On the strong universal consistency of nearest neighbor regression function estimates *

Luc Devroye †   László Györfi ‡   Adam Krzyżak §   Gábor Lugosi ¶

June 1992

Abstract

Two results are presented concerning the consistency of the k-nearest neighbor regression estimate. We show that all modes of convergence in L_1 (in probability, almost sure, complete) are equivalent if the regression variable is bounded. Under the additional condition $k/\log n \to \infty$ we also obtain the strong universal consistency of the estimate.

*Supported in part by NSERC grants A3456 and A0270, and FCAR grants EQ-1679 and EQ-2904

†School of Computer Science, McGill University, Montreal, Canada H3A 2A7

‡Department of Mathematics, Technical University of Budapest, 1521 Stoczek u. 2, Budapest, Hungary

§Department of Computer Science, Concordia University, 1455 de Maisonneuve West, Montreal, Canada H3G 1M8

¶Department of Mathematics, Technical University of Budapest, 1521 Stoczek u. 2, Budapest, Hungary

1 Introduction

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent observations of an $\mathcal{R}^d \times \mathcal{R}$-valued random vector (X, Y). Denote the probability measure of X by μ. The regression function $m(x) = \mathbf{E}(Y|X = x)$ can be estimated by the *kernel estimate*

$$m_n(x) = \frac{\sum_{i=1}^n Y_i K_h(x - X_i)}{\sum_{i=1}^n K_h(x - X_i)},$$

where $h > 0$ is a smoothing factor depending upon n, K is an absolutely integrable function (the kernel), and $K_h(x) = K(x/h)$ (Nadaraya (1964, 1970), Watson (1964)). Alternatively, one can use the k-nearest neighbor estimate,

$$m_n(x) = \sum_{i=1}^n W_{ni}(x; X_1, \ldots, X_n) Y_i \ ,$$

and $W_{ni}(x; X_1, \ldots, X_n)$ is $1/k$ if X_i is one of the k nearest neighbors of x among $X_1, \ldots, X_n$, and W_{ni} is zero otherwise. Note in particular that $\sum_{i=1}^n W_{ni} = 1$. The k-nearest neighbor estimate was studied by Cover (1968). For a survey of other estimates, see e.g. Collomb (1981, 1985) or Györfi (1981).

We are concerned with the L_1 convergence of m_n to m as measured by $J_n = \int |m_n(x) - m(x)|\mu(dx)$ where μ is the (unknown) probability measure for X. This quantity is particularly important in discrimination based on the kernel rule (see Devroye and Wagner (1980) or Stone (1977)). Stone (1977) first pointed out that there exist estimators for which $J_n \to 0$ in probability for all distributions of (X, Y) with $\mathbf{E}|Y| < \infty$. This included the nearest neighbor and histogram estimates. For example, for the k-nearest neighbors, it suffices to ask that

$$k \to \infty, \quad k/n \to 0 \ , \tag{1}$$

provided that ties among points at equal distance from x are adequately taken care of. These conditions are the best possible. In 1980, Devroye and Wagner, and independently Spiegelman and Sacks, showed that this is also the case for the kernel estimate with smoothing factor h provided that K is a bounded nonnegative function with compact support such that for a small fixed sphere S centered at the origin, $\inf_{x \in S} K(x) > 0$, and that

$$\lim_{n \to \infty} h = 0, \quad \lim_{n \to \infty} nh^d = \infty. \tag{2}$$

These results were extended and complemented by Greblicki, Krzyżak and Pawlak (1984), Krzyżak (1986) and Krzyżak and Pawlak (1984).

Interestingly, it turns out that the conditions for the "in probability" convergence of J_n are also sufficient for the strong convergence of J_n, thus rendering all modes of convergence equivalent. Difficulties arise when the X-variable does not have an absolutely continuous distribution. We summarize what is known in this respect:

- For the *k-nearest neighbor estimates*, $J_n \to 0$ almost surely under condition (1) whenever X has a density and Y is bounded (chapter X of Devroye and Györfi, 1985 and Zhao, 1987). Beck (1979) showed this result earlier under the additional constraint that m has a continuous version.

- For the *k-nearest neighbor estimate*, $J_n \to 0$ almost surely for all distributions of (X, Y) with Y bounded, provided that $k/n \to 0$ and $k/\log\log n \to \infty$ (Devroye, 1982). The unnatural condition on k arises from the proof method: the convergence of J_n to 0 is obtained by first establishing the pointwise convergence (i.e., $m_n - m \to 0$ almost surely) at almost all x (μ), and move on to L_1 convergence via a result of Glick's (1974).

- Devroye and Györfi (1983) obtained the equivalence for all distributions of (X, Y) with $|Y| \leq M < \infty$ for the *histogram regression estimate*. Györfi (1991) has pointed out that $J_n \to 0$ almost surely for a modification of *partitioning estimates* whenever $\mathbf{E}|Y| < \infty$, provided that a bin width condition similar to (2) (with the additional condition $nh^d/\log(n) \to \infty$) is satisfied.

- Assuming that Y is uniformly bounded, the *kernel estimate* is strongly consistent if (2) holds, K is a Riemann integrable kernel and $K \geq aI_S$, where $a > 0$ is a constant, and S is a ball centered at the origin that has a positive radius (Devroye and Krzyżak, 1989).

The purpose of the present paper is twofold. First we explain a simple technique based upon exponential martingale inequalities for proving the equivalence of all modes of convergence of J_n for the k-nearest neighbor estimate under no conditions on the distribution of (X, Y) other than the boundedness of Y. Thus, Stone's conditions are strong enough to imply complete and almost sure convergence. Our other result is the strong universal convergence of J_n, that is, we can replace the boundedness assumption by the natural condition $\mathbf{E}|Y| < \infty$. Here we need the additional condition $k/\log n \to \infty$ on k.

Before we can state the main results, we have to take care of the messy problem of distance ties ($\|x - X_i\| = \|x - X_j\|$). The exponential inequality used here and in Devroye and Krzyżak (1989) is basically useful whenever the removal of one data point has a limited effect on the error. Also, our covering lemma requires some sort of duality that states that if X_i is one of the near neighbors of X_j, then roughly speaking, X_j should be one of the near neighbors of X_i. Next we list three of the possible tie-breaking methods.

- **Tie breaking by indices:** If X_i and X_j are equidistant from x, then X_i is declared "closer" if $i < j$. This method has some undesirable properties. For example, if X is monoatomic, then X_1 is the nearest neighbor of all X_j's, $j > 1$, but X_j is only the $j-1$-st nearest neighbor of X_1. The influence of X_1 in such a situation is too great, making the estimate very unstable and thus undesirable. In this case, Devroye and Györfi (1985, chapter X) pointed out that when Y is not degenerate and $\epsilon > 0$ is small enough,

$$\mathbf{P}\left\{\int |\tilde{m}_n(x) - m(x)|\mu(dx) > \epsilon\right\} \geq e^{-ck}$$

  for some $c > 0$. This is in contrast to A in Theorem 1 below, where $\tilde{m}_n(x)$ is the $k-$nearest neighbor regression estimate defined by tie breaking by indices.

- **Stone's tie breaking:** Stone (1977) introduced a nearest neighbor rule, which is not a $k-$nearest neighbor rule in a strict sense, for his estimate, in general, uses more than k neighbors. If we denote the distance of the k-th nearest neighbor to x by $R_n(x)$ (note, that it is unique), then Stone's estimate is the following:

$$\hat{m}_n(x) = \frac{1}{k}\left(\sum_{i:\|x-x_i\|<R_n(x)} Y_i + \frac{k - \#\{i : \|x - x_i\| < R_n(x)\}}{\#\{i : \|x - x_i\| = R_n(x)\}} \sum_{i:\|x-x_i\|=R_n(x)} Y_i\right). \tag{3}$$

- **Tie breaking by randomization:** This is the method that we will consider. We assume that (X, Z) is a random vector independent of the data, where Z is independent of X and uniformly distributed on $[0, 1]$. The latter assumption may be replaced by the weaker assumption that Z has a density; however, as it is up to us to generate Z, we may as well pick a uniform random variable. We also artificially enlarge the data by introducing $Z_1, Z_2, \ldots, Z_n$, where the Z_i's are i.i.d. uniform $[0,1]$ as well. Thus, each (X_i, Z_i) is distributed as (X, Z). The probability measure induced by (X, Z) is denoted by ν. Given (x, z), we define

$$m_n(x,z) = \frac{1}{k}\sum_{i=1}^{k} Y_{(i)} \ ,$$

where

$$(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)}) = (X_{(1,x,z)}, Y_{(1,x,z)}), \ldots, (X_{(n,x,z)}, Y_{(n,x,z)})$$

is a reordering of the data according to increasing values of $\|x - X_{(i)}\|$. In case of distance ties, we declare (X_i, Z_i) closer to (x, z) than (X_j, Z_j) provided that

$$|Z_i - z| \leq |Z_j - z|.$$

The criterion is

$$\begin{aligned} J_n &= \mathbf{E}\left\{|m_n(X,Z) - m(X)|\Big| X_1, Z_1, Y_1, \ldots, X_n, Z_n, Y_n\right\} \ . \\ &= \int_0^1 \int |m_n(x,z) - m(x)|\mu(dx)\, dz = \int |m_n(x,z) - m(x)|\nu(d(x,z)). \end{aligned}$$

It should be stressed that if μ has a density, then tie-breaking is needed with zero probability, and becomes therefore irrelevant.

2 The equivalence theorem

The purpose of this section is to prove the following result.

Theorem 1 *Let $m_n(x,z)$ be the k-nearest neighbor estimate defined above. Then the following statements are equivalent:*

A. *For every distribution of (X, Y) with $|Y| \leq M < \infty$ and $\epsilon > 0$ there is a positive integer n_0 such that for $n > n_0$*

$$\mathbf{P}\{J_n > \epsilon\} \leq e^{-n\epsilon^2/(8M^2\gamma_d^2)},$$

where the constant γ_d is the minimal number of cones centered at the origin of angle $\pi/6$ that cover $\mathcal{R}^d$.

B. *For every distribution of (X, Y) with $|Y| \leq M < \infty$,*

$$J_n \to 0 \quad \text{with probability one as } n \to \infty \ .$$

C. *For every distribution of* (X, Y) *with* $|Y| \leq M < \infty$,

$$J_n \to 0 \quad \textit{in probability as } n \to \infty \ .$$

D.

$$\lim_{n\to\infty} k = \infty, \quad \lim_{n\to\infty} k/n = 0 \ .$$

Remark 1. A CURIOSITY. It is interesting that we can find sequences k for which $J_n \to 0$ almost surely for all distributions of (X, Y) with bounded Y, yet m_n does not tend to m in the almost sure pointwise sense (take $k \sim \log\log\log(n)$, and note that $k/\log\log(n) \to \infty$ is necessary for the almost sure pointwise convergence of the kernel estimate whenever X has a density and m is twice continuously differentiable with $m'' \neq 0$ (Devroye, 1982)).

Remark 2. NECESSITY OF THE CONDITIONS. It is not true that when $J_n \to 0$ in probability for one distribution of (X, Y), the conditions (1) on k follow: just consider the case that $Y = 0$ with probability one. Of course, the implication is true for "most" distributions of (X, Y).

Remark 3. GENERAL ESTIMATES. We won't consider smoothed versions of the k-nearest neighbor method here. For example, as in Devroye (1982), one might consider attaching weight v_{ni} to the i-th nearest neighbor, where $v_{n1} \geq v_{n2} \geq \cdots \geq v_{nn} \geq 0$ and the weights sum to one for every n. Such methods were first proposed by Royall (1966).

Remark 4. OTHER REFERENCES. For other results on k nearest neighbor convergence, see e.g., Collomb (1979, 1980), Mack (1981), Devroye (1978,1981,1982), Stute (1984) and Bhattacharya and Mack (1987).

Remark 5. RANDOM k. If k is replaced by a random variable K that is independent of the data and satisfies $K/n \to 0$ and $K \to \infty$ almost surely, then $J_n \to 0$ almost surely. Such data-based choices can be obtained by splitting the data for example.

Remark 6. DISCRIMINATION. The conditional probability of error, of the k-nearest neighbor rule in discrimination, given the data, (Cover and Hart, 1967) converges completely and strongly to the Bayes probability of error as $n \to \infty$ for all distributions of the data whenever (1) holds. This result strengthens the universal weak convergence results of Stone (1977) and Devroye and Wagner (1980).

Remark 7. L_p-CONSISTENCY. By the boundedness of Y it is easy to see that the L_p-error $J_n^{(p)} = \left(\int_0^1 \int |m_n(x,z) - m(x)|^p \mu(dx)\,dz\right)^{1/p}$ for $1 \le p < \infty$ converges to zero if and only if the L_1-error J_n does, therefore the results of Theorem 1 remain valid for L_p-errors.

Remark 8. INEQUALITIES. The inequality of part A of Theorem 1 is less useful in practice as it is only valid for $n > n_0$, where n_0 depends upon ϵ.

Proof of the theorem. Clearly, A implies B and B implies C. C implies that $\mathbf{E}J_n \to 0$, therefore, by Jensen's inequality

$$\begin{aligned}
\mathbf{E}J_n &= \mathbf{E}\left(\int |m_n(x,z) - m(x)|\ \nu(d(x,z))\right) \\
&\ge \mathbf{E}\left(\int \left|\mathbf{E}\left(\int m_n(x,z)dz | X_1, Y_1, \ldots, X_n, Y_n\right) - m(x)\right|\mu(dx)\right) \\
&= \mathbf{E}\left(\int |\hat{m}_n(x) - m(x)|\mu(dx)\right) \to 0,
\end{aligned}$$

where $\hat{m}_n$ is Stone's estimate defined by (3). But this implies D by the results of Stone (1977). The novelty in this paper is the proof that condition D implies A. We begin with an exponential inequality generalizing inequalities due to Hoeffding (1963). The generalization due to Azuma (1967) (see Stout, 1974) has led to interesting applications in combinatorics and the theory of random graphs (for a survey, see McDiarmid, 1989). We have used it in density estimation (Devroye, 1988, 1991).

Lemma 1 (MCDIARMID, 1989) *Let $X_1, \ldots, X_n$ be independent random variables taking values in a set A, and assume that $f : A^n \to R$ satisfies*

$$\sup_{\substack{x_1,\ldots,x_n \\ x_1',\ldots,x_n' \in A}} |f(x_1,\ldots,x_n) - f(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)| \le c_i \ , \ 1 \le i \le n \ .$$

Then

$$\mathbf{P}\left\{|f(X_1,\ldots,X_n) - \mathbf{E}f(X_1,\ldots,X_n)| \ge t\right\} \le 2e^{-2t^2/\sum_{i=1}^n c_i^2} \ .$$

The other tool needed to our proof is exploiting some geometric properties of the "metric" defined by the tie-breaking rule. In order to make it more transparent, we recall Lemma 10.1 from Devroye and Györfi (1985), that was used in the proof of complete consistency of $k-$nearest neighbor estimates if μ has a density. Denote by $S_{x,r}$ and $\bar{S}_{x,r}$ the open and closed balls of radius r centered at x, respectively.

Lemma 2 (Devroye, Györfi, 1985) *Let μ be an absolutely continuous probability measure on $\mathcal{R}^d$. Define*

$$B_a(x) = \left\{ x' : \mu(\bar{S}_{x',\|x-x'\|}) \leq a \right\},$$

Then for all $x \in \mathcal{R}^d$

$$\mu(B_a(x)) \leq \gamma_d a \ .$$

Since Devroye and Györfi assumed the existence of a density, they did not have to worry about tie breaking. In order to generalize Lemma 2 to our case, we need some notations. For $x \in \mathcal{R}^d$ let $C(x) \subset \mathcal{R}^d$ be a cone of angle $\pi/6$ centered at x. The cone consists of all y with the property that either $y = x$ or angle$(y - x, s) \leq \pi/6$, where s is a fixed direction. If $y, y' \in C(x)$, and $\|x - y\| < \|x - y'\|$, then $\|y - y'\| < \|x - y'\|$. Furthermore, if $\|x - y\| \leq \|x - y'\|$, then $\|y - y'\| \leq \|x - y'\|$. This follows from a simple geometric argument in the vector space spanned by x, y and y'.

For $(x, z) \in \mathcal{R}^d \times [0, 1]$ define $C_0(x, z), C_1(x, z), S_{(x,z),(r,b)} \subset \mathcal{R}^d \times [0, 1]$ as

$$C_0(x, z) = C(x) \times [0, z],$$

$$C_1(x, z) = C(x) \times [z, 1],$$

and

$$S_{(x,z),(r,b)} = S_{z,r} \times [0, 1] \cup \{(\hat{x}, \hat{z}) : \|\hat{x} - x\| = r, |\hat{z} - z| \leq b\} .$$

Clearly, $\mathcal{R}^d \times [0, 1]$ can be covered by $2\gamma_d$ sets of type $C_0(x, z)$ and $C_1(x, z)$.

The property that we need is the following.

Lemma 3 *Let*

$$B_a(x, z) = \left\{ (x', z') : \nu(S_{(x',z'),(\|x-x'\|,|z-z'|)}) \leq a \right\}.$$

Then for all $(x, z) \in \mathcal{R}^d \times [0, 1]$

$$\nu(B_a(x, z)) \leq 2\gamma_d a \ .$$

.

First of all, we prove a covering lemma, the key property of the "cones" $C_0(x, z)$ and $C_1(x, z)$.

Lemma 4 *If $(x', z') \in C_0(x, z)$ then*

$$C_0(x, z) \cap S_{(x,z),(\|x-x'\|,|z-z'|)} \subset S_{(x',z'),(\|x-x'\|,|z-z'|)}$$

and if $(x',z') \in C_1(x,z)$ *then*

$$C_1(x,z) \cap S_{(x,z),(\|x-x'\|,|z-z'|)} \subset S_{(x',z'),(\|x-x'\|,|z-z'|)}.$$

Proof Because of symmetry it is enough to prove one of the statements. We have to show that $(\hat{x},\hat{z}) \in C_0(x,z) \cap S_{(x,z),(\|x-x'\|,|z-z'|)}$, implies $(\hat{x},\hat{z}) \in S_{(x',z'),(\|x-x'\|,|z-z'|)}$.

If $\hat{x} \in C(x) \cap S_{x,\|x-x'\|}$, then from the well known property of the cone $\hat{x} \in S_{x',\|x-x'\|}$ follows, so it is enough to deal with pairs $(\hat{x},\hat{z})$ where $\|\hat{x}-x\| = \|x-x'\|$. Since $\hat{x} \in C(x)$, the only case when $\hat{x} \notin S_{x',\|x-x'\|}$ is if

$$\|\hat{x}-x\| = \|\hat{x}-x'\| = \|x'-x\|.$$

Denote the set of such $\hat{x}$'s by H. Thus, it is enough to deal with pairs in $H \times [0,1]$. Intersecting this set with the left and right hand sides of the statement we get

$$H \times [0,1] \cap C_0(x,z) \cap S_{(x,z),(\|x-x'\|,|z-z'|)} =$$
$$H \times \left([0,z] \cap \{\hat{z} : |\hat{z}-z| \leq |z-z'|\}\right)$$

and

$$H \times [0,1] \cap S_{(x',z'),(\|x-x'\|,|z-z'|)} = H \times \left(\{\hat{z} : |\hat{z}-z'| \leq |z-z'|\}\right),$$

respectively. We are ready if we show that

$$[0,z] \cap \{\hat{z} : |\hat{z}-z| \leq |z-z'|\} \subset \{\hat{z} : |\hat{z}-z'| \leq |z-z'|\}.$$

This follows easily from the assumption that $z' \in [0,z]$. □

Proof of Lemma 3 The proof is similar to that of Lemma 2. Let $C_1,\ldots,C_{2\gamma_d}$ be a collection of sets of form $C_0(x,z)$ and $C_1(x,z)$ that cover $\mathcal{R}^d \times [0,1]$. Then

$$\nu(B_a(x,z)) \leq \sum_{s=1}^{2\gamma_d} \nu(C_s \cap B_a(x,z)).$$

Let $(x',z') \in C_s \cap B_a(x,z)$. Then from Lemma 4 we have

$$\nu(C_s \cap S_{(x,z),(\|x-x'\|,|z-z'|)} \cap B_a(x,z)) \leq \nu(S_{(x',z'),(\|x-x'\|,|z-z'|)}) \leq a,$$

where we used the fact that $(x',z') \in B_a(x,z)$. Since (x',z') was arbitrary,

$$\nu(C_s \cap B_a(x,z)) \leq a,$$

which completes the proof of the lemma. $\square$

Now, we are equipped to prove that D implies A in Theorem 1. Set $r_n = r_n(x,z)$ and $b_n = b_n(x,z)$ to satisfy

$$\nu(S_{(x,z),(r_n,b_n)}) = \frac{k}{n}.$$

Note, that the solution always exists by the absolute continuity of the distribution of Z and its independence from X. Also define

$$m_n^*(x,z) = \frac{1}{k}\sum_{j=1}^{n} Y_j I_{\{(X_j,Z_j)\in S_{(x,z),(r_n,b_n)}\}}$$

Obviously,

$$\begin{aligned} |m(x) - m_n(x,z)| &\leq |m(x) - \mathbf{E}(m_n^*(x,z))| \\ &+ |\mathbf{E}(m_n^*(x,z)) - m_n^*(x,z)| + |m_n^*(x,z) - m_n(x,z)| \end{aligned} \tag{4}$$

$k/n \to 0$ implies that $r_n(x,z) \to 0$, so for the first term we have by Lebesgue's density theorem (see Wheeden and Zygmund, 1977) that

$$\begin{aligned} \mathbf{E}(m_n^*(x,z)) = \frac{1}{\nu(S_{(x,z),(r_n,b_n)})} \int_{S_{(x,z),(r_n,b_n)}} & \mathbf{E}(Y|(X,Z) = (x',z'))\ \nu(d(x',z')) \\ & \to \mathbf{E}(Y|(X,Z) = (x,z)) = m(x). \end{aligned}$$

for almost all x mod μ. By the boundedness of Y the dominated convergence theorem implies that

$$\int |m(x) - \mathbf{E}(m_n^*(x,z))|\ \nu(d(x,z)) \to 0.$$

Turning to the second term in (4) first we get an exponential bound for

$$\left| \int |\mathbf{E}(m_n^*(x,z)) - m_n^*(x,z)|\ \nu(d(x,z)) - \mathbf{E}\int |\mathbf{E}(m_n^*(x,z)) - m_n^*(x,z)|\ \nu(d(x,z)) \right|$$

by Lemma 1. Fix the data and replace (x_i, z_i, y_i) by $(\hat{x}_i, \hat{z}_i, \hat{y}_i)$ changing the value of $m_n^*(x,z)$ to $m_{ni}^*(x,z)$. Then

$$\begin{aligned} \left| \int |\mathbf{E}(m_n^*(x,z)) - m_n^*(x,z)|\ \nu(d(x,z)) - \int |\mathbf{E}(m_n^*(x,z)) - m_{ni}^*(x,z)|\ \nu(d(x,z)) \right| \\ \leq \int |m_n^*(x,z) - m_{ni}^*(x,z)|\ \nu(d(x,z)) \end{aligned}$$

But $|m_n^*(x,z) - m_{ni}^*(x,z)|$ is bounded by $2M/k$ and can differ from zero only if $(x_i,z_i) \in S_{(x,z),(r_n,b_n)}$ or $(\hat{x}_i,\hat{z}_i) \in S_{(x,z),(r_n,b_n)}$. Observe that $(x_i,z_i) \in S_{(x,z),(r_n,b_n)}$ if and only if $\nu(S_{(x,z),(\|x-x'\|,|z-z'|)}) \leq$

k/n. But the measure of such (x,z) pairs is bounded by $\gamma_d k/n$ by Lemma 2, therefore

$$\sup_{x_1,y_1,z_1,\ldots,x_n,y_n,z_n,\hat{x}_i,\hat{z}_i,\hat{y}_i} \int |m_n^*(x,z) - m_{ni}^*(x,z)| \ \nu(d(x,z)) \leq \frac{2M}{k}\frac{\gamma_d k}{n} = \frac{2M\gamma_d}{n}$$

and by Lemma 2

$$\mathbf{P}\left\{\left|\int |\mathbf{E}m_n^*(x,z) - m_n^*(x,z)| \ \nu(d(x,z)) - \mathbf{E}\int |\mathbf{E}m_n^*(x,z) - m_n^*(x,z)| \ \nu(d(x,z))\right| > \epsilon\right\} \\ \leq 2e^{-n\epsilon^2/(2M^2\gamma_d^2)}. \qquad (5)$$

So we have to show that

$$\mathbf{E}\int |\mathbf{E}m_n^*(x,z) - m_n^*(x,z)| \ \nu(d(x,z)) \to 0.$$

But using the Cauchy-Schwarz inequality we have

$$\begin{aligned}
&\mathbf{E}\int |\mathbf{E}m_n^*(x,z) - m_n^*(x,z)| \ \nu(d(x,z)) \\
&\leq \int \sqrt{\mathbf{E}|\mathbf{E}m_n^*(x,z) - m_n^*(x,z)|^2} \ \nu(d(x,z)) \\
&= \int \sqrt{\frac{1}{k^2} n Var(Y I_{\{(X,Z)\in S_{(x,z),(r_n,b_n)}\}})} \ \nu(d(x,z)) \\
&\leq \int \sqrt{\frac{M^2}{k^2} n \ \nu(S_{(x,z),(r_n,b_n)})} \ \nu(d(x,z)) \\
&= \int \sqrt{\frac{nM^2}{k^2}\frac{k}{n}} \mu(dx) \\
&= \sqrt{\frac{M^2}{k}} \to 0.
\end{aligned}$$

Finally denoting $R_n = \|X_{(k)} - x\|$, and $B_n = \|Z_{(k)} - z\|$, write the third term in (4) as

$$\begin{aligned}
&|m_n^*(x,z) - m_n(x,z)| \\
&= \frac{1}{k}\left|\sum_{j=1}^n Y_j I_{\{(X_j,Z_j)\in S_{(x,z),(r_n,b_n)}\}} - \sum_{j=1}^n Y_j I_{\{(X_j,Z_j)\in S_{(x,z),(R_n,B_n)}\}}\right| \\
&\leq \frac{M}{k}\sum_{j=1}^n \left|I_{\{(X_j,Z_j)\in S_{(x,z),(r_n,b_n)}\}} - I_{\{(X_j,Z_j)\in S_{(x,z),(R_n,B_n)}\}}\right| \\
&= M\left|\frac{1}{k}\sum_{j=1}^n I_{\{(X_j,Z_j)\in S_{(x,z),(r_n,b_n)}\}} - 1\right| = M|\hat{m}_n^*(x,z) - \mathbf{E}\hat{m}_n^*(x,z)|,
\end{aligned}$$

where $\hat{m}_n^*$ is defined as m_n^* with Y replaced by the constant random variable $\hat{Y} = 1$. Therefore the bound of (5) applies for the third term, too, and the proof is completed. $\square$

3 Strong universal consistency

In this section we demonstrate that the $k-$nearest neighbor regression estimate is consistent even if Y is not bounded, if k is chosen to satisfy $k/\log(n) \to \infty$ and $k/n \to 0$. More precisely, we prove the following theorem.

Theorem 2 *If*

$$\lim_{n\to\infty} k/\log(n) = \infty, \quad \lim_{n\to\infty} k/n = 0$$

then $J_n \to 0$ with probability one for all distributions of (X, Y) satisfying $\mathbf{E}|Y| < \infty$.

Györfi (1991) gave conditions for the strong universal consistency of a regression estimate. Translating his result to our case we get the following

Lemma 5 (Győrfi, 1991, Theorem 2) *Consider the $k-$nearest neighbor regression estimate $m_n(x, z)$. Then the L_1 error of the estimate J_n converges to zero almost surely for all distributions of (X, Y) satisfying $\mathbf{E}|Y| < \infty$ if the following two conditions are satisfied.*

(i) *$J_n \to 0$ almost surely for all distributions of (X, Y) with bounded Y.*

(ii) *There is a constant $c > 0$ such that for all distributions of (X, Y) satisfying $\mathbf{E}|Y| < \infty$*

$$\limsup_{n\to\infty} \frac{1}{k} \sum_{i=1}^{k} \int_0^1 \int |Y_{(i,x,z)}| \mu(dx) dz \leq c\mathbf{E}|Y| \quad \text{a.s..}$$

Clearly, condition (i) is satisfied by Theorem 1, so we only have to check (ii). In order to do so, we need some notation. Let A_i be the collection of all (x, z) that are such that (X_i, Z_i) is one of its k nearest neighbors. Here, we use some geometric arguments similar to those in the proof of Theorem 1. Similarly, let us define a cone $C(x, \theta, s)$, where x defines the top of the cone, s is a vector indicating a direction in $\mathcal{R}^d$, and $\theta \in (0, \pi)$ is an angle. The cone consists of all y with the property that either $y = x$ or angle$(y - x, s) \leq \theta$. For any fixed θ, there exists a finite collection S of directions such that

$$\cup_{s\in S} C(x, \theta, s) = \mathcal{R}^d$$

regardless of how x is picked. The cardinality of this set is denoted by $|S|$, and depends upon both θ and d. If $\theta < \pi/6$, and if $y, y' \in C(x, \theta, s)$, and $\|x - y\| < \|x - y'\|$, then $\|y - y'\| < \|x - y'\|$.

Furthermore, if $\|x-y\| \leq \|x-y'\|$, then $\|y-y'\| \leq \|x-y'\|$. We fix $\theta \in (0, \pi/6)$ and S as indicated above. In the space $\mathcal{R}^d \times [0,1]$, define the sets

$$C_{i,s} = C(X_i, \theta, s) \times [0,1] \ .$$

Let $B_{i,s}$ be the subset of $C_{i,s}$ consisting of all (x,z) that are among the k nearest neighbors of (X_i, Z_i) in the set

$$\{(X_1,Z_1), \ldots, (X_{i-1}, Z_{i-1}), (X_{i+1}, Z_{i+1}), \ldots, (X_n, Z_n), (x,z)\} \cap C_{i,s}$$

when distance tie-breaking is done in the described fashion. Equivalently, $B_{i,s}$ is the subset of $C_{i,s}$ consisting of all (x,z) that are closer to (X_i,Z_i) than the k-th nearest neighbor of (X_i,Z_i) in $C_{i,s}$, when distance tie-breaking is done by the fashion.

Lemma 6 *If* $(x,z) \in A_i$, *then* $(x,z) \in \cup_{s \in S} B_{i,s}$, *and thus*

$$\nu(A_i) \leq \sum_{s \in S} \nu(B_{i,s}) \ .$$

Proof To prove this claim, take $(x,z) \in A_i$. Then locate an $s \in S$ for which $(x,z) \in C_{i,s}$. We have to show that $(x,z) \in B_{i,s}$ to conclude the proof. Thus, we need to show that (x,z) is one of the k nearest neighbors of (X_i, Z_i) in the set

$$\{(X_1,Z_1), \ldots, (X_{i-1}, Z_{i-1}), (X_{i+1}, Z_{i+1}), \ldots, (X_n, Z_n), (x,z)\} \cap C_{i,s}$$

when distance tie-breaking is done appropriately.. Take (X_j, Z_j) closer to (X_i,Z_i) than (x,z) in $C_{i,s}$. If $\|X_j - X_i\| < \|x - X_i\|$, we recall that by the property of our cones that $\|x - X_j\| < \|x - X_i\|$, and thus (X_j, Z_j) is one of the $k-1$ nearest neighbors of (x,z) in $\mathcal{R}^d$. If on the other hand $\|X_j - X_i\| = \|x - X_i\|$, and z is further from Z_i than Z_j, then either $\|x - X_j\| < \|x - X_i\|$ (which shows again that (X_j,Z_j) is one of the $k-1$ nearest neighbors of (x,z) in $\mathcal{R}^d$) or $\|x - X_j\| = \|x - X_i\|$ and Z_i is further than Z_j from z. In the last case, (X_j, Z_j) is one of the $k-1$ nearest neighbors of (x,z) in $\mathcal{R}^d$. This shows that in $C_{i,s}$, there are at most $k-1$ points (X_j, Z_j) closer to (X_i, Z_i) than (x,z). Thus, with the same tie-breaking policy, (x,z) is one of the k nearest neighbors of (X_i, Z_i) in the set

$$\{(X_1,Z_1), \ldots, (X_{i-1}, Z_{i-1}), (X_{i+1}, Z_{i+1}), \ldots, (X_n, Z_n), (x,z)\} \cap C_{i,s} \ .$$

This concludes the proof of the claim. $\square$

Lemma 7 (An inequality for binomial random variables.) *Let B be a binomial random variable with parameters n and p. Then*

$$\mathbf{P}\{B > \epsilon\} \leq e^{\epsilon - np - \epsilon \log(\epsilon/np)} \ (\epsilon > np) \ ;$$

$$\mathbf{P}\{B < \epsilon\} \leq e^{\epsilon - np - \epsilon \log(\epsilon/np)} \ (\epsilon < np) \ .$$

Proof We proceed by Chernoff's exponential bounding method (Chernoff, 1952). In particular, for arbitrary $\lambda > 0$,

$$\mathbf{P}\{B > \epsilon\} \leq \mathbf{E}\{e^{\lambda B - \lambda\epsilon}\} = e^{-\lambda\epsilon}(e^{\lambda}p + 1 - p)^n \leq e^{-\lambda\epsilon + np(e^{\lambda}-1)} \ .$$

The right-hand-side is minimal for $\lambda = \log(\epsilon/np)$. Resubstitution of this value gives the first bound. The proof of the other bound is similar. □

Lemma 8 (Property of $\nu(B_{i,s})$.) *If $k/\log(n) \to \infty$ and $k/n \to 0$ then*

$$\limsup_{n\to\infty} \frac{n}{k} \max_i \nu(B_{i,s}) \leq 2 \quad \text{a.s.}.$$

Proof We prove that for every $s \in S$

$$\sum_n \mathbf{P}\left\{\frac{n}{k} \max_i \nu(B_{i,s}) > 2\right\} < \infty.$$

In order to do this we give a bound for

$$\mathbf{P}\{\nu(B_{i,s}) > \epsilon | X_i, Z_i\}.$$

If $\nu(C_{is}) \leq \epsilon$ then since $B_{is} \subseteq C_{is}$, we have $\mathbf{P}\{\nu(B_{i,s}) > \epsilon | X_i, Z_i\} = 0$, therefore we assume that $\nu(C_{is}) > \epsilon$. Fix X_i, Z_i. The distance ordering and tie-breaking method induces a total ordering of all (x, z) with respect to closeness to (X_i, Z_i). Find a pair $(x, z) \in C_{is}$ such that if B_ϵ is the collection of all $(x', z') \in C_{is}$ that are nearer to (X_i, Z_i) than (x, z), then $\nu(B_\epsilon) = \epsilon$. By our method of tie-breaking, such a pair (x, z) exists. We have the following dual relationship:

$$\mathbf{P}\{\nu(B_{is}) > \epsilon | X_i, Z_i\} = \mathbf{P}\{B_\epsilon \text{ captures } < k \text{ of the points } (X_j, Z_j), j \neq i | X_i, Z_i\} \ .$$

But if B is a binomial $(n-1, \epsilon)$ random variable, then the last probability is equal to

$$\mathbf{P}\{B < k\} \leq e^{k - (n-1)\epsilon - k\log(k/(n-1)\epsilon)} \quad \text{if} \quad k < (n-1)\epsilon \ .$$

Finally, with $\epsilon = 2k/n$ we have $k < (n-1)\epsilon$, therefore

$$\begin{aligned}
\mathbf{P}\left\{\max_{1\le i\le n} \nu(B_{is}) > \epsilon\right\} &\le n\mathbf{P}\{\nu(B_{1s}) > \epsilon\} \\
&\le ne^{k-(n-1)\epsilon - k\log(k/(n-1)\epsilon)} \\
&\le ne^{k-2k(n-1)/n - k\log(n/2(n-1))} \\
&\le ne^{-k+2k/n+k\log 2}\ ,
\end{aligned}$$

which is summable in n when $k \ge \frac{2}{1-\log 2}\log n$. $\square$

Proof of Theorem 2 By Lemma 5 and Theorem 1 it is enough to prove that there is a constant $c > 0$

$$\limsup_{n\to\infty} \frac{1}{k}\sum_{i=1}^{k}\int_0^1\int |Y_{(i,x,z)}|\mu(dx)dz \le c\mathbf{E}|Y| \quad \text{a.s..}$$

Observe that

$$\frac{1}{k}\sum_{i=1}^{k}\int_0^1\int |Y_{(i,x,z)}|\mu(dx)dz = \frac{1}{k}\sum_{i=1}^{n}|Y_{(i)}|\nu(A_i) \le \frac{1}{n}\sum_{i=1}^{n}|Y_{(i)}|\left(\frac{n}{k}\max_i \nu(A_i)\right).$$

If we can show that

$$\limsup_{n\to\infty}\frac{n}{k}\max_i \nu(A_i) \le c \quad \text{a.s.} \tag{6}$$

for some constant $c > 0$, then by the law of large numbers

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}|Y_{(i)}|\left(\frac{n}{k}\max_i \nu(A_i)\right) \le \limsup_{n\to\infty} c\frac{1}{n}\sum_{i=1}^{n}|Y_{(i)}| = c\mathbf{E}|Y|,$$

so we have to prove (6). But by Lemma 6,

$$\nu(A_i) \le \sum_{s\in S}\nu(B_{i,s}),$$

therefore, Lemma 8 implies that (6) is satisfied with $c = 2|S|$, so the proof of the theorem is completed. $\square$

Acknowledgment.

We would like to thank both referees.

References

[1] K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Mathematical Journal*, vol. 37, pp. 357–367, 1967.

[2] J. Beck, "The exponential rate of convergence of error for k_n NN nonparametric regression and decision," *Problems of Control and Information Theory*, vol. 8, pp. 303–311, 1979.

[3] P. K. Bhattacharya and Y. P. Mack, "Weak convergence of k–NN density and regression estimators with varying k and applications," *Annals of Statistics*, vol. 15, pp. 976–994, 1987.

[4] H. Chernoff, "A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations," *Annals of Mathematical Statistics*, vol. 23, pp. 493–507, 1952.

[5] G. Collomb, "Estimation de la regression par la méthode des k points les plus proches: propriétés de convergence ponctuelle," *Comptes Rendus de l'Académie des Sciences de Paris*, vol. 289, pp. 245–247, 1979.

[6] G. Collomb, "Estimation de la regression par la méthode des k points les plus proches avec noyau," pp. 159–175, Lecture Notes in Mathematics # 821, Springer-Verlag, Berlin, 1980.

[7] G. Collomb, "Estimation non paramétrique de la regression: revue bibliographique," *International Statistical Review*, vol. 49, pp. 75–93, 1981.

[8] G. Collomb, "Nonparametric regression: an up-to-date bibliography," *Statistics*, vol. 16, pp. 300–324, 1985.

[9] T. M. Cover, "Estimation by the nearest neighbor rule," *IEEE Transactions on Information Theory*, vol. IT-14, pp. 50–55, 1968.

[10] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. IT-13, pp. 21–27, 1967.

[11] L. Devroye, "A universal k-nearest neighbor procedure in discrimination," in: *Proceedings of the 1978 IEEE Computer Society Conference on Pattern Recognition and Image Processing*, pp. 142–147, 1978.

[12] L. Devroye, "On the almost everywhere convergence of nonparametric regression function estimates," *Annals of Statistics*, vol. 9, pp. 1310–1319, 1981.

[13] L. Devroye, "Necessary and sufficient conditions for the almost everywhere convergence of nearest neighbor regression function estimates," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 61, pp. 467–481, 1982.

[14] L. Devroye, "The equivalence of weak, strong and complete convergence in L1 for kernel density estimates," *Annals of Statistics*, vol. 11, pp. 896–904, 1983.

[15] L. Devroye, "The kernel estimate is relatively stable," *Probability Theory and Related Fields*, vol. 77, pp. 521–536, 1988.

[16] L. Devroye, "Exponential inequalities in nonparametric estimation," in: *Nonparametric Functional Estimation*, ed. G. Roussas, pp. 31–44, NATO ASI Series, Springer-Verlag, Berlin, 1991.

[17] L. Devroye and L. Györfi, "Distribution-free exponential bound on the L_1 error of partitioning estimates of a regression function," in: *Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics*, ed. F. Konecny, J. Mogyorodi, W. Wertz, pp. 67–76, Akademiai Kiado, Budapest, Hungary, 1983.

[18] L. Devroye and L. Györfi, *Nonparametric Density Estimation: The L_1 View*, John Wiley, New York, 1985.

[19] L. Devroye and A. Krzyżak, "An equivalence theorem for L1 convergence of the kernel regression estimate," *Journal of Statistical Planning and Inference*, vol. 23, pp. 71–82, 1989.

[20] L. Devroye and T. J. Wagner, "Distribution-free consistency results in nonparametric discrimination and regression function estimation," *Annals of Statistics*, vol. 8, pp. 231–239, 1980.

[21] N. Glick, "Consistency conditions for probability estimators and integrals of density estimators," *Utilitas Mathematica*, vol. 6, pp. 61–74, 1974.

[22] W. Greblicki, A. Krzyżak, and M. Pawlak, "Distribution-free pointwise consistency of kernel regression estimate," *Annals of Statistics*, vol. 12, pp. 1570–1575, 1984.

[23] L. Györfi, "Recent results on nonparametric regression estimate and multiple classification," *Problems of Control and Information Theory*, vol. 10, pp. 43–52, 1981.

[24] L. Györfi, "Universal consistencies of a regression estimate for unbounded regression functions," in: *Nonparametric Functional Estimation*, ed. G. Roussas, pp. 329–338, NATO ASI Series, Springer-Verlag, Berlin, 1991.

[25] P. E. Hart, "The condensed nearest neighbor rule," *IEEE Transactions on Information Theory*, vol. IT-14, pp. 515–516, 1968.

[26] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, pp. 13–30, 1963.

[27] A. Krzyżak, "The rates of convergence of kernel regression estimates and classification rules," *IEEE Transactions on Information Theory*, vol. IT-32, pp. 668–679, 1986.

[28] A. Krzyżak and M. Pawlak, "Distribution-free consistency of a nonparametric kernel regression estimate and classification," *IEEE Transactions on Information Theory*, vol. IT-30, pp. 78–81, 1984.

[29] Y. P. Mack, "Local properties of k–nearest neighbor regression estimates," *SIAM Journal on Algebraic and Discrete Methods*, vol. 2, pp. 311–323, 1981.

[30] C. McDiarmid, "On the method of bounded differences," in: *Surveys in Combinatorics 1989*, vol. 141, pp. 148–188, London Mathematical Society Lecture Notes Series, Cambridge University Press, Cambridge, 1989.

[31] E. A. Nadaraya, "On estimating regression," *Theory of Probability and its Applications*, vol. 9, pp. 141–142, 1964.

[32] E. A. Nadaraya, "Remarks on nonparametric estimates for density functions and regression curves," *Theory of Probability and its Applications*, vol. 15, pp. 134–137, 1970.

[33] R. M. Royall, "A Class of Nonparametric Estimators of a Smooth Regression Function," Ph.D. Dissertation, Stanford University, Stanford, CA, 1966.

[34] C. Spiegelman and J. Sacks, "Consistent window estimation in nonparametric regression," *Annals of Statistics*, vol. 8, pp. 240–246, 1980.

[35] C. J. Stone, "Consistent nonparametric regression," *Annals of Statistics*, vol. 8, pp. 1348–1360, 1977.

[36] W. F. Stout, *Almost Sure Convergence*, Academic Press, New York, 1974.

[37] W. Stute, "Asymptotic normality of nearest neighbor regression function estimates," *Annals of Statistics*, vol. 12, pp. 917–926, 1984.

[38] G. S. Watson, "Smooth regression analysis," *Sankhya Series A*, vol. 26, pp. 359–372, 1964.

[39] R. L. Wheeden and A. Zygmund, *Measure and Integral*, Marcel Dekker, New York, 1977.

[40] L.C.Zhao, "Exponential bounds of mean error for the nearest neighbor estimates of regression functions," *J. Multivariate Analysis*, vol.21, pp.168–178, 1987