Paper Presentation: Easy Differentially Private Linear Regression

Caio Azevedo, Zhe Huang, Danil Savine

IASD - Privacy for Machine Learning

March 26th, 2024

1 Introduction

2 Related Work

3 Methodology

4 Critical Analysis

5 Experiments

6 Contributions

7 Conclusion

Motivation

- Previous DP linear regression methods have limitations:
  - Assume that end users can specify accurate bounds on the data
  - Assume that users can conduct extensive hyperparameter tuning that will not be counted toward the privacy budget
- Authors address these limitations:
  - Requires only privacy level and one hyperparameter

Related Work

- DP-SGD [Abadi et al., 2016] requires several hyperparameters (learning rate, noise scale, group size, gradient norm bound) tuning
- AdaSSP (adaptive prediction sufficient statistics perturbation) only releases differentially privately λ_{min} [Wang, 2018] and performs a ridge regression - non-private bounds on data and labels are required
- Boosted AdaSSP [Tang et al., 2023] later outperforms both with ensemble of weak learners in the case where bounds are chosen arbitrarily
- Liu et al. [Liu et al., 2021] introduce Propose-Test-Release step, adapted from Brown et al. [Brown et al., 2021]
- Cumings-Menon [Cumings-Menon, 2022] explore independently the use of Tukey depth for DP

Main Algorithm: TukeyEM

Procedure:

1. Partition dataset, compute OLS estimators.
2. Compute volumes of regions with different approximate Tukey depths.
3. Run Propose-Test-Release(PTR) algorithm using these volumes.
4. Apply exponential mechanism to select a model from a high-depth region if PTR check passes.

TukeyEM:

- provides practical application without needing Gaussian data assumptions.
- is (ϵ, δ)-DP.

Details(1/2)

- **Definition of volume**: Given database D, the volume of the region of points in $\mathbb{R}^d$ with approximate Tukey depth at least i in D is: $V_{i,D} = \text{vol}\left(\{y \mid y \in \mathbb{R}^d \text{ and } \widetilde{T}_D(y) \geq i\}\right)$
- **Propose-Test-Release (PTR):**
  - A privacy-preserving algorithm that checks if it is safe to release data.

Algorithm 1 PTRCheck

Input: Tukey depth region volumes V, privacy parameters ε and δ
Compute lower bound k for distance to unsafe database
if $k + \text{Lap}\left(\frac{1}{\varepsilon}\right) \geq \frac{2}{\varepsilon}\log\left(\frac{1}{2\delta}\right)$ **then**
    return True
else
    return False

Details(2/2)

- If the check passes, run the exponential mechanism to pick a point of high approximate Tukey depth from the domain of points with moderately high approximate Tukey depth.
- **Exponential Mechanism:** Employed to ensure differential privacy by:
  1. Sampling a depth with a probability weighted by the exponential of the depth times the privacy budget ϵ.
     - Computing $W_{i,D}$ for depths $i \in \{m/4, m/4+1, \ldots, m/2\}$, representing the volume of the region with exact depth i.
     - Selecting a depth i with probability proportional to $W_{i,D} \cdot \exp(\epsilon \cdot i)$.
  2. Selecting a random model uniformly from the chosen depth region.
     - For $d = 1$, drawing a sample uniformly from the interval of points at depth $\hat{i}$.
     - For $d > 1$, partitioning the space and uniformly sampling from a randomly selected partition, ensuring that each point within the volume is represented.

Critical Analysis

Merits:

- User-Friendly Algorithm
- Empirical Effectiveness
- Innovative Use of Tukey Depth

Limitations:

- Lack of Theoretical Utility Guarantees
- Dependency on the Number of Models (m)
- Potential for Model Failure

Main Results

Dataset	NonDP	AdaSSP	TukeyEM	DPSGD (tuned)
Synthetic	0.997	0.991	**0.997**	**0.997**
California	0.637	-1.285	**0.099**	0.085
Diamonds	0.907	0.216	0.307	**0.828**
Traffic	0.966	0.944	**0.965**	0.938
NBA	0.621	0.018	**0.618**	0.531
Beijing	0.702	0.209	**0.698**	0.475
Garbage	0.542	0.119	**0.534**	0.215
MLB	0.722	0.519	**0.721**	0.718

Table 1: Results of the main paper: R^2 scores for different methods and datasets.

Other Results

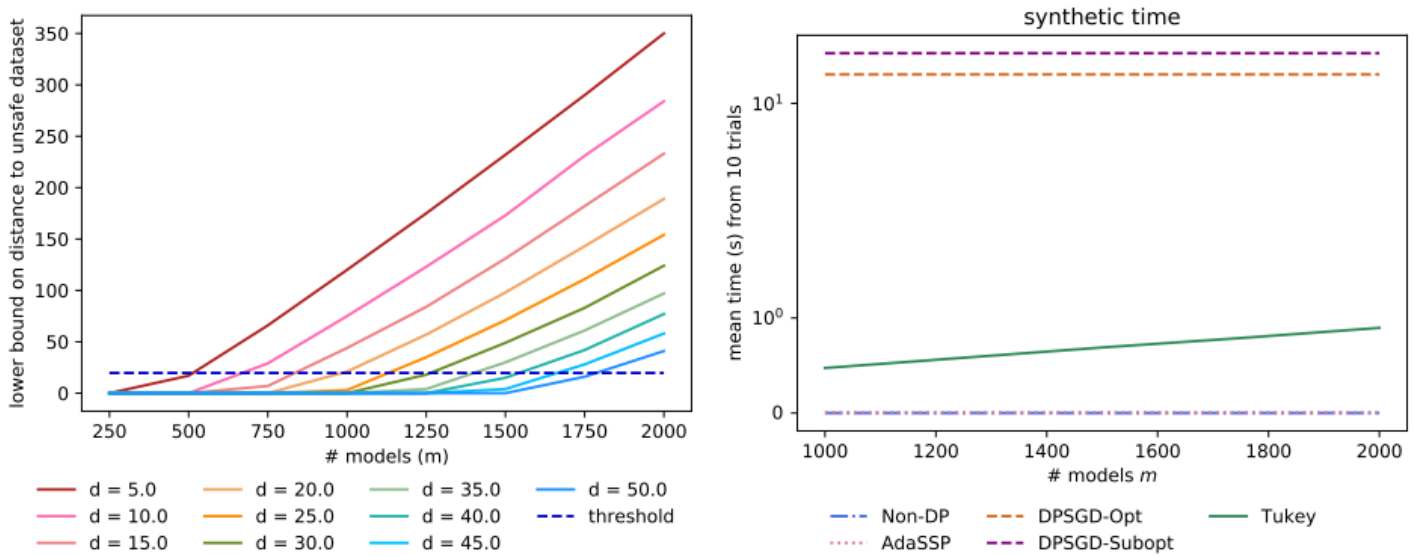

Figure 1: On the left, visualization of how number of models impacts PTR check. On the right, runtime comparison between methods.

Reproducing Results

In order to understand each step of the algorithm, we implement it ourselves and check R^2 values to see if they match the paper.

We fix $m = 1000$, $\epsilon = \log 3$, $\delta = 10^{-5}$ for our experiments.

Table 2: R^2 values for the standard OLS estimator, those reported in the paper for TukeyEM, and those in our implementation.

Dataset	**NonDP**	**TukeyEM** (paper)	**TukeyEM** (ours)
Synthetic	0.997	0.997	0.997
California	0.637	0.099	0.159
Diamonds	0.907	0.307	0.558
Traffic	0.966	0.965	0.947
NBA	0.621	0.618	0.620
Garbage	0.542	0.534	0.534
MLB	0.722	0.721	0.721

Tukey Depth Sampling

We further show that our implementation is correct by reproducing the toy example given in the paper's appendix, using the set of points $\{(1,1),(7,3),(5,7),(3,3),(5,5),(6,3)\}$.

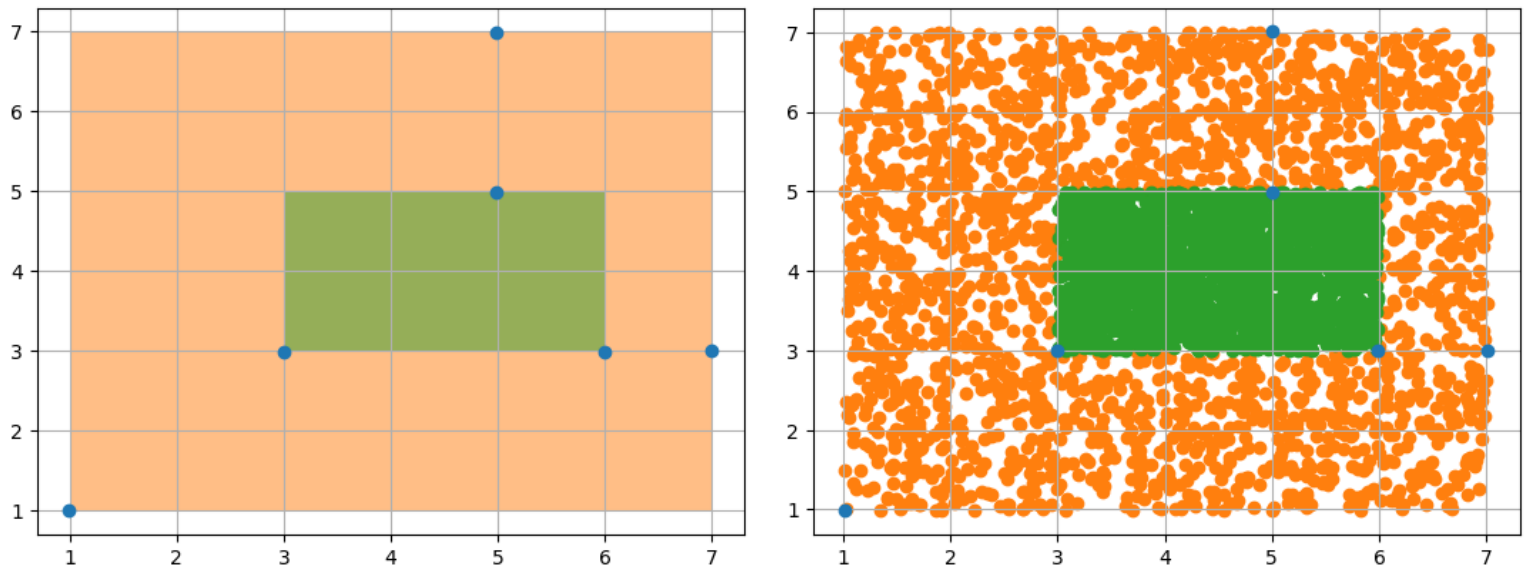

Figure 2: Illustration of approximate Tukey depth sampling.

Effect of Number of Models

As mentioned, performance can vary depending on m. We attempt to reproduce the authors' plot of median R^2 score per number of models.

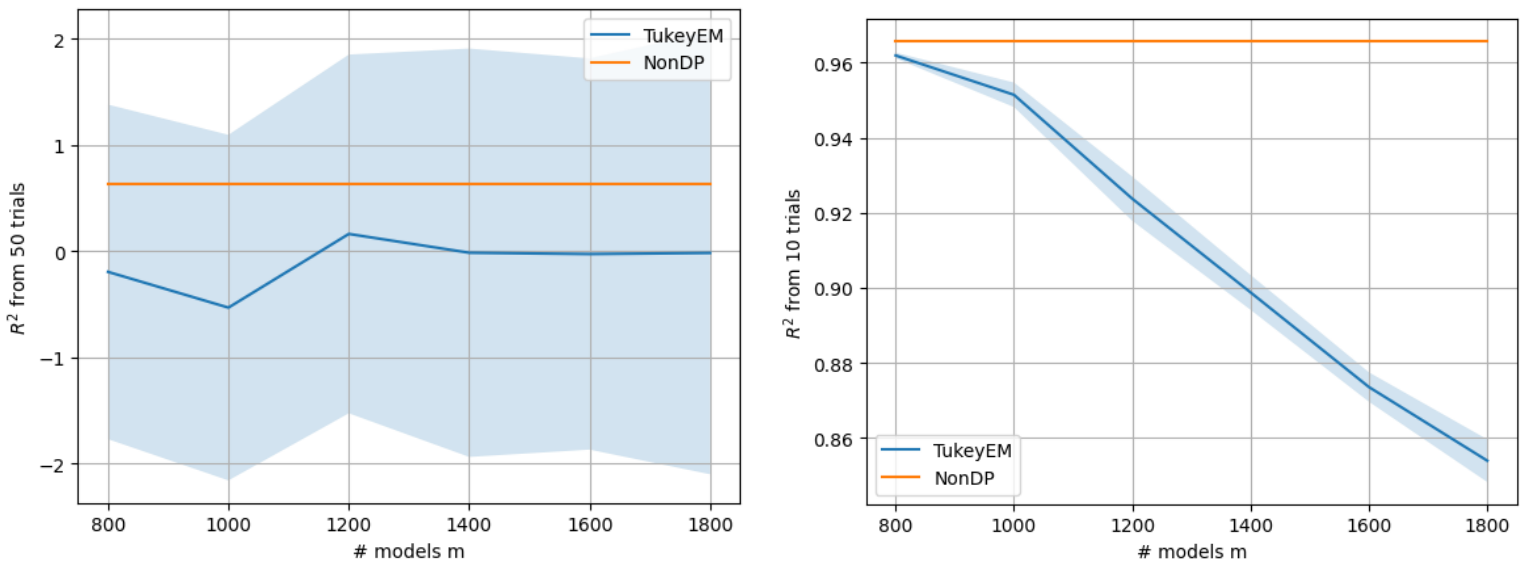

Figure 3: Dependence of R^2 score on number of models for California (left) and Traffic (right) datasets.

Extension to Logistic Regression

The authors guarantee that any method that outputs d dimensional vectors as our models can be used in TukeyEM, while keeping privacy guarantees.
We check if utility would also be kept when adapting TukeyEM to logistic regression, using a synthetic dataset.

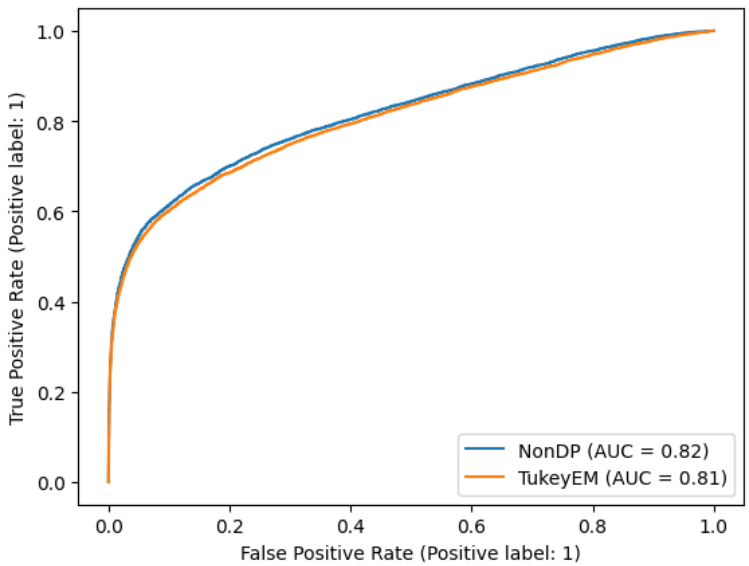

Figure 4: ROC curve for non DP (standard logistic regression) and TukeyEM logistic regression.

Conclusion

- The TukeyEM algorithm simplifies differentially private linear regression, eliminating the need for domain-specific knowledge or complex hyperparameter tuning.
- The author also demonstrated its high utility with competitive R^2 scores across various datasets.
- Integration of Tukey depth as a utility function within the exponential mechanism is a brilliant idea.
- Despite some limitations, such as the absence of theoretical utility guarantees and sensitivity to model count, its robustness in practical scenarios is commendable.

Bibliography


Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016).
Deep learning with differential privacy.
In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.


Alabi, D., McMillan, A., Sarathy, J., Smith, A., and Vadhan, S. (2022).
Differentially private simple linear regression.
Proceedings on Privacy Enhancing Technologies.


Amin, K., Joseph, M., Ribero, M., and Vassilvitskii, S. (2023).
Easy differentially private linear regression.
In *The Eleventh International Conference on Learning Representations*.


Brown, G., Gaboardi, M., Smith, A., Ullman, J., and Zakynthinou, L. (2021).
Covariance- aware private mean estimation without private covariance estimation.
Neural Information Processing Systems (NeurIPS).