Asymptotic normality of the L_1- error of a boundary estimator

Jean Geffroy[1], Stéphane Girard[2], Pierre Jacob[3]

[1]Ancien professeur à l'université de Paris VI

7, Grande Rue, 76130 Mont Saint Aignan, France.

[2]SMS/LMC, Université Grenoble 1,

BP 53, 38041 Grenoble cedex 9, France.

Stephane.Girard@imag.fr

[3] EPS/I3M, Université de Montpellier 2

Place Eugène Bataillon, 34095 Montpellier cedex 5, France.

jacob@math.univ-montp2.fr

Abstract: We present a result on the asymptotic normality of the L_1- error of the sup-piecewise constant estimator of frontier functions. This result is obtained by means of an original squeezing method of the empirical point process between two Poisson point processes. The result is available under the best rate of convergence, but becomes fully useful under sub-optimal conditions.

Key words: Boundaries, frontier function, extreme values, Poisson approximations, asymptotic normality.

1 Introduction

In an early paper, Geffroy [6] introduced the problem of estimating a subset D of $\mathbb{R}^2$ given random sample of points $\Sigma_n = \{(X_i, Y_i)\,; i = 1, ..., n\}$ drawn from the interior. He considered a set

$$D = \left\{(x, y) \in \mathbb{R}^2 : 0 \leq x \leq 1; 0 \leq y \leq f(x)\right\} \tag{1}$$

where f is a strictly positive function. Given an increasing sequence of integers k_n, for $r = 1, ..., k_n$, let $I_{n,r} = [(r-1)/k_n, r/k_n)$ and

$$U_{n,r} = \max\{Y_i : 1 \leq i \leq n; X_i \in I_{n,r}\} \tag{2}$$

where it is conveniently understood that $\max\varnothing = 0$. The Geffroy's estimate of f is the step function f_n defined by

$$f_n(x) = U_{n,r} \text{ if } x \in I_{n,r},\ r = 1, ..., k_n, \tag{3}$$

$$U_{n,k_n} \text{ if } x = 1. \tag{4}$$

Under various conditions on f and on the probability density of the $Z_i = (X_i, Y_i)$, Geffroy produced convergence theorems and limit laws for the L_∞-norm $\sup_{0\leq x\leq 1} |f_n(x) - f(x)|$. Much later, Korostelev and Tsybakov [14] embedded this very simple estimate in the more sophisticated class of piecewise-polynomial estimates. But despite of its simplicity, the piecewise-constant Geffroy's estimate remains the most appropriate when little is known about the properties of f. This is the case when estimating boundary fragments in images [15]. More precisely, in the case of a uniform sample on D, when it is only known that f possesses a bounded derivative, Korostelev and Tsybakov [14] proved that f_n is minimax for the L_∞- norm if $k_n = (\log n/n)^{-1/2}$ and for the L_1- norm $\int_0^1 |f_n(x) - f(x)|dx$ if $k_n = n^{1/2}$. Since the results for the L_∞- norm are already known from the original paper of Geffroy, solving the question of the limit law for the L_1- norm was a primary motivation of the present study.

The interest of the L_1- error in functional estimation was mainly highlighted by Devroye and Györfi [4]. One of the most attractive features of the L_1- error is to be visualized by the surface marked off by the estimator and the function to be estimated. No doubt that in a frontier estimation problem this criterion should be of particular interest.

As for the asymptotic normality of the histogram estimates of density [1, 2, 3] it seems that no direct proof is tractable. We also proceed by Poisson approximations, but in a different and original way. Girard and Jacob [12] first solved the Poissonian case which necessitate only current methods. Next, in an unpublished paper, Geffroy [7] brought out the key idea for extending their result by a squeezing method of the empirical point process between two Poisson point processes. However, this first step was achieved by Geffroy only in the elementary case where $f = 1$ on $[0, 1]$. Much more refinements for a complete treatment of the general case were necessary, and the whole development is presented here.

Of course, many works exist about estimates of boundaries but few of them provide asymptotic distributions. In the framework of production frontier estimation, the value $f(x)$ can be interpreted as the maximum level of output which is attainable for the level of input x. Then, from economical considerations, f is supposed to be increasing and concave which suggests an adapted estimator, called the DEA (Data Envelopment Analysis) estimator [5]. Its asymptotic distribution is established in [8]. When no monotonicity assumption is made on f, without intending to be exhaustive, we can cite [9, 10, 11, 14] for other asymptotic distribution studies.

But apart from the early paper [6] of Geffroy, which gives Gumbel limit laws for uniform errors, we have not found attempts on limit laws for global errors. As far as we know, the present paper is the first to give a solution for the L_1- error in the case of independent and uniformly distributed random points.

Before stating the main result, we need some further definitions.

Given two sequences of positive numbers (u_n) and (v_n), we write:

- $u_n = o(v_n)$ or equivalently $u_n \ll v_n$ if $u_n/v_n \to 0$ as $n \to \infty$,
- $u_n \sim v_n$ if $u_n/v_n \to 1$ as $n \to \infty$,
- $u_n = O(v_n)$ if $\limsup u_n/v_n < \infty$,
- $u_n \asymp v_n$ if $0 < \liminf u_n/v_n \leq \limsup u_n/v_n < \infty$.

Both conditions $u_n \asymp v_n$ and $u_n = o\left(v_n\right)$ imply $u_n = O\left(v_n\right)$. Conversely, $u_n = O\left(v_n\right)$ implies either $u_n \asymp v_n$ or $u_n = o\left(v_n\right)$ at least for subsequences. Denoting by λ the Lebesgue measure restricted on D, we put $\lambda(D) = c^{-1}$ and we also note

$$\Delta_n = \int_0^1 |f_n(x) - f(x)| dx,$$

the L_1- error, where f_n is defined by (2)–(4).

Theorem 1 *Assume f is $\alpha-$Lipschitzian. If $k_n = o(n/\log n)$ and $n = O\left(k_n^{1+\alpha}\right)$, then there exists a bounded sequence (s_n) such that*

$$\frac{nc}{s_n\sqrt{k_n}}\left(\Delta_n - E\left(\Delta_n\right)\right) \to \mathcal{N}\left(0,1\right) \text{ as } n \to \infty. \tag{5}$$

More precisely, if $n \asymp k_n^{1+\alpha}$ then $s_n \asymp 1$, and if $n = o\left(k_n^{1+\alpha}\right), s_n = 1$.

The paper is organized as follows: in section 2 are brought together the preliminary results useful for the construction of the Poisson approximation and for preparing the discussion. In section 3 are quoted, with some explanations, the known results on L_1- error for Poisson processes. The section 4 is devoted to the proof of the above theorem. Finally, in the discussion section 5 we show that the unknown term c can be replaced by a simple estimate, and that the term $E\left(\Delta_n\right)$ can be replaced by a centering sequence independent of f. This gain, which makes the result applicable, is paid by the loss of optimal speed, so we can take $s_n = 1$.

2 Some preliminaries

In what follows, (Z_n) is a sequence of independent random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ and uniformly distributed on the set D given by (1), and f_n is the estimate of f defined by the $U_{n,r}$, $r = 1, ..., k_n$ as in (2)–(4). We quote and prove quickly a exponential inequality for which we have no precise reference to propose.

Lemma 1 *Let X a Poisson random variable with parameter $\mu > 0$. For $\varepsilon/2\mu$ small enough,*

$$P\left(X - \mu > \varepsilon\right) < \exp\left(-\varepsilon^2/4\mu\right).$$

Proof. Let $0 < \delta < t$ such that $e^t - 1 - t < t^2$. Then,

$$E\left(e^{t(X-\mu)}\right) = \exp\left(\mu\left(e^t - 1 - t\right)\right) \leq e^{\mu t^2}$$

and thus, for every $\varepsilon > 0$,

$$P\left(X - \mu > \varepsilon\right) < e^{-t\varepsilon} e^{\mu t^2}.$$

If $t = \varepsilon/2\mu < \delta$, the result holds. Of course, a similar bound is available for $P\left(X - \mu < -\varepsilon\right)$. ■

Now, following the idea of [7], the proof of the main theorem requires a *random sandwiching* of the sample $\Sigma_n = \{Z_1, ..., Z_n\}$ between two Poisson point processes $\Sigma_{1,n}$ and $\Sigma_{2,n}$. Define $c_{1,n} = c\left(1 - \gamma_n\right)$ and $c_{2,n} = c\left(1 + \gamma_n\right)$, with $\gamma_n = \left(\frac{\log n}{nk_n}\right)^{1/4}$. Then consider, for every n, two Poisson random variables $N_{1,n}$ and $N_{2,n}$ on $(\Omega, \mathcal{F}, P)$, with respective parameters $n\left(1 - \gamma_n\right)$ and $n\left(1 + \gamma_n\right)$, each of them being independent of the sequence (Z_n), and satisfying $N_{1,n} < N_{2,n}$. Such a construction is achieved by taking $N_{2,n} = N_{1,n} + M_n$ where M_n is a Poisson random variable with parameter $2n\gamma_n$, independent of $N_{1,n}$. The two announced Poisson point processes are the random sets $\Sigma_{1,n} = \left\{Z_1, ..., Z_{N_{1,n}}\right\}$ and $\Sigma_{2,n} = \left\{Z_1, ..., Z_{N_{2,n}}\right\}$ and the sandwich event is $E_n = \left(\Sigma_{1,n} \subseteq \Sigma_n \subseteq \Sigma_{2,n}\right)$.

Note that the mean measure of the point process $\Sigma_{j,n}$ is $nc_{j,n}\lambda$ for $j = 1, 2$, while the mean measure of the empirical point process Σ_n is $nc\lambda$.

Lemma 2 *If $k_n \to \infty$ and $k_n = o(n/\log n)$, then*

$$\lim_{n\to\infty} \frac{n}{\sqrt{k_n}} P\left(\Omega \setminus E_n\right) = 0.$$

Proof. Clearly, $\Omega \setminus E_n = (N_{1,n} > n) \cup (N_{2,n} < n)$. From Lemma 1, we obtain for n large enough

$$P(\Omega \setminus E_n) < \exp\left(-\frac{n\gamma_n^2}{4(1-\gamma_n)}\right) + \exp\left(-\frac{n\gamma_n^2}{4(1+\gamma_n)}\right).$$

Consequently, it suffices to verify that $nk_n^{-1/2}\exp(-n\gamma_n^2)$ goes to 0 under the condition $k_n = o(n/\log n)$. Writing $k_n = n\varepsilon_n/\log n$ with $\varepsilon_n \to 0$, it amounts to checking that

$$\log n - \log\sqrt{k_n} - n\gamma_n^2 = \log n - \log\sqrt{k_n} - \frac{\log n}{\sqrt{\varepsilon_n}} \to -\infty,$$

which is immediate. ■

Now, for $j = 1, 2$ define

$$U_{j,n,r} = \max\left\{Y_i : 1 \le i \le N_{j,n}; X_i \in I_{n,r}\right\},$$

and introduce

$$\begin{aligned} m_{n,r} &= \inf\left\{f(x) : x \in I_{n,r}\right\}, \\ M_{n,r} &= \sup\left\{f(x) : x \in I_{n,r}\right\}. \end{aligned}$$

For $0 \le y \le M_{n,r}$ set also,

$$D_{n,r}(y) = \left\{(x,t) \in \mathbb{R}^2 : x \in I_{n,r}; y \le t \le f(x)\right\}.$$

Remark that $(U_{j,n,r} \le y) = (\Sigma_{j,n} \cap D_{n,r}(y) = \varnothing)$. Since the mean measure of the point process $\Sigma_{j,n}$ is $nc_{j,n}\lambda$, the distribution function of $U_{j,n,r}$ can be written

$$G_{j,n,r}(y) = P\left(U_{j,n,r} \le y\right) = \exp\left(-nc_{j,n}\lambda\left(D_{n,r}(y)\right)\right).$$

In particular, for $0 \leq y \leq m_{n,r}$, we have

$$G_{j,n,r}\left(y\right) = \exp\left(nc_{j,n}\left(y/k_n - \lambda_{n,r}\right)\right),$$

where we have defined

$$\lambda_{n,r} = \lambda\left(D_{n,r}\left(0\right)\right). \tag{6}$$

Lemma 3 *Assume f is $\alpha-$Lipschitzian. If $k_n = o(n/\log n)$ and $n = O\left(k_n^{1+\alpha}\right)$, then*

$$\lim_{n\to\infty} \frac{n}{k_n^{3/2}} \sum_{r=1}^{k_n} E\left(U_{2,n,r} - U_{1,n,r}\right) = 0.$$

Proof. For $j = 1, 2$ we have

$$E\left(U_{j,n,r}\right) = \int_0^{M_{n,r}} \left(1 - G_{j,n,r}\left(y\right)\right) dy,$$

so that $E\left(U_{2,n,r} - U_{1,n,r}\right) = A + B$, with

$$\begin{aligned} A &= \int_0^{m_{n,r}} \exp\left(\frac{nc_{1,n}}{k_n}\left(y - k_n\lambda_{n,r}\right)\right) dy - \exp\left(\frac{nc_{2,n}}{k_n}\left(y - k_n\lambda_{n,r}\right)\right) dy \\ B &= \int_{m_{n,r}}^{M_{n,r}} \left(G_{1,n,r}\left(y\right) - G_{2,n,r}\left(y\right)\right) dy. \end{aligned}$$

Now, A is calculated as a sum $A_1 + A_2$ with

$$\begin{aligned} A_1 &= \frac{k_n}{nc_{1,n}} \exp\left(\frac{nc_{1,n}}{k_n}\left(m_{n,r} - k_n\lambda_{n,r}\right)\right) - \frac{k_n}{nc_{2,n}} \exp\left(\frac{nc_{2,n}}{k_n}\left(m_{n,r} - k_n\lambda_{n,r}\right)\right) \\ A_2 &= \frac{k_n}{nc_{2,n}} \exp\left(-nc_{2,n}\lambda_{n,r}\right) - \frac{k_n}{nc_{1,n}} \exp\left(-nc_{1,n}\lambda_{n,r}\right). \end{aligned}$$

The part A_2 is easily seen to be a $o\left(n^{-s}\right)$ where s is an arbitrarily large exponent under the condition $k_n = o(n/\log n)$. Moreover, if a, b, x, y are real numbers such that $x < y < 0 < b < a$, we have

$$0 < ae^y - be^x = \left(a - b\right)e^y + b\left(e^y - e^x\right) < \left(a - b\right) + b\left(y - x\right). \tag{7}$$

Applying to A_1 the inequality (7) gives

$$\begin{aligned} A_1 &\leq \frac{k_n}{nc_{1,n}} - \frac{k_n}{nc_{2,n}} + \frac{k_n}{nc_{2,n}}\left(\frac{nc_{1,n}}{k_n}\left(m_{n,r} - k_n\lambda_{n,r}\right) - \frac{nc_{2,n}}{k_n}\left(m_{n,r} - k_n\lambda_{n,r}\right)\right) \\ &\leq \frac{k_n}{n}\left(\frac{c_{2,n} - c_{1,n}}{c_{2,n}c_{1,n}}\right) + \left(M_{n,r} - m_{n,r}\right)\frac{c_{2,n} - c_{1,n}}{c_{2,n}} \\ &\sim \frac{k_n}{n}\frac{2\gamma_n}{c\left(1 - \gamma_n^2\right)} + \left(M_{n,r} - m_{n,r}\right)\frac{2\gamma_n}{c\left(1 + \gamma_n^2\right)}. \end{aligned}$$

Under the condition $n = O\left(k_n^{1+\alpha}\right)$ and the hypothesis that f is $\alpha-$Lipschitzian, $(M_{n,r} - m_{n,r}) = O(k_n/n)$, so that $A_1 = O\left(k_n/(n\gamma_n)\right)$. Finally,

$$A = o\left(n^{-s}\right) + O\left(\frac{k_n}{n}\gamma_n\right) = O\left(\frac{k_n}{n}\gamma_n\right). \qquad (8)$$

Now, for $m_{n,r} \leq y \leq M_{n,r}$,

$$G_{1,n,r}\left(y\right) - G_{2,n,r}\left(y\right) \leq n\left(c_{2,n} - c_{1,n}\right)\lambda(D_{n,r}\left(y\right)),$$

so that

$$B \leq 2\gamma_n \frac{nc}{k_n}\left(M_{n,r} - m_{n,r}\right)^2 = O\left(\frac{k_n}{n}\gamma_n\right). \qquad (9)$$

Summarizing (8) and (9) we obtain

$$E\left(U_{2,n,r} - U_{1,n,r}\right) = O\left(\frac{k_n^{3/4}}{n^{5/4}}\left(\log n\right)^{1/4}\right).$$

Of course this last result is uniform in $r = 1, ..., k_n$, thus

$$\frac{n}{k_n^{3/2}}\sum_{r=1}^{k_n} E\left(U_{2,n,r} - U_{1,n,r}\right) = O\left(n^{-1/4}k_n^{1/4}\left(\log n\right)^{1/4}\right) = o\left(1\right),$$

and the result is proved. ■

A precise evaluation of $E\left(\Delta_n\right)$ is not useful to obtain the main result. However, we cannot avoid it in the discussion of section 5.

Lemma 4 *Assume f is $\alpha-$Lipschitzian. If $k_n = o(n/\log n)$ and $n = O\left(k_n^{1+\alpha}\right)$, then*

$$E\left(\Delta_n\right) = \frac{k_n}{\left(n+1\right)c} + O\left(\frac{n}{k_n^{1+2\alpha}}\right).$$

More precisely, if $n \asymp k_n^{1+\alpha}$ then $E\left(\Delta_n\right) = O(n^{-\frac{\alpha}{1+a}})$, and if $n = o\left(k_n^{1+\alpha}\right)$ then $E\left(\Delta_n\right) = \frac{k_n}{nc}\left(1 + o\left(1\right)\right)$.

Proof. The distribution function of $U_{n,r}$ can be written

$$G_{n,r}\left(y\right) = P\left(U_{n,r} \leq y\right) = \left(1 - c\lambda(D_{n,r}\left(y\right))\right)^n, 0 \leq y \leq M_{n,r}.$$

More precisely, for $0 \leq y \leq m_{n,r}$ we have

$$G_{n,r}\left(y\right) = \left(1 - c\left(\lambda_{n,r} - y/k_n\right)\right)^n,$$

where $\lambda_{n,r}$ is defined by (6). Let us write $E(\Delta_n) = \sum_{r=1}^{k_n} E(A_{n,r})$, where

$$\begin{aligned} A_{n,r} &= \int_{I_{n,r}} |f(x) - f_n(x)|\, dx \\ &= \int_{I_{n,r}} (f(x) - U_{n,r})\, dx + 2\int_{I_{n,r}} (U_{n,r} - f(x)) \mathbf{1}_{\{U_{n,r} \geq f(x)\}} dx. \end{aligned}$$

A straightforward calculation gives, uniformly in r:

$$\begin{aligned} E\left(\int_{I_{n,r}} (f(x) - U_{n,r})\, dx\right) &= \int_0^{M_{n,r}} (\lambda_{n,r} - y/k_n)\, dG_{n,r}(y) \\ &= \frac{1}{(n+1)c} + O\left(\frac{n}{k_n^{2+2\alpha}}\right). \end{aligned}$$

Moreover, $\max_{1\leq r\leq k_n}(1 - G_{n,r}(m_{n,r})) = O\left(n/k_n^{1+\alpha}\right)$ since f is $\alpha-$Lipschitzian, thus

$$\begin{aligned} E\left(\int_{I_{n,r}} (U_{n,r} - f(x)) \mathbf{1}_{\{U_{n,r}\geq f(x)\}} dx\right) &\leq \frac{(M_{n,r} - m_{n,r})}{k_n}(1 - G_{n,r}(m_{n,r})) \\ &= O\left(\frac{n}{k_n^{2+2\alpha}}\right). \end{aligned}$$

The discussion between the cases $n \asymp k_n^{1+\alpha}$ *and* $n = o\left(k_n^{1+\alpha}\right)$ is obvious. ■

3 The Poissonian case

Now, for $j = 1, 2$, let

$$\Delta_{j,n} = \int_0^1 |f_{j,n}(x) - f(x)|\, dx$$

denote the L_1- error for the estimate $f_{j,n}$, defined by $f_{j,n}(x) = U_{j,n,r}$, for every $x \in I_{n,r}$, $r = 1, ..., k_n$, and $f_{j,n}(1) = U_{j,n,k_n}$. We summarize below the main results of the paper [12]. The proof of the Lemma 5 is just a matter of patient calculus of moments of order $1, 2, 3$ for each random variable

$$A_{j,n,r} = \int_{I_{n,r}} |f(x) - f_{j,n}(x)|\, dx$$

based upon the distribution functions $G_{j,n,r}$ and following the same lines as the proof of Lemma 4. See [12], Lemma 3.2 for further details.

Lemma 5 *Assume f is $\alpha-$Lipschitzian. If $k_n = o(n/\log n)$ and $n = O\left(k_n^{1+\alpha}\right)$, then for $j = 1, 2$,*

(i) $\max\limits_{1\leq r\leq k_n} \left| E\left(A_{j,n,r}\right) - \dfrac{1}{nc_{j,n}} \right| = O\left(\dfrac{n}{k_n^{2+2\alpha}}\right)$

(ii) $\max\limits_{1\leq r\leq k_n} \left| E\left(A_{j,n,r}^2\right) - \dfrac{2}{n^2c_{j,n}^2} \right| = O\left(\dfrac{n}{k_n^{3+3\alpha}}\right)$

(iii) $\max\limits_{1\leq r\leq k_n} E\left(A_{j,n,r}^3\right) \leq \dfrac{6}{n^3c_{j,n}^3} + O\left(\dfrac{n}{k_n^{4+4\alpha}}\right).$

However, the following lemma is crucial when dealing with the case $n \asymp k_n^{1+\alpha}$, so we give a more detailed proof.

Lemma 6 *Assume f is $\alpha-$Lipschitzian. If $n \asymp k_n^{1+\alpha}$, then there exists $K > 0$ such that for $j = 1, 2$,*

$$\min_{1\leq r\leq k_n} \mathit{Var}(A_{j,n,r}) \geq \frac{K}{n^2c_{j,n}^2}(1 + o(1)).$$

Proof. As a consequence of the variance decomposition formula,

$$\begin{aligned} \mathrm{Var}\left(A_{j,n,r}\right) &\geq P\left(U_{j,n,r} \leq m_{n,r}\right) \mathrm{Var}(A_{j,n,r}|U_{n,r} \leq m_{n,r}) && (10)\\ &= G_{j,n,r}\left(m_{n,r}\right) \frac{1}{k_n^2} \mathrm{Var}\left(U_{j,n,r}|U_{j,n,r} \leq m_{n,r}\right) && (11) \end{aligned}$$

since $A_{j,n,r} = \lambda_{n,r} - U_{j,n,r}/k_n$ when $U_{j,n,r} \leq m_{n,r}$. Remarking that $G_{j,n,r}\left(m_{n,r}\right)$ is uniformly bounded from below, it remains to control $\mathrm{Var}\left(U_{j,n,r}|U_{j,n,r} \leq m_{n,r}\right)$, using the following steps

$$E(U_{j,n,r}|U_{j,n,r} \leq m_{n,r}) = \left(m_{n,r} - \frac{k_n}{nc_{j,n}} + o\left(\frac{k_n^2}{n^2}\right)\right) G_{j,nr}(m_{n,r})$$

$$E\left(U_{j,n,r}^2|U_{j,n,r} \leq m_{n,r}\right) = \left(m_{n,r}^2 - 2\frac{m_{n,r}k_n}{nc_{j,n}} + 2\frac{k_n^2}{n^2c_{j,n}^2} + o\left(\frac{k_n^2}{n^2}\right)\right) G_{j,nr}(m_{n,r}),$$

to get the result. ■

Then, the independence of the $\{A_{j,n,r}; r = 1, ..., k_n\}$, due to the fundamental independence property of the Poisson point process $\Sigma_{j,n}$ on disjoint subsets of D allows an easy application of Lindeberg's central limit theorem. For more details on the calculus, we refer to [12], Theorem 1.

Lemma 7 *Assume f is $\alpha-$Lipschitzian. If $k_n = o(n/\log n)$ and $n = O\left(k_n^{1+\alpha}\right)$, then for $j = 1, 2$, there exists a bounded sequence $(s_{j,n})$ such that*

$$\frac{nc_{j,n}}{s_{j,n}\sqrt{k_n}}\left(\Delta_{j,n} - E\left(\Delta_{j,n}\right)\right) \to \mathcal{N}\left(0, 1\right) \text{ as } n \to \infty.$$

More precisely, if $n \asymp k_n^{1+\alpha}$ then $s_{j,n} \asymp 1$, and if $n = o\left(k_n^{1+\alpha}\right)$, $s_{j,n} = 1$.

4 Proof of the main result

For $j = 1, 2$, $\lim_{n\to\infty} c_{j,n} = c$, thus the result of Lemma 7 can be rewritten

$$\frac{n}{s_{j,n}\sqrt{k_n}}\left(\Delta_{j,n} - E\left(\Delta_{j,n}\right)\right) \to \mathcal{N}\left(0, c^{-2}\right), j = 1, 2.$$

We are now in position for obtaining Theorem 1, by proving that the difference between $\frac{n}{s_{1,n}\sqrt{k_n}}\left(\Delta_{1,n} - E\left(\Delta_{1,n}\right)\right)$ and $\frac{n}{s_{1,n}\sqrt{k_n}}\left(\Delta_{n} - E\left(\Delta_{n}\right)\right)$ converges in probability to zero under the hypothesis of Lemma 7. First remark that

$$E_n = \left(\Sigma_{1,n} \subseteq \Sigma_n \subseteq \Sigma_{2,n}\right) \subseteq \left(f_{1,n} \leq f_n \leq f_{2,n}\right)$$

so that

$$\begin{aligned}\frac{n}{s_{1,n}\sqrt{k_n}}E\left(\mathbf{1}_{E_n}\left|\Delta_n - \Delta_{1,n}\right|\right) &\leq \frac{n}{s_{1,n}\sqrt{k_n}}E\left(\int_0^1 \left(f_{2,n}\left(x\right) - f_{1,n}\left(x\right)\right)dx\right) \\ &= \frac{n}{s_{1,n}k_n^{3/2}}\sum_{r=1}^{k_n} E\left(U_{2,n,r} - U_{1,n,r}\right).\end{aligned}$$

Then, let $M = \sup\left\{f\left(x\right) : x \in [0, 1]\right\}$: since $f_n \leq M$ and $f_{1,n} \leq M$, we have

$$\frac{n}{s_{1,n}\sqrt{k_n}}E\left(\mathbf{1}_{\Omega\smallsetminus E_n}\left|\Delta_n - \Delta_{1,n}\right|\right) \leq 2M\frac{n}{s_{1,n}\sqrt{k_n}}P\left(\Omega \smallsetminus E_n\right).$$

Lemma 2 and Lemma 3 yield

$$\lim_{n\to\infty}\frac{n}{s_{1,n}\sqrt{k_n}}E\left(\left|\Delta_n - \Delta_{1,n}\right|\right) = 0, \tag{12}$$

so that

$$\frac{n}{s_{1,n}\sqrt{k_n}}\left|\Delta_n - \Delta_{1,n}\right| \xrightarrow{P} 0. \tag{13}$$

From (12) and (13) we finally conclude

$$\frac{n}{s_{1,n}\sqrt{k_n}}\left(\Delta_{1,n}-E\left(\Delta_{1,n}\right)\right)-\frac{n}{s_{1,n}\sqrt{k_n}}\left(\Delta_n-E\left(\Delta_n\right)\right)\overset{P}{\to}0.$$

Taking $s_n = s_{1,n}$, we obtain the result of Theorem 1.

5 Discussion

1) Optimal speed: Theorem 1 is valid under the rate condition $n \asymp k_n^{1+\alpha}$. Of course this gives the best speed of convergence of Δ_n to $\mathcal{N}(0,1)$ in terms of variance. The awkward fact is that, with $n \asymp k_n^{1+\alpha}$, the variance of Δ_n can just be controlled, but not calculated, for a general f.

Note that the choice $n \asymp k_n^{1+\alpha}$ is also optimal for the L_1- error within the class of estimators f_n based upon a sequence satisfying $n = O(k_n^{1+\alpha})$. To see that, let $k_{a,n}$ and $k_{b,n}$ be two sequences of positive integers satisfying the conditions of Lemma 7. Suppose that $n \asymp k_{a,n}^{1+\alpha}$ and that $n = o(k_{b,n}^{1+\alpha})$, and write $\Delta_n^{(a)}$ and $\Delta_n^{(b)}$ for the two corresponding sequences Δ_n.

Following Lemma 4, we have $E(\Delta_n^{(a)}) = O(n^{-\frac{\alpha}{1+\alpha}})$ and $E(\Delta_n^{(b)}) = \frac{k_{b,n}}{nc}(1+o(1))$, thus

$$E(\Delta_n^{(a)})/E(\Delta_n^{(b)}) = O\left(\left(n^{-\frac{\alpha}{1+\alpha}}\right)n/k_{b,n}\right) = O\left(\left(n/k_{b,n}^{1+\alpha}\right)^{\frac{1}{1+\alpha}}\right) = o(1).$$

In the case where f has a bounded derivative, $\alpha = 1$ and the best speed is obtained with $n \asymp k_{a,n}^2$, a result which is in accordance with those of [14].

2) Estimating c: In the result (5), c is unknown. A natural estimate of c is

$$\widetilde{c}_n = \frac{k_n}{\sum_{r=1}^{k_n} U_{n,r}}.$$

Since

$$E\left(\left|\widetilde{c}_n^{-1} - c^{-1}\right|\right) = E\left(\left|\int_0^1 f_n(x)\,dx - \int_0^1 f(x)\,dx\right|\right) \leq E\left(\Delta_n\right),$$

it is readily seen that $\widetilde{c}_n$ converges in probability to c, so that $\widetilde{c}_n$ can be plugged in place of c in (5) without perturbing the result.

Corollary 1 *Assume f is $\alpha-$Lipschitzian. If $k_n = o(n/\log n)$ and $n = O\left(k_n^{1+\alpha}\right)$, then*

$$\frac{n\widetilde{c}_n}{s_n\sqrt{k_n}}\left(\Delta_n - E\left(\Delta_n\right)\right) \to \mathcal{N}\left(0,1\right) \ \ \textit{as } n\to\infty.$$

with $s_n \asymp 1$ if $n \asymp k_n^{1+\alpha}$, and $s_n = 1$ if $n = o\left(k_n^{1+\alpha}\right)$.

3) Centering the sequence: From Lemma 4, the convergence to zero of

$$\frac{nc}{\sqrt{k_n}}E\left(\Delta_n\right) - \sqrt{k_n} = O\left(\frac{n^2}{k_n^{3/2+2\alpha}}\right)$$

strongly depends of the regularity of the unknown function f. Thus, in order to replace $E\left(\Delta_n\right)$ by the sequence $k_n/(nc)$, we must renounce the optimal choice of k_n. In return, we can take $s_n = 1$.

Corollary 2 *Assume f is $\alpha-$Lipschitzian. If $k_n = o(n/\log n)$ and $n = o\left(k_n^{3/4+\alpha}\right)$, then*

$$\frac{nc}{\sqrt{k_n}}\Delta_n - \sqrt{k_n} \to \mathcal{N}\left(0,1\right) \ \ \textit{as } n\to\infty.$$

4) A cleaned result: In order to replace both c and $E\left(\Delta_n\right)$, write

$$\frac{n\widetilde{c}_n}{\sqrt{k_n}}\Delta_n - \sqrt{k_n} = \frac{\widetilde{c}_n}{c}\left(\frac{nc}{\sqrt{k_n}}\Delta_n - \sqrt{k_n}\right) + \widetilde{c}_n\sqrt{k_n}\left(c^{-1} - \widetilde{c}_n^{-1}\right).$$

Under the conditions of Corollary 2, the first term in the above sum converges in distribution to $\mathcal{N}\left(0,1\right)$, but the second term is a $o_P\left(1\right)$ only under stronger conditions. In view of Lemma 4,

$$\sqrt{k_n}E\left(\left|c^{-1} - \widetilde{c}_n^{-1}\right|\right) \leq \sqrt{k_n}E\left(\Delta_n\right) = \frac{k_n^{3/2}}{\left(n+1\right)c} + O\left(\frac{n}{k_n^{1/2+2\alpha}}\right).$$

Note that $k_n = o(n^{2/3})$ and $n = o(k_n^{3/4+\alpha})$ are incompatible for $\alpha \leq 3/4$. Moreover, for $\alpha > 3/4$, $n = o(k_n^{1/2+2\alpha})$ is a weaker condition than $n = o(k_n^{3/4+\alpha})$. Thus we have obtained

Corollary 3 *Assume f is $\alpha-$Lipschitzian. If $k_n = o(n^{2/3})$ and $n = o\left(k_n^{3/4+\alpha}\right)$, then*

$$\frac{n\widetilde{c}_n}{\sqrt{k_n}}\Delta_n - \sqrt{k_n} \to \mathcal{N}(0,1) \text{ as } n \to \infty.$$

For $\alpha = 1$, the possible choices are given by the relations $n^{2/3} \gg k_n \gg n^{4/7}$. This result shows that the asymptotic bias of the L_1- error can be estimated by $k_n/(n\widetilde{c}_n)$, suggesting a bias corrected estimator of f_n defined by

$$g_n(x) = f_n(x) + \frac{1}{n}\sum_{r=1}^{k_n} U_{n,r}.$$

Acknowledgement

The authors are indebted to the anonymous referee and the associate editor for their helpful comments and suggestions. They have contributed to a greatly improved presentation of the results of this paper.

References

[1] Beirlant, J., Györfi, L. and Lugosi, G. (1994) On the asymptotic normality of the L_1- and L_2- errors in histogram density estimation. *Canad. J. Statist.*, **22**(3), 309-318.

[2] Beirlant, J., Györfi, L (1998) On the L_1- error in histogram density estimation: the multidimensional case. *J. Nonparametr. Statist.*, **9**(2),197-216.

[3] Berlinet, A., Devroye, L. and Györfi, L. (1995) Asymptotic normality of L_1- error in density estimation. *Statistics*, 26(4), 329-343.

[4] Devroye, L. and Györfi, L. (1985) *Nonparametric density estimation: the L1 view*, Wiley.

[5] Farrell, M.J. (1957) The measurement of productive efficiency. *Journal of the Royal Statistical Society B*, **120**, 253-281.

[6] Geffroy J. (1964) Sur un problème d'estimation géométrique. *Publications de l'Institut de Statistique de l'Université de Paris*, XIII, 191-200.

[7] Geffroy, J. (2002) Sur un problème de loi-limite pour la distance L_1 entre une fonction et une approximation stochastique de celle-ci. *Technical report ENSAM-INRA-UM2*, 02-01.

[8] Gijbels, I., Mammen, E., Park, B.U. and Simar, L. (1999) On estimation of monotone and concave frontier functions, *Journal of the American Statistical Association,* **94**, 220-228.

[9] Girard, S. and Jacob, P. (2003) Extreme values and Haar series estimates of point processes boundaries. *Scandinavian Journal of Statistics*, **30**(2), 369-384.

[10] Girard, S. and Jacob, P. (2003) Projection estimates of point processes boundaries. *Journal of Statistical Planning and Inference*, **116**(1), 1-15.

[11] Girard, S. and Jacob, P. (2004) Extreme values and kernel estimates of point processes boundaries. *ESAIM: Probability and Statistics*, **8**, 150–168.

[12] Girard, S. and Jacob, P. (2005) Asymptotic normality of the L_1- error for Geffroy's estimate of point process boundaries. *Publications de l'Institut de Statistique de l'Université de Paris*, **XLIX**, 3–17.

[13] Knight, K. (2001) Limiting distributions of linear programming estimators. Extremes, **4**(2), 87–103.

[14] Korostelev, A.P. and Tsybakov, A.B. (1993) Minimax theory of image reconstruction. In *Lecture Notes in Statistics*, 82, Springer-Verlag, New-York.

[15] Mammen, E. and Tsybakov, A. B. (1995) Asymptotical minimax recovery of set with smooth boundaries. *The Annals of Statistics*, **23**(2), 502–524.