Electronic Journal of Statistics
Vol. 15 (2021) 2430–2453
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1845

Strongly universally consistent nonparametric regression and classification with privatised data

Thomas B. Berrett

Department of Statistics,
University of Warwick,
Coventry, CV4 7AL,
United Kingdom
e-mail: tom.berrett@warwick.ac.uk

László Györfi

Department of Computer Science and Information Theory,
Budapest University of Technology and Economics,
Magyar Tudósok krt. 2., Budapest, H-1117,
Hungary
e-mail: gyorfi@cs.bme.hu

Harro Walk

Institute of Stochastics and Applications,
University of Stuttgart,
Pfaffenwaldring 57, D-70569 Stuttgart,
Germany
e-mail: harro.walk@mathematik.uni-stuttgart.de

Abstract: In this paper we revisit the classical problem of nonparametric regression, but impose local differential privacy constraints. Under such constraints, the raw data $(X_1, Y_1), \ldots, (X_n, Y_n)$, taking values in $\mathbb{R}^d \times \mathbb{R}$, cannot be directly observed, and all estimators are functions of the randomised output from a suitable privacy mechanism. The statistician is free to choose the form of the privacy mechanism, and here we add Laplace distributed noise to a discretisation of the location of a feature vector X_i and to the value of its response variable Y_i. Based on this randomised data, we design a novel estimator of the regression function, which can be viewed as a privatised version of the well-studied partitioning regression estimator. The main result is that the estimator is strongly universally consistent, and we further establish an upper bound on the rate of convergence. Our methods and analysis also give rise to a strongly universally consistent binary classification rule for locally differentially private data.

AMS 2000 subject classifications: 62G08, 62G20, 68P27.

Keywords and phrases: Regression estimate, classification, local differential privacy, universal consistency.

Received November 2020.

1. Introduction

In recent years there has been a surge of interest in data analysis methodology that is able to achieve strong statistical performance without comprimising the privacy and security of individual data holders. This has often been driven by applications in modern technology, for example by Google [16], Apple [30], and Microsoft [11], but the study goes at least as far back as [35] and is often used in more traditional fields of clinical trials [32, 8] and census data [25, 14]. While there has long been an awareness that sensitive data must be anonymised, it has become apparent only relatively recently that simply removing names and addresses is insufficient in many cases [e.g. 29, 26]. The concept of differential privacy [15] was introduced to provide a rigorous notion of the amount of private information on individuals published statistics contain. Statistical treatments of this framework include [36, 23, 2, 6].

Although it is a suitable constraint for many problems, procedures that are differentially private often require the presence of a third party, who may be trusted to handle the raw data before statistics are published. To address this shortcoming, the local differential privacy constraint [see, for example, 21, 12, and the references therein] was introduced to provide a setting where analysis must be carried out in such a way that each raw data point is only ever seen by the original data holder. The simplest example of a locally differentially private mechanism is the randomised response [35] used with binary data, but mechanisms have also been developed for tasks such as classification [3], generalised linear modelling [12], empirical risk minimisation [33], density estimation [5], functional estimation [27] and goodness-of-fit testing [4].

Regression is a cornerstone of modern statistical analysis, routinely used across the sciences and beyond. We recall that, in a standard stochastic model, a regression estimator predicts for an observed d dimensional random feature vector an unknown random response, with finite second moment. The regression function, given by the conditional expectation of the response given the feature vector, achieves minimum mean squared error. Typically, the statistician does not know the underlying stochastic structure, but has access to a corresponding finite sample of independent identically distributed design-response vectors in $\mathbb{R}^d \times \mathbb{R}$, and on this basis estimates the regression function. The background will be given below at the beginning of Section 2, and in the following we shall refer several times to the monograph of [19]. A binary classification (pattern recognition) rule predicts for a feature vector an unknown random response taking values in $\{-1, 1\}$. The so-called Bayes decision rule achieves minimum error probability (Bayes error). Given a finite sample of i.i.d. design-response vectors in $\mathbb{R}^d \times \{-1, 1\}$, the Bayes rule is approximated. We formulate the setup in Section 5, while the monograph of [9] contains a detailed theory of nonparametric classification.

While regression has been relatively well-studied in the non-local model of differential privacy [e.g. 6], results in the local model are scarce. [37] studies sparse linear regression, kernel ridge regression and GLMs. [28, 33] study parametric empirical risk minimisation. [34] studies sparse linear regression. [12, 13]

study GLMs. The recent work [17] concerns a relaxed version of the locally private regression model where responses can be observed exactly, and empirically studies a Nadaraya–Watson-type estimator, but we are unaware of any other work on locally private nonparametric regression. The simpler problem of binary classification is studied in [3], but there are significant additional challenges in designing a suitable estimator for the regression problem.

In this paper we introduce and investigate a new method for nonparametric regression under α-local differential privacy constraints and also present a corresponding classification rule. For regression our procedure combines a simple non-interactive privacy mechanism with a cubic partitioning regression estimate modifying the regressogram, which was originally introduced by [31] and has been well-studied since [see, e.g., 19, Chapter 4 and Section 23.1, and the references therein]. In Section 3 we describe the procedure and state that the sequence of estimates is strongly universally consistent, in that the L_2-risk converges almost surely to zero in the large sample limit for any data-generating distribution for which the response has a finite second moment. Moreover, we give an upper bound on the rate of convergence of this estimator. Let us mention that in the degenerate case without privacy the estimator reduces to the strongly universally consistent partitioning estimator of [18]. The problem of classification is strictly easier than regression, therefore our methods and analysis also give rise to a strongly universally consistent binary classification rule for locally differentially private data.

The remainder of the paper is organised as follows. In Section 2 we introduce the necessary background on regression and local differential privacy. In Section 3 we introduce our privacy mechanism and estimators, and state our main results in the regression setting, discussing their implications for local differential privacy in Section 4. In Section 5 we study the consequences of the results for binary classification. All proofs will be deferred to Section 6.

2. Preliminaries

Let (X, Y) be a pair of random variables such that the feature vector X takes values in $\mathbb{R}^d$ and its response variable Y is a real-valued random variable with $\mathbb{E}[Y^2] < \infty$. We denote by μ the distribution of the feature vector X, that is, for all measurable sets $A \subset \mathbb{R}^d$, we have $\mu(A) = \mathbb{P}\{X \in A\}$. Then the *regression function*

$$m(x) = \mathbb{E}[Y \mid X = x] \tag{1}$$

is well defined for μ-almost all x. For each measurable function $g : \mathbb{R}^d \to \mathbb{R}$ one has

$$\mathbb{E}\left[\{g(X) - Y\}^2\right] = \mathbb{E}\left[\{m(X) - Y\}^2\right] + \mathbb{E}\left[\{m(X) - g(X)\}^2\right],$$

therefore, with the notation

$$L^* = \mathbb{E}\left[\{m(X) - Y\}^2\right] \ ,$$

we have

$$\mathbb{E}\left[\{g(X)-Y\}^2\right] = L^* + \int \{m(x)-g(x)\}^2 \mu(dx). \tag{2}$$

We measure the performance of an estimator $\hat{m}$ of m through the loss function

$$L(m,\hat{m}) := \int \{m(x)-\hat{m}(x)\}^2 \mu(dx),$$

which, by (2), may be interpreted as the excess prediction risk for a new observation X.

In this paper we are mainly concerned with regression estimates $\hat{m}$ based on partitions of the sample space, which were originally studied by [31]. Let $\mathcal{P}_h = \{A_{h,1}, A_{h,2}, \ldots\}$ be a cubic partition of $\mathbb{R}^d$ such that the cells $A_{h,j}$ are cubes of volume h^d. If $x_{h,j}$ denotes the center of the cube $A_{h,j}$, then introduce the discretization of x by the quantiser

$$Q_h(x) := x_{h,j}, \text{ if } x \in A_{h,j}.$$

The raw data will be independent and identically distributed copies

$$\mathcal{D}_n := \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$$

of the random vector (X,Y), and the estimators that we consider will be (randomised) functions of the binned data, defined by

$$\{(Q_h(X_1),Y_1),\ldots,(Q_h(X_n),Y_n)\}.$$

Using this binned data, when we do not have to satisfy privacy constraints, one may create a scheme for a public data set as follows: there are n individuals in the study such that individual i generates the sample pair (X_i,Y_i) and he submits the discretised version $(Q_h(X_i),Y_i)$ to a data collector. The data collector calculates the empirical distributions

$$\nu_n(A_{h,j}) = \frac{1}{n} \sum_{i;Q_h(X_i)=x_{h,j}} Y_i$$

and

$$\mu_n(A_{h,j}) = \frac{1}{n} \sum_{i;Q_h(X_i)=x_{h,j}} 1.$$

Then, the public data set

$$D_{n,h} = \{(j,\nu_n(A_{h,j}),\mu_n(A_{h,j})); \mu_n(A_{h,j}) > 0\}$$

is published. The data set $D_{n,h}$ has the favourable property that with high probability the size $\#(D_{n,h})$ is much less than n [cf. 24].

Using this binned data and allowing $h = h_n$ to depend on the sample size, the partitioning regression estimate is defined by

$$\check{m}_n(x) = \frac{\nu_n(A_{h_n,j})}{\mu_n(A_{h_n,j})} = \frac{\sum_{i=1}^n Y_i \mathbb{I}_{\{X_i \in A_{h_n,j}\}}}{\sum_{i=1}^n \mathbb{I}_{\{X_i \in A_{h_n,j}\}}} \qquad \text{if } x \in A_{h_n,j}, \tag{3}$$

where $0/0$ is 0 by definition and $\mathbb{I}$ denotes the indicator function. In order to have strong universal consistency, [19] modify the partitioning regression estimate as follows:

$$m_n(x) = \frac{\nu_n(A_{h_n,j})}{\mu_n(A_{h_n,j})} \mathbb{I}_{\{\mu_n(A_{h_n,j}) \geq \log n/n\}} \qquad \text{if } x \in A_{h_n,j}.$$

Theorem 1 (Theorem 23.3 in [19]). *If*

$$\lim_{n\to\infty} h_n = 0 \ \text{ and } \ \lim_{n\to\infty} nh_n^d / \log n = \infty,$$

then the estimate m_n is strongly universally consistent, i.e.,

$$\lim_{n\to\infty} \int (m(x) - m_n(x))^2 \mu(dx) = 0 \tag{4}$$

a.s. for any distribution of (X, Y) with $\mathbb{E}Y^2 < \infty$.

There is a huge literature on weak and strong universal consistency of regression estimates. Weak universal consistency means convergence

$$\lim_{n\to\infty} \int \mathbb{E}\left[\{m(x) - m_n(x)\}^2\right] \mu(dx) = 0$$

for any distribution of (X, Y) with $\mathbb{E}Y^2 < \infty$. For the weak universal consistency of local averaging regression estimates m_n, which includes partitioning estimates, kernel estimates and nearest neighbor estimates, we refer to Chapters 4–6 in [19].

3. Our regression estimation method and its strong universal consistency

Similarly to [3] we consider locally privatised data given as follows: the privacy mechanism is formulated by independent double arrays $\{\epsilon_{i,j}\}$ and $\{\zeta_{i,j}\}$ such that the elements of the arrays are i.i.d. with centred, unit-variance Laplace distributions. For $i = 1, \ldots, n$ and for $0 < M_n \leq \infty$, we write $[Y_i]_{-M_n}^{M_n} = \min\{M_n, \max(Y_i, -M_n)\}$ for the truncated response; it will be sometimes be convenient to write $[Y_i]_{-\infty}^{\infty} = Y_i$ for no truncation. Choose a sphere S_n centered at the origin. Assume that the cells $A_{h,j}$ are numbered such that $A_{h,j} \cap S_n \neq \emptyset$ when $j \leq N_n$ for some integer $N_n > 0$, and $A_{h,j} \cap S_n = \emptyset$ otherwise. Individual $i \leq n$ generates and transmits the data

$$Z_{i,j} := [Y_i]_{-M_n}^{M_n} \mathbb{I}_{\{X_i \in A_{h,j}\}} + \sigma_Z \epsilon_{i,j}, \quad j \leq N_n \tag{5}$$

and

$$W_{i,j} := \mathbb{I}_{\{X_i \in A_{h,j}\}} + \sigma_W \zeta_{i,j}, \quad j \le N_n, \tag{6}$$

where $\sigma_Z > 0$ and $\sigma_W > 0$. This means that individual i generates noisy data for any cell $A_{h,j}$ with $j \le N_n$. Proposition 1 in Section 4 shows that, for suitable choices of σ_W and σ_Z, this mechanism satisfies the α-LDP constraint. For such σ_W, σ_Z, the data set

$$\tilde{D}_{n,h} = \{(j, \tilde{\nu}_n(A_{h,j}), \tilde{\mu}_n(A_{h,j})) : j = 1, \dots, N_n\}$$

may be published without violating the α-LDP constraint, where

$$\tilde{\nu}_n(A_{h,j}) = \frac{1}{n}\sum_{i=1}^{n} Z_{i,j}\mathbb{I}_{\{j \le N_n\}} \quad \text{and} \quad \tilde{\mu}_n(A_{h,j}) = \frac{1}{n}\sum_{i=1}^{n} W_{i,j}\mathbb{I}_{\{j \le N_n\}}. \tag{7}$$

Now that we have introduced our privacy mechanism we may define our estimator of m based on $\tilde{D}_{n,h}$. For $c_n > 0$ we define

$$\tilde{m}_n(x) = \frac{\tilde{\nu}_n(A_{h_n,j})}{\tilde{\mu}_n(A_{h_n,j})}\mathbb{I}_{\{\tilde{\mu}_n(A_{h_n,j}) \ge c_n h_n^d\}}\mathbb{I}_{\{j \le N_n\}} \quad \text{when } x \in A_{h_n,j}.$$

This is a novel estimator that extends the classical partitioning regression estimate to the LDP setting. In non-private settings such estimators may be seen as averaging the value of the response over each element of the partition, but here we are unable to retain this interpretation as we cannot know exactly how many data points fall in each cell. This lack of knowledge is particularly problematic in low-density regions, where the estimate of μ is necessarily especially noisy, and where our estimator must be carefully defined. A crucial component of the estimate is the way it detects the empty cells and truncates. If X has a density, then $\mu(A_{h_n,j})$ is of order h_n^d. Furthermore, on the support of an arbitrary μ, $\mu(A_{h_n,j})/h_n^d$ is bounded away from zero. More precisely, if $A_n(x)$ stands for the cube $A_{h_n,j}$ containing x, then

$$\liminf_n \mu(A_n(x))/h_n^d > 0$$

for μ-almost all x, at least when we have a nested sequence of partitions, see Lemma 24.10 in [19]. Thus, for arbitrary μ, the order of $\mu(A_{h_n,j})$ is at least h_n^d. Therefore, $c_n \to 0$ implies that $\mu(A_{h_n,j}) > c_n h_n^d$, for large enough n.

When $\sigma_W = \sigma_Z = 0$ and $c_n = \log n/(nh_n^d)$ then we recover the non-private partitioning estimator, which has access to the raw data, discussed above.

Our first main new result extends Theorem 1 to the private setting where $\sigma_W, \sigma_Z > 0$ are fixed, and establishes the strong universal consistency of $\tilde{m}_n$.

Theorem 2. *If $S_n \uparrow \mathbb{R}^d$, $c_n \to 0$, $h_n \to 0$, $M_n \to \infty$ and*

$$\frac{(\log n)^3}{nc_n^2 h_n^{2d}} \to 0 \tag{8}$$

then

$$\lim_{n\to\infty} \int \{m(x) - \tilde{m}_n(x)\}^2 \mu(dx) = 0 \quad a.s., \tag{9}$$

for any distribution of (X, Y) *with* $\mathbb{E}Y^2 < \infty$.

The proof of Theorem 2 shows that replacement of (8) by $nc_n^2 h_n^{2d} \to \infty$ yields the weak universal consistency of $\tilde{m}_n$.

Comparing with Theorem 1, we see that that the usual condition $nh_n^d \to \infty$ has been replaced by $nh_n^{2d} \to \infty$. Heuristically, this difference can be understood by considering the properties of $\tilde{\nu}_n(A_{h_n,j})$. Writing $\nu(A) := \int_A m(x)\mu(dx)$, we have

$$\mathbb{E}\{\tilde{\nu}_n(A_{h_n,j})\} = \nu(A_{h_n,j}),$$

which is the same as in the non-private case. However, we see a difference when we consider that

$$n\mathbb{V}ar\{\tilde{\nu}_n(A_{h_n,j})\} = n\mathbb{V}ar\{\nu_n(A_{h_n,j})\} + \sigma_Z^2. \tag{10}$$

In the non-private case, the only contribution is from the first term, which can be seen to typically be $O(h_n^d)$. However, in the private case we will usually take σ_Z to be large, and hence the variance in (10) is dominated by the second term, which does not vanish with h_n. This occurs in other LDP problems [e.g. 4]; the privacy constraint introduces an unavoidable homoscedastic term into the variance of our estimator, which results in very different behaviour, including a curse-of-dimensionality that is often more severe than in non-private problems.

The proof techniques used for Theorem 2 can be used to derive upper bounds on the rates of convergence of our estimator for suitable data-generating mechanisms.

Theorem 3. *If* $S_n \uparrow \mathbb{R}^d$, $c_n \to 0$, $h_n \to 0$, $M_n \to \infty$, m *is Lipschitz continuous,* Y *is bounded and* X *has a density, which is bounded away from zero, then*

$$\mathbb{E}\int (m(x) - \tilde{m}_n(x))^2 \mu(dx) = O\left(\frac{1}{nc_n^2 h_n^{2d}}\right) + O(h_n^2).$$

For the choices

$$h_n = c' n^{-1/(2(d+1))}$$

and

$$c_n = 1/\sqrt{\log n}, \tag{11}$$

this upper bound results in

$$\mathbb{E}\int (m(x) - \tilde{m}_n(x))^2 \mu(dx) = O\left(\frac{\log n}{n^{1/(d+1)}}\right).$$

We conjecture that

$$O\left(\frac{1}{n^{1/(d+1)}}\right)$$

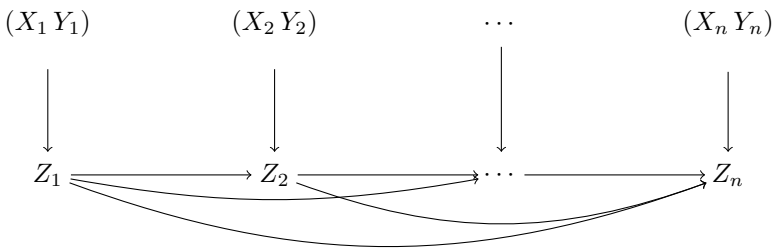
is the minimax lower bound over all α-LDP privacy mechanisms for Lipschitz continuous regression function, which would imply that our estimate is minimax optimal up to a factor of $\log n$. Furthermore, the lower bound on the density appears to be crucial; we speculate that if the density is not bounded away from zero, then the rate of convergence of any estimate can be arbitrarily slow.

4. Local differential privacy

As discussed above, when working under privacy constraints, no estimator can have direct access to the raw data $\mathcal{D}_n$, or even the binned data $\mathcal{D}_{n,h}$. Instead, it will only be allowed to depend on randomised data $(Z_1, \ldots, Z_n)$, defined on some measurable space $(\mathcal{Z}^n, \mathcal{B}^n)$, that has been generated conditional on $\mathcal{D}_n$. Formally, a *privacy mechanism* is a conditional distribution $Q : \mathcal{B}^n \times (\mathbb{R}^d \times \mathbb{R})^{\otimes n} \to [0,1]$ with the interpretation that

$$(Z_1, \ldots, Z_n) | \{\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}\} \sim Q(\cdot | (x_1, y_1), \ldots, (x_n, y_n)).$$

This privacy mechanism will be said to be *sequentially interactive* [12] if it respects the graphical structure



In particular, this requires that $Z_i \perp\!\!\!\perp (X_j, Y_j) | \{X_i, Y_i, Z_1, \ldots, Z_{i-1}\}$ for any $j \neq i$, so that Z_i is generated with only the knowledge of (X_i, Y_i) and $Z_1, \ldots, Z_{i-1}$. For this reason, such privacy mechanisms are said to be locally private. Sequentially interactive privacy mechanisms may be specified by a sequence of conditional distributions $(Q_1, \ldots, Q_n)$ with $Q_i : \mathcal{B} \times (\mathbb{R}^d \times \mathbb{R}) \times \mathcal{Z}^{i-1} \to [0,1]$ and with the interpretation that

$$Z_i | \{(X_i, Y_i) = (x_i, y_i), Z_1 = z_1, \ldots, Z_{i-1} = z_{i-1}\} \sim Q_i(\cdot | (x_i, y_i), z_1, \ldots, z_{i-1}).$$

Given $\alpha > 0$, a sequentially interactive mechanism specified by $(Q_1, \ldots, Q_n)$ is said to be α-*locally differentially private* (α-LDP) if

$$\sup_{A \in \mathcal{B}} \sup_{z_1, \ldots, z_{i-1} \in \mathcal{Z}} \sup_{(x_i, y_i), (x_i', y_i') \in \mathbb{R}^d \times \mathbb{R}} \frac{Q_i(A | (x_i, y_i), z_1, \ldots, z_{i-1})}{Q_i(A | (x_i', y_i'), z_1, \ldots, z_{i-1})} \leq e^{\alpha}$$

for each $i = 1, \dots, n$. Let $\mathcal{Q}_\alpha$ denote the set of all α-LDP privacy mechanisms.

Our privacy mechanisms, given by (5) and (6), are actually of a simpler, *non-interactive* form, where, with $Z_i = (W_{i,j}, Z_{i,j})_{j=1}^{N_n}$ for $i = 1, \dots, n$, we also have

$$Z_i \perp\!\!\!\perp (X_j, Y_j, Z_j)$$

for all $j \neq i$. In this case we have

$$Q(A_1, \dots, A_n | (x_1, y_1), \dots, (x_n, y_n)) = \prod_{i=1}^{n} Q_i(A_i | (x_i, y_i))$$

for all $(A_1, \dots, A_n) \in \mathcal{B}^n$. Such mechanisms satisfy the α-LDP constraint if and only if

$$\sup_{A \in \mathcal{B}} \sup_{(x_i, y_i), (x_i', y_i') \in \mathbb{R}^d d \times \mathbb{R}} \frac{Q_i(A | x_i, y_i)}{Q_i(A | x_i', y_i')} \leq e^\alpha$$

for each $i = 1, \dots, n$. Non-interactive mechanisms are computationally attractive in practice as they require minimal communication between the statistician and the orginal data holders, and in large-scale applications there are many practical barriers to interactivity [20].

The following result studies the local differential privacy of the mechanism given by (5) and (6) in the case that $N_n = \infty$, but it is a straightforward consequence of this that the mechanism satisfies the same bound when $N_n < \infty$.

Proposition 1. *Consider the privacy mechanism defined in* (5) *and* (6) *when* $\varepsilon_{1,1}$ *and* $\zeta_{1,1}$ *have unit-variance Laplace distribution with probability density* $x \mapsto \exp(-\sqrt{2}|x|)/\sqrt{2}$. *Writing* $q_{W,Z|X,Y}(w, z|x, y)$ *for the probability density function of* $((W_{1,j})_{j=1}^{\infty}, (Z_{1,j})_{j=1}^{\infty})$ *conditional on* $X_1 = x, Y_1 = y$, *we have*

$$\sup_{w,z \in \mathbb{R}^{\mathbb{N}}} \sup_{x,x' \in \mathbb{R}^d} \sup_{y,y' \in [-M,M]} \frac{q_{W,Z|X,Y}(w, z|x, y)}{q_{W,Z|X,Y}(w, z|x', y')} \leq \exp\Big(2^{3/2}/\sigma_W + 2^{3/2} M/\sigma_Z\Big).$$

Given $\alpha > 0$, we can therefore ensure that our privacy mechanism is α-LDP by choosing M, σ_W, σ_Z such that $2^{3/2}(1/\sigma_W + M/\sigma_Z) \leq \alpha$. This is satisfied if, for example, we take $\sigma_W^2 = 32/\alpha^2$ and $\sigma_Z^2 = 32M^2/\alpha^2$.

In problems of differential privacy one often wants to work in a high-privacy regime, where we have $\alpha \to 0$ as $n \to \infty$. With our privacy mechanism, this requires that $\min(\sigma_W, \sigma_Z/M_n) \to \infty$, and so we remark that Theorem 2 can easily be extended to the setting in which the variances σ_Z^2 and σ_W^2 may depend on the sample size n. Replacing the condition (8) with

$$\frac{(\log n)^3 (1 + \sigma_{Z,n}^2 + \sigma_{W,n}^2)}{n c_n^2 h_n^{2d}} \to 0,$$

a straightforward extension of the proof of Theorem 2 implies the strong universal consistency. Choosing $\sigma_Z \asymp M_n/\alpha$, with $M_n \to \infty$ and

$$\frac{(\log n)^3 \max(1, M_n^2/\alpha^2)}{n c_n^2 h_n^{2d}} \to 0,$$

then our mechanism satisfies the α-LDP constraint and the strong universal consistency holds.

5. Consequences in classification

For the setup of binary classification, let the feature vector X take values in $\mathbb{R}^d$, and let its label Y be ± 1 valued. If g is an arbitrary decision function then its error probability is denoted by

$$L(g) = \mathbb{P}\{g(X) \neq Y\}.$$

The Bayes decision rule g^*, given by

$$g^*(x) = sign\, m(x),$$

where $sign(z) = 1$ for $z > 0$ and $sign(z) = -1$ for $z \leq 0$, minimises the error probability. Let

$$L^* = \mathbb{P}\{g^*(X) \neq Y\}$$

denotes its error probability.

For privatised data, the partitioning classification rule is defined by

$$g_n(x) = sign\,(\tilde{\nu}_n(A_{h_n,j})) \quad \text{when } x \in A_{h_n,j}.$$

Note that this rule does not use the data $\{W_{i,j}\}$. Under the conditions

$$\lim_{n\to\infty} h_n = 0 \text{ and } \lim_{n\to\infty} nh_n^{2d} = \infty,$$

[3] showed that the partitioning classification rule g_n is weakly universally consistent, i.e.,

$$\lim_{n\to\infty} \mathbb{E}\{L(g_n)\} = L^*$$

for any distribution of (X, Y). Our work here allows us to strengthen this result to the following theorem on strong universal consistency:

Theorem 4. *If*

$$\lim_{n\to\infty} h_n = 0 \text{ and } \lim_{n\to\infty} nh_n^{2d}/\log n = \infty,$$

then the classification rule g_n is strongly universally consistent, i.e.,

$$\lim_{n\to\infty} L(g_n) = L^*$$

a.s. for any distribution of (X, Y).

The rates of convergence of the classification rule g_n, over classes of data-generating mechanisms satisying Hölder continuity and a strong density assumption, were established in [3], and were moreover shown to match a minimax lower bound. We remark that, even in the non-private case, the absence of the strong density assumption leads to slower rates of convergence [22, 1, 7].

6. Proofs and auxiliary results

The proof of Theorem 2 uses two lemmas.

Lemma 1. *For $0 < \varepsilon < 2$ and $\zeta_1, \ldots, \zeta_n$ i.i.d. with mean-zero, unit-variance Laplace distribution, one has*

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\zeta_i\right| \geq \varepsilon\right\} \leq 2e^{-n\varepsilon^2/4}.$$

Proof. Taking $t = n\varepsilon/2$ and using the fact that $\log(1-x) \geq -2x$ for $x \in [0, 1/2]$, we have

$$\begin{aligned}\mathbb{P}\Big(n^{-1}\sum_{i=1}^{n}\zeta_i \geq \varepsilon\Big) &\leq e^{-t\varepsilon}\mathbb{E}[\exp(t\zeta_1/n)]^n \\ &= \exp\Big(-t\varepsilon - n\log\Big(1 - \frac{t^2}{2n^2}\Big)\Big) \\ &\leq \exp\Big(-t\varepsilon + \frac{t^2}{n}\Big) \\ &= e^{-n\varepsilon^2/4}.\end{aligned}$$

An analogous bound holds for the lower tail of the distribution, and the result follows. □

Lemma 2. *Let $Z = (Z_1, \ldots, Z_n)$ be a collection of i.i.d. random variables taking values in some measurable set A. Let $f : A^n \to \mathbb{R}$ be a measurable, symmetric, real-valued function, such that $f(Z_1, \ldots, Z_n)$ is integrable, let $g : A^{n-1} \to \mathbb{R}$ be the function obtained from f by dropping the first argument. Then for any integer $q \geq 1$,*

$$\begin{aligned}&\mathbb{E}\left[(f(Z_1, \ldots, Z_n) - \mathbb{E}f(Z_1, \ldots, Z_n))^{2q}\right] \\ &\leq 2(c^*q)^q n^q \mathbb{E}\left[(f(Z_1, \ldots, Z_n) - g(Z_2, \ldots, Z_n))^{2q}\right] .\end{aligned}$$

with a universal constant $c^* < 5.1$.

Proof. Applying Jensen's inequality, this lemma is a special case of Lemma 4.4 in [10]. □

Proof of Theorem 2. We use the decomposition

$$\tilde{m}_n = m'_n + m^*_n,$$

where for $x \in A_{h_n,j}$ we write

$$m'_n(x) = \frac{\frac{\sigma_Z}{n}\sum_{i=1}^{n}\epsilon_{i,j}}{\tilde{\mu}_n(A_{h_n,j})}\mathbb{I}_{\{\tilde{\mu}_n(A_{h_n,j}) \geq c_n h_n^d\}}\mathbb{I}_{\{j \leq N_n\}},$$

and

$$m_n^*(x) = \frac{\nu_n(A_{h_n,j})}{\tilde{\mu}_n(A_{h_n,j})}\mathbb{I}_{\{\tilde{\mu}_n(A_{h_n,j})\geq c_n h_n^d\}}\mathbb{I}_{\{j\leq N_n\}}.$$

It suffices to show that

$$\lim_{n\to\infty}\int m_n'(x)^2\mu(dx) = 0 \quad \text{a.s.}, \tag{12}$$

and

$$\lim_{n\to\infty}\int \{m(x) - m_n^*(x)\}^2\mu(dx) = 0 \quad \text{a.s.} \tag{13}$$

But (4) implies that

$$\lim_{n\to\infty}\int \{m(x) - m_n(x)\}^2\mu(dx) = 0 \quad \text{a.s.}, \tag{14}$$

for any distribution of (X, Y) with $\mathbb{E}(Y^2) < \infty$, and in order to prove (13) it therefore suffices to show that

$$\lim_{n\to\infty}\int \{m_n(x) - m_n^*(x)\}^2\mu(dx) = 0 \quad \text{a.s.}, \tag{15}$$

for any distribution of (X, Y) with $\mathbb{E}(Y^2) < \infty$.

Proof of (12). Because of

$$\begin{aligned}\int m_n'(x)^2\mu(dx) &= \sum_j \frac{\left(\frac{\sigma_Z}{n}\sum_{i=1}^n \epsilon_{i,j}\right)^2}{\tilde{\mu}_n(A_{h_n,j})^2}\mathbb{I}_{\{\tilde{\mu}_n(A_{h_n,j})\geq c_n h_n^d\}}\mathbb{I}_{\{j\leq N_n\}}\mu(A_{h_n,j}) \\ &\leq \sigma_Z^2\sum_j \frac{\left(\frac{1}{n}\sum_{i=1}^n \epsilon_{i,j}\right)^2}{c_n^2 h_n^{2d}}\mu(A_{h_n,j}),\end{aligned}$$

it suffices to show that

$$\lim_{n\to\infty}\sigma_Z^2\sum_j \frac{\left(\frac{1}{n}\sum_{i=1}^n \epsilon_{i,j}\right)^2}{c_n^2 h_n^{2d}}\mu(A_{h_n,j}) = 0 \quad \text{a.s.} \tag{16}$$

We note

$$\mathbb{E}\{\epsilon_{1,1}^{2q}\} = 2^{-q}(2q)! \leq 2^{-q}(2q)^{2q}e^{-2q/3} = 2^q q^{2q}e^{-2q/3},$$

which together with Lemma 2 implies

$$\mathbb{E}\left\{\left(\sum_{i=1}^n \epsilon_{i,1}\right)^{2q}\right\} \leq 2(c^*q)^q n^q \mathbb{E}\{\epsilon_{1,1}^{2q}\} \leq 2^{q+1}c^{*q}q^{3q}e^{-2q/3}n^q.$$

Jensen's inequality yields

$$\begin{aligned}
&\mathbb{P}\left\{\sum_j \frac{\left(\frac{1}{n}\sum_{i=1}^n \epsilon_{i,j}\right)^2}{c_n^2 h_n^{2d}}\mu(A_{h_n,j}) > \varepsilon\right\} \\
&= \mathbb{P}\left\{\left(\sum_j \frac{\left(\frac{1}{n}\sum_{i=1}^n \epsilon_{i,j}\right)^2}{c_n^2 h_n^{2d}}\mu(A_{h_n,j})\right)^q > \varepsilon^q\right\} \\
&\le \mathbb{P}\left\{\sum_j \frac{\left(\frac{1}{n}\sum_{i=1}^n \epsilon_{i,j}\right)^{2q}}{c_n^{2q} h_n^{2qd}}\mu(A_{h_n,j}) > \varepsilon^q\right\} \\
&\le \varepsilon^{-q}\frac{\mathbb{E}\left\{\left(\frac{1}{n}\sum_{i=1}^n \epsilon_{i,1}\right)^{2q}\right\}}{c_n^{2q} h_n^{2qd}}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{P}\left\{\sum_j \frac{\left(\frac{1}{n}\sum_{i=1}^n \epsilon_{i,j}\right)^2}{c_n^2 h_n^{2d}}\mu(A_{h_n,j}) > \varepsilon\right\} &\le \frac{\varepsilon^{-q}}{c_n^{2q} h_n^{2qd} n^{2q}}\mathbb{E}\left\{\left(\sum_{i=1}^n \epsilon_{i,1}\right)^{2q}\right\} \\
&\le 2\frac{\varepsilon^{-q} 2^q c^{*q} q^{3q} e^{-2q/3}}{c_n^{2q} h_n^{2qd} n^q} \\
&= 2\left(\frac{q^3}{n c_n^2 h_n^{2d}\varepsilon/(2c^* e^{-2/3})}\right)^q.
\end{aligned}$$

Choose

$$q := \lfloor (n c_n^2 h_n^{2d}\varepsilon/(2c^* e^{1/3}))^{1/3}\rfloor.$$

Then

$$\mathbb{P}\left\{\sum_j \frac{\left(\frac{1}{n}\sum_{i=1}^n \epsilon_{i,j}\right)^2}{c_n^2 h_n^{2d}}\mu(A_{h_n,j}) > \varepsilon\right\} \le 2e^{-(n c_n^2 h_n^{2d}\varepsilon)^{1/3}/(2c^* e^{1/3})^{1/3}+1}.$$

Condition (8) yields

$$\sum_n \mathbb{P}\left\{\sum_j \frac{\left(\frac{1}{n}\sum_{i=1}^n \epsilon_{i,j}\right)^2}{c_n^2 h_n^{2d}}\mu(A_{h_n,j}) > \varepsilon\right\} < \infty$$

and thus the Borel-Cantelli lemma results in (16).

Proof of (15). If in the definition of m_n we modify ν_n such that

$$\nu_n(A_{h,j}) = \frac{1}{n}\sum_{i=1}^n [Y_i]_{-M_n}^{M_n}\mathbb{I}_{\{X_i\in A_{h,j}\}},$$

then a slight modification of the proof of Theorem 23.3 in [19] together with the condition $M_n \to \infty$ implies (4), too. We have that

$$\begin{aligned}
&\int \{m_n(x) - m_n^*(x)\}^2 \mu(dx) \\
&= \sum_{j=1}^{N_n} \left\{ \frac{\nu_n(A_{h_n,j})}{\mu_n(A_{h_n,j})} \mathbb{I}_{\{\mu_n(A_{h_n,j}) \geq \log n/n\}} - \frac{\nu_n(A_{h_n,j})}{\tilde{\mu}_n(A_{h_n,j})} \mathbb{I}_{D_n(A_{h_n,j})} \right\}^2 \mu(A_{h_n,j}) \\
&+ \sum_{j=N_n+1}^{\infty} \left\{ \frac{\nu_n(A_{h_n,j})}{\mu_n(A_{h_n,j})} \mathbb{I}_{\{\mu_n(A_{h_n,j}) \geq \log n/n\}} \right\}^2 \mu(A_{h_n,j}) \\
&\leq \sum_{j} \left\{ \frac{\nu_n(A_{h_n,j})}{\mu_n(A_{h_n,j})} \mathbb{I}_{\{\mu_n(A_{h_n,j}) \geq \log n/n\}} - \frac{\nu_n(A_{h_n,j})}{\tilde{\mu}_n(A_{h_n,j})} \mathbb{I}_{D_n(A_{h_n,j})} \right\}^2 \mu(A_{h_n,j}) \\
&+ \int_{S_n^c} m_n(x)^2 \mu(dx),
\end{aligned}$$

where

$$D_n(A_{h_n,j}) = \{\tilde{\mu}_n(A_{h_n,j}) \geq c_n h_n^d\} = \{\mu_n(A_{h_n,j}) + \tau_n(A_{h_n,j}) \geq c_n h_n^d\}$$

with $\tau_n(A_{h_n,j}) = \frac{\sigma_W}{n} \sum_{i=1}^n \zeta_{i,j}$. Since we have $\int m(x)^2 \mu(dx) < \infty$, then (4) together with $S_n \uparrow \mathbb{R}^d$ yields that

$$\int_{S_n^c} m_n(x)^2 \mu(dx) \to 0$$

a.s. Now (8) implies that

$$c_n h_n^d \geq \log n/n$$

if n is large enough. For such large n, set

$$\begin{aligned}
E_n &:= \sum_{j} \left\{ \frac{\nu_n(A_{h_n,j})}{\mu_n(A_{h_n,j})} \mathbb{I}_{\{\mu_n(A_{h_n,j}) \geq \log n/n\}} - \frac{\nu_n(A_{h_n,j})}{\tilde{\mu}_n(A_{h_n,j})} \mathbb{I}_{D_n(A_{h_n,j})} \right\}^2 \mu(A_{h_n,j}) \\
&= \sum_{j} \frac{\nu_n(A_{h_n,j})^2}{\mu_n(A_{h_n,j})^2} \mathbb{I}_{\{\mu_n(A_{h_n,j}) \geq \log n/n\}} \left\{ 1 - \frac{\mu_n(A_{h_n,j})}{\tilde{\mu}_n(A_{h_n,j})} \mathbb{I}_{D_n(A_{h_n,j})} \right\}^2 \mu(A_{h_n,j}).
\end{aligned}$$

Note that

$$\begin{aligned}
\left| 1 - \frac{\mu_n(A_{h_n,j})}{\tilde{\mu}_n(A_{h_n,j})} \mathbb{I}_{D_n(A_{h_n,j})} \right| &= \left| 1 - \frac{\mu_n(A_{h_n,j})}{\tilde{\mu}_n(A_{h_n,j})} \right| \mathbb{I}_{D_n(A_{h_n,j})} + \mathbb{I}_{D_n(A_{h_n,j})^c} \\
&= \frac{|\tau_n(A_{h_n,j})|}{\tilde{\mu}_n(A_{h_n,j})} \mathbb{I}_{D_n(A_{h_n,j})} + \mathbb{I}_{D_n(A_{h_n,j})^c} \\
&= \frac{|\tau_n(A_{h_n,j})|}{\tilde{\mu}_n(A_{h_n,j})} \mathbb{I}_{\{\tilde{\mu}_n(A_{h_n,j}) \geq c_n h_n^d\}} + \mathbb{I}_{\{\tilde{\mu}_n(A_{h_n,j}) < c_n h_n^d\}}.
\end{aligned}$$

Let $A_n(x)$ denote the cube $A_{h_n,j}$, which contains x. Then,

$$\begin{aligned}
E_n &= \int m_n(x)^2 \frac{\tau_n(A_n(x))^2}{\tilde{\mu}_n(A_n(x))^2} \mathbb{I}_{D_n(A_n(x))} \mu(dx) \\
&\qquad + \int m_n(x)^2 \mathbb{I}_{\{\tilde{\mu}_n(A_n(x)) < c_n h_n^d\}} \mu(dx) \\
&\leq \int m_n(x)^2 \frac{\tau_n(A_n(x))^2}{\tilde{\mu}_n(A_n(x))^2} \mathbb{I}_{D_n(A_n(x))} \mu(dx) \\
&\qquad + 2 \int \{m_n(x) - m(x)\}^2 \mu(dx) \\
&\qquad + 2 \int m(x)^2 \mathbb{I}_{\{\tilde{\mu}_n(A_n(x)) < c_n h_n^d\}} \mu(dx) \\
&=: F_n + G_n + H_n.
\end{aligned}$$

Define the notation

$$\mu^*(A) := \int_A m(x)^2 \mu(dx)$$

and

$$\mu_n^*(A) := \int_A m_n(x)^2 \mu(dx).$$

Since $\mathbb{E}(Y^2) < \infty$ we have $\int m(x)^2 \mu(dx) < \infty$ and hence we also have

$$\int m_n(x)^2 \mu(dx) \leq \int m(x)^2 \mu(dx) + o(1) < \infty,$$

so that μ^* and μ_n^* are bounded measures. Thus, a very similar argument to that used to prove (16) shows that

$$\lim_{n\to\infty} \sum_j \frac{\tau_n(A_{h,j})^4}{c_n^4 h_n^{4d}} \mu_n^*(A_{h,j}) = 0 \quad \text{a.s.} \tag{17}$$

where we use the fact that $\{\epsilon_{i,j}\}$, $\{\zeta_{i,j}\}$, $\{(X_i, Y_i)\}$ are independent. Then the Cauchy-Schwarz inequality, (14) and (17) imply

$$\begin{aligned}
F_n &= \int \frac{\tau_n(A_n(x))^2}{\tilde{\mu}_n(A_n(x))^2} \mathbb{I}_{\{\tilde{\mu}_n(A_n(x)) \geq c_n h_n^d\}} \mu_n^*(dx) \\
&= \sum_j \frac{\tau_n(A_{h_n,j})^2}{\tilde{\mu}_n(A_{h_n,j})^2} \mathbb{I}_{\{\tilde{\mu}_n(A_{h_n,j}) \geq c_n h_n^d\}} \mu_n^*(A_{h_n,j}) \\
&\leq \sum_j \frac{\tau_n(A_{h_n,j})^2}{c_n^2 h_n^{2d}} \mu_n^*(A_{h_n,j}) \\
&\leq \sqrt{\sum_j \frac{\tau_n(A_{h_n,j})^4}{c_n^4 h_n^{4d}} \mu_n^*(A_{h_n,j})} \sqrt{\int m_n(x)^2 \mu(dx)} \\
&\to 0 \qquad \text{a.s.}
\end{aligned}$$

The fact that $G_n \to 0$ a.s. follows from (14). We now turn to H_n. Since we have $\int m(x)^2\mu(dx) < \infty$, it suffices to show that

$$\int \mathbb{I}_{\{\tilde{\mu}_n(A_n(x))<c_n h_n^d\}}\mu(dx) \to 0 \qquad \text{a.s.},$$

i.e.,

$$\sum_j \mathbb{I}_{\{\tilde{\mu}_n(A_{h_n,j})<c_n h_n^d\}}\mu(A_{h_n,j}) \to 0 \qquad \text{a.s.}$$

By the inequality

$$\begin{aligned}&\mathbb{I}_{\{\tilde{\mu}_n(A_{h_n,j})<c_n h_n^d\}}\\ &\le \mathbb{I}_{\{|\tau_n(A_{h_n,j})|\ge c_n h_n^d/2\}} + \mathbb{I}_{\{|\mu_n(A_{h_n,j})-\mu(A_{h_n,j})|\ge c_n h_n^d/2\}} + \mathbb{I}_{\{\mu(A_{h,j})<2c_n h_n^d\}},\end{aligned}$$

one gets

$$\begin{aligned}H_n &\le 8\sum_j \frac{\tau_n(A_{h_n,j})^2}{c_n^2 h_n^{2d}}\mu(A_{h_n,j})\\ &\quad + \frac{8}{c_n^2 h_n^{2d}}\sum_j (\mu_n(A_{h_n,j}) - \mu(A_{h_n,j}))^2\mu(A_{h_n,j})\\ &\quad + 2\sum_j \mathbb{I}_{\{\mu(A_{h,j})<2c_n h_n^d\}}\mu(A_{h_n,j}).\end{aligned}$$

(16) implies that the first term tends to 0 a.s. Concerning the second term, we observe that, by the fact that Bernoulli random variables are subgaussian with variance proxy bounded by $1/4$, there exists $L > 0$ such that for any $q \in \mathbb{N}$ we have

$$\mathbb{E}\big[\big(\mu_n(A_{h_n,j}) - \mu(A_{h_n,j})\big)^{2q}\big] \le n^{-q}(Lq^{1/2})^{2q}.$$

Thus, the second term tends to zero a.s. by using a very similar argument to that used to prove (16). Finally, the third term is non-random. Let S be a sphere centred at the origin such that $\mu(S^c) \le \varepsilon$, and set

$$B_n := \bigcup_{j:\mu(A_{h_n,j})<2c_n h_n^d, A_{h_n,j}\cap S\neq\emptyset} A_{h_n,j}.$$

If λ denotes the Lebesgue measure, then

$$\sum_j \mathbb{I}_{\{\mu(A_{h,j})<2c_n h_n^d\}}\mu(A_{h_n,j}) \le \mu(B_n) + \mu(S^c) \le \mu(B_n) + \varepsilon$$

and

$$\mu(B_n) \le \sum_{j:\mu(A_{h_n,j})<2c_n h_n^d, A_{h_n,j}\cap S\neq\emptyset} 2c_n\lambda(A_{h_n,j})$$

$$\leq 2c_n \sum_{j:A_{h_n,j}\cap S\neq\emptyset} \lambda(A_{h_n,j})$$
$$\to 0.$$

Thus, we proved that $H_n \to 0$ a.s. □

Proof of Theorem 3. Since Y is bounded, we may assume that n is sufficiently large that $[Y]_{-M_n}^{M_n} = Y$ almost surely. We use the decomposition

$$\tilde{m}_n = m'_n + m^*_n,$$

where for $x \in A_{h_n,j}$ we recall from the proof of Theorem 2 that

$$m'_n(x) = \frac{\frac{\sigma_Z}{n}\sum_{i=1}^n \epsilon_{i,j}}{\tilde{\mu}_n(A_{h_n,j})} \mathbb{I}_{\{\tilde{\mu}_n(A_{h_n,j})\geq c_n h_n^d\}} \mathbb{I}_{\{j\leq N_n\}},$$

and

$$m^*_n(x) = \frac{\nu_n(A_{h_n,j})}{\tilde{\mu}_n(A_{h_n,j})} \mathbb{I}_{\{\tilde{\mu}_n(A_{h_n,j})\geq c_n h_n^d\}} \mathbb{I}_{\{j\leq N_n\}}.$$

It suffices to show that

$$\int \mathbb{E}m'_n(x)^2 \mu(dx) = O\left(\frac{1}{nc_n^2 h_n^{2d}}\right) \tag{18}$$

and

$$\int \mathbb{E}\{m(x) - m^*_n(x)\}^2 \mu(dx) = O\left(\frac{1}{nc_n^2 h_n^{2d}}\right) + O(h_n^2) \tag{19}$$

But, recalling the definition of $\check{m}_n$ from (3), Theorem 4.3 of [19] implies that

$$\int \mathbb{E}\{m(x) - \check{m}_n(x)\}^2 \mu(dx) = O\left(\frac{1}{nh_n^d}\right) + O(h_n^2), \tag{20}$$

and in order to prove (19) it therefore suffices to show that

$$\int \mathbb{E}\{\check{m}_n(x) - m^*_n(x)\}^2 \mu(dx) = O\left(\frac{1}{nc_n^2 h_n^{2d}}\right). \tag{21}$$

Proof of (18). We have that

$$\int \mathbb{E}m'_n(x)^2 \mu(dx) = \sum_j \mathbb{E}\left\{\frac{\left(\frac{\sigma_Z}{n}\sum_{i=1}^n \epsilon_{i,j}\right)^2}{\tilde{\mu}_n(A_{h_n,j})^2} \mathbb{I}_{\{\tilde{\mu}_n(A_{h_n,j})\geq c_n h_n^d\}} \mathbb{I}_{\{j\leq N_n\}}\right\} \mu(A_{h_n,j})$$
$$\leq \sigma_Z^2 \sum_j \frac{\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \epsilon_{i,j}\right)^2}{c_n^2 h_n^{2d}} \mu(A_{h_n,j})$$
$$= \sigma_Z^2 \frac{1}{nc_n^2 h_n^{2d}}.$$

Proof of (21). We have that

$$\begin{aligned}
&\int \{\check{m}_n(x) - m_n^*(x)\}^2 \mu(dx) \\
&= \sum_{j=1}^{N_n} \left\{ \frac{\nu_n(A_{h_n,j})}{\mu_n(A_{h_n,j})} - \frac{\nu_n(A_{h_n,j})}{\tilde{\mu}_n(A_{h_n,j})} \mathbb{I}_{D_n(A_{h_n,j})} \right\}^2 \mu(A_{h_n,j}) \\
&+ \sum_{j=N_n+1}^{\infty} \left\{ \frac{\nu_n(A_{h_n,j})}{\mu_n(A_{h_n,j})} \right\}^2 \mu(A_{h_n,j}) \\
&\le \sum_{j} \left\{ \frac{\nu_n(A_{h_n,j})}{\mu_n(A_{h_n,j})} - \frac{\nu_n(A_{h_n,j})}{\tilde{\mu}_n(A_{h_n,j})} \mathbb{I}_{D_n(A_{h_n,j})} \right\}^2 \mu(A_{h_n,j}) \\
&+ \int_{S_n^c} \check{m}_n(x)^2 \mu(dx),
\end{aligned}$$

where we recall that

$$D_n(A_{h_n,j}) = \left\{ \tilde{\mu}_n(A_{h_n,j}) \ge c_n h_n^d \right\} = \left\{ \mu_n(A_{h_n,j}) + \tau_n(A_{h_n,j}) \ge c_n h_n^d \right\}$$

with $\tau_n(A_{h_n,j}) = \frac{\sigma_W}{n} \sum_{i=1}^n \zeta_{i,j}$. Since X is bounded, then $S_n \uparrow \mathbb{R}^d$ yields that

$$\int_{S_n^c} \check{m}_n(x)^2 \mu(dx) = 0$$

if n is large enough. Set

$$E_n := \sum_{j} \left\{ \frac{\nu_n(A_{h_n,j})}{\mu_n(A_{h_n,j})} - \frac{\nu_n(A_{h_n,j})}{\tilde{\mu}_n(A_{h_n,j})} \mathbb{I}_{D_n(A_{h_n,j})} \right\}^2 \mu(A_{h_n,j}).$$

Letting L denote a bound of $|Y|$, we have

$$E_n \le \sum_{j} L^2 \left\{ 1 - \frac{\mu_n(A_{h_n,j})}{\tilde{\mu}_n(A_{h_n,j})} \mathbb{I}_{D_n(A_{h_n,j})} \right\}^2 \mu(A_{h_n,j}).$$

Note that

$$\begin{aligned}
\left| 1 - \frac{\mu_n(A_{h_n,j})}{\tilde{\mu}_n(A_{h_n,j})} \mathbb{I}_{D_n(A_{h_n,j})} \right| &= \left| 1 - \frac{\mu_n(A_{h_n,j})}{\tilde{\mu}_n(A_{h_n,j})} \right| \mathbb{I}_{D_n(A_{h_n,j})} + \mathbb{I}_{D_n(A_{h_n,j})^c} \\
&= \frac{|\tau_n(A_{h_n,j})|}{\tilde{\mu}_n(A_{h_n,j})} \mathbb{I}_{D_n(A_{h_n,j})} + \mathbb{I}_{D_n(A_{h_n,j})^c} \\
&= \frac{|\tau_n(A_{h_n,j})|}{\tilde{\mu}_n(A_{h_n,j})} \mathbb{I}_{\{\tilde{\mu}_n(A_{h_n,j}) \ge c_n h_n^d\}} + \mathbb{I}_{\{\tilde{\mu}_n(A_{h_n,j}) < c_n h_n^d\}}.
\end{aligned}$$

Let $A_n(x)$ denote the cube $A_{h_n,j}$, which contains x. Then,

$$E_n \le L^2 \int \frac{\tau_n(A_n(x))^2}{\tilde{\mu}_n(A_n(x))^2} \mathbb{I}_{\{\tilde{\mu}_n(A_{h_n,j}) \ge c_n h_n^d\}} \mu(dx) + L^2 \int \mathbb{I}_{\{\tilde{\mu}_n(A_n(x)) < c_n h_n^d\}} \mu(dx)$$

$$\leq L^2 \int \frac{\tau_n(A_n(x))^2}{c_n^2 h_n^{2d}} \mu(dx) + L^2 \int \mathbb{I}_{\{\tilde{\mu}_n(A_n(x)) < c_n h_n^d\}} \mu(dx).$$

Therefore

$$\begin{aligned}\mathbb{E}\{E_n\} &\leq L^2 \int \frac{\mathbb{E}\{\tau_n(A_n(x))^2\}}{c_n^2 h_n^{2d}} \mu(dx) + L^2 \int \mathbb{P}\{\tilde{\mu}_n(A_n(x)) < c_n h_n^d\} \mu(dx) \\ &=: F_n + G_n.\end{aligned}$$

We have that

$$F_n = L^2 \frac{\sigma_W^2}{n c_n^2 h_n^{2d}} \tag{22}$$

We now turn to G_n. By the inequality

$$\begin{aligned}&\mathbb{I}_{\{\tilde{\mu}_n(A_{h_n,j}) < c_n h_n^d\}} \\ &\leq \mathbb{I}_{\{|\tau_n(A_{h_n,j})| \geq c_n h_n^d/2\}} + \mathbb{I}_{\{|\mu_n(A_{h_n,j}) - \mu(A_{h_n,j})| \geq c_n h_n^d/2\}} + \mathbb{I}_{\{\mu(A_{h,j}) < 2c_n h_n^d\}},\end{aligned}$$

one gets

$$\begin{aligned}G_n &\leq 4L^2 \sum_j \frac{\mathbb{E}\{\tau_n(A_{h_n,j})^2\}}{c_n^2 h_n^{2d}} \mu(A_{h_n,j}) \\ &\quad + \frac{4L^2}{c_n^2 h_n^{2d}} \sum_j \mathbb{E}\{(\mu_n(A_{h_n,j}) - \mu(A_{h_n,j}))^2\} \mu(A_{h_n,j}) \\ &\quad + L^2 \sum_j \mathbb{I}_{\{\mu(A_{h,j}) < 2c_n h_n^d\}} \mu(A_{h_n,j}) \\ &\leq \frac{4L^2 \sigma_W^2}{n c_n^2 h_n^{2d}} + \frac{L^2}{n c_n^2 h_n^{2d}} + L^2 \sum_j \mathbb{I}_{\{\mu(A_{h,j}) < 2c_n h_n^d\}} \mu(A_{h_n,j}).\end{aligned}$$

Finally, the third term is non-random. X has a density, which is bounded away from zero, therefore there exists $f_{min} > 0$ such that $\mu(A_{h_n,j}) > 0$ implies $\mu(A_{h_n,j}) \geq f_{min} h_n^d$. Thus,

$$\sum_j \mathbb{I}_{\{\mu(A_{h,j}) < 2c_n h_n^d\}} \mu(A_{h_n,j}) = \sum_j \mathbb{I}_{\{f_{min} h_n^d \leq \mu(A_{h,j}) < 2c_n h_n^d\}} \mu(A_{h_n,j}) = 0$$

if $f_{min} > 2c_n$. □

Proof of Proposition 1. Fix any $w, z \in \mathbb{R}^{\mathbb{N}}, x, x' \in \mathbb{R}^d, y, y' \in [-M, M]$. We have

$$\begin{aligned}&\frac{f_{W,Z|X,Y}(w, z|x, y)}{f_{W,Z|X,Y}(w, z|x', y')} \\ &= \exp\Big((\sqrt{2}/\sigma_W) \sum_{j=1}^{\infty} (|w_j - \mathbb{1}_{\{x' \in A_{h,j}\}}| - |w_j - \mathbb{1}_{\{x \in A_{h,j}\}}|) \\ &\qquad + (\sqrt{2}/\sigma_Z) \sum_{j=1}^{\infty} (|z_j - y' \mathbb{1}_{\{x' \in A_{h,j}\}}| - |z_j - y \mathbb{1}_{\{x \in A_{h,j}\}}|) \Big).\end{aligned}$$

Now, if there exists $j \in \mathbb{N}$ such that $x, x' \in A_{h,j}$, we have

$$\sum_{j'=1}^{\infty} (|w_{j'} - \mathbb{1}_{\{x' \in A_{h,j'}\}}| - |w_{j'} - \mathbb{1}_{\{x \in A_{h,j'}\}}|) = 0$$

$$\sum_{j'=1}^{\infty} (|z_{j'} - y' \mathbb{1}_{\{x' \in A_{h,j'}\}}| - |z_{j'} - y \mathbb{1}_{\{x \in A_{h,j'}\}}|) = |z_j - y'| - |z_j - y| \le 2M.$$

On the other hand, if $x \in A_{h,j}$ and $x' \in A_{h,j'}$ with $j \neq j'$, then we have

$$\begin{aligned}
&\sum_{j''=1}^{\infty} (|w_{j''} - \mathbb{1}_{\{x' \in A_{h,j''}\}}| - |w_{j''} - \mathbb{1}_{\{x \in A_{h,j''}\}}|) \\
&\quad = |w_j| - |w_j - 1| + |w_{j'} - 1| - |w_{j'}| \le 2, \\
&\sum_{j''=1}^{\infty} (|z_{j''} - y' \mathbb{1}_{\{x' \in A_{h,j''}\}}| - |z_{j''} - y \mathbb{1}_{\{x \in A_{h,j''}\}}|) \\
&\quad = |z_j| - |z_j - y| + |z_{j'} - y'| - |z_{j'}| \le 2M.
\end{aligned}$$

It therefore follows that

$$\frac{f_{W,Z|X,Y}(w, z|x, y)}{f_{W,Z|X,Y}(w, z|x', y')} \le \exp\Big(2^{3/2}/\sigma_W + 2^{3/2} M/\sigma_Z\Big),$$

as required. □

Proof of Theorem 4. For the notation

$$\bar{m}_n(x) = \frac{\tilde{\nu}_n(A_{h_n,j})}{\mu(A_{h_n,j})} \quad \text{when } x \in A_{h_n,j},$$

the rule g_n has the equivalent form

$$g_n(x) = sign\, \bar{m}_n(x).$$

Theorem 2.2 in [9] implies that

$$\begin{aligned}
L(g_n) - L^* &= \int \mathbb{I}_{\{g_n(x) \neq g^*(x)\}} |m(x)| \mu(dx) \\
&= \int \mathbb{I}_{\{sign\, \bar{m}_n(x) \neq sign\, m(x)\}} |m(x)| \mu(dx) \\
&\le \int [|m(x) - \bar{m}_n(x)|]_0^1 \mu(dx).
\end{aligned}$$

Write

$$m_n(x) = \frac{\nu_n(A_{h_n,j})}{\mu(A_{h_n,j})} \quad \text{when } x \in A_{h_n,j}.$$

Then,

$$\begin{aligned}&\int [|m(x) - \bar{m}_n(x)|]_0^1 \mu(dx)\\ &\leq \int |m(x) - m_n(x)|\mu(dx) + \int [|m_n(x) - \bar{m}_n(x)|]_0^1 \mu(dx).\end{aligned}$$

By Theorem 23.1 in [19], the first term tends to 0 a.s. Similarly to the previous proof, given $\epsilon > 0$ let S be a sphere centred at the origin such that $\mu(S^c) \leq \varepsilon$, and set

$$B_n := \bigcup_{j : A_{h_n,j} \cap S \neq \emptyset} A_{h_n,j}.$$

Then,

$$\begin{aligned}\int [|m_n(x) - \bar{m}_n(x)|]_0^1 \mu(dx) &\leq \sum_{j \in B_n} [|\nu_n(A_{h_n,j}) - \tilde{\nu}_n(A_{h_n,j})|]_0^1 + \mu(S^c)\\ &\leq \sum_{j \in B_n} \left[\left|\frac{\sigma_Z}{n}\sum_{i=1}^n \epsilon_{i,j}\right|\right]_0^1 + \varepsilon\\ &\leq \sum_{j \in B_n} \left(\varepsilon h_n^d + \mathbb{I}_{\left[\left|\frac{\sigma_Z}{n}\sum_{i=1}^n \epsilon_{i,j}\right|\right]_0^1 \geq \varepsilon h_n^d}\right) + \varepsilon.\end{aligned}$$

For n sufficiently large,

$$\sum_{j \in B_n} \varepsilon h_n^d \leq 2\lambda(S)\varepsilon$$

and Lemma 1 implies

$$\begin{aligned}\sum_n \mathbb{E}\left\{\sum_{j \in B_n} \mathbb{I}_{\left|\frac{\sigma_Z}{n}\sum_{i=1}^n \epsilon_{i,j}\right| \geq \varepsilon h_n^d}\right\} &= \sum_n |B_n| \mathbb{P}\left\{\left|\frac{\sigma_Z}{n}\sum_{i=1}^n \epsilon_{i,1}\right| \geq \varepsilon h_n^d\right\}\\ &= \sum_n \frac{4\lambda(S)}{h_n^d} e^{-n(\varepsilon h_n^d/\sigma_Z)^2/4}\\ &< \infty,\end{aligned}$$

where the last step follows from the condition $nh_n^{2d}/\log n \to \infty$. Therefore, by Markov's inequality and the Borel-Cantelli lemma, we have proved that

$$\limsup_n \int [|m_n(x) - \bar{m}_n(x)|]_0^1 \mu(dx) \leq 2\lambda(S)\varepsilon + \varepsilon$$

a.s. Since $\epsilon > 0$ was arbitrary, this completes the proof. □

References

[1] Audibert, J.-Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *The Annals of Statistics* **35** 608–633. MR2336861

[2] Avella-Medina, M. and Brunel, V.-E. (2019). Differentially private sub-Gaussian location estimators. *arXiv preprint arXiv:1906.11923*.

[3] Berrett, T. B. and Butucea, C. (2019). Classification under local differential privacy. *Annales de l'ISUP* **63–80 ans de Denis Bosq** 191–205.

[4] Berrett, T. B. and Butucea, C. (2020). Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms. In *Advances in Neural Information Processing Systems* **33** 3164–3173.

[5] Butucea, C., Dubois, A., Kroll, M. and Saumard, A. (2020). Local differential privacy: Elbow effect in optimal density estimation and adaptation over Besov ellipsoids. *Bernoulli* **26** 1727–1764. MR4091090

[6] Cai, T. T., Wang, Y. and Zhang, L. (2019). The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *arXiv preprint arXiv:1902.04495*.

[7] Cannings, T. I., Berrett, T. B. and Samworth, R. J. (2020). Local nearest neighbour classification with applications to semi-supervised learning. *The Annals of Statistics* **48** 1789–1814. MR4124344

[8] Dankar, F. K. and El Emam, K. (2013). Practicing differential privacy in health care: A review. *Transactions on Data Privacy* **6** 35–67. MR3067642

[9] Devroye, L., Györfi, L. and Lugosi, G. (2013). *A Probabilistic Theory of Pattern Recognition* **31**. Springer Science & Business Media. MR1383093

[10] Devroye, L., Györfi, L., Lugosi, G. and Walk, H. (2018). A nearest neighbor estimate of the residual variance. *Electronic Journal of Statistics* **12** 1752–1778. MR3811758

[11] Ding, B., Kulkarni, J. and Yekhanin, S. (2017). Collecting telemetry data privately. In *Advances in Neural Information Processing Systems* **30** 3571–3580.

[12] Duchi, J. C., Jordan, M. I. and Wainwright, M. J. (2018). Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association* **113** 182–201. MR3803452

[13] Duchi, J. C. and Ruan, F. (2018). The right complexity measure in locally private estimation: It is not the Fisher information. *arXiv preprint arXiv:1806.05756*.

[14] Dwork, C. (2019). Differential privacy and the US census. In *Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems* 1–1.

[15] Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference* 265–284. Springer. MR2241676

[16] Erlingsson, Ú., Pihur, V. and Korolova, A. (2014). Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* 1054–1067.

[17] Farokhi, F. (2020). Deconvoluting kernel density estimation and regression for locally differentially private data. *arXiv preprint arXiv:2008.12466*.
[18] Györfi, L. (1991). Universal consistencies of a regression estimate for unbounded regression functions. In *Nonparametric Functional Estimation and Related Topics* 329–338. Springer. MR1154338
[19] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2006). *A Distribution-Free Theory of Nonparametric Regression*. Springer Science & Business Media.
[20] Joseph, M., Mao, J., Neel, S. and Roth, A. (2019). The role of interactivity in local differential privacy. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)* 94–105. IEEE.
[21] Kairouz, P., Oh, S. and Viswanath, P. (2014). Extremal Mechanisms for Local Differential Privacy. In *Advances in Neural Information Processing Systems* **27** 2879–2887. MR3491111
[22] Kohler, M. and Krzyżak, A. (2007). On the rate of convergence of local averaging plug-in classification rules under a margin condition. *IEEE Transactions on Information Theory* **53** 1735–1742. MR2317135
[23] Lei, J. (2011). Differentially private M-estimators. In *Advances in Neural Information Processing Systems* **24** 361–369.
[24] Lugosi, G. and Nobel, A. B. (1999). Adaptive model selection using empirical complexities. *The Annals of Statistics* **27** 1830–1864. MR1765619
[25] Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J. and Vilhuber, L. (2008). Privacy: Theory meets practice on the map. In *2008 IEEE 24th International Conference on Data Engineering* 277–286. IEEE.
[26] Rocher, L., Hendrickx, J. M. and De Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications* **10** 1–9.
[27] Rohde, A. and Steinberger, L. (2029). Geometrizing rates of convergence under differential privacy constraints. *The Annals of Statistics* **48** 2646–2670. MR4152116
[28] Smith, A., Thakurta, A. and Upadhyay, J. (2017). Is interaction necessary for distributed private learning? In *2017 IEEE Symposium on Security and Privacy (SP)* 58–77. IEEE.
[29] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **10** 557–570. MR1948199
[30] Tang, J., Korolova, A., Bai, X., Wang, X. and Wang, X. (2017). Privacy loss in Apple's implementation of differential privacy on MacOS 10.12. *arXiv preprint arXiv:1709.02753*. MR3549464
[31] Tukey, J. W. (1947). Non-parametric estimation II: Statistically equivalent blocks and tolerance regions – the continuous case. *The Annals of Mathematical Statistics* 529–539. MR0023033
[32] Vu, D. and Slavkovic, A. (2009). Differential privacy for clinical trial data: Preliminary evaluations. In *2009 IEEE International Conference on Data Mining Workshops* 138–143. IEEE.
[33] Wang, D., Gaboardi, M. and Xu, J. (2018). Empirical risk minimization

in non-interactive local differential privacy revisited. In *Advances in Neural Information Processing Systems* **31** 965–974.

[34] Wang, D., Smith, A. and Xu, J. (2018). High dimensional sparse linear regression under local differential privacy: Power and limitations. In *2018 NIPS Workshop in Privacy-Preserving Machine Learning* **235**.

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

[36] Wasserman, L. and Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association* **105** 375–389. MR2656057

[37] Zheng, K., Mou, W. and Wang, L. (2017). Collect at once, use effectively: Making non-interactive locally private learning possible. *arXiv preprint arXiv:1706.03316*.