Decision Tree for Locally Private Estimation with Public Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

We propose conducting locally differentially private (LDP) estimation with the aid of a small amount of public data to enhance the performance of private estimation. Specifically, we introduce an efficient algorithm called *Locally differentially Private Decision Tree* (LPDT) for LDP regression. We first use the public data to grow a decision tree partition and then fit an estimator according to the partition privately. From a theoretical perspective, we show that LPDT is ε-LDP and has a mini-max optimal convergence rate under a mild assumption of similarity between public and private data, whereas the lower bound of the convergence rate of LPDT without public data is strictly slower, which implies that the public data helps to improve the convergence rates of LDP estimation. We conduct experiments on both synthetic and real-world data to demonstrate the superior performance of LPDT compared with other state-of-the-art LDP regression methods. Moreover, we show that LPDT remains effective despite considerable disparities between public and private data.

1 Introduction

Differential privacy (DP) [15] has become a standard approach to prevent information leakage and is widely used in various fields, such as medical trials [11], recommendation systems [30], and census data releasing [2]. Local differential privacy (LDP) [24, 14], which is a variant of DP, has gained considerable attention in recent years, particularly among industrial developers [16, 21]. Unlike central differential privacy, which relies on a trusted curator who has access to the raw data, LDP assumes that data is distributed among many users, and each user privatizes their data before it is collected by the curator. Although this setting provides stronger privacy protection, training with perturbed data usually requires far more samples [14] compared to the central setting. Moreover, many basic techniques such as principal component analysis [33, 8], data standardization [8], and tree partition [37] are troublesome or even prohibited with LDP. As a result, LDP introduces challenges for various machine learning tasks that are otherwise considered relatively simple, including density estimation [13], mean estimation [14], Gaussian estimation [22], and change-point detection [7].

Fortunately, in some scenarios, private estimation performance can be enhanced with an additional public dataset [3, 5]. The public dataset can be either in-distribution, consisting of data from users who agree to share their personal information [4], or out-of-distribution, such as data from another source [20]. From a central DP perspective, an increasing amount of research has focused on leveraging public data to facilitate private learning, where public data mainly serves two purposes. On one hand, the knowledge learned from public data is implicitly transferred into the private model. Empirical investigations have demonstrated the effectiveness of pretraining on public data and fine-tuning privately on sensitive data [40, 26, 25, 38]. By gradient pre-conditioning with a subspace computed by public data, [41, 39, 23] managed to reduce the required amount of noise in differentially private gradient descent and accelerate its convergence. Through unlabeled public data, [27, 28] fed knowledge privately into student models. On the other hand, on public data, we can conduct procedures that would be infeasible without access to the raw private data. For example, [8] used parameters computed by public data to standardize private data, which can augment the sample complexity of private mean estimation. Recently, [34] employed unlabeled public data to estimate the leading eigenvalue of the covariance matrix, resulting in an improved sample complexity for generalized linear models with non-interactive local differential privacy.

The paper focuses on the problem of nonparametric regression with LDP. While regression has been extensively studied in the central setting [31, 1, 10], the LDP case remains rarely explored. A notable reason is that most gradient-based methods [31, 1] are prohibited. In order to protect privacy, each data holder needs to compute the gradient of parameters locally, which requires a large amount of memory, computing power, and communication capacity on the terminal machine [32]. [18] proposed to impose Laplace noise on the data directly to provide privacy. However, this method is known to converge slowly [17] and suffer from the curse of dimensionality. More recently, [6, 19] investigated histogram-based approaches. Though theoretically optimal, histograms may perform poorly in practice, especially when the dimension of feature space is large. Thus, both methods proposed in [6, 19] face challenges when applied to real-world problems.

Under such background, using the idea of borrowing public data information, we propose an LDP non-parametric regression algorithm called the *Locally differentially Private Decision Tree* (LPDT) that achieves both optimal convergence rate and superior empirical performance. We first create a tree partition on the public dataset using the proposed *max-edge* rule. According to the partition, each data holder encodes the private data and releases the encoding which is processed using the proposed privacy mechanism. Finally, the curator aggregates the information in each partition cell and outputs a decision tree estimator. LPDT is advantageous from at least two perspectives: (i) LPDT integrates both benefits to leverage public data. It enables adaptive partitioning procedures which can eliminate some redundant cells and can transfer information from public data to private estimation through the tree partition. (ii) It inherits the merit of the decision tree model, such as interpretability, efficiency, stability, extensiveness to multiple feature types, and resistance to the curse of dimensionality.

We summarize our contributions. (i) For the first time, we propose to use public data in locally differentially private non-parametric regression. (ii) We propose a novel LDP regression algorithm called the locally differentially private decision tree that achieves theoretical optimality while maintaining satisfying practical performance. (iii) Under mild assumptions on the similarity between the distribution of public and private data, we establish the optimal convergence rate of LPDT, whereas the supremum of excess risk of LPDT without public fails to converge to zero. This demonstrates the theoretical advantage of incorporating public data. (iv) In experiments, we compare LPDT with other existing non-parametric LDP regression methods using both synthetic and real-world datasets. Our results demonstrate the overwhelming performance of LPDT, which illustrates the empirical improvement brought by public data. Moreover, we show that LPDT performs well even in the presence of significant disparities between public and private data.

2 Methodology

This section is dedicated to the methodology of LPDT. In Section 2.1, we first present notations and preliminaries related to regression problems, followed by a recap of the definition of local differential privacy. Next, we introduce our hybrid privacy mechanism for general partition-based estimation in Section 2.2. In Section 2.3, we propose our partition rule. Finally, in Section 2.4, we provide a comprehensive description of LPDT.

2.1 Preliminaries

Notations For any vector x, let x^i denote the i-th element of x. Recall that for $1 \leq p < \infty$, the L_p-norm of $x = (x^1, \ldots, x^d)$ is defined by $\|x\|_p := (|x^1|^p + \cdots + |x^d|^p)^{1/p}$. Throughout this paper, we use the notation $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ to denote that there exist positive constant c and c' such that $a_n \leq cb_n$ and $a_n \geq c'b_n$, for all $n \in \mathbb{N}$. In addition, we denote $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Let $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. Besides, for any set $A \subset \mathbb{R}^d$, the dimameter of A is defined by $\mathrm{diam}(A) := \sup_{x,x' \in A} \|x - x'\|_2$. Let the standard Laplace random variable have the continuous probability density function $p(x) = \frac{1}{2}e^{-|x|}$ for $x \in \mathbb{R}$.

Regression is to predict the value of an unobserved output variable Y based on the observed input variable X, based on a dataset $D := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ consisting of n i.i.d. observations drawn from an unknown probability measure P on $\mathcal{X} \times \mathcal{Y} = [0,1]^d \times [-M, M]$. The density function of P is denoted as p. In addition, we have a public dataset $D^{pub} := \{(X_1^{pub}, Y_1^{pub}), \ldots, (X_{n_q}^{pub}, Y_{n_q}^{pub})\}$ drawn from distribution Q on $\mathcal{X} \times \mathcal{Y}$ with sample size n_q. Its density function is denoted as q.

It is legitimate to consider the least square loss $L : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ defined by $L(x, y, f(x)) := (y - f(x))^2$ for our target of regression. Then, for a measurable decision function $f : \mathcal{X} \to \mathcal{Y}$, the risk is defined by $\mathcal{R}_{L,\mathrm{P}}(f) := \int_{\mathcal{X}\times\mathcal{Y}} L(x, y, f(x))\, d\mathrm{P}(x, y)$. The Bayes risk, which is the smallest possible risk with respect to P and L, is given by $\mathcal{R}^*_{L,\mathrm{P}} := \inf\{\mathcal{R}_{L,\mathrm{P}}(f) | f : \mathcal{X} \to \mathcal{Y} \text{ measurable}\}$. The function that achieves the Bayes risk is called Bayes function, namely, $f^*(x) := \mathbb{E}\big(Y|X = x\big)$.

Definition 2.1 (Local Differential Privacy). Given data $\{(X_i, Y_i)\}_{i=1}^n$, each (X_i, Y_i) is mapped to a piece of privatized information s_i which is a random variable on $\mathcal{S}$. Let $\sigma(\mathcal{S})$ be the σ-field on $\mathcal{S}$. s_i is drawn conditional on (X_i, Y_i) via the distribution $\mathrm{R}\,(S \mid X_i = x, Y_i = y)$ for $S \in \sigma(\mathcal{S})$. Then the mechanism R provides *ε-local differential privacy* (ε-LDP) if

$$\sup\left\{\frac{\mathrm{R}\,(S \mid X_i = x, Y_i = y)}{\mathrm{R}\,(S \mid X_i = x', Y_i = y')} \mid S \in \sigma(\mathcal{S}), \text{ and } x, x' \in \mathcal{X}, \quad y, y' \in \mathcal{Y}\right\} \le e^{\varepsilon}.$$

This formulation of local privacy is widely adopted [13, 6]. In contrast to central DP where the likelihood ratio is taken with respect to some statistics of all data, LDP requires individuals to guarantee their own privacy by considering the likelihood ratio with respect to each (X_i, Y_i). Once the view s is provided, no further processing can reduce the deniability about taking a value (x, y) since any outcome s is nearly as likely to have come from some other initial value (x', y').

2.2 Privacy mechanism for tree partition

This section focuses on the hybrid privacy mechanism for general tree partitions. We first introduce the standard regression tree and then present our privacy mechanism based on the random response and Laplacian mechanism.

For index set $\mathcal{I}$, let $\pi = \{A_j\}_{j\in\mathcal{I}}$ be any tree partition of $\mathcal{X}$ with $\cup_{j\in\mathcal{I}} A_j = \mathcal{X}$ and $A_i \cap A_j = \emptyset$, $i \neq j$. For any $x \in \mathcal{X}$, let the cell containing x be $A(x)$. A *population decision tree regressor* with partition π is defined as

$$\overline{f}_{\pi}(x) = \frac{\sum_{j\in\mathcal{I}} \mathbf{1}\{x \in A_j\} \int_{A_j} f^*(x')\, d\mathrm{P}(x')}{\sum_{j\in\mathcal{I}} \mathbf{1}\{x \in A_j\} \int_{A_j} d\mathrm{P}(x')}. \tag{1}$$

Here, we let $0/0 = 0$ by definition. To get a empirical estimator given the data set $D = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, we estimate the numerator and the denominator of (1) respectively. To estimate the denominator, each sample (X_i, Y_i) contributes a one-hot vector $U_i \in \{0,1\}^{|\mathcal{I}|}$ where the j-th element of U_i is $\mathbf{1}\{X_i \in A_j\}$. Then an estimation of $\int_{A_j} d\mathrm{P}(x)$ is $\frac{1}{n}\sum_{i=1}^n U_i^j$, which is the number of samples in A_j divided by n. Analogously, an estimation of $\int_{A_j} f^*(x) d\mathrm{P}(x)$ is $\frac{1}{n}\sum_{i=1}^n Y_i \cdot U_i^j$. Combining the pieces, a *decision tree regressor* is defined as

$$f_{\pi}(x) = \frac{\sum_{j\in\mathcal{I}} \left(\mathbf{1}\{x \in A_j\} \sum_{i=1}^n Y_i \cdot U_i^j\right)}{\sum_{j\in\mathcal{I}} \left(\mathbf{1}\{x \in A_j\} \sum_{i=1}^n U_i^j\right)}. \tag{2}$$

In other words, $f_{\pi}(x)$ estimates $f(x)$ by the average of the responses in the cell $A(x)$. In the non-private setting, each data holder prepares U_i and Y_i according to the partition π and sends it to the curator. Then the curator aggregates the transmission following (2).

To protect the privacy of each data, we propose to estimate the numerator and denominator of the population regression tree using a privatized method. Specifically, given U_i^j, we independently sample $\tilde{U}_i^j$ using the random response technique [35]

$$\tilde{U}_i^j = \begin{cases} U_i^j - \frac{1}{1+e^{\varepsilon/4}} & \text{with probability } \frac{e^{\varepsilon/4}}{1+e^{\varepsilon/4}} \\ 1 - U_i^j - \frac{1}{1+e^{\varepsilon/4}} & \text{with probability } \frac{1}{1+e^{\varepsilon/4}}. \end{cases} \tag{3}$$

Since $\mathbb{E}_{\mathrm{R}}\left[\frac{1}{n}\sum_{i=1}^n \tilde{U}_i^j\right] = \frac{e^{\varepsilon/4}-1}{e^{\varepsilon/4}+1}\frac{1}{n}\sum_{i=1}^n U_i^j$, we take $\frac{e^{\varepsilon/4}+1}{e^{\varepsilon/4}-1}\frac{1}{n}\sum_{i=1}^n \tilde{U}_i^j$ as the estimator of $\int_{A_j} d\mathrm{P}(x)$. To privatize $Y_1, \cdots, Y_n$, we use the standard Laplace mechanism [15]. Namely, we let

$$\tilde{Y}_i = Y_i + \frac{4M}{\varepsilon}\xi_i \tag{4}$$

where ξ_i are i.i.d. standard Laplace random variables. Similarly, $\frac{e^{\varepsilon/4}+1}{e^{\varepsilon/4}-1}\frac{1}{n}\sum_{i=1}^n \tilde{Y}_i \cdot \tilde{U}_i^j$ can be used to estimate $\int_{A_j} f^*(x)d\mathrm{P}(x)$. Then using the privatized information $(\tilde{U}_i, \tilde{Y}_i), i = 1, \cdots, n$, we define the *locally differentially private decision tree regressor* as

$$f_\pi^{\mathrm{DP}}(x) = \frac{\sum_{j\in\mathcal{I}}\left(\mathbf{1}\{x\in A_j\}\sum_{i=1}^n \tilde{Y}_i \cdot \tilde{U}_i^j\right)}{\sum_{j\in\mathcal{I}}\left(\mathbf{1}\{x\in A_j\}\sum_{i=1}^n \tilde{U}_i^j\right)}. \tag{5}$$

Compared to [6, 19] which used the Laplacian mechanism to protect both U_i and Y_i, our mechanism (3) considers the fact that U is a binary vector. When $|\mathcal{I}|$ is large, (3) can be more efficient than the Laplace mechanism which has a heavier-tailed distribution [13, 14].

2.3 Max-edge partition with variance reduction

While our privacy mechanism applies to any tree partition, it can be challenging to use general partitions such as the original CART [9] for theoretical analysis. In the following, we propose a new splitting rule called the *max-edge partition rule* using the variance reduction criterion. This rule is amenable to theoretical analysis and can also achieve satisfactory practical performance. Given public dataset $\{(X_i^{pub}, Y_i^{pub})\}_{i=1}^{n_q}$, the partition rule is stated as follows:

- Let $A_0^1 := [0,1]^d$ be the initial rectangular cell and $\pi_0 := \{A_0^j\}_{j\in\mathcal{I}_0}$ be the initialized cell partition. $\mathcal{I}_0 = \{1\}$ stands for the initialized index set. In addition, let $p \in \mathbb{N}$ represent the maximum depth of the tree and let n_l represent the minimum sample size in each leaf. These parameters are fixed beforehand by the user and possibly depend on n.
- Suppose we have obtained a partition π_{i-1} of $\mathcal{X}$ after $i-1$ steps of the recursion. Let $\pi_i = \emptyset$. In the i-th step, for each $A_{i-1}^j \in \pi_{i-1}, j \in \mathcal{I}_{i-1}$, suppose it is $\times_{\ell=1}^d [a_\ell, b_\ell]$. We choose the edge to be split among the longest edges. The index set of longest edges is defined as

$$\mathcal{M}_{i-1}^j = \left\{k \mid |b_k - a_k| = \max_{\ell=1,\cdots,d} |b_\ell - a_\ell|,\ k = 1, \cdots, d\right\}.$$

- Assume we split along the ℓ-th dimension for $\ell \in \mathcal{M}_{i-1}^j$, A_{i-1}^j is then partitioned into a left sub-cell $A_{i-1}^{j,0}(\ell)$ and a right sub-cell $A_{i-1}^{j,1}(\ell)$ along the midpoint of the chosen dimension, where $A_{i-1}^{j,0}(\ell) = \left\{x \mid x \in A_{i-1}^j, x^\ell < \frac{a_\ell+b_\ell}{2}\right\}$ and $A_{i-1}^{j,1}(\ell) = A_{i-1}^j / A_{i-1}^{j,0}(\ell)$. Then the dimension to be split is chosen using the variance reduction criterion:

$$\underset{\ell\in\mathcal{M}_{i-1}^j}{\arg\min} \sum_{i=1}^{n_q}\left(Y_i^{pub} - f_{\pi_{i-1}\cup A_{i-1}^{j,0}(\ell)\cup A_{i-1}^{j,1}(\ell)/A_{i-1}^j}(X_i^{pub})\right)^2. \tag{6}$$

- Once ℓ is selected, We count the number of samples in the sub-cells $\sum_{i=1}^n \mathbf{1}(X_i^{pub} \in A_{i-1}^{j,k}(\ell)), k = 0, 1$. If either of the cells contains fewer than n_l samples, the splitting is pruned and we let $\pi_i = \pi_i \cup A_{i-1}^j$. Otherwise, let $\pi_i = \pi_i \cup \{A_{i-1}^{j,0}(\ell), A_{i-1}^{j,1}(\ell)\}$.

The complete process is presented in Algorithm 1 in the appendix. For each grid, the partition rule selects the midpoint of the longest edges that achieves the largest variance reduction. This procedure continues until there are not enough samples contained in any leaf node, or the depth of the tree reaches its limit.

2.4 Decision Tree with local differential privacy

With these preparations, we finally present the full procedure of LPDT in Algorithm 1.

Algorithm 1: Locally differentially private decision tree (LPDT)

Input: Private data $D = \{(X_i, Y_i)\}_{i=1}^n$, public data $D^{pub} = \{(X_i^{pub}, Y_i^{pub})\}_{i=1}^{n_q}$
Parameters: Depth p, minimum leaf sample size n_l.
Curator create tree partition π following max-edge rule in Section 2.3 on public data D^{pub}.
Data holders of D create privatized information (3) and (4) according to π.
Curator aggregates the privatized information and compute f_π^{DP} by (5).
Output: The LPDT estimator f_π^{DP}.

3 Theoretical results

In this section, we present our theoretical results and related comments. We first provide the ε-LDP guarantee of LPDT in Section 3.1. In Section 3.2, we establish the optimal convergence rate of LPDT with max-edge partition and the excess risk lower bound of LPDT without public data. Finally, we discuss the complexity of LPDT in Section 3.3.

3.1 Privacy guarantee for LPDT

Theorem 3.1. *Let $\pi = \{A_j\}_{j \in \mathcal{I}}$ be any partition of $\mathcal{X}$ with $\cup_{j\in\mathcal{I}} A_j = \mathcal{X}$ and $A_i \cap A_j = \emptyset$, $i \neq j$. Then the privacy mechanism $\mathrm{R}(\tilde{U}, \tilde{Y} | X, Y)$ defined in* (3) *and* (4) *is ε-LDP. Consequently, the LPDT estimator f_π^{DP} in Algorithm 1 is ε-LDP.*

3.2 Convergence rate of LPDT

We first present the necessary assumptions on the distribution P and Q.

Assumption 3.2. Let $\alpha \in (0, 1]$. Assume the regression function $f : \mathcal{X} \to \mathbb{R}$ is α-Hölder continuous, i.e. there exists a constant $c_L > 0$ such that for all $x_1, x_2 \in \mathcal{X}$, $|f(x_1) - f(x_2)| \leq c_L \|x_1 - x_2\|^\alpha$. Also, assume that the density function of P is upper bounded, i.e. $p(x) \leq \overline{c}$ for some $\overline{c} > 0$.

Assumption 3.3. We assume that there exists some constant $\tau > 1$ such that for all cells $A \in \pi$, there holds $\tau^{-1} \int_A d\mathrm{Q}_X(x) \leq \int_A d\mathrm{P}_X(x) \leq \tau \int_A d\mathrm{Q}_X(x)$.

Assumption 3.2 is a standard condition widely used in non-parametric statistics. Assumption 3.3 depicts the similarity between the distribution of public data and private data. It is also a mild assumption and requires only the probabilities in each cell under P_X and Q_X to be similar. When $p(x)$ and $q(x)$ are both bounded from 0, this assumption is satisfied. Alternatively, it suffices to require that $p(x)/q(x)$ is upper and lower bounded.

Theorem 3.4. *Let f_π^{DP} be the LPDT estimator in Algorithm 1. Suppose Assumption 3.2 and 3.3 hold. Then, for $n_q \gtrsim n^{\frac{d}{2\alpha+2d}}$, if we set $p \asymp \log n\varepsilon^2$ and $n_l \asymp n_q/2^p$, there holds*

$$\mathcal{R}_{L,\mathrm{P}}(f_\pi^{\mathrm{DP}}) - \mathcal{R}_{L,\mathrm{P}}^* \lesssim \left(\frac{\log n}{n\varepsilon^2} \right)^{\frac{\alpha}{\alpha+d} \wedge \frac{1}{3}}$$

with probability $1 - 2/n_q^2 - 5/n^2$ with respect to $\mathrm{P}^n \otimes \mathrm{Q}^{n_q} \otimes \mathrm{R}^n$ where R^n is the joint distribution of privacy mechanisms in (3) *and* (4).

Note that we only require $n_q \gtrsim n^{\frac{d}{2\alpha+2d}}$, which means the sample size of public data can be much smaller than private data. As illustrated in [19], the minimax convergence rate over Hölder function space is $(n(e^\varepsilon - 1)^2)^{-\frac{\alpha}{\alpha+d}}$, indicating that LPDT attains optimal rate when $\alpha/(\alpha + d) \leq 1/3$. In the case $\alpha/(\alpha + d) > 1/3$, or equivalently $2\alpha > d$, LPDT achieves fast yet sub-optimal convergence rate $n^{-\frac{1}{3}}$. Note that $2\alpha > d$ only when $d = 1$ and $\alpha \geq 1/2$, which rarely occurs. The next statement shows that LPDT fails without public data.

Theorem 3.5. *Let f_π^{DP} be the LPDT estimator in Algorithm 1 and $\mathcal{P}$ be the class of distributions satisfying Assumption 3.2. For $n_q = 0$ i.e. there is no public data, for any choice of p, n_l, and ε, there holds*

$$\sup_{\mathrm{P} \in \mathcal{P}} \left(\mathbb{E}\left[\mathcal{R}_{L,\mathrm{P}}(f_\pi^{\mathrm{DP}}) \right] - \mathcal{R}_{L,\mathrm{P}}^* \right) \gtrsim 1.$$

Under the same hypothesis function space, the supremum of excess risk of LPDT does not even converge without public data. Together with Theorem 3.4, this shows that the prior information contained in public data can greatly enhance the quality of the private estimation.

We compare our results with those of others. LPDT converges faster than deconvolution-based method [18] whose rate is $n^{-\frac{2\alpha}{2\alpha+5d}}$ [17]. As for histogram-based methods, [6] achieves the optimal rate only when the density function is lower bounded, which is a strong condition. To avoid the condition, [19] derived an *ad hoc* estimator by adding a regularization to the marginal density estimation. LPDT takes another approach to avoid the condition. It does not apply any regularization or truncation on the estimator in each cell. Instead, as long as Assumption 3.3 holds, the low-density regions can be identified and treated with larger cells automatically by the parameter n_l. As a sacrifice, the large cells restrict the approximation ability of LPDT and the convergence rate is no more than $n^{-1/3}$. In addition, our theoretical results hold in the sense of "with high probability", which is more closely related to practical needs than "in expectation" as addressed in [6, 19].

3.3 Complexity analysis

We demonstrate that LPDT is an efficient method. We first consider the average computation complexity of LPDT. The training stage consists of two parts. The partition procedure takes $\mathcal{O}(pn_q d)$ time and the computation of (5) takes $\mathcal{O}(pn)$ time. From the proof of Theorem 3.4, we know that $2^p \asymp \left(n\varepsilon^2/\log n\right)^{-\frac{d}{2\alpha+2d}}$. Thus the training stage complexity is around $\mathcal{O}(n\log n\varepsilon^2 + n_q d\log n\varepsilon^2)$. Since each prediction of the decision tree takes $\mathcal{O}(p)$ time, the test time for each test instance is around $\mathcal{O}(\log n\varepsilon^2)$. As for storage complexity, since LPDT only requires the storage of the tree structure and the prediction value at each node, the space complexity of LPDT is $\mathcal{O}(\left(n\varepsilon^2/\log n\right)^{-\frac{d}{2\alpha+2d}})$. In short, LPDT is an efficient method with a small number of parameters.

Table 1: Comparison of complexities of LDP regression methods.

	LPDT	PHIST [6]	DECONV [18]
Training Time Complexity	$\mathcal{O}(n\log n\varepsilon^2 + n_q d\log n\varepsilon^2)$	$\mathcal{O}(nd\log n\varepsilon^2)$	-
Testing Time Complexity	$\mathcal{O}(\log n\varepsilon^2)$	$\mathcal{O}(\log n\varepsilon^2)$	$\mathcal{O}(nd)$
Space Complexity	$\mathcal{O}(\left(n\varepsilon^2/\log n\right)^{\frac{d}{2\alpha+2d}})$	$\mathcal{O}(\left(n\varepsilon^2/\log n\right)^{\frac{d}{2\alpha+2d}})$	$\mathcal{O}(nd)$

We compare the complexities of LPDT with other LDP regression methods in Table 1. Notably, [18] is inefficient due to its unacceptable test and space complexity. Also, the dominant term of training complexity of [6] is $\mathcal{O}(nd\log n\varepsilon^2)$. When d is large, we can choose a small n_q such that LPDT yields a strictly lower complexity than [6]. In addition, although [6] enjoys the same order of space complexity as LPDT, the memory of histogram-based methods suffers from the curse of dimensionality in practice. Since the storage of $\mathcal{O}(h^{-d})$ values is required, even $h = 1/2$ requires allocating an array of size 2^d, which is problematic for large d. In contrast, LPDT can resist the curse of dimensionality by only splitting along the relevant features and keeping a small number of nodes.

4 Experiments

In the experiments, we first validate our theoretical findings with synthetic data in Section 4.1. Then, in Section 4.2, we show the superior performance of LPDT on real-world datasets with identically distributed public data. In Section 4.3, we apply LPDT to Chicago taxi data to show that LPDT performs well even with considerable differences between private and public data. Also, we analyze the influence of the distribution shift between private and public data on LPDT.

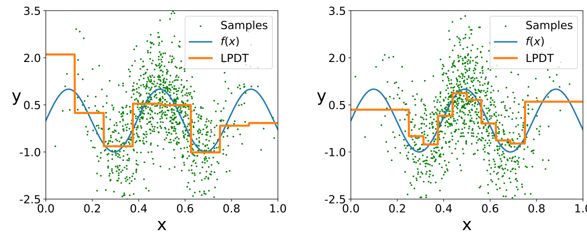
Splitting rule Note that most tree methods design their partition rules based on the information gained from the data. To boost the performance of LPDT, we also incorporate the variance reduction scheme from the original CART [9] to the tree construction. We denote the LPDT estimator using the max-edge partition rule with variance reduction in Section 2.3 as LPDT-M and denote the estimator using the standard variance reduction rule in [9] as LPDT-V. Since Theorem 3.1 holds for any partition, LPDT-V is also ε-LDP.

Experiment setup Following [21], we choose the privacy budget $\varepsilon \in [2, 8]$. We compare LPDT-M and LPDT-V with the following methods: (*i*) Private Histogram (PHIST) [6]. (*ii*) Adjusted Private

Histogram (APHIST) [19]. (*iii*) Deconvolution Kernel (DECONV) [18]. Introduction to the methods and all implementation details are presented in Appendix E.1. We employ 5-fold cross-validation for parameter selection, and techniques for tuning parameters under LDP are discussed in Section E.2. The evaluation metric is the mean squared error (MSE).

4.1 Synthetic experiments

Necessity of public data To demonstrate intuitively why public data is essential for LPDT, we first visualize its estimation on a synthetic model, $Y = \sin(16X) + \epsilon$ where $X \sim \mathcal{N}(0.5, 0.025)$ and $\varepsilon \sim \mathcal{N}(0, 1)$. In this case, the marginal distribution is highly imbalanced with the majority of samples located in the middle part of $[0, 1]$ and a few samples on the sides. For $\epsilon = 8$, we fit two LPDT models: one with 500 public data and 7,000 private data, and another with 8,000 private data.

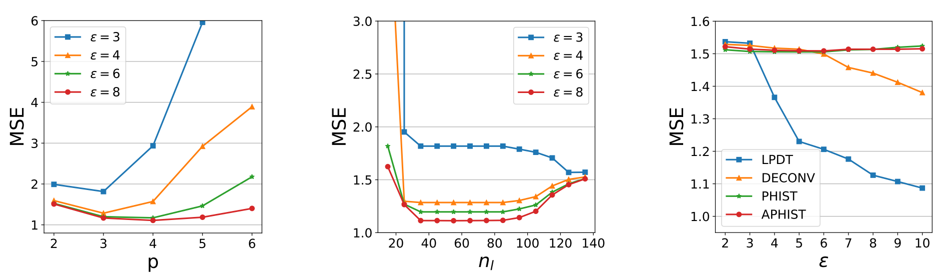


(a) Without public data (b) With public data

Figure 1: The estimated regression curve of LPDT with and without public data. 1,000 samples are displayed in green.

As shown in Figure 1(a), without public data, LPDT struggles with the imbalanced marginal. The grids on the side produce unstable predictions since only a few samples fall into them. As a result, LPDT tends to decrease depth p to stabilize its estimation. This leads to underfitting in the middle so that the predicted curve fails to capture the variation of the ground truth. In contrast, with the aid of public data, LPDT solves the issue as shown in Figure 1(b). For the middle zone where samples are abundant, LPDT creates small grids to enlarge approximation capacity. Meanwhile, it prunes the grids on the sides to ensure stability. Even with fewer private data, the MSE of LPDT is reduced from 1.19 to 1.08 thanks to the additional public data. In summary, the experiment provides empirical evidence supporting the theoretical findings in Theorems 3.4 and 3.5, which highlight the necessity of public data for the effective performance of LPDT.

Parameter analysis of depth p We conduct experiments to investigate the selection of partition depth p in terms of MSE. We generate 6,000 training samples, 2,000 test samples, and 2,000 public samples following the synthetic model described above. We pick $\varepsilon \in \{3, 4, 6, 8\}$ and $p \in \{2, 3, 4, 5, 6\}$. For each pair of p and ε, we plot the 20 times averaged MSE versus p. The result is displayed in Figure 2(a). Apparently, for each ε, as p increases, MSE first decreases until p reaches a certain value. Then MSE begins to increase as p grows. This further confirms the trade-off observed in Theorem 3.4. Moreover, the depth p at which the test error is minimized increases as ε increases. This is compatible with theory since the optimal choice of $p \asymp \log n\varepsilon^2$ is monotonically increasing with respect to ε.



(a) p - MSE (b) n_l - MSE (c) ε - MSE

Figure 2: Different parameters versus MSE.

Parameter analysis of minimum leaf sample size n_l We conduct experiments to investigate the choice of n_l in terms of MSE. Following the same generating scheme, we choose $\varepsilon \in \{3, 4, 6, 8\}$ and plot MSE of LPDT versus n_l for $n_l \in \{15, 25, \cdots, 135\}$. In Figure 2(b), the relation between MSE and n_l is U-shaped under each ε, which indicates that a properly chosen n_l is necessary as stated in Theorem 3.4. Furthermore, LPDT achieves the best MSE for $n_l \in [35, 75]$ when $\varepsilon = 4, 6, 8$, while the minimum MSE occurred at $n_l = 125$ when $\varepsilon = 3$. This finding is compatible with Theorem 3.4 which states that the optimal choice of n_l is monotonically decreasing with respect to ε.

Our parameter analyses indicate that decreasing ε favors smaller values of p and larger values of n_l. These choices lead to a decision tree partition with fewer grids. In summary, when facing higher levels of privacy demand, LPDT cuts down the number of grids to stabilize its estimation.

Privacy utility trade-off We analyze how privacy budget ε influences the quality of prediction. Under the same setup as above, we evaluate LPDT and other methods for $\varepsilon \in \{2, 3, \cdots, 10\}$ with 50 repetitions. The results are displayed in Figure 2(c). When ε increases, the MSE of LPDT decreases much faster than the other methods. Note that the MSE of both PHIST and APHIST remains high, suggesting that their performances are limited by the histogram instead of the privacy mechanism.

4.2 Real data comparison with identically distributed public data

Experiment setup We conduct experiments on 12 real datasets, each repeated 50 times with a ratio of 1:7:2 for public data, training data, and testing data in each trial. The dataset details and pre-processing steps are summarized in Appendix E.3. To ensure significance, we adopt the Wilcoxon signed-rank test [36] with a significance level of 0.05 to check if a result is significantly better. For better comparison, we also train a decision tree (denoted as DT) on the original training data with no privacy protection, whose result will serve as a lower bound.

Table 2: Average MSE over real data sets for LDP regression methods. The best results are **bolded** and the second best results are underlined. The marked results with significance towards the rest results are marked with $*$. Due to memory limitation, PHIST and APHIST are corrupted on two datasets which are marked with -.

	DT	$\varepsilon = 2$					$\varepsilon = 6$				
		LPDT-M	LPDT-V	APHIST	PHIST	DECONV	LPDT-M	LPDT-V	APHIST	PHIST	DECONV
ABA	5.67e+0	**1.01e+1**	1.01e+1	1.89e+1	1.06e+1	1.01e+7	8.38e+0*	**7.34e+0***	2.05e+1	1.05e+1	1.09e+1
AIR	2.26e+1	4.80e+1*	**4.69e+1***	1.31e+3	6.80e+1	3.00e+2	4.49e+1*	**3.60e+1***	1.60e+3	4.98e+1	4.72e+1
ALG	2.12e-2	2.57e-1	**2.43e-1**	2.52e-1	2.52e-1	9.26e+4	**2.44e-1**	2.46e-1	2.63e-1	2.47e-1	3.14e-1
AQU	1.92e+0	2.99e+0*	2.99e+0*	4.01e+0	**2.93e+0***	5.74e+3	2.73e+0*	**2.67e+0***	4.75e+0	2.83e+0	2.96e+0
BUI	1.75e+5	**1.50e+6***	1.64e+6*	-	-	1.20e+9	1.44e+6*	**1.31e+6***	-	-	2.04e+7
CBM	4.08e-27	2.12e+0*	**1.65e+0***	9.53e+0	6.97e+0	2.37e+3	7.62e-1*	**1.23e-1***	4.94e+0	3.21e+0	1.23e+5
CCP	2.19e+1	1.50e+2*	**1.06e+2***	2.07e+4	3.64e+2	3.03e+2	8.42e+1*	**5.18e+1***	2.24e+4	3.28e+2	2.56e+2
CON	9.38e+1	2.94e+2*	**2.89e+2***	3.81e+2	3.00e+2	2.24e+7	2.44e+2*	**2.13e+2***	4.16e+2	2.96e+2	3.13e+2
CPU	2.15e+1	3.41e+2	**9.00e+1***	9.26e+2	3.42e+2	2.15e+5	3.02e+2*	**6.15e+1***	9.98e+2	3.40e+2	3.98e+2
FIS	1.07e+0	2.15e+0*	**2.14e+0***	3.14e+0	2.22e+0	3.47e+3	**1.65e+0***	1.76e+0*	3.60e+0	2.16e+0	2.21e+0
HOU	2.11e+1	**8.10e+1***	8.22e+1*	1.06e+2	8.52e+1	1.92e+4	7.43e+1*	**7.10e+1***	1.23e+2	8.21e+1	2.44e+2
MUS	3.00e+2	3.47e+2*	**3.46e+2***	-	-	9.50e+3	**3.27e+2***	3.27e+2*	-	-	8.09e+3
RED	4.76e-1	7.08e-1*	**7.03e-1***	3.18e+0	7.57e-1	1.23e+8	6.75e-1*	**6.12e-1***	3.80e+0	7.12e-1	8.66e-1
WHI	5.77e-1	8.30e-1	8.42e-1	4.01e+0	**8.15e-1**	1.64e+7	7.03e-1*	**6.61e-1***	4.45e+0	8.03e-1	1.47e+0

Performance of accuracy and running time The representative results for $\varepsilon = 2, 6$ are displayed in Table 2. It can be seen that LPDT-M and LPDT-V both significantly outperform their competitors. All methods achieve a higher MSE than DT, while the results for LPDT-M and LPDT-V are reasonably close to DT. Due to memory limitations, PHIST and APHIST fail on two datasets. We also compare the total running time in Table 3 in Appendix E.4. In general, both LPDT-M and LPDT-V achieve less running time than PHIST and APHIST, and are significantly faster than DECONV.

4.3 Real data comparison with non-identically distributed public data

We apply LPDT to the *Chicago Taxi Dataset*, a collection of taxi trips in Chicago provided by the Differential Privacy Temporal Map Challenge and contains sensitive information [12]. We use the fare of each trip as labels and other information as features. Then we regard trips paid by PR card and credit card as public data and private data, respectively. In Appendix E.5, we show the two parts are distributed differently. After preprocessing, the dataset has 101 features with 2,150,565 samples in private data and 24,436 samples in public data.

Performance We report the averaged MSE of LPDT over 20 repetitions for $\varepsilon = 4, 6$, and 8. As a comparison to LPDT, we train two decision trees separately using the public and private datasets with no privacy protection. As for the comparison methods, DECONV fails due to the large sample size while PHIST and APHIST fail due to the dimensionality. However, after reducing the dimensionality by retaining only the continuous features, we are able to apply PHIST and APHIST. The results are displayed in Table 3. A first observation is that the decision tree trained on public data yields considerably worse results than the decision tree trained on private data. This suggests that relying solely on public data leads to biased predictions. Learning solely from public data achieves an MSE higher than LPDT for all three values of ε. Even for $\varepsilon = 4$, LPDT significantly outperforms histogram-based methods with $\varepsilon = 8$. This indicates that LPDT remains effective even when substantial disparities exist between the distributions of public and private data.

Table 3: Average MSE and standard deviation over Chicago taxi data.

DT		LPDT-M			PHIST	APHIST
Public	Private	$\varepsilon = 4$	$\varepsilon = 6$	$\varepsilon = 8$	$\varepsilon = 8$	$\varepsilon = 8$
3.71	0.80	3.35 (0.33)	2.86 (0.10)	2.70 (0.10)	17.22 (0.00)	38.22 (0.01)

How does non-identically distributed public data help? It is counterintuitive that non-identically distributed public data can benefit private estimation. We show the logic behind using such public data by investigating the first split feature in the tree partition. LPDT identifies whether a trip ends in Zone 32 as an important feature for predicting fare and initiates the recursive partitioning process by splitting along this feature. Figure 3 illustrates that trips ending in this zone generally have lower fares for both public and private data, although the actual fares differ significantly between the two datasets. This observation demonstrates that despite having different distributions, public and private data may exhibit similar patterns, thereby allowing the partition created on public data to still be effective on private data. Following this line of reasoning, to determine whether public data is suitable for a specific private task, we can examine whether the qualitative relationships between labels and features remain consistent across both datasets.

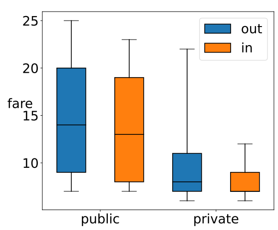
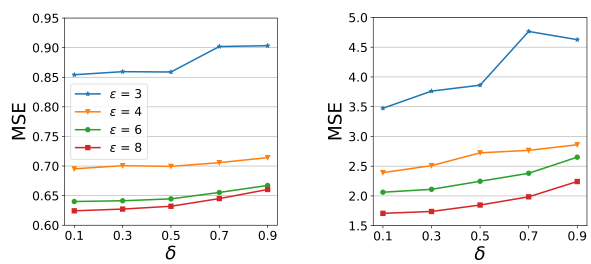


Figure 3: Box plot for fares of trips dropping in and out of Zone 32.



(a) Wine data

(b) Chicago taxi data

Figure 4: Portion of public data versus MSE of LPDT.

Analysis of distribution shift In Table 3, the MSE of LPDT reduces as ε increases but remains higher than the MSE achieved by training a decision tree directly on private data. The performance gap can be attributed to the distribution shift between private data and public data. In the following, we investigate how the difference between the two datasets affects the performance of LPDT. Besides the Chicago taxi data, we also adopt *White Wine* and *Red Wine* data in Section 4.2 as private and public data, respectively. The datasets contain the same variables but are distributed differently, as is investigated in transfer learning literature [29]. On both datasets, we combine part of the private samples with public data and perform the partition procedure on the combined dataset, with the portion of public data denoted as δ. When δ is small, there is less difference between data used for partition and training. We report the average MSE of LPDT for $\delta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ after 20 repetitions. Figure 4 shows that a small δ leads to a lower MSE for both datasets and all values of ε. Thus, LPDT is more powerful when the public and private data are distributed similarly, which also justifies Assumption 3.3.

5 Conclusion

This paper addresses the challenge of effectively performing LDP regression given both public data and private data by introducing the locally private decision tree. Due to the novel idea of leveraging public data, LPDT is accurate, efficient, and interpretable. Theoretically, we establish the privacy guarantee and optimal convergence rate of LPDT. Compared with the supremum of convergence rate without public data, we show the theoretical advantage of adopting public data. In experiments, we show the superior performance of LPDT regardless of the disparities between private and public data.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] John M Abowd. The us census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2867–2867, 2018.

[3] Noga Alon, Raef Bassily, and Shay Moran. Limits of private learning with access to public data. *Advances in neural information processing systems*, 32, 2019.

[4] Brendan Avent, Yatharth Dubey, and Aleksandra Korolova. The power of the hybrid model for mean estimation. *Proceedings on Privacy Enhancing Technologies*, (4):48–68, 2020.

[5] Raef Bassily, Albert Cheu, Shay Moran, Aleksandar Nikolov, Jonathan Ullman, and Steven Wu. Private query release assisted by public data. In *International Conference on Machine Learning*, pages 695–703. PMLR, 2020.

[6] Thomas B Berrett, László Györfi, and Harro Walk. Strongly universally consistent nonparametric regression and classification with privatised data. *Electronic Journal of Statistics*, 15:2430–2453, 2021.

[7] Tom Berrett and Yi Yu. Locally private online change point detection. *Advances in Neural Information Processing Systems*, 34:3425–3437, 2021.

[8] Alex Bie, Gautam Kamath, and Vikrant Singhal. Private estimation with public data. In *Advances in Neural Information Processing Systems*, 2022.

[9] L Breiman, JH Friedman, R Olshen, and CJ Stone. Classification and regression trees. 1984.

[10] T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021.

[11] Fida Kamal Dankar and Khaled El Emam. Practicing differential privacy in health care: A review. *Trans. Data Priv.*, 6(1):35–67, 2013.

[12] Marie Douriez, Harish Doraiswamy, Juliana Freire, and Cláudio T Silva. Anonymizing nyc taxi data: Does it matter? In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pages 140–148. IEEE, 2016.

[13] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.

[14] John C Duchi, Michael I Jordan, and Martin J Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

[15] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[16] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

[17] Jianqing Fan and Young K Truong. Nonparametric regression with errors in variables. *The Annals of Statistics*, pages 1900–1925, 1993.

[18] Farhad Farokhi. Deconvoluting kernel density estimation and regression for locally differentially private data. *Scientific Reports*, 10(1):21361, 2020.

[19] László Györfi and Martin Kroll. On rate optimal private regression under local differential privacy. *arXiv preprint arXiv:2206.00114*, 2022.

[20] Jihun Hamm, Yingjun Cao, and Mikhail Belkin. Learning privately from multiparty data. In *International Conference on Machine Learning*, pages 555–563. PMLR, 2016.

[21] Apple Inc. Differential privacy technical overview. Technical Report Apple Inc., 2017.

[22] Matthew Joseph, Janardhan Kulkarni, Jieming Mao, and Steven Z Wu. Locally private gaussian estimation. *Advances in Neural Information Processing Systems*, 32, 2019.

[23] Peter Kairouz, Monica Ribero Diaz, Keith Rush, and Abhradeep Thakurta. (nearly) dimension independent private erm with adagrad rates via publicly estimated subspaces. In *Conference on Learning Theory*, pages 2717–2746. PMLR, 2021.

[24] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. *Advances in neural information processing systems*, 27, 2014.

[25] Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[26] Harsh Mehta, Abhradeep Thakurta, Alexey Kurakin, and Ashok Cutkosky. Large scale transfer learning for differentially private image classification. *arXiv preprint arXiv:2205.02973*, 2022.

[27] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representations*, 2017.

[28] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Ulfar Erlingsson. Scalable private learning with pate. In *International Conference on Learning Representations*, 2018.

[29] Noam Segev, Maayan Harel, Shie Mannor, Koby Crammer, and Ran El-Yaniv. Learn on source, refine on target: A model transfer learning framework with random forests. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1811–1824, 2016.

[30] Hyejin Shin, Sungwook Kim, Junbum Shin, and Xiaokui Xiao. Privacy enhanced matrix factorization for recommendation with local differential privacy. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1770–1782, 2018.

[31] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pages 245–248. IEEE, 2013.

[32] Florian Tramèr, Gautam Kamath, and Nicholas Carlini. Considerations for differentially private learning with large-scale public pretraining. *arXiv preprint arXiv:2212.06470*, 2022.

[33] Di Wang and Jinhui Xu. Principal component analysis in the local differential privacy model. *Theoretical computer science*, 809:296–312, 2020.

[34] Di Wang, Huangyu Zhang, Marco Gaboardi, and Jinhui Xu. Estimating smooth GLM in non-interactive local differential privacy model with public unlabeled data. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 1207–1213. PMLR, 16–19 Mar 2021.

[35] Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[36] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

[37] Xiaotong Wu, Lianyong Qi, Jiaquan Gao, Genlin Ji, and Xiaolong Xu. An ensemble of random decision trees with local differential privacy in edge computing. *Neurocomputing*, 485:181–195, 2022.

[38] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[39] Da Yu, Huishuai Zhang, Wei Chen, and Tie-Yan Liu. Do not let privacy overbill utility: Gradient embedding perturbation for private learning. In *International Conference on Learning Representations*, 2021.

[40] Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Large scale private learning via low-rank reparametrization. In *International Conference on Machine Learning*, pages 12208–12218. PMLR, 2021.

[41] Yingxue Zhou, Steven Wu, and Arindam Banerjee. Bypassing the ambient dimension: Private sgd with gradient subspace identification. In *International Conference on Learning Representations*, 2021.