

- [4] I. Cížova, "Test of a histogram estimator for the differential entropy," Master's thesis, Czech Tech. Univ. (ČVUT), Prague, 1997 (in Czech).
- [5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [6] G. A. Darbellay, "An adaptive histogram estimator for the mutual information," Res. Rep. 1889, UTIA, Academy of Sciences, Prague, Czech Republic, 1996. Also, *Computat. Statist. Data Anal.*, to be published.
- [7] —, "The mutual information as a measure of statistical dependence," in *Proc. Int. Symp. Information Theory* (Ulm, Germany, June 29–July 4, 1997). Piscataway, NJ: IEEE Press, 1997, p. 405.
- [8] —, "Predictability: An information-theoretic perspective," in *Signal Analysis and Prediction*, A. Procházka, J. Uhlř, P. J. W. Rayner, and N. G. Kingsbury, Eds. Boston, MA: Birkhäuser-Verlag, 1998, pp. 249–262.
- [9] —, "Statistical dependences in  $\mathbb{R}^d$ : An information-theoretic approach," in *Proc. 3rd European IEEE Workshop Computationally Intensive Methods in Control and Data Processing* (Prague, Czech Republic, Sept. 7–9, 1998). Available via e-mail at library@utia.cas.cz.
- [10] G. A. Darbellay and I. Vajda, "Entropy expressions for multivariate continuous distributions," Res. Rep. 1920, UTIA, Academy of Sciences, Prague, Czech Republic, 1998. Available via e-mail at library@utia.cas.cz.
- [11] —, "Estimation of the mutual information with data-dependent partitions," Res. Rep. 1921, UTIA, Academy of Sciences, Prague, Czech Republic, 1998. Available via e-mail at library@utia.cas.cz.
- [12] R. L. Dobrushin, "General formulation of Shannon's main theorem in information theory," *Usp. Mat. Nauk*, vol. 14, pp. 3–104, 1959 (in Russian). Translated in *Amer. Math. Soc. Trans.*, vol. 33, pp. 323–438.
- [13] L. Györfi and E. C. van der Meulen, "Density-free convergence properties of various estimators of entropy," *Comput. Statist. Data Anal.*, vol. 5, pp. 425–436, 1987.
- [14] F. Liese and I. Vajda, *Convex Statistical Distances*. Leipzig, Germany: Teubner, 1987.
- [15] D. W. Scott, *Multivariate Density Estimation*. New York: Wiley, 1992.
- [16] Y. Shao and M. G. Hahn, "Limit theorems for the logarithm of sample spacings," *Statist. Probab. Lett.*, vol. 24, pp. 121–132, 1995.

## Best Asymptotic Normality of the Kernel Density Entropy Estimator for Smooth Densities

Paul P. B. Eggermont and Vincent N. LaRiccia

**Abstract**—In the random sampling setting we estimate the entropy of a probability density distribution by the entropy of a kernel density estimator using the double exponential kernel. Under mild smoothness and moment conditions we show that the entropy of the kernel density estimator equals a sum of independent and identically distributed (i.i.d.) random variables plus a perturbation which is asymptotically negligible compared to the parametric rate  $n^{-1/2}$ . An essential part in the proof is obtained by exhibiting almost sure bounds for the Kullback–Leibler divergence between the kernel density estimator and its expected value. The basic technical tools are Doob's submartingale inequality and convexity (Jensen's inequality).

**Index Terms**—Convexity, entropy estimation, kernel density estimators, Kullback–Leibler divergence, submartingales.

### I. INTRODUCTION

Let  $X_1, X_2, \dots, X_n$  be independent and identically distributed (i.i.d.) univariate random variables, with common probability density function  $g(x)$ . Let  $g^{nh}$  be the kernel density estimator

$$g^{nh}(x) = \frac{1}{n} \sum_{i=1}^n s_h(x - X_i), \quad x \in \mathbb{R} \quad (1.1)$$

with the *double exponential kernel*  $s_h(x) = (2h)^{-1} \exp(-h^{-1}|x|)$ . We are interested for practical and theoretical reasons in the estimation of the negative entropy of  $g$

$$H(g) = \int_{\mathbb{R}} g(x) \log g(x) dx \quad (1.2)$$

by the *natural estimator*  $H(g^{nh})$  with  $h \asymp n^{-\beta}$ , for some  $\beta$  with  $\frac{1}{4} < \beta < \frac{1}{2}$ , depending on the smoothness and decay of  $g$ . For some practical applications, see Györfi and van der Meulen [12], Joe [16], and references therein. Our interest in the entropy estimation problem ties in with our attempt at understanding likelihood discrepancy principles for the automatic selection of the window parameter in nonparametric deconvolution problems, see Eggermont and LaRiccia [8].

Under suitable assumptions we prove that

$$H(g^{nh}) = \frac{1}{n} \sum_{i=1}^n \log g(X_i) + \varepsilon_{nh} \quad (1.3)$$

with  $\varepsilon_{nh} = o(n^{-1/2})$  almost surely.

The conditions on  $g$  involve smoothness and that  $g$  has a finite moment of order  $>2$ . If  $\mathbb{E}[\{\log g\}^2] < \infty$  then the asymptotic normality of  $H(g^{nh})$  is assured by the central limit theorem,

$$\sqrt{n} \{H(g^{nh}) - H(g)\} \rightarrow_d Y \sim N(0, \text{Var}[\log g]) \quad (1.4)$$

Manuscript received March 13, 1996; revised November 24, 1998. The material in this correspondence was presented in part at the Conference on Nonparametric Function Estimation, Montreal, Que., Canada, October 13–24, 1997.

The authors are with the Department of Mathematical Sciences, University of Delaware, Newark, DE 19716 USA.

Communicated by K. Zeger, Associate Editor at Large.

Publisher Item Identifier S 0018-9448(99)03764-5.

as is the law of the iterated logarithm concerning its almost sure behavior

$$\limsup_{n \rightarrow \infty} (n/\log \log n)^{1/2} |H(g^{nh}) - H(g)| =_{\text{as}} \{2 \text{Var}[\log g]\}^{1/2}. \quad (1.5)$$

Moreover, it is easily verified that then  $H(g^{nh})$  is best asymptotically normal, see Levit [17], Tsybakov and van der Meulen [20].

The result (1.3) brings to mind the estimator of Ahmad and Lin [1], given by

$$\tilde{H}(g^{nh}) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \log g^{nh}(X_i). \quad (1.6)$$

It is rather surprising that with a little extra work we get best asymptotic normality for this estimator as well, using the double exponential kernel. Of course, this estimator is much easier to calculate than  $H(g^{nh})$ . It remains to be seen (in simulations) what the small sample consequences are.

On the practical side it should be mentioned that the actual calculation of  $H(g^{nh})$  for a fixed realization of  $X_1, X_2, \dots, X_n$  and a fixed value of  $h$  requires some care, but is otherwise “just” a problem of numerical integration, i.e., it does not involve any stochastic aspects. In this and other problems we would recommend the approximation of  $g^{nh}$  by piecewise-constant functions on a fine enough partition, as in Györfi and van der Meulen [11]. The approximations are easily calculated since the distribution function corresponding to the kernel  $s_h$  is available. An interesting aspect of the estimator  $H(g^{nh})$  is that the window parameter  $h$  is not very influential since it only affects the  $o(n^{-1/2})$  term. In particular, data-driven procedures for choosing  $h$  ought to be possible, see also Hall and Morton [15].

Entropy estimators based on spacings enjoy popularity as well, see van Es [21], Tsybakov and van der Meulen [20], and references therein. Asymptotically normal estimators of entropy of the form (1.1), (1.6), or based on spacings have been exhibited before but usually under very strong tail conditions, e.g., under what appears to be exponential decay, see Joe [16] and Tsybakov and van der Meulen [20], and best asymptotically normal estimators under the assumption that the density has compact support, and is bounded away from zero on its support, see Hall [13], van Es [21]. For a thorough review of the issues in entropy estimation, in particular also (the lack of) the reasonability of various assumptions, see Beirlant, Dudewicz, Györfi, and van der Meulen [3].

The principal tool for the proof of (1.3) is provided by submartingales and submartingale inequalities, see Breiman [4]. This comes about as follows. We let  $G$  denote the cumulative distribution function corresponding to the probability density function (pdf)  $g$ , and let  $G_n$  denote the empirical distribution function corresponding to the observations  $X_1, X_2, \dots, X_n$ . Recall the definition of the Kullback–Leibler information divergence between two pdf's  $\varphi, \psi$ , with  $\psi > 0$  whenever  $\varphi > 0$ , defined as

$$D(\varphi, \psi) = \int_{\mathbf{R}} \left\{ \varphi(y) \log \frac{\varphi(y)}{\psi(y)} + \psi(y) - \varphi(y) \right\} dy \quad (1.7)$$

(the term  $\psi(y) - \varphi(y)$  integrates to zero, but including it into (1.7) has the advantage that then the integrand is nonnegative) and introduce the notation  $F_{nh}$  as shorthand for

$$F_{nh} = \int_{\mathbf{R}} (s_h * \log g_h - \log g) dG_n \quad (1.8)$$

where  $*$  denotes convolution, and  $g_h = s_h * dG$  is the expected value of  $g^{nh} = s_h * dG_n$ . It is implicit that  $g_h$  and  $g^{nh}$  are both pdf's. It is straightforward to verify, using (1.3) as the definition of  $\varepsilon_{nh}$ , that

$$\varepsilon_{nh} = D(g^{nh}, g_h) + F_{nh}. \quad (1.9)$$

Obviously, the last term (1.8) of (1.9), properly scaled, is a martingale, and so by the (sub)martingale inequality its almost sure behavior is pretty much determined by the behavior of the expected absolute value. The Kullback–Leibler divergence  $D(g^{nh}, g_h)$  is more cumbersome. We show that after proper scaling it is dominated by a pair of submartingales, whose expected values are under control, and so their (approximate) almost sure rates of decay are under control as well. It might be possible to apply the same ideas directly to  $H(g^{nh})$ , but the route via the Kullback–Leibler distance is interesting in its own right.

In the above the choice of the double exponential kernel is maybe not essential, but it certainly makes some of the details come out quite elegantly. On the one hand,  $s_h(x)$  is the Green's function of a boundary value problem for a second-order differential equation. If  $\varphi$  is a continuous probability density function, then  $\psi = s_h * \varphi$  is also a pdf which satisfies

$$\begin{aligned} -h^2 \psi''(x) + \psi(x) &= \varphi(x), & x \in \mathbf{R} \\ \psi &\rightarrow 0, & x \rightarrow \pm\infty. \end{aligned} \quad (1.10)$$

On the other hand, the collection  $\{s_h\}_{h>0}$  has a nice semigroup-like property, which is useful in eliminating the blocking phenomenon that customarily arises when applying submartingale inequalities to obtain almost sure bounds. The double exponential is no stranger to density estimation: it also arises via the roughness penalization approach of Good and Gaskins [10], see Thompson and Tapia [19], and Eggermont and LaRiccia [9].

## II. ASSUMPTIONS AND THEOREMS

We assume that  $g$  has a finite moment of order  $>2$ , that is, there exist  $M > \kappa > 2$  such that

$$\mathbb{E}[|X|^M] < \infty. \quad (2.1)$$

(The reason for introducing  $\kappa$  will become clear later.) The smoothness assumptions on  $g$  involve what we call a finite Fisher information numbers, viz.,  $g$  is twice differentiable with

$$\int_{\mathbf{R}} \frac{|g'(x)|^2}{g(x)} dx < \infty \quad \int_{\mathbf{R}} \frac{|g''(x)|^2}{g(x)} dx < \infty. \quad (2.2)$$

These conditions do not have much to do with decay, as the example  $g(x) = x^2 + 1$  shows.

*Main Theorem 1:* Under the Assumptions (2.1) and (2.2), for all  $s > 1$

$$H(g^{nh}) = \int_{\mathbf{R}} \log g dG_n + \varepsilon_{nh}$$

with

$$\varepsilon_{nh} = O\left(n^{-2\kappa/(3\kappa+2)} (\log n)^s\right)$$

almost surely, provided  $h = h_n \asymp n^{-\kappa/(3\kappa+2)}$ . In particular then  $\varepsilon_{nh} = o(n^{-1/2})$  almost surely.

The ideal proof would be to suitably bound the terms in the expression (1.9) for  $\varepsilon_{nh}$ . With foresight/hindsight it is necessary to consider

$$\delta_{nh} = \sup_{\lambda \in [h/2, h]} |\varepsilon_{n, \lambda}|, \quad (2.3)$$

where  $h$  will eventually be chosen to depend on  $n$ . Dealing with the supremum in (2.3) is easy since  $H(g^{nh})$  is decreasing in  $h > 0$ , as the following lemma shows.

*Lemma 1:* For  $\Phi, \Psi$  distribution functions,  $H(s_h * d\Psi)$  and  $D(s_h * \Psi, s_h * d\Phi)$  are decreasing functions of  $h$ .

It follows that  $\varepsilon_{nh}$  is decreasing in  $h$ , whence

$$\delta_{nh} = \max\{|\varepsilon_{nh}|, |\varepsilon_{n,h/2}|\} \quad (2.4)$$

and so to bound  $\delta_{nh}$  it suffices to study  $\varepsilon_{nh}$  for fixed values of  $h$ . The required bounds for  $\varepsilon_{nh}$  are given via (1.9) by the following lemmas.

*Lemma 2:* Under the Assumptions (2.1) and (2.2), we have for every  $s > 1$

$$D(g^{nh}, g_h) =_{\text{as}} \mathcal{O}((nh)^{-\kappa/(\kappa+1)} (\log n)^s)$$

provided  $h = h_n \asymp n^{-\beta}$  for some  $0 < \beta < 1$ .

*Lemma 3:* Under the Assumptions (2.2), we have for every  $s > 1$

$$F_{nh} =_{\text{as}} \mathcal{O}(h^2 (\log n)^s)$$

provided  $h = h_n \asymp n^{-\beta}$  for some  $0 < \beta < 1$ , and  $h_n$  is constant on blocks  $2^{k-1} < n \leq 2^k$  ( $k \in \mathbb{N}$ ).

We emphasize again that the blocking phenomenon disappears by virtue of (2.4).

Combining these results with  $h_n \asymp n^{-\kappa/(3\kappa+2)}$  proves that for all  $r > 1$

$$\delta_{nh} =_{\text{as}} \mathcal{O}(n^{-2\kappa/(3\kappa+2)} (\log n)^r). \quad (2.5)$$

Since  $\kappa > 2$ , then the Main Theorem 1 is proven.

Results for the estimator of Ahmad and Lin [1] analogous to the Main Theorem are discussed in Section VII.

We finish with some general comments. Note that Lemma 2 does not deal with  $D(g^{nh}, g)$ . Indeed,  $D(g^{nh}, g_h)$  is much better behaved than  $D(g^{nh}, g)$ , see Barron *et al.* [2], Hall [14]. Also note that the result of Lemma 2 is perhaps not what one would hope for. Ideally, one would like under minimal conditions that  $D(g^{nh}, g_h)$  behaves like  $(nh)^{-1}$ . We can prove the bound  $\mathcal{O}((nh)^{-1} \log(nh))$ , but only under the condition

$$\mathbb{E}[e^{r|X|}] < \infty, \quad \text{for some } r > 0. \quad (2.6)$$

This condition implies that  $g$  decays exponentially, and so is quite strong. We omit the details.

### III. MONOTONICITY IN THE SMOOTHING PARAMETER

In this section we prove Lemma 1, regarding the monotone behavior in  $h$ . The main ingredient of the proof is the convexity of  $H(\varphi)$  in  $\varphi$ , and the convexity of  $D(\varphi, \psi)$  in  $\varphi$  and  $\psi$  jointly, see Devroye and Györfi [6], combined with the semigroup-like property of  $s_h$ , see (3.1) below.

*Proof of Lemma 1:* It is easily verified using Fourier transforms that

$$s_\lambda = (h/\lambda)^2 s_h + [1 - (h/\lambda)^2] s_\lambda * s_h \quad (3.1)$$

and hence that for  $\lambda > h$  the right-hand side is a convex combination. Since the function  $\text{nent}(t) = t \log t$  is convex, applying Jensen's inequality (twice) to  $\text{nent}([s_\lambda * d\Psi](x))$  gives (dropping the argument  $x$  everywhere)

$$\begin{aligned} \text{nent}(s_\lambda * d\Psi) &\leq (h/\lambda)^2 \text{nent}(s_h * d\Psi) \\ &\quad + [1 - (h/\lambda)^2] \text{nent}(s_\lambda * s_h * d\Psi) \\ &\leq (h/\lambda)^2 \text{nent}(s_h * d\Psi) \\ &\quad + [1 - (h/\lambda)^2] s_\lambda * \text{nent}(s_h * d\Psi). \end{aligned}$$

Upon integration over  $\mathbb{R}$  it follows that

$$H(s_\lambda * d\Psi) \leq H(s_h * d\Psi).$$

The proof for  $D(s_\lambda * d\Psi, s_\lambda * d\Phi)$  is similar, by the afore mentioned convexity of  $D(\varphi, \psi)$  in  $\varphi, \psi$  jointly. Q.E.D.

### IV. EXHIBITING SUBMARTINGALES

It is obvious that for fixed  $h > 0$ ,  $\{nF_{nh}\}_n$  with  $F_{nh}$  defined in (1.8), is a martingale. Regarding  $\{D(g^{nh}, g_h)\}_n$  we observe that it does not appear to be a submartingale after proper scaling, but upon splitting  $\mathbb{R}$  into  $\{g^{nh} > g_h\}$  and  $\{g^{nh} < g_h\}$  one verifies that

$$D(g^{nh}, g_h) = D(g^{nh} \wedge g_h, g_h) + D(g^{nh} \vee g_h, g_h) \quad (4.1)$$

where

$$\varphi \wedge \psi(x) = \min\{\varphi(x), \psi(x)\} \quad (4.2)$$

and

$$\varphi \vee \psi(x) = \max\{\varphi(x), \psi(x)\}. \quad (4.3)$$

And now

*Lemma 4:* For fixed  $h > 0$ ,  $\{n^2 D(g^{nh} \vee g_h, g_h)\}_n$  is a submartingale.

Next we observe that  $D(g^{nh} \wedge g_h, g_h) \leq \chi(g^{nh} \wedge g_h, g_h)$ , where  $\chi$  is Pearson's  $\chi^2$  distance, for pdf's  $\varphi, \psi$  defined as

$$\chi(\varphi, \psi) \stackrel{\text{def}}{=} \int_{\mathbb{R}} \frac{|\varphi - \psi|^2}{\psi}. \quad (4.4)$$

*Lemma 5:* For fixed  $h > 0$ ,  $\{n^2 \chi(g^{nh} \wedge g_h, g_h)\}_n$  is a submartingale.

To use the submartingale inequalities, we need to suitably bound the expected values. Surely,  $\mathbb{E}[D(g^{nh} \vee g_h, g_h)] \leq \mathbb{E}[D(g^{nh}, g_h)]$ . Also note that with  $\varphi \equiv g^{nh} \wedge g_h$

$$\frac{|\varphi - g_h|^2}{g_h} = |\sqrt{\varphi} - \sqrt{g_h}|^2 \frac{|\sqrt{\varphi} + \sqrt{g_h}|^2}{g_h} \leq 4|\sqrt{\varphi} - \sqrt{g_h}|^2$$

so that we get a bound in terms of the Hellinger distance

$$\begin{aligned} \chi(g^{nh} \wedge g_h, g_h) &\leq 4 \|\sqrt{\varphi} - \sqrt{g_h}\|_2^2 \\ &\leq 4 D(\varphi, g_h) \leq 4 D(g^{nh}, g_h). \end{aligned} \quad (4.5)$$

For the next to last inequality a good reference is Devroye [5]. Thus we have

*Lemma 6:* Let  $h > 0$  be fixed. Then

$$D(g^{nh}, g_h) \leq D(g^{nh} \vee g_h, g_h) + \chi(g^{nh} \wedge g_h, g_h)$$

and

$$\begin{aligned} \mathbb{E}[D(g^{nh} \vee g_h, g_h)] &\leq \mathbb{E}[D(g^{nh}, g_h)] \\ \mathbb{E}[\chi(g^{nh} \wedge g_h, g_h)] &\leq 4 \mathbb{E}[D(g^{nh}, g_h)]. \end{aligned}$$

So, in effect, the expected values of the submartingales dominating  $D(g^{nh}, g_h)$  are at most a factor 4 bigger than what they should have been.

We finish this section by proving the submartingale properties.

*Proof of Lemma 4:* Let  $S_n = D(\varphi_n, g_h)$ , where  $\varphi_n = g^{nh} \vee g_h$ . We need to show that

$$(n+1)^2 \mathbb{E}[S_{n+1} | X_1, X_2, \dots, X_n] \geq n^2 S_n.$$

To simplify notation, let  $\mathbb{E}_n \equiv \mathbb{E}[\cdot | X_1, X_2, \dots, X_n]$ . The convexity of  $D(\varphi, g_h)$  as function of  $\varphi$  gives that  $\mathbb{E}_n[S_{n+1}] \geq D(\mathbb{E}_n[\varphi_{n+1}], g_h)$ . Now

$$\begin{aligned} \mathbb{E}_n[\varphi_{n+1}] &= \mathbb{E}_n[g^{n+1, h} \vee g_h] \\ &\geq (\mathbb{E}_n[g^{n+1, h}]) \vee g_h = \psi_n \vee g_h \end{aligned} \quad (4.6)$$

where

$$\psi_n \stackrel{\text{def}}{=} \theta g^{nh} + (1-\theta)g_h$$

with  $\theta = n/(n+1)$ . Since  $p \log(p/q) + q - p$  is an increasing function of  $p$  for  $p > q$ , it follows that

$$\mathbb{E}_n[S_{n+1}] \geq D(\mathbb{E}_n[\varphi_{n+1}], g_h) \geq D(\psi_n \vee g_h, g_h).$$

Next we observe that  $(\theta g^{nh} + (1-\theta)g_h) > g_h$  precisely on the set where  $g^{nh} > g_h$ , so that

$$\mathbb{E}_n[S_{n+1}] \geq \int_{\{g^{nh} > g_h\}} \psi_n \log \frac{\psi_n}{g_h} + g_h - \psi_n.$$

At this point we need the elementary inequality, with  $A(r) = r \log r + 1 - r$  for  $0 \leq \theta \leq 1$

$$\frac{A(\theta r + 1 - \theta)}{A(r)} \geq \theta^2, \quad r > 1. \quad (4.7)$$

To see this introduce the function  $B(r) = A(\theta r + 1 - \theta) - \theta^2 A(r)$ . One verifies that  $B(1) = 0$ , and that

$$B'(r) = \theta \log(\theta r + 1 - \theta) - \theta^2 \log r$$

which is nonnegative by the concavity of the logarithm. So  $B(r) > 0$  for  $r > 1$ , thus proving (4.7).

Now use (4.7) with  $r = g^{nh}/g_h$  to obtain

$$\begin{aligned} \mathbb{E}_n[S_{n+1}] &\geq \theta^2 \int_{\{g^{nh} > g_h\}} g^{nh} \log \frac{g^{nh}}{g_h} + g_h - g^{nh} \\ &= \theta^2 D(g^{nh} \vee g_h, g_h) = \theta^2 S_n \end{aligned}$$

which concludes the proof. Q.E.D.

*Proof of Lemma 5:* With  $\mathbb{E}_n$  as in the previous proof, the convexity of  $\chi(\varphi, g_h)$  as function of  $\varphi$  gives that

$$\mathbb{E}_n[\chi(g^{n+1, h} \wedge g_h, g_h)] \geq \chi(\mathbb{E}_n[g^{n+1, h} \wedge g_h], g_h).$$

Now we have

$$\mathbb{E}_n[g^{n+1, h} \wedge g_h] \leq (\mathbb{E}_n[g^{n+1, h}]) \wedge g_h = \psi_n \wedge g_h$$

with  $\psi_n \stackrel{\text{def}}{=} \theta g^{nh} + (1-\theta)g_h$  and with  $\theta = n/(n+1)$  as before. Now  $|x-z| > |y-z|$  for  $x < y < z$ , so that

$$\chi(\mathbb{E}_n[g^{n+1, h} \wedge g_h], g_h) \geq \chi(\psi_n \wedge g_h, g_h).$$

Again, the set where  $\psi_n < g_h$  is precisely the set where  $g^{nh} < g_h$ , so that

$$\begin{aligned} \chi(\psi_n \wedge g_h, g_h) &= \int_{\{g^{nh} < g_h\}} \frac{|\theta g^{nh} + (1-\theta)g_h - g_h|^2}{g_h} \\ &= \theta^2 \int_{\{g^{nh} < g_h\}} \frac{|g^{nh} - g_h|^2}{g_h} \\ &= \theta^2 \chi(g^{nh} \wedge g_h, g_h). \end{aligned}$$

This shows that

$$\mathbb{E}_n[\chi(g^{n+1, h} \wedge g_h, g_h)] \geq \theta^2 \chi(g^{nh} \wedge g_h, g_h)$$

and the lemma follows. Q.E.D.

## V. EXPECTED VALUES

For the submartingale inequality, the expected values  $\mathbb{E}[D(g^{nh}, g_h)]$  and  $\mathbb{E}[|F_{nh}|]$  are required.

*Lemma 7:* Under the condition (2.1)

$$\mathbb{E}[D(g^{nh}, g_h)] = \mathcal{O}((nh)^{-\kappa/(\kappa+1)}) \quad (5.1)$$

provided  $nh \rightarrow \infty, h \rightarrow 0$ .

*Proof:* In Eggermont and LaRiccia [7] it is shown that

$$\mathbb{E}[D(g^{nh}, g_h)] \leq \int_{\mathbf{R} \times \mathbf{R}} g(y) s_h(x-y) \log \left( 1 + \frac{s_h(x-y)}{ng_h(x)} \right) dx dy$$

so that

$$\begin{aligned} \mathbb{E}[D(g^{nh}, g_h)] &\leq \int_{\mathbf{R} \times \mathbf{R}} g(y) s_h(x-y) \log \left( 1 + \frac{\varepsilon}{g_h(x)} \right) dx dy \\ &\leq \int_{\mathbf{R}} g_h(x) \log \left( 1 + \frac{\varepsilon}{g_h(x)} \right) dx \end{aligned}$$

where  $\varepsilon = n^{-1} \max s_h(x) = (2nh)^{-1}$ . At this point we require the following elementary inequality: for every  $p > 1$ :

$$\begin{aligned} \log(1+t) &= p \log\{(1+t)^{1/p}\} \\ &\leq p\{[1+t]^{1/p} - 1\} \leq p t^{1/p}. \end{aligned}$$

Thus we obtain the bound

$$\mathbb{E}[D(g^{nh}, g_h)] \leq p \varepsilon^{1/p} \int_{\mathbf{R}} (g_h)^{1/q} \quad (5.2)$$

with  $(1/p) + (1/q) = 1$ . Now recall that  $g_h = s_h * g$ . Since  $s_h$  has moments of all orders, and  $g$  has a moment of order  $M > \kappa > 2$ , then from Devroye [5] we have that for  $M > q - 1$  there exists a constant  $c(M)$  such that

$$\int_{\mathbf{R}} (g_h)^{1/q} \leq c(M) \{ \mathbb{E}[|X|^M + h^M |Y|^M] \}^{1/q} \quad (5.3)$$

where  $X$  has pdf  $g$  and  $Y$  has the standard double exponential distribution. Taking  $q = \kappa + 1, p = (\kappa + 1)/\kappa$ , in (5.2) gives

$$\mathbb{E}[D(g^{nh}, g_h)] = \mathcal{O}((nh)^{-\kappa/(\kappa+1)}).$$

This is the lemma. Q.E.D.

To bound  $\mathbb{E}[|F_{nh}|]$  the following lemma concerning smoothed versions of the Fisher-like information numbers (2.2) is useful. This result does *not* depend on the kernel being the double exponential density.

*Lemma 8:* If the density  $g$  satisfies the smoothness condition (2.2), then

$$\begin{aligned} \int_{\mathbf{R}} \frac{|(g_h)'|^2}{g_h} &= \mathcal{O}(1) \\ \int_{\mathbf{R}} \frac{|(g_h)''|^2}{g_h} &= \mathcal{O}(1), \quad h \rightarrow 0. \end{aligned}$$

*Proof:* Since  $(u, v) \mapsto u^2/v$  is convex in  $(u, v) \in \mathbb{R} \times (0, \infty)$  (the Hessian is semipositive definite), Jensen's inequality gives that

$$\int_{\mathbf{R}} \frac{|s_h * (g'')|^2}{s_h * g} \leq \int_{\mathbf{R}} s_h * \frac{|g''|^2}{g} = \int_{\mathbf{R}} \frac{|g''|^2}{g} < \infty$$

by Assumption (2.2). The other result is similar. Q.E.D.

We are now ready for  $\mathbb{E}[|F_{nh}|]$ .

*Lemma 9:* Under the Assumptions (2.2) we have for fixed  $h$

$$\mathbb{E}[|F_{nh}|] = \mathbb{E} \left[ \left| \int_{\mathbf{R}} (s_h * \log g_h - \log g) dG_n \right| \right] = \mathcal{O}(h^2).$$

*Proof:* The expected value is dominated by  $\int_{\mathbf{R}} g|s_h * \log g_h - \log g|$ . Writing

$$\begin{aligned} |s_h * \log g_h - \log g| &= |s_h * \log g_h - \log g_h + \log(g_h/g)| \\ &\leq |s_h * \log g_h - \log g_h| + |\log(g_h/g)| \end{aligned}$$

gives

$$\begin{aligned} \int_{\mathbf{R}} g|s_h * \log g_h - \log g| \\ \leq \int_{\mathbf{R}} g|s_h * \log g_h - \log g_h| + \int_{\mathbf{R}} g|\log(g_h/g)|. \end{aligned} \quad (5.4)$$

The use of the Green's function property (1.10) yields

$$s_h * \log g_h - \log g_h = h^2(s_h * \log g_h)'' = h^2 s_h * (\log g_h)''$$

so that the first integral on the right in (5.4) is dominated by

$$h^2 \int_{\mathbf{R}} g\{s_h * |(\log g_h)''|\} = h^2 \int_{\mathbf{R}} g_h |(\log g_h)''|. \quad (5.5)$$

But since

$$(\log g_h)'' = \frac{(g_h)''}{g_h} - \left\{ \frac{(g_h)'}{g_h} \right\}^2$$

the expression in (5.5) is dominated by

$$h^2 \left\{ \int_{\mathbf{R}} |(g_h)''| + \int_{\mathbf{R}} \frac{|(g_h)'|^2}{g_h} \right\}.$$

With Cauchy-Schwarz, the first integral is bounded by the square root of

$$\int_{\mathbf{R}} \frac{|(g_h)''|^2}{g_h},$$

and so by Lemma 8 all this is  $\mathcal{O}(h^2)$ .

For the second integral in (5.4) we first use Cauchy-Schwarz, and then the elementary inequality

$$|\log t| = 2|\log \sqrt{t}| \leq 2 \frac{|t-1|}{\sqrt{t}}, \quad t > 0$$

with  $t = g/g_h$ , to get

$$\left\{ \int_{\mathbf{R}} g|\log(g_h/g)| \right\}^2 \leq \int_{\mathbf{R}} g|\log(g_h/g)|^2 \leq 4 \int_{\mathbf{R}} \frac{|g-g_h|^2}{g_h}.$$

Once more by the Green's function property (1.10) the right-hand side equals

$$4h^4 \int_{\mathbf{R}} \frac{|(g_h)''|^2}{g_h}$$

and by Lemma 8 this is  $\mathcal{O}(h^4)$ .

Q.E.D.

## VI. USING SUBMARTINGALES TO PROVE LEMMAS 2 AND 3

In this section we prove the a.s. behavior of  $D(g^{nh}, g_h)$  and  $F_{nh}$ , as stated in Lemmas 2 and 3. The rates are not optimal, but are close enough for the present purpose. The results follow from the standard (sub)martingale inequalities combined with some standard trickery.

Suppose for each  $h > 0$  that  $\{S_n(h)\}_n$  is a submartingale. The submartingale inequality, see, e.g., Breiman [4], implies that for all  $\lambda > 0$

$$\text{Prob}[M_n(h) > \lambda^{-1}] \leq \lambda \mathbb{E}[|S_n(h)|] \quad (6.1)$$

where

$$M_n(h) = \max_{n/2 < k \leq n} S_k(h). \quad (6.2)$$

Here  $h$  is still fixed, but we may take  $h = h_n$  varying with  $n$ . Taking

$$\lambda = \lambda_n = (\log n)^s \mathbb{E}[|S_n(h_n)|] \quad (6.3)$$

where  $s > 1$  is arbitrary but fixed, then gives that

$$\text{Prob}[M_n(h_n) > \lambda_n] < (\log n)^{-s}.$$

If we replace  $n$  by  $2^n$ , then  $(\log n)^{-s}$  is replaced by a term of order  $n^{-s}$ , and so the Borel-Cantelli lemma implies that with  $m(n) \equiv 2^n$

$$\limsup_{n \rightarrow \infty} \{\lambda_{m(n)}\}^{-1} M_{m(n)}(h_{m(n)}) \leq 1, \quad \text{almost surely}$$

and so

$$\max_{m(n-1) < k \leq m(n)} S_k(h_{m(n)}) =_{\text{as}} \mathcal{O}(n^s \mathbb{E}[|S_{m(n)}(h_{m(n)})|]). \quad (6.4)$$

Now apply this to

$$S_k(h) = k^2 D(g^{kh} \vee g_h, g_h)$$

with  $h_k = (m(n))^{-\beta}$  for  $m(n-1) < k \leq m(n)$ , with  $0 < \beta < 1$ . Then

$$M_{m(n)} =_{\text{as}} \mathcal{O}(n^s (m(n))^2 a_{m(n)}) \quad (6.5)$$

with

$$a_k = (kh_k)^{-\kappa/(\kappa+1)} \quad (6.6)$$

where we used Lemma 7. It follows that for  $n \rightarrow \infty$

$$\begin{aligned} k^2 D(g^{k, h_{m(n)}} \vee g_{h_{m(n)}}, g_{h_{m(n)}}) &=_{\text{as}} \mathcal{O}(n^s (m(n))^2 a_{m(n)}), \\ m(n-1) < k \leq m(n). \end{aligned}$$

Repeating the above for the submartingale  $\{n^2 \chi(g^{nh} \wedge g_h, g_h)\}_n$ , gives the same bound, and then adding these two inequalities gives

$$k^2 D(g^{k, h_{m(n)}}, g_{h_{m(n)}}) =_{\text{as}} \mathcal{O}(n^s (m(n))^2 a_{m(n)}).$$

Now observe that for  $m(n-1) < k \leq m(n)$

$$\begin{aligned} \frac{n^s (m(n))^2 a_{m(n)}}{(\log k)^s k^2 a_k} &\leq \frac{n^s (m(n))^2 a_{m(n-1)}}{(n-1)^s \log 2 (m(n-1))^2 a_{m(n)}} \\ &\leq \text{Constant} \end{aligned}$$

so that

$$D(g^{k, h_{m(n)}}, g_{h_{m(n)}}) =_{\text{as}} \mathcal{O}((\log k)^s a_k).$$

Together with the monotonicity of  $D(g^{nh}, g_h)$  in  $h$ , this proves Lemma 2. The proof of Lemma 3 goes the same way, and is omitted.

## VII. THE ESTIMATOR OF AHMAD AND LIN [1]

In this section we study the estimator of Ahmad and Lin [1], still based on the double exponential kernel. Again this makes the details come out quite palatable. As a matter of fact, all the previous arguments are recycled here.

*Theorem 2:* Under the Assumptions (2.1) and (2.2), for all  $s > 1$

$$\tilde{H}(g^{nh}) = \int_{\mathbf{R}} \log g dG_n + \varepsilon_{nh}$$

with  $\varepsilon_{nh} = \mathcal{O}(n^{-2\kappa/(3\kappa+2)} (\log n)^s)$  almost surely, provided  $h = h_n \asymp n^{-\kappa/(3\kappa+2)}$ . In particular then  $\varepsilon_{nh} = o(n^{-1/2})$  almost surely.

For the proof it suffices to show

*Lemma 10:*

$$0 \leq \tilde{H}(g^{nh}) - H(g^{nh}) \leq 2h^2 \int_{\mathbf{R}} \frac{|(g_h)'|^2}{g_h} + 128D(g^{nh}, g_h).$$

*Proof of Lemma 10:* We note that using the Green's function property (1.10)

$$\begin{aligned} \tilde{H}(g^{nh}) - H(g^{nh}) &= \int_{\mathbf{R}} \{\log g^{nh} - s_h * \log g^{nh}\} dG_n \\ &= -h^2 \int_{\mathbf{R}} s_h * \{(\log g^{nh})'\} dG_n \\ &= -h^2 \int_{\mathbf{R}} \{(\log g^{nh})'\} s_h * dG_n \\ &= h^2 \int_{\mathbf{R}} \frac{|(g^{nh})'|^2}{g^{nh}} \\ &= 4h^2 \int_{\mathbf{R}} \{|(g^{nh})^{1/2}\}'|^2 \end{aligned} \quad (7.1)$$

where to obtain the next to last line we used integration by parts. The positivity of  $\tilde{H}(g^{nh}) - H(g^{nh})$  is thus proved. The expression (7.1) may be bounded further as

$$\leq 8h^2 \int_{\mathbf{R}} \{|(g_h)^{1/2}\}'|^2 + 8 \int_{\mathbf{R}} T$$

where

$$T = \{|(g^{nh})^{1/2} - (g_h)^{1/2}\}'|^2.$$

Now the integrand may be written in one of two ways. The first one is

$$T = \frac{|(g^{nh})'(g^{nh} - g_h) + g^{nh}(g^{nh} - g_h)'|^2}{g^{nh}g_h}.$$

Now, since  $|(s_h)'| \leq h^{-1}s_h$  is pointwise, we get  $|(g^{nh})'| \leq g^{nh}$ , and likewise for  $(g_h)'$ . Consequently,

$$T \leq 4h^{-2} \frac{|g^{nh} - g_h|^2}{g_h}.$$

But  $T$  is symmetric in  $g^{nh}$  and  $g_h$ , so we also get the same bound with  $g^{nh}$  rather than  $g_h$  in the denominator. Taking the minimum of these two bounds we get

$$T \leq 4h^{-2} \frac{|g^{nh} - g_h|^2}{g^{nh} \vee g_h}.$$

Now with the same trick as in the bound for  $\chi(g^{nh} \wedge g_h, g_h)$ , see Section IV, we get that

$$\int_{\mathbf{R}} T \leq 16h^2 D(g^{nh}, g_h)$$

thus proving the lemma. Q.E.D.

### VIII. CONCLUDING REMARKS

The results from Section II and their proofs extend to the multivariate case  $X \in \mathbb{R}^d$ , with suitable modifications, but the main result is much less impressive. The smoothing operator must be replaced by, say, the  $d$ th power of the Green's function operator for the  $d$ -dimensional boundary value problem

$$\begin{aligned} -h^2 \Delta \psi(x) + \psi(x) &= \varphi(x), & x \in \mathbb{R}^d \\ \psi &\rightarrow 0, & |x| \rightarrow \infty \end{aligned} \quad (8.1)$$

where  $\Delta$  is the Laplacian, the smoothness conditions (2.2) by

$$\int_{\mathbb{R}^d} \frac{|\nabla g(x)|^2}{g(x)} dx < \infty \quad \int_{\mathbb{R}^d} \frac{|\Delta g(x)|^2}{g(x)} dx < \infty. \quad (8.2)$$

Finally, the moment condition (2.1) must then be replaced by

$$\mathbb{E}[\|X\|^{Md}] < \infty \quad (8.3)$$

for some  $M > \kappa > 2$ . (Then (5.3) has an analog.) The Main Theorem changes though, because in Lemma 2  $h$  must be replaced by  $h^d$ , so

$$D(g^{nh}, g_h) =_{\text{as}} \mathcal{O}((nh^d)^{-\kappa/(\kappa+1)}(\log n)^s) \quad (8.4)$$

and the main result is that

$$\varepsilon_{nh} =_{\text{as}} o(n^{-2\kappa/((2+d)\kappa+2)}) \quad (8.5)$$

and for  $d \geq 2$  this is not  $o(n^{-1/2})$ . So we do get consistent estimators, but not asymptotic normality. The same applies to the estimator of Ahmad and Lin [1].

### ACKNOWLEDGMENT

The authors wish to thank L. Györfi for providing us with a preprint of the review paper Beirlant, Dudewicz, Györfi, van der Meulen [3].

### REFERENCES

- [1] I. A. Ahmad and P. E. Lin, "A nonparametric estimation of the entropy for absolutely continuous distributions," *IEEE Trans Inform. Theory*, vol. 36, pp. 688–692, 1989.
- [2] 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 Inform. Theory*, vol. 38, pp. 1437–1454, 1992.
- [3] J. Beirlant, E. Dudewicz, L. Györfi, and E. G. van der Meulen, "Nonparametric entropy estimation: An overview," *Int. J. Math Stat Sci.*, vol. 6, pp. 17–39, 1997.
- [4] L. Breiman, *Probability*. Reading, MA: Addison-Wesley, 1968.
- [5] L. Devroye, *A Course in Density Estimation*. Boston, MA: Birkhäuser, 1987.
- [6] L. Devroye and L. Györfi, *Nonparametric Density Estimation*. New York: Wiley, 1985.
- [7] P. P. B. Eggermont and V. N. LaRiccia, "Maximum smoothed likelihood density estimation," *J Nonparam. Statist.*, vol. 4, pp. 211–222, 1995.
- [8] —, "Nonlinearly smoothed EM density estimation with automatic smoothing parameter selection for nonparametric deconvolution problems," *J Amer. Statist. Assoc.*, vol. 97 pp. 1451–1458, 1997.
- [9] —, "Optimal convergence rates for Good's nonparametric maximum likelihood density estimator," unpublished manuscript, Jan. 1998.
- [10] I. J. Good and R. A. Gaskins, "Nonparametric roughness penalties for probability densities," *Biometrika*, vol. 58, pp. 255–277, 1971.
- [11] L. Györfi and E. C. van der Meulen, "Density-free convergence properties of various estimators of the entropy," *Comput. Statist. Data Anal.*, vol. 5, pp. 425–436, 1987.
- [12] —, "An entropy estimate based on a kernel density estimation," *Colloq. Math. Soc. J. Bolyai*, no. 57: *Limit Theorems in Probability and Statistics*, I. Berkes, E. Csáki, P. Révész, Eds. Amsterdam, The Netherlands: North-Holland, 1990.
- [13] P. Hall, "On powerful distributional tests based on sample spacings," *J. Multivar. Statist.*, vol. 19, pp. 201–225, 1986.
- [14] —, "On Kullback–Leibler loss and density estimation," *Ann. Statist.*, vol. 14, pp. 1491–1495, 1987.
- [15] P. Hall and S. C. Morton, "On the estimation of entropy," *Ann. Inst. Statist. Math.*, vol. 45, pp. 69–88, 1993.
- [16] H. Joe, "Estimation of entropy and other functionals of a multivariate density," *Ann. Inst. Statist. Math.*, vol. 41, pp. 683–697, 1989.
- [17] B. Ya. Levit, "Asymptotically efficient estimation of nonlinear functionals," *Probl. Inform. Transm.*, vol. 14, pp. 204–209, 1978.
- [18] A. Mokkadem, "Estimation of entropy and information of absolutely continuous random variables," *IEEE Trans Inform. Theory*, vol. 35, pp. 193–196, 1989.
- [19] J. R. Thompson and R. A. Tapia, *Nonparametric Function Estimation, Modeling, and Simulation*. Philadelphia, PA: SIAM, 1990.
- [20] 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, pp. 75–83, 1994.
- [21] B. van Es, "Estimating functionals related to a density by a class of statistics based on spacings," *Scand. J. Statist.*, vol. 19, pp. 61–72, 1992.