Estimation of the density of regression residual

László Györfi
Department of Computer Science and Information Theory
Budapest University of Technology and Economics
1521 Stoczek u. 2, Budapest, Hungary
gyorfi@cs.bme.hu

February 10, 2012

Abstract

Consider the regression problem with a response variable Y and with a feature vector X. For the regression function $m(x) = \mathbb{E}\{Y|X = x\}$, this paper investigates methods for estimating the density of the residual $Y - m(X)$ from i.i.d. data. We prove the strong universal (density-free) L_1-consistency of a recursive and a nonrecursive density estimate based on a regression estimate, and bound the rate of convergence of the nonrecursive estimate.

1 Introduction

Let Y be a real valued random variable and let $X = (X^{(1)}, \dots, X^{(d)})$ be a d-dimensional random vector. The coordinates of X may have different types of distributions, some of them may be discrete (for example binary), others may be absolutely continuous. In the sequel we do not assume anything about the distribution of X. The task of regression analysis is to estimate Y given X, i.e., one aims to find a function F defined on the range of X such

AMS 2000 subject classifications. Primary: 62G08 secondary: 62G09.
Key words and phrases: regression residual, nonparametric density estimation, nonparametric regression estimation.

that $F(X)$ is "close" to Y. Typically, closeness is measured in terms of the *mean squared error* of F,

$$\mathbb{E}\{(F(X)-Y)^2\}.$$

It is well-known that the mean squared error is minimized by the regression function m with

$$m(x)=\mathbb{E}\{Y \mid X=x\} \tag{1}$$

and a minimum mean squared error is

$$L^* := \mathbb{E}\{(Y-m(X))^2\} = \min_F \mathbb{E}\{(Y-F(X))^2\},$$

since, for each measurable function F, the mean squared error can be decomposed into

$$\begin{aligned}\mathbb{E}\{(F(X)-Y)^2\} &= \mathbb{E}\{(m(X)-Y)^2\}+\mathbb{E}\{(m(X)-F(X))^2\}\\ &= \mathbb{E}\{(m(X)-Y)^2\}+\int_{\mathbb{R}^d}(m(x)-F(x))^2\mu(dx),\end{aligned}$$

where μ denotes the distribution of X. The second term on the right hand side is called *excess error* or integrated squared error of the function F. Clearly, the mean squared error of F is close to its minimum if and only if the excess error $\int_{\mathbb{R}^d}(m(x)-F(x))^2\mu(dx)$ is close to zero.

The regression function cannot be calculated as long as the distribution of (X,Y) is unknown. Assume, however, that we observed data

$$D_n=\{(X_1,Y_1),\dots,(X_n,Y_n)\} \tag{2}$$

consisting of independent and identically distributed copies of (X,Y). D_n can be used to produce an estimate $m_n=m_n(\cdot,D_n)$ of the regression function m. Since m arises from L_2 considerations, it is natural to study $L_2(\mu)$ convergence of the regression estimate m_n to m. In particular, the estimator m_n is called *strongly universally consistent* if its excess error satisfies

$$\int_{\mathbb{R}^d}(m(x)-m_n(x))^2\mu(dx)\to 0 \text{ a.s.}$$

for all distributions of (X,Y) with $\mathbb{E}|Y|^2<\infty$.

It is of great importance to be able to estimate the various characteristics of the residual

$$Y - m(X).$$

For nonparametric estimates of the minimum mean squared error

$$L^* = \mathbb{E}\{(Y - m(X))^2\}$$

see, e.g., Dudoit and van der Laan [15], Kohler [22], Liitiäinen, Corona, and Lendasse [23], [24], Liitiäinen, Verleysen, Corona and Lendasse [25], Müller and Stadtmüller [27], Neumann [29], Pelckmans, De Brabanter, Suykens and De Moor [31], Stadtmüller and Tsybakov [33] and the literature cited there. Devroye, Györfi, Schäfer and Walk [12] proved that without any tail and smoothness condition L^* cannot be estimated with guaranteed rate of convergence, and showed a first nearest neighbor based estimate, which for Lipschitz continuous m has faster rate of convergence than that of the usual plug-in estimators. Müller, Schick and Wefelmeyer [28] estimate L^* as the variance of an independent measurement error Z in the model

$$Y = m(X) + Z \tag{3}$$

such that $\mathbb{E}\{Z\} = 0$, and X and Z are independent. Sometimes it is called additive noise model.

2 A recursive estimate

In this paper we deal with the problem how to estimate the density f of the residual

$$Y - m(X)$$

assuming that the density f exists. Our aim is to estimate f from i.i.d. data (2).

Under some smoothness conditions on the density f, Ahmad [1], Cheng [5], [4], Efromovich [16], [17], Akritas and Van Keilegom [2], Neumeyer and Van Keilgom [30] studied the estimate the density of the residual. Under the additive noise model (3), Devroye, Felber, Kohler and Krzyzak [8] introduced a density estimate of the residual, and proved its universal (density free) strong consistency in L_1.

In this paper we extend this result such that don't assume the additive noise model (3). We only assume that, for given $X = x$, the conditional density of the residual $Y - m(X)$ exists. This conditional density is denoted by $f(z \mid x)$. Then

$$f(z) = \int_{\mathbb{R}^d} f(z \mid x)\mu(dx).$$

Suppose that based on the data $(X_1, Y_1), \ldots, (X_n, Y_n)$, we are given a strongly universally consistent regression estimate m_n. We introduce a recursive density estimate of the residual, which is a slight modification of the recursive kernel density estimate due to Wolverton and Wagner [38] and Yamoto [36]. Let K be a density on $\mathbb{R}$, called kernel, $\{h_i\}$ is the bandwidth sequence. For a bandwidth $h > 0$, introduce the notation

$$K_h(z) = \frac{1}{h} K(z/h).$$

Then the recursive estimate is defined by

$$f_n(z) := \frac{1}{n} \sum_{i=1}^{n} K_{h_i}(z - Z_i), \tag{4}$$

where in the i-th term we plug-in the approximation of the i-th residual

$$Z_i := Y_i - m_{i-1}(X_i).$$

Theorem 1 *Assume that Y is square integrable. Suppose that we are given a strongly universally consistent regression estimate m_n, i.e.,*

$$\int_{\mathbb{R}^d} (m(x) - m_n(x))^2 \mu(dx) \to 0 \ a.s.$$

and for given $X = x$, the conditional density of the residual $Y - m(X)$ exists. Assume that the kernel function K is a square integrable density, and

$$\lim_{n\to\infty} h_n = 0 \ and \ \sum_{n=1}^{\infty} \frac{1}{n^2 h_n} < \infty. \tag{5}$$

Then

$$\lim_{n\to\infty} \int_{\mathbb{R}} |f_n(z) - f(z)| dz = 0$$

a.s.

PROOF For given $X = x$ and for given $(X_1, Y_1), \ldots, (X_n, Y_n)$, the approximate residual

$$Y - m_n(X) = Y - m(X) + m(X) - m_n(X)$$

has the conditional density $f(z + m_n(x) - m(x) \mid x)$ and so the density $g_n(z)$ of $Y - m_n(X)$ can be calculated as follows:

$$g_n(z) = \int_{\mathbb{R}^d} f(z + m_n(x) - m(x) \mid x)\mu(dx).$$

Next we show that

$$\lim_{n\to\infty} \int_{\mathbb{R}} |g_n(z) - f(z)|dz = 0 \tag{6}$$

a.s. For $\delta > 0$, introduce the notation

$$\Delta_x(\delta) := \sup_{|u|\le\delta} \int_{\mathbb{R}} |f(z + u \mid x) - f(z \mid x)|dz.$$

Thus,

$$\begin{aligned}
& \int_{\mathbb{R}} |g_n(z) - f(z)|dz \\
= & \int_{\mathbb{R}} |\int_{\mathbb{R}^d} f(z + m_n(x) - m(x) \mid x)\mu(dx) - \int_{\mathbb{R}^d} f(z \mid x)\mu(dx)|dz \\
\le & \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}} |f(z + m_n(x) - m(x) \mid x) - f(z \mid x)|dz \right) \mu(dx) \\
= & \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}} |f(z + m_n(x) - m(x) \mid x) - f(z \mid x)|dz \right) I_{\{|m_n(x)-m(x)|\le\delta\}}\mu(dx) \\
& + \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}} |f(z + m_n(x) - m(x) \mid x) - f(z \mid x)|dz \right) I_{\{|m_n(x)-m(x)|>\delta\}}\mu(dx) \\
\le & \int_{\mathbb{R}^d} \Delta_x(\delta)\mu(dx) + 2\mathbb{P}\{|m(X) - m_n(X)| > \delta \mid (X_1, Y_1), \ldots, (X_n, Y_n)\} \\
= & \int_{\mathbb{R}^d} \Delta_x(\delta)\mu(dx) + 2\frac{\int_{\mathbb{R}^d}(m(x) - m_n(x))^2\mu(dx)}{\delta^2} \\
\to & \int_{\mathbb{R}^d} \Delta_x(\delta)\mu(dx)
\end{aligned}$$

a.s. as $n \to \infty$. $\Delta_x(\delta) \le 2$ and for any fixed x, $\Delta_x(\delta) \to 0$ as $\delta \to 0$, therefore the dominated convergence theorem implies that

$$\int_{\mathbb{R}^d} \Delta_x(\delta)\mu(dx) \to 0$$

as $\delta \to 0$. Apply the decomposition

$$f_n(z) - f(z) = V_n(z) + B_n(z),$$

where the variation term is

$$V_n(z) = \frac{1}{n}\sum_{i=1}^{n} \left[K_{h_i}(z - Z_i) - \mathbb{E}\left\{K_{h_i}(z - Z_i) \mid (X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1})\right\}\right],$$

while the (conditional) bias term is

$$B_n(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left\{K_{h_i}(z - Z_i) \mid (X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1})\right\} - f(z).$$

Concerning the bias term, $\lim_{n\to\infty} h_n = 0$ and (6) imply that

$$\begin{aligned}
\int_{\mathbb{R}} |B_n(z)|dz &= \int_{\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{R}} K_{h_i}(z-u) g_{i-1}(u) du - f(z) \right| dz \\
&\le \int_{\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{R}} K_{h_i}(z-u) f(u) du - f(z) \right| dz \\
&\quad + \int_{\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{R}} K_{h_i}(z-u) |g_{i-1}(u) - f(u)| du dz \\
&\le \int_{\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{R}} K_{h_i}(z-u) f(u) du - f(z) \right| dz \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{R}} |g_{i-1}(u) - f(u)| du \\
&\to 0
\end{aligned}$$

a.s., because of Toeplitz Lemma and Theorem 2.4 in Devroye, Györfi [10]. $V_n(\cdot)$ is an average of L_2-valued sequence of martingale differences. We apply the generalized Chow theorem [6]: let U_n, $n = 1, 2, \ldots$ be an L_2-valued sequence of martingale differences such that

$$\sum_{n=1}^{\infty} \frac{\mathbb{E}\{\|U_n\|_2^2\}}{n^2}, < \infty$$

where $\|\cdot\|_2$ denotes the L_2 norm. Then

$$\lim_{n\to\infty} \left\| \frac{1}{n} \sum_{i=1}^{n} U_i \right\|_2 = 0$$

a.s. (cf. Györfi, Györfi, Vajda [19]). One has to verify the condition of the generalized Chow theorem:

$$\begin{aligned}
& \sum_{n=1}^{\infty} \frac{\mathbb{E}\left\{\|K_{h_i}(\cdot - Z_i) - \mathbb{E}\left\{K_{h_i}(\cdot - Z_i) \mid (X_1, Y_1), \ldots, (X_{n-1}, Y_{n-1})\right\}\|_2^2\right\}}{n^2} \\
\leq & \sum_{n=1}^{\infty} \frac{\mathbb{E}\left\{\|K_{h_i}(\cdot - Z_i)\|_2^2\right\}}{n^2} \\
\leq & \sum_{n=1}^{\infty} \frac{\|K\|_2^2}{n^2 h_n} \\
< & \infty,
\end{aligned}$$

by the condition of the theorem, and so

$$\|V_n\|_2 \to 0$$

a.s. Put

$$\hat{f}_n(z) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left\{K_{h_i}(z - Z_i) \mid (X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1})\right\}.$$

then we proved that

$$\|\hat{f}_n - f\|_1 = \|B_n\|_1 \to 0$$

a.s., where $\|\cdot\|_1$ denotes the L_1 norm, and

$$\|\hat{f}_n - f_n\|_2 = \|V_n\|_2 \to 0$$

a.s. From Lemma 3.1 in Györfi, Masry [20] we get that these two limit relations imply

$$\|f_n - f\|_1 \to 0$$

a.s.

□

3 A non-recursive estimate

Next we introduce a data splitting scheme. Assume that we are given two independent samples:

$$D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$$

and

$$D'_n = \{(X'_1, Y'_1), \dots, (X'_n, Y'_n)\}.$$

From sample D_n we generate a strongly universally consistent regression estimate m_n. Then the non-recursive estimate is defined by

$$f_n(z) := \frac{1}{n} \sum_{i=1}^{n} K_{h_n}(z - Z_i), \tag{7}$$

where in the i-th term we plug-in the approximation of the i-th residual

$$Z_i := Y'_i - m_n(X'_i).$$

Given D_n, the common density of Z_i's is g_n.

Under the additive noise model (3), Devroye, Felber, Kohler and Krzyzak [8] proved its universal strong consistency in L_1.

Theorem 2 *Suppose that we are given a strongly universally consistent regression estimate m_n, i.e.,*

$$\int_{\mathbb{R}^d} (m(x) - m_n(x))^2 \mu(dx) \to 0 \ \textit{a.s.}$$

and for given $X = x$, the conditional density of the residual $Y - m(X)$ exists. Assume that the kernel function K is a square integrable density, and

$$\lim_{n\to\infty} h_n = 0 \ \textit{and} \ \lim_{n\to\infty} n h_n = \infty. \tag{8}$$

Then

$$\lim_{n\to\infty} \int_{\mathbb{R}} |f_n(z) - f(z)| dz = 0$$

a.s.

Proof. Applying the argument is Devroye [7], we get that

$$\mathbb{P}\left\{\Big|\int_{\mathbb{R}}|f_n-f|-\mathbb{E}\left\{\int_{\mathbb{R}}|f_n-f|\mid D_n\right\}\Big|\geq\epsilon\mid D_n\right\}\leq 2e^{-n\epsilon^2/2},$$

therefore one has to prove that

$$\mathbb{E}\left\{\int_{\mathbb{R}}|f_n-f|\mid D_n\right\}\to 0$$

a.s. Concerning the conditional bias term, we have that

$$\begin{aligned}
&\int_{\mathbb{R}}|\mathbb{E}\{f_n(z)\mid D_n\}-f(z)|dz\\
=&\int_{\mathbb{R}}\Big|\int_{\mathbb{R}}K_{h_n}(z-u)g_n(u)du-f(z)\Big|dz\\
\leq&\int_{\mathbb{R}}\Big|\int_{\mathbb{R}}K_{h_n}(z-u)f(u)du-f(z)\Big|dz+\int_{\mathbb{R}}\int_{\mathbb{R}}K_{h_n}(z-u)|g_n(u)-f(u)|dudz\\
\leq&\int_{\mathbb{R}}\Big|\int_{\mathbb{R}}K_{h_n}(z-u)f(u)du-f(z)\Big|dz+\int_{\mathbb{R}}|g_n(u)-f(u)|du\\
\to&\ 0
\end{aligned}$$

a.s. For the conditional variation term, let I be an arbitrary interval, then we have that

$$\begin{aligned}
&\mathbb{E}\left\{\int_{\mathbb{R}}|\mathbb{E}\{f_n(z)\mid D_n\}-f_n(z)|dz\mid D_n\right\}\\
\leq&\int_I\mathbb{E}\left\{|\mathbb{E}\{f_n(z)\mid D_n\}-f_n(z)|\mid D_n\right\}dz+2\int_{I^c}\mathbb{E}\{f_n(z)\mid D_n\}dz\\
\leq&\int_I\sqrt{\mathbb{E}\left\{|\mathbb{E}\{f_n(z)\mid D_n\}-f_n(z)|^2\mid D_n\right\}}dz\\
&+2\int_{\mathbb{R}}|\mathbb{E}\{f_n(z)\mid D_n\}-f(z)|dz+2\int_{I^c}f(z)dz.
\end{aligned}$$

For $\epsilon>0$, choose I and n such that

$$2\int_{\mathbb{R}}|\mathbb{E}\{f_n(z)\mid D_n\}-f(z)|dz+2\int_{I^c}f(z)dz<\epsilon.$$

Thus,

$$
\begin{aligned}
& \mathbb{E}\left\{ \int_{\mathbb{R}} |\mathbb{E}\{f_n(z) \mid D_n\} - f_n(z)| dz \mid D_n \right\} \\
\leq & \int_I \sqrt{\frac{\mathbb{E}\left\{ |\mathbb{E}\{K_{h_n}(z - Z_1) \mid D_n\} - K_{h_n}(z - Z_1)|^2 \mid D_n \right\}}{n}} dz + \epsilon \\
\leq & \int_I \sqrt{\frac{\mathbb{E}\left\{ K_{h_n}(z - Z_1)^2 \mid D_n \right\}}{n}} dz + \epsilon \\
\leq & \sqrt{\frac{\|K\|_2^2 |I|}{n h_n}} + \epsilon \\
\to & \ \epsilon
\end{aligned}
$$

a.s., where $|I|$ denotes the length of the interval I. □

Remark 2. Using a tricky counter example, Devroye, Felber, Kohler and Krzyzak [8] showed that the condition of the existence of conditional densities of the residual cannot be weakened, if for the regression estimate merely strong universal consistency is assumed. The example is follows: Choose X uniformly distributed on $[0, 1]$, let U be independent of X take on values 1 and -1 with probability $1/2$, resp., and set $Y = U \cdot X$. Then Y is uniformly distributed on $[-1, 1]$ and has a density, the regression function is 0. However, $Y = Y - m(X)$ is conditioned on the value of $X = x$ concentrated on $-x$ and x and has no density. Then they constructed an approximation m_n of the regression function such that $\max_x |m_n(x)| \leq \sqrt{h_n} \to 0$ and

$$
\liminf_n \int_{\mathbb{R}} |f_n(z) - f(z)| dz \geq 1
$$

a.s., where the kernel K is the window kernel.

Remark 3. Instead of considering the density of the residual $Y - m(X)$, one may want to estimate the density of $m(X)$ from data. Luc Devroye noticed that in this scheme it is not enough to have universally consistent regression estimate, as the following example shows. Put

$$
m_n(X_i) := j h_n \text{ if } m(X_i) \in [(j - 1/2)h_n, (j + 1/2)h_n),
$$

then

$$
|m(X_i) - m_n(X_i)| \leq h_n/2 \to 0.
$$

Introduce the kernel density estimate

$$f_n(z) := \frac{1}{n}\sum_{i=1}^{n} K_{h_n}(z - m_n(X_i))$$

and assume that the support of the kernel K is contained in $[-1/4, 1/4)$. Then the support of f_n is contained in

$$A_n := \cup_{j=\infty}^{\infty}[(j-1/4)h_n, (j+1/4)h_n),$$

therefore

$$\int_{\mathbb{R}} |f_n(z) - f(z)|dz \geq \int_{A_n^c} |f_n(z) - f(z)|dz = \int_{A_n^c} f(z)dz \approx 1/2$$

if h_n is small enough.

4 The rate of convergence of the nonrecursive estimate

An important problem is to bound the rate of convergence of

$$\mathbb{E}\left\{\int_{\mathbb{R}} |f_n(z) - f(z)|dz\right\},$$

where f_n is the nonrecursive estimate. The main question is the size of degradation with respect to the case when $Y_i - m(X_i)$ is available, i.e., what is the influence of the regression estimate in the rate of convergence of the density estimate.

Theorem 3 *Under the model of additive noise (3), assume that the density f is twice differentiable and has a compact support contained in the interval I. Moreover, suppose that the kernel K is symmetric ($K(x) = K(-x)$), bounded and has compact support. Then*

$$\begin{aligned}\mathbb{E}\left\{\int_{\mathbb{R}} |f_n(z) - f(z)|dz\right\} \;\leq\; & c_1 h_n^2 + \frac{c_2}{\sqrt{nh_n}} \\ & + c_3\mathbb{E}\left\{\left|\int_{\mathbb{R}^d} m_n(x)\mu(dx) - \mathbb{E}\{Y\}\right|\right\} \\ & + c_4\mathbb{E}\left\{\int_{\mathbb{R}^d} (m_n(x) - m(x))^2\mu(dx)\right\}.\end{aligned}$$

PROOF. The proof of Theorem 2 implies that

$$\begin{aligned}
& \mathbb{E}\left\{\int_{\mathbb{R}} |f_n(z) - f(z)| dz\right\} \\
\leq\ & \mathbb{E}\left\{\int_{\mathbb{R}} |\mathbb{E}\{f_n(z) \mid D_n\} - f(z)| dz\right\} + \mathbb{E}\left\{\int_{\mathbb{R}} |\mathbb{E}\{f_n(z) \mid D_n\} - f_n(z)| dz\right\} \\
\leq\ & \int_{\mathbb{R}} \left| \int_{\mathbb{R}} K_{h_n}(z-u) f(u) du - f(z) \right| dz + \mathbb{E}\left\{\int_{\mathbb{R}} |g_n(z) - f(z)| dz\right\} \\
& + \frac{\|K\|_2 \sqrt{|I|}}{\sqrt{n h_n}} \\
\leq\ & c_1 h_n^2 + \frac{c_2}{\sqrt{n h_n}} + \mathbb{E}\left\{\int_{\mathbb{R}} |g_n(u) - f(u)| du\right\},
\end{aligned}$$

where we applied Lemma 5.4 in Devroye, Györfi [10]. The sum of the first and the second term in the right hand side is the same as that of the rate of convergence of the standard kernel estimate (cf. Theorem 5.1 in Devroye, Györfi [10]), so the excess error can be bounded by $\mathbb{E}\left\{\int_{\mathbb{R}} |g_n(z) - f(z)| dz\right\}$. In the special case (3) of additive noise we have that

$$f(z \mid x) = f(z).$$

For twice differentiable density f, let's calculate the second order Taylor expansion of $f(z + m_n(x) - m(x))$ at z:

$$f(z+m_n(x)-m(x)) = f(z)+f'(z)(m_n(x)-m(x))+\frac{f''(z_{n,x})}{2}(m_n(x)-m(x))^2$$

with some $z_{n,x}$. Then

$$\begin{aligned}
& \int_{\mathbb{R}} |g_n(z) - f(z)| dz \\
=\ & \int_{\mathbb{R}} \left| \int_{\mathbb{R}^d} f(z + m_n(x) - m(x)) \mu(dx) - f(z) \right| dz \\
=\ & \int_{\mathbb{R}} \left| \int_{\mathbb{R}^d} (f'(z)(m_n(x) - m(x)) + \frac{f''(z_{n,x})}{2}(m_n(x) - m(x))^2) \mu(dx) \right| dz \\
\leq\ & |I| \max_z |f'(z)| \left| \int_{\mathbb{R}^d} (m_n(x) - m(x)) \mu(dx) \right| \\
& + |I| \max_z |f''(z)| \int_{\mathbb{R}^d} (m_n(x) - m(x))^2 \mu(dx) \\
=\ & c_3 \left| \int_{\mathbb{R}^d} m_n(x) \mu(dx) - \mathbb{E}\{m(X)\} \right| + c_4 \int_{\mathbb{R}^d} (m_n(x) - m(x))^2 \mu(dx).
\end{aligned}$$

□

Remark 4. If $h_n = c_5 n^{-1/5}$ then

$$c_1 h_n^2 + \frac{c_2}{\sqrt{n h_n}} = c_6 n^{-2/5}.$$

If the regression function m is Lipschitz continuous and X is bounded then the partitioning, the kernel and the nearest neighbor regression estimates have rate of convergence

$$\mathbb{E}\left\{\int_{\mathbb{R}^d} (m_n(x) - m(x))^2 \mu(dx)\right\} \le c_7 n^{-2/(d+2)}, \tag{9}$$

(cf. Chapters 4, 5, 6 in Györfi et al [21]). Next we show that under some situations,

$$\mathbb{E}\left\{\left|\int_{\mathbb{R}^d} m_n(x)\mu(dx) - \mathbb{E}\{m(X)\}\right|\right\} \le c_8 n^{-2/(d+2)}, \tag{10}$$

which would imply that

$$\mathbb{E}\left\{\int_{\mathbb{R}} |f_n(z) - f(z)| dz\right\} \le c_6 n^{-2/5} + c_7 n^{-2/(d+2)},$$

and so for $d \le 3$ the rate of convergence is the same as that of standard kernel estimate.

Stone [34] first pointed out that there exist universally consistent estimators. He considered local averaging estimates, i.e., estimates of the form

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

where $W_{ni}(x)$ are the data-dependent weights governing the local averaging about x.

The *partitioning estimate* is defined by a partition $\mathcal{P}_n = \{A_{n,1}, A_{n,2} \dots\}$ of $\mathbb{R}^d$ and

$$m_n(x) = \frac{\sum_{i=1}^n Y_i I_{\{X_i \in A_n(x)\}}}{\sum_{i=1}^n I_{\{X_i \in A_n(x)\}}},$$

where $A_n(x)$ denotes the cell $A_{n,j}$ into which x falls, and $0/0 = 0$, by definition. Results on universal consistency can be found in Devroye and Györfi [9] and Györfi [18].

Corollary 1 *For the non-recursive estimate f_n, choose $h_n = c_5 n^{-1/5}$. Let the regression estimate m_n be the partitioning estimate. In addition to the conditions of Theorem 3, assume that the partition is cubic with side length*

$$h'_n = c_{13} n^{-1/(d+2)},$$

Y and X are bounded, and m satisfies the Lipschitz condition:

$$|m(x) - m(z) \leq C\|x - z\|. \tag{11}$$

Then

$$\mathbb{E}\left\{\int_{\mathbb{R}} |f_n(z) - f(z)|dz\right\} \leq c_6 n^{-2/5} + c_7 n^{-2/(d+2)}.$$

Proof. Theorem 4.3 in Györfi et al [21] implies (9), so because of Theorem 3 and Remark 4, we have to show (10). From the definition of the estimate we get that

$$\begin{aligned}
\int_{\mathbb{R}^d} m_n(x)\mu(dx) &= \int_{\mathbb{R}^d} \frac{\sum_{i=1}^n Y_i I_{\{X_i \in A_n(x)\}}}{\sum_{i=1}^n I_{\{X_i \in A_n(x)\}}}\mu(dx) \\
&= \sum_{A\in\mathcal{P}_n} \int_{\mathbb{R}^d} \frac{\sum_{i=1}^n Y_i I_{\{X_i \in A\}}}{\sum_{i=1}^n I_{\{X_i \in A\}}}\mu(dx) \\
&= \sum_{A\in\mathcal{P}_n} \frac{\sum_{i=1}^n Y_i I_{\{X_i \in A\}}}{\sum_{i=1}^n I_{\{X_i \in A\}}}\mu(A) \\
&= \frac{1}{n}\sum_{i=1}^n Y_i \frac{\mu(A_n(X_i))}{\mu_n(A_n(X_i))},
\end{aligned}$$

therefore

$$\int_{\mathbb{R}^d} m_n(x)\mu(dx) = \frac{1}{n}\sum_{i=1}^n (Y_i - m(X_i))\frac{\mu(A_n(X_i))}{\mu_n(A_n(X_i))} + \frac{1}{n}\sum_{i=1}^n m(X_i)\frac{\mu(A_n(X_i))}{\mu_n(A_n(X_i))},$$

and so

$$\begin{aligned}
&\mathbb{E}\left\{\Big|\int_{\mathbb{R}^d} m_n(x)\mu(dx) - \mathbb{E}\{m(X)\}\Big|\right\} \\
\leq\ &\mathbb{E}\left\{\Big|\frac{1}{n}\sum_{i=1}^n (Y_i - m(X_i))\frac{\mu(A_n(X_i))}{\mu_n(A_n(X_i))}\Big|\right\}
\end{aligned}$$

$$+\mathbb{E}\left\{\Big|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mu(A_n(X_i))}{\mu_n(A_n(X_i))}-1\right)(m(X_i)+L)\Big|\right\}$$
$$+\mathbb{E}\left\{\Big|\frac{1}{n}\sum_{i=1}^{n}m(X_i)-\mathbb{E}\{m(X)\}\Big|\right\}.$$

The first term of the right hand side is easy to manage, since

$$\begin{aligned}
&\mathbb{E}\left\{\Big|\frac{1}{n}\sum_{i=1}^{n}(Y_i-m(X_i))\frac{\mu(A_n(X_i))}{\mu_n(A_n(X_i))}\Big|\right\}\\
\leq\ & \sqrt{\mathbb{E}\left\{\Big|\frac{1}{n}\sum_{i=1}^{n}(Y_i-m(X_i))\frac{\mu(A_n(X_i))}{\mu_n(A_n(X_i))}\Big|^2\right\}}\\
\leq\ & \frac{2L}{\sqrt{n}}\sqrt{\mathbb{E}\left\{\frac{\mu(A_n(X_1))^2}{\mu_n(A_n(X_1))^2}\right\}}\\
=\ & \frac{2L}{\sqrt{n}}\sqrt{\sum_{A\in\mathcal{P}_n}\mathbb{P}\{X_1\in A\}\mathbb{E}\left\{\frac{\mu(A)^2}{\left(\frac{1}{n}(\sum_{i=2}^{n}I_{\{X_i\in A\}}+1)\right)^2}\right\}}\\
\leq\ & \frac{2L}{\sqrt{n}}\sqrt{2\sum_{A\in\mathcal{P}_n}\mu(A)}\\
=\ & \frac{2^{3/2}L}{\sqrt{n}},
\end{aligned}$$

where L denotes the bound of $|Y|$. For the third term of the right hand side, we get that

$$\begin{aligned}
\mathbb{E}\left\{\Big|\frac{1}{n}\sum_{i=1}^{n}m(X_i)-\mathbb{E}\{m(X)\}\Big|\right\} &\leq \sqrt{\mathbb{E}\left\{\Big|\frac{1}{n}\sum_{i=1}^{n}m(X_i)-\mathbb{E}\{m(X)\}\Big|^2\right\}}\\
&= \sqrt{\frac{\mathbb{V}ar(m(X))}{n}}\\
&\leq \frac{L}{\sqrt{n}}.
\end{aligned}$$

Concerning the second term of the right hand side, introduce the notations

$$\nu_n(A)=\frac{1}{n}\sum_{i=1}^{n}(m(X_i)+L)I_{\{X_i\in A\}}$$

and

$$\nu(A) = \int_A (m(x) + L)\mu(dx).$$

Then

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mu(A_n(X_i))}{\mu_n(A_n(X_i))} - 1\right)(m(X_i) + L) \\
= & \frac{1}{n}\sum_{i=1}^{n}\sum_{A\in\mathcal{P}_n} I_{\{X_i\in A\}}\left(\frac{\mu(A)}{\mu_n(A)} - 1\right)(m(X_i) + L) \\
= & \sum_{A\in\mathcal{P}_n}\left(\frac{\mu(A)}{\mu_n(A)} - 1\right)\frac{1}{n}\sum_{i=1}^{n} I_{\{X_i\in A\}}(m(X_i) + L) \\
= & \sum_{A\in\mathcal{P}_n}\frac{\nu_n(A)}{\mu_n(A)}(\mu(A) - \mu_n(A))I_{\{\mu_n(A)>0\}},
\end{aligned}$$

therefore

$$\begin{aligned}
& \mathbb{E}\left\{\Big|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mu(A_n(X_i))}{\mu_n(A_n(X_i))} - 1\right)(m(X_i) + L)\Big|\right\} \\
\leq & \mathbb{E}\left\{\Big|\sum_{A\in\mathcal{P}_n}\left(\frac{\nu_n(A)}{\mu_n(A)} - \frac{\nu(A)}{\mu(A)}\right)(\mu(A) - \mu_n(A))I_{\{\mu_n(A)>0\}}\Big|\right\} \\
& + \mathbb{E}\left\{\Big|\sum_{A\in\mathcal{P}_n}\frac{\nu(A)}{\mu(A)}(\mu(A) - \mu_n(A))\Big|\right\} \\
& + \mathbb{E}\left\{\Big|\sum_{A\in\mathcal{P}_n}\nu(A)I_{\{\mu_n(A)=0\}}\Big|\right\}.
\end{aligned}$$

Without loss of generality assume that $\mu(A_{n,j}) > 0$ for $j \leq M_n$, and $\mu(A_{n,j}) = 0$ otherwise. Then $M_n \leq c_{20}/h_n'^d$. The Lipschitz condition implies that

$$\left|\frac{\nu_n(A)}{\mu_n(A)} - \frac{\nu(A)}{\mu(A)}\right| I_{\{\mu_n(A)>0\}} \leq C\sqrt{d}h_n',$$

therefore

$$\mathbb{E}\left\{\Big|\sum_{A\in\mathcal{P}_n}\left(\frac{\nu_n(A)}{\mu_n(A)} - \frac{\nu(A)}{\mu(A)}\right)(\mu(A) - \mu_n(A))I_{\{\mu_n(A)>0\}}\Big|\right\}$$

$$
\begin{aligned}
&\le C\sqrt{d}h_n' \sum_{A\in\mathcal{P}_n} \mathbb{E}\left\{\Big|\mu(A)-\mu_n(A)\Big|\right\}\\
&\le C\sqrt{d}h_n'\sqrt{\frac{M_n}{n}}\\
&\le c_{21}n^{-2/(d+2)}.
\end{aligned}
$$

Moreover,

$$
\mathbb{E}\left\{\Big|\sum_{A\in\mathcal{P}_n}\nu(A)I_{\{\mu_n(A)=0\}}\Big|\right\} \le L\sum_{A\in\mathcal{P}_n}\mu(A)(1-(1-\mu(A))^n) \le \frac{LM_n}{n} \le c_{22}n^{-2/(d+2)},
$$

and

$$
\begin{aligned}
&\mathbb{E}\left\{\Big|\sum_{A\in\mathcal{P}_n}\frac{\nu(A)}{\mu(A)}(\mu(A)-\mu_n(A))\Big|\right\}\\
\le&\sqrt{\mathbb{E}\left\{\Big|\sum_{A\in\mathcal{P}_n}\frac{\nu(A)}{\mu(A)}(\mu(A)-\mu_n(A))\Big|^2\right\}}\\
\le&\sqrt{\sum_{A\in\mathcal{P}_n}\frac{\nu(A)^2}{\mu(A)^2}\mathbb{E}\left\{(\mu(A)-\mu_n(A))^2\right\}+\sum_{A\ne B\in P_n}\mathbb{C}ov\left(\frac{\nu(A)}{\mu(A)}\mu_n(A),\frac{\nu(B)}{\mu(B)}\mu_n(B)\right)}\\
\le&\sqrt{\sum_{A\in\mathcal{P}_n}\frac{\nu(A)^2}{\mu(A)^2}\frac{\mu(A)}{n}}\\
\le&\frac{L}{\sqrt{n}},
\end{aligned}
$$

where we applied that $\nu(A)\ge 0$ and $\nu(B)\ge 0$ and

$$
\mathbb{C}ov\left(\frac{\nu(A)}{\mu(A)}\mu_n(A),\frac{\nu(B)}{\mu(B)}\mu_n(B)\right) = \frac{\nu(A)}{\mu(A)}\frac{\nu(B)}{\mu(B)}\mathbb{C}ov\left(\mu_n(A),\mu_n(B)\right)\le 0,
$$

(cf. Mallows [26], Berlinet, Györfi, van der Meulen [3]). Summarizing these inequalities, the proof of the corollary is complete. □

References

[1] Ahmad, I. A. Residuals density estimation in nonparametric regression. *Statistics and Probability Letters*, 14, pp. 133-139, 1992.

[2] Akritas, M. G. and Van Keilegom, I. Non-parametric Estimation of the Residual Distribution. *Board of the Foundation of the Scandinavian Journal of Statistics*, Blackwell Publishers Ltd, 28, pp. 549-567, 2001.

[3] Berlinet, A., Györfi, L. and van der Meulen, E. C. Asymptotic normality of relative entropy in multivariate density estimation. *Publications de l'Institut de Statistique de l'Université de Paris*, 41, pp. 3-27, 1997.

[4] Cheng, F. Consistency of error density and distribution function estimators in nonparametric regression. *Statistics and Probability Letters*, 59, pp. 257-270, 2002.

[5] Cheng, F. Weak and strong uniform consistency of a kernel error density estimator in nonparametric regression. *Journal of Statistical Planning and Inference*, 119, pp. 95-107, 2004.

[6] Chow, Y. S. Local convergence of martingales and the law of large numbers. *Annals of Mathematical Statistics*, vol. 36, pp. 552–558, 1965.

[7] Devroye, L. Exponential inequalities in nonparametric estimation. In *Nonparametric functional estimation and related topics*, G. Roussas (Ed.), NATO ASI Series, Kluwer Academic Publishers, Dordrecht, pp. 31–44, 1991.

[8] Devroye, L., Felber, T., Kohler, M. and Krzyzak, A. $L1$-consistent estimation of the density of residuals in random design regression models. *Statistics and Probability Letters*, 82:173-179, 2012.

[9] Devroye, L. and Györfi, L. Distribution-free exponential upper bound on the L_1 error of partitioning estimates of a regression function. In *Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics*, eds. Konecny F., Mogyoródi, J. and Wertz, W. , pp. 67-76. Akadémiai Kiadó, Budapest, 1983.

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

[11] Devroye, L., Györfi, L., Krzyżak, A. and Lugosi, G. On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, 22, pp. 1371–1385, 1994.

[12] Devroye, L., Schäfer, D., Györfi, L. and Walk, H. The estimation problem of minimum mean squared error. *Statistics and Decisions*, 21, pp. 15-28, 2003.

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

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

[15] Dudoit, S. and van der Laan, M.J. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. Statistical Methodology, 2, pp. 131-154, 2005.

[16] Efromovich, S. Estimation of the density of regression errors. *Annals of Statistics*, 33, pp. 2194-2227, 2005.

[17] Efromovich, S. Optimal nonparametric estimation of the density of regression errors with finite support. *AISM*, 59, pp. 617-654, 2006.

[18] Györfi, L. Universal consistencies of regression estimate for unbounded regression functions. In *Nonparametric Functional Estimation and Related Topics*, ed. G. Roussas, pp. 329–338. Kluwer Academic Publishers, Dordrecht, 1991.

[19] Györfi, L., Györfi, Z. and Vajda, I. A strong law of large numbers and some applications, *Studia Sci. Math. Hungar.*, 12, pp. 233-244, 1977.

[20] Györfi, L. and Masry, E. The L_1 and L_2 strong consistency of recursive kernel density estimation from dependent samples, *IEEE Trans. Information Theory*, 36, pp. 531-539, 1990.

[21] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002.

[22] Kohler, M. Nonparametric regression with additional measurement errors in the dependent variable, *Journal of Statistical Planning and Inference*, 136, pp. 3339-3361, 2006.

[23] Liitiäinen, E., Corona, F. and Lendasse, A. On nonparametric residual variance estimation. *Neural Processing Letters*, 28, 155-167, 2009.

[24] Liitiäinen, E., Corona, F. and Lendasse, A. Residual variance estimation using a nearest neighbor statistic. *Journal of Multivariate Analysis*, 101, pp. 811-823, 2010.

[25] Liitiäinen, E., Verleysen, M, Corona, F. and Lendasse, A. Residual variance estimation in machine learning. *Neurocomputing*, 72, pp. 3692-3703, 2009.

[26] Mallows, C. L. An inequality involving multinomial probabilities. *Biometrika*, 55, p 422-424, 1968.

[27] Müller, H.-G. and Stadtmüller, U. Estimation of heteroscedasticity in regression analysis, *Annals of Statistics*, 15, pp. 610-625, 1987.

[28] Müller, U., Schick, A. and Wefelmeyer, W. Estimating the error variance in nonparametric regression by a covariate-matched U-statistic, *Statistics*, 37, pp. 179-188, 2003.

[29] Neumann, M.-H. Fully data-driven nonparametric variance estimators, *Statistics*, 25, pp. 189-212, 1994.

[30] Neumeyer, N. and Van Keilegom, I. Estimating the error distribution in nonparametric multiple regression with applications to model testing. *Journal of Multivariate Analysis*, 101, pp. 1067-1078, 2010.

[31] Pelckmans, K., De Brabanter, J., Suykens, J. A. K. and De Moor, B. The differogram: Non-parametric noise variance estimation and its use for model selection. Neurocomputing, 69, pp. 100-122, 2005.

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

[33] Stadtmüller, U. and Tsybakov, A. Nonparametric recursive variance estimation, *Statistics*, 27, pp. 55-63, 1995.

[34] Stone, C. J. Consistent nonparametric regression. *Annals of Statistics*, 5, pp. 595–645, 1977.

[35] Stout, W. F. *Almost sure convergence.* New York: Academic Press, 1974.

[36] Yamoto, H. Sequential estimation of a continuous probability density function and mode, *Bull. Math. Statist.*, 14, pp. 1–12, 1971.

[37] Walk, H. Almost sure convergence properties of Nadaraya-Watson regression estimates. In *Modeling Uncertainty. An Examination of its Theory, Methods and Applications*, eds. M. Dror, P. L'Ecuyer and F. Szidarovszky, pp. 201-223. Kluwer Academic Publishers, Dordrecht, 2002.

[38] Wolverton, C. T. and Wagner, T. J. Asymptotically optimal discriminant functions for pattern classification. *IEEE Trans. Information Theory*, IT-15, pp. 258–265, 1969.