Insufficient Statistics Perturbation: Stable Estimators for Private Least Squares

Gavin Brown[*] GRBROWN@CS.WASHINGTON.EDU
Jonathan Hayase[*] JHAYASE@CS.WASHINGTON.EDU
Samuel Hopkins[†] SAMHOP@MIT.EDU
Weihao Kong[‡] KWEIHAO@GMAIL.COM
Xiyang Liu[*] XIYANGL@CS.WASHINGTON.EDU
Sewoong Oh[*] SEWOONG@CS.WASHINGTON.EDU
Juan C. Perdomo[§] JCPERDOMO@G.HARVARD.EDU
Adam Smith[¶] ADS22@BU.EDU

Editors: Shipra Agrawal and Aaron Roth

We present a sample- and time-efficient differentially private algorithm for ordinary least squares, with error that depends linearly on the dimension and is independent of the condition number of $X^\top X$, where X is the design matrix.[1] Given covariates $X \in \mathbb{R}^{n\times d}$ and responses $y \in \mathbb{R}^n$, the OLS estimator is defined as

$$\beta_{\text{ols}} \;=\; \left(X^\top X\right)^{-1} X^\top y \;.$$

Among the many reasons for the popularity of OLS is the fact that it is a statistically and computationally efficient way of solving linear regression. Speaking informally, OLS has low excess error whenever the number of samples n is as large as the problem dimension d. Crucially, its statistical performance does not depend on the condition number $\kappa(X^\top X)$, the ratio between the maximum and minimum eigenvalues. Our algorithm has near-optimal accuracy guarantees and, for modest levels of privacy, introduces less error than the error due to sampling. We provide accuracy guarantees for any dataset with bounded statistical leverage and bounded residuals.

Given its widespread use in the analysis of personal data, there is a long line of work giving differentially private algorithms to approximate OLS. However, designing practical and efficient algorithms for this problem has been a particularly challenging endeavor. Existing algorithms for DP regression suffer from one of three limitations: poor dimension dependence (e.g., Wang, 2018; Sheffet, 2019), poor dependence on the condition number $\kappa(X^\top X)$ (Varshney et al., 2022), or run in exponential time (Liu et al., 2022).

Technically, we build on the approach of Brown et al. (2023) for private mean estimation, adding scaled noise to a carefully designed estimator of the empirical regression vector. Our central contribution is the design and analysis of a nonprivate subroutine that removes high-residual examples in a stable manner.

[*] Paul G. Allen School of Computer Science and Engineering, University of Washington. Part of this work was done while G.B. was at Boston University.
[†] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.
[‡] Google Research
[§] Harvard University
[¶] Department of Computer Science, Boston University.

1. Extended abstract. Full version appears as [arxiv:2404.15409, v1]

References

Gavin Brown, Samuel Hopkins, and Adam Smith. Fast, sample-efficient, affine-invariant private mean and covariance estimation for subgaussian distributions. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5578–5579. Proceedings of Machine Learning Research, 2023.

Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In *Conference on Learning Theory*, pages 1167–1246. Proceedings of Machine Learning Research, 2022.

Or Sheffet. Old techniques in differentially private linear regression. In *Algorithmic Learning Theory*, pages 789–827. PMLR, 2019.

Prateek Varshney, Abhradeep Thakurta, and Prateek Jain. (nearly) optimal private linear regression for sub-gaussian data via adaptive clipping. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1126–1166. PMLR, 02–05 Jul 2022.

Yu-Xiang Wang. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. *arXiv preprint arXiv:1803.02596*, 2018.