frac{\sigma^2 p}{L} \quad (41)$$

$$= A^s[F^r(x_0) - F^r(x_*)] + O(\frac{pG^2 Tm\ln(1/\delta)}{n^2\epsilon^2 L}) \quad (42)$$

$$= A^T[F^r(x_0) - F^r(x_*)] + O(\frac{pG^2 T\ln(1/\delta)}{n^2\epsilon^2\mu}). \quad (43)$$

Thus if we take T such that $A^T[F^r(x_0) - F^r(x_*)] = O(\frac{pG^2\ln(1/\delta)}{n^2\epsilon^2\mu})$, *i.e.,*

$$T = O\left(\log(\frac{n^2\epsilon^2\mu}{pG^2\ln(1/\delta)})\right).$$

We have

$$\mathbb{E}[F^r(\tilde{x}^T)] - F^r(x_*) \leq O(\frac{pG^2\ln(n\epsilon\mu/pG)\ln(1/\delta)}{n^2\epsilon^2\mu}).$$

where the big-O notation omitted the other $\ln$ term. □

9.5 Proof of Theorem 4.4

Proof.

$$\mathbb{E}_{i_t^s, u_t^s}[F^r(x_t^s) - F^r(x_*)] = \mathbb{E}_{i_t^s, u_t^s}[F(x_t^s) - F(x_*) + r(x_t^s) - r(x_*)] \quad (44)$$

$$\leq \mathbb{E}_{i_t^s, u_t^s}[F(x_{t-1}^s) + \langle\nabla F(x_{t-1}^s), x_t^s - x_{t-1}^s\rangle + \frac{L}{2}||x_t^s - x_{t-1}^s||^2 - F(x_*) + r(x_t^s) - r(x_*)] \quad (45)$$

$$\leq \mathbb{E}_{i_t^s, u_t^s}[\langle\nabla F(x_{t-1}^s), x_{t-1}^s - x_*\rangle] + \langle\nabla F(x_{t-1}^s), x_t^s - x_{t-1}^s\rangle + \frac{L}{2}||x_t^s - x_{t-1}^s||^2 + r(x_t^s) - r(x_*)] \quad (46)$$

$$= \mathbb{E}_{i_t^s, u_t^s}[\langle v_t^s, x_{t-1}^s - x_*\rangle] + \langle\nabla F(x_{t-1}^s), x_t^s - x_{t-1}^s\rangle + \frac{L}{2}||x_t^s - x_{t-1}^s||^2 + r(x_t^s) - r(x_*)]. \quad (47)$$

The last equality is due to the fact that $\mathbb{E}_{i_t^s,u_t^s}[v_t^s] = \nabla F(x_{t-1}^s)$. Since we have ([5])

$$\langle v_t^s, x_{t-1}^s - x_* \rangle + r(x_t^s) - r(x_*) \leq \langle v_t^s, x_{t-1}^s - x_t^s \rangle + \frac{||x_{t-1}^s - x_*||^2}{2\eta} - \frac{||x_t^s - x_*||^2)}{2\eta} - \frac{||x_t^s - x_{t-1}^s||^2}{2\eta}. \tag{48}$$

Plugging (48) into (33), we have

$$\begin{aligned} LHS \leq & \mathbb{E}_{i_t^s,u_t^s}[\langle v_t^s - \nabla F(x_{t-1}^s), x_{t-1}^s - x_t^s \rangle - \frac{1-\eta L}{2\eta}||x_t^s - x_{t-1}^s||^2 \\ & + \frac{||x_{t-1}^s - x_*||^2 - ||x_t^s - x_*||^2}{2\eta}] \qquad (49) \\ & \leq \mathbb{E}_{i_t^s,u_t^s} \frac{\eta}{2(1-\eta L)}||v_t^s - \nabla F(x_{t-1}^s)||^2 + \frac{||x_{t-1}^s - x_*||^2 - \mathbb{E}_{i_t^s,u_t^s}[||x_t^s - x_*||^2]}{2\eta} \qquad (50) \\ & \leq \frac{4\eta L}{1-\eta L}[F^r(x_{t-1}^s) - F^r(x_*) + F^r(\tilde{x}_{s-1}) - F^r(x_*)] + \frac{\eta}{1-\eta L} p\sigma^2 \\ & + \frac{||x_{t-1}^s - x_*||^2 - \mathbb{E}_{i_t^s,u_t^s}[||x_t^s - x_*||^2]}{2\eta}. \qquad (51) \end{aligned}$$

Choosing $\eta = \frac{1}{13L}$, summing over $t = 1, \cdots, m_s$, dividing m_s, and taking expectation, we have

$$\begin{aligned} \mathbb{E}[\frac{1}{m_s}\sum_{t=1}^{m_s} F^r(x_t^s) - F^r(x_*)] \leq \frac{1}{3}\mathbb{E}[\frac{1}{m_s}\sum_{t=0}^{m_s-1}[F^r(x_t^s) - F^r(x_*) + F^r(\tilde{x}_{s-1}) - F^r(x_*)] + \\ \frac{||x_0^s - x_*||^2 - \mathbb{E}[||x_{m_s}^s - x_*||^2]}{2\eta m_s} + \frac{1}{12L}\sigma^2 p. \end{aligned} \tag{52}$$

By the definitions of x_0^{s+1} and $\tilde{x}_s$, we have

$$\begin{aligned} 2\mathbb{E}[F^r(\tilde{x}_s) - F^r(x_*)] \leq \mathbb{E}[\frac{F^r(x_0^s) - F^r(x_*) - (F^r(x_0^{s+1}) - F(x_*))}{m_s} + \\ F^r(\tilde{x}_{s-1}) - F^r(x_*) + \frac{||x_0^s - x_*||^2 - ||x_0^{s+1} - x_*||^2}{2\eta/3m_s}] + \frac{1}{4L}\sigma^2 p, \end{aligned} \tag{53}$$

which implies that

$$\begin{aligned} & 2(\mathbb{E}[F^r(\tilde{x}_s) - F^r(x_*) + \frac{||x_0^{s+1} - x_*||^2}{4\eta/3m_s} + \frac{F^r(x_0^{s+1}) - F^r(x_*)}{2m_s}]) \qquad (54) \\ & \leq \mathbb{E}[F^r(\tilde{x}_{s-1}) - F^r(x_*) + \frac{||x_0^s - x_*||^2}{4\eta/3m_{s-1}} + \frac{F^r(x_0^s) - F^r(x_*)}{2m_{s-1}}] + \frac{1}{4L}\sigma^2 p. \qquad (55) \end{aligned}$$

Summing over $s = 1, \cdots, T$, we get

$$\begin{aligned} & \mathbb{E}[F^r(\tilde{x}_T) - F^r(x_*)] \qquad (56) \\ & \leq \frac{F^r(\tilde{x}_0) - F^r(x_*)}{2^{T-1}} + \frac{||\tilde{x}_0 - x_*||^2}{2^T 4\eta/3m} + \frac{1}{4L}\sigma^2 p. \qquad (57) \end{aligned}$$

Thus, if we take $m = \Theta(L)$ to make $A = 2F^r(\tilde{x}_0) - F^r(x_*) + \frac{||\tilde{x}_0 - x_*||^2}{4\eta/3m}$ independent of T, n, p, σ, L, plug σ into (43) we have

$$\mathbb{E}[F^r(\tilde{x}_T)] - F^r(x_*) \leq \frac{A}{2^T} + O(\frac{G^2 p 2^T m \ln 2/\delta}{n^2\epsilon^2 L}) = \frac{A}{2^T} + O(\frac{G^2 p 2^T \ln(1/\delta)}{n^2\epsilon^2}). \tag{58}$$

Let $T = O(\log(\frac{n\epsilon}{G\sqrt{p}\sqrt{1/\delta}}))$. We have

$$\mathbb{E}[F^r(\tilde{x}_s)] - F^r(x_*) \leq O(\frac{G\sqrt{p\ln(1/\delta)})}{n\epsilon}).$$

The gradient complexity is $O(2^s m + Tn) = O(\frac{nL\epsilon}{G\sqrt{p}} + n\log(\frac{n\epsilon}{G\sqrt{p}}))$. □

9.6 Proof of lemma 5.1

Proof. If $v = 0$, this is true. If not, we will show that $\frac{||v||_2}{||\mathcal{C}||_2} \leq ||v||_{\mathcal{C}}$. This is equivalent to show that $v \notin \frac{||v||_2}{||\mathcal{C}||_2}\mathcal{C}$. Take any $y \in \mathcal{C}$. Since $||\frac{||v||_2}{||\mathcal{C}||_2}y||_2 = \frac{||v||_2}{||\mathcal{C}||_2}||y||_2$, we know that $||y||_2 < ||\mathcal{C}||_2$. Thus $||\frac{||v||_2}{||\mathcal{C}||_2}y||_2 < ||v||_2$. We have $v \notin \frac{||v||_2}{||\mathcal{C}||_2}\mathcal{C}$. □

9.7 Proof of Theorem 5.4

Proof. We use $||\cdot||$ and $||\cdot||_*$ instead of $||\cdot||_{\mathcal{C}}$ and $||\cdot||_{\mathcal{C}^*}$. Also, w.l.o.g we assume that $||\mathcal{C}||_2 = 1$ (for the general case, just replace L by $L||\mathcal{C}||_2^2$). Since b_{k+1} is independent of x_{k+1}, we have for any u

$$\mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1}\nabla F(x_{k+1}), z_k - u\rangle] = \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1}(\nabla F(x_{k+1}) + b_{k+1}), z_k - u\rangle]$$
$$= \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1}(\nabla F(x_{k+1}) + b_{k+1}), z_k - z_{k+1}\rangle] + \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1}(\nabla F(x_{k+1}) + b_{k+1}), z_{k+1} - u\rangle]. \tag{59}$$

Since $z_{k+1} = \arg\min_{z\in\mathcal{C}}\{\mathcal{B}_w(z, z_k) + \alpha_{k+1}\langle \nabla F(x_{k+1}) + b_{k+1}, z - z_k\rangle\}$, which implies that $\langle \nabla\mathcal{B}_w(z_{k+1}, z_k) + \alpha_{k+1}(\nabla F(x_{k+1} + b_{k+1}), u - z_{k+1}\rangle \geq 0$ for every $u \in \mathcal{C}$. So we can get

$$\mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1}(\nabla F(x_{k+1}) + b_{k+1}), z_{k+1} - u\rangle] \tag{60}$$
$$\leq \mathbb{E}_{b_{k+1}}[\langle -\nabla\mathcal{B}_w(z_{k+1}, z_k), z_{k+1} - u\rangle] = \mathbb{E}_{b_{k+1}}[\mathcal{B}_w(u, z_k) - \mathcal{B}_w(u, z_{k+1}) - \mathcal{B}_w(z_{k+1}, z_k)], \tag{61}$$

where the equality is due to the triangle equality of Bregman divergence. Since w is 1-strong convex with respect to $||\cdot||$, we have $-\mathcal{B}_w(z_{k+1}, z_k) \leq -\frac{1}{2}||z_{k+1} - z_k||^2$. Plugging this into (44), we have

$$\mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1}\nabla F(x_{k+1}), z_k - u\rangle] \tag{62}$$
$$\leq \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1}(\nabla F(x_{k+1}) + b_{k+1}), z_k - z_{k+1}\rangle - \frac{1}{2}||z_{k+1} - z_k||^2] +$$
$$\mathcal{B}_w(u, z_k) - \mathbb{E}_{b_{k+1}}[\mathcal{B}_w(u, z_{k+1})] \tag{63}$$
$$\leq \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1}\nabla F(x_{k+1}), z_k - z_{k+1}\rangle - \frac{1}{4}||z_{k+1} - z_k||^2] + \alpha_{k+1}^2\mathbb{E}_{b_{k+1}}[||b_{k+1}||_*^2] \tag{64}$$
$$+ \mathcal{B}_w(u, z_k) - \mathbb{E}_{b_{k+1}}[\mathcal{B}_w(u, z_{k+1})]. \tag{65}$$

The last inequality is due to Cauchy-Shwartz Inequality. Thus we have $\langle \alpha_{k+1}b_{k+1}, z_k - z_{k+1}\rangle \leq \alpha_{k+1}^2||b_{k+1}||_*^2 + \frac{1}{4}||z_k - z_{k+1}||^2$. Now we want to bound $\mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1}\nabla F(x_{k+1}), z_k - z_{k+1}\rangle - \frac{1}{4}||z_{k+1} - z_k||^2]$. Define $v = r_k z_{k+1} + (1 - r_k)y_k \in \mathcal{C}$ so that $x_{k+1} - v = r_k(z_k - z_{k+1})$. We have

$$\langle \alpha_{k+1}\nabla F(x_{k+1}), z_k - z_{k+1}\rangle - \frac{1}{4}||z_{k+1} - z_k||^2 = \langle \frac{\alpha_{k+1}}{r_k}\nabla F(x_{k+1}), x_{k+1} - v\rangle$$
$$- \frac{1}{4r_k^2}||x_{k+1} - v||^2 \tag{66}$$
$$= 2\alpha_{k+1}^2 L(\langle F(x_{k+1}), x_{k+1} - v\rangle - \frac{L}{2}||x_{k+1} - v||^2) \tag{67}$$
$$\leq 2\alpha_{k+1}^2 L(-\min_{y\in\mathcal{C}}\{\frac{L}{2}||y - x_{k+1}||^2 + \langle F(x_{k+1}), y - x_{k+1}\rangle\}) \tag{68}$$
$$= 2\alpha_{k+1}^2 L(-\{\frac{L}{2}||y_{k+1} - x_{k+1}||^2 + \langle F(x_{k+1}), y_{k+1} - x_{k+1}\rangle\}) \tag{69}$$
$$\leq 2\alpha_{k+1}^2 L(F(x_{k+1}) - F(y_{k+1})). \tag{70}$$

The last inequality is due to the fact that F is $L||\mathcal{C}||_2^2$-smooth (note that $||\mathcal{C}||_2 = 1$) in $||\cdot||$ norm and the definition of y_{k+1}. Thus, we get the following

$$\mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1}\nabla F(x_{k+1}), z_k - u\rangle] = \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1}(\nabla F(x_{k+1}) + b_{k+1}), z_k - u\rangle]$$
$$\leq 2\alpha_{k+1}^2 L(F(x_{k+1}) - F(y_{k+1})) + \mathcal{B}_w(u, z_k) - \mathbb{E}_{b_{k+1}}[\mathcal{B}_w(u, z_{k+1})] + \alpha_{k+1}^2\mathbb{E}_{b_{k+1}}||b_{k+1}||_*^2. \tag{71}$$

By using the Concentration of Gaussian Width, Lemma 3.3 in [28] shows that $\mathbb{E}_{b_{k+1}}||b_{k+1}||_*^2 = \sigma^2 O(G_{\mathcal{C}}^2 + ||\mathcal{C}||_2^2)$, where $G_{\mathcal{C}}$ is the Gaussian Width of $\mathcal{C}$. From this, we have

$$\begin{aligned}
&E_{b_{k+1}}[\alpha_{k+1}(F(x_{k+1}) - F(u)] \leq \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1}\nabla F(x_{k+1}), x_{k+1} - u\rangle] \\
&= \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1}\nabla F(x_{k+1}), x_{k+1} - z_k\rangle] + \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1}\nabla F(x_{k+1}), z_k - u\rangle] \\
&\leq \frac{\alpha_{k+1}(1-r_k)}{r_k}\langle \nabla F(x_{k+1}), y_k - x_{k+1}\rangle + \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1}\nabla F(x_{k+1}), z_k - u\rangle] \\
&\leq \frac{\alpha_{k+1}(1-r_k)}{r_k}(F(y_k) - F(x_{k+1}) + \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1}\nabla F(x_{k+1}), z_k - u\rangle] \\
&\leq (2\alpha_{k+1}^2 L - \alpha_{k+1})(F(y_k) - F(x_{k+1}) + 2\alpha_{k+1}^2 L(F(x_{k+1}) - F(y_{k+1})) \\
&+ \mathcal{B}_w(u, z_k) - \mathbb{E}_{b_{k+1}}[\mathcal{B}_w(u, z_{k+1})] + \alpha_{k+1}^2 \mathbb{E}_{b_{k+1}}||b_{k+1}||_*^2.
\end{aligned}$$

Thus we obtain

$$2\alpha_{k+1}^2 L F(y_{k+1}) - (2\alpha_{k+1}^2 L - \alpha_{k+1})F(y_k) + \mathbb{E}(\mathcal{B}_w(u, z_{k+1}) - \mathcal{B}_w(u, z_k)) \tag{72}$$

$$\leq \alpha_{k+1}F(u) + \alpha_{k+1}^2\sigma^2 O(G_{\mathcal{C}}^2 + ||\mathcal{C}||_2^2). \tag{73}$$

By the definition of α_{k+1}, we have $2\alpha_k^2 L = 2\alpha_{k+1}^2 L - \alpha_{k+1} + \frac{1}{8L}$. Summing over $k = 0 \cdots, T-1$ and setting $u = x_*$, by the definition of α_k we have $\sum_{k=1}^T \alpha_k^2 = O(T^3)$. After taking the expectation we get

$$2\alpha_T^2 L\mathbb{E}[F(y_T)] + \frac{1}{8L}\mathbb{E}[\sum_{k=1}^{T-1} F(y_k)] + \mathbb{E}[\mathcal{B}_w(x_*, z_{T-1})] - \mathcal{B}_w(x_*, z_0) \tag{74}$$

$$\leq \sum_{k=1}^T \alpha_k F(x_*) + O(T^3\sigma^2(G_{\mathcal{C}}^2 + ||\mathcal{C}||_2^2)/L^2). \tag{75}$$

Plugging $\alpha_k = \frac{k+1}{4L}$ into (59), (60) and dividing both sides by a factor of $2\alpha_T^2 L$, by the fact that $\mathcal{B}_w \geq 0$ we finally get

$$\mathbb{E}[F(y_T)] - F[x_*] \leq \frac{8L\mathcal{B}_w(x_*, x_0)}{(T+1)^2} + O(T\sigma^2(G_{\mathcal{C}}^2 + ||\mathcal{C}||_2^2)/L). \tag{76}$$

Since $\sigma^2 = O(\frac{G^2 T\ln(1/\delta)}{n^2\epsilon^2})$, if choose

$$T^2 = O(\frac{L\sqrt{\mathcal{B}_w(x_*, x_0)}n\epsilon}{G\sqrt{\ln(1/\delta)}\sqrt{G_{\mathcal{C}}^2 + ||\mathcal{C}||_2^2}}), \tag{77}$$

we have the bound

$$\mathbb{E}[F(y_T)] - F(x_*) \leq O(\frac{\sqrt{\mathcal{B}_w(x_*, x_0)}\sqrt{G_{\mathcal{C}}^2 + ||\mathcal{C}||_2^2}G\sqrt{\ln(1/\delta)}}{n\epsilon}).$$

□

9.8 Proof of Theorem 6.2

Proof. First of all, we have

$$\mathbb{E}_{z_k}[F(x_{k+1}) - F(x_k)] \leq \mathbb{E}_{z_k}[-\frac{1}{L}\langle \nabla F(x_k), \nabla F(x_k) + z_k\rangle + \frac{1}{2L}||\nabla F(x_k) + z_k||^2] \tag{78}$$

$$= -\frac{1}{2L}||\nabla F(x_k)||^2 + \frac{1}{2L}\mathbb{E}_{z_k}||z_k||^2 \tag{79}$$

$$\leq -\frac{\mu}{L}(F(x_k) - F^*) + \frac{p\sigma^2}{2L}. \tag{80}$$

Re-arranging the terms, we get

$$\mathbb{E}[F(x_{k+1})] - F^* \leq (1 - \frac{\mu}{L})(F(x_k) - F^*) + \frac{p\sigma^2}{2L}.$$

Summing over $k = 0, \cdots, T$ and taking expectation, we obtain

$$\mathbb{E}[F(x_T)] - F^* \leq (1 - \frac{\mu}{L})^T (F(x_0) - F*) + \frac{Tp\sigma^2}{2L}. \tag{81}$$

Thus, when $T = O(\log(\frac{n^2\epsilon^2}{pG^2\log(1/\delta)}))$

$$\mathbb{E}[F(x_T)] - F^* \leq O(\frac{\log^2(n)pG^2\log(1/\delta)}{n^2\epsilon^2}), \tag{82}$$

where the big-O notation neglects other $\log, L, \mu$ terms. □

9.9 Proof of Theorem 6.3

Proof. The proof is similar to that of Theorem 6.2. Let $F^* = \min_{x\in\mathbb{R}^p} F(x, D)$. We have

$$\mathbb{E}_{z_k} F(x_{k+1}) - F(x_k) \leq \mathbb{E}_{z_k}[-\frac{1}{L}\langle \nabla F(x_k), \nabla F(x_k) + z_k\rangle] + \frac{1}{2L}\mathbb{E}_{z_k}||\nabla F(x_k) + z_k||^2 \tag{83}$$

$$\leq -\frac{1}{2L}||\nabla F(x_k)||^2 + \frac{p\sigma^2}{2L}. \tag{84}$$

From this, we get

$$\frac{1}{2L}||\nabla F(x_k)||^2 \leq F(x_k) - E_{z_k} F(x_{k+1}) + \frac{p\sigma^2}{2L}. \tag{85}$$

Thus, $\mathbb{E}_{m,\{z_i\}}[\|\nabla F(x_m)\|^2] = \frac{1}{T}\sum_{i=0}^{T-1}\mathbb{E}_{\{z_i\}}[\|\nabla F(x_i)\|^2]$. By (85), summing over $k = 0, \cdots T-1$, we obtain

$$\mathbb{E}_{m,\{z_i\}}[\|\nabla F(x_m)\|^2] \leq \frac{2L(F(x_0) - \mathbb{E}[F(x_T)])}{T}] + p\sigma^2 \tag{86}$$

$$\leq \frac{2L(F(x_0) - F*)}{T} + O(\frac{pG^2\log(1/\delta)T}{n^2\epsilon^2}). \tag{87}$$

Thus, if choose $T = O(\frac{\sqrt{L}n\epsilon}{\sqrt{p\log(1/\delta)}G})$, we have $\mathbb{E}[||\nabla F(x_m)||^2] \leq O(\frac{\sqrt{L}G\sqrt{p\log(1/\delta)}}{n\epsilon})$. □