Analyzing the Differentially Private Theil-Sen Estimator for Simple Linear Regression

Jayshree Sarathy and Salil Vadhan
Harvard John A. Paulson School of Engineering and Applied Sciences

ABSTRACT

In this paper, we study differentially private point and confidence interval estimators for simple linear regression. Motivated by recent work that highlights the strong empirical performance of an algorithm based on robust statistics, `DPTheilSen`, we provide a theoretical analysis of its privacy and accuracy properties, offer guidance on setting hyperparameters, and show how to produce non-parametric, differentially private confidence intervals to accompany its point estimates.

1 INTRODUCTION

In the last several years, as differential privacy has made its way from theory to practice, we are beginning to see where further theoretical research is needed to design differentially private methods for common statistical inference tasks (see, e.g., [11]). One example is with simple (ie. one-dimensional) linear regression, which is one of the most fundamental tasks in data analysis. In 2018, the economics research group, Opportunity Insights, found that there was a lack of consensus around the best differentially private algorithms for simple linear regression on regimes commonly used in practice (e.g. small-area analysis with 40 to 400 datapoints per regression). Therefore, to release linear regression estimates in their Opportunity Atlas, the group used a heuristic method that did not satisfy the formal guarantees of differential privacy [7, 8].

Motivated by this gap between theory and practice, Alabi et al. [1] conducted an empirical evaluation of several differentially private algorithms for simple linear regression. They found that a suite of robust, median-based algorithms, `DPTheilSen`, based on the non-private Theil-Sen estimator developed by Theil [25] and Sen [23], performed better than standard OLS-based algorithms across a range of practical regimes. `DPTheilSen` has now been implemented in the open-source SmartNoise library [22] as the default algorithm for simple linear regression. While their empirical study was a valuable starting point, Alabi et al. [1] stated that further theoretical understanding of the accuracy guarantees of `DPTheilSen`, as well as design of uncertainty estimates, would be needed to make this set of algorithms fully usable in practice.

In this paper, we address these open questions. We analyze the privacy and accuracy guarantees of the `DPTheilSen` algorithms. (We note that Dwork and Lei [14] analyzed one version of `DPTheilSen` in the privacy setting, which they called the "Short-Cut Regression Method," but they did not analyze the variants that were shown by Alabi et al. [1] to have stronger performance, nor did they consider uncertainty measures.) Our work provides theoretical explanations for the results of Alabi et al., offers guidance on setting hyperparameters, and is the first to design and analyze differentially private confidence intervals for `DPTheilSen`.

1.1 Related work

This work draws on the rich connections between robust statistics and differential privacy. Dwork and Lei [14] stated that "robust estimators are a useful starting point for constructing highly accurate differentially private estimators." In their paper, they gave a theoretical, asymptotic analysis of what they called the "Short-Cut Regression Method," which is similar to one of the variants of `DPTheilSen` we consider. However, Dwork and Lei did not consider the more statistically efficient variants of `DPTheilSen` that we do in this work, nor did they offer measures of uncertainty for the estimates. Couch et al. [9] also find that robust estimators perform better than parametric estimators under differential privacy, even when the data come from a parametric model, but they focus on hypothesis testing and do not provide a theoretical utility analysis. Alabi et al. [1]'s experimental evaluations demonstrate that differentially private analogues robust algorithms for simple linear regression, such as `DPTheilSen`, perform better than non-robust methods when the dataset size, variance of the independent variables, or privacy loss parameter is small. However, Alabi et al. also do not provide theoretical analysis or construct differentially private confidence intervals for this estimator.

In general, linear regression is one of the most fundamental tasks in statistics, and thus, has attracted much attention in the DP literature. Sheffet [24] considered differentially private ordinary least squares (OLS) methods and corresponding DP confidence intervals, but unlike our work, these methods assume normality of errors, require input data bounds, and satisfy approximate, rather than pure, DP. Wang [26] studied private ridge regression and considered DP confidence intervals, but these methods require consuming additional privacy budget for estimating Hessians. Barrientos et al. [3] and Evans et al. [16] use the subsample-and-aggregate framework, but their approaches rely on normality assumptions or normal approximations that only hold for large n. Bernstein and Sheldon [4] consider a Bayesian approach, but unlike our work, they require a prior on the distribution of both the regression coefficients and the independent variables.

Our approach to confidence intervals builds on the recent work of Drechsler et al. [12], who design non-parametric DP confidence intervals for the median. One main difference is that their algorithms provide finite-sample validity for i.i.d. variables, while our work offers asymptotic validity for i.i.d., as well as some forms of non-i.i.d, variables (which, in our setting, are slopes computed by the `DPTheilSen` algorithm). We do show finite-sample confidence intervals for the `TheilSen1Half` variant. Unlike our work, Drechsler et al. do not provide theoretical utility analysis. Prior work has also considered DP confidence intervals for mean estimation [13, 18, 21], but these cannot directly be applied for the median-based estimator we consider. Recent work focuses on general approaches to DP confidence intervals using bootstrapping [3, 5, 10, 17], but

these can be expensive to compute and rely on parametric assumptions, such as normally distributed errors or point estimates, that our work avoids. To the best of our knowledge, our work is the first to theoretically analyze the different variants of DPTheilSen as well as to design and analyze non-parametric, differentially private confidence intervals to accompany its point estimates.

2 PRELIMINARIES

We will consider datasets that are multisets. The space of datasets will be denoted by Multisets($\mathcal{D}, n$), where $\mathcal{D}$ is the underlying set of elements and n is the cardinality of each multiset. We view datasets $\boldsymbol{d} \in$ Multisets($\mathcal{D}, n$) as specified by histograms $m_{\boldsymbol{d}} : \mathcal{D} \to \mathbb{N}$ where $m_{\boldsymbol{d}}(w)$ gives the multiplicity of element w (so $\sum_{w \in \mathcal{D}} m_{\boldsymbol{d}}(w) = n$). We define the distance function for multisets, $dist_{\text{ms}}$: Multisets($\mathcal{D}, n$) $\times$ Multisets($\mathcal{D}, n$) $\to \mathbb{N}$, as follows. For any two datasets $\boldsymbol{d}, \boldsymbol{d'} \in$ Multisets($\mathcal{D}, n$), $dist_{\text{ms}}(\boldsymbol{d}, \boldsymbol{d'}) = \frac{1}{2} \sum_{w \in \mathcal{D}} |m_{\boldsymbol{d}}(w) - m_{\boldsymbol{d'}}(w)|$, ie. the number of records that need to be changed to transform $\boldsymbol{d}$ into $\boldsymbol{d'}$.

2.1 Differential Privacy

The algorithms in this paper satisfy pure differential privacy (DP). Since they include hyperparameters, we state a definition of DP for algorithms that take as input not only the dataset, but also the desired privacy parameters and any required hyperparameters. Two datasets $\boldsymbol{d}, \boldsymbol{d'} \in$ Multisets($\mathcal{D}, n$) are *neighboring*, denoted $\boldsymbol{d} \sim \boldsymbol{d'}$, if $dist_{\text{ms}}(\boldsymbol{d}, \boldsymbol{d'}) = 1$. Let $\mathcal{H}$ be a hyperparameter space and $\mathcal{Y}$ be an output space.

Definition 1 (Differential Privacy [15]). For $\varepsilon \in \mathbb{R}_{\geq 0}$, a randomized algorithm M : Multisets($\mathcal{D}, n$)$\times\mathbb{R}_{\geq 0}\times\mathcal{H} \to \mathcal{Y}$ is *ε-differentially private* if and only if for all neighboring datasets $\boldsymbol{d} \sim \boldsymbol{d'} \in$ Multisets($\mathcal{D}, n$) hyperparams $\in \mathcal{H}$, and sets $E \subseteq \mathcal{Y}$,

$$\Pr[M(\boldsymbol{d}, \varepsilon, \text{hyperparams}) \in E] \leq e^{\varepsilon} \cdot \Pr[M(\boldsymbol{d'}, \varepsilon, \text{hyperparams}) \in E].$$

where the probabilities are taken over the random coins of M.

2.2 Simple linear regression

We consider the standard simple linear regression model, where we are given n fixed values, $x_1, \ldots, x_n$, which are not all equal, of the predictor variable x. For each x_i, we observe the corresponding value y_i of the response random variable y. We assume that the model is $y_i = \alpha + \beta x_i + e_i$ for $i = 1, \ldots, n$, where α and β are unknown parameters, and where each e_i is sampled independently from the same continuous distribution F. Our goal is to design and analyze ε-differentially private point and interval estimators for β.

2.3 A note on the convergence bounds

Although the DPTheilSen algorithms only require the assumptions in Section 2.2, we will use a simplified setup in order to state the convergence bounds. In particular, we assume that the $x_1, \ldots, x_n$ are constants that are spaced apart equally, and that each $e_i, i \in [n]$, is sampled i.i.d. from $\mathcal{N}(0, \sigma_e^2)$. These assumptions are not required for the validity (ie. coverage) of the confidence intervals nor for the privacy guarantees of the algorithms.

3 POINT ESTIMATORS

We begin by defining the non-private Theil-Sen estimator and one of its efficient variants.

Definition 2 (Theil-Sen estimator [23, 25]). Let $(x_1, y_1), \ldots, (x_n, y_n)$ be an arbitrary ordering of dataset $\boldsymbol{d} \in$ Multisets($\mathbb{R} \times \mathbb{R}, n$). For $1 \leq i < j \leq n$ such that $x_i \neq x_j$, compute the slope s_{ij} between the points (x_i, y_i) and (x_j, y_j) as:

$$s_{ij} = \frac{y_j - y_i}{x_j - x_i} = \beta + \frac{e_j - e_i}{x_j - x_i}$$

Let $\mathbf{s}_{\boldsymbol{d}}$ denote the multiset of the slopes. The *Theil-Sen estimator* $\hat{\beta}^{TS}$ is

$$\hat{\beta}^{\text{TS}} = \text{median}(\mathbf{s}_{\boldsymbol{d}})$$

Theil [25] defined a second estimator, which has also been called the "abbreviated" method in the literature. This variant considers only up to $\lfloor n/2 \rfloor$ of the $\binom{n}{2}$ possible slopes between pairs of points.

Definition 3 (Theil-Sen Half estimator [25]). Let $(x_1, y_1), \ldots, (x_n, y_n)$ be an ordering of $\boldsymbol{d}$ such that $x_1 \leq x_2 \leq \cdots \leq x_n$. For $j = 1, \ldots, \lfloor n/2 \rfloor$ such that $x_j \neq x_{n/2+j}$, compute the slope between the points (x_j, y_j) and $(x_{n/2+j}, y_{n/2+j})$.

$$s_j = \frac{y_{n/2+j} - y_j}{x_{n/2+j} - x_j} = \beta + \frac{e_{n/2+j} - e_j}{x_{n/2+j} - x_j}$$

Let $\mathbf{s}_{\boldsymbol{d}}$ denote the multiset of the slopes. The *Theil-Sen Half estimator* $\hat{\beta}^{TSHalf}$ is

$$\hat{\beta}^{\text{TSHalf}} = \text{median}(\mathbf{s}_{\boldsymbol{d}})$$

The accuracy guarantees of the full Theil-Sen estimator (Definition 2) are typically analyzed asymptotically. In particular, the estimator can be rewritten in terms of a U-statistic [19], whose asymptotic normality yields the following theorem by Sen [23] (stated in terms of our setting).

Theorem 4. *Given datapoints $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_1, \ldots, x_n$ are equally-spaced constants with variance σ_x^2 and $y_i = \alpha + \beta x_i + e_i, e_i \sim \mathcal{N}(0, \sigma_e^2)$, for $i \in [n]$, let $\hat{\beta}^{TS}$ be the Theil-Sen estimator from Definition 2. Then,*

$$\sqrt{n} \cdot \left(\hat{\beta}^{TS} - \beta\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\pi\sigma_e^2}{3\sigma_x^2}\right)$$

Similar to the Theil-Sen algorithm, "incomplete" versions of Theil-Sen, which require sampling some subset of the $\binom{n}{2}$ pairs of points, can be analyzed through the asymptotic normality of some forms of "incomplete U-statistics" (see, e.g., [20]). In the case of Theil-Sen Half, however, the pairs are chosen such that the $\lfloor n/2 \rfloor$ slopes are independent, which means that one can obtain a finite-sample analysis using a Hoeffding bound. In Section 3.2, we show the comparison of convergence bounds for the non-private and private versions of these algorithms.

3.1 DPTheilSen algorithms

In the differentially private analogue of Theil-Sen, called DPTheilSen (Algorithm 3.1), we compute pairwise estimates of the slope. However, we replace the computation of the median of the slopes with a differentially private median algorithm, which can be one of several algorithms. We use the *widened exponential mechanism* (denoted by DPWide), which was introduced by Alabi et al. [1] and

further analyzed by Drechsler et al. [11]. We choose this algorithm as it displayed strong performance, compared to other DP median algorithms, in both works' empirical evaluations. DPWide requires a widening parameter $\Theta > 0$, and bounds on the range of the output, $[-R, R]$.

Algorithm 3.1: DPTheilSen: ε-DP Algorithm

Data: $\boldsymbol{d} \in \text{Multisets}(\mathbb{R} \times \mathbb{R}, n)$
Privacy parameters: $\varepsilon \in \mathbb{R}_{\geq 0}$
Hyperparameters: $R, \theta \in \mathbb{R}_{\geq 0}$
$\mathbf{s}_{\boldsymbol{d}} = \{\}$
for *each of the* $\binom{n}{2}$ *unordered pairs of datapoints* $w, w' \in \boldsymbol{d}$ **do**
  $s = \text{Slope}(w, w')$
  Add s to $\mathbf{s}_{\boldsymbol{d}}$
$\widetilde{\beta}^{\text{TS}} = \text{DPWide}\left(\mathbf{s}_{\boldsymbol{d}}, \frac{\varepsilon}{n-1}, (1/2, [-R, R], \theta)\right)$
return $\widetilde{\beta}^{\text{TS}}$

Algorithm 3.2: Slope

Data: $(x, y), (x', y') \in \mathbb{R} \times \mathbb{R}$

$$s = \begin{cases} 0, & \text{if } y' - y = 0 \text{ and } x' - x = 0 \\ \text{sign}(y' - y) \cdot \infty, & \text{if } x' - x = 0 \\ (y' - y)/(x' - x), & \text{otherwise} \end{cases}$$

return s

Lemma 5 ([1]). *Algorithm 3.1 (*DPTheilSen*) is ε-DP.*

Next, we consider a more efficient algorithm, DPTheilSenkHalf (Algorithm 3.3), which is a differentially private version of Theil-Sen Half. (When $k = 1$, we denote the algorithm DPTheilSenHalf.) Recall that in Theil-Sen Half, the datapoints are ordered such that $x_1 \leq x_2 \leq \cdots \leq x_n$, and the slopes are computed between the datapoints (x_j, y_j) and $(x_{n/2+j}, y_{n/2+j})$, for $j = 1, \ldots, \lfloor n/2 \rfloor$. In the DP setting, however, sorting all of the datapoints by ascending x-values would require scaling the privacy budget down by a factor of n, which is too costly. A second approach is to instead randomly select $k \geq 1$ matchings of the n points that do not depend on any particular ordering of the datapoints. As every point is used to compute at most k slopes, the privacy parameter ε only has to be scaled down by k before being passed to the DPmed algorithm. A version of this algorithm (with $k = 1$) was previously considered in the DP setting by Dwork and Lei [14].

We adopt a blend of these two approaches, which facilitates the utility analysis. We sort the datapoints into only two bins based on their x-values (ie. $x_i \leq x_j$ for every datapoint (x_i, y_i) in the first bin and every datapoint (x_j, y_j) in the second bin). Sorting into two bins requires paying an additional factor of 2 in the privacy parameter, since a change in one datapoint may cause a change in each of the two bins. Then, we define a partition of the complete bipartite graph into $\lceil n/2 \rceil$ matchings, from which we randomly select k matchings with replacement.

Lemma 6. *Algorithm 3.3 (*DPTheilSenkHalf*) is ε-DP.*

Algorithm 3.3: DPTheilSenkHalf: ε-DP Algorithm

Data: $\boldsymbol{d} \in \text{Multisets}(\mathbb{R} \times \mathbb{R}, n)$
Privacy parameters: $\varepsilon \in \mathbb{R}_{\geq 0}$
Hyperparameters: $R, \theta \in \mathbb{R}_{\geq 0}$
$\vec{B}_1, \vec{B}_2 = \text{Partition-and-Permute}(\boldsymbol{d})$ // Partition the datapoints into two bins, B_1 and B_2, such that for all $(x_i, y_i) \in B_1, (x_j, y_j) \in B_2$, $x_i \leq x_j$, $|B_1| = \lfloor n/2 \rfloor$, and $|B_2| = \lceil n/2 \rceil$. Then, randomly permute the two bins to obtain $\vec{B}_1$ and $\vec{B}_2$.
$M = \text{Match}(\vec{B}_1, \vec{B}_2, k)$ // For every $p \in [k]$, draw a random $r_p \leftarrow \{1, \ldots, \lceil n/2 \rceil\}$ and match the ith point in $\vec{B}_1$ with the $(i + r_p) \mod \lceil n/2 \rceil$th point in $\vec{B}_2$, for $i = 1, \ldots, \lfloor n/2 \rfloor$. Return M, the set of matched points.
$\mathbf{s}_{\boldsymbol{d}} = \{\}$
for $(w_1, w_2) \in M$ **do**
  $s = \text{Slope}(w_1, w_2)$
  Add s to $\mathbf{s}_{\boldsymbol{d}}$
$\widetilde{\beta}^{\text{DPTSkHalf}} = \text{DPWide}\left(\mathbf{s}_{\boldsymbol{d}}, \frac{\varepsilon}{2k}, (1/2, [-R, R], \theta)\right)$
return $\widetilde{\beta}^{\text{DPTSHalf}}$

3.2 Convergence Bounds

A $(1-p)$-*convergence bound* for an estimator $\hat{\beta}$ of the true slope β is a value $t = t(\alpha, \beta, (x_1, \ldots, x_n), \sigma_x, \sigma_e, n, \varepsilon)$ such that with probability at least $1 - p$ (over the e_i's in the data and the coins of the estimator), we have $|\hat{\beta} - \beta| \leq t$.

Our convergence bounds for DPTheilSen and DPTheilSenkHalf combine asymptotic analysis for the non-private convergence (as in Theorem 4) and finite-sample analysis to characterize the effects of privacy. These bounds are therefore not fully rigorous, and we are working towards finite-sample analyses of these algorithms, which we believe should be derivable using Berry-Esseen-like theorems for complete and incomplete U-statistics (eg. [6]). We do currently show finite-sample bounds for DPTheilSen1Half, which is in contrast with Dwork and Lei's [14] asymptotic analysis of a similar version of this algorithm.

In Table 1, we display the $(1-p)$-convergence bounds for DPTheilSen, DPTheilSenHalf (ie. $k = 1$), and DPTheilSenkHalf. For comparison, we also include the non-private bounds of OLS, Theil-Sen, and Theil-Sen Half, as well as the bounds of an differentially private analogue of OLS, called DPSuffStats, which was analyzed by Alabi and Vadhan [2]. We let $c_{p/2,n} = \phi^{-1}(1 - p/2)/\sqrt{n}$.

The first three bounds in the table correspond to the non-private algorithms. We see that Theil-Sen nearly recovers the accuracy of OLS, up to a factor of $\sqrt{\pi/3}$. Theil-Sen Half, on the other hand, is a factor of $\sqrt{2}$ worse than Theil-Sen. The bounds for DPSuffStats, DPTheilSen, and DPTheilSenHalf have the same constant factors for the highest order term as OLS, Theil-Sen and Theil-Sen Half, respectively, but they include lower order terms corresponding to the noise due to privacy.

These bounds confirm some experimental findings of Alabi et al. [1] with respect to the differences between DPSuffStats and DPTheilSen and hyperparameter selection. First, the bounds show

Estimator	Convergence bound
OLS	$\frac{\sigma_e}{\sigma_x} \cdot \frac{c_{p/4}}{\sqrt{n}}$
Theil-Sen*	$\sqrt{\frac{\pi}{3}} \cdot \frac{\sigma_e}{\sigma_x} \cdot \frac{c^*_{p/4}}{\sqrt{n}} \cdot (1 + o(1))$
Theil-Sen Half	$\sqrt{\frac{2\pi}{3}} \cdot \frac{\sigma_e}{\sigma_x} \cdot \frac{c_{p/4}}{\sqrt{n}} \cdot (1 + o(1))$ $n > \ln(4/p)$
DPSuffStats	$\frac{\sigma_e}{\sigma_x} \cdot \frac{c_{p/12}}{\sqrt{n}} \cdot (1 + \tau) + \tau\,(1 + \tau + \lvert\beta\rvert)$, $\tau \approx \frac{(1-1/n) r_u^2 \log(3/p)}{\varepsilon \cdot n \cdot \sigma_x^2}$
DPTheilSen*	$\sqrt{\frac{\pi}{3}} \cdot \frac{\sigma_e}{\sigma_x} \cdot \left(\frac{c^*_{p/16}}{\sqrt{n}} + \tau\right)(1 + o(1)) + \theta$, $\tau = \frac{\ln(R/\sqrt{\pi} \cdot p \cdot \theta \cdot \sigma_e)}{\varepsilon n}$ suff. small
DPTheilSen Half	$\sqrt{\frac{2\pi}{3}} \cdot \frac{\sigma_e}{\sigma_x} \cdot \left(\frac{c_{p/16}}{\sqrt{n}} + \tau\right)(1 + o(1)) + \theta$, $\tau = \frac{\ln(R/\sqrt{\pi} \cdot p \cdot \theta \cdot \sigma_e)}{\varepsilon n}$ suff. small, $n > 16 \ln(16/p)$
DPTheilSen kHalf*	$\sqrt{\frac{2\pi(2k+1)}{9k}} \frac{\sigma_e}{\sigma_x} \cdot \left(\frac{c^*_{p/16}}{\sqrt{n}} + \tau\right)(1 + o(1)) + \theta$, $\tau = \frac{\ln(R/\sqrt{\pi} \cdot p \cdot \theta \cdot \sigma_e)}{\varepsilon n}$ suff. small

Table 1: Comparison of $1-p$ convergence bounds (some constraints omitted). For estimators marked with a star, bounds are asymptotic.

that the DPTheilSen variants have a logarithmic dependence on the range R of the output $\tilde{\beta}$, while DPSuffStats has a quadratic dependence on the range r_u for the input data, and a linear dependence on $|\beta|$. Second, they confirm that by using $k \approx 10$ matchings in DPTheilSenkHalf, we can gain computational efficiency without losing too much utility, although the partitioning step hinders us from recovering the full utility of DPTheilSen.

In addition, these bounds offer insight guidance on how should one set the widening parameter θ. For a given n, let τ_n be the upper bound on τ such that the normal quantile approximation is valid. (We also allow τ_n to absorb the term $c_{p/4,n}$). For fixed $p, \varepsilon, R, \sigma_e, \sigma_x$, and n, if we select θ to minimize the DPTheilSenHalf bound, for example, while satisfying the constraint that $\tau \le \tau_n$, we have that

$$\theta \approx \max\left(\frac{\sigma_e}{\varepsilon n \sigma_x}, R \exp\left(-\varepsilon n \cdot \ln(2/p) \cdot \tau_n\right)\right)$$

The factor $\sigma_e/(n\sigma_x)$ in the first term corresponds to the standard deviation of the slopes computed by DPTheilSenHalf. When the slopes are highly concentrated, the first term becomes small. The second term, however, is independent of σ_e and σ_x, which allows θ to remain bounded away from 0 and prevents a blowup in τ. (Handling the case of concentrated slopes was Alabi et al. [1]'s original motivation for designing the widened exponential mechanism.) Note that R, ε, n and p are known in practice; if the experimental design suggests that the slopes may be concentrated (eg. if the x-values are located at one of two endpoints of an interval), it may be beneficial to set θ to scale with the second term.

4 INTERVAL ESTIMATORS

Finally, we show how to produce differentially private confidence intervals for DPTheilSen. The algorithm below (Algorithm 4.1) applies the "naïve" exponential mechanism confidence interval from Drechsler et al. [12]. The idea is to run the widened exponential mechanism quantile estimator twice (with different target quantiles) such that with high probability, the two estimates capture the non-private confidence interval for the median. We use the naïve version to facilitate the utility analysis, but an open question is whether we can theoretically analyze Drechsler et al.'s more nuanced algorithms (which empirically provide tighter confidence intervals for the median) within the context of DPTheilSen.

Algorithm 4.1: DPTheilSenCI: ε-DP Algorithm

Data: $\boldsymbol{d} \in \text{Multisets}(\mathbb{R} \times \mathbb{R}, n)$
Privacy parameters: $\varepsilon \in \mathbb{R}_{>0}$
Hyperparameters: $p \in (0,1), R, \theta \in \mathbb{R}_{>0}$, alg $\in$ {DPTheilSen, DPTheilSenHalf, DPTheilSenkHalf}
Compute slopes $\mathbf{s}_{\boldsymbol{d}} = (s_1, \ldots, s_N)$ according to alg.

$$b = \begin{cases} \sqrt{4/9} \cdot c^*_{p/4,n} & \text{if alg is DPTheilSen} \\ \sqrt{(2k+1)/(3k)} \cdot c^*_{p/4,n} & \text{if alg is DPTheilSenkHalf} \\ \sqrt{2} \cdot c_{p/4,n} & \text{if alg is DPTheilSenHalf} \end{cases}$$

$t = \ln\left(\frac{4(R-\theta)}{\theta \cdot p}\right) / (\varepsilon N)$
$\widetilde{\beta}_L^{\text{TS}} = \text{DPWide}(\mathbf{s}_{\boldsymbol{d}}, \varepsilon/2, (1/2 - b - t, [-R, R], \theta)) - \theta$
$\widetilde{\beta}_U^{\text{TS}} = \text{DPWide}(\mathbf{s}_{\boldsymbol{d}}, \varepsilon/2, (1/2 + b + t, [-R, R], \theta)) + \theta$
return $[\widetilde{\beta}_L^{\text{TS}}, \widetilde{\beta}_U^{\text{TS}}]$

Lemma 7. *Algorithm 4.1 (*DPTheilSenCI*) is ε-DP.*

As DPTheilSenHalf produces slopes that are independent, the resulting confidence interval has finite-sample validity. For DPTheilSen and DPTheilSenkHalf, the slopes are not independent; however, we are able to use their limiting distributions to show asymptotic validity of the confidence intervals for these estimators. In both cases, the validity statements do not depend on any parametric assumptions.

5 CONCLUSION

In this work, we analyze the theoretical privacy and utility guarantees of DPTheilSen. We provide convergence bounds, offer insight into hyperparameter selection, and show how to produce differentially private confidence intervals. We plan to provide finite-sample analyses for all the variants of this algorithm. In the future, we hope to analyze the optimality of these algorithms and extend them to the setting of multivariate linear regression.

REFERENCES

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

[2] Alabi, D., and Vadhan, S. On the statistical convergence rate of differentially private ordinary least squares.

[3] Barrientos, A. F., Reiter, J., Machanavajjhala, A., and Chen, Y. Differentially private significance tests for regression coefficients. *Journal of Computational and Graphical Statistics 28* (2017), 440 – 453.

[4] Bernstein, G., and Sheldon, D. R. Differentially private bayesian inference for exponential families. In *Advances in Neural Information Processing Systems* (2018), S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc.

[5] Brawner, T. W., and Honaker, J. Bootstrap inference and differential privacy: Standard errors for free.

[6] Callaert, H., Janssen, P., et al. The berry-esseen theorem for u-statistics. *Annals of Statistics 6*, 2 (1978), 417–421.

[7] Chetty, R., and Friedman, J. N. A practical method to reduce privacy loss when disclosing statistics based on small samples. *American Economic Review Papers and Proceedings 109* (2019), 414–420.

[8] Chetty, R., Friedman, J. N., Hendren, N., Jones, M. R., and Porter, S. R. The opportunity atlas: Mapping the childhood roots of social mobility. Tech. rep., National Bureau of Economic Research, 2018.

[9] Couch, S., Kazan, Z., Shi, K., Bray, A., and Groce, A. Differentially private nonparametric hypothesis testing. *arXiv preprint arXiv:1903.09364* (2019).

[10] D'Orazio, V., Honaker, J., and King, G. Differential privacy for social science inference. *Alfred P. Sloan Foundation Economic Research Paper Series* (2015).

[11] Drechsler, J. Differential privacy for government agencies–are we there yet? *arXiv preprint arXiv:2102.08847* (2021).

[12] Drechsler, J., Globus-Harris, I., McMillan, A., Sarathy, J., and Smith, A. Nonparametric differentially private confidence intervals for the median. *Journal of Survey Statistics and Methodology 10*, 3 (2022), 804–829.

[13] Du, W., Foot, C., Moniot, M., Bray, A., and Groce, A. Differentially private confidence intervals. *arXiv arXiv:2001.02285* (2020).

[14] Dwork, C., and Lei, J. Differential privacy and robust statistics. In *STOC* (2009), vol. 9, pp. 371–380.

[15] Dwork, C., McSherry, F., Nissim, K., and Smith, A. D. 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* (2006), pp. 265–284.

[16] Evans, G., and King, G. Statistically valid inferences from differentially private data releases, with application to the facebook urls dataset. *Political Analysis* (2021 2021).

[17] Ferrando, C., Wang, S.-F., and Sheldon, D. General-purpose differentially-private confidence intervals. *ArXiv abs/2006.07749* (2020).

[18] Gaboardi, M., Rogers, R., and Sheffet, O. Locally private mean estimation: z-test and tight confidence intervals. In *Proceedings of Machine Learning Research* (16–18 Apr 2019), K. Chaudhuri and M. Sugiyama, Eds., vol. 89 of *Proceedings of Machine Learning Research*, PMLR, pp. 2545–2554.

[19] Hoeffding, W. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics 19* (1948), 293–325.

[20] Janson, S. The asymptotic distributions of incomplete u-statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 66*, 4 (1984), 495–505.

[21] Karwa, V., and Vadhan, S. Finite sample differentially private confidence intervals. In *ITCS* (2018).

[22] OpenDP.

[23] Sen, P. K. Estimates of the regression coefficient based on kendall's tau. *Journal of the American statistical association 63*, 324 (1968), 1379–1389.

[24] Sheffet, O. Differentially private ordinary least squares. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017* (2017), pp. 3105–3114.

[25] Theil, H. A rank-invariant method of linear and polynomial regression analysis, 3; confidence regions for the parameters of polynomial regression equations. *Indagationes Mathematicae 1*, 2 (1950), 467–482.

[26] Wang, Y. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018* (2018), pp. 93–103.