The Annals of Probability
1986, Vol. 14, No. 3, 891–901

ON ALMOST SURE CONVERGENCE OF CONDITIONAL EMPIRICAL DISTRIBUTION FUNCTIONS

BY WINFRIED STUTE

University of Giessen

We investigate the almost sure convergence of a kernel-type conditional empirical distribution function both in sup-norm and weighted sup-norms. As an application we get a strong law for the Nadaraya–Watson estimate of a regression function $m(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x})$ under a weak moment condition on Y.

0. Introduction and main results. In this paper we derive almost sure convergence results for conditional empirical distribution functions. These functions may be viewed as nonparametric estimates of a conditional distribution function. To be precise, consider a random vector $\xi = (\mathbf{X}, \mathbf{Y})$ in $\mathbb{R}^d$ defined on a probability space $(\Omega, \mathscr{A}, \mathbb{P})$, where $\mathbf{X} = (X^1, \dots, X^{d_1}) \in \mathbb{R}^{d_1}$, $\mathbf{Y} = (Y^1, \dots, Y^{d_2}) \in \mathbb{R}^{d_2}$, and $d = d_1 + d_2$. Write $m(\mathbf{y}|\mathbf{x})$ for the conditional probability that $\mathbf{Y} \le \mathbf{y}$ (componentwise) given that $\mathbf{X} = \mathbf{x}$. The function $m(\cdot|\cdot)$ contains the full information on the dependence structure of $\mathbf{X}$ and $\mathbf{Y}$.

Now, let $\xi_1, \xi_2, \dots$ denote a sequence of independent random vectors with the same distribution as ξ, say H. Let K be the naive kernel on $\mathbb{R}^{d_1}$, i.e.,

$$K(x_1, \dots, x_{d_1}) = 1 \quad \text{if } -\tfrac{1}{2} \le x_i < \tfrac{1}{2} \text{ for } i = 1, \dots, d_1$$

and zero elsewhere, and put, for given $\mathbf{x} \in \mathbb{R}^{d_1}$,

$$K_n(\mathbf{x}, \mathbf{X}_i) \equiv K_n(\mathbf{X}_i) := K\left(\frac{\mathbf{x} - \mathbf{X}_i}{a_n}\right).$$

Here $(a_n)_n$ is a sequence of bandwidths satisfying

$$(1) \qquad a_n \to 0 \quad \text{and} \quad n a_n^{d_1} \to \infty.$$

Now, put

$$(2) \qquad m_n(\mathbf{y}|\mathbf{x}) = \frac{\sum_{i=1}^n 1_{\{\mathbf{Y}_i \le \mathbf{y}\}} K_n(\mathbf{X}_i)}{\sum_{i=1}^n K_n(\mathbf{X}_i)}.$$

In other words, $m_n(\mathbf{y}|\mathbf{x})$ is a normalized number of $\mathbf{Y}_i$ such that $\mathbf{Y}_i \le \mathbf{y}$ subject to the constraint $K_n(\mathbf{X}_i) = 1$. When $a_n > 0$ is small $m_n(\cdot|\mathbf{x})$ is thus reflecting the spatial distribution of those data points $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ for which $\mathbf{X}_i$ is close to $\mathbf{x}$. This construction is also basic to many nonparametric estimates of the regression function

$$m(\mathbf{x}) \equiv \mathbb{E}(Y^1|\mathbf{X} = \mathbf{x}).$$

Received October 1984; revised April 1985.

AMS 1980 *subject classifications.* Primary 60F15, 62G05; secondary 62J02.

Key words and phrases. Conditional empirical distribution function, Glivenko–Cantelli convergence, weight functions, Nadaraya–Watson estimator.

See Collomb (1981) for a survey. In fact, replacing $1_{\{\mathbf{Y}_i \le \mathbf{y}\}}$ by Y_i^1 in (2), we arrive at the so-called Nadaraya–Watson estimate of $m(\mathbf{x})$. Cf. Nadaraya (1964) and Watson (1964). For more details see Section 1 below.

Returning to $m(\mathbf{y}|\mathbf{x})$, along with m_n, we have to consider

$$m_n^*(\mathbf{y}|\mathbf{x}) = \sum_{i=1}^{n} \frac{1_{\{\mathbf{Y}_i \le \mathbf{y}\}} K_n(\mathbf{X}_i)}{n\mathbb{E}(K_n(\mathbf{X}_1))} \tag{3}$$

whenever defined. Clearly

$$m_n(\mathbf{y}|\mathbf{x}) = \frac{m_n^*(\mathbf{y}|\mathbf{x})}{f_n(\mathbf{x})}, \tag{4}$$

where

$$\begin{aligned} f_n(\mathbf{x}) &= \sum_{i=1}^{n} \frac{K_n(\mathbf{X}_i)}{n\mathbb{E}(K_n(\mathbf{X}_1))} \\ &= m_n^*(\infty|\mathbf{x}). \end{aligned}$$

Hence, as to almost sure convergence of m_n, we may confine ourselves to m_n^*. The main results of this paper will be proved under no conditions whatsoever on the distribution of ξ.

Typically, the statements will be true for almost all $\mathbf{x}$, i.e., for all $\mathbf{x} \notin N$ where N is such that $\mathbb{P}(\mathbf{X} \in N) = 0$. Our first result establishes Glivenko–Cantelli convergence of m_n to m, under very weak assumptions on the bandwidths.

THEOREM 1. *Assume that $a_n \to 0$ in such a way that*

$$\sum_{n \ge 1} \exp\left[-\rho n a_n^{d_1}\right] < \infty \quad \text{for all } \rho > 0. \tag{5}$$

Then, as $n \to \infty$, for almost all $\mathbf{x}$

$$D_n(\mathbf{x}) \equiv D_n := \sup_{\mathbf{y} \in \mathbb{R}^{d_2}} |m_n(\mathbf{y}|\mathbf{x}) - m(\mathbf{y}|\mathbf{x})| \to 0 \tag{6}$$

with probability one.

Actually, we shall prove more than (6) in that we show that for each $\varepsilon > 0$, $\sum_{n \ge 1} \mathbb{P}(D_n > \varepsilon) < \infty$, i.e., $D_n \to 0$ completely.

Condition (5) is always satisfied whenever $\ln n = o(n a_n^{d_1})$. It holds true for all reasonable choices of a bandwidth.

The Glivenko–Cantelli convergence (6) may be used in a straightforward manner for proving consistency of estimates which can be written as a sup-norm continuous function of $m_n(\cdot|\mathbf{x})$. For example, (6) immediately implies almost sure convergence of the Nadaraya–Watson estimate whenever Y^1 is bounded.

More generally, it follows from Theorem 1, that for almost all $\mathbf{x}$, $m_n(d\mathbf{y}|\mathbf{x}) \to m(d\mathbf{y}|\mathbf{x})$ weakly, i.e.,

$$\int f(\mathbf{y}) m_n(d\mathbf{y}|\mathbf{x}) \to \int f(\mathbf{y}) m(d\mathbf{y}|\mathbf{x})$$

for all bounded continuous functions f on $\mathbb{R}^{d_2}$.

To handle also the unbounded case one has to introduce weight functions leading to stronger concepts of convergence (topologies). See Wellner (1977) and Mason (1982) for a discussion of this in the field of usual (unconditional) empirical distribution functions.

To be specific, let h be any positive function on $\mathbb{R}^{d_2}$ nondecreasing in each of its arguments. Put

$$D_n(h) \equiv D_n(h, \mathbf{x}) = \sup_{\mathbf{y} \in \mathbb{R}^{d_2}} \frac{|m_n(\mathbf{y}|\mathbf{x}) - m(\mathbf{y}|\mathbf{x})|}{h(\mathbf{y})}.$$

It turns out, that $D_n(h) \to 0$ under a weak moment condition on $h^{-1}(\mathbf{Y})$, but with more restrictions on the bandwidths.

THEOREM 2. *We have $D_n(h, \mathbf{x}) \to 0$ with probability one for almost all $\mathbf{x}$ whenever*

$$\int h^{-r}(\mathbf{Y})\, d\mathbb{P} < \infty \quad \textit{for some } r > 1 \tag{7}$$

provided that $c_n = \ln n / n a_n^{d_1}$ satisfies $c_n \downarrow 0$ and $\sum c_n^r < \infty$.

REMARK. The results of this paper are easily seen to hold also true when K is a kernel of bounded variation with bounded support. Use integration by parts to reduce the general case to that considered in (2).

1. A strong law for the Nadaraya–Watson estimate.

In this section we shall state a result on almost sure pointwise consistency of the Nadaraya–Watson estimate

$$m_n(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i K_n(\mathbf{X}_i)}{\sum_{i=1}^n K_n(\mathbf{X}_i)} = \int y m_n(dy|\mathbf{x})$$

under a weak moment assumption on $Y (\equiv Y^1$, with $d_2 = 1)$. Of course, for the true hypothetical regression function

$$m(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x})$$

to be well defined, it suffices to assume that Y has a finite first moment.

The almost sure convergence of $m_n(\mathbf{x})$ has been investigated before by many authors. A crucial assumption throughout the papers has been that $|Y| \leq c$ for some finite c. Nadaraya (1970) studied the case of a ξ admitting a Lebesgue density, the kernel K being of bounded variation. For such a K, upon integrating by parts, the convergence to zero of the stochastic error is then an easy consequence of the exponential bounds for univariate and multivariate empirical processes as given by Dvoretzky, Kiefer, and Wolfowitz (1956) and Kiefer (1961). The existence of densities was necessary only to treat the bias. Devroye and Wagner (1980) considered the case when only $\mathbf{X}$ had a Lebesgue density. The stochastic error was treated by applying a standard Bernstein exponential bound to the sum of the (bounded) i.i.d. random variables $Y_i K((\mathbf{x} - \mathbf{X}_i)/a_n)$. The assumption of absolute continuity could be dispensed with in Devroye (1981). This was possible because of consideration of a general result from Wheeden and Zygmund (1977) on differentiation of integrals w.r.t. arbitrary measures.

Greblicki, Krzyżak, and Pawlak (1984) obtained similar results, under a weaker condition on the bandsequences [our condition (5)] and on the kernel, again for bounded Y's. As we shall see, the last assumption may be substantially weakened by using the convergence of conditional empirical distribution functions in weighted sup-norm metrics. In fact, the SLLN for $m_n(x)$ will be valid under

$$\mathbb{E}(|Y|^{r+\rho}) < \infty \quad \text{for some } r > 1 \text{ and } \rho > 0, \tag{8}$$

which is only slightly stronger than requiring the existence of finite first moments. Note that no additional assumption on the distribution of ξ is needed.

Theorem 3. *Under* (8) *for almost all* $\mathbf{x}$

$$\lim_{n\to\infty} m_n(\mathbf{x}) = m(\mathbf{x}) \quad \textit{with probability one}$$

provided that $c_n \downarrow 0$ *and* $\sum c_n^r < \infty$.

2. Lemmas and proofs. A fundamental tool in our analysis will be a slight extension of Lemma 1.2 in Stute (1984). This is concerned with the local behavior of the multivariate empirical process at a fixed point $\mathbf{z}$ in $\mathbb{R}^d$. For further reference, denote with

$$H_n(\mathbf{z}) \equiv H_n(z_1,\dots,z_d) = n^{-1}\sum_{i=1}^{n} 1_{(-\infty,\mathbf{z}]}(\xi_i), \qquad \mathbf{z} \in \mathbb{R}^d,$$

the empirical distribution function (d.f.) pertaining to an independent sample with d.f. H. Write

$$\alpha_n(\mathbf{z}) = n^{1/2}[H_n(\mathbf{z}) - H(\mathbf{z})], \qquad \mathbf{z} \in \mathbb{R}^d,$$

for the corresponding empirical process. Throughout we shall adopt the representation

$$\alpha_n(\mathbf{z}) = \bar{\alpha}_n(F_1(z_1),\dots,F_d(z_d)), \tag{9}$$

where $\bar{\alpha}_n$ is the empirical process of an i.i.d. sample with underlying distribution function C, the copula function of H. This means that C is a d.f. on $[0,1]^d$ with uniform marginals factorizing H in terms of its marginals $F_1,\dots,F_d$, say:

$$H(z_1,\dots,z_d) = C(F_1(z_1),\dots,F_d(z_d)). \tag{10}$$

Hence in what follows we may and do assume (first) that H has uniform marginals. Write α_n for $\bar{\alpha}_n$ and consider the case $\mathbf{0} = \mathbf{z} \in [0,1]^d$. We then have

Lemma 4. *Given* $0 < \delta < \frac{1}{2}$, *there exist some finite* $C(\delta)$ *and* $c(\delta) > 0$ *such that for all* $\mathbf{a} \in [0,1]^d$ *with* $H(\mathbf{a}) < \delta/4$ *and* $s > 0$ *with* $s \geq c(\delta)\sqrt{H(\mathbf{a})/n}$ *and* $32 \leq (s\delta(1-2\delta))^2$

$$\mathbb{P}\Big(\sup_{\mathbf{0}\leq \mathbf{t}\leq \mathbf{a}} \alpha_n(\mathbf{t}) > s\sqrt{H(\mathbf{a})}\Big) \leq C(\delta)\mathbb{P}\Big(\alpha_n(\mathbf{a}) \geq s(1-2\delta)^d\sqrt{H(\mathbf{a})}\Big). \tag{11}$$

PROOF. Whenever $\mathbf{a} = (a_1, \ldots, a_d)$ is such that $a_i < \frac{1}{2}$ for $i = 1, \ldots, d$ and $2 \leq s\sqrt{nH(\mathbf{a})}$ the assertion is identical with that of Lemma 1.2 in Stute (1984). For $d = 1$ the first growth assumption on s is superfluous [cf. Lemma 1.1. in Stute (1984)]. When $d = 2$, observe that the crucial Lemma 6.1 in Stute (1984) is equally valid for all $a_1 \leq 1 - \rho$ and $0 \leq a_2 \leq 1$ with $0 < \rho < 1$, i.e., for all (a_1, a_2) with a_1 being bounded away from 1. Such a constraint is necessary in order to bound the probability of success p in the binomial conditioning argument appearing in the proof there. By interchanging the role of a_1 and a_2 we thus obtain that (11) holds for $\mathbf{a} = (a_1, a_2)$ such that $\min(a_1, a_2) \leq 1 - \rho$. The condition $2 \leq s\sqrt{nH(\mathbf{a})}$ entailed $p \leq \delta s\sqrt{nH(a_1, a_2)}/2$, which was crucial for bounding, e.g., the level y on page 375 of the above cited paper. The same bound is also valid, however, under the (asymptotically) weaker assumption $s \geq c(\delta)\sqrt{H(\mathbf{a})/n}$ by observing that $p \leq H(\mathbf{a})/(1 - \rho)$. The constants $c(\delta, \rho)$ and $C(\delta, \rho)$ typically increase as $\rho \downarrow$. Now, if $\rho > 0$ is so small that $H(1 - \rho, 1 - \rho) > \frac{1}{2}$, we automatically have $\min(a_1, a_2) \leq 1 - \rho$, since otherwise $H(\mathbf{a}) > \frac{1}{2}$. This proves the lemma for $d = 2$. For $d \geq 3$ use induction on d [cf. Lemma 6.2 in Stute (1984)]. □

In a sense, both growth conditions on s are of the same type since (at least for large n) $s \geq c(\delta) \cdot \sqrt{H(\mathbf{a})/n}$ is satisfied whenever $32 \leq (s(1 - 2\delta))^2$.

Also a bound corresponding to (11) is valid for the probability that $\inf_{\mathbf{0} \leq \mathbf{t} \leq \mathbf{a}} \alpha_n(\mathbf{t}) < -s\sqrt{H(\mathbf{a})}$, so that in summary

$$\mathbb{P}\Big(\sup_{\mathbf{0} \leq \mathbf{t} \leq \mathbf{a}} |\alpha_n(\mathbf{t})| > s\sqrt{H(\mathbf{a})}\Big) \leq C(\delta)\mathbb{P}\Big(|\alpha_n(\mathbf{a})| \geq s(1 - 2\delta)^d\sqrt{H(\mathbf{a})}\Big). \tag{12}$$

Now, bounding the sup over all $\mathbf{0} \leq \mathbf{t} \leq \mathbf{a}$ is only a matter of convenience. Likewise, a corresponding bound holds for the sup extended over any (small) rectangle. To be precise, denote with $I_{\mathbf{z}_1, \mathbf{z}_2} = \prod_{i=1}^{d}(z_{1i}, z_{2i}]$ the rectangle in $[0, 1]^d$ pertaining to $\mathbf{z}_1 = (z_{11}, \ldots, z_{1d}) \leq \mathbf{z}_2 = (z_{21}, \ldots, z_{2d})$. Fix $\mathbf{z} = (z_1, \ldots, z_d) \in [0, 1]^d$, and write μ and μ_n for the distributions of H and H_n, respectively. Inequality (12) now becomes, in obvious notation,

$$\begin{aligned} &\mathbb{P}\Big(\sup_{\mathbf{z}-\mathbf{a} \leq \mathbf{t} \leq \mathbf{z}+\mathbf{a}} |\alpha_n(I_{\mathbf{z}-\mathbf{a}, \mathbf{t}})| > s\sqrt{\mu(I_{\mathbf{z}-\mathbf{a}, \mathbf{z}+\mathbf{a}})}\Big) \\ &\qquad \leq C(\delta)\mathbb{P}\Big(|\alpha_n(I_{\mathbf{z}-\mathbf{a}, \mathbf{z}+\mathbf{a}})| > s(1 - 2\delta)^d\sqrt{\mu(I_{\mathbf{z}-\mathbf{a}, \mathbf{z}+\mathbf{a}})}\Big), \end{aligned} \tag{13}$$

where $\mathbf{a} = (a_1, \ldots, a_d) \geq \mathbf{0}$.

Now, recall $d = d_1 + d_2$ and fix $\mathbf{x} = (x_1, \ldots, x_{d_1}) \in [0, 1]^{d_1}$. Take $\mathbf{z} = (x_1, \ldots, x_{d_1}, \frac{1}{2}, \ldots, \frac{1}{2}) \in [0, 1]^d$ and put $\mathbf{a} = (\frac{1}{2}a, \ldots, \frac{1}{2}a, \frac{1}{2}, \ldots, \frac{1}{2})$. Observe that $\mu_n(I_{\mathbf{z}-\mathbf{a}, \mathbf{z}+\mathbf{a}})$ is the relative number of data points $\xi_i = (X_i^1, \ldots, X_i^{d_1}, Y_i^1, \ldots, Y_i^{d_2})$ such that

$$x_j - \frac{a}{2} < X_i^j \leq x_j + \frac{a}{2} \quad \text{for } 1 \leq j \leq d_1.$$

Note that

$$m_n^*(t_1,\dots,t_{d_2}|\mathbf{x}) = \mu_n\left(\prod_{j=1}^{d_1}\left(x_j - \frac{a_n}{2}, x_j + \frac{a_n}{2}\right] \times \prod_{j=1}^{d_2}[0, t_j]\right)/\mu(I_{\mathbf{z}-\mathbf{a},\mathbf{z}+\mathbf{a}}).$$

Put

$$\overline{m}_n^*(t_1,\dots,t_{d_2}|\mathbf{x}) = \mu\left(\prod_{j=1}^{d_1}\left(x_j - \frac{a_n}{2}, x_j + \frac{a_n}{2}\right] \times \prod_{j=1}^{d_2}[0, t_j]\right)/\mu(I_{\mathbf{z}-\mathbf{a},\mathbf{z}+\mathbf{a}}),$$

the expectation of m_n^*, and notice that $\sup_{\mathbf{z}-\mathbf{a}\le\mathbf{t}\le\mathbf{z}+\mathbf{a}}|\alpha_n(I_{\mathbf{z}-\mathbf{a},\mathbf{t}})|$ is an upper bound for

$$\sup_{\substack{0\le t_j\le 1\\ j=1,\dots,d_2}} n^{1/2}\mu(I_{\mathbf{z}-\mathbf{a},\mathbf{z}+\mathbf{a}})\left|m_n^*(t_1,\dots,t_{d_2}|\mathbf{x}) - \overline{m}_n^*(t_1,\dots,t_{d_2}|\mathbf{x})\right|.$$

Inequality (13) thus yields for $\varepsilon > 0$

$$(14)\qquad \begin{aligned} &\mathbb{P}\left(\sup_{\substack{0\le t_j\le 1\\ j=1,\dots,d_2}}\left|m_n^*(t_1,\dots,t_{d_2}|\mathbf{x}) - \overline{m}_n^*(t_1,\dots,t_{d_2}|\mathbf{x})\right| > \varepsilon\right)\\ &\quad\le C(\delta)\mathbb{P}\left(\left|\alpha_n(I_{\mathbf{z}-\mathbf{a},\mathbf{z}+\mathbf{a}})\right| > \varepsilon(1-2\delta)^d n^{1/2}\mu(I_{\mathbf{z}-\mathbf{a},\mathbf{z}+\mathbf{a}})\right),\end{aligned}$$

provided that $s = \varepsilon\sqrt{n\mu(I_{\mathbf{z}-\mathbf{a},\mathbf{z}+\mathbf{a}})}$ satisfies the usual growth conditions. By Lemma 2.2 of Devroye (1981) we have $s \to \infty$ for almost all $\mathbf{x}$ whenever $na_n^{d_1} \to \infty$, so that (14) is true for at least all $n \ge n_0(\mathbf{x})$. The right-hand side of (14) may be bounded from above by applying some standard exponential bound. See, e.g., inequality (1.2) in Stute (1982). In particular, since we are only interested in small $\varepsilon > 0$, we may assume w.l.o.g. that the first growth condition in (1.2) is satisfied. We have thus arrived at the following

LEMMA 5. *Given $0 < \delta < \frac{1}{2}$, we have for all $0 < \varepsilon \le \varepsilon_0(\delta)$ and $n \ge n_0(\mathbf{x}, \delta, \varepsilon)$*

$$\begin{aligned}&\mathbb{P}\left(\sup_{\substack{0\le t_j\le 1\\ j=1,\dots,d_2}}\left|m_n^*(t_1,\dots,t_{d_2}|\mathbf{x}) - \overline{m}_n^*(t_1,\dots,t_{d_2}|\mathbf{x})\right| > \varepsilon\right)\\ &\quad\le 2C(\delta)\exp\left[-(1-\delta)(1-2\delta)^d\varepsilon^2 n\mu(I_{\mathbf{z}-\mathbf{a},\mathbf{z}+\mathbf{a}})/2\right].\end{aligned}$$

Recall that for Lemma 5 we had tacitly assumed that H has uniform marginals. For an arbitrary H, remember (9) and (10). Make the transformation $t_j = F_j(x_j)$ for $1 \le j \le d_1$ and $t_j = F_j(y_j)$ for $d_1 < j \le d$. In the fundamental inequality (13) the sup now has to be extended over the set of $\mathbf{t}$'s for which

$$F_j\left(x_j - \frac{a_n}{2}\right) \le t_j \le F_j\left(x_j + \frac{a_n}{2}\right), \qquad j = 1,\dots,d_1.$$

Since in no other step of the proof has the assumption of marginal uniformity been crucial this shows that Lemma 5 remains true also in the general case, with $0 \leq t_j \leq 1$ replaced by $y_j \in \mathbb{R}$. We are now in a position to give the

PROOF OF THEOREM 1. It follows from the proof of Lemma 2.2 in Devroye (1981) that, as $a = a_n \to 0$, with $\mathbf{z} = (x_1, \ldots, x_{d_1}, \frac{1}{2}, \ldots, \frac{1}{2})$,

$$\frac{\mu(I_{\mathbf{z}-\mathbf{a},\mathbf{z}+\mathbf{a}})}{a_n^{d_1}} \to c(\mathbf{x}) \quad \text{for almost all } \mathbf{x},$$

for some positive $c(\mathbf{x})$, possibly infinite. We thus get for all small $\varepsilon > 0$ and $n \geq n_0$

$$\mathbb{P}\big(\|m_n^*(\cdot|\mathbf{x}) - \overline{m}_n^*(\cdot|\mathbf{x})\| > \varepsilon\big) \leq K \exp\big[-c(\varepsilon) n a_n^{d_1}\big],$$

where $K < \infty$ and $c(\varepsilon) > 0$. Use Borel–Cantelli to get

$$\|m_n^*(\cdot|\mathbf{x}) - \overline{m}_n^*(\cdot|\mathbf{x})\| \to 0 \quad \text{almost surely}$$

for almost all $\mathbf{x}$. To prove Theorem 1 it remains to show [use (4)] that

$$\sup_{\mathbf{y} \in \mathbb{R}^{d_2}} |\overline{m}_n^*(\mathbf{y}|\mathbf{x}) - m(\mathbf{y}|\mathbf{x})| \to 0.$$

For $\mathbf{y}$ fixed, $\overline{m}_n^*(\mathbf{y}|\mathbf{x}) \to m(\mathbf{y}|\mathbf{x})$ follows from standard results in real analysis [cf., e.g., Wheeden and Zygmund (1977), page 189], with the negligible set of $\mathbf{x}$'s depending on $\mathbf{y}$. To find a universal null set use monotonicity of $\overline{m}_n^*(\cdot|\mathbf{x})$ and $m(\cdot|\mathbf{x})$ and apply a standard Polya–Cantelli-type argument [cf. Polya and Szegö (1972), page 81, example 127]. Alternatively, one could also apply one of the available more sophisticated "uniform convergence of measures" concepts. See Billingsley and Topsøe (1967) or Gaenssler and Stute (1976). □

PROOF OF THEOREM 2. In view of Theorem 1, to treat a weighted discrepancy, it remains to show that

$$\frac{[m_n(\mathbf{y}|\mathbf{x}) - m(\mathbf{y}|\mathbf{x})]}{h(\mathbf{y})} \to 0$$

uniformly on the set of $\mathbf{y}$'s for which at least one of its coordinates is small. To be precise, fix some small $c > 0$ and observe that

$$D_n(h) \leq c^{-1} D_n(1) + \sup_{\mathbf{y}:\, h(\mathbf{y}) < c} \frac{m_n(\mathbf{y}|\mathbf{x})}{h(\mathbf{y})} + \sup_{\mathbf{y}:\, h(\mathbf{y}) < c} \frac{m(\mathbf{y}|\mathbf{x})}{h(\mathbf{y})}.$$

By Theorem 1 for almost all $\mathbf{x}$ the first term converges to zero with probability one. By Markov's inequality, the third summand is less than or equal to

$$\int_{\{h^{-1}(\mathbf{z}) \geq c^{-1}\}} h^{-1}(\mathbf{z}) m(d\mathbf{z}|\mathbf{x}). \tag{15}$$

Since $h^{-1}(\mathbf{Y})$ has finite expectation, we have that h^{-1} is $m(\cdot|\mathbf{x})$-integrable for almost all $\mathbf{x}$. For such an $\mathbf{x}$ the integral (15) may be made arbitrarily small by letting $c \downarrow 0$.

So far, the method of proof has been similar to the unconditional case, with the usual Glivenko–Cantelli theorem replaced by statement (6) of this paper [see, e.g., Wellner (1977)]. While, in the classical setup, the ordinary strong law of large numbers is then in place in order to treat the middle summand, such a complete result is not at hand in the case of m_n. Instead we shall use a factorization

$$\frac{m_n^*(\mathbf{y}|\mathbf{x})}{h(\mathbf{y})} = X_n(x, y)Z_n(x, y) \equiv X(y)Z(y),$$

where Z is nonrandom and X is a reverse martingale, to which an exponential bound of Wellner (1978) applies.

Because of (4) it remains to show that

$$\sup_{\mathbf{y}:\, h(\mathbf{y})<c} \frac{m_n^*(\mathbf{y}|\mathbf{x})}{h(\mathbf{y})} \tag{16}$$

gets small with probability one for c sufficiently small. Since, by monotonicity of h, $\mathbf{Y}_i \le \mathbf{y}$ implies $h(\mathbf{Y}_i) \le h(\mathbf{y})$ (16) is bounded from above by

$$\sup_{0<y<c} \frac{m_n^*(y|\mathbf{x})}{y}, \tag{17}$$

where m_n^* now pertains to the transformed $h(\mathbf{Y}_i)$, $i = 1, 2, \dots$. In other words, in what follows we may assume w.l.o.g. $d_2 = 1$, with $Y_i \equiv \mathbf{Y}_i$ positive and h the identity function on $(0, \infty)$. Now, for $0 < y$, write

$$\frac{m_n^*(y|\mathbf{x})}{y} = X(y)Z(y),$$

where

$$X(y) = \frac{\mu_n\left(\prod_{j=1}^{d_1}(x_j - a_n/2, x_j + a_n/2] \times [0, y]\right)}{\mu\left(\prod_{j=1}^{d_1}(x_j - a_n/2, x_j + a_n/2] \times [0, y]\right)}$$

and

$$Z(y) = \frac{\mu\left(\prod_{j=1}^{d_1}(x_j - a_n/2, x_j + a_n/2] \times [0, y]\right)}{\mu(I_{\mathbf{z}-\mathbf{a},\mathbf{z}+\mathbf{a}})y}.$$

For studying the X-process, because of (9) and (10), it suffices to consider the case when Y_i has a continuous distribution function. Given $0 < q < 1$, define y_m as to be the infimum of those y's for which

$$\mu\left(\prod_{j=1}^{d_1}\left(x_j - \frac{a_n}{2}, x_j + \frac{a_n}{2}\right] \times [0, y]\right) = q^m \mu(I_{\mathbf{z}-\mathbf{a},\mathbf{z}+\mathbf{a}}).$$

By continuity, y_m is well defined for $m \ge 0$, with y_0 possibly being infinite. Also $y_m \downarrow$. Choose m_0 so that

$$y_{m_0+1} < c \le y_{m_0},$$

and let $m_1 = m_1(n)$ be defined by

$$q^{m_1+1} < \frac{\ln n}{na_n^{d_1}} = c_n \leq q^{m_1}.$$

We then have for each $\varepsilon > 0$ and every $n \in \mathbb{N}$

$$\mathbb{P}\left(\sup_{y_{m_1+1} \leq y < c} \frac{m_n^*(y|\mathbf{x})}{y} > \varepsilon \right) \leq \sum_{m=m_0}^{m_1} \mathbb{P}\left(\sup_{y_{m+1} \leq y} X(y) \geq \varepsilon q Z^{-1}(y_{m+1}) \right).$$

Since X as a process in y has the same distribution theory as a univariate empirical distribution function (and, in particular, is a reverse martingale) we may apply Lemma 1 in Wellner (1978) to get

$$\mathbb{P}\left(\sup_{y_{m+1} \leq y} X(y) \geq \varepsilon q Z^{-1}(y_{m+1}) \right) \leq \exp\left[-nq^{m+1}\mu(I_{\mathbf{z}-\mathbf{a},\mathbf{z}+\mathbf{a}}) h\left(\varepsilon q Z^{-1}(y_{m+1}) \right) \right],$$

provided that

$$\varepsilon q Z^{-1}(y_{m+1}) \geq 1. \tag{18}$$

Here

$$h(x) = x(\ln x - 1) + 1, \qquad x > 0.$$

As to (18) note that

$$Z(y_{m+1}) \leq \frac{\mathbb{E}\left(1_{\{\xi \in I_{\mathbf{z}-\mathbf{a},\mathbf{z}+\mathbf{a}},\, Y \leq c\}} Y^{-1} \right)}{\mu(I_{\mathbf{z}-\mathbf{a},\mathbf{z}+\mathbf{a}})},$$

which converges to

$$\mathbb{E}\left(1_{\{Y \leq c\}} Y^{-1} \middle| \mathbf{X} = \mathbf{x} \right)$$

for almost all $\mathbf{x}$ as $n \to \infty$. The last term can be made arbitrarily small by letting $c \to 0$. Thus the left-hand side of (18) can in fact be made arbitrarily large uniformly in m, at least for all large n, by choosing c sufficiently small. Furthermore,

$$\frac{\mu(I_{\mathbf{z}-\mathbf{a},\mathbf{z}+\mathbf{a}})}{a_n^{d_1}} \to c(\mathbf{x}) > 0$$

for almost all $\mathbf{x}$. In summary, since $m_1 = O(\ln n)$,

$$\mathbb{P}\left(\sup_{y_{m_1+1} \leq y < c} \frac{m_n^*(y|\mathbf{x})}{y} > \varepsilon \right) = O(n^{-2})$$

for c sufficiently small. Borel–Cantelli yields

$$\limsup_{n \to \infty} \sup_{y_{m_1+1} \leq y < c} \frac{m_n^*(y|\mathbf{x})}{y} \leq \varepsilon$$

with probability one. Furthermore,

$$\sup_{c_n/\varepsilon \le y \le y_{m_1+1}} \frac{m_n^*(y|\mathbf{x})}{y} \le X(y_{m_1+1})\varepsilon,$$

which is easily seen to be of order $O(\varepsilon)$.

Next we shall show that $m_n^*(c_n/\varepsilon) = 0$ eventually with probability one, establishing the proof of the theorem. For this, since $c_n \downarrow 0$ by assumption, it suffices to show that $\mathbb{P}(Y \le c_n/\varepsilon)$, $n \ge 1$, is summable. This follows, however, immediately from the fact that $1/Y$ has a finite rth moment and c_n^r, $n \ge 1$, is summable. □

REMARK. The assumption $\sum c_n^r < \infty$ has been needed only to guarantee $m_n^*(c_n/\varepsilon) = 0$ eventually. As we shall see, under absolute continuity of $\mathbf{X}$, the growth conditions on c_n (and hence on a_n) may be substantially weakened. In fact, when $a_n \downarrow 0$ and $c_n \downarrow 0$, $m_n^*(c_n/\varepsilon) = 0$ eventually also follows from the summability of

$$\mathbb{P}\left(\mathbf{X} \in \prod_{j=1}^{d_1}\left(x_j - \frac{a_n}{2}, x_j + \frac{a_n}{2}\right], \mathbf{Y} \le c_n/\varepsilon\right), \qquad n \ge 1.$$

Under a finite rth moment assumption this is easily seen to be of order

$$c_n^r \mathbb{P}\left(\mathbf{X} \in \prod_{j=1}^{d_1}\left(x_j - \frac{a_n}{2}, x_j + \frac{a_n}{2}\right]\right).$$

From differentiation theory, the last probability is of order $a_n^{d_1}$ for Lebesgue almost $\mathbf{x} \in \mathbb{R}^{d_1}$ and thus, by absolute continuity, for $\mathbf{X}$-almost all $\mathbf{x}$. The summability criterion thus becomes

$$\sum c_n^r a_n^{d_1} < \infty.$$

PROOF OF THEOREM 3. First of all, a version of Theorem 2 also holds true with $m_n(\mathbf{y}|\mathbf{x})$ replaced by $1 - m_n(\mathbf{y}|\mathbf{x})$, with h nonincreasing. Just replace Y_i by $-Y_i$. We may also assume that the Y's are nonnegative.

The result then follows from the equation

$$m_n(\mathbf{x}) - m(\mathbf{x}) = \int_0^\infty \frac{(1 - m_n(y|\mathbf{x}) - (1 - m(y|\mathbf{x})))h(y)\,dy}{h(y)},$$

where

$$h(y) = \begin{cases} 1 & \text{for } y \le 1, \\ y^{-1} & \text{for } y > 1, \end{cases}$$

and where $p > 1$ is so small that $pr < r + \rho$. By assumption $\mathbb{E}(h^{-r}(Y)) < \infty$. Since h is Lebesgue-integrable, the result is an immediate consequence of Theorem 2. □

Acknowledgments. Thanks to a referee and an associate editor for their editorial criticism on an earlier draft of the paper.

REFERENCES

BILLINGSLEY, P. and TOPSØE, F. (1967). Uniformity in weak convergence. *Z. Wahrsch. verw. Gebiete* **7** 1–16.

COLLOMB, G. (1981). Estimation non paramétrique de la regression: revue bibliographique. *Internat. Statist. Rev.* **49** 75–93.

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

DEVROYE, L. and WAGNER, T. J. (1980). On the L_1 convergence of kernel regression function estimators with applications in discrimination. *Z. Wahrsch. verw. Gebiete* **51** 15–25.

DVORETZKY, A., KIEFER, J. and WOLFOWITZ, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.* **27** 642–669.

GAENSSLER, P. and STUTE, W. (1976). On uniform convergence of measures with applications to uniform convergence of empirical distributions. *Lecture Notes in Math.* **566** 45–56. Springer, Berlin.

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

KIEFER, J. (1961). On large deviations of the empiric d.f. of vector chance variables and a law of the iterated logarithm. *Pacific J. Math.* **11** 649–660.

MASON, D. M. (1982). Some characterizations of almost sure bounds for weighted multidimensional empirical distributions and a Glivenko–Cantelli theorem for sample quantiles. *Z. Wahrsch. verw. Gebiete* **59** 505–513.

NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141–142.

NADARAYA, E. A. (1970). Remarks on nonparametric estimates for density functions and regression curves. *Theory Probab. Appl.* **15** 134–137.

POLYA, G. and SZEGÖ, G. (1972). *Problems and Theorems in Analysis* **1**. Springer, Berlin.

STUTE, W. (1982). The oscillation behavior of empirical processes. *Ann. Probab.* **10** 86–107.

STUTE, W. (1984). The oscillation behavior of empirical processes: the multivariate case. *Ann. Probab.* **12** 361–379.

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

WELLNER, J. A. (1977). A Glivenko–Cantelli theorem and strong laws of large numbers for functions of order statistics. *Ann. Statist.* **5** 473–480.

WELLNER, J. A. (1978). Limit theorems for the ratio of the empirical distribution function to the true distribution function. *Z. Wahrsch. verw. Gebiete* **45** 73–88.

WHEEDEN, R. L. and ZYGMUND, A. (1977). *Measure and Integral.* Marcel Dekker, New York.

MATHEMATISCHES INSTITUT DER
JUSTUS-LIEBIG-UNIVERSITÄT
ARNDTSTRAßE 2
D-6300 GIESSEN
FEDERAL REPUBLIC OF GERMANY