Dissertation Plan

Algorithms for Privacy and Representations in the Generalized Linear Model

Daniel Alabi*

The use of statistical methods by social scientists is very common. Often, these methods are most useful when the computed statistics can be released to researchers and policymakers to inform important decision-making. Naively releasing these statistics raises concerns about the disclosure of private information about individuals in the dataset. The possibility of such attacks led to *differential privacy*, a rigorous approach to quantifying privacy loss, introduced by Dwork, McSherry, Nissim, and Smith [DMNS06]. Statistical estimators, such as regression estimators, can be made to satisfy this rigorous, worst-case notion of privacy. A common approach for such activity is to reduce the general estimation problem to other problems (e.g., median estimation or least squares minimization). As a result, however, the utility of such estimators would depend on the reduction used and the underlying differentially private mechanism. We aim to quantify the utility of such estimators for solving problems in the generalized linear model.

One of the most widely used and studied statistical methods available to scientists is linear regression. Its power for predictive analysis in myriad domains is well known [BLLT20, BMR21, BHMM19]. Unfortunately, privacy-preserving solutions for this problem are not well established. For example, the Opportunity Insights Lab at Harvard have data from census tracts across the United States [CF19]. One of the use cases of the Opportunity Atlas tool developed by their lab is to predict social and income mobility via univariate regressions on small census tracts. For each tract and demographic group, data about parent and child national income ranks are collected. Then a single-independent-variable univariate regression slope $\hat{\alpha}$ can be computed as a measure of economic mobility. Opportunity Insights [CF19] provided a practical method – which they term the "Maximum Observed Sensitivity" (MOS) algorithm – to reduce privacy loss of their released estimates but their method, so far, is not formally private.

We are thus motivated by the following question:

> *Can we release regression statistics, methods, or procedures that satisfy the formal guarantees of privacy (such as differential privacy) and simultaneously have good utility?*

We have, thus far, answered this question affirmatively by providing some theoretical and empirical results for single-independent-variable univariate regression point estimates [AMS+20]. These results are a first step towards providing formally private and practical releases that could be used by economics and social science labs. **Most notably, our work – partially funded by the U.S. Census Bureau – has produced high-utility methods for analyzing data from the 2020 United States Census.** Since we begun our research, we have developed methods based

*School of Engineering and Applied Sciences, Harvard University. Email: alabid@g.harvard.edu.

on robust statistics that sometimes match or exceed the utility guarantees of the MOS algorithm on small datasets. Much work remains to be done and some of the research questions we hope to explore are:

1. **Convergence Rates**: What properties of the dataset are required for point or interval differentially private estimators to converge – in terms of either computational and statistical efficiency – to the non-private estimators?

2. **Distributed Settings**: In the cases where the dataset is horizontally or vertically distributed amongst several parties, what privacy-preserving solutions exist for linear regression models and what can we say about the utility of these models?

3. **High-Dimensional Settings**: On high-dimensional datasets, can we still provide high-utility private linear regression models?

4. **Multi-Objective Settings**: In the case where, in addition to average accuracy, other objectives are to be optimized, what statistical and computational guarantees can we provide?

The questions above will be approached methodologically via empirical and theoretical analysis. The theoretical analysis will be aimed at providing both upper bounds and lower bounds on the statistical and computational complexity of private linear regression.

Thesis Committee

1. Boaz Barak.

2. Cynthia Dwork.

3. Gary King.

4. Advisor: Salil Vadhan.

Convergence Rates

On the theoretical side, we hope to explore the convergence (in the finite sample regime) rates of these estimators and the parameter settings that are most helpful for faster convergence. What properties of the dataset ensure faster convergence? And does the convergence behaviour differ significantly in the finite-sample and central-limit regimes? In addition, we will explore the release of private confidence intervals (instead of point estimates) for certain estimators. Previous work in this space made assumptions on the underlying data generating process (e.g., normal assumptions on the independent variables made by [She19, KV17]). We will aim to make minimal assumptions on the independent variables. Since the current landscape of the literature on differentially private linear regression is focused on prediction rather than explanation (or establishing correlations), we will study the utility of estimators for use in hypothesis testing [AV20a]. In particular, we will establish connections between testing and estimation for linear regression, delineating what conditions are suitable for statistically efficient tests. Because of the tradeoffs (i.e., space, memory, time) between different estimators, we will design estimators that satisfy certain constraints (e.g., bounded memory or general approximations to function classes [ABX08, KM21]). For example, in

[ABEC21], we design bounded-space estimators for (single or all) quantile estimation, a fundamental operation for robust estimators.

We show in [AMS+20] that there exist differentially private algorithms for simple linear regression that perform well on small datasets (size ranging from small tens to small hundreds). Some datasets analyzed in [AMS+20] were obtained from the Opportunity Insights team and are of the form $\{(x_i, y_i)\}_{i=1}^n \in [0,1]^n \times [0,1]^n$ where $\mathbf{x} = (x_1, \ldots, x_n)^T$ is the independent variable in one dimension. The accuracy is measured in terms of the empirical and population error of the point estimators. Differentially private analogues of both robust and non-robust estimators are presented in [AMS+20] and the estimators are compared against one another. [1] An advantage of the robust estimators in [AMS+20] (e.g., median-based estimators) is that they work well on small-sized datasets whereas a disadvantage is that they're generally computationally inefficient to optimize whereas non-robust estimators that optimize the ordinary least squares (OLS) objective are generally more efficient and provably run in linear time. As a result, we have theoretically compared non-robust estimators against one another and will perform an accompanying empirical comparison of these non-robust estimators for the ordinary least squares objective [AV20b]. Furthermore, we hope to empirically tease out what properties of the data distribution enable certain estimators to perform well. Through our experiments, we hope to compare the robust and non-robust estimators in terms of both statistical and computational performance. In [AMS+20], the main application is for use in the release of public data from for the Opportunity Insights team. We will explore use by other social science and economics labs. In addition, for future work, we wish to explore the use of Bayesian inference methods for differentially private linear regression.

Distributed Settings

We will also explore the guarantees of privacy and utility in settings where the data is either vertically distributed (columns split amongst several parties) or horizontally distributed (rows split amongst parties). [GSB+17] explore privacy-preserving vertically distributed linear regression using hybrid multi-party protocols. We will explore achieving their results using differential privacy instead and compare the resulting utility guarantees from both works. We will also extend the work to horizontally distributed linear regression. Preliminary work on distributed linear regression via first-order optimization is presented in [STU17]. For a fixed design matrix, [DJW13] also consider information-theoretic limits of imposing local differential privacy for distributed linear regression. Extending this work to general design matrices (and other forms of statistical inference) is important future work.

High-Dimensional Settings

The study of sparse linear regression with differential privacy has been studied in the local setting [WX19]. We will explore this model in both the central and local models and perhaps derive lower bounds based on previous work [CWZ19, CWZ20, BS15].

So far, our experimental work on differentially private linear regression [AMS+20] has focused on the one-dimensional case where there is a single independent variable. In many settings, there are multiple independent variables (for example, the work of [Wan18] where datasets have at

[1] Informally, robust estimators generally refer to statistical methods that are relatively (compared to other estimators) insensitive to outliers in the dataset. We build upon previous work connecting robust estimators and differential privacy [DL09].

least 13 independent variables). We will extend our work in [AMS+20] to handle more than one independent variable. Furthermore, in some cases, the number of independent variables exceeds the number of examples in the dataset (referred to as *sparse linear regression*). Examples of this situation occur often in computational genomics where the predictors are gene-expression measurements from thousands of possible genes and there might only be a few patients in the dataset. We hope to experimentally explore private linear regression in high-dimensional settings.

Multi-Objective Settings

Ordinary least squares and, more generally, linear regression methods are one of the most used class of key techniques for statistical inference. Because of their versatility, these techniques can be embedded as sub-routines in other applications. For example, via the use of Lagrangian duality [AAW+20b], fair regression can be achieved via (polynomial-time) reductions to ordinary least squares [AIK18, ADW19]. We have achieved analogous results but with differential privacy guarantees [Ala19]. Additionally, we provided a *general framework* for multi-objective optimization with privacy guarantees via approximately solving bounded-divergence linear sub-problems. Through experiments, we also show that some formulations of multi-objective regression can be solved via differentially private linear regression and perform well statistically (with use of additional techniques) [AAW20a]. In addition, we can use bounded divergence linear optimizers for other goals such as model correction/projection and hypothesis testing.

References

[AAW20a] Julius Adebayo, Daniel Alabi, and Chris Wiggins. Prioritizing minority groups when applying differential privacy. Working paper, 2020.

[AAW+20b] Wael Alghamdi, Shahab Asoodeh, Hao Wang, Flávio P. Calmon, Dennis Wei, and Karthikeyan Natesan Ramamurthy. Model projection: Theory and applications to fair machine learning. In *IEEE International Symposium on Information Theory, ISIT 2020, Los Angeles, CA, USA, June 21-26, 2020*, pages 2711–2716. IEEE, 2020.

[ABEC21] Daniel Alabi, Omri Ben-Elezier, and Anamay Chaturvedi. Learning differentially private quantiles with bounded space. Working paper, 2021.

[ABX08] Benny Applebaum, Boaz Barak, and David Xiao. On basing lower-bounds for learning on worst-case assumptions. In *49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008, October 25-28, 2008, Philadelphia, PA, USA*, pages 211–220, 2008.

[ADW19] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 120–129, 2019.

[AIK18] Daniel Alabi, Nicole Immorlica, and Adam Kalai. Unleashing linear optimizers for group-fair learning and optimization. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, pages 2043–2066, 2018.

[Ala19] Daniel Alabi. The cost of a reductions approach to private fair optimization. *CoRR*, abs/1906.09613, 2019.

[AMS+20] Daniel Alabi, Audra McMillan, Jayshree Sarathy, Adam Smith, and Salil Vadhan. Differentially private simple linear regression. *CoRR*, abs/2007.05157, 2020.

[AV20a] Daniel Alabi and Salil Vadhan. Hypothesis testing for differentially private linear regression. Working paper, 2020.

[AV20b] Daniel Alabi and Salil Vadhan. Infinite sensitivity finite sample differentially private convergence analysis. Working paper, 2020.

[BHMM19] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[BLLT20] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[BMR21] Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *CoRR*, abs/2103.09177, 2021.

[BS15] Raef Bassily and Adam Smith. Local, private, efficient protocols for succinct histograms. In *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*, STOC '15, pages 127–135, 2015.

[CF19] Raj Chetty and John N. Friedman. A practical method to reduce privacy loss when disclosing statistics based on small samples. *American Economic Review Papers and Proceedings*, 109:414–420, 2019.

[CWZ19] T. Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *CoRR*, abs/1902.04495, 2019.

[CWZ20] T. Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy in generalized linear models: Algorithms and minimax lower bounds. *CoRR*, abs/2011.03900, 2020.

[DJW13] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pages 429–438, 2013.

[DL09] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, May 31 - June 2, 2009*, pages 371–380, 2009.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, pages 265–284, 2006.

[GSB+17] Adrià Gascón, Phillipp Schoppmann, Borja Balle, Mariana Raykova, Jack Doerner, Samee Zahur, and David Evans. Privacy-preserving distributed linear regression on high-dimensional data. *PoPETs*, 2017(4):345–364, 2017.

[KM21] Zander Kelley and Raghu Meka. Random restrictions and prgs for ptfs in gaussian space. *Electron. Colloquium Comput. Complex.*, 2021.

[KV17] Vishesh Karwa and Salil P. Vadhan. Finite sample differentially private confidence intervals. *CoRR*, abs/1711.03908, 2017.

[She19] Or Sheffet. Differentially private ordinary least squares. *J. Priv. Confidentiality*, 9(1), 2019.

[STU17] Adam D. Smith, Abhradeep Thakurta, and Jalaj Upadhyay. Is interaction necessary for distributed private learning? In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 58–77, 2017.

[Wan18] Yu-Xiang Wang. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. *CoRR*, abs/1803.02596, 2018.

[WX19] 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, 9-15 June 2019, Long Beach, California, USA*, pages 6628–6637, 2019.