

On the Consistency of the Kozachenko-Leonenko Entropy Estimate

Luc Devroye and László Györfi 

Abstract—We revisit the problem of the estimation of the differential entropy $H(f)$ of a random vector X in $\mathbb{R}^d$ with density f , assuming that $H(f)$ exists and is finite. In this note, we study the consistency of the popular nearest neighbor estimate H_n of Kozachenko and Leonenko. Without any smoothness condition we show that the estimate is consistent ($\mathbb{E}\{|H_n - H(f)|\} \rightarrow 0$ as $n \rightarrow \infty$) if and only if $\mathbb{E}\{\log(\|X\| + 1)\} < \infty$. Furthermore, if X has compact support, then $H_n \rightarrow H(f)$ almost surely.

Index Terms—Differential entropy estimate, consistency conditions.

I. INTRODUCTION

THE differential entropy of a random $\mathbb{R}^d$ -valued vector X with probability density function f is

$$H(f) = - \int f(x) \ln f(x) dx = -\mathbb{E}\{\ln f(X)\} \quad (1)$$

when this integral exists.

The objective of this paper is to study an estimate of (1) based on independent and identically distributed samples $X_1, \dots, X_n$, with density f .

Estimation of the differential entropy has a long history. The early part of that story was described by Beirlant *et al.* [2]. Since then, there has been considerable activity on the topic. There are estimates based on kernel methods (Joe [21]; Györfi and van der Meulen [17]; Hall and Morton [18]; Shwartz *et al.* [28]; Paninski and Yajima [27]; Krishnamurthy *et al.* [24]; Kandasamy *et al.* [22]), on wavelet functions, (Donoho *et al.* [10]; Delyon and Juditsky [8], and Chesneau *et al.* [6]), on partitioning methods (Stowell and Plumbley [34]), and on nearest neighbour strategies (Kozachenko and Leonenko [23]; Tsybakov and Van der Meulen [35]; Singh *et al.* [30]; Sricharan, Raich, and Hero [32]; Sricharan, Wei, and Hero [33]; Singh and Póczos [31]; Gao, Oh, and Viswanath [13] and [12]; Delattre and Fournier [7]; Lord, Sun, and Boltt [25]; Berrett, Samworth, and Yuan [3]). Other related work deals with general properties

of functional estimation (see, e.g., Birgé and Massart [5]) or sufficient conditions of consistency of the so-called integral estimate $\int f_n \log(1/f_n)$ when f_n is a general density estimate (see, e.g., Godavarti [14]).

There has been particular interest in minimax rates of convergence, culminating in the paper by Han, Jiao, Weissman and Wu [19] who obtained the minimax rates for classes of densities on the unit cube of $\mathbb{R}^d$ that are Besov balls or Hölder (Lipschitz) balls of smoothness parameters s . Their rates are of exact order $\max(1/(n \log n)^{s/(s+d)}, 1/\sqrt{n})$. For standard Lipschitz densities, the minimax rate is of exact order $1/\sqrt{n}$ (achieving the parametric rate) for dimension one, but it is $1/(n \log n)^{-1/(d+1)}$ when $d > 1$. They also construct a kernel-based estimate that is minimax optimal for these classes.

Most of the previous work assumes compact support, and thus sidesteps the thorny issue of infinite tails. The objective of this note is merely to give a complete characterization of the consistency of the popular Kozachenko-Leonenko estimate [23] for estimating the differential entropy. Without any smoothness condition on the density f we study three types of consistency: strong, L_1 and weak.

It is an open research problem, whether there exists an entropy estimate such that it is consistent under the only condition that $H(f)$ is finite. The most obvious way of estimating the differential entropy is the partitioning-based estimate. If the corresponding partition is deterministic, then we conjecture that there exist no deterministic partitions such that the relating differential entropy estimate is a.s. consistent for any finite differential entropy. The significant breakthrough in this respect is due to Wang, Kulkarni, Verdu [36] and Silva, Narayanan [29]. They suggested data-dependent partitioning, for which the partitioning-based estimates of Kullback-Leibler (KL) divergence and of mutual information are strongly consistent under the only condition that the KL-divergence and the mutual information are finite, respectively. We guess that a universally consistent entropy estimate can be derived from data dependent partitioning.

For the cross-validation estimate or leave-one-out entropy estimate, $f_{n,i}$ denotes a density estimate based on $X_1, \dots, X_n$ leaving X_i out, and the corresponding entropy estimate is of the form

$$H_n = -\frac{1}{n} \sum_{i=1}^n \ln f_{n,i}(X_i). \quad (2)$$

Kozachenko and Leonenko [23] introduced the nearest neighbor entropy estimate as follows. Let $R_{n,i}(x)$, $x \in \mathbb{R}^d$,

Manuscript received February 25, 2021; accepted November 8, 2021. Date of publication November 12, 2021; date of current version January 20, 2022. The work of Luc Devroye was supported by an National Science and Engineering Research Council of Canada (NSERC) Discovery Grant. (Corresponding author: László Györfi.)

Luc Devroye is with the School of Computer Science, McGill University, Montreal, QC H3A 0G4, Canada (e-mail: lucdevroye@gmail.com).

László Györfi is with the Department of Computer Science and Information Theory, Budapest University of Technology and Economics, 1111 Budapest, Hungary (e-mail: gyorfi@cs.bme.hu).

Communicated by S. Boucheron, Associate Editor for Statistical Learning. Digital Object Identifier 10.1109/TIT.2021.3127938

be defined by

$$R_{n,i}(x) = \min_{j \neq i, j \leq n} \|x - X_j\|,$$

where $\|\cdot\|$ denotes the Euclidean norm. Then the nearest neighbor entropy estimate is

$$H_n = \frac{1}{n} \sum_{i=1}^n \ln((n-1)R_{n,i}(X_i)^d v_d) + C_E, \quad (3)$$

where $C_E = -\int_0^\infty e^{-t} \ln t dt = 0.5772\dots$ is the Euler-Mascheroni constant and v_d denotes the volume of the unit sphere in $\mathbb{R}^d$. The estimate in (3) has the form of (2) if $f_{n,i}$ is a constant multiple of the first-nearest-neighbor (1-NN) density estimate:

$$f_{n,i}(x) = \frac{1}{(n-1) \min_{j \neq i, j \leq n} \|x - X_j\|^d v_d e^{C_E}}.$$

Notice that this particular $f_{n,i}$ is not consistent in L_1 , because

$$\int f_{n,i}(x) dx = \infty.$$

Furthermore, $f_{n,i}(x)$ is unbounded at the $X_j, j \neq i$. Also, $f_{n,i}(x)$ does not in general tend to $f(x)$ in probability, i.e., the density estimates $f_{n,i}$ are not weakly consistent.

Under some mild conditions on the density f , Kozachenko and Leonenko [23] proved the mean square consistency. Biau and Devroye [4] showed that for bounded X , if $\int f(x) \ln^2(f(x) + 1) dx < \infty$, then $H_n \rightarrow H(f)$ in probability.

For smooth densities, Berrett, Samworth and Yuan [3], Delattre and Fournier [7] and Tsybakov and van der Meulen [35] studied the rate of convergence of H_n and of its extensions to many nearest neighbors. Jiao, Gao, and Han [20] establish the upper bound

$$\mathbb{E}\{|H_n - H(f)|\} = O\left(\frac{\ln n}{n^{1/(d+1)}} + \frac{1}{\sqrt{n}}\right)$$

for the Kozachenko-Leonenko estimate H_n (see (3) above), uniformly over all Lipschitz densities f with Lipschitz constant $\leq L$ and support in $[0, 1]^d$. This is within a polylogarithmic factor of the minimax lower bound of Han *et al.* [19]. We believe that this rate is tight.

In this paper we show that for bounded X the Kozachenko-Leonenko estimate is strongly consistent. Furthermore, we give a necessary and sufficient tail condition for L_1 consistency. In addition, we construct a counterexample for the weak consistency for a uniform density on an unbounded set. The proofs are presented in the last section.

II. CONSISTENCY RESULTS

For bounded X , without any smoothness condition on the density f the Kozachenko-Leonenko estimate is strongly consistent:

Theorem 1: If the support of f is bounded and

$$\int f(x) \ln(f(x) + 1) dx < \infty, \quad (4)$$

then

$$\lim_{n \rightarrow \infty} H_n = H(f) \quad (5)$$

a.s.

Next, we present a necessary and sufficient tail condition on the consistency in L_1 :

Theorem 2: Assume that $H(f)$ is finite. For any density f ,

$$\lim_{n \rightarrow \infty} \mathbb{E}\{|H_n - H(f)|\} = 0 \quad (6)$$

if and only if

$$\mathbb{E}\{(\ln \|X\|)^+\} < \infty. \quad (7)$$

If $\mathbb{E}\{(\ln \|X\|)^+\} = \infty$, then maybe the expectations $\mathbb{E}\{\ln \|X\|\}$ and $\mathbb{E}\{\ln R_{n,1}(X_1)\}$ don't exist. However, for finite $H(f)$, we show that the expectation $\mathbb{E}\{H_n\}$ is well defined such that it is larger than $-\infty$.

As sufficient condition, (7) appeared in the studies of distribution and density estimates consistent in KL-divergence. For discrete distributions concentrated to the set of positive integers, Györfi, Páli and van der Meulen [16] proved that a distribution with finite Shannon entropy cannot be estimated consistently in KL-divergence. It means that for any distribution estimate $p_n = (p_{n,1}, p_{n,2}, \dots)$ there exist a distribution $p = (p_1, p_2, \dots)$ with finite Shannon entropy such that for the KL-divergence

$$KL(p, p_n) = \sum_{i=1}^{\infty} p_i \ln \frac{p_i}{p_{n,i}} = \infty$$

for all n a.s. However, under (7) one can construct a distribution estimate p_n such that $KL(p, p_n) \rightarrow 0$ a.s. This positive finding has been generalized to density estimation consistent in KL-divergence. For the condition (7), one can create a density g of power tail such that the mixture of the ordinary histogram and of this g is consistent in KL-divergence, see Barron, Györfi and van der Meulen [1].

Theorem 2 is a complete characterization of the consistency in L_1 . In fact, we show that there are two cases:

- either $\mathbb{E}\{(\ln \|X\|)^+\} = \infty$ and then $\mathbb{E}\{H_n\} = \infty$, for all n ,
- or $\mathbb{E}\{(\ln \|X\|)^+\} < \infty$ and then $\lim_{n \rightarrow \infty} \mathbb{E}\{|H_n - H(f)|\} = 0$.

Remark: In the sequel, we show an example, where the density is uniform on an unbounded set and $H_n \rightarrow \infty$ in probability.

Denote by λ the Lebesgue measure and by μ the distribution of X . For $d = 1$, let

$$f = \mathbb{I}_A,$$

where $A = \cup_j A_j$ with disjoint intervals A_j such that $\lambda(A) = 1$. Then, X is uniformly distributed on A and therefore $H(f) = 0$. For $j \geq 1$, let $\Delta_j = \frac{1}{j(j+1)}$ and $a_j = 2^{2^j}$. Set $A = \cup_{j \geq 1} [a_j, a_j + \Delta_j]$. Then $\lambda(A) = 1$. Note that in this example $\mathbb{E}\{(\ln \|X\|)^+\} = \infty$, and therefore $\mathbb{E}\{H_n\} = \infty$, for all n . We will show:

Fact 1: In this setup, $\lim_{n \rightarrow \infty} H_n = \infty$ in probability.

III. PROOFS

Proof of Theorem 1: Put

$$\bar{f}_h(x) = \frac{\mu(B(x, h))}{\lambda(B(x, h))} = \frac{\int_{B(x, h)} f(z) dz}{\lambda(B(x, h))},$$

where $B(x, h)$ stands for the sphere centered at x and having the radius r . Then,

$$\begin{aligned} H_n - H(f) &= \frac{1}{n} \sum_{i=1}^n \ln((n-1)\lambda(B(X_i, R_{n,i}(X_i)))) + C_E - H(f) \\ &= -\frac{1}{n} \sum_{i=1}^n \ln \frac{\mu(B(X_i, R_{n,i}(X_i)))}{\lambda(B(X_i, R_{n,i}(X_i)))} - H(f) \\ &+ \frac{1}{n} \sum_{i=1}^n \ln((n-1)\mu(B(X_i, R_{n,i}(X_i)))) + C_E \\ &= \tilde{H}_n - H(f) + M_n + C_E, \end{aligned}$$

where

$$\tilde{H}_n = -\frac{1}{n} \sum_{i=1}^n \ln \bar{f}_{R_{n,i}(X_i)}(X_i),$$

and

$$M_n = \frac{1}{n} \sum_{i=1}^n \ln((n-1)\mu(B(X_i, R_{n,i}(X_i)))).$$

Biau and Devroye [4] showed that the distribution of M_n does not depend on the density f , and $\mathbb{E}\{M_n\} = -C_E + O(1/n)$ and $\text{Var}(M_n) = O(1/n)$. The problem left is to show

$$M_n - \mathbb{E}\{M_n\} \rightarrow 0 \quad (8)$$

a.s. and

$$\tilde{H}_n \rightarrow H(f) \quad (9)$$

a.s.

The proof of (8) relies on the following extension of the Efron-Stein inequality for the centered higher moments:

Lemma 1: (Devroye *et al.* [9]) Let $Z = (Z_1, \dots, Z_n)$ be a collection of independent random variables taking values in some measurable set A and denote by $Z^{(i)} = (Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_n)$ the collection with the i -th random variable dropped. Let $f : A^n \rightarrow \mathbb{R}$ be a measurable real-valued function and the function $g_i : A^{n-1} \rightarrow \mathbb{R}$ is obtained from f by dropping the i -th argument, $i = 1, \dots, n$. Then for any integer $q \geq 1$,

$$\begin{aligned} &\mathbb{E}[(f(Z) - \mathbb{E}f(Z))^{2q}] \\ &\leq (cq)^q \mathbb{E} \left[\left(\sum_{i=1}^n (f(Z) - g_i(Z^{(i)}))^2 \right)^q \right] \\ &\quad + (cq)^q \mathbb{E} \left[\left(\sum_{i=1}^n \mathbb{E} \left[(f(Z) - g_i(Z^{(i)}))^2 \mid Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_n \right] \right)^q \right], \end{aligned} \quad (10)$$

with a universal constant $c < 5.1$.

For the term $M_n - \mathbb{E}\{M_n\}$, define $M_n^{(i)}$ as M_n without the i -th term. We apply Lemma 1 with $q = 2$:

$$\begin{aligned} &\mathbb{E}[(M_n - \mathbb{E}\{M_n\})^4] \\ &\leq (c2)^2 \mathbb{E} \left[\left(\sum_{i=1}^n (M_n - M_n^{(i)})^2 \right)^2 \right] \\ &\quad + (c2)^2 \mathbb{E} \left[\left(\sum_{i=1}^n \mathbb{E} \left[(M_n - M_n^{(i)})^2 \mid X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n \right] \right)^2 \right] \\ &\leq 2^3 c^2 n^2 \mathbb{E} \left[(M_n - M_n^{(n)})^4 \right] \\ &= 2^3 c^2 n^2 \mathbb{E} \left[\left(M_n - \frac{n-1}{n} M_{n-1} \right)^4 \right]. \end{aligned}$$

Noting that

$$\begin{aligned} &\left(M_n - \frac{n-1}{n} M_{n-1} \right)^4 \\ &= \frac{1}{n^4} \left(\sum_{i=1}^n \ln((n-1)\mu(B(X_i, R_{n,i}(X_i)))) \right. \\ &\quad \left. - \sum_{i=1}^{n-1} \ln((n-2)\mu(B(X_i, R_{n-1,i}(X_i)))) \right)^4 \\ &= \frac{1}{n^4} \left(-\ln((n-1)\mu(B(X_n, R_{n,n}(X_n)))) \right. \\ &\quad \left. + (n-1) \ln \frac{n-2}{n-1} \right. \\ &\quad \left. + \sum_{i=1}^{n-1} \ln \frac{\mu(B(X_i, R_{n-1,i}(X_i)))}{\mu(B(X_i, R_{n,i}(X_i)))} \mathbb{I}_{R_{n,i}(X_i) < R_{n-1,i}(X_i)} \right)^4, \end{aligned}$$

the c_r -inequality and Jensen's inequality imply that

$$\begin{aligned} &\mathbb{E}[(M_n - \mathbb{E}\{M_n\})^4] \\ &\leq \frac{2^3 3^3 c^2}{n^2} \left(\mathbb{E} \left[(\ln((n-1)\mu(B(X_n, R_{n,n}(X_n))))^4 \right] \right. \\ &\quad \left. + 1 + \mathbb{E} \left[\left(\sum_{i=1}^{n-1} \ln \frac{\mu(B(X_i, R_{n-1,i}(X_i)))}{\mu(B(X_i, R_{n,i}(X_i)))} \right. \right. \right. \\ &\quad \left. \left. \cdot \mathbb{I}_{R_{n,i}(X_i) < R_{n-1,i}(X_i)} \right)^4 \right] \right) \\ &\leq \frac{6^3 c^2}{n^2} \left(\mathbb{E} \left[(\ln((n-1)\mu(B(X_n, R_{n,n}(X_n))))^4 \right] + 1 \right. \\ &\quad \left. + \mathbb{E} \left[\left(\sum_{i=1}^{n-1} \mathbb{I}_{R_{n,i}(X_i) < R_{n-1,i}(X_i)} \right)^3 \right. \right. \\ &\quad \left. \left. \cdot \sum_{i=1}^{n-1} \left(\ln \frac{\mu(B(X_i, R_{n-1,i}(X_i)))}{\mu(B(X_i, R_{n,i}(X_i)))} \right)^4 \right] \right). \end{aligned}$$

For any x , $\mu(B(x, \|x - X_i\|))$ is uniformly distributed on $[0, 1]$ (see Section 1.2 in [4]), and therefore

$$\begin{aligned} &(n-1)\mu(B(X_n, R_{n,n}(X_n))) \\ &= (n-1)\mu(B(X_n, \min_{1 \leq i \leq n-1} \|X_i - X_n\|)) \\ &= (n-1) \min_{1 \leq i \leq n-1} \mu(B(X_n, \|X_i - X_n\|)). \end{aligned}$$

It implies, that for given X_n

$$(n-1)\mu(B(X_n, R_{n,n}(X_n))) \stackrel{\mathcal{L}}{=} (n-1) \min_{1 \leq i \leq n-1} U_i,$$

where $\stackrel{\mathcal{L}}{=}$ denotes equality in distribution, and $U_1, \dots, U_{n-1}$ are i.i.d. uniform on $[0, 1]$. Thus,

$$\begin{aligned} & \mathbb{E} \left[(\ln((n-1)\mu(B(X_n, R_{n,n}(X_n))))^4 \right] \\ &= \mathbb{E} \left[\left(\ln \left((n-1) \min_{1 \leq i \leq n-1} U_i \right) \right)^4 \right] \\ &= O(1). \end{aligned}$$

Lemma 20.6 in [4] yields

$$\begin{aligned} \sum_{i=1}^{n-1} \mathbb{I}_{R_{n,i}(X_i) < R_{n-1,i}(X_i)} &= \sum_{i=1}^{n-1} \mathbb{I}_{\|X_i - X_n\| < R_{n-1,i}(X_i)} \\ &\leq \gamma_d \end{aligned}$$

a.s., where γ_d is the minimal number of cones of angle $\pi/6$ that cover $\mathbb{R}^d$. Thus,

$$\begin{aligned} & \mathbb{E} \left[(M_n - \mathbb{E}\{M_n\})^4 \right] \\ &\leq \frac{6^3 c^2}{n^2} (O(1)) \\ &+ \gamma_d^3 (n-1) \mathbb{E} \left[\left(\ln \frac{\mu(B(X_1, R_{n-1,1}(X_1)))}{\mu(B(X_1, R_{n,1}(X_1)))} \right)^4 \right]. \end{aligned}$$

One has

$$\begin{aligned} & \mathbb{E} \left[\left(\ln \frac{\mu(B(X_1, R_{n-1,1}(X_1)))}{\mu(B(X_1, R_{n,1}(X_1)))} \right)^4 \right] \\ &= \int_0^\infty \mathbb{P} \left\{ \left(\ln \frac{\mu(B(X_1, R_{n-1,1}(X_1)))}{\mu(B(X_1, R_{n,1}(X_1)))} \right)^4 \geq s \right\} ds \\ &= \int_0^\infty \mathbb{E} \left[\mathbb{P} \left\{ \frac{\mu(B(X_1, R_{n-1,1}(X_1)))}{\mu(B(X_1, \|X_1 - X_n\|))} \geq e^{s^{1/4}} \right\} \right. \\ &\quad \left. | X_1, \dots, X_{n-1} \right] ds \\ &= \int_0^\infty \mathbb{E} \left[\mathbb{P} \left\{ \mu(B(X_1, R_{n-1,1}(X_1))) e^{-s^{1/4}} \right. \right. \\ &\quad \left. \left. \geq \mu(B(X_1, \|X_1 - X_n\|)) | X_1, \dots, X_{n-1} \right\} \right] ds \\ &= \int_0^\infty \mathbb{E} [\mu(B(X_1, R_{n-1,1}(X_1))) e^{-s^{1/4}}] ds \\ &\leq \int_0^\infty e^{-s^{1/4}} ds / (n-1). \end{aligned}$$

These limit relations imply

$$\mathbb{E} \left[(M_n - \mathbb{E}\{M_n\})^4 \right] = O(1/n^2),$$

which together with the Borel-Cantelli Lemma yields (8).

In the proof of (9) we apply Breiman's generalized ergodic theorem (see Lemma 27.2 in [15]):

Lemma 2: Let $X_1, X_2, \dots$ be a stationary and ergodic sequence. If $F_n = F_n(\{X_i\})$, $n = 1, 2, \dots$ are random functions such that

$$F_n(X_1) \rightarrow F(X_1) \quad (11)$$

a.s. and

$$\mathbb{E}\{\sup_n |F_n(X_1)|\} < \infty, \quad (12)$$

then

$$\frac{1}{n} \sum_{i=1}^n F_n(X_i) \rightarrow \mathbb{E}\{F(X_1)\} \quad (13)$$

a.s.

If $X'_n(x)$ stands for the second nearest neighbor of x among $X_1, \dots, X_n$, then $R_{n,i}(X_i) = \|X'_n(X_i) - X_i\|$ and so

$$\bar{f}_{R_{n,i}(X_i)}(X_i) = \bar{f}_{\|X'_n(X_i) - X_i\|}(X_i)$$

and

$$\tilde{H}_n = -\frac{1}{n} \sum_{i=1}^n \ln \bar{f}_{\|X'_n(X_i) - X_i\|}(X_i).$$

Therefore, (9) means that

$$-\frac{1}{n} \sum_{i=1}^n \ln \bar{f}_{\|X'_n(X_i) - X_i\|}(X_i) \rightarrow H(f) \quad (14)$$

a.s. Defining

$$F_n(x) = -\ln \bar{f}_{\|X'_n(x) - x\|}(x)$$

and

$$F(x) = -\ln f(x),$$

we verify the conditions of Lemma 2. The Lebesgue differentiation theorem (cf. Theorem 20.18 in [4]) yields that

$$\lim_{r \downarrow 0} \bar{f}_r(x) = f(x)$$

for λ -almost all x . The Cover-Hart theorem (cf. Lemma 2.2 in [4]) implies

$$\|X'_n(x) - x\| \rightarrow 0$$

a.s. for μ -almost all x . As μ is absolutely continuous with respect to λ , these limit relations result in

$$\bar{f}_{\|X'_n(x) - x\|}(x) \rightarrow f(x)$$

a.s. for μ -almost all x , from which (11) follows. Let L denote an upper bound on $\|X\|$. Introduce the Hardy-Littlewood maximal functions

$$f^*(x) = \sup_{h>0} \bar{f}_h(x)$$

and

$$g^*(x) = \sup_{2L>h>0} \frac{1}{\bar{f}_h(x)}$$

(4) implies that

$$\int f(x) \ln(f^*(x) + 1) dx < \infty, \quad (15)$$

while if, in addition, X is bounded, then

$$\int f(x) \ln(g^*(x) + 1) dx < \infty, \quad (16)$$

see page 82 in [4]. Note that (16) is the only item in the proof, where the boundedness of X is used. Thus,

$$\begin{aligned} & |\ln \bar{f}_{\|X'_n(x)-x\|}(x)| \\ &= (\ln \bar{f}_{\|X'_n(x)-x\|}(x))^+ + \left(\ln \frac{1}{\bar{f}_{\|X'_n(x)-x\|}(x)} \right)^+ \\ &\leq \ln(f^*(x) + 1) + \ln(g^*(x) + 1), \end{aligned}$$

and so (15) and (16) result in

$$\begin{aligned} & \mathbb{E}\{\sup_n |\ln \bar{f}_{\|X'_n(X_1)-X_1\|}(X_1)|\} \\ &\leq \int f(x) \ln(f^*(x) + 1) dx + \int f(x) \ln(g^*(x) + 1) dx \\ &< \infty, \end{aligned}$$

which yields (12), and the conditions of Lemma 2 are verified.

Note, that similarly to the proof of (8), we can show the universal strong law of the sum nearest neighbor balls: for any density f ,

$$\sum_{i=1}^n \mu \left(B \left(X_i, \min_{j \neq i, j \leq n} \|X_i - X_j\| \right) \right) \rightarrow 1$$

a.s. $\square$

Proof of Theorem 2: First we have to show that the expectation $\mathbb{E}\{H_n\}$ and equivalently $\mathbb{E}\{\tilde{H}_n\}$ exist. Jensen's inequality implies that

$$\begin{aligned} \mathbb{E}\{(\tilde{H}_n)^-\} &= \mathbb{E} \left\{ \left(-\frac{1}{n} \sum_{i=1}^n \ln \bar{f}_{R_{n,i}(X_i)}(X_i) \right)^- \right\} \\ &\geq \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left\{ (-\ln \bar{f}_{R_{n,i}(X_i)}(X_i))^- \right\} \\ &= \mathbb{E} \left\{ (-\ln \bar{f}_{R_{n,1}(X_1)}(X_1))^- \right\} \\ &\geq - \int f(x) (\ln f^*(x))^+ dx \\ &> -\infty, \end{aligned}$$

when

$$\int f(x) \ln \frac{f^*(x)}{f(x)} dx < \infty. \quad (17)$$

Choose $0 < L < \text{ess sup}_x f(x)$. Jensen's inequality implies

$$\begin{aligned} & \int f(x) \ln \frac{f^*(x)}{f(x)} \mathbb{I}_{f(x) > L} dx \\ &\leq \left(\int f(x) \mathbb{I}_{f(x) > L} dx \right) \ln \frac{\int f^*(x) \mathbb{I}_{f(x) > L} dx}{\int f(x) \mathbb{I}_{f(x) > L} dx} \\ &< \infty, \end{aligned}$$

because (4) together with

$$\lambda(\{x : f(x) > L\}) < \infty$$

yields

$$\int f^*(x) \mathbb{I}_{f(x) > L} dx < \infty,$$

see Fefferman and Stein [11]. Furthermore,

$$\begin{aligned} & \int f(x) \ln \frac{f^*(x)}{f(x)} \mathbb{I}_{f(x) \leq L} dx \\ &= \int f(x) \ln f^*(x) \mathbb{I}_{f(x) \leq L} dx \\ &\quad - \int f(x) \ln f(x) \mathbb{I}_{f(x) \leq L} dx \\ &\leq L \int \ln \max\{f^*(x), 1\} dx \\ &\quad - \int f(x) \ln f(x) \mathbb{I}_{f(x) \leq L} dx. \end{aligned}$$

Fefferman and Stein [11] proved that

$$\lambda(\{x : f^*(x) > t\}) \leq \frac{c}{t}, \quad (18)$$

with $t > 0$ such that c depends only on the dimension d , see also (a) of Lemma 10.47 in [37]. By (18),

$$\begin{aligned} & \int \ln \max\{f^*(x), 1\} dx \\ &= \int_0^\infty \lambda(\{x : \ln \max\{f^*(x), 1\} > t\}) dt \\ &= \int_0^\infty \lambda(\{x : \max\{f^*(x), 1\} > e^t\}) dt \\ &= \int_0^\infty \lambda(\{x : f^*(x) > e^t\}) dt \\ &\leq \int_0^\infty \frac{c}{e^t} dt \\ &= c. \end{aligned}$$

Thus, (17) is verified and so we proved that $\mathbb{E}\{\tilde{H}_n\}$ exists.

ASSUME THAT $\mathbb{E}\{(\ln \|X\|)^+\} = \infty$. Then, for any $x \in \mathbb{R}^d$, $\mathbb{E}\{(\ln \|X - x\|)^+\} = \infty$ and therefore $\mathbb{E}\{(\ln R_{2,1}(X_1))^+\} = \mathbb{E}\{(\ln \|X_1 - X_2\|)^+\} = \infty$. Next we show that $\mathbb{E}\{(\ln R_{n,1}(X_1))^+\} = \mathbb{E}\{\min_{2 \leq i \leq n} (\ln \|X_1 - X_i\|)^+\} = \infty$, too: Find r such that $\mu(B(0, r)) = 1/2$. Then

$$\begin{aligned} & \mathbb{E}\{(\ln R_{n,1}(X_1))^+\} \\ &\geq \mathbb{E}\left\{ \left(\ln \frac{\|X_1\|}{2} \right)^+ \mathbb{I}_{\|X_1\| \geq 2r} \mathbb{I}_{X_2, \dots, X_n \in B(0, r)} \right\} \\ &= \frac{1}{2^{n-1}} \mathbb{E}\left\{ \left(\ln \frac{\|X_1\|}{2} \right)^+ \mathbb{I}_{\|X_1\| \geq 2r} \right\} \\ &= \infty. \end{aligned}$$

Thus,

$$\begin{aligned} \mathbb{E}\{\tilde{H}_n\} &= - \int \mathbb{E}\{\ln \bar{f}_{R_{n,1}(x)}(x)\} f(x) dx \\ &\geq \int \mathbb{E}\left\{ \left(\ln \frac{\lambda(B(x, R_{n,1}(x)))}{\mu(B(x, R_{n,1}(x)))} \right)^+ \right\} f(x) dx \\ &\geq \int \mathbb{E}\{(\ln \lambda(B(x, R_{n,1}(x))))^+\} f(x) dx \\ &\geq d \mathbb{E}\{(\ln R_{n,1}(X_1))^+\} \\ &= \infty, \end{aligned} \quad (19)$$

which yields the necessary part of the theorem:

$$\mathbb{E}\{|\tilde{H}_n - H(f)|\} = \infty.$$

ASSUME THAT $\mathbb{E}\{(\ln \|X\|)^+\} < \infty$. Notice that

$$\begin{aligned} \mathbb{E}\{H_n\} &= \mathbb{E}\{\ln((n-1)R_{n,1}(X_1)^d v_d)\} + C_E \\ &\leq \ln(n-1) + \mathbb{E}\{\ln(R_{2,1}(X_1)^d v_d)\} + C_E \\ &\leq \ln(n-1) + \mathbb{E}\{(\ln(\|X_2 - X_1\|^d v_d))^+\} + C_E \\ &< \infty. \end{aligned} \quad (20)$$

(19) and (20) means that $\mathbb{E}\{H_n\} < \infty$ iff (7) holds. By the proof of Theorem 1, (6) is equivalent to

$$\mathbb{E}\{|\tilde{H}_n - H(f)|\} \rightarrow 0. \quad (21)$$

Note that for $a \in \mathbb{R}$, $|a| = 2a^+ - a$ and thus, one has

$$\begin{aligned} \mathbb{E}\{|\tilde{H}_n - H(f)|\} &\leq \int \mathbb{E}\left\{\left|\ln \frac{\bar{f}_{R_{n,1}}(x)}{f(x)}\right|\right\} f(x) dx \\ &= 2 \int \mathbb{E}\left\{\left(\ln \frac{\bar{f}_{R_{n,1}}(x)}{f(x)}\right)^+\right\} f(x) dx \\ &\quad + \mathbb{E}\{\tilde{H}_n\} - H(f). \end{aligned}$$

We show that

$$\int \mathbb{E}\left\{\left(\ln \frac{\bar{f}_{R_{n,1}}(x)}{f(x)}\right)^+\right\} f(x) dx \rightarrow 0 \quad (22)$$

and

$$\mathbb{E}\{\tilde{H}_n\} = - \int \mathbb{E}\{\ln \bar{f}_{R_{n,1}}(x)\} f(x) dx \rightarrow H(f). \quad (23)$$

Concerning (22), we have a domination:

$$\left(\ln \frac{\bar{f}_{R_{n,1}}(x)}{f(x)}\right)^+ \leq \left(\ln \frac{f^*(x)}{f(x)}\right)^+ = \ln \frac{f^*(x)}{f(x)},$$

and therefore (17) and the pointwise convergence yield (22).

With respect to (23), note that (22) implies that

$$\begin{aligned} (H(f) - \mathbb{E}\{\tilde{H}_n\})^+ &\leq \int \mathbb{E}\left\{\left(\ln \frac{\bar{f}_{R_{n,1}}(x)}{f(x)}\right)^+\right\} f(x) dx \\ &\rightarrow 0 \end{aligned}$$

and therefore

$$\liminf_n \mathbb{E}\{\tilde{H}_n\} \geq H(f).$$

So, we need to show

$$\limsup_n \mathbb{E}\{\tilde{H}_n\} \leq H(f). \quad (24)$$

For

$$g_r^*(x) = \inf_{0 < h \leq r} \bar{f}_h(x),$$

we have that

$$\begin{aligned} \mathbb{E}\{\tilde{H}_n\} &= -\mathbb{E}\left\{\int f(x) \ln \bar{f}_{R_{n,1}}(x) dx\right\} \\ &= -\mathbb{E}\left\{\int \mathbb{I}_{R_{n,1}(x) \leq r} f(x) \ln \bar{f}_{R_{n,1}}(x) dx\right\} \\ &\quad - \mathbb{E}\left\{\int \mathbb{I}_{R_{n,1}(x) > r} f(x) \ln \bar{f}_{R_{n,1}}(x) dx\right\} \\ &\leq - \int f(x) \ln g_r^*(x) dx \\ &\quad + \mathbb{E}\left\{\int f(x) \left(\ln \frac{\lambda(B(x, R_{n,1}(x)))}{\mu(B(x, r))}\right)^+ dx\right\}. \end{aligned}$$

The integrand of the last term tends to 0 a.s. for μ -almost all x and the convergence is dominated by

$$\mathbb{E}\left\{\int f(x) \left(\ln \frac{\lambda(B(x, R_{2,1}(x)))}{\mu(B(x, r))}\right)^+ dx\right\} < \infty,$$

which follows from

$$\mathbb{E}\{\tilde{H}_2\} = \mathbb{E}\left\{\int f(x) \ln \frac{\lambda(B(x, R_{2,1}(x)))}{\mu(B(x, R_{2,1}(x)))} dx\right\} < \infty.$$

Thus, by the dominated convergence theorem

$$\mathbb{E}\left\{\int f(x) \left(\ln \frac{\lambda(B(x, R_{n,1}(x)))}{\mu(B(x, r))}\right)^+ dx\right\} \rightarrow 0,$$

and so

$$\limsup_n \mathbb{E}\{\tilde{H}_n\} \leq - \int f(x) \ln g_r^*(x) dx.$$

Therefore,

$$\begin{aligned} \limsup_n \mathbb{E}\{\tilde{H}_n\} &\leq - \sup_{0 < r} \int f(x) \ln g_r^*(x) dx \\ &= - \lim_{r \downarrow 0} \int f(x) \ln g_r^*(x) dx \\ &= H(f). \end{aligned}$$

Thus, (24) is verified and so the proof of the sufficient part of the theorem is complete. $\square$

Proof of Fact 1: The data can be represented as follows: let $Y_1, \dots, Y_n$ be i.i.d. such that

$$\mathbb{P}\{Y_i = j\} = \Delta_j.$$

Let $U_1, \dots, U_n$ be i.i.d. uniform on $[0, 1]$. Set

$$X_i = a_{Y_i} + \Delta_{Y_i} U_i.$$

We will recall two things from the theory of order statistics:

(i) If $U_1, \dots, U_n$ are i.i.d. uniform on $[0, 1]$, then the smallest neighbor distance Z_n^* (the smallest 1-spacings) satisfies

$$Z_n^* \rightarrow E$$

in distribution, where E is standard exponential.

(ii) With probability one,

$$\max_{1 \leq i \leq n} Y_i \geq n^{1/2}$$

except finitely often.

Proof:

$$\mathbb{P}\left\{\max_{1 \leq i \leq n} Y_i \leq n^{1/2}\right\} = (1 - \mathbb{P}\{Y_i > n^{1/2}\})^n \leq e^{-n^{1/2}}.$$

Apply Borel-Cantelli.

(iii) If Y^{**}, Y^* are the largest and second largest Y_i 's, then

$$\mathbb{P}\{Y^{**} = Y^*\} \rightarrow 0.$$

Proof:

$$\begin{aligned} \mathbb{P}\{Y^{**} = Y^*\} &\leq \binom{n}{2} \mathbb{P}\left\{Y_1 = Y_2 \geq \max_{3 \leq i \leq n} Y_i\right\} \\ &\leq n^2 \sum_{i=1}^{\infty} \Delta_i^2 (1 - 1/(i+1))^{n-2} \\ &\leq n^2 \sum_{i \leq n^{5/6}} \Delta_i^2 e^{-(n-2)/(i+1)} \\ &\quad + n^2 \sum_{i > n^{5/6}} \Delta_i^2. \end{aligned}$$

The first term on the right hand side is less than

$$n^2 \sum_{i \leq n^{5/6}} e^{-(n-2)/(n^{5/6}+1)} = e^{-n^{1/6}(1+o(1))} \rightarrow 0,$$

while the second term on the right hand side is

$$O(n^2/n^{5/6}) = O(1/\sqrt{n}) \rightarrow 0.$$

For the sake of simplicity, consider the negative of the estimate in (3) without the bias correction:

$$\ell_n = \frac{1}{n} \sum_{i=1}^n \log_2 \frac{1}{nZ_i}, \quad (25)$$

where $Z_i = \min_{j \neq i, j \leq n} |X_i - X_j|$. We show that $\ell_n \rightarrow -\infty$ in probability, which implies the fact. Let Y^{**} and Y^* be the Y -values of the largest and second largest X_i 's. Let Z^{**} be the nearest neighbor distance for to largest X_i 's. This is also the distance between the two largest X_i 's. If $Y^{**} \neq Y^*$, then

$$Z^{**} \geq a_{Y^{**}} - a_{Y^*} - \Delta_{Y^*} \geq \frac{1}{2} a_{Y^{**}} = \frac{1}{2} 2^{Y^{**}}.$$

Furthermore, all other nearest neighbor distances are $\geq$ (in distribution) the minimal distance between n uniform order statistics (by pushing the intervals of A together), and this is $(1/n^2)$ times something that tends in law to a standard exponential E . So, by (ii)

$$\begin{aligned} \ell_n &\leq \frac{1}{n} \log_2 \left[\frac{2}{2^{2Y^{**}}} \right] + \log_2 \left[\frac{n}{E + o_P(1)} \right] \\ &= -\frac{2^{Y^{**}} - 1}{n} + O_P(\log_2 n) \\ &\leq -\frac{1}{n} \mathbb{I}_{Y^{**} \leq \sqrt{n}} - \frac{2\sqrt{n} - 1}{n} \mathbb{I}_{Y^{**} > \sqrt{n}} + O_P(\log_2 n). \end{aligned}$$

Thus,

$$\begin{aligned} \mathbb{P}\left\{\ell_n > -\frac{2\sqrt{n}}{2n}\right\} &\leq \mathbb{P}\{Y^{**} = Y^*\} + \mathbb{P}\{Y^{**} \leq \sqrt{n}\} \\ &\rightarrow 0, \end{aligned}$$

where we applied (iii). $\square$

ACKNOWLEDGMENT

The authors would like to thank the referees and the associate editor for their help.

REFERENCES

- [1] A. R. Barron, L. Györfi, and E. C. van der Meulen, "Distribution estimation consistent in total variation and in two types of information divergence," *IEEE Trans. Inf. Theory*, vol. 38, no. 5, pp. 1437–1454, Sep. 1992.
- [2] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. van der Meulen, "Nonparametric entropy estimation: An overview," *Int. J. Math. Stat. Sci.*, vol. 6, pp. 17–39, Jun. 1997.
- [3] T. B. Berrett, R. J. Samworth, and M. Yuan, "Efficient multivariate entropy estimation via k -nearest neighbour distances," *Ann. Statist.*, vol. 47, no. 1, pp. 288–318, 2019.
- [4] G. Biau and L. Devroye, *Lectures on the Nearest Neighbor Method*. Cham, Switzerland: Springer-Verlag, 2015.
- [5] L. Birgé and P. Massart, "Estimation of integral functionals of a density," *Ann. Statist.*, vol. 23, no. 1, pp. 11–29, 1995.
- [6] C. Christophe, N. Fabien, and S. O. Silvia, "A note on the adaptive estimation of the differential entropy by wavelet methods," *Commentationes Math. Univ. Carolinae*, vol. 58, no. 1, pp. 87–100, Jun. 2017.
- [7] S. Delattre and N. Fournier, "On the Kozachenko-Leonenko entropy estimator," *J. Statist. Planning Inference*, vol. 185, pp. 69–93, Jun. 2017.
- [8] B. Delyon and A. Juditsky, "On minimax wavelet estimators," *Appl. Comput. Harmon. Anal.*, vol. 3, no. 3, pp. 215–228, Jul. 1996.
- [9] L. Devroye, L. Györfi, G. Lugosi, and H. Walk, "A nearest neighbor estimate of the residual variance," *Electron. J. Statist.*, vol. 12, no. 1, pp. 1752–1778, 2018.
- [10] D. L. Donoho, I. M. Johnstone, G. Kerkycharian, and D. Picard, "Density estimation by wavelet thresholding," *Ann. Statist.*, vol. 24, no. 2, pp. 508–539, Apr. 1996.
- [11] C. Fefferman and E. M. Stein, "Some maximal inequalities," *Amer. J. Math.*, vol. 93, no. 1, pp. 107–115, 1971.
- [12] W. Gao, S. Oh, and P. Viswanath, "Density functional estimators with k -nearest neighbor bandwidths," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 1351–1355.
- [13] W. Gao, S. Oh, and P. Viswanath, "Breaking the bandwidth barrier: Geometrical adaptive entropy estimation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 29, Barcelona, Spain, Dec. 2016, pp. 2460–2468.
- [14] M. Godavarti, "Convergence of differential entropies," *IEEE Trans. Inf. Theory*, vol. 50, no. 1, pp. 171–176, Jan. 2004.
- [15] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. New York, NY, USA: Springer-Verlag, 2002.
- [16] L. Györfi, I. Páli, and E. C. van der Meulen, "There is no universal source code for an infinite source alphabet," *IEEE Trans. Inf. Theory*, vol. 40, no. 1, pp. 267–271, Jan. 1994.
- [17] L. Györfi and E. C. van der Meulen, "On the nonparametric estimation of the entropy functional," in *Nonparametric Functional Estimation and Related Topics*, G. Roussas, Ed. Norwell, MA, USA: Kluwer, 1991, pp. 81–95.
- [18] P. Hall and S. C. Morton, "On the estimation of entropy," *Ann. Inst. Statist. Math.*, vol. 45, no. 1, pp. 69–88, 1993.
- [19] Y. Han, J. Jiao, T. Weissman, and Y. Wu, "Optimal rates of entropy estimation over Lipschitz balls," *Ann. Statist.*, vol. 48, no. 6, pp. 3228–3250, Dec. 2020.
- [20] J. Jiao, W. Gao, and Y. Han, "The nearest neighbor information estimator is adaptively near minimax rate-optimal," 2017, *arXiv:1711.08824*.
- [21] H. Joe, "Estimation of entropy and other functionals of a multivariate density," *Ann. Inst. Statist. Math.*, vol. 41, no. 4, pp. 683–697, 1989.
- [22] K. Kandasamy, A. Krishnamurthy, B. Póczos, and L. Wasserman, "Nonparametric von Mises estimators for entropies, divergences and mutual informations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 397–405.
- [23] L. F. Kozachenko and N. N. Leonenko, "Sample estimate of entropy of a random vector," *Problems Inf. Transmiss.*, vol. 23, no. 2, pp. 95–101, 1987.
- [24] A. Krishnamurthy, K. Kandasamy, B. Póczos, and L. Wasserman, "Nonparametric estimation of Rényi divergence and friends," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 919–927.
- [25] W. M. Lord, J. Sun, and E. M. Bolitt, "Geometric k -nearest neighbor estimation of entropy and mutual information," *Chaos, Interdiscipl. J. Nonlinear Sci.*, vol. 28, no. 3, Mar. 2018, Art. no. 033114.

- [26] U. Ozertem, I. Uysal, and D. Erdogmus, "Continuously differentiable sample-spacing entropy estimation," *IEEE Trans. Neural Netw.*, vol. 19, no. 11, pp. 1978–1984, Nov. 2008.
- [27] L. Paninski and M. Yajima, "Undersmoothed kernel entropy estimators," *IEEE Trans. Inf. Theory*, vol. 54, no. 9, pp. 4384–4388, Sep. 2008.
- [28] S. Shwartz, M. Zibulevsky, and Y. Y. Schechner, "Fast kernel entropy estimation and optimization," *Signal Process.*, vol. 85, no. 5, pp. 1045–1058, May 2005.
- [29] J. Silva and S. Narayanan, "Nonproduct data-dependent partitions for mutual information estimation: Strong consistency and applications," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3497–3511, Jul. 2010.
- [30] H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk, "Nearest neighbor estimates of entropy," *Amer. J. Math. Manage. Sci.*, vol. 23, nos. 3–4, pp. 301–321, 2003.
- [31] S. Singh and B. Póczos, "Finite-sample analysis of fixed- k nearest neighbor density functional estimators," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1217–1225.
- [32] K. Sricharan, R. Raich, and A. O. Hero, "Estimation of nonlinear functionals of densities with confidence," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4135–4159, Jul. 2012.
- [33] K. Sricharan, D. Wei, and A. O. Hero, "Ensemble estimators for multivariate entropy estimation," *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4374–4388, Jul. 2013.
- [34] D. Stowell and M. D. Plumbley, "Fast multidimensional entropy estimation by k -d partitioning," *IEEE Signal Process. Lett.*, vol. 16, no. 6, pp. 537–540, Jun. 2009.
- [35] A. B. Tsybakov and E. C. van der Meulen, "Root- n consistent estimators of entropy for densities with unbounded support," *Scand. J. Statist.*, vol. 23, no. 1, pp. 75–83, 1996.
- [36] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Trans. Inf. Theory*, vol. 51, no. 9, pp. 3064–3074, Sep. 2005.
- [37] R. L. Wheeden and A. Zygmund, *Measure and Integral*. New York, NY, USA: Marcel Dekker, 1977.

Luc Devroye was born in Belgium. He received the Ph.D. degree from the University of Texas, in 1976. He joined the School of Computer Science, McGill University, Montreal, Canada, in 1977. His research interests include probability theory as applied to the analysis of algorithms, mathematical statistics, machine learning, pattern recognition, and random number generation.

László Györfi was born in Hungary. He received the degree in mathematics and physics from the Loránd Eötvös University, Budapest, in 1970. He joined the Department of Computer Science and Information Theory, Budapest University of Technology and Economics, Budapest, Hungary. He became an Ordinary Member of the Hungarian Academy of Sciences, in 2001. His main interests are nonparametric statistics and multiple access communication.