Noninteractive Locally Private Learning of Linear Models via Polynomial Approximations

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

Editors: Aurélien Garivier and Satyen Kale

Abstract

Minimizing a convex risk function is the main step in many basic learning algorithms. We study protocols for convex optimization which provably leak very little about the individual data points that constitute the loss function. Specifically, we consider differentially private algorithms that operate in the local model, where each data record is stored on a separate user device and randomization is performed locally by those devices. We give new protocols for *noninteractive* LDP convex optimization—i.e., protocols that require only a single randomized report from each user to an untrusted aggregator.

We study our algorithms' performance with respect to expected loss—either over the data set at hand (empirical risk) or a larger population from which our data set is assumed to be drawn. Our error bounds depend on the form of individuals' contribution to the expected loss. For the case of *generalized linear losses* (such as hinge and logistic losses), we give an LDP algorithm whose sample complexity is only linear in the dimensionality p and quasipolynomial in other terms (the privacy parameters ϵ and δ, and the desired excess risk α). This is the first algorithm for nonsmooth losses with sub-exponential dependence on p.

For the Euclidean median problem, where the loss is given by the Euclidean distance to a given data point, we give a protocol whose sample complexity grows quasipolynomially in p. This is the first protocol with sub-exponential dependence on p for a loss that is not a generalized linear loss .

Our result for the hinge loss is based on a technique, dubbed polynomial of inner product approximation, which may be applicable to other problems. Our results for generalized linear losses and the Euclidean median are based on new reductions to the case of hinge loss. [1]

Keywords: Differential Privacy, Empirical Risk Minimization, Round Complexity

[1] Extended Abstract. Full version appears at https://arxiv.org/abs/1812.06825

© 2019 D. Wang, A. Smith & J. Xu.

Introduction

In this paper, we study differentially private convex risk minimization via noninteractive, locally differentially private (LDP) protocols.

Differential privacy in the local model (??). Consider n players with each holding a private data record x_i in a data universe $\mathcal{D}$, and a server coordinating the protocol. An LDP protocol executes for some number T of *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. For simplicity, we only consider protocols where each player participates in only one subset.

Definition 1 *An algorithm Q is (ϵ, δ)-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*

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

A multi-player protocol is (ϵ, δ)-LDP if for all players i, for all possible inputs and behaviors of the server (and the other players), the transcript of player i's interaction with the server is (ϵ, δ)-LDP. If $T = 1$, we say that the protocol is noninteractive.

? gave a separation between interactive and noninteractive protocols. Specifically, they showed that there is a concept class, similarity to parity, which can be efficiently learned by interactive algorithms but which requires sample size exponential in the dimension to be learned by noninteractive local algorithms.

Convex risk minimization 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 $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\} \in \mathcal{D}^n$ 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 settings, such as mean estimation, there may not be separate labels). 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 measured at a particular output w_{priv}: The empirical risk,

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

and the 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 is to design noninteractive LDP protocols that minimize the empirical and/or population excess risks. Alternatively, we can express our goal this problem in terms of *sample complexity*: find the smallest n for which we can design protocols that achieve error at most α (in the worst case over data sets, or over generating distributions, depending on how we measure risk).

? first considered worst-case error bounds for LDP convex optimization. For 1-Lipchitz convex losses over a bounded constraint set, they gave a highly interactive SGD-based protocol with sample complexity $n = O(p/\epsilon^2\alpha^2)$; moreover, they showed that no LDP protocol which interacts with each player only once can achieve asymptotically better sample complexity, even for linear losses.

? considered the round complexity of LDP protocols for convex optimization. They observed that known methods perform poorly when constrained to be run noninteractively. They gave new protocols that improved on the state of the art but nevertheless required sample complexity exponential in p. Specifically, they showed:

Theorem 2 (?) *Under the assumptions above, there is a noninteractive ϵ-LDP algorithm that for all distribution $\mathcal{P}$ on $\mathcal{D}$, with probability $1-\beta$, returns a solution with population error at most α as long as $n = \tilde{O}(c^p \log(1/\beta)/\epsilon^2\alpha^{p+1})$, where c is an absolute constant. A similar result holds for empirical risk Err_D.*

Furthermore, lower bounds on the parallel query complexity of stochastic optimization (e.g., **??**) mean that, for natural classes of LDP optimization protocols (based on measuring noisy gradients), the exponential dependence of the sample size on the dimension p (in the terms of $\alpha^{-(p+1)}$ and c^p) is, in general, unavoidable (**?**).

This situation is challenging: when the dimensionality p is high, the sample complexity (at least $\alpha^{-(p+1)}$) is enormous even for a very modest target error. However, several results have already shown that for some specific loss functions, the exponential dependency on the dimensionality can be avoided. For example, **?** show that, in the case of linear regression, there is a noninteractive (ϵ,δ)-LDP algorithm[2] with expected empirical error α and sample complexity $n = \tilde{O}(p\epsilon^{-2}\alpha^{-2})$. This indicates that there is a gap between the general case and what is achievable for some specific, commonly used loss functions.

Our Contributions The results above motivate the following basic question:

> *Are there natural conditions on the loss function which allow for noninteractive ϵ-LDP algorithms with sample complexity growing sub-exponentially (ideally, polynomially or even linearly) on the dimensionality p?*

To answer this question, we first consider the case of hinge loss functions, which are "plus functions" of an inner product: $\ell(w;x,y) = [y\langle w,x\rangle]_+$ where $[a]_+ = \max\{0,a\}$. Hinge loss arises, for example, when fitting support vector machines. We construct our noninteractive LDP algorithm by using Chebyshev polynomials to approximate the loss's derivative after smoothing. Players randomize their inputs by randomizing the coefficients of a polynomial approximation. The aggregator uses the noisy reports to provide biased gradient estimates when running Stochastic Inexact Gradient Descent (**?**).

We show that a variant of the same algorithm can be applied to convex, 1-Lipschitz generalized linear loss function, any loss function where each records's contribution has the form $\ell(w;x,y) = f(y_i\langle w,x_i\rangle)$ for some 1-Lipschitz convex function f.

Our algorithm has sample complexity that depends only linearly, instead of exponentially, on the dimensionality p and quasipolynomially on α,ϵ and $\log(1/\delta)$. The protocol exploits the fact that any 1-dimensional 1-Lipschitz convex function can be expressed as a convex combination of linear functions and hinge loss functions. We state its properties succinctly:

[2] Note that these two results are for noninteractive (ϵ,δ)-LDP, a variant of ϵ-LDP. We omit quasipolynomial terms related to $\log(1/\delta)$ in this paper.

Theorem 3 *For any $\epsilon, \delta, \alpha \in (0, 1/2)$, there is a noninteractive local (ϵ, δ)-differentially private algorithm that, to achieve expected empirical (resp., population) error α in the worst case over data sets (resp., distributions) and* 1-Lipschitz, convex generalized linear loss functions*, requires sample size $n = \tilde{O}(p \cdot \frac{d^d}{\epsilon^d})$ (where the $\tilde{O}$ notation hides factors quasipolynomial in $\log(1/\delta)$), where $d = c\log(1/\alpha)$ for some constant $c > 0$.*

We also apply our method to other loss functions. In particular, we show that in the *Euclidean median problem*, where the loss function is the ℓ_2 norm $L(w; D) = \frac{1}{2n}\sum_{i=1}^{n} \|w - x_i\|_2$, the sample complexity is only quasipolynomial in $p, \alpha, \delta, \epsilon$. This is the first noninteractive LDP protocol with sub-exponential dependence on p for a natural loss function that is not a generalized linear loss. Our result is based on the observation that the ℓ_2 norm function can be approximated by a convex combination of appropriately-scaled hinge losses. We obtain:

Theorem 4 *For any $\epsilon, \delta, \alpha \in (0, 1/2)$, there is a noninteractive local (ϵ, δ)-differentially private algorithm that, to achieve expected empirical (resp., population) error α* for the Euclidean median problem *in the worst case over data sets (resp., distributions), requires sample size $n = \tilde{O}(\frac{d^d}{\epsilon^d})$ where $d = c\log(C/\alpha)$ for some constant $c > 0$, $C = \frac{4\sqrt{\pi}p\Gamma(\frac{p-1}{2}+1)}{2\Gamma(\frac{p}{2}+1)} = O(\sqrt{p})$, and $\tilde{O}(\cdot)$ hides factors quasipolynomial in* $\log(1/\delta)$.

Related Work

Differentially private convex optimization, first formulated by **?** and **?**, has been the focus of an active line of work for the past decade. We discuss here only those results which are related to the local model.

? initiated the study of learning under local differential privacy. Specifically, they showed a general equivalence between learning in the local model and learning in the statistical query model. **?** gave the first lower bounds for the accuracy of LDP protocols, for the special case of counting queries (equivalently, binomial parameter estimation). The general problem of LDP convex risk minimization was first studied by **?**, which provided tight upper and lower bounds for a range of settings. Subsequent work considered a range of statistical problems in the LDP setting, providing upper and lower bounds—we omit a complete list here.

? initiated the study of the round complexity of LDP convex optimization, connecting it to the parallel complexity of (nonprivate) stochastic optimization.

Convex risk minimization in the *noninteractive* LDP received considerable recent attention (**???**) (see Table 1 for details). **?** first studied the problem with general convex loss functions and showed that the exponential dependence on the dimensionality is unavoidable for a class of noninteractive algorithms. **?** demonstrated that such an exponential dependence in the term of α is avoidable if the loss function is 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 quasipolynomial in p.

The work most related to ours is that of (**?**), which also considered some specific loss functions in high dimensions, such as sparse linear regression and kernel ridge regression. They first propose a method based on Chebyshev polynomial approximation to the gradient function. Their idea is a key ingredient in our algorithms. There are still several differences. First, their analysis requires

Methods	Sample Complexity	Assumption on the Loss Function
(**?**, Claim 4)	$\tilde{O}(4^p\alpha^{-(p+2)}\epsilon^{-2})$	1-Lipschitz
(**?**, Theorem 10)	$\tilde{O}(2^p\alpha^{-(p+1)}\epsilon^{-2})$	1-Lipschitz and Convex
?	$\Theta(p\epsilon^{-2}\alpha^{-2})$	Linear Regression
?	$\tilde{O}\big((cp^{\frac{1}{4}})^p\alpha^{-(2+\frac{p}{2})}\epsilon^{-2}\big)$	$(8, T)$-smooth
?	$\tilde{O}(4^{p(p+1)}D_p^2\epsilon^{-2}\alpha^{-4})$	(∞, T)-smooth
?	$p\cdot\big(\frac{1}{\alpha}\big)^{O(\log\log(1/\alpha)+\log(1/\epsilon))}$	Smooth Generalized Linear
This Paper	$p\cdot\big(\frac{1}{\alpha}\big)^{O(\log\log(1/\alpha)+\log(1/\epsilon))}$	1-Lipschitz Convex Generalized Linear
This Paper	$\Big(\frac{\sqrt{p}}{\alpha}\Big)^{O(\log\log(\sqrt{p}/\alpha)+\log(1/\epsilon))}$	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 $\|\mathcal{C}\|_2 \leq 1$. Asymptotic statements assume $\epsilon, \delta, \alpha \in (0, 1/2)$ and ignore quasipolynomial dependencies on $\log(1/\delta)$.

additional assumptions on the loss function, such as smoothness and boundedness of higher order derivatives, which are not satisfied by the hinge loss. In contrast, our approach applies to any convex, 1-Lipschitz generalized linear loss. Second, we introduce a novel argument to ”lift” our hinge loss algorithms to more general linear losses and the Euclidean median.

We defer proofs and more detailed descriptions to the online full version.

Acknowledgments

D.W. and J.X. were supported in part by NSF through grants CCF-1422324 and CCF-1716400. A.S. was supported by NSF awards IIS-1447700 and AF-1763786 and a Sloan Foundation Research Award. Part of this research was done while D.W. was visiting Boston University and Harvard University’s Privacy Tools Project.