

Differentially Private Empirical Risk Minimization in Non-interactive Local Model via Polynomial of Inner Product Approximation

Di Wang

DWANG45@BUFFALO.EDU

*Department of Computer Science and Engineering
State University of New York at Buffalo
Buffalo, NY 14260, USA*

Adam Smith

ADS22@BU.EDU

*Department of Computer Science
Boston University
Boston, MA 02215, USA*

Jinhui Xu

JINHUI@BUFFALO.EDU

*Department of Computer Science and Engineering
State University of New York at Buffalo
Buffalo, NY 14260, USA*

Editor:

Abstract

In this paper, we study the Empirical Risk Minimization problem in the non-interactive Local Differential Privacy (LDP) model. First, we show that for the hinge loss function, there is an (ϵ, δ) -LDP algorithm whose sample complexity for achieving an error of α is only linear in the dimensionality p and quasi-polynomial in other terms. Then, we extend the result to any 1-Lipschitz generalized linear convex loss functions by showing that every such function can be approximated by a linear combination of hinge loss functions and some linear functions. Finally, we apply our technique to the Euclidean median problem and show that its sample complexity needs only to be quasi-polynomial in p , which is the first result with a sub-exponential sample complexity in p for non-generalized linear loss functions. Our results are based on a technique, called polynomial of inner product approximation, which may be applicable to other problems.

1. Introduction

In the big data era, a tremendous amount of individual data are generated every day. Such data, if properly used, could greatly improve many aspects of our daily lives. However, due to the sensitive nature of such data, a great deal of care needs to be taken while analyzing them. Private data analysis seeks to enable the benefits of learning from data with the guarantee of privacy-preservation. Differential privacy (Dwork et al., 2006) has emerged as a rigorous notion for privacy which allows accurate data analysis with a guaranteed bound on the increase in harm for each individual to contribute her data. Methods to guarantee differential privacy have been widely studied, and recently adopted in industry (Near, 2018; Erlingsson et al., 2014).

Two main user models have emerged for differential privacy: the central model and the local one. In the central model, data are managed by a trusted central entity which is responsible for collecting them and for deciding which differentially private data analysis to perform and to release. A classical use case for this model is the one of census data (Haney et al., 2017). In the local model instead, each individual manages his/her proper data and discloses them to a server through some differentially private mechanisms. The server collects the (now private) data of each individual and combines them into a resulting data analysis. A classical use case for this model is the one aiming at collecting statistics from user devices like in the case of Google’s Chrome browser (Erlingsson et al., 2014), and Apple’s iOS-10 (Near, 2018; Tang et al., 2017).

In the local model, two basic kinds of protocols exist: interactive and non-interactive. Smith et al. (2017) have recently investigated the power of non-interactive differentially private protocols. These protocols are more natural for the classical use cases of the local model, e.g., both the projects from Google and Apple use the non-interactive model. Moreover, implementing efficient interactive protocols in such applications is more challenging due to the latency of the network. Despite its applications in industry, the local model has been much less studied than the central one. Part of the reason for this is that there are intrinsic limitations in what one can do in the local model. As a consequence, many basic questions, that are well studied in the central model, have not been completely understood in the local model, yet.

In this paper, we study differentially private Empirical Risk Minimization in the non-interactive local model. Before showing our contributions and discussing comparisons with previous work, we first discuss our motivations.

Problem setting (Smith et al., 2017; Kasiviswanathan et al., 2011) Given a convex, closed and bounded constraint set $\mathcal{C} \subseteq \mathbb{R}^p$, a data universe \mathcal{D} , and a loss function $\ell : \mathcal{C} \times \mathcal{D} \mapsto \mathbb{R}$, a dataset $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\} \in \mathcal{D}^n$ with data records $\{x_i\}_{i=1}^n \subset \mathbb{R}^p$ and labels (responses) $\{y_i\}_{i=1}^n \subset \mathbb{R}$ defines an *empirical risk* function: $L(w; D) = \frac{1}{n} \sum_{i=1}^n \ell(w; x_i, y_i)$ (note that in some case there could be no labels $\{y_i\}_{i=1}^n$, such as mean estimation). When the inputs are drawn i.i.d from an unknown underlying distribution \mathcal{P} on \mathcal{D} , we can also define the *population risk* function: $L_{\mathcal{P}}(w) = \mathbb{E}_{D \sim \mathcal{P}^n}[\ell(w; D)]$.

Thus, we have the following two types of excess risk, empirical risk

$$\text{Err}_D(w_{\text{priv}}) = L(w_{\text{priv}}; D) - \min_{w \in \mathcal{C}} L(w; D)$$

and population risk

$$\text{Err}_{\mathcal{P}}(w_{\text{priv}}) = L_{\mathcal{P}}(w_{\text{priv}}) - \min_{w \in \mathcal{C}} L_{\mathcal{P}}(w).$$

The problem considered in this paper to find $w_{\text{priv}} \in \mathcal{C}$ under non-interactive local differential privacy (see Definition 2) so as to minimize the empirical and/or population excess risks. Alternatively, when the dimensionality p is a constant or low, we can express this problem in terms of *sample complexity* and find the smallest n that achieves $\text{Err}_D \leq \alpha$ and $\text{Err}_{\mathcal{P}} \leq \alpha$, where α is the user specified error tolerance (or simply called error).

Motivation Smith et al. (2017) proved the following result concerning the problem for general convex 1-Lipschitz loss functions over a bounded constraint set.

Theorem 1 ((Smith et al., 2017)) *Under the assumptions above, there is a non-interactive ϵ -LDP algorithm such that for all distribution \mathcal{P} on \mathcal{D} , with probability $1 - \beta$, we have*

$$\text{Err}_{\mathcal{P}}(w_{\text{priv}}) \leq \tilde{O}\left(\frac{(\sqrt{p}\log^2(1/\beta))^{\frac{1}{p+1}}}{\epsilon^2 n}\right). \quad (1)$$

A similar result holds for empirical risk $\text{Err}_{\mathcal{D}}$. Equivalently, to ensure a no more than α error, the sample complexity needs to be $n = \tilde{\Omega}(\sqrt{p}c^p\epsilon^{-2}\alpha^{-(p+1)})$, where c is some constant (approximately 2). Furthermore, the exponential dependence of the sample size on the dimensionality p (in the terms of $\alpha^{-(p+1)}$ and c^p) is, in general, unavoidable.

This situation is somehow undesirable: when the dimensionality p is high and the target error is low, the dependency on $\alpha^{-(p+1)}$ could make the sample size quite large. However, several results have already shown that for some specific loss functions, the exponential dependency on the dimensionality can be avoided. For example, Smith et al. (2017) show that, in the case of linear regression, there is a non-interactive (ϵ, δ) -LDP algorithm¹ whose sample complexity for achieving error α in the empirical risk is $n = \Omega(p \log(1/\delta)\epsilon^{-2}\alpha^{-2})$. This indicates that there is a gap between the general case and the cases of some specific loss functions. This motivates us to consider the following basic question.

Is it possible to introduce some natural conditions on the loss function which ensures the existence of a non-interactive ϵ -LDP algorithm with sample complexity sub-exponentially (ideally, it should be polynomially or even linearly) depending on the dimensionality p ?

To answer this question, we first show that for any 1-Lipschitz generalized linear convex loss function, *i.e.*, $\ell(w; x, y) = f(y_i < w, x_i >)$ for some 1-Lipschitz convex function f , there is an (ϵ, δ) non-interactive LDP algorithm, whose sample complexity for achieving error α in empirical risk depends only linearly, instead of exponentially, on the dimensionality p and quasi-polynomially on α, δ, ϵ . Our idea is based on some results in Approximation Theory. We first consider the case of hinge loss functions by using Chebyshev polynomials to approximate its derivative function after smoothing, and then performing the algorithm of Stochastic Inexact Gradient Descent (Dvurechensky and Gasnikov, 2016). Next we extend to all convex general linear functions, the key idea is showing that any 1-Lipschitz convex function in \mathbb{R} can be expressed as a linear combination of some linear functions and hinge loss functions, *i.e.*, plus functions of inner product $[< w, s >]_+ = \max\{0, < w, s >\}$. Based on this we propose a general method which is called the polynomial of inner product approximation.

We also apply our method to other type of loss functions. Particularly, we show that in the Euclidean median problem, where the loss function is the ℓ_2 norm, the sample complexity is only quasi-polynomial in $p, \alpha, \delta, \epsilon$. **Note that this is the first result for a non-generalized linear function with a sample complexity sub-exponentially depending on p .** Our result is based on the observation that the ℓ_2 norm function can be approximated by a linear combination of the absolute inner product functions, which can also be expressed as a combination of linear functions and plus functions of inner product.

1. Note that these two results are for non-interactive (ϵ, δ) -LDP, and we mainly focus on non-interactive ϵ -LDP algorithms. Thus, we omit terms related to $\log(1/\delta)$ in this paper.

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, such as (Wang et al., 2017; Bassily et al., 2014; Kifer et al., 2012; Chaudhuri et al., 2011; Talwar et al., 2015). We consider in this paper only those results which are related to DP-ERM under the local model.

ERM in the interactive local model of differential privacy has been studied for quite some time (Kasiviswanathan et al., 2011; Beimel et al., 2008; Duchi et al., 2017, 2013; Zheng et al., 2017; Smith et al., 2017). Kasiviswanathan et al. (2011) showed a general equivalence between learning in the local model and learning in the statistical query model. Beimel et al. (2008) gave the lower bound of the squared error of distributed protocols for mean estimation. Duchi et al. (2017, 2013) obtained the lower bound $O(\frac{\sqrt{d}}{\epsilon\sqrt{n}})$ and optimal algorithms for general convex optimization, which require many rounds of interactions.

Methods	Sample Complexity	Assumption on the Loss Function
Claim 4 in (Smith et al., 2017)	$\tilde{\Omega}(4^p \alpha^{-(p+2)} \epsilon^{-2})$	1-Lipschitz
Theorem 10 in (Smith et al., 2017)	$\Omega(2^p \alpha^{-(p+1)} \epsilon^{-2})$	1-Lipschitz and Convex
Smith et al. (2017)	$\Theta(p \epsilon^{-2} \alpha^{-2})$	Linear Regression
Wang et al. (2018)	$\tilde{\Omega}((c_0 p^{\frac{1}{4}})^p \alpha^{-(2+\frac{p}{2})} \epsilon^{-2})$	$(8, T)$ -smooth
Wang et al. (2018)	$\tilde{\Omega}(4^{p(p+1)} D_p^2 \epsilon^{-2} \alpha^{-4})$	(∞, T) -smooth
Zheng et al. (2017)	$\Omega(p(\frac{8}{\alpha})^{4 \log \log(8/\alpha)} (\frac{4}{\epsilon})^{2c \log(8/\alpha)+2} (\frac{1}{\alpha^2 \epsilon^2}))$	Smooth Generalized Linear
This Paper	$\Omega(\frac{(c \log(4/\alpha))^{4c \log(4/\alpha)+4} 16^c \log(4/\alpha) p}{\epsilon^{4c \log(4/\alpha)+4} \alpha^4})$	1-Lipschitz Convex Generalized Linear
This Paper	$\Omega((\frac{c \log(4\sqrt{p}/\alpha)}{\epsilon^{2c \log(4\sqrt{p}/\alpha)} \alpha^4})^{2c \log(4\sqrt{p}/\alpha)+2} 8^c \log(4\sqrt{p}/\alpha) p^3)$	Euclidean Median

Table 1: Comparisons on the sample complexities for achieving error α in the empirical risk, where c is a constant. We assume that $\|x_i\|_2, \|y_i\| \leq 1$ for every $i \in [n]$ and the constraint set $\|\mathcal{C}\|_2 \leq 1$.

ERM in the non-interactive local model of differential privacy has received considerable attentions recently (Zheng et al., 2017; Smith et al., 2017; Wang et al., 2018) (see Table 1 for details). Smith et al. (2017) studied the general convex loss functions for population excess risk and showed that the exponential dependence on the dimensionality is unavoidable. In (Wang et al., 2018), the authors demonstrated that such an exponential dependence in the term of α is actually avoidable if the loss function smooth enough (*i.e.*, (∞, T) -smooth). Their result even holds for non-convex loss functions. However, there is still another term c^{p^2} in the sample complexity. In this paper, we investigate the conditions which allow us to avoid this issue and obtain sample complexity which is linear or quasi-polynomial in p .

Perhaps, the work most related to ours is the one in (Zheng et al., 2017), which also considered some specific loss functions in high dimensions, such as sparse linear regression and kernel ridge regression. The major differences with ours are the follows. Firstly, although it studied a similar class of loss functions (*i.e.*, Smooth Generalized Linear Loss functions) and

used the polynomial approximation approach, it needs quite a few assumptions on the loss function additional to the smoothness condition, such as Lipschitz smoothness and boundness on the higher order derivative functions, which are clearly not satisfied by the hinge loss functions. Contrarily, ours only assumes the 1-Lipschitz convex condition on the loss function. Secondly, even though the idea in our algorithm for the hinge loss functions is similar to theirs, its generalization to any generalized linear loss function is completely different and mainly based on some techniques from approximation theory. Thirdly, our approach can even be extended to some non-generalized linear function and achieved the first sample complexity for the Euclidean median problem with sub-exponential (*i.e.*, quasi-polynomial) dependence on p .

3. Preliminaries

Assumption 1 We assume that $\|x_i\|_2 \leq 1$ and $|y_i| \leq 1$ for each $i \in [n]$ and the constraint set $\|\mathcal{C}\|_2 \leq 1$. Unless specified otherwise, the loss function is assumed to be general linear, that is, the loss function $\ell(\theta; x_i, y_i) \equiv f(y_i \langle x_i, \theta \rangle)$ for some 1-Lipschitz convex function.

We note that the above assumptions on x_i, y_i and \mathcal{C} are quite common for the studies of DP-ERM (Smith et al., 2017; Wang et al., 2018; Zheng et al., 2017). The general linear assumption holds for a large class of functions such as Generalized Linear Model and SVM. We also note that there is another definition for general linear functions, $\ell(w; x, y) = f(\langle w, x \rangle, y)$, which is more general than our definition. This class of functions has been studied in (Kasiviswanathan and Jin, 2016; Wang et al., 2018); we leave as future research to extend our work to this class of loss functions.

Differential privacy in the local model. In LDP, we have a data universe \mathcal{D} , n players with each holding a private data record $x_i \in \mathcal{D}$, and a server coordinating the protocol. An LDP protocol executes a total of T rounds. In each round, the server sends a message, which is also called a query, to a subset of the players requesting them to run a particular algorithm. Based on the query, each player i in the subset selects an algorithm Q_i , runs it on her own data, and sends the output back to the server.

Definition 2 (Kasiviswanathan et al., 2011; Smith et al., 2017) *An algorithm Q is ϵ -locally differentially private (LDP) if for all pairs $x, x' \in \mathcal{D}$, and for all events E in the output space of Q , we have*

$$\Pr[Q(x) \in E] \leq e^\epsilon \Pr[Q(x') \in E].$$

A multi-player protocol is ϵ -LDP if for all possible inputs and runs of the protocol, the transcript of player i 's interaction with the server is ϵ -LDP. If $T = 1$, we say that the protocol is ϵ non-interactive LDP.

In the following, we will rephrase some basic definitions and lemmas on Chebyshev polynomial approximation.

Definition 3 *The Chebyshev polynomials $\{\mathcal{T}(x)_n\}_{n \geq 0}$ are recursively defined as follows*

$$\mathcal{T}_0(x) \equiv 1, \mathcal{T}_1(x) \equiv x \text{ and } \mathcal{T}_{n+1}(x) = 2x\mathcal{T}_n(x) - \mathcal{T}_{n-1}(x).$$

It satisfies that for any $n \geq 0$

$$\mathcal{T}_n(x) = \begin{cases} \cos(n \arccos(x)), & \text{if } |x| \leq 1 \\ \cosh(n \operatorname{arccosh}(x)), & \text{if } x \geq 1 \\ (-1)^n \cosh(n \operatorname{arccosh}(-x)), & \text{if } x \leq -1 \end{cases}$$

Definition 4 For every $\rho > 0$, let Γ_ρ be the ellipse Γ of foci ± 1 with major radius $1 + \rho$.

Definition 5 For a function f with a domain containing in $[-1, 1]$, its degree- n Chebyshev truncated series is denoted by $P_n(x) = \sum_{k=0}^n a_k \mathcal{T}_k(x)$, where the coefficient $a_k = \frac{2-1[k=0]}{\pi} \int_{-1}^1 \frac{f(x) \mathcal{T}_k(x)}{\sqrt{1-x^2}} dx$.

Lemma 6 (Chebyshev Approximation Theorem (Trefethen, 2013)) Let $f(z)$ be a function that is analytic on Γ_ρ and has $|f(z)| \leq M$ on Γ_ρ . Let $P_n(x)$ be the degree- n Chebyshev truncated series of $f(x)$ on $[-1, 1]$. Then, we have

$$\max_{x \in [-1, 1]} |f(x) - P_n(x)| \leq \frac{2M}{\rho + \sqrt{2\rho + \rho^2}} (1 + \rho + \sqrt{2\rho + \rho^2})^{-n},$$

$$|a_0| \leq M, \text{ and } |a_k| \leq 2M(1 + \rho + \sqrt{2\rho + \rho^2})^{-k}.$$

The following theorem shows the convergence rate of the Stochastic Inexact Gradient Method (Dvurechensky and Gasnikov, 2016), which will be used in our algorithm. We first give the definition of inexact oracle.

Definition 7 For an objective function, a (γ, β, σ) stochastic oracle returns a tuple $(F_{\gamma, \beta, \sigma}(w; \xi), G_{\gamma, \beta, \sigma}(w; \xi))$ such that

$$\begin{aligned} \mathbb{E}_\xi[F_{\gamma, \beta, \sigma}(w; \xi)] &= f_{\gamma, \beta, \sigma}(w), \\ \mathbb{E}_\xi[G_{\gamma, \beta, \sigma}(w; \xi)] &= g_{\gamma, \beta, \sigma}(w), \\ \mathbb{E}_\xi[\|G_{\gamma, \beta, \sigma}(w; \xi) - g_{\gamma, \beta, \sigma}(w)\|_2^2] &\leq \sigma^2, \\ 0 \leq f(v) - f_{\gamma, \beta, \sigma}(w) - \langle g_{\gamma, \beta, \sigma}(w), v - w \rangle &\leq \frac{\beta}{2} \|v - w\|^2 + \gamma, \forall v, w \in \mathcal{C}. \end{aligned}$$

Lemma 8 (Convergence Rate of SIGM (Dvurechensky and Gasnikov, 2016)) Assume that $f(w)$ is endowed with a (γ, β, σ) stochastic oracle with $\beta \geq O(1)$. Then, the sequence w_k generated by SIGM algorithm satisfies the following inequality

$$\mathbb{E}[f(w_k)] - \min_{w \in \mathcal{C}} f(w) \leq \Theta\left(\frac{\beta \sigma \|\mathcal{C}\|_2^2}{\sqrt{k}} + \gamma\right).$$

4. Main Results

In this section, we present our main results for LDP-ERM.

4.1 Sample Complexity for Hinge Loss Function

We first consider LDP-ERM with hinge loss function and then extend the obtained result to general convex linear functions.

The hinge loss function is defined as $\ell(w; x_i, y_i) = f(y_i \langle x_i, w \rangle) = [\frac{1}{2} - y_i \langle w, x_i \rangle]_+$, where the plus function $[x]_+ = \max\{0, x\}$, i.e., $f(x) = \max\{0, \frac{1}{2} - x\}$ for $x \in [-1, 1]$. Note that to avoid the scenario that $1 - y_i \langle w, x_i \rangle$ is always greater than or equal to 0, we use $\frac{1}{2}$, instead of 1 as in the classical setting.

Before showing our idea, we first smoothen the function $f(x)$. The following lemma shows one of the smooth functions that is close to f in the domain of $[-1, 1]$ (note that there are other ways to smoothen f ; see (Chen and Mangasarian, 1996) for details).

Lemma 9 *Let $f_\beta(x) = \frac{\frac{1}{2}-x+\sqrt{(\frac{1}{2}-x)^2+\beta^2}}{2}$ be a function with parameter $\beta > 0$. Then, we have*

1. $f_\beta(x)$ is analytic on $x \in \mathbb{R}$.
2. $|f_\beta(x) - f(x)|_\infty \leq \frac{\beta}{2}, \forall x \in \mathbb{R}$.
3. $f_\beta(x)$ is 1-Lipschitz, that is, $f'(x)$ is bounded by 1 for $x \in \mathbb{R}$.
4. f_β is $\frac{1}{\beta}$ -smooth and convex.

The above lemma indicates that $f_\beta(x)$ is a smooth and convex function which well approximates $f(x)$. This suggests that we can focus on $f_\beta(y_i \langle w, x_i \rangle)$, instead of f . Our idea is to construct a locally private (γ, β, σ) stochastic oracle for some γ, β, σ to approximate $f'_\beta(y_i \langle w, x_i \rangle)$ in each iteration, and then run the SIGM step of (Dvurechensky and Gasnikov, 2016). By Lemma 9, we know that f'_β is bounded and analytic; thus, we can use Lemma 6 to approximate f'_β via Chebyshev polynomials. Let $P_d(x) = \sum_{i=0}^d a_i \mathcal{T}_i(x) = \sum_{i=0}^d c_i x^i$, where $\max_{x \in [-1, 1]} |P_d(x) - f'(x)| \leq \frac{\alpha}{4}$ (i.e., $d = c \log(4/\alpha)$ for some constant $c > 0$) and $\sum_{i=0}^d c_i x^i$ is the polynomial expansion of $\sum_{i=0}^d a_i \mathcal{T}_i(x)$. Then, we have $\nabla_w \ell(w; x, y) = f'(y \langle w, x \rangle) y x^T$, which can be approximated by $[\sum_{i=0}^d c_i (y \langle w, x \rangle)^i] y x^T$. The idea is that if $(y \langle w, x \rangle)^i$ and $y x^T$ can be approximated locally differentially privately by directly adding $i+1$ numbers of independent Gaussian noises, which means it is possible to form an unbiased estimator of the term $[\sum_{i=0}^d c_i (y_i \langle w, x_i \rangle)^i] y_i x_i^T$. The error of this procedure can be estimated by Lemma 8. Details of the algorithm are given in Algorithm 1.

Theorem 10 *For each $i \in [n]$, the term $G(w_t, i)$ generated by Algorithm 1 is an $(\frac{\alpha}{2}, \frac{1}{\beta}, O(\frac{d^{2d+2} 4^d \sqrt{p}}{\epsilon^{2d+2}} + \alpha + 1))$ stochastic oracle for function $L_\beta(w; D) = \frac{1}{n} \sum_{i=1}^n f_\beta(y_i \langle x_i, w \rangle)$, where f_β is the function in Lemma 9.*

From Lemmas 8, 9 and Theorem 10, we have the following sample complexity bound for the hinge loss function under the non-interactive local model.

Algorithm 1 Hinge Loss-LDP

```

1: Input: Player  $i \in [n]$  holds data  $(x_i, y_i) \in \mathcal{D}$ , where  $\|x_i\|_2 \leq 1, \|y_i\|_2 \leq 1$ ; privacy
   parameters  $\epsilon, \delta$ ;  $d$  is the degree of Chebyshev truncated series of  $f'_\beta$  to achieve the
   approximation error of  $\frac{\alpha}{4}$  and  $P_d(x) = \sum_{i=0}^d a_i \mathcal{T}_i(x) = \sum_{i=0}^d c_i x^i$  is its Chebyshev
   polynomial approximation.
2: for Each Player  $i \in [n]$  do
3:   Calculate  $x_{i,0} = x_i + \sigma_{i,0}$  and  $y_{i,0} = y_i + z_{i,0}$ , where  $\sigma_{i,0} \sim \mathcal{N}(0, \frac{32 \log(1.25/\delta)}{\epsilon^2} I_p)$  and
       $z_{i,0} \sim \mathcal{N}(0, \frac{32 \log(1.25/\delta)}{\epsilon^2})$ .
4:   for  $j = 1, \dots, \frac{d(d+1)}{2}$  do
5:      $x_{i,j} = x_i + \sigma_{i,j}$ , where  $\sigma_{i,j} \sim \mathcal{N}(0, \frac{8 \log(1.25/\delta) d^2 (d+1)^2}{\epsilon^2} I_p)$ 
6:      $y_{i,j} = y_i + z_{i,j}$ , where  $z_{i,j} \sim \mathcal{N}(0, \frac{8 \log(1.25/\delta) d^2 (d+1)^2}{\epsilon^2})$ 
7:   end for
8:   Send  $\{x_{i,j}\}_{j=0}^{\frac{d(d+1)}{2}}$  and  $\{y_{i,j}\}_{j=0}^{\frac{d(d+1)}{2}}$  to the server.
9: end for
10: for the Server side do
11:   for  $t = 1, 2, \dots, n$  do
12:     Randomly sample  $i \in [n]$  uniformly.
13:     Set  $t_{i,0} = 1$ 
14:     for  $j = 1, \dots, d$  do
15:        $t_{i,j} = \prod_{k=j(j-1)/2+1}^{j(j+1)/2} y_{i,k} < w_t, x_{i,k} >$ 
16:     end for
17:     Denote  $G(w_t, i) = (\sum_{j=0}^d c_j t_{i,j}) y_{i,0} x_{i,0}^T$ .
18:     Update SIGM in (Dvurechensky and Gasnikov, 2016) by  $G(w_t, i)$ 
19:   end for
20: end for
21: return  $w_n$ 

```

Theorem 11 For any $\epsilon > 0$ and $0 < \delta < 1$, Algorithm 1 is (ϵ, δ) non-interactively locally differentially private². Furthermore, for the target error α , if choosing sample size $n = \Omega(\frac{d^{4d+4} 16^d p}{\epsilon^{4d+4} \alpha^4})$ and setting $\beta = \Theta(\frac{d^{d+1} 2^d \sqrt{4/p}}{\epsilon^{d+1} \sqrt{n}})$, the output w_n satisfies the following inequality

$$\mathbb{E}L(w_n, D) - \min_{w \in \mathcal{C}} L(w, D) \leq \alpha,$$

where $d = c \log(4/\alpha)$ for some universal constant $c > 0$.

Remark 12 Note that the sample complexity bound in Theorem 11 is quite loose for parameters other than p . This is mainly due to the fact that we use only the basic composition theorem to ensure local differential privacy. It is possible to obtain a tighter bound by using Advanced Composition Theorem (Dwork et al., 2010) (same for other algorithms in this paper). Details of the improvement are omit from this version. We can also extend to the population

2. Note that in the non-interactive local model, (ϵ, δ) -LDP is equivalent to ϵ -LDP by using some protocol given in Bun et al. (2018); this allows us to omit the term of δ . The full sample complexity of n is quasi-polynomial in $\ln(1/\delta)$.

risk by the same algorithm, the main difference is that now $G(w, i)$ is a $(\frac{\alpha}{2}, \frac{1}{\beta}, O(\frac{d^{2d+2}4^d\sqrt{p}}{\epsilon^{2d+2}} + \alpha + \sigma))$ stochastic oracle, where $\sigma^2 = \mathbb{E}_{(x,y) \sim \mathcal{P}} \|\ell(w; x, y) - \mathbb{E}_{(x,y) \sim \mathcal{P}} \ell(w; x, y)\|_2^2$. For simplicity of presentation, we omit the details here.

4.2 Extension to Generalized Linear Convex Loss Functions

In this section, we extend our results for the hinge loss function to generalized linear convex loss functions $L(w, D) = \frac{1}{n} \sum_{i=1}^n f(y_i \langle x_i, w \rangle)$ for any 1-Lipschitz convex function f .

One possible way (for the extension) is to follow the same approach used in previous section. That is, we first smoothen the function f by some function f_β . Then, we use Chebyshev polynomials to approximate the derivative function f'_β , and apply an algorithm similar to Algorithm 1. One of the main issues of this approach is that we do not know whether Chebyshev polynomials (*i.e.*, Lemma 6) can be directly used for every smooth convex function. Instead, we will use some ideas in Approximation Theory, which says that every 1-Lipschitz convex function can be expressed by a linear combination of the absolute functions and some linear functions.

To implement this approach, we first note that for the plus function $f(x) \equiv \max\{0, x\}$, by using Algorithm 1 we can get the same result as in Theorem 11. Since the absolute function $|x| = 2 \max\{0, x\} - x$, Theorem 11 clearly also holds for the absolute function. The following key lemma shows that every 1-dimensional 1-Lipschitz convex function $f : [-1, 1] \mapsto [-1, 1]$ is contained in the convex hull of the set of absolute and identity functions. We need to point out that Smith et al. (2017) gave a similar lemma. Their proof is, however, somewhat incomplete and thus we give a complete one in this paper.

Lemma 13 *Let $f : [-1, 1] \mapsto [-1, 1]$ be a 1-Lipschitz convex function. If we define the distribution \mathcal{Q} which is supported on $[-1, 1]$ as the output of the following algorithm:*

1. first sample $u \in [f'(-1), f'(1)]$ uniformly,
2. then output s such that $u \in \partial f(s)$ (note that such an s always exists due to the fact that f is convex and thus f' is non-decreasing); if multiple number of such as s exist, return the maximal one,

then, there exists a constant c such that

$$\forall \theta \in [-1, 1], f(\theta) = \frac{f'(1) - f'(-1)}{2} \mathbb{E}_{s \sim \mathcal{Q}} |\theta - s| + \frac{f'(1) + f'(-1)}{2} \theta + c.$$

Using Lemma 13 and the ideas discussed in the previous section, we can now show that the sample complexity in Theorem 11 also holds for any general linear convex function. See Algorithm 2 for the details.

Theorem 14 *Under Assumption 1, where the loss function ℓ is $\ell(w; x, y) = f(y < w, x >)$ for any 1-Lipschitz convex function f , for any $\epsilon, \delta \in (0, 1]$, Algorithm 2 is (ϵ, δ) non-interactively differentially private. Moreover, given the target error α , if choosing n and β such that $n = \Omega(\frac{d^{4d+4}16^d p}{\epsilon^{4d+4}\alpha^4})$ and $\beta = \Theta(\frac{d^{d+1}2^d \sqrt[4]{p}}{\epsilon^{d+1}\sqrt[4]{n}})$, the output w_n satisfies the following inequality*

$$\mathbb{E}L(w_n, D) - \min_{w \in \mathcal{C}} L(w, D) \leq \alpha,$$

where $d = c \log(4/\alpha)$ for some universal constant $c > 0$ independent of f .

Algorithm 2 General Linear-LDP

```

1: Input: Player  $i \in [n]$  holds raw data record  $(x_i, y_i) \in \mathcal{D}$ , where  $\|x_i\|_2 \leq 1$  and  $\|y_i\|_2 \leq 1$ ;
   privacy parameters  $\epsilon, \delta$ ; degree  $d$  of the Chebyshev truncated series of  $h'_\beta$  to achieve the
   approximation error  $\frac{\alpha}{4}$ , where  $h_\beta = \frac{x + \sqrt{x^2 + \beta^2}}{2}$  and  $P_d(x) = \sum_{i=0}^d a_i \mathcal{T}_i(x) = \sum_{i=0}^d c_i x^i$ 
   is its Chebyshev polynomial approximation. Loss function  $\ell$  can be represented by
    $\ell(w; x, y) = f(y < w, x >)$ .
2: for Each Player  $i \in [n]$  do
3:   Calculate  $x_{i,0} = x_i + \sigma_{i,0}$  and  $y_{i,0} = y_i + z_{i,0}$ , where  $\sigma_{i,0} \sim \mathcal{N}(0, \frac{32 \log(1.25/\delta)}{\epsilon^2} I_p)$  and
    $z_{i,0} \sim \mathcal{N}(0, \frac{32 \log(1.25/\delta)}{\epsilon^2})$ 
4:   for  $j = 1, \dots, \frac{d(d+1)}{2}$  do
5:      $x_{i,j} = x_i + \sigma_{i,j}$ , where  $\sigma_{i,j} \sim \mathcal{N}(0, \frac{8 \log(1.25/\delta) d^2 (d+1)^2}{\epsilon^2} I_p)$ 
6:      $y_{i,j} = y_i + z_{i,j}$ , where  $z_{i,j} \sim \mathcal{N}(0, \frac{8 \log(1.25/\delta) d^2 (d+1)^2}{\epsilon^2})$ 
7:   end for
8:   Send  $\{x_{i,j}\}_{j=0}^{\frac{d(d+1)}{2}}$  and  $\{y_{i,j}\}_{j=0}^{\frac{d(d+1)}{2}}$  to the server.
9: end for
10: for the Server side do
11:   for  $t = 1, 2, \dots, n$  do
12:     Randomly sample  $i \in [n]$  uniformly.
13:     Randomly sample  $\frac{d(d+1)}{2}$  numbers of i.i.d  $s = \{s_k\}_{k=1}^{\frac{d(d+1)}{2}} \in [-1, 1]$  based on the
     distribution  $\mathcal{Q}$  in Lemma 13.
14:     Set  $t_{i,0} = 1$ 
15:     for  $j = 1, \dots, d$  do
16:        $t_{i,j} = \prod_{k=j(j-1)/2+1}^{j(j+1)/2} \left( \frac{y_{i,k} < w_t, x_{i,k} > -s_k}{2} \right)$ 
17:     end for
18:     Denote  $G(w_t, i, s) = (f'(1) - f'(-1))(\sum_{j=0}^d c_j t_{i,j}) y_{i,0} x_{i,0}^T + f'(-1)$ .
19:     Update SIGM in (Dvurechensky and Gasnikov, 2016) by  $G(w_t, i, s)$ 
20:   end for
21: end for
22: return  $w_n$ 

```

Remark 15 The above theorem suggests that the sample complexity for any generalized linear loss function depends only linearly on p . However, there are still some not so desirable issues. Firstly, the dependence on α is quasi-polynomial, while previous work (Wang et al., 2018) has already shown that it is only polynomial (i.e., α^{-4}) for sufficiently smooth loss functions. Secondly, the term of ϵ is not optimal in the sample complexity, since it is $\epsilon^{-\Omega(\ln(1/\alpha))}$, while the optimal one is ϵ^{-2} . We leave it as an open problem to remove the quasi-polynomial dependency. Thirdly, the assumption on the loss function is that $\ell(w; x, y) = f(y < w, x >)$, which includes the generalized linear models and SVM. However, as mentioned earlier, there is another slightly more general function class $\ell(w; x, y) = f(< w, x >, y)$ which does not always satisfy our assumption, e.g., linear regression and ℓ_1 regression. For linear regression, we have already known its optimal bound $\Theta(p\alpha^{-2}\epsilon^{-2})$; for ℓ_1 regression, we can use a method similar to Algorithm 1 to achieve

a sample complexity which is linear in p . Thus, a natural question is whether the sample complexity is still linear in p for all loss functions $\ell(w; x, y)$ that can be written as $f(\langle w, x \rangle, y)$.

4.3 Further Extension to Euclidean Median Problem

Last section has showed that using the approximation of hinge loss function and polynomials of inner product functions, we can extend our approach to generalized linear convex loss functions for LDP-ERM. To show the power of this method, we consider in this section the Euclidean median problem, which cannot be written as a function of inner product $\langle w, x \rangle$. Euclidean median problem is one of the classic problem in optimization and has been studied for many years (Cohen et al., 2016) :

$$L(w; D) = \frac{1}{2n} \sum_{i=1}^n \|w - x_i\|_2.$$

Note that we need $2n$, instead of n , data points to ensure that the loss function $\frac{\|w - x_i\|_2}{2}$ is 1-Lipschitz and the term $\langle \frac{w - x_i}{2}, u \rangle$ is bounded by 1 in $\|C\|_2 \leq 1$. It is obvious that the ℓ_2 -norm loss function cannot be written as a function of inner product. However, the following key lemma tells us that it can actually be well approximated by a linear combination of the absolute inner product functions.

Lemma 16 *Let P be the distribution of uniformly sampling from $(p-1)$ -dimensional unit sphere \mathbb{S}^{p-1} . Then, we have*

$$\|x\|_2 = \frac{\sqrt{\pi} p \Gamma(\frac{p-1}{2})}{2\Gamma(\frac{p}{2})} \mathbb{E}_{u \sim P} |\langle u, x \rangle|.$$

Note that the term $\frac{\sqrt{\pi} p \Gamma(\frac{p-1}{2})}{2\Gamma(\frac{p}{2})} = O(\sqrt{p})$.

With Lemma 16, we have Algorithm 3 and the following theorem for the Euclidean median problem based on the ideas in previous sections.

Theorem 17 *For any $\epsilon > 0$ and $0 < \delta < 1$, Algorithm 3 is (ϵ, δ) non-interactively locally differentially private. Furthermore, for the target error α , if choosing the sample size n and β such that $n = \Omega(\frac{d^{2d+2} 8^d p^3}{\epsilon^{2d} \alpha^4})$ and $\beta^2 = \Theta(\frac{C d^{d+2} \sqrt{8^d}}{\epsilon^d \sqrt[2]{n}})$, the output w_n satisfies the following inequality*

$$\mathbb{E} L(w_n; D) - \min_{w \in \mathcal{C}} L(w; D) \leq \alpha,$$

where $d = c \log(4C/\alpha)$ for some constant $c > 0$ and $C = \frac{\sqrt{\pi} p \Gamma(\frac{p-1}{2} + 1)}{2\Gamma(\frac{p}{2} + 1)} = O(\sqrt{p})$.

From previous sections, we can see that for any convex generalized linear loss function, the sample complexity needs only linearly depending on the dimensionality p . So far, we know that all loss functions have a sample complexity which is either linear in p (i.e., all known loss functions can be written as $f(\langle w, x \rangle, y)$) or exponential in p (such as the

Algorithm 3 Euclidean Median-LDP

```

1: Input: Player  $i \in [n]$  holding data  $\{x_i\}_{i=1}^n \in \mathcal{D}$ , where  $\|x_i\|_2 \leq 1$ ; privacy parameters
    $\epsilon, \delta$ ; degree  $d$  of the Chebyshev truncated series of  $h'_\beta$  to achieve the approximation error
    $\frac{\alpha}{2C}$ , where  $h_\beta = \frac{x + \sqrt{x^2 + \beta^2}}{2}$  and  $P_d(x) = \sum_{i=0}^d a_i T_i(x) = \sum_{i=0}^d c_i x^i$  is its Chebyshev
   polynomial approximation. Loss function  $\ell(w; x) = \frac{1}{2}\|w - x\|_2$  and  $C = \frac{\sqrt{\pi} p \Gamma(\frac{p-1}{2})}{2\Gamma(\frac{p}{2})}$ .
2: for Each Player  $i \in [n]$  do
3:   for  $j = 1, \dots, \frac{d(d+1)}{2}$  do
4:      $x_{i,j} = x_i + \sigma_{i,j}$ , where  $\sigma_{i,j} \sim \mathcal{N}(0, \frac{2 \log(1.25/\delta) d^2 (d+1)^2}{\epsilon^2} I_p)$ 
5:     Send  $\{x_{i,j}\}_{j=1}^{\frac{d(d+1)}{2}}$  to the server.
6:   end for
7: end for
8: for the Server side do
9:   for  $t = 1, 2, \dots, n$  do
10:    Randomly sample  $i \in [n]$  uniformly.
11:    Randomly sample  $\frac{d(d+1)}{2}$  number of i.i.d  $u = \{u_k\}_{k=0}^{\frac{d(d+1)}{2}} \in \mathbb{S}^{p-1}$  which follow the
    uniform distribution on the surface  $\mathbb{S}^{p-1}$ .
12:    Set  $t_{i,0} = 1$ 
13:    for  $j = 1, \dots, d$  do
14:       $t_{i,j} = \Pi_{k=j(j-1)/2+1}^{j(j+1)/2} \left( \frac{\langle u_k, w_t - x_{i,k} \rangle}{2} \right)$ 
15:    end for
16:    Denote  $G(w_t, i, u) = C \times u_0^T [\sum_{j=1}^d c_j t_{i,j} - \frac{1}{2}]$ .
17:    Update SIGM in (Dvurechensky and Gasnikov, 2016) by  $G(w_t, i, u)$ 
18:  end for
19: end for
20: return  $w_n$ 

```

example given in (Smith et al., 2017)). Thus, to our best knowledge, the Euclidean median problem (or ERM with loss function $\ell(w, x) = \frac{1}{2}\|w - x\|_2$) is the first result which is not generalized linear, but still has a sample complexity sub-exponential in p .

Compared with the result for generalized linear loss functions, the quasi-polynomial dependency in the sample complexity of the Euclidean median problem comes from the multiplicative factor $O(\sqrt{p})$ in Lemma 16, which forces us to use Chebyshev polynomial to achieve the error of $O(\frac{\alpha}{\sqrt{p}})$, instead of $O(\alpha)$ as in the previous sections. It remains as an open problem to determine whether this dependency is necessary. Also, extending our method to other loss functions is another direction for future research.

5. Discussion

In this paper, we propose a general method for Empirical Risk Minimization in non-interactive differentially private model by using polynomial of inner product approximation. Compared with the method of directly using polynomial approximation, such as the one in (Wang et al., 2018), which needs exponential (in p) number of grids to estimate the function

privately, our method avoid this undesirable issue. Using this method, we show that the sample complexity for any 1-Lipschitz generalized linear convex function is only linear in p . Moreover, we show that our method can be extended to the Euclidean median problem and achieve a sample complexity that is quasi-polynomial in p .

Appendix A. Detailed Proofs

Proof [Proof of Lemma 9] It is easy to see that items 1 and 2 are true. Item 3 is

due to the following $|f'_\beta(x)| = \left| \frac{-1 + \frac{x - \frac{1}{2}}{\sqrt{(x - \frac{1}{2})^2 + \beta^2}}}{2} \right| \leq 1$. Item 4 is because of the following $0 \leq f''_\beta(x) = \frac{\beta^2}{((x - \frac{1}{2})^2 + \beta^2)^{\frac{3}{2}}} \leq \frac{1}{\beta}$. \blacksquare

Proof [Proof of Theorem 10] For simplicity, we omit the term of δ , which will not affect the linear dependency. Let

$$\hat{G}(w, i) = \left[\sum_{j=0}^d c_j (y_i \langle w, x_i \rangle)^j \right] y_i x_i^T,$$

$$\mathbb{E}_i \hat{G}(w, i) = \frac{1}{n} \sum_{i=1}^n \hat{G}(w, i) = \hat{G}(w).$$

For the term of $G(w, i)$, the randomness comes from sampling the index i and the Gaussian noises added for preserving local privacy.

Note that in total $\mathbb{E}_{\sigma, z, i} G(w, i) = \hat{G}(w)$, where $\sigma = \{\sigma_{i,j}\}_{j=0}^{\frac{d(d+1)}{2}}$ and $z = \{z_{i,j}\}_{j=0}^{\frac{d(d+1)}{2}}$.

It is easy to see that $\mathbb{E}_{\sigma, z} G(w, i) = \mathbb{E}[(\sum_{j=0}^d c_j t_{i,j}) y_{i,0} x_{i,0}^T \mid i] = \hat{G}(w, i)$, which is due to the fact that $\mathbb{E} t_{i,j} = (y_i \langle w, x_i \rangle)^j$ and each $t_{i,j}$ is independent. We now calculate the variance for this term with fixed i . Firstly, we have $\text{Var}(y_{i,0} x_{i,0}^T) = O(\frac{p}{\epsilon^4})$. For each $t_{i,j}$, we get

$$\text{Var}(t_{i,j}) \leq \Pi_{k=j(j-1)/2+1}^{j(j+1)/2} \text{Var}(y_{i,k}) (\text{Var}(\langle w_i, x_{i,k} \rangle) + (\mathbb{E}(w_i^T x_{i,k}))^2) \leq \tilde{O}((\frac{d(d+1)}{\epsilon^2})^{2j}).$$

Since function f'_β is bounded by 1 and analytic, by Lemma 6 we know that $|a_i| \leq 1$ for each i . Also note that $c_k = \sum_{m=k}^d a_m b_{mk}$, where $|a_m| \leq 1$ is the Chebyshev coefficient of the original function f'_β and b_{mk} is the coefficient of order k monomial in Chebyshev basis $\mathcal{T}_m(x)$. By (Qazi and Rahman, 2007), we have

$$|b_{mk}| \leq \max_{\theta \in (0, \frac{1}{2})} O(\sqrt{m} [\frac{(1-\theta)^{1-\theta}}{\theta^\theta (1-2\theta)^{1-2\theta}}]^m) \leq O(\sqrt{m} 2^m).$$

This tells that $|c_k| \leq O(d^{\frac{3}{2}} 2^d)$ for each i . In total, we have

$$\text{Var}(G(w_t, i) \mid i) \leq O(d \cdot d^3 4^d \cdot (\frac{d(d+1)}{\epsilon^2})^{2d} \cdot \frac{p}{\epsilon^4}) = \tilde{O}(\frac{d^{4d+4} 16^d p}{\epsilon^{4d+4}}).$$

Next we consider $\text{Var}(\hat{G}(w, i))$. Since

$$\|\hat{G}(w, i) - f'_\beta(y_i x_i^T w) y_i x_i^T\|_2^2 = \|\sum_{j=0}^d c_j (y_i \langle w, x_i \rangle)^j - f'_\beta(w) y_i x_i^T\|_2^2 \leq \left(\frac{\alpha}{4}\right)^2,$$

we get

$$\begin{aligned} \text{Var}(\hat{G}(w, i)) &\leq O(\mathbb{E}[\|\hat{G}(w, i) - f'_\beta(y_i x_i^T w) y_i x_i^T\|_2^2] + \mathbb{E}[\|\hat{G}(w) - \nabla L_\beta(w; D)\|_2^2] \\ &\quad + \mathbb{E}[\|f'_\beta(y_i x_i^T w) y_i x_i^T - \nabla L_\beta(w; D)\|_2^2]) \leq O((\alpha + 1)^2). \end{aligned}$$

In total, we have $\mathbb{E}[\|G(w, i) - \hat{G}(w)\|_2^2] \leq \mathbb{E}[\|G(w, i) - \hat{G}(w, i)\|_2^2] + \mathbb{E}[\|\hat{G}(w, i) - \hat{G}(w)\|_2^2] \leq \tilde{O}\left(\left(\frac{d^{2d+2} 4^d \sqrt{p}}{\epsilon^{2d+2}} + \alpha + 1\right)^2\right)$.

Also, we know that

$$\begin{aligned} L_\beta(v; D) - L_\beta(w; D) - \langle \hat{G}(w), v - w \rangle &= \\ L_\beta(v; D) - L_\beta(w; D) - \langle \nabla L_\beta(w; D), v - w \rangle + \langle \nabla L_\beta(w; D) - G(w), v - w \rangle & \\ \leq \frac{1}{2\beta} \|v - w\|_2^2 + \frac{\alpha}{2}, & \end{aligned}$$

since L_β is $\frac{1}{\beta}$ -smooth and $|\langle \nabla L_\beta(w) - G(w), v - w \rangle| \leq \frac{\alpha}{2}$.

Thus, $G(w, i)$ is an $(\frac{\alpha}{2}, \frac{1}{\beta}, O(\frac{d^{2d+2} 4^d \sqrt{p}}{\epsilon^{2d+2}} + \alpha + 1))$ stochastic oracle of L_β . ■

Proof [Proof of Theorem 11]

The guarantee of differential privacy is by Gaussian mechanism and composition theorem.

By Theorem 10 and Lemma 8, we have

$$\mathbb{E}L_\beta(w_n, D) - \min_{w \in \mathcal{C}} L_\beta(w, D) \leq O\left(\frac{(\frac{d^{2d+2} 4^d \sqrt{p}}{\epsilon^{2d+2}} + \alpha + 1)}{\beta \sqrt{n}} + \frac{\alpha}{2}\right) = O\left(\frac{d^{2d+2} 4^d \sqrt{p}}{\epsilon^{2d+2} \beta \sqrt{n}} + \frac{\alpha}{2}\right).$$

By Lemma 9, we know that

$$\mathbb{E}L(w_n, D) - \min_{w \in \mathcal{C}} L(w, D) \leq O\left(\beta + \frac{d^{2d+2} 4^d \sqrt{p}}{\epsilon^{2d+2} \beta \sqrt{n}} + \frac{\alpha}{2}\right).$$

Thus, if we take $\beta = \Theta(\frac{d^{d+1} 2^d \sqrt{4p}}{\epsilon^{d+1} \sqrt{n}})$ and $n = \Omega(\frac{d^{4d+4} 16^d p}{\epsilon^{4d+4} \alpha^4})$, we have

$$\mathbb{E}L(w_n, D) - \min_{w \in \mathcal{C}} L(w, D) \leq \alpha.$$

■

Proof [Proof of Lemma 13] Let $g(\theta) = \mathbb{E}_{s \sim \mathcal{Q}} |s - \theta|$. Then, we have the following for every θ , where $f'(\theta)$ is well defined,

$$\begin{aligned} g'(\theta) &= \mathbb{E}_{s \sim \mathcal{Q}} [1_{s \leq \theta}] - \mathbb{E}_{s \sim \mathcal{Q}} [1_{s > \theta}] \\ &= \frac{[f'(\theta) - f'(-1)] - [f'(1) - f'(\theta)]}{f'(1) - f'(-1)} \\ &= \frac{2f'(\theta) - (f'(1) + f'(-1))}{f'(1) - f'(-1)}. \end{aligned}$$

Thus, we get

$$F'(\theta) = \frac{f'(1) - f'(-1)}{2} g'(\theta) + \frac{f'(1) + f'(-1)}{2} = f'(\theta).$$

Next, we show that if $F'(\theta) = f'(\theta)$ for every $\theta \in [0, 1]$, where $f'(\theta)$ is well defined, there is a constant c which satisfies the condition of $F(\theta) = f(\theta) + c$ for all $\theta \in [0, 1]$.

Lemma 18 *If f is convex and 1-Lipschitz, then f is differentiable at all but countably many points. That is, f' has only countable many discontinuous points.*

Proof [Proof of Lemma 18] Since f is convex, we have the following for $0 \leq s < u \leq v < t \leq 1$

$$\frac{f(u) - f(s)}{u - s} \leq \frac{f(t) - f(v)}{t - v},$$

This is due to the property of 3-point convexity, where

$$\frac{f(u) - f(s)}{u - s} \leq \frac{f(t) - f(u)}{t - u} \leq \frac{f(t) - f(v)}{t - v}.$$

Thus, we can obtain the following inequality of one-sided derivation, that is,

$$f'_-(x) \leq f'_+(x) \leq f'_-(y) \leq f'_+(y)$$

for every $x < y$. For each point where $f'_-(x) < f'_+(x)$, we pick a rational number $q(x)$ which satisfies the condition of $f'_-(x) < q(x) < f'_+(x)$. From the above discussion, we can see that all these $q(x)$ are different. Thus, there are at most countable many points where f is non-differentiable. \blacksquare

From the above lemma, we can see that the Lebesgue measure of these discontinuous points is 0. Thus, f' is Riemann Integrable on $[-1, 1]$. By Newton-Leibniz formula, we have the following for any $\theta \in [0, 1]$,

$$\int_{-1}^{\theta} f'(x) dx = f(\theta) - f(-1) = \int_{-1}^{\theta} F'(x) dx = F(\theta) - F(-1).$$

Therefore, we get $F(\theta) = f(\theta) + c$ and complete the proof. \blacksquare

Proof [Proof of Theorem 14]

Let h_β denote the function $h_\beta(x) = \frac{x + \sqrt{x^2 + \beta^2}}{2}$. By Lemma 13 we have

$$f(\theta) = (f'(1) - f'(-1))\mathbb{E}_{s \sim \mathcal{Q}} \frac{|s - \theta|}{2} + \frac{f'(1) + f'(-1)}{2}\theta + c.$$

Now, we consider function $F_\beta(\theta)$, which is

$$F_\beta(\theta) = (f'(1) - f'(-1))\mathbb{E}_{s \sim \mathcal{Q}} [2h_\beta(\frac{\theta - s}{2}) - \frac{\theta - s}{2}] + \frac{f'(1) + f'(-1)}{2}\theta + c.$$

From this, we have

$$\nabla F_\beta(\theta) = (f'(1) - f'(-1))\mathbb{E}_{s \sim \mathcal{Q}} [\nabla h_\beta(\frac{\theta - s}{2})] + \frac{f'(1) + f'(-1)}{2} - \frac{f'(1) - f'(-1)}{2}.$$

Note that since $|x| = 2 \max\{x, 0\} - x$, we can get 1) $|F_\beta(\theta) - f(\theta)| \leq O(\beta)$ for any $\theta \in \mathbb{R}$, 2) $F_\beta(x)$ is $O(\frac{1}{\beta})$ -smooth and convex since $h_\beta(\theta - s)$ is $\frac{1}{\beta}$ -smooth and convex, and 3) $F_\beta(\theta)$ is $O(1)$ -Lipschitz. Now, we optimize the following problem in the non-interactive local model:

$$F_\beta(w; D) = \frac{1}{n} \sum_{i=1}^n F_\beta(y_i < x_i, w >).$$

For each fixed i and s , we let

$$\hat{G}(w, i, s) = (f'(1) - f'(-1)) \left[\sum_{j=1}^d c_j \Pi_{k=j(j-1)/2+1}^{j(j+1)/2} \left(\frac{y_i < w_t, x_i > -s_k}{2} \right) \right] y_i x_i^T + f'(-1).$$

Then, we have $\mathbb{E}_{\sigma, z} G(w, i, s) = \hat{G}(w, i, s)$. By using a similar argument given in the proof of Theorem 10, we get

$$\text{Var}(\hat{G}(w, i, s) | i, s) \leq \tilde{O}\left(\frac{d^{4d+4} 16^d p}{\epsilon^{4d+4}}\right).$$

Thus, for each fixed i we have

$$\mathbb{E}_s \hat{G}(w, i, s) = \bar{G}(w, i) = (f'(1) - f'(-1)) \left[\mathbb{E}_{s \sim \mathcal{Q}} \sum_{j=1}^d c_j \left(\frac{y_i < w, x_i > -s}{2} \right)^j \right] y_i x_i^T + f'(-1).$$

Next, we bound the term of $\text{Var}(\hat{G}(w, i, s) | i) \leq O(d)$.

Let $t_j = \Pi_{k=j(j-1)/2+1}^{j(j+1)/2} \left(\frac{y_i < w_t, x_i > -s_k}{2} \right)$. Then, we have

$$\text{Var}(t_j) \leq \Pi_{k=j(j-1)/2+1}^{j(j+1)/2} |y_i|^2 \text{Var}(< w_t, x_i > -s_k) \leq O(1).$$

Thus, we get

$$\text{Var}(\hat{G}(w, i, s) | i) \leq O\left(\sum_{j=1}^d c_j^2 \text{Var}(t_j)\right) = O(d \times d^3 \times 4^d = O(d^4 4^d)).$$

Since $\mathbb{E}_i \bar{G}(w, i) = \hat{G} = \frac{1}{n} \sum_{i=1}^n \bar{G}(w, i)$, we have $\text{Var}(\bar{G}(w, i)) \leq O((\alpha + 1)^2)$ by a similar argument given in the proof of Theorem 10. Thus, in total we have

$$\mathbb{E} \|G(w, i, s) - \hat{G}\| \leq \tilde{O}\left(\left(\frac{d^{2d+2} 4^d \sqrt{p}}{\epsilon^{2d+2}} + \alpha + 1 + d^2 2^d\right)^2\right) = \tilde{O}\left(\left(\frac{d^{2d+2} 4^d \sqrt{p}}{\epsilon^{2d+2}}\right)^2\right).$$

The other part of the proof is the same as that of Theorem 10. ■

Proof [Proof of Lemma 16] Let $g(x) = \mathbb{E}_{u \sim P} |< u, x >|$. Then, we have the following properties:

- For every x, y if $\|x\|_2 = \|y\|_2$, then $g(x) = g(y)$. This is due to the rotational symmetry of the ℓ_2 -norm ball.
- For any constant α , we have $g(\alpha x) = |\alpha|g(x)$.

Thus, for every $x \in \mathbb{R}^p$, we have $g(x) = \|x\|_2 g(\frac{x}{\|x\|_2}) = \|x\|_2 g(e_1)$, where $e_1 = (1, 0, \dots, 0)$. Next we calculate $g(e_1) = \mathbb{E}_{x \sim P} |x_1|$.

Let $s_p(r)$ denote the area of a $p-1$ -dimensional sphere with radius r . Then, we have $s_p(r) = \frac{p\pi^{\frac{p}{2}}}{\Gamma(\frac{p}{2}+1)} r^{p-1}$. Thus, we get

$$\mathbb{E}_{x \sim P} |x_1| = \frac{2}{s_p(1)} \int_0^1 s_{p-1}(\sqrt{1-r^2}) r dr.$$

By changing the coordinate $r = \sin(\theta)$, we then have

$$\mathbb{E}_{x \sim P} |x_1| = \frac{2}{s_p(1)} \int_0^{\frac{\pi}{2}} s_{p-1}(\cos \theta) \sin(\theta) \cos(\theta) d\theta = \frac{2s_{p-1}(1)}{s_p(1)} \int_0^{\frac{\pi}{2}} \cos^{p-1}(\theta) \sin(\theta) d\theta.$$

Also, since $\int_0^{\frac{\pi}{2}} \cos^{p-1}(\theta) \sin(\theta) d\theta = \frac{1}{p}$, we obtain

$$\mathbb{E}_{x \sim P} |x_1| = \frac{2s_{p-1}(1)}{s_p(1)p} = \frac{2(p-1)\pi^{\frac{p-1}{2}} \Gamma(\frac{p}{2}+1)}{p\pi^{\frac{p}{2}} \Gamma(\frac{p-1}{2}+1)} \cdot \frac{1}{p} = \frac{2\Gamma(\frac{p}{2})}{\sqrt{\pi}p\Gamma(\frac{p-1}{2})} = O\left(\frac{1}{\sqrt{p}}\right),$$

where the last inequality comes from the Stirling's approximation of the Γ -function. Hence, we have $\|x\|_2 = \frac{\sqrt{\pi}p\Gamma(\frac{p-1}{2})}{2\Gamma(\frac{p}{2})} g(x)$. ■

Proof [Proof of Theorem 17]

By Lemma 16, we can see that the optimization problem becomes the following

$$L(w; D) = \frac{C}{n} \sum_{i=1}^n \mathbb{E}_{u \sim P} |< u, \frac{w - x_i}{2} >|.$$

Let $\tilde{L}_\beta(w; D)$ denote the following function

$$\tilde{L}_\beta(w; D) = \frac{C}{n} \sum_{i=1}^n \mathbb{E}_{u \sim P} [2h_\beta\left(\frac{< u, w - x_i >}{2}\right) - < u, \frac{w - x_i}{2} >].$$

Then, we have

$$\nabla \tilde{L}_\beta(w; D) = \frac{C}{n} \sum_{i=1}^n \mathbb{E}_{u \sim P} [u^T h_\beta(\frac{\langle u, w - x_i \rangle}{2}) - \frac{u^T}{2}].$$

Thus, we know that $|\tilde{L}_\beta(w; D) - L(w; D)|_\infty \leq O(C\beta)$, and $\tilde{L}(w; D)$ is $O(\frac{C}{\beta})$ -smooth and convex. Now, consider the term $\hat{G}(w, i, u) = u_0^T [\sum_{j=1}^d c_j t_{i,j} - \frac{1}{2}]$.

For each fixed i, u , we know that

$$\mathbb{E}_\sigma \hat{G}(w, i, u) = \bar{G}(w, i, u) = u_0^T [\sum_{j=1}^d c_j \Pi_{k=j(j-1)/2+1}^{j(j+1)/2} (\frac{\langle u_k, w - x_i \rangle}{2}) - \frac{1}{2}].$$

Thus, by a similar argument given in the proof of Theorem 11 and the fact that $\|u_k\|_2 \leq 1$, we have

$$\text{Var}(\hat{G}(w, i, u) | i, u) \leq \tilde{O}(d \times d^3 \times 4^d (\frac{d(d+1)}{\epsilon^2})^d) = \tilde{O}(\frac{8^d d^{d+4}}{\epsilon^{2d}}).$$

Next, for each fixed i , we have

$$\mathbb{E}_u \bar{G}(w, i, u) = \check{G}(w, i) = \mathbb{E}_{u \sim P} [u^T (\sum_{j=1}^d c_j (\frac{\langle u, w - x_i \rangle}{2})^j - \frac{1}{2})].$$

Thus, we get $\text{Var}(\hat{G}(w, i, u)) \leq O(d^4 4^d)$.

For the term $\check{G}(w, i)$, by a similar argument given in the proof of Theorem 11, we know that

$$\mathbb{E}_i \check{G}(w, i) = \check{G}(w) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{u \sim P} [u^T (\sum_{j=1}^d c_j (\frac{\langle u, w - x_i \rangle}{2})^j - \frac{1}{2})].$$

Thus, we have $\text{Var}(\check{G}(w, i)) \leq O((\frac{\alpha}{2C} + 1)^2)$.

In total, we have $\text{Var}(G(w, i, u)) \leq \tilde{O}((\frac{C\sqrt{8^d d^{d+2}}}{\epsilon^d} + C)^2)$. This means that $G(w, i, u)$ is an $(\frac{\alpha}{2}, O(\frac{C}{\beta}), O(\frac{C\sqrt{8^d d^{d+2}}}{\epsilon^d} + C))$ stochastic oracle of $\hat{L}(w; D)$.

By Lemma 8, we know that after n iterations, the following holds

$$\mathbb{E}[\hat{L}(w_n; D)] - \min_{w \in C} \hat{L}(w; D) \leq \Theta(\frac{C}{\beta} \times \frac{C\sqrt{8^d d^{d+2}}}{\sqrt{n}\epsilon^d} + \frac{\alpha}{2}).$$

By the relation between $\hat{L}(w; D)$ and $L(w; D)$, we finally get

$$\mathbb{E}[L(w_n; D)] - \min_{w \in C} L(w; D) \leq \Theta(\frac{C}{\beta} \times \frac{C\sqrt{8^d d^{d+2}}}{\sqrt{n}\epsilon^d} + \frac{\alpha}{2} + C\beta).$$

Taking $\beta^2 = \Theta(\frac{C(2\sqrt{2})^d d^{d+2}}{\sqrt{n}\epsilon^d})$, we get the proof. ■

References

- Raef Bassily, Adam Smith, and Abhradeep 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.
- Amos Beimel, Kobbi Nissim, and Eran Omri. Distributed private data analysis: Simultaneously solving how and what. In *CRYPTO*, volume 5157, pages 451–468. Springer, 2008.
- Mark Bun, Jelani Nelson, and Uri Stemmer. Heavy hitters and the structure of local privacy. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 435–447. ACM, 2018.
- Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.
- Chunhui Chen and Olvi L Mangasarian. A class of smoothing functions for nonlinear and mixed complementarity problems. *Computational Optimization and Applications*, 5(2): 97–138, 1996.
- Michael B Cohen, Yin Tat Lee, Gary Miller, Jakub Pachocki, and Aaron Sidford. Geometric median in nearly linear time. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 9–21. ACM, 2016.
- John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, 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*, (just-accepted), 2017.
- Pavel Dvurechensky and Alexander Gasnikov. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. *Journal of Optimization Theory and Applications*, 171(1):121–145, 2016.
- Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, volume 3876, pages 265–284. Springer, 2006.
- Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60. IEEE, 2010.
- Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067. ACM, 2014.

- Samuel Haney, Ashwin Machanavajjhala, John M. Abowd, Matthew Graham, Mark Kutzbach, and Lars Vilhuber. Utility cost of formal privacy for releasing national employer-employee statistics. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, pages 1339–1354, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4197-4. doi: 10.1145/3035918.3035940. URL <http://doi.acm.org/10.1145/3035918.3035940>.
- 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.
- Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.
- Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. *Journal of Machine Learning Research*, 1(41):3–1, 2012.
- Joe Near. Differential privacy at scale: Uber and berkeley collaboration. In *Enigma 2018 (Enigma 2018)*, Santa Clara, CA, 2018. USENIX Association.
- MA Qazi and QI Rahman. Some coefficient estimates for polynomials on the unit interval. *Serdica Mathematical Journal*, 33(4):449p–474p, 2007.
- Adam Smith, Abhradeep Thakurta, and Jalaj Upadhyay. Is interaction necessary for distributed private learning? In *IEEE Symposium on Security and Privacy*, 2017.
- Kunal Talwar, Abhradeep Guha Thakurta, and Li Zhang. Nearly optimal private lasso. In *Advances in Neural Information Processing Systems*, 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.
- Lloyd N Trefethen. *Approximation theory and approximation practice*, volume 128. Siam, 2013.
- Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 2719–2728, 2017.
- Di Wang, Marco Gaboardi, and Jinhui Xu. Empirical risk minimization in non-interactive local differential privacy revisited. *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, 3-8 December 2018, Montreal, QC, Canada*, 2018. URL <http://arxiv.org/abs/1802.04085>.

Kai Zheng, Wenlong Mou, and Liwei Wang. Collect at once, use effectively: Making non-interactive locally private learning possible. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 4130–4139, 2017.