Critical Analysis of: Easy Differentially Private Linear Regression

Caio Azevedo, Zhe Huang, Danil Savine

1. Introduction

Previous attempts at differentially private (DP) linear regression have serious limitations: they either require the user to provide bounds on feature and label norms so that privacy guarantees may be satisfied (Kifer et al., 2012; Dwork et al., 2014), which in turn needs extensive domain knowledge with regards to the data; or they require extensive hyperparameter tuning (Song et al., 2013; Abadi et al., 2016).

With this problem in mind, the authors propose an "easy" differentially private algorithm that does not have either of these limitations (Amin et al., 2023). The ease here refers to the experience of end users, as the only thing required is the dataset itself and the desired level of privacy. The authors further show by empirical results that the algorithm also has high utility, by testing it on several different datasets and achieving R^2 scores on par with or higher than other DP methods.

The algorithm itself, TukeyEM, is based on the exponential mechanism, using as utility function the notion of Tukey depth. After collecting several ordinary least squares (OLS) estimators from random subsets of the data, we perform a propose-test-release (PTR) check to see if our set of estimators has high Hamming distance to any "unsafe" set. If this is the case, we may apply the exponential mechanism restricted to regions of high approximate Tukey depth. The algorithm is (ϵ, δ)-DP because a "safe" database–one that passes the PTR check–is one that gives the restricted exponential mechanism similar outputs to those of neighboring databases.

Tukey depth is a measure of centrality of a point relative to others in multidimensional space. Since exact Tukey depth is NP-hard to compute for arbitrary feature dimensions (Johnson & Preparata, 1978), the authors propose a fast approximation that does not interfere with privacy guarantees. The final algorithm is not only private and high-utility, but also computationally efficient.

In the following sections we delve deeper into the technical details of the algorithm, analyze its merits and limitations, reimplement it ourselves and try to provide a slight contribution by expanding it to logistic regression.

2. Related Work

The TukeyEM algorithm extends prior research proposed by (Alabi et al., 2022), which introduces a differentially private version of the Theil-Sen estimator for one-dimensional linear regression. Another method that requires minimum user input is Boosted AdaSSP (Tang et al., 2023), which applies gradient boosting to the AdaSSP algorithm introduced by (Wang, 2018). Key limitation of these methods is that they obtain somewhat weaker performance on a larger class of datasets. Additionally, (Liu et al., 2021) presents relevant work by incorporating a PTR step, adapted from (Brown et al., 2021), alongside a restricted exponential mechanism. However, both steps suffer from inefficiency. Independently, (Cumings-Menon, 2022) explore the use of Tukey depth and regression depth for privately selecting non-private regression models, offering a separate perspective.

3. Methodology

3.1. Preliminaries

As the original paper, we begin by outlining the concept of differential privacy, adopting the 'add-remove' variation.

Definition 3.1. (Dwork et al., 2006) Databases D, D' from data domain $\mathcal{X}$ are neighbors, denoted $D \sim D'$, if they differ in the presence or absence of a single record. A randomized mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ is (ϵ, δ)-differentially private (DP) if for all $D \sim D'$ and any $S \subseteq \mathcal{Y}$

$$P_M[M(D) \in S] \leq e^{\epsilon} P_M[M(D') \in S] + \delta.$$

When $\delta = 0$, $\mathcal{M}$ is ϵ-DP. One general ϵ-DP algorithm is the exponential mechanism.

Definition 3.2. (McSherry & Talwar, 2007) Given database D and utility function $u : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ mapping (database, output) pairs to scores with sensitivity

$$\Delta_u = \max_{D \sim D', y \in \mathcal{Y}} |u(D, y) - u(D', y)|,$$

the exponential mechanism selects item $y \in \mathcal{Y}$ with probability proportional to $\exp(\frac{\epsilon u(D,y)}{2\Delta_u})$. We say the utility function is u is monotonic if, for $D_1 \subset D_2$, for any y, $u(D_1, y) \leq u(D_2, y)$. Given monotonic u, the 2 inside the exponent denominator can be dropped.

Lemma 3.3. *(McSherry & Talwar, 2007) The exponential mechanism is ϵ-DP.*

Then, we define Tukey depth.

Definition 3.4. (Tukey, 1975) A halfspace h_v is defined by a vector $v \in \mathbb{R}^d, h_v = \{y \in \mathbb{R}^d | \langle v, y \rangle \geq 0\}$. Let $D \subset \mathbb{R}^d$ be a collection of n points. The Tukey depth $T_D(y)$ of a point $y \in \mathbb{R}^d$ with respect to D is the minimum number of points in D in any halfspace containing y,

$$T_D(y) = \min_{h_v | y \in h_v} \sum_{x \in D} 1_{x \in h_v}.$$

Note that for a collection of n points, the maximum possible Tukey depth is $n/2$. The authors proved a theoretical utility result for a version of their algorithm that uses exact Tukey depth. However, Tukey depth is NP-hard to compute for arbitrary d (Johnson & Preparata, 1978), so their experiments instead use a notion of approximate Tukey depth that can be computed efficiently. The approximate notion of Tukey depth only takes a minimum over the $2d$ halfspaces corresponding to the canonical basis.

Definition 3.5. Let $E = \{e_1, ..., e_d\}$ be the canonical basis for $\mathbb{R}^d$ and let $D \subset \mathbb{R}^d$. The approximate Tukey depth of a point $y \in \mathbb{R}^d$ with respect to D, denoted $\widetilde{T}_D(y)$, is the minimum number of points in D in any of the $2d$ halfspaces determined by E containing y,

$$\widetilde{T}_D(y) = \min_{e_| e_j \in \pm E, y \in h_{y_i \cdot e_j}} \sum_{x \in D} 1_{h_{y_i \cdot e_j}}.$$

Stated more plainly, approximate Tukey depth only evaluates depth with respect to the d axis-aligned directions.

3.2. Main Algorithm

Utilizing the exponential mechanism in conjunction with Tukey depth, the algorithm put forward, named TukeyEM, unfolds through a sequence of four distinct phases:

1. **Step 1:** Randomly partition the dataset into m subsets, non-privately compute the OLS estimator on each subset, and collect the m estimators into set $\{\hat{\beta}_i\}_{i=1}^m$.
2. **Step 2:** Compute the volumes of regions of different approximate Tukey depths with respect to $\{\hat{\beta}_i\}_{i=1}^m$.
3. **Step 3:** Run a propose-test-release (PTR) algorithm using these volumes. If it passes, set B to be the region of $\mathbb{R}^d$ with approximate Tukey depth at least $m/4$ in $\{\hat{\beta}_i\}_{i=1}^m$ and proceed to the next step. If not, release $\perp$ (failure).
4. **Step 4:** If the previous step succeeds, apply the exponential mechanism, using approximate Tukey depth as the utility function, to privately select a point from B.

A basic utility result for the version of TukeyEM using exact Tukey depth appears below. The result is a direct application of work from (Brown et al., 2021).

Theorem 3.6. *(Brown et al., 2021) Let $0 < \alpha, \gamma < 1$ and let $S = \{\beta_1, \ldots, \beta_m\}$ be an i.i.d. sample from the multivariate normal distribution $\mathcal{N}(\beta^*, \Sigma)$ with covariance $\Sigma \in \mathbb{R}^{d \times d}$ and mean $\mathbb{E}[\beta_i] = \beta^* \in \mathbb{R}^d$. Given $\hat{\beta} \in \mathbb{R}^d$ with Tukey depth at least p with respect to S, there exists a constant $c > 0$ such that when $m \geq c\left(\frac{d + \log(1/\gamma)}{\alpha^2}\right)$ with probability $1 - \gamma$, $\|\hat{\beta} - \beta^*\|_\Sigma \leq \Phi^{-1}(1 - p/m + \alpha)$, where Φ denotes the CDF of the standard univariate Gaussian.*

In practical applications, models derived from real datasets frequently exhibit characteristics akin to Gaussian distributions, such as concentrated peaks, rapid decay in the tails, and symmetrical shapes. However, it is important to note as per the authors' clarification that Theorem 3.6 serves to establish conditions that are adequate but not obligatory. The TukeyEM algorithm does not hinge on any specific distributional characteristics for maintaining privacy, and deviations from Gaussian traits do not hinder the precision of its estimations.

Subsequent sections provide a detailed exposition of the authors' approach employing approximate Tukey depth, culminating with the comprehensive pseudocode presented in Algorithm 2 and the key findings encapsulated in Theorem 3.15.

3.3. Computing Volumes

Next, as the original paper, we commence with the process of calculating volumes associated with varying levels of Tukey depths. As elucidated in the following subsection, these calculated volumes are integral to the execution of the PTR subroutine.

Definition 3.7. Given database D, define $V_{i,D} = \text{vol}\left(\{y \mid y \in \mathbb{R}^d \text{ and } \widetilde{T}_D(y) \geq i\}\right)$, the volume of the region of points in $\mathbb{R}^d$ with approximate Tukey depth at least i in D. When D is clear from context, we write V_i for brevity.

Lemma 3.8. *Lines 6 to 13 of Algorithm 2 compute $\{V_i\}_{i=1}^{m/2}$ in time $O(dm \log(m))$.*

3.4. Applying Propose-Test-Release

The subsequent phase in TukeyEM's procedure utilizes PTR to delineate the output region for the exponential mechanism. This phase is encapsulated within the PTRCheck subroutine.

The overall strategy applies work done by (Brown et al., 2021). Their technique conducts a private evaluation to determine if a database is sufficiently removed—via Hamming distance—from any database deemed "unsafe." Upon passing this PTR assessment, it activates an exponential

mechanism circumscribed to a high Tukey depth territory. The classification of a "safe" database is one in which the confined exponential mechanism mirrors the output distribution across any adjacent database, thereby rendering the comprehensive algorithm differentially private (DP). In the ambit of their utility analysis, they present a lemma that equates a volumetric condition across various Tukey depth regions to a minimal Hamming distance from an "unsafe" database (refer to Lemma 3.8 in (Brown et al., 2021)). This foundational lemma posits that the PTR check generally succeeds when the algorithm is fed a sufficient amount of Gaussian data, which then secures the utility assurance. Nonetheless, their method necessitates the calculation of both the precise Tukey depths for the samples and the exact Hamming distance from peril for the current database. The computational demand for these calculations escalates exponentially with the dimension d (see their Section C.2 ((Brown et al., 2021)).

The authors employ the concept of approximate Tukey depth(Definition 3.5) to address both of the previously mentioned challenges. First, as the previous section demonstrated, computing the approximate Tukey depths of a collection of m d-dimensional points only takes $O(dm\log(m))$. Second, they refine the lower bound provided by Brown to achieve a 1-sensitive lower bound on the Hamming distance between the current database and any unsafe database, thereby establishing an effective alternative to the precise Hamming distance computation described in (Brown et al., 2021).

The overall structure of PTRCheck is therefore as follows: use the volume condition to compute a 1-sensitive lower bound on the given database's distance to unsafety; add noise to the lower bound and compare it to a threshold calibrated so that an unsafe dataset has probability $\leq \delta$ of passing; and 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.

Algorithm 1 PTRCheck

1: **Input**: Tukey depth region volumes V, privacy parameters ε and δ
2: Use Lemma 3.6 with $t = \frac{|V|}{2}$ and $\frac{\delta}{8e^{\varepsilon}}$ to compute lower bound k for distance to unsafe database
3: **if** $k + \text{Lap}\left(\frac{1}{\varepsilon}\right) \geq \frac{\log(1/2\delta)}{\varepsilon}$ **then**
4: **return** True
5: **else**
6: **return** False

Definition 3.9. (Brown et al., 2021) Two distributions P and Q over domain $\mathcal{W}$ are (ε, δ)-indistinguishable, denoted $P \approx_{\varepsilon,\delta} Q$, if for any measurable subset $W \subseteq \mathcal{W}$,

$$P[w \in W] \leq e^{\varepsilon} Q[w \in W] + \delta$$

$$Q[w \in W] \leq e^{\varepsilon} P[w \in W] + \delta.$$

Note that (ε, δ)-DP is equivalent to (ε, δ)-indistinguishability between output distributions on arbitrary neighboring databases. Given database D, let A denote the exponential mechanism with utility function T_D (see Definition 2.5). Given nonnegative integer t, let A_t denote the same mechanism that assigns score $-\infty$ to any point with score $< t$, i.e., only samples from points of score $\geq t$. We will say a database is "safe" if A_t is indistinguishable between neighbors.

Definition 3.10 (Definition 3.1 (Brown et al., 2021))**.** Database D is (ε, δ, t)-safe if for all neighboring D' we have $A_t(D) \approx_{\varepsilon,\delta} A_t(D')$. Let $\text{Safe}_{(\varepsilon,\delta,t)}$ be the set of safe databases, and let $\text{Unsafe}_{(\varepsilon,\delta,t)}$ be its complement.

Below is the main result of this section, Lemma 3.11. Briefly, it modifies Lemma 3.8 from (Brown et al., 2021) to construct a 1-sensitive lower bound on distance to unsafety.

Lemma 3.11. *Define $M(D)$ to be a mechanism that receives as input database D and computes the largest k from $\{0, \ldots, t-1\}$ such that there exists $g > 0$ where, for volumes V defined using a monotonic utility function,*

$$\frac{V_{t-k-1,D}}{V_{t+k+g+1,D}} \cdot e^{-\epsilon \cdot g/2} \leq \delta$$

or outputs -1 if the inequality does not hold for any such k. Then for arbitrary D

1. *M is 1-sensitive, and*
2. *for all $z \in \text{Unsafe}_{(\varepsilon, 4e^{\epsilon}\delta, t)}$, $d_H(D, z) \geq M(D)$.*

PTRCheck therefore runs the mechanism defined by Lemma 3.11, add Laplace noise to the result, and proceeds to the restricted exponential mechanism if the noisy statistic crosses a threshold.

Lemma 3.12. *Given the depth volumes V computed in Lines 11 to 12 of Algorithm 2, PTRCheck(V, ε, δ) is ε-DP and takes time $O(m\log(m))$.*

3.5. Sampling

If PTRCheck passes, TukeyEM then calls the exponential mechanism restricted to points of approximate Tukey depth at least $t = m/4$, a subroutine denoted RestrictedTukeyEM (Line 15 in Algorithm 2). Note that the passage of PTR ensures that with probability at least $1 - \delta$, running RestrictedTukeyEM is (ϵ, δ)-DP. The authors use a common two step process for sampling from an exponential mechanism over a continuous space: 1) sample a depth using the exponential mechanism, then 2) return a uniform sample from the region corresponding to the sampled depth.

3.5.1. Sampling a Depth

Definition 3.13. Given database D, define $W_{i,D} = \text{vol}(\{\mathbf{y} \in \mathbb{R}^d \mid \widetilde{T}_D(\mathbf{y}) = i\})$, the volume of the region of points in $\mathbb{R}^d$ with approximate Tukey depth exactly i in D.

To execute the first step of sampling, for $i \in \{m/4, m/4+1, \ldots, m/2\}$, $W_{i,D} = V_{i,D} - V_{i+1,D}$, so we can compute $\{W_{i,D}\}_{i=m/4}^{m/2}$ from the V computed earlier in time $O(m)$. The restricted exponential mechanism then selects approximate Tukey depth i with probability

$$P[i] \propto W_{i,D} \cdot \exp(\epsilon \cdot i)$$

3.5.2. Uniformly Sampling From a Region

Once an approximate Tukey depth $\hat{i}$ is obtained, the objective shifts to procuring a uniformly distributed random point at this depth. Constructed volume $W_{\hat{i},D}$ represents the collection of points $y = (y_1, \ldots, y_d)$ where each dimension j is no less than $\hat{i}$ in depth, and at least one dimension j' matches $\hat{i}$ exactly. The result is straightforward when $d = 1$: draw a uniform sample from the union of the two intervals of points of depth exactly $\hat{i}$ (depth from the "left" and "right").

For $d > 1$, the basic idea of the sampling process is to partition the overall volume into disjoint subsets, compute each subset volume, sample a subset according to its proportion in the whole volume, and then sample uniformly from that subset. The partitioning approach, as proposed by the authors, cleaves the entire depth region at i precisely along the initial dimension bearing the exact depth i. This approach guarantees a valid partitioning since any point within the aggregate volume possesses at least one dimension at this specific depth, simplifying the partition volumes' computation with the prior calculated S. Finally, the last sampling step will be easy because the final subset will simply be a pair of (hyper)rectangles.

Lemma 3.14. *SamplePointWithDepth(S, i) returns a uniform random sample from the region of points with approximate Tukey depth i in S in time $O(d)$.*

3.6. Final Result

All of the necessary material before lead to the main result, Theorem 3.15, stated below.

Theorem 3.15. *TukeyEM, given in Algorithm 2, is (ϵ, δ)-DP and takes time $O(d^2 n + dm \log(m))$.*

A more detailed proof of the theorem can be found in the paper itself.

Algorithm 2 TukeyEM

Input: Features matrix $X \in \mathbb{R}^{n \times d}$, label vector $y \in \mathbb{R}^n$, number of models m, privacy parameters ε and δ
Evenly and randomly partition X and y into subsets $\{(X_i, y_i)\}_{i=1}^m$
for $i = 1, \ldots, m$ **do**
    Compute OLS estimator $\beta_i \leftarrow (X_i^T X_i)^{-1} X_i^T y_i$
end for
for dimension $j \in [d]$ **do**
    $\{\beta_{i,j}\}_{i=1}^m \leftarrow$ projection of $\{\beta_i\}_{i=1}^m$ onto dimension j
    $(S_{j,1}, \ldots, S_{j,m}) \leftarrow \{\beta_{i,j}\}_{i=1}^m$ sorted in nondecreasing order
end for
Collect projected estimators into $S \in \mathbb{R}^{d \times m}$, where each row is nondecreasing.
for $i \in [m/2]$ **do**
    Compute volume of region of depth $\geq i$, $V_i \leftarrow \prod_{j=1}^d (S_{j,m-(i-1)} - S_{j,i})$
end for
if $PTRCheck(V, \varepsilon/2, \delta)$ **then**
    $\hat{\beta} \leftarrow RestrictedTukeyEM(V, S, m/4, \varepsilon/2)$
    return $\hat{\beta}$
else
    return $\perp$

4. Critical Analysis

The innovative approach proposed in the paper advances the field of differentially private linear regression by addressing significant limitations present in previous methodologies. The authors' development of the TukeyEM algorithm marks a pivotal shift towards simplifying the application of DP in linear regression, ensuring ease of use for end users without requiring extensive domain knowledge or hyperparameter tuning. This section critically examines the merits and limitations of the proposed method, informed by the authors' findings.

4.1. Merits

1. **User-Friendly Algorithm:** One of the most salient features of the TukeyEM algorithm is its user-friendly design, requiring only the dataset and desired privacy level as inputs. This represents a significant improvement over prior methods that demanded intricate domain knowledge or hyperparameter specifications, making DP more accessible and practical for a broader audience.

2. **Empirical Effectiveness:** The empirical validation of TukeyEM showcases its high utility across various datasets, achieving comparable or superior R^2 scores against other DP methods. This not only demonstrates

the algorithm's robust performance but also its adaptability to different data characteristics, underscoring its practical significance.

3. **Innovative Use of Tukey Depth:** The application of Tukey depth as a utility function within the exponential mechanism is a novel approach that leverages centrality measures for DP. By opting for an approximate Tukey depth, the authors navigate the computational challenges associated with its exact calculation, preserving both privacy guarantees and computational efficiency.

4.2. Limitations

1. **Lack of Theoretical Utility Guarantees:** Unlike previous works, the paper does not provide theoretical utility guarantees for Gaussian data, which constitutes a gap in the robustness of the proposed method. While empirical evidence supports the algorithm's effectiveness, the absence of theoretical underpinnings may limit its applicability to broader contexts where such guarantees are crucial.

2. **Dependency on the Number of Models (m):** The algorithm's performance hinges on the correct specification of m, the number of data subsets. Although the authors offer heuristics for choosing m, this requirement introduces an element of uncertainty and could pose challenges for end users without a deep understanding of the algorithm's inner workings.

3. **Potential for Model Failure:** The possibility that TukeyEM may fail to return a model under reasonable specifications of m is a notable concern. This characteristic could lead to unpredictability in its application, especially in scenarios where model output is critical.

4. **High-Dimensional Data Applicability:** The time complexity analysis suggests that TukeyEM might not be well-suited for high-dimensional data, a limitation that could restrict its utility in increasingly common big data applications.

5. Replication of Experiments

5.1. Replicating R^2 Scores

As an effort to better understand each step of the algorithm, we implement it ourselves using only the paper's instructions as guidelines [1]. Our first experiment tries to replicate the R^2 scores on seven of the datasets used originally. As the authors point out, $m = 1000$ is generally enough to ensure that the PTR check always passes, and therefore we fix this value for all datasets. Details on the datasets and their features can be found in the paper. Table 1 shows our results.

Table 1. 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

One might see that our results closely follow the ones in the paper, pointing to a correct implementation of the algorithm. Note that, for datasets which tend to give lower R^2 scores in DP methods, such as California and Diamonds, there is also a higher variance for the results, which explains the difference in R^2 values in our implementation.

5.2. Dependence of R^2 on Number of Models

Next, we would like to verify whether or not performance is indeed stable with the choice of m. We attempt to demonstrate this by letting m vary from 800 to 1800, and plotting median R^2 score for TukeyEM compared to that of OLS. Results can be seen in Figure 1.

Note the similarity of profile between our curves and the ones presented by the authors in the paper's appendix. The difference from our median values to the ones in the paper for the California dataset may be explained not only by randomness but also by slight implementation differences which do not compromise on utility nor privacy.

Also remark that for some datasets, such as Traffic, performance can significantly decrease as we increase the number of models.

5.3. Sampling from Approximate Tukey Depth

We also wish to further give foundation to the fact we correctly implemented the algorithm. We do this by reproducing the small toy experiment shown in the paper's appendix, where regions of equal approximate Tukey depth are shown for a set of two-dimensional points, in contrast with exact Tukey depth.

We use the same set of points $\{(1,1),(7,3),(5,7),(3,3),(5,5),(6,3)\}$ and show that our algorithm for uniform sampling on regions of same approximate Tukey depth is functional in Figure 2.

[1] Code available at https://shorturl.at/cfstL

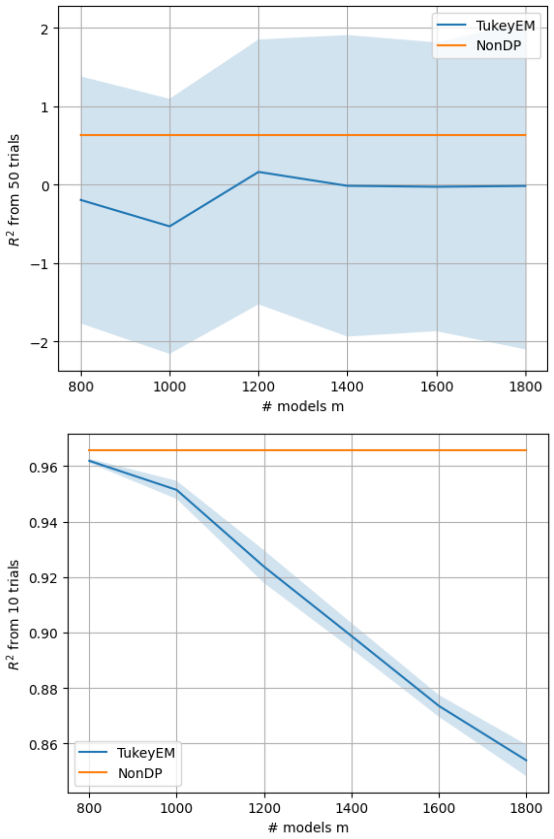

Figure 1. R^2 scores as a function of m for two datasets: California (top) and Traffic (bottom).

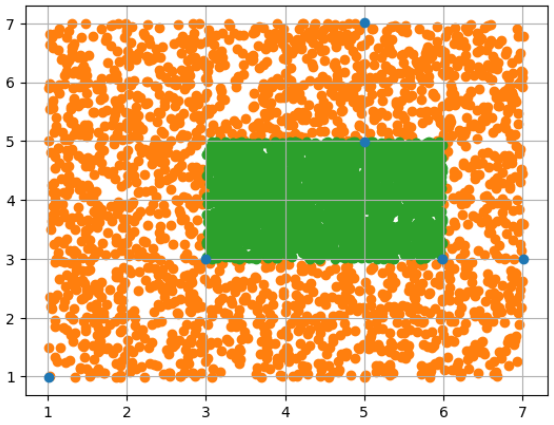

Figure 2. Uniformly distributed samples from regions with the same approximate Tukey depth.

6. Contributions

Finally, the authors focus on linear regression, but suggest that any method that produces d-dimensional vector models would keep the privacy guarantees, such that adapting TukeyEM to these other scenarios is straightforward. We therefore follow their lead and adapt TukeyEM to logistic regression.

Using `make_classification` from `sklearn`, we produce a dataset similar to Synthetic, but for binary classification. We use 22000 samples, with 10 features, not including the intercept. We adapt specifically line 4 of Algorithm 2 to make β_i receive a standard logistic regression estimator.

A standard logistic regression estimator gives an area under the ROC curve (AUC) of 0.974, while TukeyEM gives an AUC of 0.971. This shows that indeed, TukeyEM is able to achieve high accuracy while requiring no special knowledge about the data and providing important privacy guarantees.

7. Conclusion

In conclusion, the paper "Easy Differentially Private Linear Regression" presents a significant advancement in the field of differentially private linear regression by introducing the TukeyEM algorithm, which overcomes key limitations of previous methods. The algorithm offers a user-friendly approach that eliminates the need for extensive domain knowledge or hyperparameter tuning, making it accessible to a broader range of users. Empirical evaluations demonstrate its high utility across various datasets, achieving competitive R2 scores compared to other differentially private methods.

A notable innovation of TukeyEM lies in its utilization of Tukey depth as a utility function within the exponential mechanism, providing a novel approach that balances privacy guarantees with computational efficiency. Despite some limitations, such as the lack of theoretical utility guarantees and sensitivity to the choice of the number of models, the algorithm's robust performance and ease of use make it a valuable contribution to the field.

Through rigorous experimentation and replication of results, we have validated the effectiveness and correctness of the TukeyEM algorithm, confirming its utility in practical applications. Overall, the paper provides a solid foundation for further research and development in the realm of differentially private linear regression, with TukeyEM serving as a promising tool for privacy-preserving data analysis.

References

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

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

Amin, K., Joseph, M., Ribero, M., and Vassilvitskii, S. Easy differentially private linear regression. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=rSUCajhLsQ.

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

Cumings-Menon, R. Differentially private estimation via statistical depth, 2022.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography Conference (TCC)*, 2006.

Dwork, C., Talwar, K., Thakurta, A., and Zhang, L. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pp. 11–20, 2014.

Johnson, D. S. and Preparata, F. P. The densest hemisphere problem. *Theoretical Computer Science*, 6(1):93–107, 1978.

Kifer, D., Smith, A., and Thakurta, A. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pp. 25–1. JMLR Workshop and Conference Proceedings, 2012.

Liu, X., Kong, W., and Oh, S. Differential privacy and robust statistics in high dimensions, 2021.

McSherry, F. and Talwar, K. Mechanism design via differential privacy. *Foundations of Computer Science (FOCS)*, 2007.

Milionis, J., Kalavasis, A., Fotakis, D., and Ioannidis, S. Differentially private regression with unbounded covariates, 2022.

Song, S., Chaudhuri, K., and Sarwate, A. D. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pp. 245–248. IEEE, 2013.

Tang, S., Aydore, S., Kearns, M., Rho, S., Roth, A., Wang, Y., Wang, Y.-X., and Wu, Z. S. Improved differentially private regression via gradient boosting, 2023.

Tukey, J. W. Mathematics and the picturing of data. *International Congress of Mathematicians(IMC)*, 1975.

Wang, Y.-X. Revisiting differentially private linear regression: optimal and adaptive prediction estimation in unbounded domain, 2018.