Bernoulli **4**(2), 1998, 143–165

On a universal strong law of large numbers for conditional expectations

ANDRZEJ S. KOZEK,[1]* JULIAN R. LESLIE[1] and
EUGENE F. SCHUSTER[2]

[1]*Macquarie University, Department of Statistics C5C, Sydney, NSW 2109, Australia*
[2]*Department of Mathematical Sciences, University of Texas at El Paso, El Paso TX 79968, USA*

A number of generalizations of the Kolmogorov strong law of large numbers are known including convex combinations of random variables (rvs) with random coefficients. In the case of pairs of i.i.d. rvs $(X_1, Y_1), \ldots, (X_n, Y_n)$, with μ being the probability distribution of the xs, the averages of the Ys for which the accompanying Xs are in a vicinity of a given point x may converge with probability 1 (w.p. 1) and for μ-almost everywhere (μ a.e.) x to conditional expectation $r(x) = \mathrm{E}(Y|X = x)$. We consider the Nadaraya–Watson estimator of $\mathrm{E}(Y|X = x)$ where the vicinities of x are determined by window widths h_n. Its convergence towards $r(x)$ w.p. 1 and for μ a.e. x under the condition $\mathrm{E}|Y| < \infty$ is called a strong law of large numbers for conditional expectations (SLLNCE). If no other assumptions on μ except that implied by $\mathrm{E}|Y| < \infty$ are required then the SLLNCE is called universal. In the present paper we investigate the minimal assumptions for the SLLNCE and for the universal SLLNCE. We improve the best-known results in this direction.

Keywords: conditional expectation; kernel estimator; Nadaraya–Watson estimator; nonparametric regression; strong convergence; strong law of large numbers; universal convergence

1. Introduction and summary

Let (X, Y) be a $(d+1)$-dimensional random vector (rv), $X \in \mathbb{R}^d$, $Y \in \mathbb{R}^1$, and let μ stand for the probability distribution of X. Throughout the paper we shall use the abbreviation r.v. to denote either a random vector or a random variable. The regression function $r(x)$ is defined by

$$r(x) = \mathrm{E}(Y|X = x). \tag{1.1}$$

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent, identically distributed (i.i.d.) copies of (X, Y). One of the simplest and most investigated estimators of $r(x)$, the Nadaraya–Watson (NW) estimator (Nadaraya 1964; Watson 1964), is given by

$$\hat{r}_n(x) = \textstyle\sum_{i=1}^n Y_i K\left(\frac{x - X_i}{h_n}\right) \Big/ \textstyle\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) \tag{1.2}$$

*To whom correspondence should be addressed. e-mail: akozek@zen.efs.mq.edu.au where $K(x)$ is a kernel and h_n is a window width. In this paper we restrict our attention to kernel $K(x)$ being an indicator function of a unit ball in $\mathbb{R}^d$.

The NW estimator has the form of an arithmetical mean of those rvs Y which have the accompanying X component in the h_n vicinity of x. Hence its strong convergence under the condition $\mathrm{E}|Y| < \infty$ corresponds to the strong law of large numbers for conditional expectations (SLLNCE) and goes beyond the classical extensions of the Kolmogorov strong law of large numbers (SLLN) in which the arithmetic mean has been replaced with convex combinations, also random, but in a way preserving convergence to the expected value; we refer the reader to Howell *et al.* (1981) and Taylor and Calhoun (1983) for results in this direction. If no other assumptions on μ except that implied by $\mathrm{E}|Y| < \infty$ are required then the SLLNCE is called *universal*.

The existing theory of the strong convergence of the NW estimator requires either the existence of Lebesgue density of X – and many other technical assumptions (see Mukerjee 1989) – but only the first finite moment of Y, or assumptions on the existence of s-moments with $s > 1$; cf. Zhao and Fang (1985) and Stute (1986). It is natural to ask if the universal SLLNCE holds true and in the present paper we investigate this problem.

The study of the universal convergence, where only moment conditions on Y and no other conditions on μ are required, began in Stone (1977), where the convergence in probability and in L^p were proved. The universal convergence in probability and for μ a.e. xs of the NW estimator to the regression function and under the assumption of the first moment finite only, was proved in Greblicki *et al.* (1984). Research towards the universal SLLNCE has a long history; however, we mention only those papers that are most relevant for the present approach: Devroye (1981), Cheng (1983), Greblicki *et al.* (1984), Cheng and Zhao (1985), Zhao and Fang (1985), Stute (1986) and Mukerjee (1989). Our main result towards the universal SLLNCE is given in Theorem 1, where, under weak regularity assumptions on the marginal distribution of X, we prove an SLLNCE for $\hat{r}_n(x)$. Example 1 suggests that Theorem 1 cannot be essentially improved. Let us note that, as in proofs of the classical SLLNCE, to avoid assumptions of higher moments in Theorem 1, we consider truncated random variables. Our proof of the equivalence of the estimators based on the truncated and non-truncated random variables requires, however, Assumption κ (see Section 2) on the distribution of X to be met.

The existence of slightly higher moments than the first moment makes it possible to avoid the truncation step (cf. Theorem 2) so that at the cost of some moment conditions one can achieve a universal strong consistency of the NW estimator. Alternatively, one can consider just estimators based on the truncated random variables. Estimators of this type were first considered in Schuster (1968, pp. 45 and 59), where, under additional conditions, their uniform and strong convergence was proved, and next in Cheng (1983), where their strong and universal consistency was obtained. By skipping the truncation step in the proof of Theorem 1 we obtain in Theorem 3 a strong and universal consistency of a version of these estimators. The proof of Theorem 3 and the truncation used therein differ from those in Schuster (1968) and Cheng (1983).

Hence we have the following three options available.

- If 'logarithmic' moments (2.12) are finite then the universal and strong convergence of the NW estimator holds true (cf. Theorem 2).

- If the marginal distribution of X meets Assumption κ then the SLLNCE is valid (cf. Theorem 1). In Example 1 and Remark 5 we show that Assumption κ may not be fulfilled by some atomless Lebesgue singular probability measures.
- In the case of the NW estimator based on truncated random variables, the universal strong consistency obtains (cf. Theorem 3).

Let us note that the moment conditions required in Theorem 2 are weaker than those in Greblicki *et al.* (1984), Zhao and Fang (1985) and Stute (1986). Assumption κ, used in Theorem 1, is in turn much weaker than the requirement of a differentiable probability density function of X in Mukerjee (1989). Our method of proof also avoids other restrictive technical assumptions in Mukerjee (1989).

The paper is organized in the following way. In Section 2 we present the main results of the paper. Section 3 contains the proofs of the theorems. In the Appendix (Section 4) we collect some known technical results and extend them to the form needed in the paper.

2. The main results

Let $I_{\mathscr{S}}(x)$ stand for the indicator function of set $\mathscr{S}$ and $K(x)$ for the kernel equal to the indicator function of $B(0, 1)$, the unit ball in $\mathbb{R}^d$ centred at $0 \in \mathbb{R}^d$. The expected number of X_is in an h_n-vicinity of x is given by

$$\gamma_x(n) = n \cdot \mu_x(h_n), \tag{2.1}$$

where $h_n = C \cdot n^{-\delta}$, $\mu_x(r)$ is the μ-measure of the ball of radius r centred at $x \in \mathbb{R}^d$, $0 < \delta < d^{-1}$, and μ is a probability distribution of X. Since the argument x of the NW estimator will be considered fixed we shall suppress x whenever it does not lead to any confusion. Hence we shall write $\gamma(n)$ for $\gamma_x(n)$. We shall see that a truncated NW estimator (2.15) is equal w.p. 1 for all but finitely many ns to the original NW estimator (1.2) provided there exist constants $c_1 = c_1(x)$, $c_2 = c_2(x)$, and $\kappa = \kappa(x) \in [1 - d\delta, 1]$ such that

$$c_1 \leqslant \liminf_{n\to\infty} \frac{\gamma(n)}{n^\kappa} \leqslant \limsup_{n\to\infty} \frac{\gamma(n)}{n^\kappa} \leqslant c_2. \tag{2.2}$$

Let us note that by Corollary 2 of Lemma 4 in the Appendix there exists a μ a.e. positive function $c(x)$ such that $c(x)n^{1-d\delta} \leqslant \gamma(n) \leqslant n$. It is clear that (2.2) is met with $\kappa = 1$ for every atom x of μ. Moreover, by the Lebesgue theorem (Theorem 10.49 in Wheeden and Zygmund 1977) condition (2.2) is met with $\kappa = 1 - d\delta$ on the set on which μ is Lebesgue absolutely continuous.

Assumption κ. *We shall assume that there is a countable collection of $\kappa_i \in [1 - d\delta, 1]$ such that for μ a.e. x there exists a corresponding κ_i such that (2.2) holds true for $\kappa = \kappa_i$.*

Not every probability measure on $\mathbb{R}^d$ meets Assumption κ which requires, in effect, a constant rate of decrease for the μ-measure of an h_n-ball when $h_n \to 0$. It is met in the case of the Lebesgue absolutely continuous densities, discrete distributions, rectifiable measures (which correspond to a smooth transport of a Lebesgue absolutely continuous measure onto lower than d-dimensional manifolds; cf. Preiss 1987), and also in the case of a range of Lebesgue singular measures including the Cantor measure (i.e. a 'uniform' measure on the Cantor set; see Example 1 below for more details). It is also met by probability distributions μ which can be represented as countable sums of components, each component being a finite measure of one of the listed types. Example 1 below shows, however, that there exists a probability measure (which we will call for convenience a Cantor–Preiss measure) which does not meet Assumption κ for μ-almost every x. Our main result on the SLLNCE requires Assumption κ and can be formulated in the following way.

Theorem 1. *Let the kernel $K(x) = I_{B(0,1)}(x)$ and the window width $h_n = C \cdot n^{-\delta}$ for some $\delta \in (0, 1/d)$. If $(X_1, Y_1), \dots, (X_n, Y_n)$ are i.i.d. rvs in $\mathbb{R}^d \times \mathbb{R}$, $E|Y| < \infty$, and Assumption κ is met then for μ a.e. $x \in \mathbb{R}^d$ and with probability 1*

$$\lim_{n\to\infty} \hat{r}_n(x) = r(x),$$

where $\hat{r}_n(x)$ is the NW estimator given by (1.2).

Example 1. Let μ be a probability distribution of a random variable $X = 2\sum_{i=1}^{\infty} U_i/3^i$, where U_i are independent $\{0, 1\}$ random variables such that $p_i = \Pr(U_i = 1) = 1/(\tau_i + 1)$. If we set $\tau_i = 1$ and $\mu_C = \mu$ we call μ_C a uniform Cantor measure. In the case of $\tau_i = i$ we put $\mu_{CP} = \mu$ and shall call μ_{CP} a Cantor–Preiss measure. In Lemma 3 we shall prove that the uniform Cantor measure meets Assumption κ with $\kappa = 1 - (\ln 2/\ln 3) \cdot \delta$, while the Cantor–Preiss measure does not. In the latter case for μ_{CP} a.e. x inequality (2.2) obtains only either with $c_1 = 0$ or with $c_2 = \infty$. Note that in the case of the μ_C-measure, $\ln 2/\ln 3$ is the Hausdorff–Besicovitch dimension of the Cantor set. (Probability measure μ_{CP} and some of its pathological properties (e.g. $\limsup_{h\to 0}(\mu_{CP}(\beta h)/\mu_{CP}(h)) = \infty$, μ_{CP} a.e. and, $\forall \beta > 1$) have been communicated to us by Professor David Preiss in a private letter in 1993 to one of the authors.)

The existence of slightly higher moments than the first moment implies the universal convergence for the NW estimator without referring to Assumption κ and without laborious truncation of variables in the proof of the theorem. To this end, we shall consider a class of convex functions which have properties similar to the square function and which contains functions increasing to infinity slower than any $(1 + \epsilon)$-power function. The symbol C will be used in several different settings to indicate a non-negative constant.

Let $\mathscr{K}$ consist of all symmetric convex functions $\Phi(t)$ with the following properties:

$$\phi(2t) < C\phi(t), \text{ where } \phi(t) = \Phi'(t),\ t \geqslant 0, \tag{2.3}$$

$$\Phi(0) = 0, \tag{2.4}$$

$$\Phi(\sqrt{t}) \text{ is subadditive.} \tag{2.5}$$

It is easy to see that if $\Phi \in \mathscr{K}$, then for every $a > 0$ there exist positive constants $C_1(a)$ and $C_2(a)$ such that for every t

$$C_1(a)\Phi(t) \leqslant \Phi(at) \leqslant C_2(a)\Phi(t). \tag{2.6}$$

Let us recall that by convexity of Φ the function $\phi(t)$ is non-decreasing. Moreover, functions on $\mathbb{R}^+$ with a non-increasing positive derivative are subadditive. Hence, whenever $\phi(t)/t$ is non-increasing, $\Phi(\sqrt{t})$ is subadditive, i.e. condition (2.5) is satisfied. It is easy to see that the following functions belong to class $\mathscr{K}$:

$$\Phi_r(t) = |t|^r, \qquad r \in (1, 2],$$

$$\Phi_{(k,r)}(t) = |t|(C_{(k,r)} + \mathrm{lon}_{(k,r)}(t)),$$

where $k \geqslant 1$, $r > 1$, $C_{(k,r)}$ is a sufficiently large constant and $\mathrm{lon}_{(k,r)}(t)$ is given by

$$\mathrm{lon}_{(1)}(t) = \ln(1+t), \qquad t > 0,$$

$$\mathrm{lon}_{(k)}(t) = \mathrm{lon}_{(1)}(\mathrm{lon}_{(k-1)}(t)), \qquad k > 1,\ t > 0,$$

$$\mathrm{lon}_{(k,r)}(t) = \mathrm{lon}_{(1)}(t)\mathrm{lon}_{(2)}(t)\ldots\mathrm{lon}_{(k-1)}(t)(\mathrm{lon}_{(k)}(t))^r \tag{2.7}$$

for integer $k \geqslant 1$, $r > 1$ and $t > 0$.

The importance of the class $\mathscr{K}$ follows from the Burkholder–Gundy inequality (cf. Burkholder and Gundy 1970, Corollary 5.4, p. 283),

$$\mathrm{E}\Phi\left(\sum_{i=1}^{n} Z_i\right) \leqslant C\mathrm{E}\Phi\left(\left(\sum_{i=1}^{n} Z_i^2\right)^{1/2}\right),$$

valid for independent random variables Z_i, each with expectation zero, and for Φ symmetric, convex, and satisfying (2.3) and (2.4). The Burkholder–Gundy inequality and (2.5) imply immediately that if $Z_1, Z_2, \ldots, Z_n$ are independent centred random variables, and $\Phi \in \mathscr{K}$, then there exists a constant C such that

$$\mathrm{E}\Phi\left(\sum_{i=1}^{n} Z_i\right) \leqslant C\sum_{i=1}^{n}\mathrm{E}\Phi(Z_i). \tag{2.8}$$

Let us note that a class of functions similar to $\mathscr{K}$ is often used in the classical proofs of the strong convergence of weighted sums of independent rvs using Kolmogorov's three series theorem (cf., e.g., Petrov 1975). In the present case we shall apply Kolmogorov's maximal inequality on the level of conditional probability distributions; the three series theorem technique does not seem easily applicable here.

Theorem 2. *Let the kernel $K(x) = I_{B(0,1)}(x)$, the window width $h_n \searrow 0$, $\Phi \in \mathscr{K}$, and*

$$\psi(t) = \Phi(t)/|t|. \tag{2.9}$$

If $(X_1, Y_1), \ldots, (X_n, Y_n)$ *are i.i.d. rvs in* $\mathbb{R}^d \times \mathbb{R}$,

$$\mathrm{E}\Phi(|Y|) < \infty, \tag{2.10}$$

and

$$\sum_{n=1}^{\infty} 1 \Big/ (n\psi(nh_n^d)) < \infty, \tag{2.11}$$

then for μ *a.e.* $x \in \mathbb{R}^d$ *and with probability 1*

$$\lim_{n\to\infty} \hat{r}_n(x) = r(x),$$

where $\hat{r}_n(x)$ *is the NW estimator given by (1.2).*

Theorem 2 implies readily the following particular case showing the extent to which the requirement of finite moments can be weakened in the case of universal convergence. Let us note that assumption (2.12) is weaker than the condition $\mathrm{E}|Y|^s < \infty$ for $s > 1$, used in Zhao and Fang (1985) and in Stute (1986).

Corollary 1. *Let the kernel* $K(x) = I_{B(0,1)}(x)$ *and the window width* $h_n = C \cdot n^{-\delta}$ *for some* $\delta \in (0, 1/d)$. *If* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *are i.i.d. rvs in* $\mathbb{R}^d \times \mathbb{R}$ *and*

$$\mathrm{E}|Y|\mathrm{lon}_{(k,r)}(|Y|) < \infty \tag{2.12}$$

for some $k \geqslant 1$ *and* $r > 1$, *then for* μ *a.e.* $x \in \mathbb{R}^d$ *and with probability 1*

$$\lim_{n\to\infty} \hat{r}_n(x) = r(x).$$

Remark 1. Condition (2.11) provides a link between the moments (2.10) and the window width allowed for the convergence. The higher the moments of Y that are finite the narrower the window widths h_n that are allowed. Let us note that if for some positive ϵ $\mathrm{E}|Y|^{1+\epsilon} < \infty$, then condition (2.11) is met with

$$h_n = \left(\frac{(\log n)^{2/\epsilon}}{n}\right)^{1/d}. \tag{2.13}$$

This is a weaker requirement than the corresponding conditions $\inf_{n>0}\{n^{\epsilon/2} \cdot h_n^d > 0\}$ in Zhao and Fang (1985, Theorem 1) and $\sum_{n\geqslant 1} \exp(-\rho n h_n^d) < \infty$ for all $\rho > 0$ in Stute (1985, Theorem 1). Zhao and Fang (1985, Theorem 2) require the exponential moments $\mathrm{E}\exp(t|Y|^\lambda) < \infty$ to allow window width (2.13).

Remark 2. Etemadi (1981) presented the beautiful idea of using monotonicity arguments to simplify Kolmogorov's proof of the strong law of large numbers. This was in fact a starting point for the present paper. However, elementary arguments can be used along these lines to prove consistency of the Nadaraya–Watson estimate only under the generalized moment assumption, like in our Theorem 2, as well as assuming that the X-variable has a Lebesgue density. Our efforts to get around this obstacle resulted in the present version of the paper.

Finally, we shall consider an estimator based on truncated random variables where the truncation depends on the sample size. Let $k(p) = 2^p$ for non-negative integers p. For each positive integer n there is a unique integer $p = p(n)$ such that $n \in (k(p-1), k(p)]$, so that without any confusion we can omit the argument n at $p(n)$. Let C_1 be a constant independent of n and

$$\tilde{Y}_{i,n} = \begin{cases} Y_i & \text{if } |Y_i| \leqslant C_1 \cdot q_0^p,\ i \leqslant n, \text{ and } n \in (k(p-1), k(p)] \\ 0 & \text{otherwise,} \end{cases} \tag{2.14}$$

where $q_0 = 2^{1-d\delta}$ and δ is a parameter of the window width.

Theorem 3. *Let the kernel $K(x) = I_{B(0,1)}(x)$, and the window width $h_n = C \cdot n^{-\delta}$ for some $\delta \in (0, 1/d)$. If $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. rvs in $\mathbb{R}^d \times \mathbb{R}$ and*

$$\hat{r}_n(x) = \frac{\sum_{i=1}^n \tilde{Y}_{i,n} K\left(\dfrac{x - X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\dfrac{x - X_i}{h_n}\right)}, \tag{2.15}$$

where $\tilde{Y}_{i,n}$ is given by (2.14), then for μ a.e. $x \in \mathbb{R}^d$ and with probability 1

$$\lim_{n\to\infty} \hat{r}_n(x) = r(x). \tag{2.16}$$

Remark 3. Theorem 3 does not require moments of Y to be finite or any assumption on the distribution of X, and hence it may be called a universal consistency of the estimator given by (2.15). However, because of truncation (2.14), it cannot be called a law of large numbers. In a sense it is similar in structure to robust estimators and suggests that robust estimators may be strongly universally consistent without the rather strong technical assumptions required for related results obtained elsewhere (cf. Härdle *et al.* 1988; Hall and Jones 1990). Consistency of the robust modifications of the NW estimator is, however, beyond the scope of the present paper.

3. Proofs of the theorems

To simplify presentation, we introduce notation which lets us avoid overloading symbols with numerous parameters. Let us recall that for integer p we write $k(p) = 2^p$ and that for every integer n there is a unique $p = p(n)$ such that

$$k(p-1) < n \leqslant k(p). \tag{3.1}$$

So, without any confusion, we can write p instead of $p(n)$. For most of the proof the argument x of the NW estimator will be considered fixed and, when there is no confusion, we shall suppress x altogether. Hence, with the sequence of window widths h_n given, we shall write

$$K_{i,m} = K_m(X_i) = K\left(\frac{x - X_i}{h_m}\right), \qquad \mathbf{K}_{k(p-1)} = (K_{1,k(p-1)}, \ldots, K_{k(p),k(p-1)}), \tag{3.2}$$

$$\mu(m) = \mu_x(h_m) = \mathrm{E}K_m(X_1), \tag{3.3}$$

and

$$w(\nu, m) = \sum_{i=1}^{\nu} K\left(\frac{x - X_i}{h_m}\right). \tag{3.4}$$

Given $K_{i,m} = K((x - X_i)/h_m)$ for $i = 1, \ldots, N$, we use this to define an ordering of the random variables $X_1, \ldots, X_N$ and an induced ordering of the matching $Y_1, \ldots, Y_N$, and of their centred counterparts of a truncated version defined in (3.19). This ordering assigns lower ranks to variables for which the matching $K_{i,m}$ variables have value 1, higher ranks to variables for which the matching $K_{i,m}$ variables have value 0, and in each of these two groups the ranks are assigned in chronological order (i.e. in the order in which they originally appeared in the sample). We shall consider two particular rankings corresponding to different pairs (m, N): (n)-ranking corresponding to $m = n$ and $N = n$; and (p)-ranking corresponding to $m = k(p-1)$ and $N = k(p)$.

Clearly, in general (n)- and (p)-rankings are different. We shall write N_ν for the (p)-rank of the element having rank ν in the (n)-ranking. Moreover, let $(j)_n$ and $(j)_p$ stand for the indices i and k of the Y_i and Y_k for which in the (n)-ranking $\mathrm{rank}_{(n)}(Y_i) = j$ and in the (p)-ranking $\mathrm{rank}_{(p)}(Y_k) = j$, respectively.

Since $K_{i,n} \leqslant K_{i,k(p-1)}$, we obtain that for $\nu \leqslant w(n, n)$

$$\nu \leqslant N_\nu = w((\nu)_n, k(p-1)).$$

Let us note that the following relation holds true for $\nu \in \{1, \ldots, w(n, n)\}$:

$$\sum_{j=1}^{\nu} Y_{(j)_n} = \sum_{j=1}^{N_\nu} Y_{(j)_p} - \sum_{j=1}^{N_\nu} Y_{(j)_p} \cdot Q_{(j)_p,n}, \tag{3.5}$$

where

$$Q_{(j)_p,n} = K_{(j)_p,k(p-1)} - K_{(j)_p,n} = K\left(\frac{x - X_{(j)_p}}{h_{k(p-1)}}\right) - K\left(\frac{x - X_{(j)_p}}{h_n}\right).$$

Clearly, relation (3.5) is also valid for the truncated random variables given by (3.19).

Proof of Theorem 1. Let us note that Bernstein's inequality implies easily (cf. Devroye 1981, Theorem 4.2; or Greblicki *et al.* 1984, Theorem 2) that almost surely and for μ a.e. x

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} K_n(X_i)}{n \cdot \mu(n)} = 1. \tag{3.6}$$

Hence, to prove Theorem 1 it is enough to prove the μ a.s. convergence of $\tilde{r}_n(x)$ given by

$$\tilde{r}_n(x) = \frac{\sum_{i=1}^{n} Y_i K_n(X_i)}{n \cdot \mu(n)}. \tag{3.7}$$

(Notice that (3.7) implies (3.6); however, we shall need (3.6) in the proof of (3.7).)

We shall consider truncated random variables and the truncation will depend both on p and x, the point at which we are estimating $r(x)$.

Let

$$\bar{Y}_p = \begin{cases} Y & \text{if } |Y| \leqslant q^p, \\ 0 & \text{otherwise,} \end{cases} \tag{3.8}$$

and

$$\bar{Y}_{i,p} = \begin{cases} Y_i & \text{if } |Y_i| \leqslant q^p \text{ and } i \leqslant k(p), \\ 0 & \text{otherwise,} \end{cases} \tag{3.9}$$

where $q = 2^{\kappa_j}$ and κ_j depends on x (cf. Assumption κ). For simplicity we shall, however, consider the case when one κ meets Assumption κ for μ a.e. x. To obtain the general case one can apply the argument that follows for every κ_j separately.

Consider $\hat{r}_n^{(p)}(x)$ corresponding to $\tilde{r}_n(x)$ but built up of the truncated random variables and given by

$$\hat{r}_n^{(p)}(x) = \frac{\sum_{i=1}^{n} \bar{Y}_{i,p} K_n(X_i)}{n \cdot \mu(n)}. \tag{3.10}$$

For $n \in (k(p-1),\, k(p)]$ we have

$$\hat{r}_n^{(p)}(x) \neq \tilde{r}_n(x) \Rightarrow \exists\, i \in (1,\, k(p)]\colon\ Y_i \cdot K_{i,n} \neq \bar{Y}_{i,p} \cdot K_{i,n}, \tag{3.11}$$

and hence

$$\Pr(\exists\, n \in (k(p-1),\, k(p)]\colon\ \hat{r}_n^{(p)}(x) \neq \tilde{r}_n(x)) \leqslant k(p) \cdot \Pr(|Y| > q^p,\ K_{1,k(p-1)} = 1). \tag{3.12}$$

Let us note that

$$\sum_{p=1}^{\infty} k(p) \cdot \Pr(|Y| > q^p,\ K_{1,k(p-1)} = 1) = \sum_{p=1}^{\infty} \gamma(k(p)) \cdot \frac{\Pr(|Y| > q^p,\ K_{1,k(p-1)} = 1)}{\mu(k(p))}. \tag{3.13}$$

By (2.2) and Lemma 7 applied to each component of the series we obtain, with M_2 as in Lemma 5,

$$\int \sum_{p=1}^{\infty} \gamma(k(p)) \cdot \frac{\Pr(|Y| > q^p,\ K_{1,k(p-1)} = 1)}{\mu(k(p))} \mu(\mathrm{d}x) \leqslant c_2 \cdot M_2 \sum_{p=1}^{\infty} q^p \Pr(|Y| > q^p)$$

$$\leqslant c_2 \cdot \frac{q}{q-1} \cdot M_2 \mathrm{E}|Y| < \infty. \tag{3.14}$$

Hence, by the Borel-Cantelli lemma it is enough to show that w.p. 1 and for μ a.e. x,

$$\lim_{n\to\infty} \hat{r}_n^{(p)}(x) = r(x). \tag{3.15}$$

Let us note that for $i = 1, \ldots, k(p)$ we have

$$\mathrm{E}(\bar{Y}_{i,p}|K_{i,n}) = \mathrm{E}(\bar{Y}_{i,p}|K_{1,n}, \ldots, K_{n,n}) = \begin{cases} \dfrac{\int \bar{Y}_p \cdot K_{1,n}\,\mathrm{d}P}{\int K_{1,n}\,\mathrm{d}P} & \text{if } K_{i,n} = 1, \\[2ex] \dfrac{\int \bar{Y}_p \cdot (1 - K_{1,n})\,\mathrm{d}P}{\int (1 - K_{1,n})\,\mathrm{d}P} & \text{if } K_{i,n} = 0. \end{cases} \tag{3.16}$$

Putting

$$\bar{r}_{k(p-1)}(x) = \int \bar{Y}_p \cdot K_{1,k(p-1)}\,\mathrm{d}P \Big/ \int K_{1,k(p-1)}\,\mathrm{d}P \tag{3.17}$$

we obtain

$$\hat{r}_n^{(p)}(x) - r(x) = \frac{\sum_{i=1}^n (\bar{Y}_{i,p} - \bar{r}_{k(p-1)}(x)) K_n(X_i)}{n \cdot \mu(n)} + \frac{\sum_{i=1}^n K_n(X_i)}{n \cdot \mu(n)} \cdot \bar{r}_{k(p-1)}(x) - r(x)$$

$$= I_n^{(1)}(x) + I_n^{(2)}(x). \tag{3.18}$$

Convergence of $I_n^{(1)}(x)$. Let $n \in (k(p-1), k(p)]$. For $j = 1, \ldots, k(p)$ let

$$\bar{Z}_{j,p} = \bar{Y}_{j,p} - \mathrm{E}(\bar{Y}_{j,p}|K_{j,k(p-1)}); \tag{3.19}$$

for $j = 1, \ldots, w(n, n)$ let

$$\bar{Z}_{(j)_n} = \bar{Y}_{(j)_n,p} - \bar{r}_{k(p-1)}; \tag{3.20}$$

and for $j = 1, \ldots, w(k(p), k(p-1))$ let

$$\bar{Z}_{(j)_p} = \bar{Y}_{(j)_p,p} - \bar{r}_{k(p-1)}. \tag{3.21}$$

By (3.5) we obtain for $\nu = 1, \ldots, w(n, n)$

$$\sum_{j=1}^{\nu} \bar{Z}_{(j)_n} = \sum_{j=1}^{N_\nu} \bar{Z}_{(j)_p} - \sum_{j=1}^{N_\nu} \bar{Z}_{(j)_p} \cdot Q_{(j)_p,n}. \tag{3.22}$$

Let events $\mathscr{A}_n$, $\mathscr{B}_n$ and $\mathscr{D}_p$ be given by

$$\mathscr{A}_n = \left\{ \max_{\nu=1,\ldots,w(n,n)} \left| \sum_{j=1}^{\nu} \bar{Z}_{(j)_n} \right| > 4\epsilon \cdot k(p-1) \cdot \mu(k(p)) \right\} \tag{3.23}$$

$$\mathscr{B}_n = \left\{ \max_{\nu=1,\ldots,w(k(p),k(p-1))} \left| \sum_{j=1}^{\nu} \bar{Z}_{(j)_p} \cdot Q_{(j)_p,n} \right| \leqslant \epsilon \cdot k(p) \cdot \mu(k(p)) \right\} \tag{3.24}$$

$$\mathscr{D}_p = \left\{ \max_{\nu=1,\ldots,w(k(p),k(p-1))} \left| \sum_{j=1}^{\nu} \bar{Z}_{(j)_p} \right| > \epsilon \cdot k(p) \cdot \mu(k(p)) \right\}. \tag{3.25}$$

By (3.22) we have $\mathscr{A}_n \cap \mathscr{B}_n \subset \mathscr{D}_p$. Moreover,

$$\begin{aligned} \{|I_n^{(1)}(x)| > 4\epsilon\} &= \left\{ \left| \sum_{i=1}^{n} \bar{Z}_{i,p} K_n(X_i) \right| > 4\epsilon \cdot n \cdot \mu(n) \right\} \\ &\subset \left\{ \max_{\nu=1,\ldots,w(n,n)} \left| \sum_{j=1}^{\nu} \bar{Z}_{(j)_n} \right| > 4\epsilon \cdot n \cdot \mu(n) \right\} \subset \mathscr{A}_n. \end{aligned} \tag{3.26}$$

We will prove in Lemma 1 that $\sum_{p=1}^{\infty} \Pr(\mathscr{D}_p) < \infty$ and in Lemma 2 that $\lim_{n\to\infty} \Pr(\mathscr{B}_n) = 1$. Since events $\mathscr{A}_n$ and $\mathscr{B}_n$ are independent, we obtain by the events lemma (Loève 1977, Section 18.1.c, p. 258) that

$$0 = \Pr(\limsup_p \mathscr{D}_p) \geqslant \Pr(\limsup_n \mathscr{A}_n \mathscr{B}_n) \geqslant \tfrac{1}{2} \Pr(\limsup_n \mathscr{A}_n), \tag{3.27}$$

which implies the convergence of $I_n^{(1)}(x)$. Hence, it remains to prove Lemmas 1 and 2.

Lemma 1. $\sum_{p=1}^{\infty} \Pr(\mathscr{D}_p) < \infty$.

Proof. Let β_p denote the joint probability distribution $\mathbf{K}_{k(p-1)}$ given by (3.2) and notice that by independence of $Y_1, \ldots, Y_{k(p)}$ we have that for each $q_1, \ldots, q_{k(p)}$ the conditional probability distribution of $(\bar{Z}_{(1)_p}, \ldots, \bar{Z}_{(k(p))_p})$ given $K_{1,k(p-1)} = q_1, \ldots, K_{k(p),k(p-1)} = q_{k(p)}$ is a product measure of its components, i.e., $\bar{Z}_{(1)_p}, \ldots, \bar{Z}_{(k(p))_p}$ are conditionally independent. By (3.16), (3.17) and (3.19)–(3.21) they are also centred at the conditional expectation. Hence, by Lemma 9, we can apply Kolmogorov's maximal inequality to the conditional distributions and to the particular permutation of $\bar{Z}$ variables determined by the (p)-ordering. We obtain

$$\begin{aligned}
\Pr(\mathscr{D}_p) &= \int \Pr\left(\max_{\nu=1,\dots,w(k(p),k(p-1))} \left| \sum_{j=1}^{\nu} \bar{Z}_{(j)_p} \right| > \epsilon \cdot k(p) \cdot \mu(k(p)) | \mathbf{K}_{k(p-1)} \right) \mathrm{d}\beta_p \\
&= \int \Pr\left(\max_{\nu=1,\dots,k(p)} \left| \sum_{j=1}^{\nu} \bar{Z}_{(j)_p} \cdot K_{(j)_p,k(p-1)} \right| > \epsilon \cdot k(p) \cdot \mu(k(p)) | \mathbf{K}_{k(p-1)} \right) \mathrm{d}\beta_p \\
&\leqslant \int \frac{\sum_{j=1}^{k(p)} \mathrm{var}\,(\bar{Z}_{(j)_p} K_{(j)_p,k(p-1)} | \mathbf{K}_{k(p-1)})}{(\epsilon \cdot k(p) \cdot \mu(k(p)))^2} \,\mathrm{d}\beta_p \qquad (3.28) \\
&\leqslant \int \frac{\sum_{j=1}^{k(p)} \mathrm{E}(\bar{Y}^2_{(j)_p} K_{(j)_p,k(p-1)} | \mathbf{K}_{k(p-1)})}{(\epsilon \cdot k(p) \cdot \mu(k(p)))^2} \,\mathrm{d}\beta_p \\
&= \frac{k(p)\mathrm{E}\bar{Y}^2_{1,p} K_{k(p-1)}(X_1)}{(\epsilon \cdot k(p) \cdot \mu(k(p)))^2}. \qquad (3.29)
\end{aligned}$$

Hence, by (2.2) and for $q = 2^{\kappa}$ we obtain that, for all but a finite number of p,

$$\Pr(\mathscr{D}_p) \leqslant \frac{\mathrm{E}\bar{Y}^2_{1,p} K_{k(p-1)}(X_1)}{\epsilon^2 c_1 q^p \cdot \mu(k(p))}. \qquad (3.30)$$

Let $\lceil t \rceil$ stand for the smallest integer greater than or equal to t. Notice that

$$\begin{aligned}
\mathrm{E}\bar{Y}^2_{1,p} K_{k(p-1)}(X_1) &\leqslant \sum_{j=1}^{\lceil q^p \rceil} j^2 \Pr(j-1 < |Y| \leqslant j,\ K_{k(p-1)}(X) = 1) \\
&\leqslant 2 \sum_{j=1}^{\lceil q^p \rceil} \sum_{\alpha=1}^{j} \alpha \Pr(j-1 < |Y| \leqslant j,\ K_{k(p-1)}(X) = 1) \\
&\leqslant 2 \sum_{\alpha=1}^{\lceil q^p \rceil} \sum_{j=\alpha}^{\lceil q^p \rceil} \alpha \Pr(j-1 < |Y| \leqslant j,\ K_{k(p-1)}(X) = 1) \\
&\leqslant 2 \sum_{\alpha=1}^{\lceil q^p \rceil} \alpha \Pr(\alpha-1 < |Y| \leqslant \lceil q^p \rceil,\ K_{k(p-1)}(X) = 1) \\
&\leqslant 2 \sum_{\nu=1}^{p} \sum_{\alpha=\lceil q^{\nu-1} \rceil+1}^{\lceil q^{\nu} \rceil} \lceil q^{\nu} \rceil \Pr(\alpha-1 < |Y| \leqslant \lceil q^p \rceil,\ K_{k(p-1)}(X) = 1) \\
&\quad + 2\Pr(0 < |Y| \leqslant \lceil q^p \rceil,\ K_{k(p-1)}(X) = 1) = 2(J_1(p) + J_2(p)). \qquad (3.31)
\end{aligned}$$

Passing to summation over p, we have

$$\sum_{p=1}^{\infty} \Pr(\mathscr{D}_p) \leqslant \frac{1}{\epsilon^2 c_1^2} \cdot \left(\sum_{p=1}^{\infty} \frac{J_1(p)}{q^p \cdot \mu(k(p))} + \sum_{p=1}^{\infty} \frac{J_2(p)}{q^p \cdot \mu(k(p))} \right). \tag{3.32}$$

Now, we shall prove that both series on the right-hand side of (3.32) are finite μ a.e. First, let us change the order of summation:

$$\begin{aligned} \sum_{p=1}^{\infty} \frac{J_1(p)}{q^p \cdot \mu(k(p))} &= \sum_{\nu=1}^{\infty} \sum_{p=\nu}^{\infty} \sum_{\alpha=\lceil q^{\nu-1}\rceil+1}^{\lceil q^{\nu}\rceil} \frac{\lceil q^{\nu}\rceil \Pr(\alpha - 1 < |Y| \leqslant \lceil q^p \rceil, K_{k(p-1)}(X) = 1)}{q^p \cdot \mu(k(p))} \\ &= \sum_{\nu=1}^{\infty} \sum_{\alpha=\lceil q^{\nu-1}\rceil+1}^{\lceil q^{\nu}\rceil} \sum_{p=\nu}^{\infty} \frac{\lceil q^{\nu}\rceil \Pr(\alpha - 1 < |Y| \leqslant \lceil q^p \rceil, K_{k(p-1)}(X) = 1)}{q^p \cdot \mu(k(p))}. \end{aligned} \tag{3.33}$$

Notice that by Lemma 7 we have

$$\int \frac{\Pr(\alpha - 1 < |Y| \leqslant \lceil q^p \rceil, K_{k(p-1)}(X) = 1)}{\mu(k(p))} \mu(\mathrm{d}x) \leqslant M_2 \Pr(|Y| > \alpha - 1). \tag{3.34}$$

Integrating both sides of (3.33) with respect to x and using (3.34), we obtain

$$\begin{aligned} \int \left(\sum_{p=1}^{\infty} \frac{J_1(p)}{q^p \cdot \mu(k(p))} \right) \mu(\mathrm{d}x) &\leqslant M_2 \sum_{\nu=1}^{\infty} \sum_{\alpha=\lceil q^{\nu-1}\rceil+1}^{\lceil q^{\nu}\rceil} \sum_{p=\nu}^{\infty} \frac{\lceil q^{\nu}\rceil}{q^p} \Pr(|Y| > \alpha - 1) \\ &\leqslant 2M_2 \frac{q}{q-1} \sum_{\nu=1}^{\infty} \Pr(|Y| > \nu - 1) \\ &\leqslant 2M_2 \frac{q}{q-1} \mathrm{E}|Y| < \infty. \end{aligned} \tag{3.35}$$

Hence, the integrand must be finite μ a.e. In a similar way (3.34) implies that the second series at (3.32) is finite μ a.e., completing the proof that $\sum_{p=1}^{\infty} \Pr(\mathscr{D}_p) < \infty$. □

Lemma 2. *We have*

$$\lim_{n\to\infty} \Pr(\mathscr{B}_n) = 1.$$

Proof. From the proof of Lemma 1 it is easy to see that for $n \in (k(p-1), k(p)]$ the probability $\Pr(\mathscr{B}_n^{\mathrm{c}})$ is bounded by (3.28) and (3.29). Since (3.29) sums up over p to a finite number, $\Pr(\mathscr{B}_n^{\mathrm{c}})$ must converge to 0. This completes the proof of convergence of $I_n^{(1)}(x)$. □

Convergence of $I_n^{(2)}(x)$. By (3.6) it is enough to prove that

$$\lim_{p\to\infty} \overline{r}_{k(p-1)}(x) = r(x) \qquad \mu \text{ a.e.} \tag{3.36}$$

We have

$$\begin{aligned}\overline{r}_{k(p-1)}(x)-r(x)&=\frac{\int \overline{Y}_p\cdot K_{1,k(p-1)}\,\mathrm{d}P}{\int K_{1,k(p-1)}\,\mathrm{d}P}-\frac{\int Y\cdot K_{1,k(p-1)}\,\mathrm{d}P}{\int K_{1,k(p-1)}\,\mathrm{d}P}+\frac{\int Y\cdot K_{1,k(p-1)}\,\mathrm{d}P}{\int K_{1,k(p-1)}\,\mathrm{d}P}-r(x)\\&=G_p^{(1)}+G_p^{(2)}.\end{aligned}\tag{3.37}$$

By Theorem 10.49 in Wheeden and Zygmund (1977) $G_p^{(2)}$ converges to 0. Let

$$H_{l,m}=\int|Y|\cdot I(|Y|>q^l)\cdot K_{1,k(m)}\,\mathrm{d}P\bigg/\int K_{1,k(m)}\,\mathrm{d}P.\tag{3.38}$$

It is clear that $|G_p^{(1)}|\leqslant H_{p,p-1}$. Letting $r_l(x)=\mathrm{E}(|Y|\cdot I(|Y|>q^l)|X=x)$, we obtain

$$H_{l,m}=\int r_l(x_1)\cdot K_{1,k(m)}\mu(\mathrm{d}x_1)\bigg/\int K_{1,k(m)}\mu(\mathrm{d}x_1).\tag{3.39}$$

Again by Theorem 10.49 in Wheeden and Zygmund (1977), we have for every l

$$\lim_{m\to\infty}H_{l,m}(x)=r_l(x)\qquad\mu\text{ a.e.}\tag{3.40}$$

Choose x such that (3.40) holds true for all $l=1,2,\ldots$, $r_l(x)\geqslant r_s(x)$ for $s>l$, and that $r_l(x)$ converges to zero. The set of such xs has μ-measure 1. Choose L such that $r_l(x)<\epsilon$ for all $l\geqslant L$. Next, take $M\geqslant L$ such that $|H_{L,m}-r_L(x)|<\epsilon$ for $m\geqslant M$. Since $H_{l,m}$ is decreasing in l we conclude that for $p>M$ we have

$$|G_p^{(1)}|<H(p,p-1)<H(L,p-1)\leqslant|H(L,p-1)-r_L(x)|+r_L(x)\leqslant 2\epsilon.$$

This proves that $G_p^{(1)}$ is converging to 0 and completes the proof of Theorem 1. □

Remark 4. An alternative but more technical way of proving Lemmas 1 and 2 consists in proving a version of Lemma 1 in the case of an arbitrary ranking dependent on random variables $X_1,\ldots,X_N$ only, and then drawing conclusions for the particular cases of (n)- and (p)-rankings defined at the beginning of this section.

Proof of Theorem 2. By (3.6) it is enough to prove convergence of $\tilde{r}_n(x)$ defined in (3.7). For $i=1,\ldots,n$ we have

$$\mathrm{E}(Y_i|K_{i,n})=\mathrm{E}(r(X_i)|K_{i,n})=\mathrm{E}(Y_i|K_{1,n},\ldots,K_{n,n})=\begin{cases}\dfrac{\int Y\cdot K_{1,n}\,\mathrm{d}P}{\int K_{1,n}\,\mathrm{d}P} & \text{if } K_{i,n}=1\\[2ex]\dfrac{\int Y\cdot(1-K_{1,n})\,\mathrm{d}P}{\int(1-K_{1,n})\,\mathrm{d}P} & \text{if } K_{i,n}=0.\end{cases}\tag{3.41}$$

Putting

$$r_{k(p-1)}(x) = \int Y \cdot K_{1,k(p-1)}\, \mathrm{d}P \bigg/ \int K_{1,k(p-1)}\, \mathrm{d}P, \tag{3.42}$$

we obtain

$$\begin{aligned}\tilde{r}_n(x) - r(x) &= \frac{\sum_{i=1}^{n}(Y_i - r_{k(p-1)}(x))K_n(X_i)}{n \cdot \mu(n)} + \frac{\sum_{i=1}^{n} K_n(X_i)}{n \cdot \mu(n)} \cdot r_{k(p-1)}(x) - r(x) \\ &= I_n^{(1)}(x) + I_n^{(2)}(x). \end{aligned} \tag{3.43}$$

Convergence of $I_n^{(2)}(x)$ follows easily from the Lebesgue theorem (cf. Theorem 10.49 in Wheeden and Zygmund 1977) and from (3.6). It remains to prove a.s. convergence to zero of $I_n^{(1)}(x)$.

Let $n \in (k(p-1), k(p)]$. For $j = 1, \ldots, k(p)$ let

$$Z_{j,p} = Y_j - \mathrm{E}(Y_j | K_{j,k(p-1)}); \tag{3.44}$$

for $j = 1, \ldots, w(n, n)$ let

$$Z_{(j)_n} = Y_{(j)_n} - r_{k(p-1)}; \tag{3.45}$$

and for $j = 1, \ldots, w(k(p), k(p-1))$ let

$$Z_{(j)_p} = Y_{(j)_p} - r_{k(p-1)}. \tag{3.46}$$

By (3.5) we obtain for $\nu = 1, \ldots, w(n, n)$

$$\sum_{j=1}^{\nu} Z_{(j)_n} = \sum_{j=1}^{N_\nu} Z_{(j)_p} - \sum_{j=1}^{N_\nu} Z_{(j)_p} \cdot Q_{(j)_p,n}. \tag{3.47}$$

Let events $\mathscr{A}_n$, $\mathscr{B}_n$, and $\mathscr{D}_p$ be given by (3.23)–(3.25) but with $\bar{Z}$s replaced by Zs given by (3.44)–(3.46). By (3.47) we have $\mathscr{A}_n \cap \mathscr{B}_n \subset \mathscr{D}_p$ and

$$\begin{aligned}\{|I_n^{(1)}(x)| > 4\epsilon\} &= \left\{ \left| \sum_{i=1}^{n} Z_{i,p} K_n(X_i) \right| > 4\epsilon \cdot n \cdot \mu(n) \right\} \\ &\subset \left\{ \max_{\nu=1,\ldots,w(n,n)} \left| \sum_{j=1}^{\nu} Z_{(j)_n} \right| > 4\epsilon \cdot n \cdot \mu(n) \right\} \subset \mathscr{A}_n. \end{aligned} \tag{3.48}$$

We will prove below (much more briefly than in Lemma 1 but using moment assumption (2.10)) that $\sum_{p=1}^{\infty} \Pr(\mathscr{D}_p) < \infty$ and that $\lim_{n\to\infty} \Pr(\mathscr{B}_n) = 1$. Hence as events $\mathscr{A}_n$ and $\mathscr{B}_n$ are independent, we obtain inequalities (3.27), thereby concluding the proof.

So, it remains to prove that $\sum_{p=1}^{\infty} \Pr(\mathscr{D}_p) < \infty$ and that $\lim_{n\to\infty} \Pr(\mathscr{B}_n) = 1$. Let $\psi(t)$ be given by (2.9). By Lemma 8 and applying inequalities (2.6) and (2.8) we obtain

$$\begin{aligned}
\Pr(\mathscr{D}_p) &= \int \Pr\left(\max_{\nu=1,\ldots,w(k(p),k(p-1))} \left| \sum_{j=1}^{\nu} Z_{(j)_p} \right| > \epsilon \cdot k(p) \cdot \mu(k(p)) | \mathbf{K}_{k(p-1)} \right) \mathrm{d}\beta_p \\
&= \int \Pr\left(\max_{\nu=1,\ldots,k(p)} \left| \sum_{j=1}^{\nu} Z_{(j)_p} \cdot K_{(j)_p,k(p-1)} \right| > \epsilon \cdot k(p) \cdot \mu(k(p)) | \mathbf{K}_{k(p-1)} \right) \mathrm{d}\beta_p \\
&\leqslant \int \frac{\mathrm{E}\left(\Phi\left(\sum_{j=1}^{k(p)} Z_{(j)_p} K_{(j)_p,k(p-1)} \right) | \mathbf{K}_{k(p-1)} \right)}{\Phi(\epsilon \cdot k(p) \cdot \mu(k(p)))} \mathrm{d}\beta_p \qquad (3.49) \\
&\leqslant C_0 \mathrm{E}\Phi\left(\sum_{j=1}^{k(p)} (Z_j K_{k(p-1)}(X_j)) \right) \Big/ \Phi(k(p) \cdot \mu(k(p))) \\
&\leqslant C_0 \sum_{j=1}^{k(p)} \mathrm{E}\Phi(Z_j K_{k(p-1)}(X_j)) / \Phi(k(p) \cdot \mu(k(p))) \\
&= C_0 \cdot k(p) \cdot \mathrm{E}\Phi(Z_1 K_{k(p-1)}(X_1)) / \Phi(k(p) \cdot \mu(k(p))) \\
&= C_0 (1/\psi(k(p) \cdot \mu(k(p)))) (\mathrm{E}\Phi(|Z_1| K_{k(p-1)}(X_j)) / \mu(k(p))). \qquad (3.50)
\end{aligned}$$

Hence, by Lemma 7, by (2.11) and by the Cauchy condensation theorem (cf. Knopp 1948, p. 120) we obtain

$$\begin{aligned}
\int \sum_{p=1}^{\infty} \Pr(\mathscr{D}_p) \mu(\mathrm{d}x) &\leqslant C_0 \cdot M_2 \cdot \mathrm{E}\Phi(|Z_1|) \sum_{p=1}^{\infty} 1 \Big/ \psi(k(p) \cdot \mu(k(p))) \\
&\leqslant C_0 \cdot M_2 \cdot \mathrm{E}\Phi(Z_1) \sum_{p=1}^{\infty} 1 \Big/ \psi(k(p) \cdot h_{k(p)}^d) < \infty.
\end{aligned}$$

Thus, with probability 1 events $\mathscr{D}_p$ may occur only finitely many times. As in the proof of Lemma 2 let us note that $\Pr(\mathscr{B}_n^{\mathrm{c}})$ is bounded by (3.49). This completes the proof of Theorem 2. □

Proof of Theorem 3. By (3.6), and since $Y_{i,n} = \bar{Y}_{i,p}$ for $n \in (k(p-1), k(p)]$ (cf. (2.14) and (3.9)), the convergence of $\hat{r}_n(x)$ given by (2.15) is implied by the convergence of $\hat{r}_n^{(p)}(x)$ given by (3.10). However, in contrast with Theorem 1, we do not need to prove that the limits of (2.15) and of (1.2) are equal μ a.e. Hence, we can use just the left-hand side of inequality (2.2) (cf. (3.29)) with $\kappa = 1 - d\delta$. By Lemma 4 it is the lowest possible κ for any probability measure μ. The right-hand side of inequality (2.2) was used in (3.14), which is not required in the present case of the truncated estimator (2.15). In this way, and since the convergence of $\hat{r}_n^{(p)}(x)$ was proved while proving Theorem 1, we infer Theorem 3. □

We now prove the properties of the Cantor and Cantor–Preiss measures referred to in Example 1.

Lemma 3. *Let $X = 2 \cdot \sum_{i=0}^{\infty} U_i/3^i$, where U_i are independent $\{0, 1\}$ random variables, and let μ_{C} and μ_{CP} be the Cantor and Cantor–Preiss measures, respectively, defined in Example 1. Let $h_n = Cn^{-\delta}$ for $n = 1, 2, \ldots, \delta \in (0, 1)$, and let $\gamma_{\mathrm{C}}(n)$ and $\gamma_{\mathrm{CP}}(n)$ be functions γ defined by (2.2) and corresponding to the Cantor and Cantor–Preiss measures, respectively. Then*

(*a*) *for $\kappa_{\mathrm{C}} = 1 - (\ln 2/\ln 3)\delta$ we have*

$$\frac{1}{2} \leqslant \frac{\gamma_{\mathrm{C}}(h_n)}{n^{\kappa_{\mathrm{C}}}} \leqslant 2 \qquad \mu_{\mathrm{C}} \text{ a.e.,} \tag{3.51}$$

(*b*) *for the probability distribution μ_{CP},*

$$\limsup_{n\to\infty} \frac{\gamma_{\mathrm{CP}}(h_n)}{n^{\kappa_{\mathrm{CP}}}} = \begin{cases} \infty & \text{if } \kappa_{\mathrm{CP}} \in [1 - d\delta, 1) \\ 0 & \text{if } \kappa_{\mathrm{CP}} = 1 \end{cases} \tag{3.52}$$

μ_{CP} a.e.

Proof. Let $\mathscr{C}$ be the Cantor set and $x_0 \in \mathscr{C}$, i.e., $x_0 = 2\sum_{i=1}^{\infty} u_{0i}/3^i$ and $u_{0i} \in \{0, 1\}$. For simplicity, let $h_n = n^{-\delta}$ and let k_n be positive integers such that

$$3^{-k_n} \leqslant n^{-\delta} < 3^{-(k_n - 1)}.$$

Let

$$A_k = A_k(x_0) = \left\{ x \in \mathscr{C}\colon x = 2\sum_{i=0}^{\infty} u_i \Big/ 3^i,\ u_i \in \{0, 1\}, \text{ and } u_i = u_{0i}, \text{ for } i = 1, 2, \ldots, k \right\}.$$

Hence

$$\mu(A_{k_n}) \leqslant \mu(h_n) \leqslant \mu(A_{k_n - 1}) \tag{3.53}$$

and

$$\mu(A_k) = \prod_{i=1}^{k} p_i^{u_{0i}} (1 - p_i)^{1 - u_{0i}},$$

where $p_i = \frac{1}{2}$ for $\mu = \mu_{\mathrm{C}}$ or $p_i = 1/(i+1)$ for $\mu = \mu_{\mathrm{CP}}$. In the case of the Cantor measure we have

$$\mu_{\mathrm{C}}(A_k) = 2^{-k} = \mathrm{e}^{-k \ln 2},$$

and hence by (3.53) we obtain

$$\tfrac{1}{2} n^{-\delta(\ln 2/\ln 3)} \leqslant \mu_{\mathrm{C}}(h_n) \leqslant 2n^{-\delta(\ln 2/\ln 3)}.$$

Setting $\kappa = 1 - \delta(\ln 2/\ln 3)$, we have

$$\frac{1}{2} \leqslant \frac{n\mu_{\mathrm{C}}(h_n)}{n^{\kappa}} \leqslant 2,$$

i.e., μ_{C} meets Assumption κ with $\kappa = 1 - \delta(\ln 2/\ln 3)$.

In the case of the Cantor–Preiss measure we consider a random index X_0 of the sets A_k, where $X_0 = 2\sum_{i=1}^{\infty} U_{0i}/3^i$ is distributed according to the Cantor–Preiss distribution with U_{0i} independent $\{0, 1\}$ rvs and $p_i = \Pr(U_{0i} = 1) = 1/(i+1)$. We have

$$\mu_{\mathrm{CP}}(A_{k_n}) = \prod_{i=1}^{k_n} \frac{i^{(1-U_{0i})}}{i+1} = \frac{1}{k_n+1} \exp\left(-\sum_{i=1}^{k_n} Y_i\right),$$

where $Y_i = U_{0i} \ln i$. Let

$$a_n = (\ln n)^{(3+\delta)/2}$$

and note that by the strong law of large numbers (cf. Petrov 1975, ch. 9, Theorem 14, p. 272),

$$\frac{1}{a_n}\sum_{i=1}^{n}(Y_i - \mathrm{E}Y_i) \to 0 \qquad \text{a.s.}$$

Thus

$$\lim_{n\to\infty} (\ln k_n)^{-(3+\delta)/2}\left(\sum_{i=1}^{k_n} Y_i - \sum_{i=1}^{k_n}\frac{\ln i}{i+1}\right) \to 0 \qquad \mu_{\mathrm{CP}} \text{ a.e.}$$

Now, by the Euler summation formula (cf. Knopp 1948, p. 523) we obtain

$$\begin{aligned}\sum_{i=1}^{k_n} Y_i &= \sum_{i=1}^{k_n}\frac{\ln i}{i+1} + o((\ln k_n)^{(3+\delta)/2}) \\ &= \frac{1}{2}(\ln k_n)^2 + o((\ln k_n)^{(3+\delta)/2}).\end{aligned}$$

We can now evaluate $\mu_{\mathrm{CP}}(A_{k_n})$:

$$\ln(\mu_{\mathrm{CP}}(A_{k_n})) = -\tfrac{1}{2}(\ln k_n)^2 + o((\ln k_n)^{(3+\delta)/2}),$$

and similarly

$$\ln(\mu_{\mathrm{CP}}(A_{k_n-1})) = -\tfrac{1}{2}(\ln k_n)^2 + o((\ln k_n)^{(3+\delta)/2}).$$

Hence

$$\ln(\mu_{\mathrm{CP}}(h_n)) = -\tfrac{1}{2}(\ln k_n)^2 + o((\ln k_n)^{(3+\delta)/2}),$$

which implies easily that for μ_{CP} a.e. x_0

$$\limsup_{n\to\infty} \frac{n \cdot \mu_{\mathrm{CP}}(h_n)}{n^{\kappa}} = \begin{cases} +\infty & \text{if } \kappa < 1, \\ 0 & \text{if } \kappa = 1. \end{cases}$$

Hence Assumption κ is not met for any $\kappa \in [1 - d\delta, 1]$. □

Remark 5. Let us note that it is straightforward to extend Example 1 and Lemma 3 and consider a class of probability distributions on the Cantor set $\mathscr{C}$. To this end it is enough to allow p_i to be arbitrary functions of i with values in [0, 1]. One can show that if

$$\lim_{i\to\infty} p_i = p \text{ and } p \in (0, 1)$$

then the corresponding probability distribution on $\mathscr{C}$ meets Assumption κ. If one allows p_i to fluctuate between two different values, say p^- and p^+, both in (0, 1), then one can construct a probability distribution for which there exist distinct κ_1 and κ_2 with

$$0 < c_1 \leqslant \liminf_{n\to\infty} \frac{\gamma(n)}{n^{\kappa_1}} \leqslant \liminf_{n\to\infty} \frac{\gamma(n)}{n^{\kappa_2}} \leqslant c_2 < \infty.$$

4. Appendix

Lemma 4 (Wheeden and Zygmund 1977, Lemma 10.50). *For any complete probability measure μ on the Borel subsets of $\mathbb{R}^d$ there exist a finite, non-negative, μ-measurable function $\phi(x)$, and a set $\mathscr{A}$ such that $\mu(\mathscr{A}) = 1$ and*

$$\frac{h^d}{\mu_x(h)} \leqslant \phi(x) \text{ for every } h > 0 \text{ and } x \in \mathscr{A},$$

where $\mu_x(h)$ stands for a probability measure μ of a ball in $\mathbb{R}^d$ centred at x and of radius h.

Corollary 2. *If $h_n = Cn^{-\delta}$, $0 < \delta < d^{-1}$, and $\gamma(n)$ is given by (2.1), then there exists a function $c(x) > 0$ such that*

$$c(x) \cdot n^{\kappa_0} \leqslant \gamma(n) \leqslant n,$$

where $\kappa_0 = 1 - d\delta$.

The following lemma extends Lemma 1 in Devroye and Wagner (1980).

Lemma 5. *Let*

$$1 \leqslant \frac{R}{r} \leqslant \rho < \infty$$

and let M_ρ be the minimal number of balls of radius $1/(2\rho)$ necessary to cover the unit ball $\mathscr{B} = \mathscr{B}_{(0,1)}$ centred at $0 \in \mathbb{R}^d$ and of radius 1. Then for every $z \in \mathbb{R}^d$ and every measurable set A

$$\int_A \frac{K\left(\frac{z-x}{R}\right)}{\int K\left(\frac{y-x}{r}\right)\mu(\mathrm{d}y)}\mu(\mathrm{d}x) \leqslant M_\rho.$$

Proof. Let $\{A_1, \ldots, A_{M_\rho}\}$ be the cover of $\mathscr{B}$ by balls of radius $1/(2\rho)$. Since $z - R.\mathscr{B} \subset z - R \cup A_i$ we have

$$K\left(\frac{z-x}{R}\right) \leqslant \sum_{i=1}^{M_\rho} I_{z-RA_i}(x).$$

Let $x, w \in z - RA_i$. Then $\|x - w\| \leqslant 2R/(2\rho) \leqslant r$, where $\|x\|$ stands for the Euclidean norm of vector $x \in \mathbb{R}^d$. Thus the following implication holds true:

$$\text{if } x \in z - RA_i \text{ then } w \in z - RA_i \text{ implies } w \in x + r.\mathscr{B},$$

and, moreover, for $x \in z - RA_i$ we have

$$\int K\left(\frac{y-x}{r}\right)\mu(\mathrm{d}y) = \mu(x + r.\mathscr{B}) \geqslant \mu(z - RA_i).$$

Hence

$$\int_A \frac{K\left(\frac{z-x}{R}\right)}{\int K\left(\frac{y-x}{r}\right)\mu(\mathrm{d}y)}\mu(\mathrm{d}x) \leqslant \int \frac{K\left(\frac{z-x}{R}\right)}{\int K\left(\frac{y-x}{r}\right)\mu(\mathrm{d}y)}\mu(\mathrm{d}x) \leqslant \sum_{i=1}^{M_\rho} \frac{\mu(z-RA_i)}{\mu(z-RA_i)} = M_\rho,$$

and the proof is complete. □

Lemma 5 easily implies the following lemma.

Lemma 6. *Let $V(z)$ be a non-negative function and let*

$$\int V(z)\mu(\mathrm{d}z) = V < \infty.$$

If the ratio $H/h \in [1, \rho)$ then

$$\int \frac{\int V(z)K\left(\frac{z-x}{H}\right)\mu(\mathrm{d}z)}{\int K\left(\frac{y-x}{h}\right)\mu(\mathrm{d}y)}\mu(\mathrm{d}x) \leqslant M_\rho \cdot V$$

and the bound does not depend on the particular choice of H or h.

Applying Lemma 6 to the conditional distribution of Z given $X = x$, we obtain:

Lemma 7. *Let* (Z, X) *be a random vector in* $\mathbb{R} \times \mathbb{R}^d$ *and let* μ *be the probability distribution of* X. *If the ratio* $H/h \in [1, \rho)$ *then for any Borel set* $\mathscr{A}$ *we have*

$$\int \frac{\Pr\left(Z \in \mathscr{A}, K\left(\dfrac{x-X}{H}\right) = 1\right)}{\mathrm{E}K\left(\dfrac{x-X}{h}\right)} \mu(\mathrm{d}x) \leqslant M_\rho \cdot \Pr(Z \in \mathscr{A})$$

and the bound does not depend on the particular choice of H, h *or* μ.

Lemma 8. *Let* $Y_1, \ldots, Y_n$ *be independent rvs centred at expectations,* $i_1, \ldots, i_n$ *a fixed permutation of* $1, \ldots, n$, *and* Φ *a symmetric, convex function. Then*

$$\Pr\left(\max_{\nu=1,\ldots,n} \left|\sum_{j=1}^{\nu} Y_{i_j}\right| > \epsilon\right) \leqslant \frac{\mathrm{E}\Phi\left(\sum_{j=1}^{n} Y_j\right)}{\Phi(\epsilon)}.$$

Proof. The extension of the Kolmogorov maximal inequality to the case of convex functions follows easily by virtually the same argument as in Problems 2 and 3 in Loève (1977, p. 275). The only change required is a replacement of the power function with the general convex function, but by the Jensen inequality this does not present any difficulty. The classical Kolmogorov inequality is usually stated in the case of the identity permutation $1, \ldots, n$. However, the proof goes through without change in the case of any deterministic permutation $i_1, \ldots, i_n$ of $1, \ldots, n$. □

Lemma 9. *Let* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *be independent, identically distributed rvs with* $\mathrm{E}|Y| < \infty$ *and* Φ *a symmetric, convex function. Then for every permutation* $\{Y_{i_j}\}$ *of* $Y_1, \ldots, Y_n$ *depending only on* $\mathbf{X} = (X_1, \ldots, X_n)$, *the conditional maximal inequality*

$$\Pr\left(\max_{\nu=1,\ldots,n} \left|\sum_{j=1}^{\nu} Y_{i_j} - \mathrm{E}(Y|\mathbf{X})\right| > \epsilon|\mathbf{X}\right) \leqslant \frac{\mathrm{E}\left(\Phi\left(\sum_{j=1}^{n} Y_j\right)\middle|\mathbf{X}\right)}{\Phi(\epsilon)}$$

holds true.

Proof. Let us note that the a.e. existence of the regular conditional probability is implied by Theorem 8.1 of Parthasarathy (1967, p. 147). The conditional independence of the Y_i is implied easily by independence of $(X_1, Y_1), \ldots, (X_n, Y_n)$ (cf. Lemma 1 in Cheng 1984). Hence, and since the relevant permutations are functions of the conditional variables, Lemma 8 can be applied to the conditional probability distributions. So, the maximal inequality holds conditionally. □

Remark 6. It is clear that the permutation in Lemma 8 cannot depend on rvs $Y_1, \ldots, Y_n$. In particular, it can easily be shown that Kolmogorov's maximal inequality does not hold true for partial sums of $Y_{(1)}, \ldots, Y_{(n)}$, where $Y_{(1)} \leqslant Y_{(2)} \leqslant \ldots, Y_{(n)}$ are the increasing order statistics. However, the case of permutations corresponding to the random variables ordered according to their increasing values is not excluded from the class of admissible permutations considered in Mukerjee (1989, Section 2, p. 19) and in papers on monotonic regression quoted by Mukerjee (1989). Hence Theorem 2.1 in Mukerjee (1989), the proof of which is based on the Kolmogorov maximal inequality, does not seem to be valid for all his admissible permutations.

Acknowledgments

We wish to thank both Professor David Preiss for personal correspondence and comments on pathological properties of measures which inspired the present Example 1 and Professor Hari Mukerjee for discussions and comments on his paper. We also wish to thank the referees for their careful reading of the paper which led to improvements in the clarity of presentation.

References

Burkholder, D.L. and Gundy, R.F. (1970) Extrapolation and interpolation of quasi-linear operators on martingales. *Acta Math.*, **124**, 249–304.

Cheng, P. (1983) On the strong consistency and convergence rate of improved kernel estimates for the regression function. *J. Systems Sci. Math. Sci.*, **3**, 304–315.

Cheng, P. and Zhao, L. (1985) Strong consistency of the improved nearest neighbor estimates of regression functions. *Kexue Tongbao*, **30**, 717–722.

Cheng, P.E. (1984) Strong consistency of nearest neighbor regression function estimators. *J. Multivariate Anal.*, **15**, 63–72.

Devroye, L.P. (1981) On the almost everywhere convergence of nonparametric regression function estimates. *Ann. Statist.*, **9**, 1310–1319.

Devroye, L.P. and Wagner, T.J. (1980) Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Ann. Statist.*, **8**, 231–239.

Etemadi, N. (1981) An elementary proof of the strong law of large numbers. *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, **55**, 119–122.

Greblicki, W., Krzyżak, A. and Pawlak, M. (1984) Distribution-free pointwise consistency of kernel regression estimates. *Ann. Statist.*, **12**, 1570–1575.

Hall, P. and Jones, M.C. (1990) Adaptive *M*-estimators in nonparametric regression. *Ann. Statist.*, **18**, 1712–1728.

Härdle, W., Janssen, P. and Serfling, R. (1988) Strong uniform consistency rates for estimators of conditional functionals. *Ann. Statist.*, **16**, 1428–1449.

Howell, J.O., Taylor, R.L. and Woyczynski, W.A. (1981) Stability of linear forms in independent random variables in Banach spaces. *Probability in Banach Spaces III*, Lecture Notes in Math. **860**, pp. 231–245. Berlin: Springer-Verlag.

Knopp, K. (1948) *Theory and Application of Infinite Series*. London: Blackie & Son.

Loève, M. (1977) *Probability Theory I.* New York: Springer-Verlag.

Mukerjee, H. (1989) A strong law of large numbers for nonparametric regression. *J. Multivariate Anal.*, **30**, 17–26.

Nadaraya, E.A. (1964) On estimating regression. *Theory Probab. Appl.*, **9**, 157–159.

Parthasarathy, K.R. (1967) *Probability Measures on Metric Spaces*. New York: Academic Press.

Petrov, V.V. (1975) *Sums of Independent Random Variables*. New York: Springer-Verlag.

Preiss, D. (1987) Geometry of measures in R^d: distribution, rectifiability, and densities. *Ann. Math.*, **125**, 537–643.

Schuster, E.F. (1968) Estimation of a probability density function with applications in statistical inference. Ph.D. thesis, University of Arizona, Tucson.

Stone, C.J. (1977) Consistent nonparametric regression. *Ann. Statist.*, **5**, 595–645.

Stute, W. (1986) On almost sure convergence of conditional empirical distribution functions. *Ann. Probab.*, **14**, 891–901.

Taylor, R.L. and Calhoun, C.A. (1983) On the almost sure convergence of randomly weighted sums of random elements. *Ann. Probab.*, **11**, 795–797.

Watson, G.S. (1964) Smooth regression analysis. *Sankhya Ser. A*, **26**, 359–372.

Wheeden, R.L. and Zygmund, A. (1977) *Measure and Integral*. New York: Marcel Dekker.

Zhao, L.C. and Fang, Z.B. (1985) Strong convergence of kernel estimates of nonparametric regression functions. *Chinese Ann. Math. Ser. B*, **6**, 147–155.

Received October 1995 and revised December 1996.