Locally Private Estimation with Public Features

Yuheng Ma
School of Statistics,
Renmin University of China

Ke Jia
School of Statistics,
Renmin University of China

Hanfang Yang
Center for Applied Statistics,
School of Statistics,
Renmin University of China

Abstract

We initiate the study of locally differentially private (LDP) learning with public features. We define semi-feature LDP, where some features are publicly available while the remaining ones, along with the label, require protection under local differential privacy. Under semi-feature LDP, we demonstrate that the mini-max convergence rate for non-parametric regression is significantly reduced compared to that of classical LDP. Then we propose `HistOfTree`, an estimator that fully leverages the information contained in both public and private features. Theoretically, `HistOfTree` reaches the mini-max optimal convergence rate. Empirically, `HistOfTree` achieves superior performance on both synthetic and real data. We also explore scenarios where users have the flexibility to select features for protection manually. In such cases, we propose an estimator and a data-driven parameter tuning strategy, leading to analogous theoretical and empirical results.

1 Introduction

Data privacy regulations such as Europe's General Data Protection Regulation (GDPR) (European Parliament and Council of the European Union) have led to a notable rise in the significance of privacy-preserving machine learning (Cummings and Desai, 2018). As a golden standard, local differential privacy (LDP) (Kairouz et al., 2014; Duchi et al., 2018), a variant of differential privacy (DP) (Dwork et al., 2006), has gained considerable attention in recent years, especially among industry experts (Erlingsson et al., 2014; Inc., 2017). LDP assumes that each sample is possessed by a data holder, who privatizes their data before it is collected by the curator. Offering a stronger sense of privacy protection compared to central DP, LDP often encounters much more sophisticated challenges which obstruct both the theoretical analysis and practical implementation of LDP learning.

Fortunately, in some scenarios, a protocol with weakened protection can be adopted to privatize only the sensitive part of each sample. For instance, label differential privacy (Ghazi et al., 2021; Malek Esmaeili et al., 2021; Badanidiyuru Varadaraja et al., 2024) assumes only the labels contain sensitive information, while the features are freely accessible to the data users. Moreover, clear theoretical advantages have been established (Wang and Xu, 2019; Xu et al., 2023; Zhao et al., 2024b) for label LDP over classical LDP. Recent works demonstrate the effectiveness of partially protecting a small portion of features (Curmei et al., 2023; Shen et al., 2023; Krichene et al., 2024; Chua et al., 2024), mostly from a methodological perspective. Yet, no theoretical framework has been established under this setting, which is the primary aim of this work.

As is addressed by GDPR, a properly functioning consent management process should provide granular consent options regarding the specific types of data being collected and the purposes for which it will be processed (Nouwens et al., 2020). In our case, to meet the requirements, users must be able to choose which parts of their data are collected under privacy protections and which are not. This necessitates the design of estimators with personalized privacy preferences, i.e., the private and public features for each user are manually selected rather than preset.

Given this context, we study the problem of statistical estimation under local differential privacy with user-personalized public features. We choose non-parametric regression as the task. An extension to classification is straightforward, and our mechanism also directly implies an optimal protocol for density estimation. Our

Proceedings of the 28^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

contributions are summarized as follows:

- We formalize the problem of LDP learning with public features. Specifically, we define semi-feature LDP, where a portion of the features is publicly available, while the remaining ones, along with the label, require protection. Moreover, we establish the first minimax lower bound under semi-feature LDP for non-parametric regression.
- We propose the `HistOfTree` estimator for both aligned (where the position of public features for all users is identical) and personalized (where users select their own public features) privacy preferences. Additionally, we provide a data-driven parameter selection rule.
- We demonstrate the superiority of `HistOfTree`. Theoretically, we provide an excess risk upper bound for personalized privacy preferences, indicating a mini-max optimal convergence rate in the aligned case. Empirically, we show that `HistOfTree` significantly outperforms naive competitors on both synthetic and real datasets.

2 Related Work

Public Data and Features There has been a long line of work about the benefit of public data (Papernot et al., 2017, 2018; Liu et al., 2021a,b; Yu et al., 2021, 2022; Nasr et al., 2023; Gu et al., 2023; Fuentes et al., 2024; Wang et al., 2024). Public features are less considered. Recently, a line of work has addressed the reasonableness of providing privacy protection w.r.t. a small portion of sensitive features (Curmei et al., 2023; Ghazi et al., 2023; Shen et al., 2023; Krichene et al., 2024; Chua et al., 2024). Effective methodologies were proposed under central differential privacy, although no theoretical framework quantifying their utility has been provided. A more extreme case is label differential privacy (Chaudhuri and Hsu, 2011; Busa-Fekete et al., 2021; Ghazi et al., 2021; Malek Esmaeili et al., 2021; Cunningham et al., 2022; Ghazi et al., 2022; Badanidiyuru Varadaraja et al., 2024; Zhao et al., 2025), where all the features are regarded as public. Under central DP, this relaxation is shown to improve the generalization error with only a constant related to privacy budget (Badanidiyuru Varadaraja et al., 2024). However, under LDP, the advantage of label LDP over LDP becomes more apparent. See the established generalization gap for both parametric (Wang and Xu, 2019) and non-parametric estimation (Xu et al., 2023; Zhao et al., 2024b).

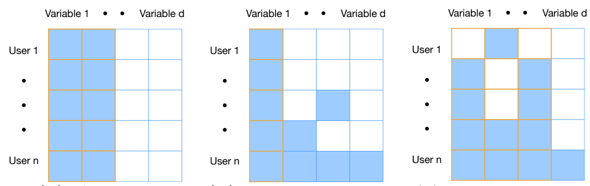
Personalized Privacy Our personalized privacy preferences exhibit heterogeneity w.r.t. both users and attributes, i.e., different attributes of the same user can be either private or public, and different users have different choices of which features to protect. Various works have studied the problem of learning under heterogeneous privacy preference w.r.t. users. Wang et al. (2015); Yang et al. (2021); Li et al. (2022) considered locally private learning, with each data holder possessing their own privacy budget. Sun et al. (2014); Song et al. (2019); Zhang et al. (2022, 2024) assume that the privacy budget of each sample is related to its value. These studies neither provide theoretical guarantees nor cover our problem setting since there is no heterogeneity w.r.t. different features in their case. Similar to our aligned setting, Amorino and Gloter (2023) proposed Componentwise LDP, where each attribute is released through independent private channels with different budgets. This is a more restrictive case than ours, as we require private features to be jointly protected. They also do not consider heterogeneous privacy w.r.t. users. Moreover, their analysis does not allow for public features, i.e., features with infinitely large budgets. More recently, Aliakbarpour et al. (2024) proposed a Bayesian Coordinate LDP framework, which is more general than ours. This framework also presents a promising approach to addressing potential risks arising from inner correlations between features, a topic not covered in this paper.

3 Methodology

This section is dedicated to the methodology of `HistOfTree`. In Section 3.1, we first present notations and preliminaries related to regression problems. We also provide the definition of semi-feature local differential privacy. Next, we introduce the `HistOfTree` estimator in Section 3.2 under aligned privacy preference. In Section 3.3, we consider the case where the privacy preference of each user is different and propose a corresponding estimator.

3.1 Preliminaries

Notations 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 $a \wedge b = \min(a, b)$. 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 $\|x\|_p := (|x^1|^p + \cdots + |x^d|^p)^{1/p}$. Let $x_{i:j} = (x_i, \cdots, x_j)$ be the slicing view of x in the $i, \cdots, j$ position. For any set $A \subset \mathbb{R}^d$, the diameter of A is defined by $\mathrm{diam}(A) := \sup_{x,x' \in A} \|x - x'\|_2$. Let $A \times B$ be the Cartesian product of sets where



(a) Aligned (b) Concentrated (c) Heavy tailed

Figure 1: Illustration of different W, where blue means $W_i^j = 1$. In the aligned case (a), all users protect the first two features. In the personalized case, users specify different features, with the protected features being concentrated in (b) and spread in (c). The yellow boundaries represent the s selected private features.

$A \in \mathcal{X}_1$ and $B \in \mathcal{X}_2$. For measure P on X_1 and Q on $\mathcal{X}_2$, define the product measure $\mathrm{P} \otimes \mathrm{Q}$ on $\mathcal{X}_1 \times \mathcal{X}_2$ as $\mathrm{P} \otimes \mathrm{Q}(A \times B) = \mathrm{P}(A) \cdot \mathrm{Q}(B)$. For an integer k, denote the k-fold product measure on $\mathcal{X}_1^k$ as P^k.

The goal of regression is to predict the value of an unobserved label of a given input X, based on a dataset $D := \{(X_i, Y_i)\}_{i=1}^n$ consisting of n i.i.d. observations drawn from an unknown probability measure P on $\mathcal{X} \times \mathcal{Y} = [0,1]^d \times [-M, M]$. 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 is the smallest possible risk with respect to P and L. The function that achieves the Bayes risk is called Bayes function, namely, $f^*(x) := \mathbb{E}\big(Y | X = x\big)$.

Semi-feature LDP We formally set up private learning with public features. For each (X_i, Y_i), the label Y_i and a subset of features of X_i require privacy protection, while the remaining features can be released freely. Let $W_i^\ell \in \{0, 1\}$ be the indicator of whether the ℓ-th feature of X_i should be protected. We assume the collection of private features is independent of P and W_i^ℓs are not necessarily protected and are publicly known. The potential risk and solutions are discussed in Section 3.4. We consider the aligned (resp. personalized) case, where the private features are identical (resp. different) for each sample. The corresponding W are illustrated in Figure 1. Denote the number of private features of X_i by s_i^*.

With a slight abuse of notation, we write $(X_i, Y_i) = (X_i^{\mathrm{priv}}, X_i^{\mathrm{pub}}, Y_i)$, where X_i^{priv} and X_i^{pub} are the collection of private and public features for X_i, respectively. Let $\mathcal{X}_i^{\mathrm{priv}}$ and $\mathcal{X}_i^{\mathrm{pub}}$ be their corresponding domain space. Note that X_i^{priv} and X_i^{pub} are not necessarily ordered. For instance, we allow $X_i^{\mathrm{priv}} = (X_i^1, X_i^3)$ and $X_i^{\mathrm{pub}} = (X_i^2, X_i^4)$. Given the notations, we rigorously define our privacy constraint.

Definition 3.1 (Semi-feature local differential privacy). Given data $\{(X_i, Y_i)\}_{i=1}^n$, a privatized information Z_i, which is a random variable on $\mathcal{S}$, is released based on (X_i, Y_i) and $Z_1, \cdots, Z_{i-1}$. Let $\sigma(\mathcal{S})$ be the σ-field on $\mathcal{S}$. Z_i is drawn conditional on (X_i, Y_i) and $Z_{1:i-1}$ via the distribution $\mathrm{R}\left(S \mid X_i^{\mathrm{priv}}, X_i^{\mathrm{pub}}, Y_i, Z_{1:i-1}\right)$ for $S \in \sigma(\mathcal{S})$. Then the mechanism $\mathrm{R} = \{\mathrm{R}_i\}_{i=1}^n$ is *ε-semi-feature local differential privacy* (*ε-semi-feature LDP*) if for all $1 \leq i \leq n, S \in \sigma(\mathcal{S})$, all $x^{\mathrm{priv}}, x^{\mathrm{priv}\prime} \in \mathcal{X}_i^{\mathrm{priv}}$, $x^{\mathrm{pub}} \in \mathcal{X}_i^{\mathrm{pub}}$, $y, y' \in \mathcal{Y}$, and $z_{1:i-1} \in \mathcal{S}^{i-1}$, there holds

$$\frac{\mathrm{R}_i\left(S \mid x^{\mathrm{priv}}, x^{\mathrm{pub}}, y, z_{1:i-1}\right)}{\mathrm{R}_i\left(S \mid x^{\mathrm{priv}\prime}, x^{\mathrm{pub}}, y', z_{1:i-1}\right)} \leq e^{\varepsilon} \tag{1}$$

Moreover, if for all $1 \leq i \leq n, S \in \sigma(\mathcal{S})$, all $x^{\mathrm{priv}}, x^{\mathrm{priv}\prime} \in \mathcal{X}_i^{\mathrm{priv}}$, $x^{\mathrm{pub}} \in \mathcal{X}_i^{\mathrm{pub}}$, $y, y' \in \mathcal{Y}$, there holds

$$\frac{\mathrm{R}_i\left(S \mid x^{\mathrm{priv}}, x^{\mathrm{pub}}, y\right)}{\mathrm{R}_i\left(S \mid x^{\mathrm{priv}\prime}, x^{\mathrm{pub}}, y'\right)} \leq e^{\varepsilon}, \tag{2}$$

then R is *non-interactive ε-semi-feature LDP*. Here, we let $0/0 = 1$.

Similar definitions are proposed under central DP setting (Shen et al., 2023; Krichene et al., 2024; Chua et al., 2024). Semi-feature LDP requires individuals to guarantee their own privacy by considering the likelihood ratio of each $(X_i^{\mathrm{priv}}, Y_i)$ as in (1) and (2). The value of X_i^{pub} is regarded as known. Once a view z is provided, no further processing can reduce the deniability about taking a value (x^{priv}, y) since any z is nearly as likely to come from other initial value $(x^{\mathrm{priv}\prime}, y')$. The definition is meaningful when no prior knowledge exists about the relationship between x^{priv} and x^{pub}. Otherwise, if x^{priv} can be inferred using x^{pub}, a mechanism satisfying Definition 3.1 may not necessarily provide sufficient protection for x^{priv}. See, for example, Aliakbarpour et al. (2024).

If $s_i^* = d$, i.e. all features are private, our definition reduces to the common LDP (Kairouz et al., 2014; Duchi et al., 2018) where non-parametric regression has been well-explored (Berrett et al., 2021; Györfi and Kroll, 2022). If $s_i^* = 0$, i.e. only the labels are private, the definition reduces to label local differential privacy (Busa-Fekete et al., 2021; Cunningham et al., 2022), which is shown to be an easier problem than pure LDP (Wang and Xu, 2019; Xu et al., 2023; Zhao et al., 2024b). Our definition explores the intermediate phase, which is practically meaningful (Krichene et al., 2024; Shen et al., 2023; Chua et al., 2024).

3.2 HistOfTree Estimator for Aligned Private Features

In this section, we introduce the HistOfTree algorithm under the aligned private features. Let $\mathcal{X}^{\text{priv}} = [0,1]^{s^*}$ and $\mathcal{X}^{\text{pub}} = [0,1]^{d-s^*}$ be the private and public domain for all i.

Privacy Mechanism We adopt a partition-based estimator, which creates disjoint partitions and predicts via the average of training labels in each partition grid. Specifically, let $\pi^{\text{priv}} = \{A_j\}$ be a partition of $\mathcal{X}^{\text{priv}}$, i.e. $\cup A_j = \mathcal{X}^{\text{priv}}$ and $A_{j_1} \cap A_{j_2} = \emptyset$, $j_1 \neq j_2$. Similarly, $\pi^{\text{pub}} = \{B_k\}$ on $\mathcal{X}^{\text{pub}}$. The partition-based estimator (which is non-private) of x in $A_j \times B_k$ is

$$f(x) = \frac{\sum_{i=1}^n Y_i \cdot \mathbf{1}_{A_j}(X_i^{\text{priv}})\mathbf{1}_{B_k}(X_i^{\text{pub}})}{\sum_{i=1}^n \mathbf{1}_{A_j}(X_i^{\text{priv}})\mathbf{1}_{B_k}(X_i^{\text{pub}})}. \tag{3}$$

The denominator represents the sample marginal distribution in $A_j \times B_k$, while the numerator corresponds to their joint distribution. Thus, their division estimates their conditional relationship. To perform such an estimation, three pieces of information are necessary from each data holder: Y_i, $\mathbf{1}_{A_j}(X_i^{\text{priv}})$, and $\mathbf{1}_{B_k}(X_i^{\text{pub}})$. The first two of these require protection. We employ the standard Laplace mechanism (Dwork et al., 2006) to protect Y_i, specifically:

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

where ξ_is are i.i.d. standard Laplace random variables. For indicator functions, we use the randomized response mechanism (Warner, 1965) and let

$$\tilde{U}_i^j = \begin{cases} C_\varepsilon\left(\mathbf{1}_{A_j}(X_i^{\text{priv}}) - \frac{1}{1+e^{\varepsilon/4}}\right) & \text{w.p. } \frac{e^{\varepsilon/4}}{1+e^{\varepsilon/4}}, \\ C_\varepsilon\left(\mathbf{1}_{A_j^c}(X_i^{\text{priv}}) - \frac{1}{1+e^{\varepsilon/4}}\right) & \text{w.p. } \frac{1}{1+e^{\varepsilon/4}}, \end{cases} \tag{5}$$

where $C_\varepsilon = \frac{e^{\varepsilon/4}+1}{e^{\varepsilon/4}-1}$. Here, A_j^c denotes the complement of A_j. Note that $\mathbb{E}_{\text{R}}\left[\tilde{U}_i^j\right] = \mathbf{1}_{A_j}(X_i^{\text{priv}})$ and thus we have an unbiased estimator of $\mathbf{1}_{A_j}(X_i^{\text{priv}})$.

Partition To formalize the private estimator, we also need to formalize the partitions π^{priv} and π^{pub}. The choice of π^{priv} is restrictive due to privacy constraints. We adopt the histogram partition, a classical technique in LDP learning (Berrett et al., 2021; Györfi and Kroll, 2022). Specifically, consider $0 = a_0 < a_1 < \cdots < a_t = 1$, a equal length partition of $[0,1]$. Then a histogram partition of $\mathcal{X}^{\text{priv}} = [0,1]^{s^*}$ is $\left\{\otimes_{k=1}^{s^*}[a_{\sigma_k}, a_{\sigma_k+1}) | 0 \leq \sigma_k \leq t\right\}$. For π^{pub}, we can fully utilize the information contained in the dataset. Given $\tilde{Y}_i$s and X_i^{pub}s, various candidate methods are available depending on the sample size, computational power, and prior information. We choose a decision tree partition for its merit of interpretability, efficiency, stability, extensiveness to multiple feature types, and resistance to the curse of dimensionality. A decision tree partition is obtained by recursively split grids into subgrids using criteria such as variance reduction. Since it can be challenging to use the original CART rule (Breiman, 1984) for theoretical analysis, we adopt the *max-edge* rule following Cai et al. (2023); Ma et al. (2023); Ma and Yang (2024). This rule is amenable to theoretical analysis and can also achieve satisfactory practical performance. For each grid, the partition rule selects the midpoint of the longest edges that achieves the largest variance reduction. This procedure continues until the depth of the tree reaches its predetermined limit. Together, the HistOfTree estimator is

$$\tilde{f}(x) = \sum_{A_j, B_k} \mathbf{1}_{A_j \times B_k}(x) \frac{\sum_{i=1}^n \tilde{Y}_i \tilde{U}_i^j \mathbf{1}_{B_k}(X_i^{\text{pub}})}{\sum_{i=1}^n \tilde{U}_i^j \mathbf{1}_{B_k}(X_i^{\text{pub}})}. \tag{6}$$

3.3 Personalized Private Features

In practice, it is unlikely that all users share the same preferences regarding which features' privacy to prioritize. For instance, while most people may consider their favorite sports as insensitive information, some individuals feel uneasy about releasing such details due to concerns about targeted advertising, particularly from e-commerce platforms. We consider W_i to be different for each i. We assume the privacy budget ε are identical across users.

Partition The first step is to manually select out the features to be treated as private. Suppose $\sum_{i=1}^n W_i^\ell$ is ranked in decreasing order, i.e., the first feature is the most protected. We select out the first s features with the largest $\sum_{i=1}^n W_i^\ell$, i.e. $1, \cdots, s$. The selection of parameter s is discussed in Section 4.3. We create a histogram partition π^{priv} on these dimensions. Then on the rest of dimensions, we use all public items to create a decision tree partition π^{pub}. Specifically, when splitting each grid, the variance reduction along the ℓ-th dimension (7) is calculated by samples with $W_i^\ell = 0$. We give the detailed algorithm for such partition rule in Algorithm 1.

Privacy Mechanism Injecting noise into the grid indices is trickier. We first define the potential grids for a sample. For X_i, let

$$\mathcal{V}_i = \Big\{ A \times B \in \pi^{\text{priv}} \otimes \pi^{\text{pub}} \mid \exists \overline{X} \in A \times B \text{ s.t. } X_i^\ell = \overline{X}^\ell \text{ if } W_i^\ell = 0 \Big\}.$$

Intuitively, $\mathcal{V}_i$ is the collection of grids where X_i could potentially be located, given no information about its

Algorithm 1: Max-edge partition rule

Input: Public data $\{X_i^{\text{pub}}\}_{i=1}^n$, privatized labels $\{\tilde{Y}_i\}_{i=1}^n$, depth p.
Initialization: $B_{0,0} = [0,1]^d$, $D_{0,0} = \{X_i^{\text{pub}}\}_{i=1}^n$, $\pi_k = \emptyset$ for $1 \le k \le p$.
for $k = 1$ ***to*** p **do**
  for $B_{k-1,j}$ **do**
    Suppose $B_{k-1,j} = \times_{\ell=1}^d [a_\ell, b_\ell]$, let $\mathcal{M}_{i-1}^j = \{k \mid |b_k - a_k| = \max_{\ell=1,\cdots,d} |b_\ell - a_\ell|\}$. `# Longest edges.`
    Let $B_{k-1,j}^0(\ell) = \left\{x \mid x \in B_{k-1,j}, x^\ell < \frac{a_\ell + b_\ell}{2}\right\}$ and $B_{k-1,j}^1(\ell) = B_{k-1,j}/B_{k-1,j}^0(\ell)$.
    Let $D_{k-1,j}(\ell) = \left\{X_i^{\text{pub}} \in D_{k-1,j} \mid W_i^\ell = 0\right\}$. `# Identify available samples along each axis.`
    Let $g(A, D)$ be the sample variance of $\tilde{Y}_i$s whose X_i^{pub} is in both A and D. Select ℓ as

$$\arg\min_{\ell \in \mathcal{M}_{i-1}^j} \; g(B_{k-1,j}^0(\ell), D_{k-1,j}(\ell)) + g(B_{k-1,j}^1(\ell), D_{k-1,j}(\ell)). \tag{7}$$

    $B_{k,2j-1} = B_{k-1,j}^0(\ell)$, $B_{k,2j} = B_{k-1,j}^1(\ell)$, $\pi_k = \pi_k \cup \{B_{k,2j-1}, B_{k,2j}\}$.
    $D_{k,2j-1} = \{X_i^{\text{pub}} \in D_{k-1,j}(\ell) \mid X_i^{\text{pub}\ell} < \frac{a_\ell + b_\ell}{2}\}$, $D_{k,2j} = D_{k-1,j}(\ell)/D_{k,2j-1}$. `# Allocate samples.`
  end
end
Output: Partition π_p

private features. For simplicity, let j be the index of $A \times B$ in the combined partition $\pi^{\text{priv}} \otimes \pi^{\text{pub}}$. We then define the privacy mechanism:

$$\tilde{Y}_i = Y_i + \frac{4M}{\varepsilon}\xi_i, \text{ and}$$
$$\tilde{V}_i^j = \begin{cases} 0 & \text{if } A \times B \notin \mathcal{V}_i, \\ C_\varepsilon \left(\mathbf{1}_{A\times B}(X_i) - \frac{1}{1+e^{\varepsilon/4}}\right) & \text{else w.p. } \frac{e^{\varepsilon/4}}{1+e^{\varepsilon/4}}, \\ C_\varepsilon \left(\mathbf{1}_{(A\times B)^c}(X_i) - \frac{1}{1+e^{\varepsilon/4}}\right) & \text{else w.p. } \frac{1}{1+e^{\varepsilon/4}}. \end{cases} \tag{8}$$

In this case, the indices of potential grids are estimated analogously to $\tilde{U}_i^j$, while the indices for the remaining grids, which are already known without private features, are zeroed out. Furthermore, we have $\mathbb{E}_{\text{P,R}}\left[\tilde{V}_i^j\right] = \mathbb{E}_{\text{P}}\left[\mathbf{1}_{A\times B}(X_i)\right]$. Consequently, we define the personalized estimator:

$$\tilde{f}(x) = \sum_{A\times B} \mathbf{1}_{A\times B}(x) \frac{\sum_{i=1}^n \tilde{Y}_i \tilde{V}_i^j}{\sum_{i=1}^n \tilde{V}_i^j} \tag{9}$$

where j is the index of $A\times B$. The following proposition shows the privacy guarantee of (8).

Proposition 3.2. *Let $\pi = \{A \times B \mid A \in \pi^{priv}, B \in \pi^{pub}\}$ be any partition of $\mathcal{X}$. Then the privacy mechanism (8) is non-interactively ε-semi-feature LDP.*

When the public features are shared across users, i.e. $\sum_{i=1}^n W_i^\ell = n$ for $1 \le \ell \le s$ and $\sum_{i=1}^n W_i^\ell = 0$ otherwise, mechanism (8) reduces to that in (4) and (5). The privacy guarantees of these mechanisms are thus implied by Proposition 3.2. In this case, the estimator (9) also recovers (6).

3.4 Potential Privacy Risk from W

In some cases, W and (X, y) are dependent, which brings additional privacy leakage. For instance, AIDS carriers may choose to conceal their status if they are diagnosed with the condition. If the carriers are aware of the mechanism and believes it could leak information, they would provide incorrect results (choose not to prevent this feature and claim no AIDS). We propose two solutions. (i) If the curator wishes to prevent such wrong data collection, it should classify highly correlated features as private beforehand, including these features among the first s^* private features. This approach requires certain prior knowledge and does not completely resolve the privacy accounting issue. (ii) Observing the estimator (9), there is no need to release W to the server. The server only needs to receive a binary vector of V. The only remaining issue is the tree partition, which could be mitigated by using random selection instead of (7). Although this reduces the effectiveness of the partition, it entirely eliminates the risk associated with W.

4 Theoretical Results

In this section, we present our theoretical results and related comments. We first provide the mini-max lower bound under semi-feature LDP with aligned privacy preference in Section 4.1. In Section 4.2, we establish the optimal convergence rate of `HistOfTree` estimator. Based on the theoretical findings, we discuss the choice of the number of private dimensions s in Section 4.3.

4.1 A Mini-max Lower Bound

We first present a necessary assumption on the distribution P, which is a standard condition widely used in non-parametric statistics.

Assumption 4.1. Assume that the density function of P is upper and lower bounded, i.e. $\underline{c} \leq d\mathrm{P}(x)/dx \leq \overline{c}$ for some $\overline{c} \geq \underline{c} > 0$. Let $\alpha \in (0, 1]$. Assume $f^* : \mathcal{X} \to \mathbb{R}$ is α-Hölder continuous, i.e. there exists $c_L > 0$ s.t. for all $x_1, x_2 \in \mathcal{X}$, $|f^*(x_1) - f^*(x_2)| \leq c_L \|x_1 - x_2\|^\alpha$.

The bounded density assumption can greatly simply the analysis, yet can be removed without hurting the conclusion, see Györfi and Kroll (2022). We present the following theorem that specifies the mini-max lower bound with semi-feature LDP under the above assumptions. The proof is based on first constructing a finite class of hypotheses and then applying information inequalities for local privacy (Duchi et al., 2018) and Assouad's Lemma (Tsybakov, 2009).

Theorem 4.2. *Denote the function class of* P *satisfying Assumption 4.1 by* $\mathcal{F}$. *Consider the aligned privacy preference where* $W_i^\ell = 1$ *if* $\ell \leq s^*$ *and* $W_i^\ell = 0$ *otherwise. Then for any estimator* $\widehat{f}$ *that is sequentially-interactive* ε*-semi-feature LDP, there holds*

$$\inf_{\widehat{f}} \sup_{\mathcal{F}} \mathbb{E}_{\mathrm{P}} \left[\mathcal{R}_{\mathrm{P}}(\widehat{f}) - \mathcal{R}_{\mathrm{P}}^* \right] \gtrsim \left(n(e^\varepsilon - 1)^2 \right)^{-\frac{2\alpha}{2\alpha+d+s^*}} \vee n^{-\frac{2\alpha}{2\alpha+d}}. \tag{10}$$

The lower bound consists of two terms. The second term, $n^{-\frac{2\alpha}{2\alpha+d}}$, represents the classical mini-max lower bound for non-private learners under Assumption 4.1 (Tsybakov, 2009). Regarding the first term, if ε is sufficiently large such that $(e^\varepsilon - 1) \gtrsim n^{\frac{s^*}{4\alpha+2d}}$, it is dominated by the second term. This is the case where the level of privacy is not significant enough to degrade the estimator. For constant level ε, it approaches the second term as s^* becomes smaller, as the learning essentially becomes non-private for $s^* = 0$. If $s^* = d$, our results reduce to the special case of LDP non-parametric regression with only private data. Previous studies (Berrett et al., 2021; Györfi and Kroll, 2022) have provided estimators that achieve the mini-max optimal convergence rate. Our results depict the optimal behavior in the intermediate zone. The performance gain brought by $d - s^*$ public features grows with the rate $\frac{2\alpha}{2\alpha+2d-(d-s^*)}$, which becomes more significant for larger values of $d - s^*$.

4.2 Convergence Rate of `HistOfTree`

In this section, we present an upper bound for the `HistOfTree` estimator under personalized privacy preferences. This bound subsequently implies the optimal rate under aligned privacy preferences.

Theorem 4.3. *Let the* `HistOfTree` *estimator* $\tilde{f}$ *be defined in* (9). *For* $\pi = \pi^{priv} \otimes \pi^{pub}$, *let* π^{priv} *be a histogram partition with* t *bins and* π^{pub} *be generated by Algorithm 1 with depth* p. *Let Assumption 4.1 holds. Define the solution to the equation*

$$p^* = \arg\min_p \frac{2^{p\frac{d+s}{d-s}} \cdot \log n}{n\varepsilon^2} \cdot \delta(p) + \ 2^{-2\alpha p/(d-s)} \tag{11}$$

where we let

$$\delta(p) = \frac{1}{n} \sum_{i=1}^n 2^{\sum_{\ell=s+1}^d W_i^\ell \cdot p/(d-s)}. \tag{12}$$

Then, let $\lambda^* = \log_2(\delta(p^*))/p^*$. *Then,* $\tilde{f}$ *is* ε*-semi-feature LDP. Moreover, for any* $\varepsilon \lesssim (n/\log n)^{\frac{s}{\alpha+d-s}}$, *there exists some choice of the parameters* $p \asymp p^* \asymp \log n\varepsilon^2$ *and* $t \asymp 2^{p^*/(d-s)}$ *such that*

$$\mathcal{R}_{L,\mathrm{P}}(\tilde{f}) - \mathcal{R}_{L,\mathrm{P}}^* \lesssim \left(\frac{\log n}{n\varepsilon^2} \right)^{\frac{2\alpha}{2\alpha+d+s+\lambda^*(d-s)}} \tag{13}$$

holds with probability $1 - 8/n^2$ *w.r.t.* $\mathrm{P}^n \otimes \mathrm{R}$ *where* R *is the joint distribution of privacy mechanisms in* (8).

The upper bound (13) is some power of $\log n / n\varepsilon^2$. The presence of the $\log n$ term comes from the high probability nature of Theorem 4.3, while the privacy budget contributes a term of ε^2. Apparently, the tighter the privacy requirement (i.e., the smaller ε), the more relaxed the upper bound becomes. Moreover, we require $\varepsilon \lesssim (n/\log n)^{\frac{s}{\alpha+d-s}}$. If ε exceeds this threshold, the privacy noise becomes negligible, and the presented upper bound might be dominated by the non-private term $n^{-\frac{2\alpha}{2\alpha+d}}$.

We comment on the theorem regarding the selected p^* and the term $\delta(p^*)$. The final rate (13) involves λ^*, a manufactured quantity representing the impact of the denser tail of the privacy mask matrix W. Generally, for a fixed p, if a user i chooses to protect more features in $s+1, \cdots, d$, the quantity $\sum_{\ell=s+1}^d W_i^\ell$ becomes larger, and $\delta(p)$ also increases. In this case, the selected p^* decreases, whereas the final $\delta(p^*)$ increases (see proof for details). This suggests that, for more demanding privacy protection, the optimal depth p^* (as well as the number of histogram bins t, since $t \asymp \log n\varepsilon^2$) should be reduced, meaning fewer bins are preferable to maintain estimation stability. Additionally, since $2^{\lambda^* p^*} = \delta(p^*)$, the value of λ^* is larger, subsequently increasing the upper bound. This confirms the intuition, as protecting more features inherently necessitates a looser bound.

Under the aligned privacy preference, where $W_i^\ell = 0$ for $\ell \geq s^* + 1$, Theorem 4.3 has a direct implication as outlined in the following corollary.

Corollary 4.4. *Consider the aligned privacy preference where $W_i^\ell = 1$ if $\ell \le s^*$ and $W_i^\ell = 0$ otherwise. Then, under the same conditions and settings to Theorem 4.3, let $p \asymp p^* \asymp \log n\varepsilon^2$ and $t \asymp 2^{p^*/(d-s^*)}$. There holds*

$$\mathcal{R}_{L,\mathrm{P}}(\tilde{f}) - \mathcal{R}^*_{L,\mathrm{P}} \lesssim \left(\frac{\log n}{n\varepsilon^2}\right)^{\frac{2\alpha}{2\alpha+d+s^*}} \tag{14}$$

with probability $1 - 8/n^2$ with respect to $\mathrm{P}^n \otimes \mathrm{R}$.

Compared to Theorem 4.2, the `HistOfTree` estimator with the best parameter choice reaches the mini-max lower bound. We note that when ε is large, there is a gap between $(e^\varepsilon - 1)^2$ as in the lower bound (10) and ε^2 in the upper bound. The gap is commonly observed (Duchi et al., 2018; Györfi and Kroll, 2022; Ma et al., 2023; Ma and Yang, 2024; Ma et al., 2024; Zhao et al., 2024a).

4.3 Selection of s

In this section, we explore the selection of the parameter s within the context of personalized privacy preferences. Unlike scenarios with aligned preferences, where an intrinsic value of $s = s^*$ is implied, the personalized case requires the explicit specification of s as a hyper-parameter. When considering the final convergence rate (13), it is noted that choosing either a significantly large or small value for s inevitably leads to a large denominator of the rate, impacting the efficiency of the convergence. From the proof of Theorem 4.3, we recognize that the right-hand side of (11) is a high probability upper bound of the excess risk. As a result, we tend to choose the s that minimizes the upper bound. Together with the selection of p, the parameter tuning is specified as an optimization problem

$$s, p^* = \arg\min_{s,p} \frac{2^{p\frac{d+s}{d-s}} \cdot \log n}{n\varepsilon^2} \cdot \delta(p) + \; 2^{-2p/(d-s)} \tag{15}$$

where the unknown α is replaced by 1, i.e. the Lipschitz continuous case. Since there are at most $d \log n$ possible candidates for (s, p^*), this can be done via a brute force search. The minimum, analogous to Theorem 4.3, is $\left(\log n/n\varepsilon^2\right)^{\frac{2}{2+d+s+\lambda^*(d-s)}}$. Clearly, if the private features are concentrated across users, i.e. most $\sum_{\ell=s+1}^d W_i^\ell$ are small, λ^* will eventually be small. As a result, s leans to be small. In the opposite, if private features are less concentrated (long tail), a large s is selected to ensure a broader privacy coverage. This behavior is illustrated in Figure 1.

There are also some interesting corner cases of the selection procedure. First, the solution is consistent with the aligned case. Specifically, selecting any $s' > s^*$ leads to $\lambda^* = 0$. As a result, the rate has $\frac{2}{2+d+s'} < \frac{2}{2+d+s^*}$. For $s' < s^*$, the rate remains $\frac{2}{2+d+s^*}$, meaning that choosing small s' is theoretically equivalent, although it may result in less effective partitions in practice. (15) also covers the case of locally private estimation with public data (Ma et al., 2023), where $W_i^\ell = 1$ if $i \le n - n_q$ and $W_i^\ell = 0$ otherwise. In this case, the object in (15) yields a λ^* strictly smaller than 1, and $s = 0$ is the optimal choice. Our estimator reduces to the locally private decision tree proposed by Ma et al. (2023).

5 Experiments

To demonstrate the superiority of proposed methods and to validate our theoretical findings, we conduct experiments on both synthetic and real datasets in Section 5.1 and 5.2, respectively.

The tested methods include: **(i)** `HistOfTree` is the proposed estimator. In addition to the partition rule proposed in Algorithm 1, we boost the performance by incorporating the criterion reduction scheme from the original CART (Breiman, 1984), similar to Ma et al. (2023); Ma and Yang (2024); **(ii)** `AdHistOfTree` is identical to `HistOfTree` except that the choice of s, p, and t are selected according to Section 4.3. **(iii)** `Hist` is the locally private histogram estimation (Berrett et al., 2021; Györfi and Kroll, 2022) which treats all features and the label as private; **(iv)** `KRR` is the obfuscation approach that directly perturb X_i^j using k-generalized randomized response if $W_i^j = 1$. It then fit with a decision tree. **(v)** `ParDT` estimates by adding Laplace noise to labels and then fitting a decision tree based on the noisy label and the public part of the samples. **(vi)** `LabelDT` estimates by adding Laplace noise to labels and then using a decision tree. It serves as the label LDP benchmark. **(vi)** `DT` is the non-private decision tree and serves as the non-private benchmark. Implementation details of all these methods are provided in Appendix C.

The evaluation metric in all experiments is the mean squared error (MSE). For each model, we report the best result over its parameter grids, with the best result determined based on the average of at least 50 replications. The size of the parameter grids is selected based on running time to ensure that each method incurs an equal amount of computation. All experiments are conducted on a machine with 72-core Intel Xeon 2.60GHz and 128GB of main memory. The illustrative codes can be found on GitHub[1].

[1] `https://github.com/Karlmyh/LDP-PublicFeatures`

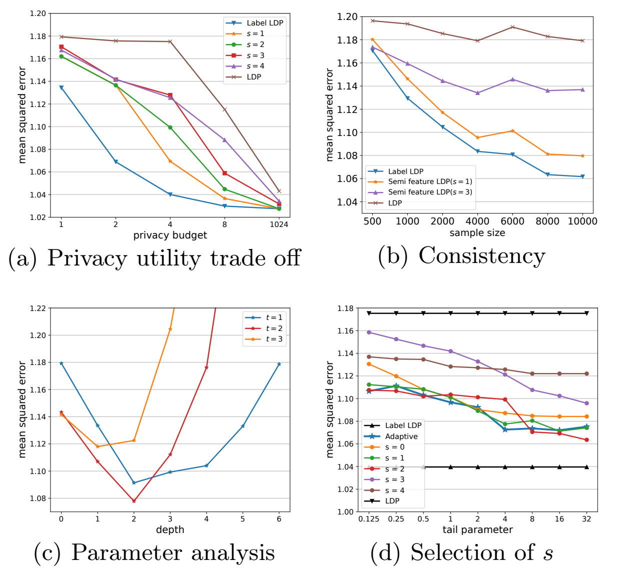


Figure 2: Experiment results on synthetic data. `LabelDT` and `Hist` are captioned as label LDP and LDP, respectively. `HistOfTree` is captioned with specific choice of parameters s and t. In 2(a), we apply uneven scaling to the x-axis to accommodate the outlying value of 1024, representing the non-private performance. In 2(d), `AdHistOfTree` is captioned as adaptive.

5.1 Simulation

In the simulation, we consider the following distribution P. The marginal distributions P_X is the uniform distributions on $\mathcal{X} = [0,1]^5$. The label is generated via $Y = \sin(X^1) + \sin(X^2) + \sin(X^5) + \sigma$, where σ is a standard Gaussian random variable. It can be easily verified that the constructed distributions above satisfy Assumption 4.1. For mask matrix W, we always let $\sum_{i=1}^n W_i^\ell \geq \sum_{i=1}^n W_i^{\ell'}$ if $\ell \leq \ell'$. In this case, useful information is contained in both public and private features, which necessitates our method.

Privacy Utility Trade-off We analyze how privacy budget ε influences the prediction quality under different numbers of public features. For $1 \leq s^* \leq 4$, we consider $n = 10000$ and vary ε. The results are displayed in Figure 2(a). When ε increases, all MSEs decrease, while the performance steadily improves as the number of public features increases. This observation is compatible with both Theorem 4.3 and the notion that semi-feature LDP serves as the intermediate stage of LDP and label LDP.

Consistency We analyze the consistency of (6) as n grows. We take $\varepsilon = 4$. The results are displayed in Figure 2(b), showing that the MSE is diminishing as n grows, while the rate is faster for small s.

Parameter Analysis We conduct experiments to investigate the influence of key parameters p and t. We consider $n = 10000$, $\varepsilon = 4$, and $s^* = 2$. As shown in Figure 2(c), for each t, as p increases, MSE exhibits a U-shaped curve. Also, the best performance is achieved with $t = 2$, indicating such a phenomenon is also true for t. This further emphasizes the importance of choosing correct hyperparameters as outlined in Theorem 4.3. Moreover, the depth p at which the test error is minimized increases as t increases. This is compatible with theory since $t \asymp 2^{p/(d-s^*)}$.

Selection of s under Different Behaviors of W We investigate how the behavior of W affects the selection of s. Specifically, we use W_γ parameterized by γ, where $W_{\gamma i}^\ell = 1$ if $i \leq n/\ell^\gamma$. A smaller γ implies a thicker tail of W. We compare the performance of `HistOfTree` with different choices of s, as well as `AdHistOfTree` whose s is selected automatically. In Figure 2(d), the observations are two-fold. First, MSE decreases w.r.t. γ, which aligns with Theorem 4.3 that a thicker tail of W leads to a larger upper bound. On the other hand, one s does not consistently outperform another under different thicknesses of the tail, which necessitates the discussions about W in Section 4.2 and 4.3. The best s lies between 1 and 2, while `AdHistOfTree` always achieves approximately the best performance. This illustrates the effectiveness of the selection rule in (15).

5.2 Real Data

We conduct experiments on 10 real-world datasets that are available online. These datasets cover a wide range of sample sizes and feature dimensions commonly encountered in real-world scenarios. Some of these datasets contain both sensitive and insensitive features, making them suitable for our case. Privacy preferences follow two paradigms. For both paradigms, we rank the features subjectively according to their level of sensitivity, from sensitive to non-sensitive. In the aligned case, we select out the first $s^* = \lceil \log \sqrt{d} \rceil$ features. In the personalized case, we set $W_i^\ell = 1$ if $i \leq n/10^{\lfloor \ell/s^* \rfloor}$ and 0 otherwise. A summary of key information for these datasets and pre-processing details is in Appendix C.2.

We first compute the mean squared error over 50 random train-test splits for $\varepsilon = 1, 2, 4$. To standardize the scale across datasets, we report the MSE ratio relative to non-private fitting a decision tree over the whole training samples. The results for $\varepsilon = 2$ are displayed in Table 1, while the others are presented in Appendix C.3. For all privacy budgets, the proposed methods significantly outperform competitors in terms of both average performance (rank sum) and the number of best results achieved. The data-driven selected pa-

Table 1: Real data performance for $\varepsilon = 2$. To ensure significance, we employ the Wilcoxon signed-rank test (Wilcoxon, 1992) with a significance level of 0.05 to determine if a result is significantly better. The best results are **bolded** and those holding significance towards the rest results are marked with $*$. ME and CART stand for max-edge and the classical CART rule, respectively.

	Aligned					Personalized						
	`HistOfTree`		`ParDT`	`Hist`	`KRR`	`HistOfTree`		`AdHistOfTree`		`ParDT`	`Hist`	`KRR`
	ME	CART				ME	CART	ME	CART			
ABA	1.85	1.65	**1.63**	1.98	1.81	**1.61***	1.80	1.61	1.78	1.65	1.98	1.77
AQU	1.55	**1.47***	1.71	1.63	2.05	**1.60***	1.60	**1.60***	1.60	1.75	1.63	2.05
BUI	1.48	**1.27***	1.51	1e4	2.43	**1.42**	1.47	1.43	1.47	1.59	1e4	2.41
DAK	6.69	**6.65***	8.03	8.34	9.57	6.73	6.73	**6.57***	**6.57***	8.85	8.34	9.57
FIS	**1.88***	2.01	2.14	2.16	2.35	2.09	2.16	**2.07**	2.14	2.23	2.16	2.42
MUS	**1.13***	**1.13***	1.26	2e5	1.44	1.13	1.13	**1.12**	1.13	1.26	2e5	1.45
POR	3.15	**3.04***	3.32	4.23	4.02	3.08	3.08	**2.98***	**2.98***	3.35	4.23	3.92
PYR	**1.48***	**1.48***	1.65	3e3	2.29	1.49	1.49	**1.42***	**1.42***	1.77	3e3	2.30
RED	1.44	**1.37***	1.45	1.55	1.67	1.44	**1.43**	1.44	1.44	1.46	1.55	1.67
WHI	1.42	**1.41**	**1.41**	1.53	1.50	1.42	**1.39***	1.44	1.44	1.42	1.53	1.49
Rank Sum	22	**15**	28	45	43	20	27	**15**	23	48	62	62

rameters achieve comparable performance to the best parameter, indicating the effectiveness of the selection rule. In some cases, `AdHistOfTree` performs slightly better, largely due to the incompleteness of parameter grids of `HistOfTree`. Moreover, we observe that the CART rule outperforms ME in the aligned setting, while the opposite holds in the personalized setting. This is attributed to the reliance on sufficient public features of CART.

6 Limitations

The study is a first step towards locally private learning with public features. There are several limitations that can serve as a guide for future work. The current lower bound can be further explored to see whether Theorem 4.3 is tight. Under the current scheme, we shall have $\sum_{i=1}^n \mathcal{E}^{\frac{2\alpha+d+\sum_{j=1}^d W_i^j}{2\alpha}} \geq \frac{1}{(e^\varepsilon-1)^2}$ for lower bound, where $\mathcal{E}$ is the excess risk. For the upper bound, we have $\sum_{i=1}^n \mathcal{E}^{-\frac{2\alpha+d+\sum_{j=1}^d W_i^j}{2\alpha}} \geq n^2\varepsilon^2$. They are matched similarly as in the article if $\sum_{j=1}^d W_i^j$ are all equal, which includes the aligned case as a special case but is far more restrictive than the personalized case.

During the rebuttal, one of the reviewers raised a concern that the privacy preferences used in the real-data experiments are subjective rather than derived from actual users. We acknowledge this limitation. Real-world data containing genuine privacy preferences is extremely rare and typically only obtainable through online testing. Such data must include both the value of a sensitive feature and the user's preference for privatizing it. However, (1) collecting such data is uncommon, and (2) missing values may arise when users choose not to disclose sensitive features. Following the reviewer's suggestion, we attempted to collect data using a public survey service. However, the collected data proved to be too noisy, and further data collection is not feasible due to limited funding. As a result, no additional experiments or datasets are included as contributions to this paper.

Acknowledgments

Hanfang Yang is the corresponding author. The authors would like to thank the reviewers for their constructive comments, which led to a significant improvement in this work. The research is supported by the Special Funds of the National Natural Science Foundation of China (Grant No. 72342010). Yuheng Ma is supported by the Outstanding Innovative Talents Cultivation Funded Programs 2024 of Renmin University of China. This research is also supported by Public Computing Cloud, Renmin University of China.

References

Maryam Aliakbarpour, Syomantak Chaudhuri, Thomas A Courtade, Alireza Fallah, and Michael I Jordan. Enhancing feature-specific data protection via bayesian coordinate differential privacy. *arXiv preprint arXiv:2410.18404*, 2024.

Chiara Amorino and Arnaud Gloter. Minimax rate for multivariate data under componentwise local differential privacy constraints. *arXiv preprint arXiv:2305.10416*, 2023.

Ashwinkumar Badanidiyuru Varadaraja, Badih Ghazi, Pritish Kamath, Ravi Kumar, Ethan Leeman, Pasin

Manurangsi, Avinash V Varadarajan, and Chiyuan Zhang. Optimal unbiased randomizers for regression with label differential privacy. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

L Breiman. Classification and regression trees. *The Wadsworth & Brooks/Cole*, 1984.

Robert Istvan Busa-Fekete, Umar Syed, Sergei Vassilvitskii, et al. Population level privacy leakage in binary classification wtih label noise. In *NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021.

Yuchao Cai, Yuheng Ma, Yiwei Dong, and Hanfang Yang. Extrapolated random tree for regression. In *Proceedings of the 40th International Conference on Machine Learning*, pages 3442–3468. PMLR, 2023.

Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 155–186. JMLR Workshop and Conference Proceedings, 2011.

Lynn Chua, Qiliang Cui, Badih Ghazi, Charlie Harrison, Pritish Kamath, Walid Krichene, Ravi Kumar, Pasin Manurangsi, Krishna Giri Narra, Amer Sinha, et al. Training differentially private ad prediction models with semi-sensitive features. *arXiv preprint arXiv:2401.15246*, 2024.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4):547–553, 2009.

Rachel Cummings and Deven Desai. The role of differential privacy in gdpr compliance. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, page 20, 2018.

Teddy Cunningham, Konstantin Klemmer, Hongkai Wen, and Hakan Ferhatosmanoglu. Geopointgan: Synthetic spatial data with local label differential privacy. *arXiv preprint arXiv:2205.08886*, 2022.

Mihaela Curmei, Walid Krichene, Li Zhang, and Mukund Sundararajan. Private matrix factorization with public item features. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 805–812, 2023.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

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

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.

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.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. URL `https://data.europa.eu/eli/reg/2016/679/oj`.

Miguel Fuentes, Brett C Mullins, Ryan McKenna, Gerome Miklau, and Daniel Sheldon. Joint selection: Adaptively incorporating public information for private synthetic data. In *International Conference on Artificial Intelligence and Statistics*, pages 2404–2412. PMLR, 2024.

Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. Deep learning with label differential privacy. *Advances in neural information processing systems*, 34:27131–27145, 2021.

Badih Ghazi, Pritish Kamath, Ravi Kumar, Ethan Leeman, Pasin Manurangsi, Avinash Varadarajan, and Chiyuan Zhang. Regression with label differential privacy. In *The Eleventh International Conference on Learning Representations*, 2022.

Badih Ghazi, Ravi Kumar, Pasin Manurangsi, and Thomas Steinke. Algorithms with more granular differential privacy guarantees. In *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251, page 54, 2023.

Xin Gu, Gautam Kamath, and Zhiwei Steven Wu. Choosing public datasets for private machine learning via gradient subspace distance. *arXiv preprint arXiv:2303.01256*, 2023.

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

Apple Inc. Differential privacy technical overview. Technical Report Apple Inc., 2017. URL `https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf`.

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

Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy.

In *International Conference on Machine Learning*, pages 2436–2444. PMLR, 2016.

Michael R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer, New York, 2008.

Walid Krichene, Nicolas E Mayoraz, Steffen Rendle, Shuang Song, Abhradeep Thakurta, and Li Zhang. Private learning with public features. In *International Conference on Artificial Intelligence and Statistics*, pages 4150–4158. PMLR, 2024.

Xiaoguang Li, Haonan Yan, Zelei Cheng, Wenhai Sun, and Hui Li. Protecting regression models with personalized local differential privacy. *IEEE Transactions on Dependable and Secure Computing*, 20(2):960–974, 2022.

Terrance Liu, Giuseppe Vietri, Thomas Steinke, Jonathan Ullman, and Steven Wu. Leveraging public data for practical private query release. In *International Conference on Machine Learning*, pages 6968–6977. PMLR, 2021a.

Terrance Liu, Giuseppe Vietri, and Steven Z Wu. Iterative methods for private synthetic data: Unifying framework and new methods. *Advances in Neural Information Processing Systems*, 34:690–702, 2021b.

Yi-Cheng Liu and I-Cheng Yeh. Using mixture design and neural networks to build stock selection decision support systems. *Neural Computing and Applications*, 28:521–535, 2017.

Yuheng Ma and Hanfang Yang. Optimal locally private nonparametric classification with public data. *Journal of Machine Learning Research*, 25(167):1–62, 2024.

Yuheng Ma, Han Zhang, Yuchao Cai, and Yang Hanfang. Decision tree for locally private estimation with public data. *Advances in Neural Information Processing Systems*, 2023.

Yuheng Ma, Ke Jia, and Hanfang Yang. Better locally private sparse estimation given multiple samples per user. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 33746–33776, 2024.

Mani Malek Esmaeili, Ilya Mironov, Karthik Prasad, Igor Shilov, and Florian Tramer. Antipodes of label differential privacy: Pate and alibi. *Advances in Neural Information Processing Systems*, 34:6934–6945, 2021.

Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn, and Wes B Ford. The population biology of abalone (haliotis species) in tasmania. i. blacklip abalone (h. rubra) from the north coast and islands of bass strait. *Sea Fisheries Division, Technical Report*, 48:p411, 1994.

Milad Nasr, Saeed Mahloujifar, Xinyu Tang, Prateek Mittal, and Amir Houmansadr. Effectively using public data in privacy preserving machine learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202. PMLR, 2023.

Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. Dark patterns after the gdpr: Scraping consent pop-ups and demonstrating their influence. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13, 2020.

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.

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.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Zeyu Shen, Anilesh Krishnaswamy, Janardhan Kulkarni, and Kamesh Munagala. Classification with partially private features. *arXiv preprint arXiv:2312.07583*, 2023.

Haina Song, Tao Luo, Xun Wang, and Jianfeng Li. Multiple sensitive values-oriented personalized privacy preservation based on randomized response. *IEEE transactions on information forensics and security*, 15:2209–2224, 2019.

Chongjing Sun, Yan Fu, Junlin Zhou, Hui Gao, et al. Personalized privacy-preserving frequent itemset mining using randomized response. *The Scientific World Journal*, 2014, 2014.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009. ISBN 978-0-387-79051-0. doi: 10.1007/b13794. URL `http://hfhhc391f4815d8064db7sqbf60x0xf60c6kpo.faxb.libproxy.ruc.edu.cn/10.1007/b13794`. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Chendi Wang, Yuqing Zhu, Weijie J Su, and Yu-Xiang Wang. Neural collapse meets differential privacy: Curious behaviors of noisygd with near-perfect representation learning. In *Forty-first International Conference on Machine Learning*, 2024.

Di Wang and Jinhui Xu. On sparse linear regression in the local differential privacy model. In *International Conference on Machine Learning*, pages 6628–6637. PMLR, 2019.

Shaowei Wang, Liusheng Huang, Miaomiao Tian, Wei Yang, Hongli Xu, and Hansong Guo. Personalized privacy-preserving data aggregation for histogram estimation. In *2015 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2015.

Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. Locally differentially private protocols for frequency estimation. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 729–745, 2017.

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

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

Xingxing Xiong, Shubo Liu, Dan Li, Zhaohui Cai, and Xiaoguang Niu. A comprehensive survey on local differential privacy. *Security and Communication Networks*, 2020:1–29, 2020.

Shirong Xu, Chendi Wang, Will Wei Sun, and Guang Cheng. Binary classification under local label differential privacy using randomized response mechanisms. *Transactions on Machine Learning Research*, 2023.

Ge Yang, Shaowei Wang, and Haijie Wang. Federated learning with personalized local differential privacy. In *2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS)*, pages 484–489. IEEE, 2021.

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.

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*, 2022.

Dongyan Zhang, Lili Zhang, Zhiyong Zhang, and Zhongya Zhang. Adaptive personalized randomized response method based on local differential privacy. *International Journal of Information Security and Privacy (IJISP)*, 18(1):1–19, 2024.

Guopeng Zhang, Xuebin Chen, Yuanli Jia, and Ran Zhai. Support personalized weighted local differential privacy skyline query. *Security and Communication Networks*, 2022, 2022.

Puning Zhao, Li Shen, Rongfei Fan, Qingming Li, Huiwen Wu, Jiafei Wu, and Zhe Liu. Learning with user-level local differential privacy. *arXiv preprint arXiv:2405.17079*, 2024a.

Puning Zhao, Huiwen Wu, Jiafei Wu, and Zhe Liu. On theoretical limits of learning with label differential privacy. 2024b.

Puning Zhao, Jiafei Wu, Zhe Liu, Li Shen, Zhikun Zhang, Rongfei Fan, Le Sun, and Qingming Li. Enhancing learning with label differential privacy by vector approximation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=IwPXYk6BV9`.

Fang Zhou, Q Claire, and Ross D King. Predicting the geographical origin of music. In *2014 IEEE International Conference on Data Mining*, pages 1115–1120. IEEE, 2014.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:

(a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. [Yes]
(b) Complete proofs of all theoretical results. [Yes]
(c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Not Applicable]
(b) The license information of the assets, if applicable. [Not Applicable]
(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
(d) Information about consent from data providers/curators. [Not Applicable]
(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]
(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

In this appendix, we provide the detailed results for methodology (Appendix A), the proofs (Appendix B), and experiments (Appendix C).

A Related Methodology of Locally Differentially Private Decision Tree

A.1 Generalized Randomized Response Mechanism

In this section, we introduce the generalized randomized response mechanism (Warner, 1965; Kairouz et al., 2016; Ghazi et al., 2022) (or direct encoding (Wang et al., 2017)). Specifically, for the aligned privacy preference case, the mechanism calculates (5) by

$$\Pr\left[U_i = j\right] = \begin{cases} \frac{e^{\varepsilon/2}}{e^{\varepsilon/2}+|\pi^{\mathrm{priv}}|-1}, & \text{if } \mathbf{1}_{A_j}(X_i^{\mathrm{priv}}) \\ \frac{1}{e^{\varepsilon/2}+|\pi^{\mathrm{priv}}|-1}, & \text{otherwise.} \end{cases} \tag{16}$$

and uses $\tilde{U}_i^j = \frac{e^{\varepsilon/2}+|\pi^{\mathrm{priv}}|-1}{e^{\varepsilon/2}-1}\left(\mathbf{1}_{U_i=j} - \frac{1}{e^{\varepsilon/2}+|\pi^{\mathrm{priv}}|-1}\right)$ will serve as an estimator of $\mathbf{1}_{A_j}(X_i^{\mathrm{priv}})$, since $\mathbb{E}_{\mathrm{P,R}}\left[\tilde{U}_i^j\right] = \mathbb{E}_{\mathrm{P}}\left[\mathbf{1}_{A_j}(X_i^{\mathrm{priv}})\right]$. The estimator (6) remains identical. For the personalized case, the mechanism computes (8) by

$$\tilde{Y}_i = Y_i + \frac{4M}{\varepsilon}\xi_i, \quad \Pr\left[V_i = j\right] = \begin{cases} 0 & \text{if } A\times B \notin \mathcal{V}_i, \\ \frac{e^{\varepsilon/2}}{e^{\varepsilon/2}+|\mathcal{V}_i|-1} & \text{elif } \mathbf{1}_{A\times B}(X_i), \\ \frac{1}{e^{\varepsilon/2}+|\mathcal{V}_i|-1}, & \text{otherwise.} \end{cases} \tag{17}$$

To estimate indicator function $\mathbf{1}_{A\times B}(X_i)$, we use

$$\tilde{V}_i^j = \frac{e^{\varepsilon/2}+|\mathcal{V}_i|-1}{e^{\varepsilon/2}-1}\left(\mathbf{1}_{V_i=j} - \frac{1}{e^{\varepsilon/2}+|\mathcal{V}_i|-1}\right)\cdot \mathbf{1}_{A\times B\in\mathcal{V}_i}.$$

In this case, the indexes of potential grids are estimated analogously to $\tilde{U}_i^j$, while the remaining grids are free of privacy concern and zeroed out. The privacy guarantee of (17) is guaranteed by the following proposition.

Proposition A.1. *Let $\pi = \{A\times B \mid A\in\pi^{priv}, B\in\pi^{pub}\}$ be any partition of $\mathcal{X}$. Then the privacy mechanism* (17) *is ε-semi-feature LDP.*

The reason we introduce this mechanism is due to the fact that when $|\mathcal{V}_i|$ is small, it has been shown to be quite efficient (Wang et al., 2017; Kairouz et al., 2016; Xiong et al., 2020). In experiments where sample sizes $n\sim 1e3$ and $|\mathcal{V}_i|\sim 1e1$, we find that the generalized randomized response mechanism performs slightly better than the randomized responses introduced in the main text. However, the variance of such a mechanism grows linearly with $|\mathcal{V}_i|$, which can be troublesome in theoretical analysis where $|\mathcal{V}_i|$ is n raised to some power. Consequently, in such cases, the convergence rate fails to attain optimality.

A.2 Complexity Analysis

We analyze the complexity of `HistOfTree` from perspectives of computation, memory, and communication. The overall communication rounds for personalized estimator (9) is two. Each user first sends its privatized label $\tilde{Y}_i$ and public features X_i^{pub}, which costs communication capacity of at most $d+1$ real numbers. Based on the created partition, it then sends the binary encoding $\tilde{V}_i^j$s, which are $t^s 2^p$ bools in total.

The computation complexity of `HistOfTree` consists of two parts, creating the partition and computing the node values. The partition procedure takes $\mathcal{O}(pnd)$ time and the computation of (9) takes $\mathcal{O}(t^s 2^p)$ time. From the proof of Theorem 4.3, we know that $t^s 2^p \lesssim n\varepsilon^2$. Thus the training stage complexity is at most $\mathcal{O}(n\varepsilon^2 + nd\log n\varepsilon^2)$, which is linear in n. Since each prediction of the decision tree takes $\mathcal{O}(p)$ time, the test time for each test instance is $\mathcal{O}(\log n\varepsilon^2)$. As for storage complexity, since `HistOfTree` only requires storage of the tree structure and the node values, the space complexity is also $\mathcal{O}(t^s 2^p)$. In short, `HistOfTree` is an efficient method in terms of computation, memory, and communication.

B Proofs

B.1 Proof of Proposition 3.2

Proof of Proposition 3.2. Since our mechanism (8) does not rely on previous private information $Z_1, \cdots, Z_{i-1}$ when drawing Z_i, it is a non-interactive mechanism. Each user only reveals two pieces of information that is private, i.e. $\tilde{Y}_i$ and $\tilde{V}_i$. Note that they are conditionally independent, thus for each conditional distribution $\mathrm{R}_i\left(\tilde{Y}_i, \tilde{V}_i, | X_i^{\text{priv}} = x_i^{\text{priv}}, X_i^{\text{pub}} = x_i^{\text{pub}}, Y_i = y\right)$, we can compute the density ratio as

$$\sup_{x_i^{\text{priv}} x_i^{\text{priv}\prime}, x_i^{\text{pub}}, y, y'} \frac{\mathrm{R}_i\left(\tilde{Y}_i, \tilde{V}_i | X_i^{\text{priv}} = x_i^{\text{priv}}, X_i^{\text{pub}} = x_i^{\text{pub}}, Y_i = y\right)}{\mathrm{R}_i\left(\tilde{Y}_i, \tilde{V}_i | X_i^{\text{priv}} = x_i^{\text{priv}\prime}, X_i^{\text{pub}} = x_i^{\text{pub}}, Y_i = y'\right)}$$
$$\leq \sup_{x_i^{\text{priv}} x_i^{\text{priv}\prime}, x_i^{\text{pub}}} \frac{\mathrm{R}_i(\tilde{V}_i | X_i^{\text{priv}} = x_i^{\text{priv}}, X_i^{\text{pub}} = x_i^{\text{pub}})}{\mathrm{R}_i(\tilde{V}_i | X_i^{\text{priv}} = x_i^{\text{priv}\prime}, X_i^{\text{pub}} = x_i^{\text{pub}})} \cdot \sup_{y, y'} \frac{\mathrm{R}_i(\tilde{Y}_i | Y_i = y)}{\mathrm{R}_i(\tilde{Y}_i | Y_i = y')}. \tag{18}$$

(i) The first part associates to the randomized response mechanism. Since the conditional density is identical if x^{priv} and $x^{\text{priv}\prime}$ belongs to a same A_j, we have

$$\frac{\mathrm{R}_i(\tilde{V}_i | X_i^{\text{priv}} = x_i^{\text{priv}}, X_i^{\text{pub}} = x_i^{\text{pub}})}{\mathrm{R}_i(\tilde{V}_i | X_i^{\text{priv}} = x_i^{\text{priv}\prime}, X_i^{\text{pub}} = x_i^{\text{pub}})} = \frac{\prod_j \mathrm{R}_i(\tilde{V}_i^j | X_i^{\text{priv}} = x_i^{\text{priv}}, X_i^{\text{pub}} = x_i^{\text{pub}})}{\prod_j \mathrm{R}_i(\tilde{V}_i^j | X_i^{\text{priv}} = x_i^{\text{priv}\prime}, X_i^{\text{pub}} = x_i^{\text{pub}})}.$$

By definition, given different x, there are only two $\mathbf{1}_{A\times B}(x)$ are different for all $A \times B$. Without loss of generality, assume they differ on the first two elements. Then we have

$$\sup_{x_i^{\text{priv}} x_i^{\text{priv}\prime}, x_i^{\text{pub}}} \frac{\mathrm{R}_i(\tilde{V}_i | X_i^{\text{priv}} = x_i^{\text{priv}}, X_i^{\text{pub}} = x_i^{\text{pub}})}{\mathrm{R}_i(\tilde{V}_i | X_i^{\text{priv}} = x_i^{\text{priv}\prime}, X_i^{\text{pub}} = x_i^{\text{pub}})}$$
$$= \sup_{x_i^{\text{priv}} x_i^{\text{priv}\prime}, x_i^{\text{pub}}} \frac{\prod_{j=1,2} \mathrm{R}_i(\tilde{V}_i^j | X_i^{\text{priv}} = x_i^{\text{priv}}, X_i^{\text{pub}} = x_i^{\text{pub}})}{\prod_{j=1,2} \mathrm{R}_i(\tilde{V}_i^j | X_i^{\text{priv}} = x_i^{\text{priv}\prime}, X_i^{\text{pub}} = x_i^{\text{pub}})} \leq e^{\varepsilon/4 + \varepsilon/4} = e^{\varepsilon/2}. \tag{19}$$

(ii) For the second part, there holds

$$\sup_{y,y'} \frac{\mathrm{R}_i(\tilde{Y}_i | Y_i = y)}{\mathrm{R}_i(\tilde{Y}_i | Y_i = y')} \leq \sup_{y,y'} \frac{d\mathrm{R}_i(\tilde{Y}_i | Y_i = y)}{d\mathrm{R}_i(\tilde{Y}_i | Y_i = y')} = \sup_{y,y'} \frac{\exp\left(-\frac{\varepsilon}{4M} |\tilde{Y}_i - y|\right)}{\exp\left(-\frac{\varepsilon}{4M} |\tilde{Y}_i - y'|\right)}$$
$$\leq \sup_{y,y'} \exp\left(\frac{\varepsilon}{4M} |y - y'|\right) \leq e^{\varepsilon/2}. \tag{20}$$

Bringing (19) and (20) into (18) yields the desired conclusion. □

Proof of Proposition A.1. We only need to modify (i) of the previous proof, which is now characterized by the generalized randomized response mechanism. For each grid $A \times B$ with index j, we consider the three cases in (8). The first case is $A \times B \notin \mathcal{V}_i$. The definition of $\mathcal{V}_i$ yields that, for any $X_i' = (x_i^{\text{priv}\prime}, x_i^{\text{pub}})$, its potential grids $\mathcal{V}_i'$ is exactly the same as $\mathcal{V}_i$. Thus, we have $\mathrm{R}_i(V_i = k | X_i^{\text{priv}} = x_i^{\text{priv}}, X_i^{\text{pub}} = x_i^{\text{pub}}) = \mathrm{R}_i(V_i = k | X_i^{\text{priv}} = x_i^{\text{priv}\prime}, X_i^{\text{pub}} = x_i^{\text{pub}}) = 0$. And thus the likelihood ratio is bounded by $e^{\varepsilon/2}$ as we define $0/0 = 1$. The second case is $A \times B \in \mathcal{V}_i$ and $X_i \in A \times B$. In this case, any $X_i' = (x_i^{\text{priv}\prime}, x_i^{\text{pub}})$ will have $A \times B \in \mathcal{V}_i'$. As a result,

$$\frac{\mathrm{R}_i(V_i = k | X_i^{\text{priv}} = x_i^{\text{priv}}, X_i^{\text{pub}} = x_i^{\text{pub}})}{\mathrm{R}_i(V_i = k | X_i^{\text{priv}} = x_i^{\text{priv}\prime}, X_i^{\text{pub}} = x_i^{\text{pub}})} \leq \frac{\frac{e^{\varepsilon/2}}{e^{\varepsilon/2} + |\mathcal{V}_i| - 1}}{\frac{1}{e^{\varepsilon/2} + |\mathcal{V}_i| - 1}} = e^{\varepsilon/2}.$$

The third case is the symmetric situation of the second case. Three cases together, we have

$$\frac{\mathrm{R}_i(V_i | X_i^{\text{priv}} = x_i^{\text{priv}}, X_i^{\text{pub}} = x_i^{\text{pub}})}{\mathrm{R}_i(V_i | X_i^{\text{priv}} = x_i^{\text{priv}\prime}, X_i^{\text{pub}} = x_i^{\text{pub}})} \leq e^{\varepsilon/2}.$$

□

B.2 Proof of Theorem 4.2

Proof of Theorem 4.2. The lower bound contains two terms. The second term $n^{-\frac{2\alpha}{2\alpha+d}}$ is the classical mini-max lower bound for non-private learners under Assumption 4.1. See (Tsybakov, 2009) for a comprehensive discussion. In the following, we prove the conclusion for the first term. The overall proof strategy mimics that of (Györfi and Kroll, 2022), while the utilization of information inequality is different. We first construct a finite set of hypotheses distribution. We then combine the information inequality under privacy constraint (Duchi et al., 2018) and Assouad's Lemma (Tsybakov, 2009).

Let $K_0 : \mathbb{R} \to [0,\infty)$ be a α-Holder continuous function with constant 1 such that $|K_0(x_1) - K_0(x_2)| \leq \|x_1 - x_2\|_2^{\alpha}$. Also, let $\operatorname{supp}(K_0) \subseteq [0,1]$. For $x = (x_1, \ldots, x_d) \in [0,1]^d$, define the function $K : [0,1]^d \to \mathbb{R}$ via $K(x) = \min_{i=1,\ldots,d} K_0(x_i)$. W.o.l.g., we let $\int_{[0,1]^d} K(x')dx' = 1$ and $\|K\|_\infty = 2$. We define the index set $\sigma = (\sigma^1, \cdots, \sigma^d) \in \{0, \cdots, k-1\}^d$. We write $\sigma = (\sigma_1, \sigma_2)$, where $\sigma_1 \in \{0, \cdots, k-1\}^{s^*}$ is the first s^* entries representing indexes on private dimensions, and $\sigma_2 \in \{0, \cdots, k-1\}^{d-s^*}$ representing indexes on public dimensions. Let $\theta^1 \in \{0,1\}^{k^{s^*}}$ be some one-hot encoding of σ_1 and $\theta^2 \in \{0,1\}^{k^{d-s^*}}$ for σ_2 accordingly. Denote $\theta = (\theta^1, \theta^2)$. In this case, we have a bijection between a θ and a σ. Define the function

$$K^\theta(x) = \frac{1}{k^\alpha} K(kx^1 - \sigma^1, \cdots, kx^d - \sigma^d)$$

for integer k and $x \in \times_{i=1}^d \left[\sigma^i/k, \left(\sigma^i + 1\right)/k\right]$. We consider the following class of distributions P^θ where $\mathrm{P}_X^\theta : [0,1]^d \to [0,\infty)$ is the uniform distribution, and $Y|X = x$ is the uniform distribution on $[K^\theta(x) - 1/2, K^\theta(x) + 1/2]$.

We first verify that the constructed P^θs satisfy Assumption 4.1. The marginal distribution is uniform and thus has bounded density. The conditional expectation K^θ has

$$|K^\theta(x_1) - K^\theta(x_2)| = \frac{1}{k^\alpha}|K(x_1) - K(x_2)| \leq \frac{1}{k^\alpha}\|kx_1 - kx_2\|_2^\alpha = \|x_1 - x_2\|_2^\alpha$$

for all θ. Thus, all P^θ is readily checked to satisfy Assumption 4.1. Let us now assume that the raw data have been anonymized by means of an arbitrary privacy mechanism R that is sequentially-interactive ε-semi-feature LDP. R maps each of (X_i, Y_i) to Z_i based on $Z_1, \cdots, Z_{i-1}$. Let $\mathrm{P}^{\theta n}$ and $\mathrm{RP}^{\theta n}$ be the joint distribution of $((X_1, Y_1), \cdots, (X_n, Y_n))$ and $(Z_1, \cdots, Z_n)$, respectively. Let $\widehat{f}$ be the estimator drawn by using $(Z_1, \cdots, Z_n)$. We are interested in the following quantity

$$\mathbb{E}_{\mathrm{RP}^{\theta n}} \mathcal{R}_{L,\mathrm{P}^\theta}(\widehat{f}) - \mathcal{R}^*_{L,\mathrm{P}^\theta} = \mathbb{E}_{\mathrm{RP}^{\theta n}} \left[\int_{[0,1]^d} \left(\widehat{f}(x') - K^\theta(x')\right)^2 d\mathrm{P}_{\mathrm{X}}^\theta(\mathrm{x}')\right].$$

For each fixed $\widehat{f}$, let

$$\widehat{\theta} = \arg\min_{\theta \in \{0,1\}^{k^d}} \left(\mathcal{R}_{L,\mathrm{P}^\theta}(\widehat{f}) - \mathcal{R}^*_{L,\mathrm{P}^\theta}\right)$$

Thus, by a triangular decomposition, the risk has

$$\begin{aligned}
\mathcal{R}_{L,\mathrm{P}^\theta}(K^{\widehat{\theta}}) - \mathcal{R}^*_{L,\mathrm{P}^\theta} \leq & \mathcal{R}_{L,\mathrm{P}^\theta}(\widehat{f}) - \mathcal{R}^*_{L,\mathrm{P}^\theta} + \mathcal{R}_{L,\mathrm{P}^{\widehat{\theta}}}(\widehat{f}) - \mathcal{R}^*_{L,\mathrm{P}^{\widehat{\theta}}} \\
\leq & 2\left(\mathcal{R}_{L,\mathrm{P}^\theta}(\widehat{f}) - \mathcal{R}^*_{L,\mathrm{P}^\theta}\right).
\end{aligned}$$

By a reordering of the decomposition of the domain $[0,1]^d = \cup_\sigma \times_{i=1}^d \left[\sigma^i/k, \left(\sigma^i + 1\right)/k\right] := \cup_\theta B^\theta$, we can reduce the expected risk as

$$\begin{aligned}
\mathbb{E}_{\mathrm{RP}^{\theta n}} \mathcal{R}_{L,\mathrm{P}^\theta}(\widehat{f}) - \mathcal{R}^*_{L,\mathrm{P}^\theta} \geq & \frac{1}{2}\left(\mathbb{E}_{\mathrm{RP}^{\theta n}} \mathcal{R}_{L,\mathrm{P}^\theta}(K^{\widehat{\theta}}) - \mathcal{R}^*_{L,\mathrm{P}^\theta}\right) \\
= & \frac{1}{2}\mathbb{E}_{\mathrm{RP}^{\theta n}} \left[\int_{[0,1]^d} \left(K^{\widehat{\theta}}(x') - K^\theta(x')\right)^2 dx'\right] \\
= & \frac{1}{k^{2\alpha+d}} \mathbb{E}_{\mathrm{RP}^{\theta n}} \left[\mathbf{1}_{\widehat{\theta}=\theta}\right] \cdot \int_{[0,1]^d} K(x')dx' = \frac{1}{k^{2\alpha+d}} \mathbb{E}_{\mathrm{RP}^{\theta n}} \left[\mathbf{1}_{\widehat{\theta}=\theta}\right].
\end{aligned} \tag{21}$$

For any pair of distribution P_1 and P_2, let KL be the Kullback-Leibler divergence of P_1 and P_2, i.e. $KL(\mathrm{P}_1\|\mathrm{P}_2) = \int \log\left(d\mathrm{P}_1(x)/d\mathrm{P}_2\right) d\mathrm{P}_1(x)$. Also, let the total variation distance of P_1 and P_2 be $V(\mathrm{P}_1,\mathrm{P}_2) = \frac{1}{2}\int |d\mathrm{P}_1(x) - d\mathrm{P}_2(x)|$. To utilize the machinery in Duchi et al. (2018), we need to bound the quantity $KL(\mathrm{RP}^{\theta_1 n}\|\mathrm{RP}^{\theta_2 n})$ for arbitrary θ_1 and θ_2. Tensorization yields

$$KL(\mathrm{RP}^{\theta_1 n}\|\mathrm{RP}^{\theta_2 n}) = \sum_{i=1}^n KL(\mathrm{R}_i\mathrm{P}^{\theta_1}\|\mathrm{R}_i\mathrm{P}^{\theta_2}). \tag{22}$$

Let $p^{\theta\text{pub}}$ be the probability density of $(X^{s^*+1}, X^{s^*+2}, \cdots, X^d)$. Let $\mathrm{P}^{\theta\text{priv}}$ be the conditional probability measure of $(X^1, X^2, \cdots, X^{s^*}, Y)$ conditional on $(X^{s^*+1}, X^{s^*+2}, \cdots, X^d)$ and $p^{\theta\text{priv}}$ be its density function. And let r_i be the density of mechanism R_i condition on (X_i, Y_i). Since the estimated $\widehat{\theta}$ is drawn on both Z_i and X_i^{pub}, w.o.l.g. we write the output of mechanism R as (Z_i, X_i^{pub}). Then the output density function is $m_i^\theta(z_i, x_i^{\text{pub}}) = \int_{\mathcal{X}^{\text{priv}}\times\mathcal{Y}} r_i(z_i|x_i^{\text{priv}}, x_i^{\text{pub}}, y_i) p^{\theta\text{priv}}(x_i^{\text{priv}}, y_i|x_i^{\text{pub}}) p^{\theta\text{pub}}(x_i^{\text{pub}}) dx_i^{\text{priv}} dy_i$. Then we have

$$\begin{aligned}
KL(\mathrm{R}_i\mathrm{P}^{\theta_1}\|\mathrm{R}_i\mathrm{P}^{\theta_2}) &= \int_{\mathcal{X}^{\text{pub}}}\int_{\mathcal{S}} \log\left(\frac{m_i^{\theta_1}(z_i, x_i^{\text{pub}})}{m_i^{\theta_2}(z_i, x_i^{\text{pub}})}\right) m_i^{\theta_1}(z_i, x_i^{\text{pub}}) dz_i dx_i^{\text{pub}} \\
&= \int_{\mathcal{X}^{\text{pub}}} \underbrace{\left[\int_{\mathcal{S}} \log\left(\frac{m_i^{\theta_1}(z_i|x_i^{\text{pub}}) p^{\theta_1\text{pub}}(x_i^{\text{pub}})}{m_i^{\theta_2}(z_i|x_i^{\text{pub}}) p^{\theta_2\text{pub}}(x_i^{\text{pub}})}\right) m_i^{\theta_1}(z_i|x_i^{\text{pub}}) dz_i\right]}_{(**)} p^{\theta_1\text{pub}}(x_i^{\text{pub}}) dx_i^{\text{pub}}.
\end{aligned} \tag{23}$$

Our construction yields that

$$(**) = \int_{\mathcal{S}} \log\left(\frac{m_i^{\theta_1}(z_i|x_i^{\text{pub}})}{m_i^{\theta_2}(z_i|x_i^{\text{pub}})}\right) m_i^{\theta_1}(z_i|x_i^{\text{pub}}) dz_i = KL\left((\mathrm{R}_i\mathrm{P}^{\theta_1})_{Z_i|X_i^{\text{pub}}}\|(\mathrm{R}_i\mathrm{P}^{\theta_2})_{Z_i|X_i^{\text{pub}}}\right)$$

By Definition 3.1, Z_i given X_i^{pub} is actually locally differentially private (as rigorously defined in Duchi et al. (2018); Györfi and Kroll (2022)) w.r.t. (X_i^{priv}, Y_i). Thus, applying Theorem 1 in Duchi et al. (2018), we have

$$(**) \leq 4\left(e^\varepsilon - 1\right)^2 V^2\left(\mathrm{P}^{\theta_1\text{priv}}, \mathrm{P}^{\theta_2\text{priv}}\right).$$

Bringing in our construction where the density on $[0,1]^d \times [K^\theta(x) - 1/2, K^\theta(x) + 1/2]$ is 1 and otherwise zero, we get

$$\begin{aligned}
(**) \leq& 4\left(e^\varepsilon - 1\right)^2 \left(\int_{\mathcal{X}^{\text{priv}}} |K^{\theta_1}(x_i^{\text{priv}}, x_i^{\text{pub}}) - K^{\theta_2}(x_i^{\text{priv}}, x_i^{\text{pub}})| dx_i^{\text{priv}}\right)^2 \\
=& 4\left(e^\varepsilon - 1\right)^2 \cdot k^{-2\alpha - 2s^*} \cdot \left(2 \cdot \int_{[0,1]^{s^*}} |K(x, kx_i^{\text{pub}} - \sigma_2)| dx\right)^2.
\end{aligned}$$

Bringing back this into (23) and apply Cauchy's inequality, there holds

$$\begin{aligned}
KL(\mathrm{R}_i\mathrm{P}^{\theta_1}\|\mathrm{R}_i\mathrm{P}^{\theta_2}) \leq& 16\left(e^\varepsilon - 1\right)^2 \cdot k^{-2\alpha - 2s^*} \int_{\mathcal{X}^{\text{pub}}}\int_{[0,1]^{s^*}} |K(x, kx_i^{\text{pub}} - \sigma_2)|^2 dx p^{\theta_1\text{pub}}(x_i^{\text{pub}}) dx_i^{\text{pub}} \\
\leq& 16\left(e^\varepsilon - 1\right)^2 \cdot k^{-2\alpha - 2s^* - (d - s^*)} \int_{[0,1]^d} |K(x)| dx \leq 32\left(e^\varepsilon - 1\right)^2 \cdot k^{-2\alpha - s^* - d}.
\end{aligned}$$

Bringing back this result into (22) leads to

$$KL(\mathrm{RP}^{\theta_1 n}\|\mathrm{RP}^{\theta_2 n}) \leq 32\left(e^\varepsilon - 1\right)^2 \cdot k^{-2\alpha - s^* - d} n.$$

Consequently, taking a $k = \left(32n(e^\varepsilon - 1)^2\right)^{\frac{1}{2\alpha + s^* + d}}$ yields a $KL(\mathrm{RP}^{\theta_1 n}\|\mathrm{RP}^{\theta_2 n}) \leq 1$. Applying Theorem 2.12, Statement (iv) of Tsybakov (2009) gives us $\mathbb{E}_{\mathrm{RP}^{\theta n}}\left[\mathbf{1}_{\widehat{\theta} = \theta}\right] \geq k^d/8$, which is combined with (21) and leads to

$$\mathbb{E}_{\mathrm{RP}^{\theta n}} \mathcal{R}_{L,\mathrm{P}^\theta}(\widehat{f}) - \mathcal{R}^*_{L,\mathrm{P}^\theta} \gtrsim \left(n(e^\varepsilon - 1)^2\right)^{\frac{2\alpha}{2\alpha + s^* + d}}.$$

$\square$

$$\sum_{i=1}^{n}\mathcal{E}^{\frac{2\alpha+d+\sum_{j=1}^{d}W_i^j}{2\alpha}} \leq \frac{1}{(e^{\varepsilon}-1)^2}$$

$$\sum_{i=1}^{n}\mathcal{E}^{-\frac{2\alpha+d+\sum_{j=1}^{d}W_i^j}{2\alpha}} = n^2\varepsilon^2$$

B.3 Proof of Theorem 4.3

We first define the intermediate quantity

$$\overline{f}(x) = \sum_{A_j,B_k}\mathbf{1}_{A_j\times B_k}(x)\frac{\int_{A_j\times B_k} f^*(x)d\mathrm{P}_X(x)}{\int_{A_j\times B_k} d\mathrm{P}_X(x)}. \tag{24}$$

We rely on the following decomposition.

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(\tilde{f}) - \mathcal{R}_{L,\mathrm{P}}(f^*) = &\underbrace{\mathcal{R}_{L,\mathrm{P}}(\tilde{f}) - \mathcal{R}_{L,\mathrm{P}}(f)}_{\textbf{Privatized Error}} + \underbrace{\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(\overline{f})}_{\textbf{Sample Error}} \\ &+ \underbrace{\mathcal{R}_{L,\mathrm{P}}(\overline{f}) - \mathcal{R}_{L,\mathrm{P}}(f^*)}_{\textbf{Approximation Error}}.\end{aligned} \tag{25}$$

where the `HistOfTree` estimator $\tilde{f}$, the non-private partition-based estimator f, and the population partition-based estimator $\overline{f}$ are defined in (9), (3), and (24), respectively. Loosely speaking, the first error term quantifies the depravation brought by adding privacy noises to the estimator, which we call the privatized error. The second term corresponds to the expected estimation error brought by the randomness of the data, which we call the sample error. The last term is called approximation error, which arises due to the limited approximation capacity of piecewise constant functions. The following three lemmas provide bounds for each of the three errors.

Lemma B.1 (**Bounding privatised error**). *Let the `HistOfTree` estimator $\tilde{f}$ and the non-private partition-based estimator f be defined in* (9) *and* (3)*, respectively. For $\pi = \pi^{priv}\otimes\pi^{pub}$, let π^{priv} be a histogram partition with t bins and π^{pub} be generated by Algorithm 1 with depth p. Let Assumption 4.1 hold. Suppose $t^s\cdot 2^p\cdot\log n \lesssim n$. Then, there holds*

$$\mathcal{R}_{L,\mathrm{P}}(\tilde{f}) - \mathcal{R}_{L,\mathrm{P}}(f) \lesssim \frac{t^{2s}\cdot 2^p\cdot \log n\cdot \frac{1}{n}\sum_{i=1}^{n}2^{\sum_{\ell=s+1}^{d}W_i^{\ell}\cdot p/(d-s)}}{n\varepsilon^2}$$

with probability $\mathrm{P}^n\otimes\mathrm{R}^n$ at least $1-4/n^2$.

Lemma B.2 (**Bounding sample error**). *Let the non-private partition-based estimator f and the population partition-based estimator $\overline{f}$ be defined in* (3) *and* (24)*, respectively. For $\pi = \pi^{priv}\otimes\pi^{pub}$, let π^{priv} be a histogram partition with t bins and π^{pub} be generated by Algorithm 1 with depth p. Let Assumption 4.1 hold. Suppose $t^s\cdot 2^p\cdot\log n\lesssim n$. Then, there holds*

$$\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(\overline{f}) \lesssim \frac{t^s\cdot 2^p\cdot\log n}{n}$$

with probability $\mathrm{P}^n\otimes\mathrm{R}^n$ at least $1-4/n^2$.

Lemma B.3 (**Bounding approximation error**). *Let the population partition-based estimator $\overline{f}$ be defined in* (24)*. For $\pi = \pi^{priv}\otimes\pi^{pub}$, let π^{priv} be a histogram partition with t bins and π^{pub} be generated by Algorithm 1 with depth p. Let Assumption 4.1 hold. Then there holds*

$$\mathcal{R}_{L,\mathrm{P}}(\overline{f}) - \mathcal{R}_{L,\mathrm{P}}(f^*) \lesssim\ t^{-2\alpha} + 2^{-2\alpha p/(d-s)}.$$

Proof of Theorem 4.3. The privacy of $\tilde{f}$ follows from Proposition 3.2. We then focus on the excess risk upper bound. Bringing Proposition B.1, B.2, and B.3 into the decomposition (25), we have

$$\begin{aligned}&\mathcal{R}_{L,\mathrm{P}}(\tilde{f})-\mathcal{R}^*_{L,\mathrm{P}}\\ \lesssim&\frac{t^{2s}\cdot 2^p\cdot\log n\cdot\frac{1}{n}\sum_{i=1}^n 2^{\sum_{\ell=s+1}^d W_i^\ell\cdot p/(d-s)}}{n\varepsilon^2}+\frac{t^s\cdot 2^p\cdot\log n}{n}+t^{-2\alpha}+2^{-2\alpha p/(d-s)}\end{aligned}\tag{26}$$

holds with probability $\mathrm{P}^n\otimes\mathrm{R}^n$ at least $1-8/n^2$. For any $\varepsilon^2\lesssim t^s\cdot\frac{1}{n}\sum_{i=1}^n 2^{\sum_{\ell=s+1}^d W_i^\ell\cdot p/(d-s)}$, the first term dominants the second term. We will specify this range of ε later. Since we take $t\asymp 2^{p/(d-s)}$, we have $t^{-2\alpha}\asymp 2^{-2\alpha p/(d-s)}$ and $t^{2s}\cdot 2^p\asymp 2^{p(d+s)/(d-s)}$. Therefore, (26) becomes

$$\mathcal{R}_{L,\mathrm{P}}(\tilde{f})-\mathcal{R}^*_{L,\mathrm{P}}\lesssim\frac{2^{p(d+s)/(d-s)}\cdot\log n\cdot\frac{1}{n}\sum_{i=1}^n 2^{\sum_{\ell=s+1}^d W_i^\ell\cdot p/(d-s)}}{n\varepsilon^2}+2^{-2\alpha p/(d-s)},$$

which recover the equation (11). Recalling the way we select p^*, the upper bound becomes

$$\mathcal{R}_{L,\mathrm{P}}(\tilde{f})-\mathcal{R}^*_{L,\mathrm{P}}\lesssim\frac{2^{p^*(d+s)/(d-s)}\cdot\log n\cdot 2^{p^*\lambda^*}}{n\varepsilon^2}+2^{-2\alpha p^*/(d-s)}.$$

The minimizer is taken at the p^* that match the two terms, which is $2^{p^*}\asymp\left(n\varepsilon^2/\log n\right)^{(d-s)/(2\alpha+d+s+\lambda^*(d-s))}$. In this case, the upper bound is

$$\mathcal{R}_{L,\mathrm{P}}(\tilde{f})-\mathcal{R}^*_{L,\mathrm{P}}\lesssim\left(\frac{\log n}{n\varepsilon^2}\right)^{\frac{2\alpha}{2\alpha+d+s+\lambda^*(d-s)}}.$$

The definition of λ^* yields that it is less than 1. Thus,

$$t^s\asymp 2^{\frac{sp^*}{d-s}}\asymp\left(\frac{\log n}{n\varepsilon^2}\right)^{-\frac{s}{2\alpha+d+s+\lambda^*(d-s)}}\gtrsim\left(\frac{\log n}{n\varepsilon^2}\right)^{-\frac{s}{2\alpha+2d}}.$$

As a result, it suffices to require $\left(\frac{\log n}{n\varepsilon^2}\right)^{-\frac{s}{2\alpha+2d}}\gtrsim\varepsilon^2$, which yields $\varepsilon\lesssim(n/\log n)^{\frac{s}{\alpha+d-s}}$. □

B.4 Proofs of Results in Section B.3

To prove lemmas in Section B.3, we first present several technical results.

Lemma B.4. *Suppose $\zeta_i, i=1,\cdots,n$ are independent random variables such that $a_i\leq\zeta_i\leq b_i$. Then there holds*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n\zeta_i-\mathbb{E}\frac{1}{n}\sum_{i=1}^n\zeta_i\right|\geq t\right]\leq 2e^{-\frac{2n^2t^2}{\sum_{i=1}^n(b_i-a_i)^2}}$$

for any $t>0$.

Proof of Lemma B.4. The conclusion follows from Example 2.4 and Proposition 2.5 in (Wainwright, 2019). □

Lemma B.5. *Suppose $\xi_i, i=1,\cdots,n$ are independent sub-exponential random variables with parameters (ν_i,β_i) (Wainwright, 2019, Definition 2.9). Then there holds*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n\xi_i-\mathbb{E}\frac{1}{n}\sum_{i=1}^n\xi_i\right|\geq t\right]\leq 2e^{-\frac{nt^2}{2\nu_*^2}}.$$

for any $0<t<\nu_*^2/(n\beta_*)$, where $\nu_*=\sqrt{\sum_{i=1}^n\nu_i^2}$ and $\beta_*=\max_{i=1,\cdots,n}\beta_i$. Moreover, a standard Laplace random variable is sub-exponential with parameters $(\sqrt{2},1)$, and $a\xi_i$ has $(a\nu_i,a\beta_i)$ for any positive a.

Proof of Lemma B.5. The conclusion follows from (2.18) in (Wainwright, 2019). □

Lemma B.6. *Let $\tilde{\mathcal{A}}$ be the collection of all cells $\times_{i=1}^d [a_i, b_i]$ in $\mathbb{R}^d$. The VC index of $\tilde{\mathcal{A}}$ equals $2d+1$. Moreover, for all $0 < \varepsilon < 1$, there exists a universal constant C such that*

$$\mathcal{N}(\mathbf{1}_{\tilde{\mathcal{A}}}, \|\cdot\|_{L_1(Q)}, \varepsilon) \le C(2d+1)(4e)^{2d+1}(1/\varepsilon)^{2d}.$$

Proof of Lemma B.6. The first result of the VC index follows from Example 2.6.1 in (van der Vaart and Wellner, 1996). The second result of covering number follows directly from Theorem 9.2 in (Kosorok, 2008). □

Lemma B.7. *For $\pi = \pi^{priv} \otimes \pi^{pub}$, let π^{priv} be a histogram partition with t bins and π^{pub} be generated by Algorithm 1 with depth p. Suppose Assumption 4.1 holds. Then for any $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ drawn i.i.d. from P and any $A \in \pi$, there holds*

$$\left|\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \in A\} - \int_A d\mathrm{P}_X(x')\right| \le \sqrt{\frac{\overline{c}\cdot t^{-s}\cdot 2^{1-p}(4d+5)\log n}{n}} + \frac{2(4d+5)\log n}{3n} + \frac{4}{n}$$

with probability P^n at least $1 - 1/n^2$. Here we use A instead of $A \times B$ for notation simplicity.

Proof of Lemma B.7. In the subsequent proof, we define the empirical measure $\mathrm{D} := \frac{1}{n}\sum_{i=1}^n \delta_{(X_i, Y_i)}$ given samples $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$, where δ is the Dirac function. Let $\tilde{\mathcal{A}}$ be the collection of all cells $\times_{i=1}^d [a_i, b_i]$ in $\mathbb{R}^d$. Applying Lemma B.6 with $Q := (\mathrm{D}_X + \mathrm{P}_X)/2$, there exists an ε-net $\{\tilde{A}_k\}_{k=1}^K \subset \tilde{\mathcal{A}}$ with

$$K \le C(2d+1)(4e)^{2d+1}(1/\varepsilon)^{2d} \tag{27}$$

such that for any $A \in \pi$, there exist some $k \in \{1, \ldots, K\}$ such that

$$\|\mathbf{1}\{x \in A\} - \mathbf{1}\{x \in \tilde{A}_k\}\|_{L_1((\mathrm{D}_X + \mathrm{P}_X)/2)} \le \varepsilon,$$

Since

$$\begin{aligned}
&\|\mathbf{1}\{x \in A\} - \mathbf{1}\{x \in \tilde{A}_k\}\|_{L_1((\mathrm{D}_X + \mathrm{P}_X)/2)} \\
=&1/2 \cdot \|\mathbf{1}\{x \in A\} - \mathbf{1}\{x \in \tilde{A}_k\}\|_{L_1(\mathrm{D}_X)} + 1/2 \cdot \|\mathbf{1}\{x \in A\} - \mathbf{1}\{x \in \tilde{A}_k\}\|_{L_1(\mathrm{P}_X)},
\end{aligned}$$

we get

$$\|\mathbf{1}\{x \in A\} - \mathbf{1}\{x \in \tilde{A}_k\}\|_{L_1(\mathrm{D}_X)} \le 2\varepsilon, \quad \|\mathbf{1}\{x \in A\} - \mathbf{1}\{x \in \tilde{A}_k\}\|_{L_1(\mathrm{P}_X)} \le 2\varepsilon. \tag{28}$$

Consequently, by the definition of the covering number and the triangle inequality, for any $A \in \pi$, there holds

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{x \in A\}(X_i) - \int_{\tilde{A}_p^j} d\mathrm{P}_X(x')\right| \\
\le&\left|\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{x \in \tilde{A}_k\}(X_i) - \int_{\tilde{A}_k} d\mathrm{P}_X(x')\right| + \|\mathbf{1}\{x \in A\} - \mathbf{1}\{x \in \tilde{A}_k\}\|_{L_1(\mathrm{D}_X)} \\
&+ \|\mathbf{1}\{x \in A\} - \mathbf{1}\{x \in \tilde{A}_k\}\|_{L_1(\mathrm{P}_X)} \le \left|\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{x \in \tilde{A}_k\}(X_i) - \int_{\tilde{A}_k} d\mathrm{P}_X(x')\right| + 4\varepsilon.
\end{aligned}$$

Therefore, we get

$$\sup_{j \in \mathcal{I}}\left|\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{x \in A\}(X_i) - \int_{\tilde{A}_p^j} d\mathrm{P}_X(x')\right| \le \sup_{1 \le k \le K}\left|\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{x \in \tilde{A}_k\}(X_i) - \int_{\tilde{A}_k} d\mathrm{P}_X(x')\right| + 4\varepsilon. \tag{29}$$

For any fixed $1 \le k \le K$, let the random variable ξ_i be defined by $\xi_i := \mathbf{1}\{X_i \in \tilde{A}_k\} - \int_{\tilde{A}_k} d\mathrm{P}_X(x')$. Then we have $\mathbb{E}_{\mathrm{P}_X}\xi_i = 0$, $\|\xi\|_\infty \le 1$, and $\mathbb{E}_{\mathrm{P}_X}\xi_i^2 \le \int_{\tilde{A}_k} d\mathrm{P}_X(x')$. According to Assumption 4.1, there holds $\mathbb{E}_{\mathrm{P}_X}\xi_i^2 \le \overline{c}\cdot t^{-s}\cdot 2^{-p}$. Applying Bernstein's inequality, we obtain

$$\left|\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \in \tilde{A}_k\} - \int_{\tilde{A}_k} d\mathrm{P}_X(x')\right| \le \sqrt{\frac{\overline{c}\cdot t^{-s}\cdot 2^{1-p}\cdot\tau}{n}} + \frac{2\tau\log n}{3n}$$

with probability P^n at least $1-2e^{-\tau}$. Then the union bound together with the covering number estimate (27) implies that

$$\sup_{1\leq k\leq K}\left|\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i\in\tilde{A}_k\}-\int_{\tilde{A}_k} d\mathrm{P}_X(x')\right|\leq\sqrt{\frac{\overline{c}\cdot t^{-s}\cdot 2^{1-p}(\tau+\log(2K))}{n}}+\frac{2(\tau+\log(2K))\log n}{3n}$$

with probability P^n at least $1-e^{-\tau}$. Let $\tau=2\log n$ and $\varepsilon=1/n$. Then for any $n>N_1:=(2C)\wedge(2d+1)\wedge(4e)$, we have $\tau+\log(2K)=2\log n+\log(2C)+\log(2d+1)+(2d+1)\log(4e)+2d\log n\leq(4d+5)\log n$. Therefore, we have

$$\sup_{1\leq k\leq K}\left|\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i\in\tilde{A}_k\}-\int_{\tilde{A}_k} d\mathrm{P}_X(x')\right|\leq\sqrt{\frac{\overline{c}\cdot t^{-s}\cdot 2^{1-p}(4d+5)\log n}{n}}+\frac{2(4d+5)\log n}{3n}\tag{30}$$

with probability P^n at least $1-1/n^2$. This together with (29) yields that

$$\sup_{j\in\mathcal{I}}\left|\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{x\in A\}-\int_{\tilde{A}_p^j} d\mathrm{P}_X(x')\right|\leq\sqrt{\frac{\overline{c}\cdot t^{-s}\cdot 2^{1-p}(4d+5)\log n}{n}}+\frac{2(4d+5)\log n}{3n}+\frac{4}{n}.$$

□

Lemma B.8. *For $\pi=\pi^{priv}\otimes\pi^{pub}$, let π^{priv} be a histogram partition with t bins and π^{pub} be generated by Algorithm 1 with depth p. Suppose Assumption 4.1 holds. Then for any $\{(X_1,Y_1),\cdots,(X_n,Y_n)\}$ drawn i.i.d. from P and any $A\in\pi$, there holds*

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^n Y_i\mathbf{1}\{X_i\in A(x)\}-\int_{A(x)} f^*(x')d\mathrm{P}_X(x')\right|\\ \leq& M\sqrt{\frac{\overline{c}\cdot t^{-s}\cdot 2^{1-p}(4d+5)\log n}{n}}+\frac{2M(4d+5)\log n}{3n}+\frac{4M}{n}\end{aligned}$$

with probability P^n at least $1-1/n^2$. Here we use A instead of $A\times B$ for notation simplicity.

Proof of Lemma B.8. Let $\tilde{\mathcal{A}}$ be the collection of all cells $\times_{i=1}^d[a_i,b_i]$ in $\mathbb{R}^d$. Then there exists an ε-net $\{\tilde{A}_k\}_{k=1}^K\subset\tilde{\mathcal{A}}$ with K bounded by (27) such that for any $j\in\mathcal{I}$, (28) holds for some $k\in\{1,\ldots,K\}$. Consequently, by the definition of the covering number and the triangle inequality, for any $A\in\pi$, there holds

$$\begin{aligned}&\left|\sum_{i=1}^n \mathbf{1}\{X_i\in A\}Y_i-\int_A f^*(x')d\mathrm{P}_X(x')\right|\\
&\leq\left|\sum_{i=1}^n \mathbf{1}\{X_i\in\tilde{A}_k\}Y_i-\int_{\tilde{A}_k} f^*(x')d\mathrm{P}_X(x')\right|\\
&+\int_{\mathbb{R}^d}\left|\mathbf{1}\{x'\in A\}-\mathbf{1}\{x'\in\tilde{A}_k\}\right|\left|f^*(x')\right|d\mathrm{P}_X(x')+\sum_{i=1}^n\left|\mathbf{1}\{X_i\in\tilde{A}_k\}-\mathbf{1}\{X_i\in A\}\right|\left|Y_i\right|\\
&\leq\left|\sum_{i=1}^n \mathbf{1}\{X_i\in\tilde{A}_k\}Y_i-\int_{\tilde{A}_k} f^*(x')d\mathrm{P}_X(x')\right|\\
&+\max_{1\leq i\leq n}|Y_i|\cdot\|\mathbf{1}\{x\in A\}-\mathbf{1}\{x\in\tilde{A}_k\}\|_{L_1(\mathrm{D}_X)}+M\cdot\|\mathbf{1}\{x\in A\}-\mathbf{1}\{x\in\tilde{A}_k\}\|_{L_1(\mathrm{P}_X)}\\
&\leq\left|\sum_{i=1}^n \mathbf{1}_{\tilde{A}_k}(X_i)Y_i-\int_{\tilde{A}_k} f^*(x')d\mathrm{P}_X(x')\right|+4M\varepsilon.\end{aligned}\tag{31}$$

where the last inequality follow from the condition $\mathcal{Y}\subset[-M,M]$.

For any fixed $1\leq k\leq K$, let the random variable $\tilde{\xi}_i$ be defined by $\tilde{\xi}_i:=\mathbf{1}\{X_i\in\tilde{A}_k\}Y_i-\int_{\tilde{A}_k}f^*(x')\,d\mathrm{P}_X(x')$. Then we have $\mathbb{E}_{\mathrm{P}}\tilde{\xi}_i=0$, $\|\xi\|_\infty\leq M+1$, and $\mathbb{E}_{\mathrm{P}}\tilde{\xi}_i^2\leq M^2\int_{\tilde{A}_k}d\mathrm{P}(x')$. According to Assumption 4.1, there holds

$\mathbb{E}_{\mathrm{P}}\tilde{\xi}_i^2 \leq M^2 \cdot \overline{c} \cdot t^{-s} \cdot 2^{-p}$. Applying Bernstein's inequality, we obtain

$$\left|\sum_{i=1}^n \mathbf{1}\{X_i \in \tilde{A}_k\}Y_i - \int_{\tilde{A}_k} f^*(x')d\mathrm{P}_X(x')\right| \leq \sqrt{\frac{M^2 \cdot \overline{c} \cdot t^{-s} \cdot 2^{1-p} \cdot \tau}{n}} + \frac{2M\tau \log n}{3n}$$

with probability P^n at least $1 - 2e^{-\tau}$. Similar to the proof of Lemma B.7, one can show that for any $n \geq N_1$, there holds

$$\sup_{1\leq k\leq K}\left|\sum_{i=1}^n \mathbf{1}\{X_i \in \tilde{A}_k\}Y_i - \int_{\tilde{A}_k} f^*(x')d\mathrm{P}_X(x')\right| \leq M\sqrt{\frac{\overline{c} \cdot t^{-s} \cdot 2^{1-p} \cdot \tau}{n}} + \frac{2M\tau \log n}{3n}$$

with probability P^n at least $1 - 1/n^2$. This together with (31) yields that

$$\begin{aligned}&\left|\sum_{i=1}^n \mathbf{1}\{X_i \in A\}Y_i - \int_{A_p^j} f^*(x')d\mathrm{P}_X(x')\right| \\ \leq & M\sqrt{\frac{\overline{c} \cdot t^{-s} \cdot 2^{1-p}(4d+5)\log n}{n}} + \frac{2M(4d+5)\log n}{3n} + \frac{4M}{n}.\end{aligned} \tag{32}$$

□

Lemma B.9. *For $\pi = \pi^{priv} \otimes \pi^{pub}$, let π^{priv} be a histogram partition with t bins and π^{pub} be generated by Algorithm 1 with depth p. Then for any $A \times B := \times_{i=1}^d [a_i, b_i] \in \pi$, there holds*

$$2^{-2}\left(\sqrt{d-s} \cdot 2^{-p/(d-s)} + \sqrt{s}\, t^{-1}\right) \leq \mathrm{diam}(A \times B) \leq 2\left(\sqrt{d-s} \cdot 2^{-p/(d-s)} + \sqrt{s}\, t^{-1}\right).$$

Proof of Lemma B.9. For each $A \times B$, s edges of A have length t^{-1}. The rest $d-s$ edges of B are decided by the max-edge partition rule. When the depth of the tree p is a multiple of dimension $d-s$, each cell of the tree partition is a high-dimensional cube with a side length $2^{-p/(d-s)}$. On the other hand, when the depth of the tree p is not a multiple of dimension $d-s$, we consider the max-edge tree partition with depth $\lfloor p/(d-s)\rfloor$ and $\lceil p/(d-s)\rceil$, whose corresponding side length of the higher dimensional cube is $2^{-\lfloor p/(d-s)\rfloor}$ and $2^{-\lceil p/(d-s)\rceil}$. Note that in the splitting procedure of max-edge partition, the side length of each sub-rectangle decreases monotonically with the increase of p, so the side length of a random tree partition cell is between $2^{-\lceil p/(d-s)\rceil}$ and $2^{-\lfloor p/(d-s)\rfloor}$. This implies that

$$\sqrt{d-s} \cdot 2^{-\lceil p/(d-s)\rceil} \leq \mathrm{diam}(B) \leq \sqrt{d-s} \cdot 2^{-\lfloor p/(d-s)\rfloor}$$

Since $p/(d-s) - 1 \leq \lfloor p/(d-s)\rfloor \leq \lceil p/(d-s)\rceil \leq p/(d-s) + 1$, we immediately get $2^{-1}\sqrt{d-s} \cdot 2^{-p/(d-s)} \leq \mathrm{diam}(B) \leq 2\sqrt{d-s}\cdot 2^{-p/(d-s)}$. Together, we have $\mathrm{diam}(A\times B) \leq \mathrm{diam}(A)+\mathrm{diam}(B) \leq \sqrt{s}t^{-1}+2\sqrt{d-s}\cdot 2^{-p/(d-s)}$ and $\mathrm{diam}(A \times B) \geq (\mathrm{diam}(A) + \mathrm{diam}(B))/2 \geq (\sqrt{d-s} \cdot 2^{-p/(d-s)} + \sqrt{s}t^{-1})/4$. □

Lemma B.10. *For $\pi = \pi^{priv} \otimes \pi^{pub}$, let π^{priv} be a histogram partition with t bins and π^{pub} be generated by Algorithm 1 with depth p. Then for each i, there are at most $t^s \cdot 2^{p-\lfloor p/(d-s)\rfloor\cdot\left(d-s-\sum_{\ell=s+1}^d W_i^\ell\right)}$ elements of $V_i^1, \cdots, V_i^{t^s\cdot 2^p}$ are non-zero.*

Proof of Lemma B.10. The number of non-zero elements is the cardinality of potential grids $|\mathcal{V}_i|$. For $A \times B$, since the first s dimensions are all private, all t^s possibility of A potentially appears in $\mathcal{V}_i$. For B, consider the process of Algorithm 1 that gradually grows 2^p grids. At any step $k = 1, \cdots, p$, the algorithm split along the ℓ_k-th dimension. If $W_i^{\ell_k}$ is 1, i.e. the feature is private, the number of potential grids will double. Otherwise, it remains the same. Each feature is split at least $\lfloor p/(d-s)\rfloor$ times, i.e. there are at most $p - \lfloor p/(d-s)\rfloor \cdot \left(d-s-\sum_{\ell=s+1}^d W_i^\ell\right)$ splits which causes the number of potential grids to double. This quantity is upper bonded by $\sum_{\ell=s+1}^d p \cdot W_i^\ell/(d-s) + d - s$. Multiplying potential possibilities of A and B leads to $t^s \cdot 2^{\sum_{\ell=s+1}^d p\cdot W_i^\ell/(d-s)+d-s}$. □

Proof of Lemma B.1. We intend to bound

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(\tilde{f}) - \mathcal{R}_{L,\mathrm{P}}(f) &= \int_{\mathcal{X}} \left|\tilde{f}(x) - f(x)\right|^2 d\mathrm{P}_X(x) \\ &= \sum_j \left|\frac{\sum_i \tilde{Y}_i \cdot \tilde{V}_i^j}{\sum_i \tilde{V}_i^j} - \frac{\sum_i Y_i \cdot V_i^j}{\sum_i V_i^j}\right|^2 \int_{A\times B} d\mathrm{P}_X(x) \\ &= \sum_j \frac{1}{t^s 2^p} \left|\frac{\sum_i \tilde{Y}_i \cdot \tilde{V}_i^j}{\sum_i \tilde{V}_i^j} - \frac{\sum_i Y_i \cdot V_i^j}{\sum_i V_i^j}\right|^2 \end{aligned} \tag{33}$$

where j is the index of $A \times B$. Here, V_i^j is defined as $\mathbf{1}_{A\times B}(X_i)$. For each term, we have decomposition

$$\begin{aligned}\left|\frac{\sum_i \tilde{Y}_i \cdot \tilde{V}_i^j}{\sum_i \tilde{V}_i^j} - \frac{\sum_i Y_i \cdot V_i^j}{\sum_i V_i^j}\right|^2 \leq 3 \cdot \Bigg(&\underbrace{\left|\frac{\sum_{i=1}^n (\tilde{Y}_i - Y_i)\tilde{V}_i^j}{\sum_{i=1}^n \tilde{V}_i^j}\right|^2}_{(I)} \\ + \underbrace{\left|\frac{\sum_{i=1}^n Y_i\tilde{V}_i^j - \sum_{i=1}^n Y_i V_i^j}{\sum_{i=1}^n \tilde{V}_i^j}\right|^2}_{(II)} &+ \underbrace{\left|\frac{\sum_{i=1}^n Y_i V_i^j \sum_{i=1}^n V_i^j - \sum_{i=1}^n Y_i V_i^j \sum_{i=1}^n \tilde{V}_i^j}{\sum_{i=1}^n \tilde{V}_i^j \sum_{i=1}^n V_i^j}\right|^2}_{(III)} \Bigg)\end{aligned}$$

using triangular inequality. We bound the three parts separately.

(i) For the numerator of (I), let v_j be the ratio of non-zero elements in $\tilde{V}_1^j, \cdots, \tilde{V}_n^j$, i.e. $v_j = \sum_{i=1}^n \mathbf{1}\{\tilde{V}_1^j \neq 0\}/n$. Note that $\tilde{Y}_i - Y_i = 4M\xi_i/\varepsilon$ and $\tilde{V}_i^j \in [-1, 1]$. Thus, $(\tilde{Y}_i - Y_i)\tilde{V}_i^j$ are either zero or sub-exponential random variables with parameter $(\frac{2M}{\sqrt{2}\varepsilon}, 1)$. Consequently, Lemma B.5 yields that

$$\left|\frac{1}{n}\sum_{i=1}^n (\tilde{Y}_i - Y_i)\tilde{V}_i^j\right| \leq \sqrt{\frac{128M^2 \cdot v_j \cdot \log n}{n\varepsilon^2}} \tag{34}$$

with probability at least $1 - 1/n^2$. For the denominator, applying Lemma B.4 yields

$$\left|\frac{1}{n}\sum_{i=1}^n \tilde{V}_i^j - \frac{1}{n}\sum_{i=1}^n V_i^j\right| \leq \sqrt{\frac{v_j \log n}{n}} \tag{35}$$

with probability at least $1 - 2/n^2$. Moreover, by Lemma B.7, we have

$$\frac{1}{n}\sum_{i=1}^n V_i^j \geq \int_{A\times B} d\mathrm{P}_X(x) - \left|\int_{A\times B} d\mathrm{P}_X(x) - \frac{1}{n}\sum_{i=1}^n V_i^j\right| \gtrsim \frac{1}{t^s \cdot 2^p} - \sqrt{\frac{\log n}{t^s \cdot 2^p \cdot n}} \gtrsim \frac{1}{t^s \cdot 2^p}$$

holds with probability at least $1 - 1/n^2$. Then, we can guarantee that

$$\left|\frac{1}{n}\sum_{i=1}^n \tilde{V}_i^j\right| \gtrsim \frac{1}{t^s \cdot 2^p} - \sqrt{\frac{v_j \log n}{n}} \gtrsim \frac{1}{t^s \cdot 2^p} \tag{36}$$

with probability $1 - 3/n^2$. This together with (34) yields

$$(I) \lesssim \frac{t^{2s} \cdot 2^{2p} \cdot v_j \cdot \log n}{n\varepsilon^2}. \tag{37}$$

(ii) For (II), we apply Lemma B.4 and get

$$\left|\frac{1}{n}\sum_{i=1}^n Y_i\tilde{V}_i^j - \frac{1}{n}\sum_{i=1}^n Y_i V_i^j\right| \leq \sqrt{\frac{v_j \cdot M \log n}{n}}$$

with probability $1-2/n^2$ since $|Y_i| \le M$. Combining this with (36), we get

$$(II) \lesssim \frac{t^{2s} \cdot 2^{2p} \cdot v_j \cdot \log n}{n\varepsilon^2}. \tag{38}$$

(iii) For (III), note that $V_i^j \in \{0,1\}$, we have $\sum_{i=1}^n Y_i V_i^j \le M \sum_{i=1}^n V_i^j$. Thus,

$$\begin{aligned}\left|\frac{\sum_{i=1}^n Y_i V_i^j \sum_{i=1}^n V_i^j - \sum_{i=1}^n Y_i V_i^j \sum_{i=1}^n \tilde{V}_i^j}{\sum_{i=1}^n \tilde{V}_i^j \sum_{i=1}^n V_i^j}\right|^2 \le & M^2 \left|\frac{\sum_{i=1}^n V_i^j - \sum_{i=1}^n \tilde{V}_i^j}{\sum_{i=1}^n \tilde{V}_i^j}\right|^2 \\ \lesssim & \frac{t^{2s} \cdot 2^{2p} \cdot v_j \cdot \log n}{n\varepsilon^2}\end{aligned}$$

where the last inequality follows from (35) and (36). Combining this with (37) and (38), we have

$$\sum_j \frac{1}{t^s 2^p} \left| \frac{\sum_i \tilde{Y}_i \cdot \tilde{V}_i^j}{\sum_i \tilde{V}_i^j} - \frac{\sum_i Y_i \cdot V_i^j}{\sum_i V_i^j} \right|^2 \lesssim \frac{t^s \cdot 2^p \cdot \sum_j v_j \cdot \log n}{n\varepsilon^2}$$

with probability $\mathrm{P}^n \times \mathrm{R}^n$ at least $1-4/n^2$. By Lemma B.10,

$$\sum_j v_j \le \frac{1}{n} \sum_{i=1}^n \sum_j \mathbf{1}\{V_i^j \neq 0\} \lesssim \frac{t^s \log n}{n\varepsilon^2} \sum_{i=1}^n 2^{\sum_{\ell=s+1}^d W_i^\ell \cdot p/(d-s)}.$$

This directly implies the desired bound of $\mathcal{R}_{L,\mathrm{P}}(f_\pi^{\mathrm{DP}}) - \mathcal{R}_{L,\mathrm{P}}(f_\pi)$. □

Proof of Lemma B.2. We intend to bound

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(\overline{f}) \le & \max_j \left| \frac{\frac{1}{n}\sum_{i=1}^n Y_i V_i^j}{\frac{1}{n}\sum_{i=1}^n V_i^j} - \frac{\int_{A\times B} f^*(x') d\mathrm{P}_X(x')}{\int_{A\times B} d\mathrm{P}_X(x)} \right|^2 \\ \le & \underbrace{\left| \frac{\frac{1}{n}\sum_{i=1}^n Y_i V_i^j \int_{A\times B} d\mathrm{P}_X(x') - \int_{A\times B} f^*(x') d\mathrm{P}(x') \int_{A\times B} d\mathrm{P}_X(x')}{\frac{1}{n}\sum_{i=1}^n V_i^j \int_{A\times B} d\mathrm{P}_X(x)} \right|^2}_{(I)} \\ & + \underbrace{\left| \frac{\int_{A\times B} f^*(x') d\mathrm{P}(x') \int_{A\times B} d\mathrm{P}_X(x') - \frac{1}{n}\sum_{i=1}^n V_i^j \int_{A\times B} f^*(x') d\mathrm{P}_X(x')}{\frac{1}{n}\sum_{i=1}^n V_i^j \int_{A\times B} d\mathrm{P}_X(x)} \right|^2}_{(II)}\end{aligned}$$

We bound two terms separately. For (I), Lemma B.7 yields

$$\frac{1}{n}\sum_{i=1}^n V_i^j \int_{A_j} d\mathrm{P}_X(x) \gtrsim t^{-2s} \cdot 2^{-2p} \tag{39}$$

with probability $1-1/n^2$. For the numerator, Lemma B.8 yields

$$\left| \frac{1}{n}\sum_{i=1}^n Y_i V_i^j - \int_{A\times B} f^*(x') d\mathrm{P}_X(x') \right|^2 \left| \int_{A\times B} d\mathrm{P}_X(x') \right|^2 \lesssim \frac{\log n}{t^s 2^p \cdot n} \cdot t^{-2s} \cdot 2^{-2p}.$$

Together, we get $(I) \lesssim t^s 2^p/n$. Analogously, by Lemma B.7, we have

$$\left| \int_{A\times B} d\mathrm{P}_X(x') - \sum_{i=1}^n V_i^j \right|^2 \left| \int_{A\times B} f^*(x') d\mathrm{P}_X(x') \right|^2 \le \frac{1}{t^s 2^p \cdot n} \cdot M^2 \cdot t^{-2s} \cdot 2^{-2p}.$$

This together with (39) yields $(II) \lesssim t^s 2^p/n$. The bound of (I) and (II) together yields the desired conclusion. □

Proof of Lemma B.3. Let $A \times B(x)$ be the grid of which x belongs to. We intend to bound

$$\mathcal{R}_{L,\mathrm{P}}(\overline{f}) - \mathcal{R}_{L,\mathrm{P}}(f^*) = \int_{x\in\mathcal{X}} \left(\frac{\int_{A\times B(x)} f(x')d\mathrm{P}_X(x')}{\int_{A\times B(x)} d\mathrm{P}_X(x')} - f(x) \right)^2 d\mathrm{P}_X(x).$$

For each x, if Assumption 4.1 holds, the point-wise error can be bounded by

$$\begin{aligned}\frac{\int_{A\times B(x)} f(x')d\mathrm{P}_X(x')}{\int_{A\times B(x)} d\mathrm{P}_X(x')} - f(x) \leq & \frac{\int_{A\times B(x)} |f(x') - f(x)|d\mathrm{P}_X(x')}{\int_{A\times B(x)} d\mathrm{P}_X(x')} \\ \leq & \frac{\int_{A\times B(x)} c_L\|x'-x\|^\alpha d\mathrm{P}_X(x')}{\int_{A\times B(x)} d\mathrm{P}_X(x')} \leq c_L \mathrm{diam}(A\times B(x))^\alpha\end{aligned}$$

Then we can bound the excess risk using Lemma B.9,

$$\mathcal{R}_{L,\mathrm{P}}(\overline{f}) - \mathcal{R}_{L,\mathrm{P}}(f^*) \lesssim \mathrm{diam}(A\times B(x))^\alpha \lesssim 2^{-2p\alpha/(d-s)} + t^{-2\alpha}.$$

$\square$

C Experiment Details

C.1 Implementation Details

For `HistOfTree`, we adopt the generalized randomized response mechanism introduced in Section A.1. We add one more parameter adjusting the allocation of the privacy budget on the numerator and denominator. Specifically, let the mechanism for $\tilde{Y}_i$ and $\tilde{U}_i$ be respectively $\rho\varepsilon$-LDP and $(1-\rho)\varepsilon$-semi-feature LDP for $\rho \in [0,1]$. In this case, the mechanisms are which means the hybrid mechanism is still ε-LDP. We select $\rho \in \{0.5, 0.7, 0.9\}$. For other parameters, we select $p \in \{1,2,4,6\}$ and $t \in \{1,2,3\}$.

For `AdHistOfTree`, we also adopt the generalized randomized response mechanism. For better empirical performance, the second term of the objective function is multiplied with a constant selected in $\{0.01, 0.1, 1\}$. Also, we allow the selection $t \in \{t^*-1, t^*, t^*+1\}$, where t^* is the default value derived in Section 4.3.

For `Hist`, we implement the private histogram proposed by (Berrett et al., 2021; Györfi and Kroll, 2022) in Python. `Hist` applies the Laplacian mechanism to privatize the estimation of marginal and joint probabilities for a cubic histogram partition with the number of bins t. In cells with a private marginal probability estimation less than ζ, estimation is truncated to 0. We let $t \in \{1,2,3,4\}$. We set the truncation parameter $\zeta \in \{0.01, 0.05\}$.

For `ParDT` and `LabelDT`, we add Laplace noise with scale ξ/ε where ξ is the range of the label. For `ParDT`, `LabelDT`, `KRR` and `DT`, we use the implementation by Scikit-Learn (Pedregosa et al., 2011). We select *max_depth* in $\{1,2,4,6,8\}$ and *min_sample_leaf* in $\{1,10,100\}$. For `KRR`, we select k in $\{2,3,4,5\}$.

C.2 Dataset Details

For all datasets, each feature is min-max scaled to the range $[0,1]$ individually.

ABA: The *Abalone* dataset originally comes from biological research (Nash et al., 1994) and now it is accessible on UCI Machine Learning Repository (Dua and Graff, 2017). ABA contains 4177 observations of one target variable and 8 attributes related to the physical measurements of abalone.

AQU: The *QSAR aquatic toxicity* dataset was used to develop quantitative regression QSAR models to predict acute aquatic toxicity towards the fish Pimephales promelas (fathead minnow) on a set of 908 chemicals. It contains 546 instances of 8 input attributes and 1 output attribute.

BUI: The *Residential Building Data Set Data Set* dataset on UCI Machine Learning Repository includes construction cost, sale prices, project variables, and economic variables corresponding to real estate single-family residential apartments in Tehran, Iran. It contains 372 instances of 103 input attributes and 2 output attributes.

DAK: The *Istanbul Stock Exchange* dataset includes returns of the Istanbul stock exchange with seven other international indexes. It has 536 instances with 10 features.

FIS: The *QSAR fish toxicity* dataset on UCI Machine Learning Repository was used to develop quantitative regression QSAR models to predict acute aquatic toxicity towards the fish Pimephales promelas (fathead minnow) on a set of 908 chemicals. It contains 908 instances of 7 features.

MUS: The *The Geographical Original of Music* dataset was built from a personal collection of 1059 tracks covering 33 countries/areas. The program MARSYAS (Zhou et al., 2014) was used to extract audio features from the wave files. We used the default MARSYAS settings in single vector format (68 features) to estimate the performance with basic timbal information covering the entire length of each track.

POR: The *Stock Portfolio Performance* data set of performances of weighted scoring stock portfolios are obtained with mixture design from the US stock market historical database (Liu and Yeh, 2017). It has 315 samples with 12 features.

PYR: The *Pyrimidines* dataset is a subset of the Qualitative Structure Activity Relationships dataset on UCI datasets. This sub-dataset has 74 instances of dimension 27

RED: This dataset contains the information on red wine of the *Wine Quality* dataset (Cortez et al., 2009) on UCI Machine Learning Repository. There are 11 input variables to predict the output variable wine quality. 4898 instances are collected in the dataset.

WHI: This dataset also originates from the *Wine Quality* dataset (Cortez et al., 2009) on the UCI Machine Learning Repository. There are 11 features related to white wine to predict the corresponding wine quality.

C.3 Additional Experiment Results

Table 2: Real data performance for $\varepsilon = 1$.

	Aligned					Personalized						
	`HistOfTree` ME	`HistOfTree` CART	`ParDT`	`Hist`	`KRR`	`HistOfTree` ME	`HistOfTree` CART	`AdHistOfTree` ME	`AdHistOfTree` CART	`ParDT`	`Hist`	`KRR`
ABA	2.04	**1.80***	2.04	2.02	2.74	**1.85***	2.04	1.86	2.00	2.26	2.02	2.71
AQU	**1.71***	**1.71***	2.86	1.80	3.35	1.71	1.73	**1.69***	**1.69***	3.03	1.80	3.35
BUI	**1.59***	**1.59***	3.18	5e5	6.70	1.64	1.64	**1.61***	**1.61***	3.28	5e5	6.57
DAK	6.69	**6.65***	8.03	8.43	22.6	7.95	7.95	**7.47***	**7.47***	19.0	8.43	22.6
FIS	**2.22***	2.32	3.22	2.45	4.60	2.32	2.32	**2.26***	**2.26***	3.49	2.45	4.59
MUS	1.17	**1.17**	1.89	9e5	2.72	1.17	1.17	**1.16**	**1.16**	1.99	9e5	2.74
POR	**4.15***	**4.15***	4.67	7.22	7.89	4.04	4.04	**3.68***	**3.68***	4.72	7.22	7.70
PYR	**2.31***	**2.31***	2.98	1e5	5.61	2.32	2.32	**1.75***	**1.75***	3.18	1e5	5.64
RED	**1.48***	**1.48***	2.02	1.81	2.53	**1.48***	1.52	**1.48***	**1.48***	2.08	1.81	2.53
WHI	**1.46***	**1.46***	1.61	1.61	1.85	1.47	1.47	**1.46**	**1.46**	1.70	1.61	1.84
Rank Sum	21	**18**	35	38	47	26	34	**11**	12	56	56	67

Table 3: Real data performance for $\varepsilon = 4$.

	Aligned					Personalized						
	`HistOfTree` ME	`HistOfTree` CART	`ParDT`	`Hist`	`KRR`	`HistOfTree` ME	`HistOfTree` CART	`AdHistOfTree` ME	`AdHistOfTree` CART	`ParDT`	`Hist`	`KRR`
ABA	1.72	1.59	**1.58**	1.97	1.64	**1.54**	1.63	1.55	1.62	1.61	1.97	1.97
AQU	1.45	**1.35***	1.56	1.58	1.64	1.50	1.47	1.54	**1.45***	1.68	1.58	1.65
BUI	1.23	**1.16***	1.32	3e4	1.30	1.26	**1.19***	1.28	1.22	1.36	3e4	1.29
DAK	4.67	**4.48***	8.03	6.32	6.57	**4.53**	4.55	5.95	5.57	6.01	6.32	6.58
FIS	1.66	**1.64**	1.65	2.12	1.82	**1.62***	1.66	1.77	1.68	1.68	2.12	1.77
MUS	**1.12**	**1.12**	1.15	6e3	1.36	**1.10**	1.12	**1.10**	1.12	1.22	6e3	1.14
POR	2.89	**2.40***	2.98	3.21	3.09	**2.66**	2.75	2.68	2.72	2.97	3.31	3.02
PYR	1.26	**1.16***	1.33	7e2	1.49	1.29	1.29	**1.27**	**1.27**	1.43	7e2	1.50
RED	**1.27**	1.47	1.29	1.49	1.44	1.27	**1.26**	1.30	1.27	1.28	1.49	1.44
WHI	**1.24**	**1.24**	1.26	1.51	1.42	1.24	**1.23**	1.43	1.37	1.30	1.51	1.42
Rank Sum	22	**16**	29	47	38	**18**	24	34	26	47	66	57