Uniform-in-bandwidth consistency for kernel-type estimators of Shannon's entropy

Salim Bouzebda and Issam Elhattab

L.S.T.A., Université Pierre et Marie Curie
4 place Jussieu 75252 Paris Cedex 05, France
e-mail: salim.bouzebda@upmc.fr
ENCG-Casablanca, Université Hassan II Mohammedia
e-mail: issam.elhattab@upmc.fr

Abstract: We establish uniform-in-bandwidth consistency for kernel-type estimators of the differential entropy. We consider two kernel-type estimators of Shannon's entropy. As a consequence, an asymptotic 100% confidence interval of entropy is provided.

Contents

1. Introduction and estimation

Let $(\mathbf{X}_n)_{n \geq 1}$ be a sequence of independent and identically distributed $\mathbb{R}^d$-valued random vectors, $d \geq 1$, with cumulative distribution function $\mathbb{F}(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^d$ and a density function $f(\cdot)$ with respect to Lebesgue measure on $\mathbb{R}^d$. Here, as usual, $\mathbf{X} = (X_1, \ldots, X_d) \leq \mathbf{x} = (x_1, \ldots, x_d)$ means that each component of $\mathbf{X}$ is less than or equal to the corresponding component of $\mathbf{x}$, that is, $X_i \leq x_i$, for all $i = 1, \ldots, d$. The differential (or Shannon) entropy of $f(\cdot)$ is defined to be

$$H(f) \quad := \quad -\int_{\mathbb{R}^d} f(\mathbf{x}) \log\left(f(\mathbf{x})\right) d\mathbf{x} \tag{1.1}$$

$$:= \quad -\int_{\mathbb{R}^d} \log\left(f(\mathbf{x})\right) d\mathbb{F}(\mathbf{x}), \tag{1.2}$$

440

whenever this integral is meaningful, and where, for $\mathbf{x} = (x_1, \ldots, x_d)$, $d\mathbf{x}$ denotes Lebesgue measure in $\mathbb{R}^d$. We will use the convention that $0 \log(0) = 0$ since $u \log(u) \to 0$ as $u \to 0$.

The concept of differential entropy was originally introduced in Shannon's paper [42]. Since this early epoch, the notion of entropy has been the subject of great theoretical and applied interest. We refer to [10, Chapter 8.] for a comprehensive overview of differential entropy and their mathematical properties. Entropy concepts and principles play an fundamental role in many applications, such as statistical communication theory [25], quantization theory [39], statistical decision theory [33], and contingency table analysis [28]. [11] introduced the concept of convergence in entropy and showed that the latter convergence concept implies convergence in $\mathcal{L}_1$. This property indicates that entropy is a useful concept to measure "*closeness in distribution*", and also justifies heuristically the usage of sample entropy as test statistics when designing entropy-based tests of goodness-of-fit. This line of research has been pursued by [45, 38, 19, 27, 20] and [24] [including the references therein]. The idea here is that many families of distributions are characterized by maximization of entropy subject to constraints (see, e.g., [32] and [34]). There is a huge literature on the Shannon's entropy and its applications. It is not the purpose of this paper to survey this extensive literature.

In the literature, various estimator for $H(f)$, based on a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from the underlying distribution, have been proposed and their asymptotic properties studied. For an exhaustive list of references in this vein, we refer to [29, 4] and the references therein.

We mention that there exist mainly two approaches to the construction of entropy estimators. The first approach is based on spacings when $d = 1$. The the second approach, to be used in this paper to estimate $H(f)$, consists in first obtaining a suitable density estimate $f_n(\cdot)$ for $f(\cdot)$, and then substituting $f(\cdot)$ by $f_n(\cdot)$ in an entropy-like functional of $f(\cdot)$.

The main contribution of the present paper is to establish an almost sure uniform in bandwidth consistency of the kernel-type estimator of the entropy functional $H(f)$. In the entropy framework, the results obtained here are believed to be novel.

We start by giving some notation and conditions that are needed for the forthcoming sections. Below, we will work under the following assumptions on $f(\cdot)$ to establish our results.

(F.1) The functional $H(f)$ is well-defined by (1.1), in the sense that

$$|H(f)| < \infty. \tag{1.3}$$

We recall from (cf. [3, p. 237], [5, p. 108]) that the finiteness of $H(f)$ is guaranteed if both $\mathbb{E}\|\mathbf{X}\|^2 < \infty$, where $\|\cdot\|$ denotes the Euclidian norm in $\mathbb{R}^d$, (in which case $H(f) < \infty$) and $f(\cdot)$ is bounded (in which case $H(f) > -\infty$). Ash gives an example of a density function on $\mathbb{R}$ for which $H(f) = \infty$ and also one for which $H(f) = -\infty$. We refer to [30, Section 4] for conditions characterizing (1.3) in terms of $f(\cdot)$.

To define our entropy estimator we define, in a first step, a kernel density estimator. Towards this aim, we introduce a measurable function $K(\cdot)$ fulfilling the following conditions.

(K.1) $K(\cdot)$ is of bounded variation on $\mathbb{R}^d$;
(K.2) $K(\cdot)$ is right continuous on $\mathbb{R}^d$, i.e., for any $\mathbf{t} = (t_1, \ldots, t_d)$, we have

$$K(t_1, \ldots, t_d) = \lim_{\varepsilon_1 \downarrow 0, \ldots, \varepsilon_d \downarrow 0} K(t_1 + \varepsilon_1, \ldots, t_d + \varepsilon_d);$$

(K.3) $\|K\|_\infty := \sup_{\mathbf{t} \in \mathbb{R}^d} |K(\mathbf{t})| =: \kappa < \infty$;
(K.4) $\int_{\mathbb{R}^d} K(\mathbf{t}) d\mathbf{t} = 1$.

The well known Akaike-Parzen-Rosenblatt (refer to [2, 36] and [40]) kernel estimator of $f(\cdot)$ is defined, for any $\mathbf{x} \in \mathbb{R}^d$, by

$$f_{n,h_n}(\mathbf{x}) := (nh_n^d)^{-1} \sum_{i=1}^{n} K((\mathbf{x} - \mathbf{X}_i)/h_n), \tag{1.4}$$

where $0 < h_n \leq 1$ is the smoothing parameter. For notational convenience, we have chosen the same bandwidth sequence for each margins. This assumption can be dropped easily. Refer for example to [23, Remark 8] for more details.

In a second step, given $f_{n,h_n}(\cdot)$, we estimate $H(f)$ using the representation (1.1), by setting

$$H_{n,h_n}^{(1)}(f) := -\int_{A_n} f_{n,h_n}(\mathbf{x}) \log \left(f_{n,h_n}(\mathbf{x}) \right) d\mathbf{x}, \tag{1.5}$$

where

$$A_n := \{\mathbf{x} : f_{n,h_n}(\mathbf{x}) \geq \gamma_n\},$$

and $\gamma_n \downarrow 0$ is a sequence of positive constant. The *plug-in* estimator $H_{n,h_n}^{(1)}(f)$ was introduced by [18] for $d = 1$ and $A_n = [-b_n, b_n]$, where b_n is a specified sequence of constants. The integral estimator $H_{n,h_n}^{(1)}(f)$ can be easily calculated if, for example, $f_n(\cdot)$ is a histogram.

In the present paper, we will consider also the *resubstitution* estimate proposed in [1]. In this case, we shall study uniform-in-bandwidth consistency of the estimator of $H(f)$ based on the representation (1.2) which is, in turn, defined by

$$H_{n,h_n}^{(2)}(f) := -\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\Omega_{n,i}} \log \left(f_{n,h_n}(\mathbf{X}_i) \right), \tag{1.6}$$

where

$$\Omega_{n,i} := \{f_{n,h_n}(\mathbf{X}_i) \geq \gamma_n\}, \quad \text{for} \quad i = 1, \ldots, n$$

The limiting behavior of $f_{n,h_n}(\cdot)$, for appropriate choices of the bandwidth h_n, has been studied by a large number statisticians over many decades. For good sources of references to research literature in this area along with statistical applications consult [16, 15, 6, 41] and [37]. In particular, under our assumptions,

the condition that $h_n \to 0$ together with $nh_n \to \infty$ is necessary and sufficient for the convergence in probability of $f_{n,h_n}(\mathbf{x})$ towards the limit $f(\mathbf{x})$, independently of $\mathbf{x} \in \mathbb{R}^d$ and the density $f(\cdot)$. Various uniform consistency results involving the estimator $f_{n,h_n}(\cdot)$ have been recently established. We refer to [12, 22, 13] and the references therein. The first seminal paper that devoted to obtaining uniform in bandwidth results for the kernel-type estimator was [23]. Since then, there is a considerable interest in obtaining so-called uniform in bandwidth results for kernel-type estimators depending on a bandwidth sequence. In this paper we will use their methods to establish convergence results for the estimates $H^{(1)}_{n,h_n}(f)$ and $H^{(2)}_{n,h_n}(f)$ of $H(f)$ in the same spirit of [7, 8].

The remainder of this paper is organized as follows. In Section 2, we state our main results concerning the limiting behavior of $H^{(1)}_{n,h_n}(f)$ and $H^{(2)}_{n,h_n}(f)$. Some concluding remarks and possible future developments are mentioned in Section 3. To avoid interrupting the flow of the presentation, all mathematical developments are relegated to Section 4.

2. Main results

To prove the strong consistency of $H^{(1)}_{n,h_n}$, we shall consider another, but more appropriate and more computationally convenient, centering factor than the expectation $\mathbb{E}H^{(1)}_{n,h_n}$, which is delicate to handle. This is given by

$$\widehat{\mathbb{E}}H^{(1)}_{n,h_n}(f) := - \int_{A_n} \mathbb{E}f_{n,h_n}(\mathbf{x}) \log\left(\mathbb{E}f_{n,h_n}(\mathbf{x})\right)d\mathbf{x}.$$

The main result, concerning $H^{(1)}_{n,h}$, to be proved here may now be stated precisely as follows.

Theorem 2.1. *Let* $K(\cdot)$ *satisfy* (K.1-2-3-4), *and let* $f(\cdot)$ *be a bounded density fulfill* (F.1). *Let* $c > 0$ *and* $\{h_n\}_{n\geq 1}$ *be a sequence of positive constants such that,* $cn^{-1}\gamma_n^{-4}(\log n) \leq h_n < 1$. *Then there exists a positive constant* Υ, *such that*

$$\limsup_{n\to\infty} \sup_{h_n \leq h \leq 1} \frac{\sqrt{nh\gamma_n^4}|H^{(1)}_{n,h}(f) - \widehat{\mathbb{E}}H^{(1)}_{n,h}(f)|}{\sqrt{(\log(1/h) \vee \log\log n)}} \leq \Upsilon \quad a.s.$$

The proof of Theorem 2.1 is postponed until §4.

Let $(h'_n)_{n\geq 1}$ and $(h''_n)_{n\geq 1}$ be two sequences of constants such that $0 < h'_n < h''_n < 1$, together with $h''_n \to 0$ and $nh'_n\gamma_n^4/\log n \to \infty$, as $n \to \infty$. A direct application of Theorem 2.1 shows that, with probability 1,

$$\sup_{h'_n \leq h \leq h''_n} |H^{(1)}_{n,h}(f) - \widehat{\mathbb{E}}H^{(1)}_{n,h}(f)| = O\left(\sqrt{\frac{(\log(1/h'_n) \vee \log\log n)}{nh'_n\gamma_n^4}}\right).$$

This, in turn, implies that

$$\lim_{n\to\infty} \sup_{h'_n \leq h \leq h''_n} |H^{(1)}_{n,h}(f) - \widehat{\mathbb{E}}H^{(1)}_{n,h}(f)| = 0 \quad a.s. \tag{2.1}$$

The following result handles the uniform deviation of the estimate $H_{n,h_n}^{(1)}(f)$ with respect to $H(f)$.

Corollary 2.2. *Let $K(\cdot)$ satisfy (K.1-2-3-4), and let $f(\cdot)$ be a uniformly Lipschitz continuous and bounded density on $\mathbb{R}^d$, fulfilling (F.1). Then for each pair of sequences $0 < h'_n < h''_n \leq 1$ with $h''_n \to 0$, $nh'_n\gamma_n^4/\log n \to \infty$ and $|\log(h''_n)|/\log\log n \to \infty$ as $n \to \infty$, we have*

$$\lim_{n\to\infty} \sup_{h'_n \leq h \leq h''_n} |H_{n,h}^{(1)}(f) - H(f)| = 0 \qquad a.s. \tag{2.2}$$

The proof of Corollary 2.2 is postponed until §4.

Remark 2.3. We note that the main problem in using entropy estimates such as (1.5) is to choose properly the smoothing parameter h_n. The uniform in bandwidth consistency result given in (2.2) shows that any choice of h between h'_n and h''_n ensures the consistency of $H_{n,h}^{(1)}(f)$. In other word, the fluctuation of the bandwidth in a small interval do not affect the consistency of the nonparametric estimator of $H(f)$.

Now, we shall establish another result in a similar direction for a class of compactly supported densities. We need the following additional conditions.

(F.2) $f(\cdot)$ has a compact support say $\mathbb{I}$ and is s-time continuously differentiable, and there exists a constant $0 < \mathfrak{M} < \infty$ such that

$$\sup_{\mathbf{x}\in\mathbb{I}} \left| \frac{\partial^s f(\mathbf{x})}{\partial x_1^{j_1} \ldots \partial x_d^{j_d}} \right| \leq \mathfrak{M}, \quad j_1 + \cdots + j_d = s.$$

(K.5) $K(\cdot)$ is of order s, i.e., for some constant $\mathfrak{S} \neq 0$,

$$\int_{\mathbb{R}^d} t_1^{j_1} \ldots t_d^{j_d} K(\mathbf{t}) d\mathbf{t} = 0, \quad j_1, \ldots, j_d \geq 0, \quad j_1 + \cdots + j_d = 1, \ldots, s-1,$$

$$\int_{\mathbb{R}^d} |t_1^{j_1} \ldots t_d^{j_d}| K(\mathbf{t}) d\mathbf{t} = \mathfrak{S}, \quad j_1, \ldots, j_d \geq 0, \quad j_1 + \cdots + j_d = s.$$

Under the condition (F.2), the differential entropy of $f(\cdot)$ may be written as follows

$$H(f) = - \int_{\mathbb{I}} f(\mathbf{x}) \log\left(f(\mathbf{x}) \right) d\mathbf{x}.$$

Theorem 2.4. *Let $K(\cdot)$ satisfy (K.1-2-3-4-5), and let $f(\cdot)$ fulfill (F.1-2). Then for each pair of sequences $0 < h'_n < h''_n \leq 1$ with $h''_n \to 0$ and $nh'_n/\log n \to \infty$ as $n \to \infty$, we have, for any $\gamma > 0$*

$$\limsup_{n\to\infty} \sup_{h'_n \leq h \leq h''_n} \frac{\sqrt{nh}|H_{n,h}^{(1)}(f) - H(f)|}{\sqrt{\log(1/h) \vee \log\log n}} \leq \zeta(\mathbb{I}) \quad a.s.,$$

where

$$\zeta(\mathbb{I}) := \left(\frac{\gamma^2 + \gamma + 1}{\gamma^2} \right)^{1/2} \sup_{\mathbf{x}\in\mathbb{I}} \left\{ f(\mathbf{x}) \int_{\mathbb{R}^d} K^2(\mathbf{u}) d\mathbf{u} \right\}^{1/2}.$$

The proof of Theorem 2.4 is postponed until §4.

To state our result concerning $H_{n,h_n}^{(2)}(f)$ we need the following additional condition.

(F.3) $\mathbb{E}\left[\log^2\left(f(\mathbf{X})\right)\right] < \infty.$

Remark 2.5. Condition (F.3) is extremely weak and is satisfied by all commonly encountered distributions including many important heavy tailed distributions for which the moments do not exists (see. e.g. [43] for more details and references on the subject.)

To prove the strong consistency of $H_{n,h_n}^{(2)}$ we consider the following centering factor

$$\widehat{\mathbb{E}} H_{n,h_n}^{(2)}(f) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\Omega_{n,i}} \log\left(\mathbb{E}(f_{n,h_n}(\mathbf{x}) \mid \mathbf{X}_i = \mathbf{x})\right).$$

The main results concerning $H_{n,h}^{(2)}(f)$ are summarized in the following Theorems.

Theorem 2.6. *Let* $K(\cdot)$ *satisfy* (K.1-2-3-4)*, and let* $f(\cdot)$ *be a bounded density fulfilling* (F.1)*. Let* $c > 0$ *and* $\{h_n\}_{n\geq 1}$ *be a sequence of positive constants such that,* $cn^{-1}\gamma_n^{-2}(\log n) \leq h_n < 1$*. Then there exists a positive constant* Υ'*, such that*

$$\limsup_{n\to\infty} \sup_{h_n \leq h \leq 1} \frac{\sqrt{nh\gamma_n^2}|H_{n,h}^{(2)}(f) - \widehat{\mathbb{E}} H_{n,h}^{(2)}(f)|}{\sqrt{(\log(1/h) \vee \log\log n)}} \leq \Upsilon' \quad a.s.$$

The proof of Theorem 2.6 is postponed until §4.

Theorem 2.7. *Assume that the kernel function* $K(\cdot)$ *is compactly supported and satisfies the conditions* (K.1-2-3-4-5)*. Let* $f(\cdot)$ *be a bounded density function fulfilling the conditions* (F.1-2-3)*. Let* $\{h'_n\}_{n\geq 1}$ *and* $\{h''_n\}_{n\geq 1}$ *such that* $h'_n = An^{-\delta}$ *and* $h''_n = Bn^{-\delta}$ *with arbitrary choices of* $0 < A < B < \infty$ *and* $(1/(d+4)) \leq \delta < 1$*. Then, for* $\gamma > 0$*, we have with probability one,*

$$\limsup_{n\to\infty} \sup_{h'_n \leq h \leq h''_n} \frac{\sqrt{nh\gamma_n^2}|H_{n,h}^{(2)}(f) - H(f)|}{\sqrt{2\log(1/h)}} \leq \sigma_{\mathbb{I}}, \qquad (2.3)$$

where

$$\sigma_{\mathbb{I}} := \frac{1}{\gamma}\left\{\sup_{\mathbf{x}\in\mathbb{I}} f(\mathbf{x}) \int_{\mathbb{R}^d} K^2(\mathbf{u})d\mathbf{u}\right\}^{1/2},$$

where $\mathbb{I}$ *is given in* (F.2)*.*

The proof of Theorem 2.7 is postponed until §4.

Remark 2.8. Theorem 2.4 leads, using the techniques developed in [13], to the construction of asymptotic 100% certainty interval for the true entropy $H(f)$, i.e., as $n \to \infty$, for each $\varepsilon > 0$

$$\mathbb{P}\left(H(f) \in \left[H_{n,h}^{(1)}(f) - A_{n,\varepsilon}, H_{n,h}^{(1)}(f) + B_{n,\varepsilon}\right]\right) \approx 100\%,$$

see (2.5) bellow for explicit expressions of $A_{n,\varepsilon}$ and $B_{n,\varepsilon}$. We give in what follows, the idea how to construct this interval. Throughout, we let $h \in [h'_n, h''_n]$, where h'_n and h''_n are as in Theorem 2.4. We infer from Theorem 2.4 that, for suitably chosen data-dependent functions $L_n = L_n(X_1, \ldots, X_n) > 0$, for each $0 < \varepsilon < 1$, we have, as $n \to \infty$,

$$\mathbb{P}\left(\frac{1}{L_n}|H_{n,h}^{(1)}(f) - H(f)| \geq 1 + \varepsilon\right) \to 0. \tag{2.4}$$

Assuming the validity of the statement (2.4), we obtain asymptotic certainty interval for $H(f)$ in the following sense. For each $0 < \varepsilon < 1$, we have, as $n \to \infty$,

$$\mathbb{P}\left(H(f) \in \left[H_{n,h}^{(1)}(f) - (1 + \varepsilon)L_n, H_{n,h}^{(1)}(f) + (1 + \varepsilon)L_n\right]\right) \to 1. \tag{2.5}$$

Whenever (2.5) holds for each $0 < \varepsilon < 1$, we will say that the interval

$$\left[H_{n,h}^{(1)}(f) - L_n, H_{n,h}^{(1)}(f) + L_n\right],$$

provides asymptotic 100% certainty interval for $H(f)$.

To construct L_n we proceed as follows. Assume that there exists a sequence $\{\mathbb{I}_n\}_{n \geq 1}$ of strictly nondecreasing compact subsets of $\mathbb{I}$, such that

$$\bigcup_{n \geq 1} \mathbb{I}_n = \mathbb{I}$$

(for the estimation of the support $\mathbb{I}$ we may refer to [17] and the references therein). Furthermore, suppose that there exists a sequence (possibly random) $\{\zeta_n(\mathbb{I}_n)\}, n = 1, 2, \ldots$, converging to $\zeta(\mathbb{I})$ in the sense that

$$\mathbb{P}\left(\left|\frac{\zeta_n(\mathbb{I}_n)}{\zeta(\mathbb{I})} - 1\right| \geq \varepsilon\right) \to 0 \qquad \text{as } n \to \infty \text{ for each } \varepsilon > 0. \tag{2.6}$$

Observe that the statement (2.6) is satisfied when the choice

$$\zeta_n(\mathbb{I}_n) := \sup_{\mathbf{x} \in \mathbb{I}_n} \sqrt{f_{n,h}(\mathbf{x}) \int_{\mathbb{R}^d} K^2(\mathbf{u}) d\mathbf{u}}$$

is considered. Consequently, we may define the quantity L_n displayed in the statement (2.4) by

$$L_n := \sqrt{\frac{\gamma_n^4 \left(\log(1/h) \vee \log \log n\right)}{nh}} \times \zeta_n(\mathbb{I}_n).$$

Remark 2.9. A practical choice of γ_n is $\beta(\log n)^{-\alpha}$ where $\beta > 0$ and $\alpha \geq 0$. In the case of the density which is bounded away from 0, α is equal to 0.

Remark 2.10. [26] establish uniform in bandwidth consistency and central limit theorems for a different but related estimator to the one proposed in the present paper. That is, [26] propose

$$\widehat{H}_{n,h_n} := -\frac{1}{n} \sum_{i=1}^{n} \log \{f_{n,h_n,-i}(X_i)\},$$

where

$$f_{n,h_n,-i}(X_i) := 1/((n-1)h_n) \sum_{1 \le j \ne i \le n} K\left((X_i - X_j)/h_n\right).$$

Their results hold subject to the condition ((C) p. 751, where we choose $\phi(x) = x \log x$ that corresponds to the negative entropy) which is satisfied when density $f(\cdot)$ is bounded away from 0 on its support, refer to Remark 1. p. 752 of [26], their approach is different from that used in this paper and is based on the notion of a local U-statistic. We mention that the estimator proposed by [26] seems to be simpler and with faster rates of convergence. The fact that we use the a "thresholding" estimator of the entropy permits us to consider a large class of density by paying the price of loss in the rate of convergence. Furthermore, if we assume that the density $f(\cdot)$ is bounded away from 0 on its support, then the rate of the strong convergence is of order $\{\{\log(1/h_n)\}/\{nh_n\}\}^{1/2}$ which is the same rate of the strong convergence for the density kernel-type estimators, this is precisely the contain of Theorem 2.4.

3. Concluding remarks and future works

We have addressed the problem of nonparametric estimation of Shannon's entropy. The results presented in this work are general, since the required conditions are fulfilled by a large class of densities.

The evaluation of the integral in (1.5) requires numerical integration and is not easy if $f_{n,h_n}(\cdot)$ is a kernel density estimator but it does not involve any stochastic aspects. The integral estimator can however be easily calculated if we approximate $f_{n,h_n}(\cdot)$ by piecewise-constant functions on a fine enough partition, for example, $f_{n,h_n}(\cdot)$ is a histogram. We mention that in some particular case ($K(\cdot)$ is a double exponential kernel), the approximations are easily calculated since the distribution function corresponding to the kernel $K(\cdot)$ is available, confer [21] for more details. An interesting aspect of the $H_{n,h_n}^{(2)}(f)$ is that its rate of convergence is faster than that of $H_{n,h_n}^{(1)}(f)$ and that is very easy to compute.

It will be interesting to enrich our results presented here by an additional uniformity in term of γ_n in the supremum appearing in all our theorems, which requires non trivial mathematics, this would go well beyond the scope of the present paper. Another direction of research is to obtain results, based on U-statistic approach, similar to that in [26] for entropy estimator under general conditions, i.e., without assuming the condition that the density $f(\cdot)$ is bounded away from 0 on its support.

4. Proofs

This section is devoted to the proofs of our results.

Proof of Theorem 2.1.

We first decompose $H^{(1)}_{n,h_n}(f) - \widehat{\mathbb{E}} H^{(1)}_{n,h_n}(f)$ into the sum of two components, by writing

$$
\begin{aligned}
H^{(1)}_{n,h_n}(f) &- \widehat{\mathbb{E}} H^{(1)}_{n,h_n}(f) \\
&= -\int_{A_n} f_{n,h_n}(\mathbf{x}) \log\left(f_{n,h_n}(\mathbf{x})\right) d\mathbf{x} \\
&\quad + \int_{A_n} \mathbb{E} f_{n,h_n}(\mathbf{x}) \log\left(\mathbb{E} f_{n,h_n}(\mathbf{x})\right) d\mathbf{x} \\
&= -\int_{A_n} \left\{\log f_{n,h_n}(\mathbf{x}) - \log \mathbb{E} f_{n,h_n}(\mathbf{x})\right\} \mathbb{E} f_{n,h_n}(\mathbf{x}) d\mathbf{x} \\
&\quad - \int_{A_n} \left\{f_{n,h_n}(\mathbf{x}) - \mathbb{E} f_{n,h_n}(\mathbf{x})\right\} \log f_{n,h_n}(\mathbf{x}) d\mathbf{x} \\
&:= \boldsymbol{\Delta}_{1,n,h_n} + \boldsymbol{\Delta}_{2,n,h_n}. \tag{4.1}
\end{aligned}
$$

We observe that for all $z > 0$, $|\log z| \leq \left|\frac{1}{z} - 1\right| + |z - 1|$. Therefore, for any $\mathbf{x} \in A_n = \{\mathbf{x} : f_{n,h_n}(\mathbf{x}) \geq \gamma_n\}$, we get

$$
\begin{aligned}
|\log f_{n,h_n}(\mathbf{x}) - \log \mathbb{E} f_{n,h_n}(\mathbf{x})| &= \left|\log \frac{f_{n,h_n}(\mathbf{x})}{\mathbb{E} f_{n,h_n}(\mathbf{x})}\right| \\
&\leq \left|\frac{\mathbb{E} f_{n,h_n}(\mathbf{x})}{f_{n,h_n}(\mathbf{x})} - 1\right| + \left|\frac{f_{n,h_n}(\mathbf{x})}{\mathbb{E} f_{n,h_n}(\mathbf{x})} - 1\right| \\
&= \frac{|\mathbb{E} f_{n,h_n}(\mathbf{x}) - f_{n,h_n}(\mathbf{x})|}{f_{n,h_n}(\mathbf{x})} + \frac{|f_{n,h_n}(\mathbf{x}) - \mathbb{E} f_{n,h_n}(\mathbf{x})|}{\mathbb{E} f_{n,h_n}(\mathbf{x})}.
\end{aligned}
$$

In the following $\|\cdot\|_\infty$ denotes, as usual, the supremum norm, i.e., $\|\phi(\mathbf{x})\|_\infty := \sup_{\mathbf{x} \in \mathbb{R}^d} \|\phi(\mathbf{x})\|$. We know (see, e.g., [23]), for each $h'_n < h < h''_n$, as $n \to \infty$, we have

$$
\|f_{n,h}(\mathbf{x}) - \mathbb{E} f_{n,h}(\mathbf{x})\|_\infty = O\left(\sqrt{\frac{(\log(1/h'_n) \vee \log\log n)}{nh'_n}}\right).
$$

For any $\mathbf{x} \in A_n$, one can see that

$$
\mathbb{E} f_{n,h_n}(\mathbf{x}) \geq \gamma_n.
$$

We readily obtain from these relations, for any $\mathbf{x} \in A_n$, that

$$
|\log f_{n,h_n}(\mathbf{x}) - \log \mathbb{E} f_{n,h_n}(\mathbf{x})| \leq \frac{2}{\gamma_n} |f_{n,h_n}(\mathbf{x}) - \mathbb{E} f_{n,h_n}(\mathbf{x})|.
$$

We can therefore write, for any $n \geq 1$, the following chain of inequalities

$$
\begin{aligned}
|\boldsymbol{\Delta}_{1,n,h_n}| &= \left| \int_{A_n} \{\log f_{n,h_n}(\mathbf{x}) - \log \mathbb{E} f_{n,h_n}(\mathbf{x})\} \, \mathbb{E} f_{n,h_n}(\mathbf{x}) d\mathbf{x} \right| \\
&\leq \int_{A_n} |\log f_{n,h_n}(\mathbf{x}) - \log \mathbb{E} f_{n,h_n}(\mathbf{x})| \, \mathbb{E} f_{n,h_n}(\mathbf{x}) d\mathbf{x} \\
&\leq \frac{2}{\gamma_n} \int_{A_n} |f_{n,h_n}(\mathbf{x}) - \mathbb{E} f_{n,h_n}(\mathbf{x})| \, \mathbb{E} f_{n,h_n}(\mathbf{x}) d\mathbf{x} \\
&\leq \frac{2}{\gamma_n} \sup_{\mathbf{x} \in A_n} |\mathbb{E} f_{n,h_n}(\mathbf{x}) - f_{n,h_n}(\mathbf{x})| \int_{A_n} \mathbb{E} f_{n,h_n}(\mathbf{x}) d\mathbf{x} \\
&\leq \frac{2}{\gamma_n} \sup_{\mathbf{x} \in \mathbb{R}^d} |\mathbb{E} f_{n,h_n}(\mathbf{x}) - f_{n,h_n}(\mathbf{x})| \int_{\mathbb{R}^d} \mathbb{E} f_{n,h_n}(\mathbf{x}) d\mathbf{x}.
\end{aligned}
$$

In view of condition (K.4), by the change of variables and an application of Fubini's theorem, we have

$$
\int_{\mathbb{R}^d} \mathbb{E} f_{n,h}(\mathbf{x}) d\mathbf{x} = 1.
$$

Thus, for any $n \geq 1$, we have the following bound

$$
|\boldsymbol{\Delta}_{1,n,h_n}| \leq \frac{2}{\gamma_n} \sup_{\mathbf{x} \in \mathbb{R}^d} |\mathbb{E} f_{n,h_n}(\mathbf{x}) - f_{n,h_n}(\mathbf{x})|. \tag{4.2}
$$

We next evaluate the second term $\boldsymbol{\Delta}_{2,n,h_n}$ in the right side of (4.1). Since $|\log z| \leq \frac{1}{z} + z$, for all $z > 0$, one can see that

$$
\begin{aligned}
|\boldsymbol{\Delta}_{2,n,h_n}| &= \left| \int_{A_n} \{f_{n,h_n}(\mathbf{x}) - \mathbb{E} f_{n,h_n}(\mathbf{x})\} \log f_{n,h_n}(\mathbf{x}) d\mathbf{x} \right| \\
&\leq \int_{A_n} |f_{n,h_n}(\mathbf{x}) - \mathbb{E} f_{n,h_n}(\mathbf{x})| \left[\frac{1}{f_{n,h_n}(\mathbf{x})} + f_{n,h_n}(\mathbf{x}) \right] d\mathbf{x}.
\end{aligned}
$$

Similarly as above, we get, for any $\mathbf{x} \in A_n$,

$$
\begin{aligned}
\frac{1}{f_{n,h_n}(\mathbf{x})} + f_{n,h_n}(\mathbf{x}) &= \left(\frac{1}{f_{n,h_n}(\mathbf{x}) f_{n,h_n}(\mathbf{x})} + 1 \right) f_{n,h_n}(\mathbf{x}) \\
&\leq \left(\frac{1}{\gamma_n^2} + 1 \right) f_{n,h_n}(\mathbf{x}).
\end{aligned}
$$

We can therefore write the following chain of inequalities, for any $n \geq 1$,

$$
\begin{aligned}
|\boldsymbol{\Delta}_{2,n,h_n}| & \\
&\leq \left(\frac{1}{\gamma_n^2} + 1 \right) \int_{A_n} |\mathbb{E} f_{n,h_n}(\mathbf{x}) - f_{n,h_n}(\mathbf{x})| \, f_{n,h_n}(\mathbf{x}) d\mathbf{x} \\
&\leq \left(\frac{1}{\gamma_n^2} + 1 \right) \sup_{\mathbf{x} \in A_n} |\mathbb{E} f_{n,h_n}(\mathbf{x}) - f_{n,h_n}(\mathbf{x})| \int_{A_n} f_{n,h_n}(\mathbf{x}) d\mathbf{x} \\
&\leq \left(\frac{1}{\gamma_n^2} + 1 \right) \sup_{\mathbf{x} \in A_n} |\mathbb{E} f_{n,h_n}(\mathbf{x}) - f_{n,h_n}(\mathbf{x})| \int_{\mathbb{R}^d} f_{n,h_n}(\mathbf{x}) d\mathbf{x}.
\end{aligned}
$$

In view of condition (K.4), by change of variables, we have

$$\int_{\mathbb{R}^d} f_{n,h}(\mathbf{x})d\mathbf{x} = 1.$$

Thus, for any $n \geq 1$, we have

$$|\mathbf{\Delta}_{2,n,h_n}| \leq \left(\frac{1}{\gamma_n^2} + 1\right) \sup_{\mathbf{x} \in \mathbb{R}^d} |\mathbb{E}f_{n,h_n}(\mathbf{x}) - f_{n,h_n}(\mathbf{x})| . \tag{4.3}$$

We now impose some slightly more general assumptions on the kernel $K(\cdot)$ than that of Theorem 2.1. Consider the class of functions

$$\mathcal{K} := \left\{ K((\mathbf{x} - \cdot)/h^{1/d}) : h > 0, \ \mathbf{x} \in \mathbb{R}^d \right\}.$$

For $\varepsilon > 0$, set $N(\varepsilon, \mathcal{K}) = \sup_Q N(\kappa\varepsilon, \mathcal{K}, d_Q)$, where the supremum is taken over all probability measures Q on $(\mathbb{R}^d, \mathcal{B})$, where $\mathcal{B}$ represents the σ-field of Borel sets of $\mathbb{R}^d$. Here, d_Q denotes the $L_2(Q)$-metric and $N(\kappa\varepsilon, \mathcal{K}, d_Q)$ is the minimal number of balls $\{g : d_Q(g, g') < \varepsilon\}$ of d_Q-radius ε needed to cover $\mathcal{K}$. We assume that $\mathcal{K}$ satisfies the following uniform entropy condition.

(K.6) for some $C > 0$ and $\nu > 0$,

$$N(\varepsilon, \mathcal{K}) \leq C\varepsilon^{-\nu}, 0 < \varepsilon < 1. \tag{4.4}$$

Finally, to avoid using outer probability measures in all of statements, we impose the following measurability assumption.

(K.7) $\mathcal{K}$ is a pointwise measurable class, that is, there exists a countable sub-class $\mathcal{K}_0$ of $\mathcal{K}$ such that we can find for any function $g \in \mathcal{K}$ a sequence of functions $\{g_m : m \geq 1\}$ in $\mathcal{K}_0$ for which

$$g_m(\mathbf{z}) \longrightarrow g(\mathbf{z}), \qquad \mathbf{z} \in \mathbb{R}^d.$$

Remark 4.1. Remark that condition (K.6) is satisfied whenever (K.1) holds, i.e., $K(\cdot)$ is of bounded variation on $\mathbb{R}^d$ (in the sense of Hardy and Kauser, see, e.g. [9, 47] and [31]). Condition (K.7) is satisfied whenever (K.2) holds, i.e., $K(\cdot)$ is right continuous (refer to [13] and [23] and the references therein).

By Theorem 1 of [23], whenever $K(\cdot)$ is measurable and satisfies (K.3-4-6-7), and when $f(\cdot)$ is bounded, we have for each $c > 0$, and for a suitable function $\Sigma(c)$, with probability 1,

$$\limsup_{n \to \infty} \sup_{cn^{-1}\log n \leq h \leq 1} \frac{\sqrt{nh}\|f_{n,h} - \mathbb{E}f_{n,h}\|_\infty}{\sqrt{\log(1/h) \vee \log\log n}} = \Sigma(c) < \infty, \tag{4.5}$$

which implies, in view of (4.2) and (4.3), that, with probability 1,

$$\limsup_{n \to \infty} \sup_{h_n \leq h < 1} \frac{\sqrt{nh\gamma_n^4}|\mathbf{\Delta}_{1,n,h}|}{\sqrt{(\log(1/h) \vee \log\log n)}} = 0, \tag{4.6}$$

and

$$\limsup_{n\to\infty} \sup_{h_n\leq h<1} \frac{\sqrt{nh\gamma_n^4}|\boldsymbol{\Delta}_{2,n,h}|}{\sqrt{(\log(1/h)\vee\log\log n)}} \leq \Upsilon(c). \tag{4.7}$$

Recalling (4.1), the proof of Theorem 2.1 is completed by combining (4.6) with (4.7). $\qquad\square$

Proof of Corollary 2.2.

Recall $A_n = \{\mathbf{x} : f_{n,h_n}(\mathbf{x}) \geq \gamma_n\}$ and let A_n^c the complement of A_n in $\mathbb{R}^d$ (i.e., $A_n^c = \{\mathbf{x} : f_{n,h_n}(\mathbf{x}) < \gamma_n\}$). Observe that

$$
\begin{aligned}
|f(\mathbf{x})| &\geq |f_{n,h_n}(\mathbf{x})| - |f_{n,h_n}(\mathbf{x}) - f(\mathbf{x})| \geq \gamma_n \\
&\quad + O\left(\sqrt{\frac{(\log(1/h_n')\vee\log\log n)}{nh_n'}}\right) + O(h_n''^{1/d}).
\end{aligned}
$$

Keep in mind that $|\log(h_n'')|/\log\log n \to \infty$ as $n \to \infty$, thus, for n enough large, the two last terms of the last inequality are dominated by the first one, then, we obtain

$$|f(\mathbf{x})| \geq \gamma_n.$$

We repeat the arguments above with the formal change of $H_{n,h_n}^{(1)}(f)$ by $H(f)$. We show that, for any $n \geq 1$,

$$
\begin{aligned}
|\widehat{\mathbb{E}}H_{n,h_n}^{(1)}&(f) - H(f)| \\
&\leq \left|\int_{A_n^c} f(\mathbf{x})\log\big(f(\mathbf{x})\big)d\mathbf{x}\right| \\
&\quad + \frac{1}{\gamma_n}\sup_{\mathbf{x}\in\mathbb{R}^d}|\mathbb{E}f_{n,h_n}(\mathbf{x}) - f(\mathbf{x})| \\
&\quad + \left(\frac{1}{\gamma_n^2} + 1\right)\sup_{\mathbf{x}\in\mathbb{R}^d}|\mathbb{E}f_{n,h_n}(\mathbf{x}) - f(\mathbf{x})|. \tag{4.8}
\end{aligned}
$$

It is obvious to see that

$$
\begin{aligned}
\int_{A_n^c} f(\mathbf{x})d\mathbf{x} &\leq \int_{\frac{1}{2}f(\mathbf{x})\leq\gamma_n} f(\mathbf{x})d\mathbf{x} + \int_{f_{n,h}(\mathbf{x})\leq\gamma_n\leq\frac{1}{2}f(\mathbf{x})} f(\mathbf{x})d\mathbf{x} \\
&\leq \int_{\frac{1}{2}f(\mathbf{x})\leq\gamma_n} f(\mathbf{x})d\mathbf{x} + 2\int_{\mathbb{R}^d}|f_{n,h}(\mathbf{x}) - f(\mathbf{x})|d\mathbf{x}.
\end{aligned}
$$

Observe that we have

$$\mathbb{1}_{\{\frac{1}{2}f(\mathbf{x})\leq\gamma_n\}}f(\mathbf{x}) \leq f(\mathbf{x})$$

and $\mathbb{1}_{\{\frac{1}{2}f(\mathbf{x})\leq\gamma_n\}}f(\mathbf{x}) \to 0$ as $n \to \infty$, thus an application of Lebesgue dominated convergence theorem gives

$$\lim_{n\to\infty} \int_{\frac{1}{2}f(\mathbf{x})\leq\gamma_n} f(\mathbf{x})d\mathbf{x} = 0. \tag{4.9}$$

Keep in mind that the conditions $h_n \to 0$ together with $nh_n \to \infty$ as $n \to \infty$, ensure that (see e.g., [15])

$$\lim_{n\to\infty} \int_{\mathbb{R}^d} |f_{n,h_n}(\mathbf{x}) - f(\mathbf{x})|d\mathbf{x} = 0 \qquad a.s.$$

We need the following instrumental fact due to [14, Lemma 3.3. p.40] and see also [35, Proof of Theorem 2.2] which for convenience and easy reference we state here.

Fact. Let $[h'_n, h''_n]$ be a sequence of deterministic interval, where $nh'_n \to \infty$ and $h''_n \to 0$, as $n \to \infty$. For every $\epsilon > 0$, then there exist $n_0 > 0$ and $r > 0$ such that

$$\mathbb{P}\left\{ \sup_{h'_n\leq h\leq h''_n} \int_{\mathbb{R}^d} |f_{n,h}(\mathbf{x}) - f(\mathbf{x})|d\mathbf{x} > \epsilon \right\} \leq \exp\left\{-rn\epsilon^2\right\}, \quad n \geq n_0.$$

A routine application of the Borel-Cantelli lemma implies, for all $h \in [h'_n, h''_n]$ such that $nh'_n \to \infty$ and $h''_n \to 0$, as $n \to \infty$, that

$$\lim_{n\to\infty} \sup_{h'_n\leq h\leq h''_n} \int_{\mathbb{R}^d} |f_{n,h}(\mathbf{x}) - f(\mathbf{x})|d\mathbf{x} = 0 \qquad a.s. \tag{4.10}$$

By combining (4.10) with (4.9), we obtain

$$\lim_{n\to\infty} \sup_{h'_n\leq h\leq h''_n} \int_{A_n^c} f(\mathbf{x})d\mathbf{x} = 0 \qquad a.s. \tag{4.11}$$

Since the entropy $H(f)$ is finite [by condition (F.1)], the measure

$$\nu(A) := \int_A |\log(f(\mathbf{x}))|d\mathbb{F}(\mathbf{x}),$$

is absolutely continuous with respect to the measure $\mu(A) = \int_A d\mathbb{F}(\mathbf{x})$, which guaranteed that

$$\lim_{n\to\infty} \sup_{h'_n\leq h\leq h''_n} \int_{A_n^c} f(\mathbf{x})\log(f(\mathbf{x}))d\mathbf{x} = 0 \qquad a.s. \tag{4.12}$$

Recall that we have for each $h'_n < h < h''_n$, as $n \to \infty$,

$$\|\mathbb{E}f_{n,h}(\mathbf{x}) - f(\mathbf{x})\|_\infty = O(h''^{1/d}_n). \tag{4.13}$$

Thus, we have

$$\lim_{n\to\infty} \sup_{h'_n \leq h \leq h''_n} \gamma_n^{-2} \|\mathbb{E} f_{n,h}(\mathbf{x}) - f(\mathbf{x})\|_\infty = 0.$$

This when combined with (4.8), entails that, as $n \to \infty$,

$$\sup_{h'_n \leq h \leq h''_n} \|\widehat{\mathbb{E}} H_{n,h}^{(1)}(f) - H(f)\| \to 0. \tag{4.14}$$

Using (4.11) and (4.14) in connection with (2.1) imply the desired conclusion (2.2). $\qquad\square$

Proof of Theorem 2.4.

Under conditions (F.2), (K.5) and using Taylor expansion of order s we get, for $\mathbf{x} \in \mathbb{I}$,

$$|\mathbb{E} f_{n,h}(\mathbf{x}) - f(\mathbf{x})| = \frac{h^{s/d}}{s!} \left| \int \sum_{k_1 + \cdots + k_d = s} t_1^{k_1} \ldots t_d^{k_d} \frac{\partial^s f(\mathbf{x} - h\theta\mathbf{t})}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}} K(\mathbf{t}) d\mathbf{t} \right|,$$

where $\theta = (\theta_1, \ldots, \theta_d)$ and $0 < \theta_i < 1$, $i =, 1, \ldots, d$. Thus a straightforward application of Lebesgue dominated convergence theorem gives, for n large enough,

$$\sup_{\mathbf{x} \in \mathbb{I}} |\mathbb{E} f_{n,h}(\mathbf{x}) - f(\mathbf{x})| = O(h_n''^{s/d}). \tag{4.15}$$

Let $\mathbb{J}$ be a nonempty compact subset of the interior of $\mathbb{I}$ (say $\mathring{\mathbb{I}}$). First, note that we have from Corollary 3.1.2. p. 62 of [46]

$$\limsup_{n\to\infty} \sup_{h'_n \leq h \leq h''_n} \sup_{\mathbf{x} \in \mathbb{J}} \frac{\sqrt{nh}|f_{n,h}(\mathbf{x}) - f(\mathbf{x})|}{\sqrt{\log(1/h) \vee \log\log n}} = \sup_{\mathbf{x} \in \mathbb{J}} \left\{ f(\mathbf{x}) \int_{\mathbb{R}^d} K^2(\mathbf{t}) d\mathbf{t} \right\}^{1/2} \tag{4.16}$$

Set, for all $n \geq 1$,

$$\pi_n(\mathbb{J}) = \left| \int_{\mathbb{J}} f_{n,h_n}(\mathbf{x}) \log \left(f_{n,h_n}(\mathbf{x}) \right) d\mathbf{x} - \int_{\mathbb{J}} f(\mathbf{x}) \log \left(f(\mathbf{x}) \right) d\mathbf{x} \right|. \tag{4.17}$$

Using condition (F.2) ($f(\cdot)$ is compactly supported), $f(\cdot)$ is bounded away from zero on its support, thus, we have for n enough large, there exists $\gamma > 0$, such that $f(\mathbf{x}) > \gamma$, for all $\mathbf{x}$ in the support of $f(\cdot)$. By the same previous arguments we have, for n enough large,

$$\begin{aligned} \pi_n(\mathbb{J}) &\leq \frac{1}{\gamma} \sup_{\mathbf{x} \in \mathbb{J}} |f_{n,h_n}(\mathbf{x}) - f(\mathbf{x})| \\ &\quad + \left(\frac{1}{\gamma^2} + 1\right) \sup_{\mathbf{x} \in \mathbb{J}} |f_{n,h_n}(\mathbf{x}) - f(\mathbf{x})|. \end{aligned}$$

One finds, by combining the last equation with (4.16),

$$\limsup_{n\to\infty} \sup_{h'_n\leq h\leq h''_n} \frac{\sqrt{nh}\ \pi_n(\mathbb{J})}{\sqrt{\{(\log(1/h)\vee\log\log n)}}$$
$$\leq\ \left(\frac{\gamma^2+\gamma+1}{\gamma^2}\right)^{1/2}\sup_{\mathbf{x}\in\mathbb{J}}\left\{f(\mathbf{x})\int_{\mathbb{R}^d}K^2(\mathbf{t})d\mathbf{t}\right\}^{1/2}. \qquad (4.18)$$

Let $\{\mathbb{J}_\ell\}$, $\ell=1,2,\ldots$, be a sequence of nondecreasing nonempty compact subsets of $\mathring{\mathbb{I}}$ such that

$$\bigcup_{\ell\geq 1}\mathbb{J}_\ell=\mathbb{I}.$$

Now, from (4.18), it is straightforward to observe that

$$\lim_{\ell\to\infty}\limsup_{n\to\infty}\sup_{h'_n\leq h\leq h''_n}\frac{\sqrt{nh\gamma_n^4}\pi_n(\mathbb{J}_\ell)}{\sqrt{(\log(1/h)\vee\log\log n)}}$$
$$\leq\ \lim_{\ell\to\infty}\left(\frac{\gamma^2+\gamma+1}{\gamma^2}\right)^{1/2}\sup_{\mathbf{x}\in\mathbb{J}_\ell}\left\{f(\mathbf{x})\int_{\mathbb{R}^d}K^2(\mathbf{t})d\mathbf{t}\right\}^{1/2}$$
$$\leq\ \left(\frac{\gamma^2+\gamma+1}{\gamma^2}\right)^{1/2}\sup_{\mathbf{x}\in\mathbb{I}}\left\{f(\mathbf{x})\int_{\mathbb{R}^d}K^2(\mathbf{t})d\mathbf{t}\right\}^{1/2}.$$

The proof of Theorem 2.4 is completed. $\qquad\qquad\qquad\qquad\qquad\square$

Proof of Theorem 2.6.

Let $\varphi_{n,h_n}(\mathbf{x}):=\mathbb{E}(f_{n,h_n}(\mathbf{x}))$. Recall that

$$H^{(2)}_{n,h_n}(f)-\widehat{\mathbb{E}}H^{(2)}_{n,h_n}(f)\ =\ -\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\Omega_{n,i}}\log(f_{n,h_n}(\mathbf{X}_i))+\mathbb{1}_{\Omega_{n,i}}\log\left(\varphi_{n,h_n}(\mathbf{X}_i)\right)$$
$$=:\ \mathbf{\Xi}_{n,h_n}.$$

Using a Taylor-Lagrange expansion of the $\log(\cdot)$ function, we have, for some random sequence $\theta_n\in(0,1)$,

$$\Xi_{n,h_n}=\frac{1}{n}\sum_{i=1}^n\mathbb{1}_{\Omega_{n,i}}\left[\frac{f_{n,h_n}(\mathbf{X}_i)-\varphi_{n,h_n}(\mathbf{X}_i)}{(1-\theta_n)f_{n,h_n}(\mathbf{X}_i)+\theta_n\varphi_{n,h_n}(\mathbf{X}_i)}\right].$$

Recalling that $\Omega_{n,i}=\left\{f_{n,h_n}(\mathbf{X}_i)\geq\gamma_n\right\}$, we readily obtain, with probability 1,

$$|\Xi_{n,h_n}|\ \leq\ \frac{1}{n\gamma_n}\sum_{i=1}^n\mathbb{1}_{\Omega_{n,i}}\left|f_{n,h_n}(\mathbf{X}_i)-\varphi_{n,h_n}(\mathbf{X}_i)\right|$$
$$\leq\ \frac{1}{\gamma_n}\sup_{\mathbf{x}\in\mathbb{I}}\left|f_{n,h_n}(\mathbf{x})-\varphi_{n,h_n}(\mathbf{x})\right|$$
$$=\ \frac{1}{\gamma_n}\sup_{\mathbf{x}\in\mathbb{I}}\left|f_{n,h_n}(\mathbf{x})-\mathbb{E}(f_{n,h_n}(\mathbf{x}))\right|.$$

Combining the last inequality with (4.5), we readily obtain the desired result.
$$\square$$

Proof of Theorem 2.7.

We have

$$H^{(2)}_{n,h_n}(f) - H(f) = \{H^{(2)}_{n,h_n}(f) - \widehat{\mathbb{E}}H^{(2)}_{n,h_n}(f)\} + \{\widehat{\mathbb{E}}H^{(2)}_{n,h_n}(f) - H(f)\}.$$

Since the first term in the right hand of the last equality is controlled in the preceding proof, it remains only to evaluate the second one. To simplify our exposition, we will decompose $\widehat{\mathbb{E}}H^{(2)}_{n,h_n}(f) - H(f)$ into the sum of three components, that is

$$
\begin{aligned}
\widehat{\mathbb{E}}H^{(2)}_{n,h_n}(f) - H(f) &= -\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\Omega_{n,i}} \log(\varphi_{n,h_n}(\mathbf{X}_i)) + \mathbb{E}\left(\log\left(f(\mathbf{X}_i)\right)\right) \\
&= -\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\Omega_{n,i}} \left(\log(\varphi_{n,h_n}(\mathbf{X}_i)) - \log(f(\mathbf{X}_i))\right) \\
&\quad -\frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{1}_{\Omega_{n,i}} \log(f(\mathbf{X}_i)) - \log(f(\mathbf{X}_i))\right) \\
&\quad -\frac{1}{n}\sum_{i=1}^{n} \left(\log(f(\mathbf{X}_i)) - \mathbb{E}\left(\log(f(\mathbf{X}_i))\right)\right) \\
&=: -\boldsymbol{\nabla}_{1,n,h_n} - \boldsymbol{\nabla}_{2,n,h_n} - \boldsymbol{\nabla}_{3,n,h_n}. \qquad (4.19)
\end{aligned}
$$

In view of (4.19), we have

$$\boldsymbol{\nabla}_{1,n,h_n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\Omega_{n,i}} \left(\log(\varphi_{n,h}(\mathbf{X}_i)) - \log(f(\mathbf{X}_i))\right).$$

Using again a Taylor-Lagrange expansion of the $\log(\cdot)$ function, we have, for some random sequence $\theta_n \in (0,1)$,

$$\boldsymbol{\nabla}_{1,n,h_n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\Omega_{n,i}} \left[\frac{\varphi_{n,h_n}(\mathbf{X}_i) - f(\mathbf{X}_i)}{(1-\theta_n)\varphi_{n,h_n}(\mathbf{X}_i) + \theta_n f(\mathbf{X}_i)}\right].$$

By condition (F.2), there exists a constant $\eta_{\mathbb{I}} > 0$, such that $f(\mathbf{x}) > \eta_{\mathbb{I}}$ for all $\mathbf{x} \in \mathbb{I}$. It follows that for n enough large that, $f(\mathbf{x}) > \gamma_n$ for all $\mathbf{x} \in \mathbb{I}$. Recalling that $\Omega_{n,i} = \{f_{n,h_n}(\mathbf{X}_i) \geq \gamma_n\}$, we readily obtain, with probability 1,

$$
\begin{aligned}
|\boldsymbol{\nabla}_{1,n,h_n}| &\leq \frac{1}{n\gamma_n}\sum_{i=1}^{n} \mathbb{1}_{\Omega_{n,i}} |\varphi_{n,h_n}(\mathbf{X}_i) - f(\mathbf{X}_i)| \\
&\leq \frac{1}{\gamma_n}\sup_{\mathbf{x}\in\mathbb{I}} |\varphi_{n,h_n}(\mathbf{x}) - f(\mathbf{x})|.
\end{aligned}
$$

We mention that the bandwidth h is to be chosen in such a way that the bias of $f_{n,h}(\mathbf{x})$ may be neglected, in the sense that

$$\lim_{n\to\infty} \sup_{h'_n \leq h \leq h''_n} \left\{\frac{nh}{2\log(1/h)}\right\}^{1/2} \sup_{\mathbf{x}\in\mathbb{I}} \left|\varphi_{n,h}(\mathbf{x}) - f(\mathbf{x})\right| = 0, \qquad (4.20)$$

which is implied by (4.15). Thus,

$$\limsup_{n \to \infty} \sup_{h'_n \le h \le h''_n} \frac{\sqrt{nh\gamma_n^2}|\boldsymbol{\nabla}_{1,n,h}|}{\sqrt{2\log(1/h)}} = 0. \tag{4.21}$$

We next evaluate the second term $\boldsymbol{\nabla}_{2,n,h_n}$ in the right side of (4.19). We have from (4.15) and (4.5)

$$\sup_{h'_n \le h \le h''_n} \sup_{\mathbf{x} \in \mathbb{I}} \left| f_{n,h}(\mathbf{x}) - f(\mathbf{x}) \right| = O\left(\sqrt{\frac{(\log(1/h'_n)}{nh'_n}} \right).$$

Thus, for n sufficiently large, almost surely, $f_{n,h}(\mathbf{x}) \ge (1/2)f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{I}$ and all $h \in [h'_n, h''_n]$. Note that under condition (F.2), the density $f(\cdot)$ is compactly supported, it is possible to find a positive constant $\eta_{\mathbb{I}}$ such as $f(\mathbf{x}) > \eta_{\mathbb{I}}$. This implies that $f_{n,h}(\mathbf{x}) \ge \eta_{\mathbb{I}}/2$, and thus, for all n enough large, we have, almost surely,

$$\mathbb{1}_{\Omega_{n,i}} = 1, \tag{4.22}$$

which implies that, for all n enough large, almost surely,

$$\boldsymbol{\nabla}_{2,n,h_n} = 0. \tag{4.23}$$

We finally evaluate the second term $\boldsymbol{\nabla}_{3,n,h_n}$ in the right side of (4.19). We have,

$$\boldsymbol{\nabla}_{3,n,h_n} = -\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\xi}_i,$$

where, for $i = 1, \ldots, n$,

$$\boldsymbol{\xi}_i := \log\{f(\mathbf{X}_i)\} - \mathbb{E}\Big(\log\{f(\mathbf{X}_i)\} \Big),$$

are a centered independent and identically distributed random variables with finite variance $\mathrm{Var}\big(\log(f(\mathbf{X}_i)) \big)$ (condition (F.3)). Observe that

$$\frac{\gamma_n}{n} \frac{\sqrt{nh_n} \sum_{i=1}^{n} \boldsymbol{\xi}_i}{\sqrt{2\log(1/h_n)}} = \frac{\gamma_n \sqrt{h_n \log \log n}}{\sqrt{\log(1/h_n)}} \frac{\sum_{i=1}^{n} \boldsymbol{\xi}_i}{\sqrt{2n \log \log n}}$$

which, by the law of the iterated logarithm, tends to 0 as n tends to infinity. Namely,

$$\lim_{n \to \infty} \sup_{h'_n \le h \le h''_n} \frac{\sqrt{nh\gamma_n^2}|\boldsymbol{\nabla}_{3,n,h}|}{\sqrt{2\log(1/h)}} = 0. \tag{4.24}$$

Using (4.24) and (4.23) in connection with (4.16) completes the proof of Theorem 2.7. $\square$

Acknowledgments

References

[1] AHMAD, I. A. and LIN, P. E. (1976). A nonparametric estimation of the entropy for absolutely continuous distributions. *IEEE Trans. Information Theory*, **IT-22**(3), 372–375. MR0408987

[2] AKAIKE, H. (1954). An approximation to the density function. *Ann. Inst. Statist. Math., Tokyo*, **6**, 127–132. MR0067412

[3] ASH, R. (1965). *Information theory*. Interscience Tracts in Pure and Applied Mathematics, No. 19. Interscience Publishers John Wiley & Sons, New York-London-Sydney. MR0229475

[4] BEIRLANT, J., DUDEWICZ, E. J., GYÖRFI, L., and VAN DER MEULEN, E. C. (1997). Nonparametric entropy estimation: an overview. *Int. J. Math. Stat. Sci.*, **6**(1), 17–39. MR1471870

[5] BERGER, T. (1971). *Rate distortion theory*. Prentice-Hall Inc., Englewood Cliffs, N. J. A mathematical basis for data compression, Prentice-Hall Series in Information and System Sciences. MR0408988

[6] BOSQ, D. and LECOUTRE, J.-P. (1987). *Théorie de l'estimation fonctionnelle*. Économie et Statistiques Avancées. Economica, Paris.

[7] BOUZEBDA, S. and ELHATTAB, I. (2009). A strong consistency of a nonparametric estimate of entropy under random censorship. *C. R. Math. Acad. Sci. Paris*, **347**(13-14), 821–826. MR2543991

[8] BOUZEBDA, S. and ELHATTAB, I. (2010). Uniform in bandwidth consistency of the kernel-type estimator of the shannon's entropy. *C. R. Math. Acad. Sci. Paris*, **348**(5-6), 317–321. MR2600131

[9] CLARKSON, J. A. and ADAMS, C. R. (1933). On definitions of bounded variation for functions of two variables. *Trans. Amer. Math. Soc.*, **35**(4), 824–854. MR1501718

[10] COVER, T. M. and THOMAS, J. A. (2006). *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition. MR2239987

[11] CSISZÁR, I. (1962). Informationstheoretische Konvergenzbegriffe im Raum der Wahrscheinlichkeitsverteilungen. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, **7**, 137–158. MR0191733

[12] DEHEUVELS, P. (2000). Uniform limit laws for kernel density estimators on possibly unbounded intervals. In *Recent advances in reliability theory (Bordeaux, 2000)*, Stat. Ind. Technol., pages 477–492. Birkhäuser Boston, Boston, MA. MR1783500

[13] DEHEUVELS, P. and MASON, D. M. (2004). General asymptotic confidence bands based on kernel-type function estimators. *Stat. Inference Stoch. Process.*, **7**(3), 225–277. MR2111291

[14] DEVROYE, L. (1987). *A Course in density estimation.* Birkhäuser, Boston-Basel-Stuttgart. MR0891874

[15] DEVROYE, L. and GYÖRFI, L. (1985). *Nonparametric density estimation.* Wiley Series in Probability and Mathematical Statistics: Tracts on Probability and Statistics. John Wiley & Sons Inc., New York. The L_1 view. MR0780746

[16] DEVROYE, L. and LUGOSI, G. (2001). *Combinatorial methods in density estimation.* Springer Series in Statistics. Springer-Verlag, New York. MR1843146

[17] DEVROYE, L. and WISE, G. L. (1980). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM J. Appl. Math.*, **38**(3), 480–488. MR0579432

[18] DMITRIEV, J. G. and TARASENKO, F. P. (1973). The estimation of functionals of a probability density and its derivatives. *Teor. Verojatnost. i Primenen.*, **18**, 662–668. MR0359157

[19] DUDEWICZ, E. J. and VAN DER MEULEN, E. C. (1981). Entropy-based tests of uniformity. *J. Amer. Statist. Assoc.*, **76**(376), 967–974. MR0650913

[20] EBRAHIMI, N., HABIBULLAH, M., and SOOFI, E. (1992). Testing exponentiality based on Kullback-Leibler information. *J. Roy. Statist. Soc. Ser. B*, **54**(3), 739–748. MR1185219

[21] EGGERMONT, P. P. B. and LARICCIA, V. N. (1999). Best asymptotic normality of the kernel density entropy estimator for smooth densities. *IEEE Trans. Inform. Theory*, **45**(4), 1321–1326. MR1686275

[22] EINMAHL, U. and MASON, D. M. (2000). An empirical process approach to the uniform consistency of kernel-type function estimators. *J. Theoret. Probab.*, **13**(1), 1–37. MR1744994

[23] EINMAHL, U. and MASON, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.*, **33**(3), 1380–1403. MR2195639

[24] ESTEBAN, M. D., CASTELLANOS, M. E., MORALES, D., and VAJDA, I. (2001). Monte Carlo comparison of four normality tests using different entropy estimates. *Comm. Statist. Simulation Comput.*, **30**(4), 761–785. MR1878453

[25] GALLAGER, R. (1968). *Information theory and reliable communication.* New York-London-Sydney-Toronto: John Wiley & Sons, Inc. XVI, 588 p.

[26] GINÉ, E. and MASON, D. M. (2008). Uniform in bandwidth estimation of integral functionals of the density function. *Scand. J. Statist.*, **35**(4), 739–761. MR2468873

[27] GOKHALE, D. (1983). On entropy-based goodness-of-fit tests. *Comput. Stat. Data Anal.*, **1**, 157–165.

[28] GOKHALE, D. V. and KULLBACK, S. (1978). *The information in contingency tables*, volume 23 of *Statistics: Textbooks and Monographs.* Marcel Dekker Inc., New York. MR0524727

[29] GYÖRFI, L. et VAN DER MEULEN, E. C. (1990). An entropy estimate based on a kernel density estimation. In *Limit theorems in probability and*

statistics (Pécs, 1989), volume 57 of *Colloq. Math. Soc. János Bolyai*, pages 229–240. North-Holland, Amsterdam. MR1116790

[30] GYÖRFI, L. and VAN DER MEULEN, E. C. (1991). On the nonparametric estimation of the entropy functional. In *Nonparametric functional estimation and related topics (Spetses, 1990)*, volume 335 of *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.*, pages 81–95. Kluwer Acad. Publ., Dordrecht. MR1154321

[31] HOBSON, E. W. (1958). *The theory of functions of a real variable and the theory of Fourier's series. Vol. I.* Dover Publications Inc., New York, N.Y.

[32] JAYNES, E. T. (1957). Information theory and statistical mechanics. *Phys. Rev. (2)*, **106**, 620–630. MR0087305

[33] KULLBACK, S. (1959). *Information theory and statistics.* John Wiley and Sons, Inc., New York. MR0103557

[34] LAZO, A. C. and RATHIE, P. N. (1978). On the entropy of continuous probability distributions. *IEEE Trans. Inf. Theory*, **24**, 120–122. MR0479686

[35] LOUANI, D. (2005). Uniform L_1-distance large deviations in nonparametric density estimation. *Test*, **14**(1), 75–98. MR2203423

[36] PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, **33**, 1065–1076. MR0143282

[37] PRAKASA RAO, B. L. S. (1983). *Nonparametric functional estimation.* Probability and Mathematical Statistics. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York. MR0740865

[38] PRESCOTT, P. (1976). On a test for normality based on sample entropy. *J. R. Stat. Soc., Ser. B*, **38**, 254–256. MR0420958

[39] RÉNYI, A. (1959). On the dimension and entropy of probability distributions. *Acta Math. Acad. Sci. Hungar.*, **10**, 193–215 (unbound insert). MR0107575

[40] ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, **27**, 832–837. MR0079873

[41] SCOTT, D. W. (1992). *Multivariate density estimation.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York. Theory, practice, and visualization, A Wiley-Interscience Publication. MR1191168

[42] SHANNON, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.*, **27**, 379–423, 623–656. MR0026286

[43] SONG, K.-S. (2000). Limit theorems for nonparametric sample entropy estimators. *Statist. Probab. Lett.*, **49**(1), 9–18. MR1789659

[44] Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *Ann. Probab.*, **22**(1), 28–76. MR1258865

[45] VASICEK, O. (1976). A test for normality based on sample entropy. *J. Roy. Statist. Soc. Ser. B*, **38**(1), 54–59. MR0420958

[46] VIALLON, V. (2006). *Processus empiriques, estimation non paramétrique et données censurées.* Ph.D. thesis, Université Paris 6.

[47] VITUŠKIN, A. G. (1955). *O mnogomernyh variaciyah.* Gosudarstv. Izdat. Tehn.-Teor. Lit., Moscow. MR0075267