Estimating Smooth GLM in Non-interactive Local Differential Privacy Model with Public Unlabeled Data

Di Wang[*][†] DI.WANG@KAUST.EDU.SA
CEMSE
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

Huanyu Zhang HZ388@CORNELL.EDU
School of Electrical and Computer Engineering
Cornell University
Ithaca, NY

Marco Gaboardi GABOARDI@BU.EDU
Department of Computer Science
Boston University
Boston, MA

Jinhui Xu JINHUI@BUFFALO.EDU
Department of Computer Science and Engineering
State University of New York at Buffalo
Buffalo, NY

Editors: Vitaly Feldman, Katrina Ligett and Sivan Sabato

Abstract

In this paper, we study the problem of estimating smooth Generalized Linear Models (GLM) in the Non-interactive Local Differential Privacy (NLDP) model. Different from its classical setting, our model allows the server to access some additional public but unlabeled data. By using Stein's lemma and its variants, we first show that there is an (ϵ, δ)-NLDP algorithm for GLM (under some mild assumptions), if each data record is i.i.d sampled from some sub-Gaussian distribution with bounded ℓ_1-norm. Then with high probability, the sample complexity of the public and private data, for the algorithm to achieve an α estimation error (in ℓ_∞-norm), is $O(p^2\alpha^{-2})$ and $O(p^2\alpha^{-2}\epsilon^{-2})$, respectively, if α is not too small (*i.e.,* $\alpha \geq \Omega(\frac{1}{\sqrt{p}})$), where p is the dimensionality of the data. This is a significant improvement over the previously known exponential or quasi-polynomial in α^{-1}, or exponential in p sample complexity of GLM with no public data. We then extend our idea to the non-linear regression problem and show a similar phenomenon for it. Finally, we demonstrate the effectiveness of our algorithms through experiments on both synthetic and real world datasets. To our best knowledge, this is the first paper showing the existence of efficient and effective algorithms for GLM and non-linear regression in the NLDP model with public unlabeled data.

Keywords: Differential Privacy, Generalized Linear Model, Local Differential Privacy

[*] The first two authors contributed equally to this paper.
[†] Extended abstract. Full version appears as (Wang et al., 2019b).

1. Introduction

Generalized Linear Model (GLM) is one of the most fundamental models in statistics and machine learning. It generalizes ordinary linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. GLM was introduced as a way of unifying various statistical models, including linear, logistic and Poisson regressions. It has a wide range of applications in various domains, such as social sciences (Warne, 2017), genomics research (Takada et al., 2017), finance (McNeil and Wendin, 2007) and medical research (Lindsey and Jones, 1998). The model can be formulated as follows.

GLM: Let $y \in [0, 1]$ be the response variable that belongs to an exponential family with natural parameter η. That is, its probability density function can be written as $p(y|\eta) = \exp(\eta y - \Phi(\eta))h(y)$, where Φ is the *cumulative generating function*. Given observations $y_1, \cdots, y_n$ such that $y_i \sim p(y_i|\eta_i)$ for $\eta = (\eta_1, \cdots, \eta_n)$, the maximum likelihood estimator (MLE) can be written as $p(y_1, y_2, \cdots|\eta) = \exp(\sum_{i=1}^{n} y_i\eta_i - \Phi(\eta_i))\Pi_{i=1}^{n} h(y_i)$. In GLM, we assume that η is modeled by linear relations, *i.e.,* $\eta_i = \langle x_i, w^*\rangle$ for some $w^* \in \mathbb{R}^p$ and feature vector x_i. Thus, maximizing MLE is equivalent to minimizing $\frac{1}{n}\sum_{i=1}^{n}[\Phi(\langle x_i, w\rangle) - y_i\langle x_i, w\rangle]$. The goal is to find w^*, which is equivalent to minimizing its population version

$$w^* = \arg\min_{w\in\mathbb{R}^p} \mathbb{E}_{(x,y)}[\Phi(\langle x, w\rangle) - y\langle x, w\rangle].$$

One often encountered challenge for using GLM in real world applications is how to handle sensitive data, such as those in social science and medical research. As a commonly-accepted approach for preserving privacy, Differential Privacy (DP) (Dwork et al., 2006) provides provable protection against identification and is resilient to arbitrary auxiliary information that might be available to attackers.

As a popular way of achieving DP, Local Differential Privacy (LDP) has received considerable attentions in recent years and been adopted in industry (Ding et al., 2017; Erlingsson et al., 2014; Tang et al., 2017). In LDP, each individual manages his/her proper data and discloses them to a server through some DP mechanisms. The server collects the (now private) data of each individual and combines them into a resulting data analysis. Information exchange between the server and each individual could be either only once or multiple times. Correspondingly, protocols for LDP are called non-interactive LDP (NLDP) or interactive LDP. Due to its ease of implementation (*e.g.* no need to deal with the network latency problem), NLDP is often preferred in practice. Formally LDP is defined as following.

Definition 1 (Local Differential Privacy (Kasiviswanathan et al., 2011)) *A randomized 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*

$$\mathbb{P}(Q(x) \in E) \leq e^{\epsilon}\mathbb{P}(Q(x') \in E) + \delta.$$

When $\delta = 0$, $\mathcal{A}$ is ϵ-LDP. A multi-player protocol is $(\epsilon, \delta)/\epsilon$-LDP if for all possible inputs and runs of the protocol, the transcript of player i's interaction with the server is $(\epsilon, \delta)/\epsilon$-LDP. If $T = 1$, we say that the protocol is $(\epsilon, \delta)/\epsilon$ ***non-interactive LDP (NLDP)****.*

While there are many results on GLM in the DP and interactive LDP models (Chaudhuri et al., 2011; Bassily et al., 2014; Jain and Thakurta, 2014; Kasiviswanathan and Jin, 2016), GLM in NLDP is still not well understood due to the limitation of the model. (Smith et al., 2017; Wang et al., 2018; Zheng et al., 2017) and (Wang et al., 2019a) comprehensively studied this problem. However, all of these results are on the negative side. More specifically, they showed that to achieve an error of α, the sample complexity needs to be quasi-polynomial or exponential in α^{-1} (based on different assumptions) (Wang et al., 2019a; Zheng et al., 2017; Wang et al., 2020) or exponential in the dimensionality p (Smith et al., 2017; Wang et al., 2018) (see Related Work section for more details). Recently, (Dagan and Feldman, 2020) showed that an exponential lower bound (either in p or α^{-1}) on the number of samples for solving the standard task of learning a large-margin linear separator in the NLDP model. Due to these negative results, there is no study on the practical performance of these algorithms.

To address this high sample complexity issue of NLDP, a possible way is to make use of some recent developments on the central DP model. Quite a few results (Bassily and Nandi, 2019; Hamm et al., 2016; Papernot et al., 2016, 2018; Bassily et al., 2018) have suggested that by allowing the server to access some public but unlabeled data in addition to the private data, it is possible to reduce the sample complexity in the central DP model, under the assumption that these public data have the same marginal distribution as the private ones. It has also shown that such a relaxed setting is likely to enable better practical performance for problems like Empirical Risk Minimization (ERM) (Hamm et al., 2016; Papernot et al., 2016). Thus, it would be interesting to know whether the relaxed setting can also help reduce sample complexity in the NLDP model.

With this thinking, our main questions now become the follows. **Can we further reduce the sample complexity of GLM in the NLDP model if the server has additional public but unlabeled data? Moreover, is there any efficient algorithm for this problem in the relaxed setting?**

In this paper, we provide positive answers to the above two questions. Specifically, we focus on the GLM estimation problem in the NLDP model with public but unlabeled data:

Our Model: Different from the above classical NLDP model where only one private dataset $\{(x_i, y_i)\}_{i=1}^n$ exists, the NLDP model in our setting allows the server to have an additional public but unlabeled dataset $D' = \{x_j\}_{j=n+1}^{n+m} \subset \mathcal{X}^m$, where each x_j is sampled from $\mathcal{P}_x$, which is the marginal distribution of $\mathcal{P}$ (*i.e.,* they have the same distribution as $\{x_i\}_{i=1}^n$). Our contributions can be summarized as follows:

1. We first show that when the feature vector x of GLM is sub-Gaussian with bounded ℓ_1-norm, there is an (ϵ, δ)-NLDP algorithm for GLM (under some mild assumptions) whose sample complexities of the private and public data, for achieving an error of α (in ℓ_∞-norm), are $O(p^2\epsilon^{-2}\alpha^{-2})$ and $O(p^2\alpha^{-2})$ (with other terms omitted), respectively, if α is not too small (*i.e.,* $\alpha \geq \Omega(\frac{1}{\sqrt{p}})$). Specifically, we show that

   Theorem 2 *Under some assumption of the data distribution and the loss function, for sufficiently large m, n such that*

$$m \geq \Omega\big(\|\Sigma\|_2\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\}\rho_2\rho_\infty^2 p^2\big),$$
$$n \geq \Omega\big(\frac{\rho_2\rho_\infty^2\|\Sigma\|_2^2 p^2\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \log\frac{1}{\delta}\log\frac{p}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}}\big),$$

there is an (ϵ, δ)-NLDP algorithm whose output $\hat{w}^{glm}$ satisfies the following with probability at least $1 - \exp(-\Omega(p)) - \xi$,

$$\|\hat{w}^{glm} - w^*\|_\infty \leq O\Big(\frac{\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\}p}{\sqrt{m}} + \frac{\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\}p\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\sqrt{n}} + \frac{\|w^*\|_\infty^3 \max\{1, \|w^*\|_\infty\}}{\sqrt{p}}\Big),$$

where $G, L, \tau, M, \bar{c}, r, \kappa_x, \frac{1}{c_\Phi}$ are assumed to be $O(1)$ and thus omitted in the Big-O notations.

We note that this is the first result that achieves a **fully polynomial** sample complexity for a general class of loss functions in the NLDP model with public unlabeled data. We also provide several examples that satisfy those assumptions.

2. We then extend our idea to the non-linear regression problem. We assume that there is an underlying vector $w^* \in \mathbb{R}^p$ with $\|w^*\|_2 \leq 1$ such that

$$y = f(\langle x, w^* \rangle) + \sigma,$$

where x is the feature vector sampled from some distribution (for simplicity, we assume that the mean is zero) and y is the response. σ is the zero-mean noise which is independent of x and bounded by some constant $C = O(1)$ (*i.e.,* $\sigma \in [-C, C]$). f is some known differentiable link function with $f(0) \neq \infty$ By using the zero-bias transformation (Goldstein et al., 1997), we show that when x is sub-Gaussian with bounded ℓ_1-norm, it exhibits the same phenomenon as GLM.

3. Finally, we provide an experimental study of our algorithms on both synthetic and real world datasets. The experimental results suggest that our methods are efficient and effective, which is consistent with our theoretical analysis.

2. Related Work

Private learning with public unlabeled data has been studied previously in (Hamm et al., 2016; Papernot et al., 2016, 2018; Bassily et al., 2018). These results differ from ours in quite a few ways. Firstly, all of them consider either the multiparty setting or the centralized model. Consequently, none of them can be used to solve our problems. Specifically, (Hamm et al., 2016) considered the multiparty setting where each party possesses several data records and each party uses their data to get a classifier, however, this approach could not be extended to local DP model since now each party only has just one data point and it is impossible to get any useful classifier based one data point. (Papernot et al., 2016, 2018) investigated the DP model, used sub-sample and aggregate to train some deep learning models, but provided no provable sample complexity. (Bassily et al., 2018) also studied the DP model by combining the distance to instability and the sparse vector techniques, and showed some theoretical guarantees. However, both the sub-sample/aggregate and the sparse vector methods cannot be used in the NLDP model. Moreover, public data in their methods are also used quite differently from ours. Secondly, all of the above results use the private data to label the public data and conduct the learning process on the public data, while we use the public data to approximate some crucial constants. Finally, all of the previous methods rely on the known model or loss functions, while in our algorithms the loss functions could be unknown to the users; also the server could use multiple loss functions with the same sample complexity.

The problems considered in this paper can be viewed as restricted versions of the general ERM problem in the NLDP model. Due to its challenging nature, ERM in NLDP has only been considered in a few papers, such as (Smith et al., 2017; Wang et al., 2018, 2019a; Zheng et al., 2017; Daniely and Feldman, 2018; Wang and Xu, 2019). (Smith et al., 2017) gave the first result on convex ERM in NLDP and provided an algorithm with a sample complexity of $O(2^p\alpha^{-(p+1)}\epsilon^{-2})$. They showed that the exponential dependency on the dimensionality p is not avoidable for general loss functions. Later, (Wang et al., 2018) showed that when the loss function is smooth enough, the exponential term of $\alpha^{-\Omega(p)}$ can be reduced to polynomial, but not the other exponential terms. Recently, (Wang et al., 2019a; **?**) further showed that the sample complexity for any 1-Lipschitz convex GLM can be reduced to linear in p and exponential in α^{-1}, which extends the work in (Zheng et al., 2017), whose sample complexity is linear in p and quasi-polynomial in α^{-1} for smooth GLM. In this paper, we show, for the first time, that the sample complexity of GLM can be reduced to fully polynomial with the help of some public but unlabeled data and some mild assumptions on GLM. There are also works on some special loss functions. For example, (Wang and Xu, 2019, 2021) studied the high dimensional sparse linear regression problem and (Daniely and Feldman, 2018) considered the problem of learning halfspaces with polynomial samples. Since these results are only for some special loss functions (instead of a family of functions), they are incomparable with ours.

Acknowledgments

Di Wang was support in part by the baseline funding of King Abdullah University of Science and Technology (KAUST). Huanyu Zhang supported was in part by the National Science Foundation (NSF) under Grant No. 1815893 and 1704443. Jinhui Xu was supported in part by the National Science Foundation (NSF) under Grant No. CCF-1716400 and IIS-1919492. Part of the work was done when Di Wang and Marco Gaboardi were visiting the Simons Institute of the Theory for Computing.

References

Raef Bassily and Anupama Nandi. Privately answering classification queries in the agnostic pac model. *arXiv preprint arXiv:1907.13553*, 2019.

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

Raef Bassily, Abhradeep Guha Thakurta, and Om Dipakbhai Thakkar. Model-agnostic private learning. In *Advances in Neural Information Processing Systems*, pages 7102–7112, 2018.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

Yuval Dagan and Vitaly Feldman. Interaction is necessary for distributed learning with privacy or communication constraints. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 450–462, 2020.

Amit Daniely and Vitaly Feldman. Learning without interaction requires separation. *arXiv preprint arXiv:1809.09165*, 2018.

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

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

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.

Larry Goldstein, Gesine Reinert, et al. Stein's method and the zero bias transformation with application to simple random sampling. *The Annals of Applied Probability*, 7(4):935–952, 1997.

Jihun Hamm, Yingjun Cao, and Mikhail Belkin. Learning privately from multiparty data. In *International Conference on Machine Learning*, pages 555–563, 2016.

Prateek Jain and Abhradeep Guha Thakurta. (near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning*, pages 476–484, 2014.

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.

James K Lindsey and Bradley Jones. Choosing among generalized linear models applied to medical data. *Statistics in medicine*, 17(1):59–68, 1998.

Alexander J McNeil and Jonathan P Wendin. Bayesian inference for generalized linear mixed models of portfolio credit risk. *Journal of Empirical Finance*, 14(2):131–149, 2007.

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

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018.

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

Yasuaki Takada, Ryutaro Miyagi, Aya Takahashi, Toshinori Endo, and Naoki Osada. A generalized linear model for decomposing cis-regulatory, parent-of-origin, and maternal effects on allele-specific gene expression. *G3: Genes, Genomes, Genetics*, 7(7):2227–2234, 2017.

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.

Di Wang and Jinhui Xu. On sparse linear regression in the local differential privacy model. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, California, USA, June 9-15, 2019*, 2019.

Di Wang and Jinhui Xu. On sparse linear regression in the local differential privacy model. *IEEE Trans. Inf. Theory*, 67(2):1182–1200, 2021.

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

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

Di Wang, Huanyu Zhang, Marco Gaboardi, and Jinhui Xu. Estimating smooth glm in non-interactive local differential privacy model with public unlabeled data. *arXiv preprint arXiv:1910.00482*, 2019b.

Di Wang, Marco Gaboardi, Adam Smith, and Jinhui Xu. Empirical risk minimization in the non-interactive local model of differential privacy. *J. Mach. Learn. Res.*, 21:200:1–200:39, 2020.

Russell T. Warne. *Statistics for the Social Sciences: A General Linear Model Approach*. Cambridge University Press, 2017. doi: 10.1017/9781316442715.

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-Volume 70*, pages 4130–4139. JMLR. org, 2017.