Estimation of a density using real and artificial data [*]

Luc Devroye[1], Tina Felber[2], and Michael Kohler[2,†]

[1] *School of Computer Science, McGill University, 3480 University Street, Montreal, Canada H3A 2K6, email: lucdevroye@gmail.com*

[2] *Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: tfelber@mathematik.tu-darmstadt.de, kohler@mathematik.tu-darmstadt.de*

January 5, 2012

Abstract

Let X, X_1, X_2, ... be independent and identically distributed $\mathbb{R}^d$-valued random variables and let $m : \mathbb{R}^d \to \mathbb{R}$ be a measurable function such that a density f of $Y = m(X)$ exists. Given a sample of the distribution of (X, Y) and additional independent observations of X we are interested in estimating f. We apply a regression estimate to the sample of (X, Y) and use this estimate to generate additional artificial observations of Y. Using these artificial observations together with the real observations of Y we construct a density estimate of f by using a convex combination of two kernel density estimates. It is shown that if the bandwidths satisfy the usual conditions and if in addition the supremum norm error of the regression estimate converges almost surely faster towards zero than the bandwidth of the kernel density estimate applied to the artificial data, then the convex combination of the two density estimates is L_1–consistent. The performance of the estimate for finite sample size is illustrated by simulated data, and the usefulness of the procedure is demonstrated by applying it to a density estimation problem in a simulation model.

AMS classification: Primary 62G07; secondary 62G20.

Key words and phrases: Density estimation, L_1–error, nonparametric regression, consistency.

1 Introduction

Let X, X_1, X_2, ... be independent and identically distributed $\mathbb{R}^d$-valued random variables and let $m : \mathbb{R}^d \to \mathbb{R}$ be an unknown measurable function such that a density f of $Y = m(X)$ exists. The distribution of X is unknown—its measure will be denoted by μ. The density f of $Y = m(X)$ must be estimated, and estimates will be compared on the basis of total variation distance.

[*] Running title: *Density estimation using real and artificial data*

[†] Corresponding author. Tel: +49-6151-16-6846

This problem is substantially different from that of the estimation of the regression function m, as will be apparent from the discussion below. Note also that X does not have to have a density. In $\mathbb{R}^2$, consider $X = (U, U)$, where U is uniform on $[0, 1]$, and set $Y = m(X) = U$. Then Y is uniformly distributed, yet X does not have a density. In $\mathbb{R}^1$, a more intricate example involving the Cantor set shows that X has in general not a density. Let the ternary expansion of $X \in (0, 1)$ be $0.b_1b_2\ldots$, where $b_1, b_2, \ldots$ are i.i.d. and uniformly drawn from $\{0, 2\}$. Then X does not possess a density. Define the mapping m by the binary expansion of $m(X)$, given by $0.(b_1/2)(b_2/2)\ldots$. Since the bits in this expansion are i.i.d. and uniform on $\{0, 1\}$, $Y = m(X)$ is uniformly distributed on $[0, 1]$. However, if Y has a density, then X is non-atomic, i.e., continuous: its distribution function is continuous. We do not wish to assume anything about the underlying distribution of X.

We distinguish between three data models:

- (i) In the classical model, we have one data size constant, n, and we observe the i.i.d. sequence

  $$X_1, \ldots, X_n,$$

  (drawn from the distribution of X), and

  $$Y_i = m(X_i), 1 \le i \le n.$$

- (ii) In the finite information model, we have two data size constants, n and N, and we observe the i.i.d. sequence

  $$X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+N}$$

  (drawn from the distribution of X), and

  $$Y_i = m(X_i), 1 \le i \le n.$$

  This model is of interest in many applications, where the source of the X_i's is cheap and readily available, but the measurements Y_i are expensive or rare. Especially internet data fit this set-up.

- (iii) In the full information model, which corresponds to $N = \infty$, we assume that μ, the distribution of X, is known, and that we have access to

  $$X_1, \ldots, X_n,$$

  (drawn from the distribution of X), and

  $$Y_i = m(X_i), 1 \le i \le n.$$

  This model is of theoretical interest, as it delineates how far one can push the boundary in the finite information model.

The present paper takes a first look at the problem at hand, namely the estimation of the density f of Y for the finite information model (ii). It is of particular interest to learn how the presence of additional X-data (case (ii) with $N > 0$) can aid with the estimation. We present a new estimator, and are broadly concerned with its consistency under the widest possible conditions, never assuming anything about the underlying distribution of X. We also point out avenues of future research on this set of problems.

The safest way to approach this matter is by ignoring the X_i's altogether. In this case f can be estimated by applying, e.g., a standard kernel density estimate (Parzen (1962), Rosenblatt (1956)) defined by

$$f_n(y) = \frac{1}{h_n} \cdot \sum_{i=1}^{n} K\left(\frac{y - Y_i}{h_n}\right)$$

with some kernel function $K : \mathbb{R} \to \mathbb{R}$ which is a density (e.g., the naive kernel $K(u) = 1/2 \cdot 1_{[-1,1]}$) and some bandwidth $h_n > 0$, which is a parameter of the estimate. For this estimate it is known that

$$h_n \to 0 \quad (n \to \infty) \quad \text{and} \quad n \cdot h_n \to \infty \quad (n \to \infty)$$

imply that the estimate is L_1–consistent for all densities (cf., Mnatsakanov and Khmaladze (1981) and Devroye (1983)):

$$\int |f_n(x) - f(x)| \, dx \to 0 \text{ almost surely}$$

as $n \to \infty$. By Scheffé's Lemma (see, e.g., Devroye and Györfi (1985)) this implies that the estimated distribution converges to the true distribution in total variation distance and hence the above L_1–consistent density estimate allows simultaneous estimation of all probabilities. For general results in density estimation we refer to the books of Devroye and Györfi (1985), Devroye (1987) and Devroye and Lugosi (2000).

Improvements in the performance can be achieved in model (i) if additional information about m is available. This will not be our focus. In model (ii), without assuming anything about m or X, there is indeed help in the form of additional X_i's. We achieve this by estimating m by m_n based on the data

$$\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\},$$

which allows us to generate artificial (approximate) observations of $Y = m(X)$ via $\widehat{Y}_i = m_n(X_{n+i})$ $(i = 1, \dots, N)$. In a second step we apply separately kernel density estimates to the data sets $Y_1, \dots, Y_n$ and $\widehat{Y}_1, \dots, \widehat{Y}_N$ and use a convex combination of the resulting two density estimates as an estimate of f. The L_1–error of this estimate depends in particular on the interplay between the error of the estimate m_n and the bandwidth of the second kernel density estimate. In our main result we give sufficient conditions for the L_1–consistency of our estimate, and under suitable smoothness conditions on m we are able to show that our density estimate is indeed L_1–consistent. Furthermore we indicate under which condiditons our estimate should achieve a better rate of convergence than the simple estimate mentioned above, and these observations are confirmed in our simulation part. As an application we consider a density estimation method in a simulation method.

Estimation of m from the data $\mathcal{D}_n$ can be done via regression estimation, a field studied already over many years. The most popular estimates include kernel regression estimate (cf., e.g., Nadaraya (1964, 1970), Watson (1964), Devroye and Wagner (1980), Stone (1977, 1982) or Devroye and Krzyżak (1989)), partitioning regression estimate (cf., e.g., Györfi (1981) or Beirlant and Györfi (1998)), nearest neighbor regression estimate (cf., e.g., Devroye (1982) or Devroye, Györfi, Krzyżak and Lugosi (1994)), least squares estimates (cf., e.g., Lugosi and Zeger (1995) or Kohler (2000)) or smoothing spline estimates (cf., e.g., Whaba (1990) or Kohler and Krzyżak (2001)). For a detailed introduction to nonparametric regression we refer to Györfi et al. (2002).

Our analysis depends critically on the connection between the error of the regression estimate and the error of the density estimate. A similar phenomenon occurs in density estimation of the density of residuals of a regression model (cf., e.g., Ahmad (1992), Cheng (2004), Efromovich (2005, 2006) or Devroye et al. (2012)), and in our proof we apply techniques related to the ones in Devroye et al. (2012).

Our data set can be considered as one data set (X_1, Y_1), $\dots$, (X_n, Y_n) with labels and one unlabelled data set X_{n+1}, X_{n+2}, $\dots$and our procedure can be considered as semi-supervised learning for density estimation, which is usually studied in the context of pattern recognition (cf., e.g., Castelli and Cover (1996), Chapelle et.al. (2006) and the wide-ranging literature cited therein.)

The outline of the paper is as follows: We start in Section 2 with a discrete analog of the problem, where we illustrate the potential usefulness of the artificial data. Then we present in Section 3 a general consistency result for a newly proposed estimate in the general finite information model, indicate in Section 4 how in a special situation the used regression estimate might be improved drastically, investigate in Section 5 the performance of the estimate from Section 3 for finite sample size by simulated data and illustrate the usefulness of the procedure by applying it to a density estimation problem in a simulation model, and give an outlook in Section 6. The proof of our main result is given in the Appendix.

2 The discrete analog

In the discrete version of this problem, X is a random variable on the positive integers with

$$p_i = \mathbf{P}\{X = i\}.$$

There is an unknown function m on the positive integers. The objective is to estimate the distribution of the atomic random variable $m(X)$. Since m itself takes only countably many values, the canonical version of the problem is such that m itself takes values in the positive integers. We define

$$q_j = \mathbf{P}\{m(X) = j\},$$

and are interested in estimating q_j from data such that the total variation error is small.

We distinguish between three data models:

- (i) In the primitive version, we have one data size constant, n, and we observe the i.i.d. sequence
$$X_1, \ldots, X_n,$$
(drawn from the distribution of X), and
$$Y_i = m(X_i), 1 \le i \le n.$$

- (ii) In the finite information version, we have two data size constants, n and N, and we observe the i.i.d. sequence
$$X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+N}$$
(drawn from the distribution of X), and
$$Y_i = m(X_i), 1 \le i \le n.$$

- (iii) In the full information version, which corresponds to $N = \infty$, we assume that $\{p_i : i \ge 1\}$ is given and that we have access to
$$X_1, \ldots, X_n,$$
(drawn from the distribution of X), and
$$Y_i = m(X_i), 1 \le i \le n.$$

We would like to investigate how the additional data $X_{n+1}, \ldots, X_{n+N}$ can help in the estimation. Since models (i) and (iii) correspond to $N = 0$ and $N = \infty$, respectively, it should be clear that we should first try to compare (i) and (iii). In case (i), it is difficult to improve on the empirical estimate,

$$q_{n,j} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{[m(X_i)=j]}, j \ge 1.$$

Note that $nq_{n,j}$ is binomial (n, q_j), and hence $q_{n,j}$ is an unbiased estimate of q_j. The expected total variation error is easily bounded—we give it explicitly for later reference:

$$\begin{aligned}
\sum_j \mathbf{E}\{|q_j - q_{n,j}|\} \quad &= 2\sum_j \mathbf{E}\{(q_j - q_{n,j})_+\} \\
&\le 2\sum_j \min\left(q_j, \sqrt{\mathrm{Var}\{q_{n,j}\}}\right) \\
&= 2\sum_j \min\left(q_j, \sqrt{q_j(1-q_j)/n}\right).
\end{aligned}$$

If $\sum_j \sqrt{q_j} < \infty$, then the upper bound is $O(1/\sqrt{n})$. In all but the trivial case that $q_j = 1$ for some j, the expected total variation error

$$\sum_j \mathbf{E}\{|q_j - q_{n,j}|\}$$

tends to 0 at the rate $1/\sqrt{n}$ or slower, because if $q_k \in (0,1)$ for some $k \in \mathbb{N}$ then

$$\frac{\mathbf{E}\{|q_k - q_{n,k}|\}}{1/\sqrt{n}} \to 0 \quad (n \to \infty)$$

implies that

$$\frac{\sqrt{n} \cdot (q_{n,k} - q_k)}{\sqrt{q_k \cdot (1 - q_k)}} \to 0 \quad \text{in } L_1$$

which is a contradiction to the central limit theorem.

In case (iii), the situation is remarkably different. Let $A = \{X_1, \ldots, X_n\}$ with duplicates removed. If $i \in A$, $m(i)$ is known. If $i \notin A$, m is unknown and cannot possibly be guessed. Since

$$q_j = \mathbf{P}\{m(X) = j\} = \sum_i p_i \mathbb{I}_{[m(i)=j]},$$

we set

$$q_{n,j} = \sum_{i \in A} p_i \mathbb{I}_{[m(i)=j]}.$$

Clearly, $0 \le q_{n,j} \le q_j$, and we do not have the unbiasedness we enjoyed in case (i). However, the expected total variation error has a simple and universal expression that is the same (!!!) for all choices of m. First note that the total variation error is

$$\begin{aligned} \sum_j (q_j - q_{n,j}) &= \sum_j \left(\sum_i p_i \mathbb{I}_{[m(i)=j]} - \sum_{i \in A} p_i \mathbb{I}_{[m(i)=j]} \right) \\ &= \sum_i p_i - \sum_{i \in A} p_i \\ &= \sum_{i \notin A} p_i. \end{aligned}$$

Thus, the expected total variation error is

$$\sum_i p_i \mathbf{P}\{i \notin A\} = \sum_i p_i (1 - p_i)^n.$$

This error tends to zero with n in all cases, and the rate of decrease depends upon the tail of $\{p_i\}$. However, it is much better than for (i). To wit, consider X with compact support. Then the expected total variation error tends to 0 at an exponential rate in n.

It is of interest to see how the finite information model interpolates between these two behaviors.

3 The general finite information model

In the sequel we consider the general finite information model, where we introduce a new density estimate and study its consistency.

In the definition of our generic estimate, we first apply a suitably defined regression estimate m_n to the data

$$\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$$

in order to estimate $m : \mathbb{R}^d \to \mathbb{R}$. Then we use the estimate m_n of m in order to define an artificial sample of Y. To do this, we choose the size N of this sample and define artificial data via

$$\widehat{Y}_1 = m_n(X_{n+1}), \dots, \widehat{Y}_N = m_n(X_{n+N}).$$

Next we apply standard kernel density estimates separately to the data $Y_1, \dots, Y_n$ and $\widehat{Y}_1, \dots, \widehat{Y}_N$. Let K be the so-called naive kernel defined by

$$K(u) = \frac{1}{2} \cdot 1_{[-1,1]}(u) \quad (u \in \mathbb{R}),$$

let $h_n > 0$ and $\widehat{h}_N > 0$ and define

$$f_n(y) = \frac{1}{n \cdot h_n} \cdot \sum_{i=1}^{n} K\left(\frac{y - Y_i}{h_n}\right)$$

and

$$\widehat{f}_N(y) = \frac{1}{N \cdot \widehat{h}_N} \cdot \sum_{i=1}^{N} K\left(\frac{y - \widehat{Y}_i}{\widehat{h}_N}\right).$$

Finally we use a convex combination of these two estimates as estimate of f, i.e., we choose a weight

$$w_n = w_n\left(\mathcal{D}_n, X_{n+1}, \dots, X_{n+N}\right) \in [0, 1]$$

and estimate f by

$$g_n = w_n \cdot f_n + (1 - w_n) \cdot \widehat{f}_N.$$

The precise choice of weight function is left open, as is the choice of regression function estimate m_n. The first business at hand is to determine the consistency of the generic estimate. Since we do not impose restrictions on w_n, the choice $w_n = 0$ and $w_n = 1$ imply that both f_n and $\widehat{f}_N$ must be consistent. The latter can only happen if $N \to \infty$. Also, the performance of m_n is critical. For example, if one lets m_n be the nearest neighbor regression function estimate ($m_n(x) = m(X_i)$ if X_i is the nearest neighbor of x), then is the atomic nature of such m_n a problem for the consistency of $\widehat{f}_N$? The consistency theorem we present in the next section takes a higher view, and gives a natural technical condition that links m_n to $\widehat{h}_N$.

Our main result is the following theorem.

Theorem 1 *Let X, X_1, X_2, $\ldots$ be independent and identically distributed $\mathbb{R}^d$-valued random variables and let $m : \mathbb{R}^d \to \mathbb{R}$ be a measurable function such that a density f of $Y = m(X)$ exists. Set $Y_i = m(X_i)$ $(i \in \mathbb{N})$. Let K be the naive kernel and for $n \in \mathbb{N}$ let $N \in \mathbb{N}$, $h_n > 0$ and $\widehat{h}_N > 0$. Given (X_1, Y_1), $\ldots$, (X_n, Y_n), X_{n+1}, X_{n+2}, $\ldots$, and a regression estimate m_n let g_n be the estimate of f as defined in Section 2. Assume that $N = N(n) \to \infty$ as $n \to \infty$, and that*

$$\max(h_n, \widehat{h}_n) \to 0 \quad (n \to \infty), \quad n \cdot \min(h_n, \widehat{h}_n) \to \infty \quad (n \to \infty). \tag{1}$$

Assume furthermore the following on m_n: for every $\epsilon > 0$, $n \in \mathbb{N}$, there exists a (random) set $A_{n,\epsilon} = A_{n,\epsilon}(\mathcal{D}_n) \subseteq \mathbb{R}$ such that

$$\lim_{n\to\infty} \mathbf{P}\{\mu\{A^c_{n,\epsilon}\} > \epsilon\} = 0 \tag{2}$$

and

$$\frac{\|m_n - m\|_{\infty, A_{n,\epsilon}}}{\widehat{h}_N} = \frac{\sup_{x \in A_{n,\epsilon}} |m_n(x) - m(x)|}{\widehat{h}_N} = o(1) \quad \textit{in probability.} \tag{3}$$

Then, regardless of how w_n is chosen,

$$\int_{\mathbb{R}} |g_n(y) - f(y)|\, dy \to 0 \quad \textit{in probability.}$$

[In particular, $\int_{\mathbb{R}} \left| \widehat{f}_N(y) - f(y) \right| dy = o(1)$ in probability as well.] If, in addition,

$$\limsup_{n\to\infty} \mu\{A^c_{n,\epsilon}\} \le \epsilon \quad \textit{almost surely} \tag{4}$$

and

$$\frac{\|m_n - m\|_{\infty, A_{n,\epsilon}}}{\widehat{h}_N} = \frac{\sup_{x \in A_{n,\epsilon}} |m_n(x) - m(x)|}{\widehat{h}_N} = o(1) \quad \textit{almost surely} \tag{5}$$

then

$$\int_{\mathbb{R}} |g_n(y) - f(y)|\, dy \to 0 \quad \textit{almost surely.}$$

Proof. The proof is given in the Appendix.

Conditions (3) and (5) can be derived from rate of convergence results for nonparametric regression estimates. A less cumbersone, but weaker, consistency result is the following

Corollary 1 *Let X, X_1, X_2, $\ldots$ be independent and identically distributed $\mathbb{R}^d$-valued random variables and let $m : \mathbb{R}^d \to \mathbb{R}$ be a measurable function such that a density f of $Y = m(X)$ exists. Let K be the naive kernel and for $n \in \mathbb{N}$ let $N \in \mathbb{N}$, $h_n > 0$ and $\widehat{h}_N > 0$. Set $Y_i = m(X_i)$ $(i \in \mathbb{N})$. Given (X_1, Y_1), $\ldots$, (X_n, Y_n), X_{n+1}, X_{n+2}, $\ldots$, and a regression estimate m_n let g_n be the estimate of f as defined in Section 2. Assume that (1) holds.*

(i) *If, in addition,*

$$\mathbf{E}\int |m_n(x)-m(x)|^2\mu(dx) = \mathbf{E}\left(|m_n(X)-m(X)|^2\right) = o\left(\widehat{h}_N^2\right), \tag{6}$$

then, regardless of the choice of w_n,

$$\int_{\mathbb{R}} |g_n(y)-f(y)|\,dy \to 0 \quad \textit{in probability.}$$

(ii) *If, in addition,*

$$\frac{\int |m_n(x)-m(x)|^2\mu(dx)}{\widehat{h}_N^2} \to 0 \quad \textit{almost surely,} \tag{7}$$

then, regardless of the choice of w_n,

$$\int_{\mathbb{R}} |g_n(y)-f(y)|\,dy \to 0 \quad \textit{almost surely.}$$

Proof. In order to prove (i), choose $a_n \in \mathbb{R}_+$ such that

$$\frac{\mathbf{E}\int |m_n(x)-m(x)|^2\mu(dx)}{a_n^2} = o(1) \quad \text{and} \quad \frac{a_n}{\widehat{h}_N} = o(1). \tag{8}$$

In order to show (ii), choose $a_n = a_n(\mathcal{D}_n) \in \mathbb{R}_+$ such that

$$\frac{\int |m_n(x)-m(x)|^2\mu(dx)}{a_n^2} = o(1) \quad \text{almost surely} \quad \text{and} \quad \frac{a_n}{\widehat{h}_N} = o(1) \quad \text{almost surely.} \tag{9}$$

In both cases set

$$A_{n,\epsilon} = \left\{x \in \mathbb{R}^d : |m_n(x)-m(x)| \le a_n\right\}.$$

Then

$$\frac{\|m_n - m\|_{\infty, A_{n,\epsilon}}}{\widehat{h}_N} \le \frac{a_n}{\widehat{h}_N} \to 0 \quad \text{almost surely}$$

by (8) or (9), respectively. Furthermore, by Markov's inequality we have

$$\mu(A_{n,\epsilon}^c) \le \frac{\int |m_n(x)-m(x)|^2\mu(dx)}{a_n^2},$$

so (2) and (4) are implied by (8) and (9), resp. Application of Theorem 1 yields the assertion. □

Remark 1. Assume that m is a linear function and that m_n is a linear regression estimate. In this case we expect

$$\limsup_{n\to\infty} \frac{\mathbf{E}\int |m_n(x)-m(x)|^2\mu(dx)}{1/n} < \infty.$$

If we require that $\lim_{n\to\infty} \sqrt{n} \cdot h_N = \infty$, then the assumption (6) is satisfied and the estimate g_n (and thus $\widehat{f_N}$) is weakly L_1–consistent. If N is very large and f is Lipschitz continuous, the L_1–error of $\widehat{f_N}$ is dominated by the bias which is of order h_N, and this can be arbitrarily close to $1/\sqrt{n}$. This is in contrast to the rate of convergence of f_n for Lipschitz continuous f, which is of order $n^{-1/(2\cdot 1+1)} = n^{-1/3}$ (Devroye and Györfi (1985)).

Remark 2. Assume that m is (p, C)-smooth, which means in case $p \in \mathbb{N}$ that all partial derivatives of order $p-1$ of m exist and are Lipschitz continuous with Lipschitz constant C. Furthermore assume that X and Y are both bounded almost surely. Then standard least squares estimates m_n satisfy

$$\limsup_{n\to\infty} \frac{\mathbf{E}\int |m_n(x) - m(x)|^2 \mu(dx)}{n^{-2p/(2p+d)}} < \infty$$

(cf., Kohler (2000) or Corollary 19.1 in Györfi et al. (2002)). The assumption (6) is satisfied if we have $\lim_{n\to\infty} n^{p/(2p+d)} \cdot h_N = \infty$. In that case, the estimate $\widehat{f_N}$ is weakly L_1-consistent. When N is very large, the L_1-error of $\widehat{f_N}$ is again dominated by the bias, and arguing as in Remark 1 we see that we can expect that for Lipschitz continuous f, the rate of convergence of $\widehat{f_N}$ is better than the one for f if

$$\frac{p}{2p+d} > \frac{1}{3}, \quad \text{which is equivalent to} \quad p > d.$$

Remark 3. The techniques introduced in the proofs of Theorem 1 and Corollary 1 can be applied in the context of estimation of the density of the residuals in a nonparametric regression model. Here we assume that $Y - m(X)$ has a density f, and we try to estimate f using the data $(X_1, Y_1), \ldots, X_{2n}, Y_{2n})$. To do this, we compute first a regression estimate $m_n(x) = m_n(x, \mathcal{D}_n)$ and use then

$$\widehat{f_n}(x) = \frac{1}{n \cdot h_n} \cdot \sum_{i=1}^{n} K\left(\frac{x - (Y_{n+i} - m_n(X_{n+i}))}{h_n}\right)$$

as estimate of f. Assume $h_n \to 0$ $(n \to \infty)$, $n \cdot h_n \to \infty$ $(n \to \infty)$ and

$$\frac{\mathbf{E}\int |m_n(x) - m(x)|^2 \mu(dx)}{\widehat{h}_n^2} \to 0 \quad (n \to \infty). \tag{10}$$

We argue as in the proof of Theorem 1, where we use

$$\mathbf{E}\{\widehat{f_n}(x)|\mathcal{D}_n\} = \int \frac{1}{h_n} \cdot K\left(\frac{x - (y - m_n(z))}{h_n}\right) \nu(dz, dy)$$

(where ν is the joint probability measure of (X, Y)) and replace $m_n(X_{n+i})$ by $Y_{n+i} - m_n(X_{n+i})$). Then we obtain weak consistency:

$$\int_{\mathbb{R}} |\widehat{f_n}(y) - f(y)|\, dy \to 0 \quad \text{in probability.}$$

It follows from Stone (1977) that there exist weakly universally consistent regression estimates, so for each distribution of (X, Y) we can find a sequence of bandwidths h_n (depending on this distribution) such that (10) holds. Observe that this result does not require that X and $Y - m(X)$ are independent.

It is an open problem, whether the same result also holds for a fixed sequence of bandwidths $\{h_n\}_n$ and a fixed sequence of regression estimates.

4 The linear interpoland

In this section, we look at a very specific choice of m_n, the linear interpoland, when $d = 1$, and in addition, $m : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable and the distribution function F of X satisfies for some $a < b$: $F(a) = 0$, $F(b) = 1$ and on (a, b), F is twice continuously differentiable with first derivative bounded away from zero and with second derivative uniformly bounded on (a, b). [Equivalently, μ has a continuously differentiable density on $[a, b]$ that is bounded away from 0 on that interval.] These assumptions imply that F^{-1} exists, has on $(0, 1)$ a uniformly bounded second derivative and that $m \circ F^{-1}$ is twice continuously differentiable with uniformly bounded second derivative on $(0, 1)$. This is an exploratory example that illustrates how one can handsomely beat the rates suggested in the remarks of the previous section when N is large.

Two observations are crucial: first of all, the total variation error is invariant under componentwise monotone transformations of X, if we wish to estimate the density of X. Secondly, if g is known, as in model (iii), then the probability integral transform can be used and X can be replaced by $G(X)$, where G is the (known) distribution function of X. In particular, for $d = 1$, we may in all cases assume, without loss of generality, that X is uniformly distributed on $[0, 1]$. This is called the canonical version of the problem.

The canonical version above corresponds to the case $N = \infty$, where we assume that F (i.e., μ) is explicitly known. In the sequel we take a slightly more realistic stance and assume instead that $N > n^4$. Let F_N be the empirical distribution function for $X_{n+1}, \ldots, X_{n+N}$, and let $Q_n m$ be the linear interpoland of $(F_N(X_i), m(X_i))$ $(i = 1, \ldots, n)$ (where the first and the last linear part are extended to $-\infty$ and ∞, respectively), let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics of $X_1, \ldots, X_n$, and set

$$m_n(x) = \begin{cases} (Q_n m)(F_N(x)), & \text{if } X_{(1)} \le x \le X_{(n)}, \\ m(X_{(1)}), & \text{if } x < X_{(1)}, \\ m(X_{(n)}), & \text{if } x > X_{(n)}. \end{cases}$$

Then

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \, \mu(dx) = O\left(\frac{1}{n^3}\right). \tag{11}$$

Consequently, if we demand that $\lim_{n\to\infty} n^{3/2} \cdot h_N = \infty$ (so that assumption (6) is satisfied), the estimate $\widehat{f}_N$ is weakly L_1–consistent. Arguing as above we see that we can expect in case of Lipschitz smooth densities rates of convergence better than n^{-r} for any $r < 3/2$.

Proof of (11). Let $\bar{Q}_n m$ be the linear interpoland of $(F(X_i), m(X_i))$ $(i = 1, \ldots, n)$. For $x \in (a, b)$ set

$$\eta_n(x) = \begin{cases} F(X_{(1)}), & \text{if } F(x) \leq F(X_{(1)}), \\ F(X_{(k)}) - F(X_{(k-1)}), & \text{if } F(X_{(k-1)}) < F(x) \leq F(X_{(k)}), \\ 1 - F(X_{(n)}), & \text{if } F(x) \geq F(X_{(n)}). \end{cases}$$

Let $x \in (a, b)$ be arbitrary. First we assume $x \in [X_{(1)}, X_{(n)}]$. In this case we have

$$\begin{aligned} |m_n(x) - m(x)| &\leq \left|(Q_n m)\,(F_N(x)) - \left(\bar{Q}_n m\right)(F_N(x))\right| \\ &\quad + \left|\left(\bar{Q}_n m\right)(F_N(x)) - \left(\bar{Q}_n m\right)(F(x))\right| \\ &\quad + \left|\left(\bar{Q}_n m\right)(F(x)) - \left(m \circ F^{-1}\right)(F(x))\right| \\ &=: T_{1,n} + T_{2,n} + T_{3,n}. \end{aligned}$$

Since

$$\frac{m(X_{(k)}) - m(X_{(k-1)})}{F(X_{(k)}) - F(X_{(k-1)})} = \frac{m \circ F^{-1}(F(X_{(k)})) - m \circ F^{-1}(F(X_{(k-1)}))}{F(X_{(k)}) - F(X_{(k-1)})} = (m \circ F^{-1})'(\xi)$$

for some $\xi \in (F(X_{(k-1)}), F(X_{(k)}))$, $\bar{Q}_n m$ is Lipschitz continuous with Lipschitz constant bounded by $\sup_{z \in (0,1)} \left|\left(m \circ F^{-1}\right)'(z)\right|$, from which we conclude

$$T_{2,n} \leq \sup_{z \in (0,1)} \left|\left(m \circ F^{-1}\right)'(z)\right| \cdot \sup_{u \in \mathbb{R}} |F_N(u) - F(u)|\,.$$

By construction of $Q_n m$ and of $\bar{Q}_n m$, for any $y \in [F_N(X_{(1)}), F_N(X_{(n)})]$ there exists $\bar{y} \in \mathbb{R}$ satisfying

$$(Q_n m)(y) = (\bar{Q}_n m)(\bar{y}) \quad \text{and} \quad |y - \bar{y}| \leq \sup_{u \in \mathbb{R}} |F_N(u) - F(u)|\,.$$

(In case $y = F_N(X_{(k)}) + \alpha \cdot (F_N(X_{(k+1)}) - F_N(X_{(k)}))$ for some $k \in \{1, \ldots, n-1\}$ and some $\alpha \in [0, 1]$ we set $\bar{y} = F(X_{(k)}) + \alpha \cdot (F(X_{(k+1)}) - F(X_{(k)}))$.) Consequently, setting $y = F_N(x)$ we get by using again the Lipschitz property of $\bar{Q}_n m$:

$$T_{1,n} = |(\bar{Q}_n m)(\bar{y}) - (\bar{Q}_n m)(y)| \leq \sup_{z \in (0,1)} \left|\left(m \circ F^{-1}\right)'(z)\right| \cdot \sup_{u \in \mathbb{R}} |F_N(u) - F(u)|\,.$$

Finally, to bound $T_{3,n}$ we observe that $\left(m \circ F^{-1}\right)(u)$ is equal to the value of the first term of the Taylor series of $m \circ F^{-1}$ around u and evaluated at u, and that this Taylor series polynomial p satisfies

$$\bar{Q}_n p = p.$$

Hence, setting $u = F(x)$, it suffices to bound

$$\left|\left(\bar{Q}_n m\right)(u) - \left(\bar{Q}_n p\right)(u)\right|.$$

Since both expressions are linear interpolands of functions at points with x-values $F(X_i)$ $(i = 1, \dots, n)$, their maximum distance for $u \in \left[F(X_{(1)}), F(X_{(n)})\right]$ is bounded by the distance between $m \circ F^{-1}$ and its Taylor series approximant evaluated at the x-points closest to u, which has distance $\eta_n(x)$ from x.

Summarizing the above results we get for $x \in [X_{(1)}, X_{(n)}]$

$$\begin{aligned} |m_n(x) - m(x)| \quad &\leq \quad 2 \cdot \sup_{z \in (0,1)} \left| \left(m \circ F^{-1}\right)'(z) \right| \cdot \sup_{u \in \mathbb{R}} |F_N(u) - F(u)| \\ &\quad + \frac{1}{2} \sup_{z \in (0,1)} \left| \left(m \circ F^{-1}\right)''(z) \right| \cdot \eta_n(x)^2. \end{aligned}$$

For $x < X_{(1)}$ we get (by using the Lipschitz property of $m \circ F^{(-1)}$)

$$|m_n(x) - m(x)| = |(m \circ F^{-1})(F(X_{(1)})) - (m \circ F^{-1})(F(x))| \leq \sup_{z \in (0,1)} \left| \left(m \circ F^{-1}\right)'(z) \right| \cdot F(X_{(1)})$$

and for $x > X_{(n)}$ we have

$$|m_n(x) - m(x)| \leq \sup_{z \in (0,1)} \left| \left(m \circ F^{-1}\right)'(z) \right| \cdot (1 - F(X_{(n)})).$$

This implies

$$\begin{aligned} &\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \\ &\leq \text{const } \cdot \mathbf{E} \left(\sup_{u \in \mathbb{R}} |F_N(u) - F(u)|^2 \right) + \text{const } \cdot \mathbf{E} \left(\eta_n(X)^4 \right) \\ &\quad + \text{const } \cdot \mathbf{E} \left(F(X_{(1)})^3 \right) + \text{const } \cdot \mathbf{E} \left((1 - F(X_{(n)}))^3 \right). \end{aligned}$$

Using

$$\mathbf{E}\left(Z^k\right) = \int_0^\infty k \cdot t^{k-1} \cdot \mathbf{P}\{Z > t\} \, dt$$

for a non-negative real random variable X we can conclude from Problem 9.5 in Györfi et al. (2002) that

$$\begin{aligned} \mathbf{E} \left(\sup_{u \in \mathbb{R}} |F_N(u) - F(u)|^2 \right) \quad &\leq \quad \frac{1}{n^3} + \int_{1/n^3}^\infty \mathbf{P} \left\{ \sup_{u \in \mathbb{R}} |F_N(u) - F(u)| > \sqrt{t} \right\} dt \\ &\leq \quad \frac{1}{n^3} + \int_{1/n^3}^1 8 \cdot (N+1) \cdot \exp\left(-\frac{N \cdot t}{128} \right) dt \\ &= \quad O\left(\frac{1}{n^3} \right). \end{aligned}$$

Furthermore using the fact that $F(X_1)$ is uniformly distributed on $[0, 1]$ we get

$$\mathbf{E}\left(\eta_n(X)^4\right)$$

$$
\begin{aligned}
&= \int_0^\infty 4 \cdot t^3 \cdot \mathbf{P}\left[\eta_n(X) > t\right] \, dt \\
&\leq 4 \cdot \int_0^1 t^3 \cdot \mathbf{P}\left[\forall i \in \{1,\dots,n\} : F(X_i) \notin \left[F(X) - \frac{t}{2},\ F(X)\right], F(X) > \frac{t}{2}\right] \, dt \\
&\quad + 4 \cdot \int_0^1 t^3 \cdot \mathbf{P}\left[\forall i \in \{1,\dots,n\} : F(X_i) \notin \left[F(X),\ F(X) + \frac{t}{2}\right], F(X) < 1 - \frac{t}{2}\right] \, dt \\
&= 4 \cdot \int_0^1 t^3 \cdot 2 \cdot \left(1 - \frac{t}{2}\right)^n \, dt \\
&\leq 8 \cdot \int_0^1 t^3 \cdot \exp\left(-\frac{n\,t}{2}\right) \, dt \\
&\leq \frac{384}{n^3} \int_0^1 \exp\left(-\frac{n\,t}{2}\right) \, dt = O\left(\frac{1}{n^4}\right),
\end{aligned}
$$

where the last inequality follows by partial integration.

Finally we observe

$$
\begin{aligned}
\mathbf{E}\left((1 - F(X_{(n)}))^3\right) = \mathbf{E}\left(F(X_{(1)})^3\right) &= \int_0^1 3 \cdot t^2 \cdot \mathbf{P}\left[\forall i \in \{1,\dots,n\} \,:\, F(X_i) > t\right] \, dt \\
&\leq \int_0^1 3 \cdot t^2 \cdot e^{-n \cdot t} \, dt = \frac{6}{n^3} = O\left(\frac{1}{n^3}\right).
\end{aligned}
$$

□

Remark 4. For $d > 1$, we may assume that X has a copula distribution (i.e., a distribution with uniform marginals). In this case it is not necessary to assume that X has a density g.

5 Application to simulated data

In this section we illustrate the finite sample size performance of our estimates by applying them to simulated data.

In our first example we set $X = (X^{(1)}, X^{(2)})$ for independent standard normally distributed random variables $X^{(1)}$ and $X^{(2)}$ and choose $Y = m(X)$ for

$$
m(x_1, x_2) = 2 \cdot x_1 + x_2 + 2.
$$

In this case Y is normally distributed with expectation 2 and variance $2^2 + 1^2 = 5$. We estimate the density of Y by the estimate introduced in Section 2, where we use a fully data-driven smoothing spline estimate to estimate the linear function m. For this purpose we use the routine *Tps()* from the library *fields* in the statistics package *R*. For the weights we use three different values: $w_n = 1$ (in which case we use only the real data), $w_n = 0$ (in which case the estimate is based only on the artificial data) and $w_n = n/(n + N)$ (in which case we use real and artificial data and all data points have the same weight). We set $n = 200$ and $N = 800$ and choose the bandwidths by minimizing the L_1–errors of the estimate via comparing the estimated density with the true density (so we assume that we have available an oracle which chooses the optimal bandwidth, so that we can ignore effects occuring because of inproper choice of the bandwidths). Figure 1 shows the three estimates and the true density in a typical simulation. Since the result of our simulation depends on the randomly occuring data points, we repeat the simulation 100 times with independent realizations of the occuring random variables and report in Figure 2 boxplots of the occuring L_1–errors (where we approximate the integrals by Riemann sums in order to compute the L_1–errors approximately). Comparing the boxplots in Figure 2 we see that the median of the L_1–errors in case of the estimate which uses only artificial data (0.1097) is nearly twice as big as the median of the L_1–errors of the estimate which uses only artificial data (0.0648). If we assign the same weight to every data point it is even more smaller (0.0612).

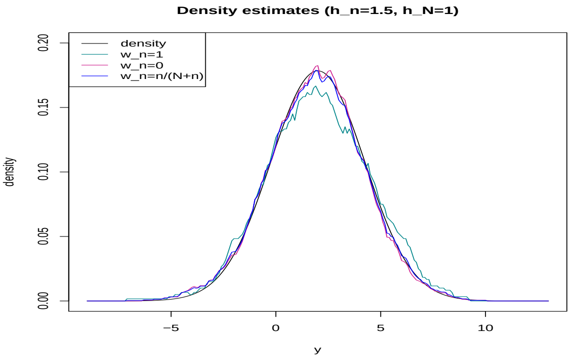

Figure 1: Density estimates in the first model

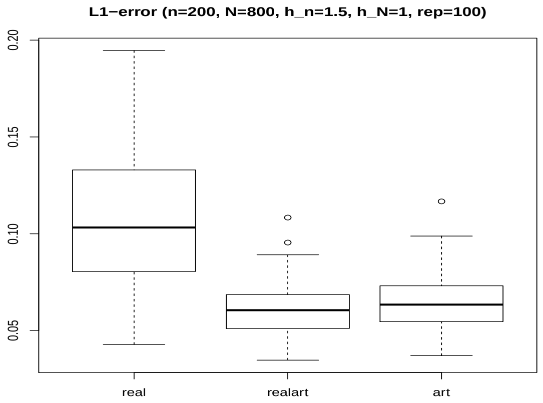

Figure 2: Boxplots in the first model

In our second example we set $X = (X^{(1)}, X^{(2)})$ for independent standard normally distributed random variables $X^{(1)}$ and $X^{(2)}$ and choose $Y = m(X)$ for

$$m(x_1, x_2) = x_1^2 + x_2^2.$$

Then Y is chi-squared distributed with two degrees of freedom. We define the estimate as in the first example. Again Figure 3 shows the three estimates and the true density in a typical simulation, and in Figure 4 we compare boxplots of the L_1–errors of the estimate. From Figure 4 we see that the mean L_1–error of the estimate with $w_n = 0$ (0.1426) is well below the first estimate with $w_n = 1$ (0.2208). The median of the estimate which uses real and artificial data is the smallest (0.1321).

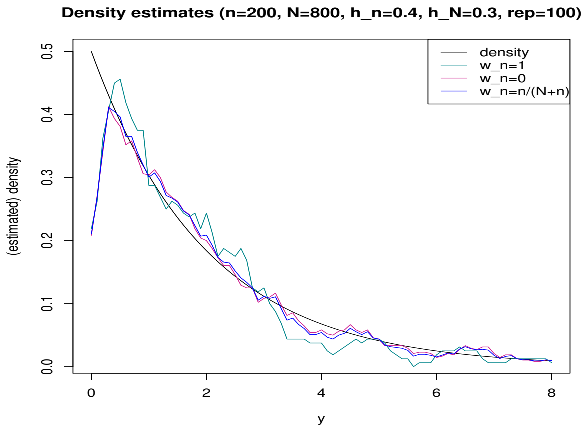

Figure 3: Density estimates in the second model

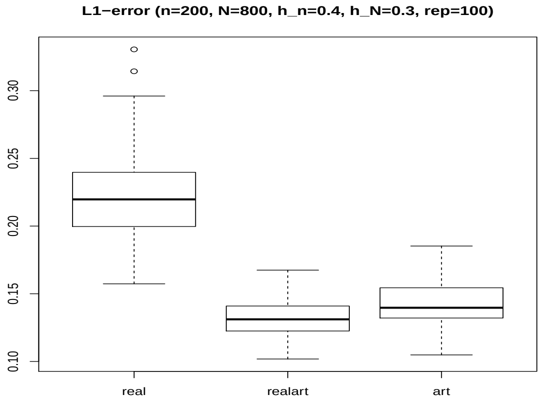

Figure 4: Boxplots in the second model

In Figures 5 and 6 we repeat the same simulation choosing X as a standard-normally distributed random variable and $m(x) = \exp(x)$. In this case $Y = m(X)$ is log-normally distributed. Figure 6 shows the same results as before. The estimate which uses only real data is again the worst (0.2221). If every data point has the same weight the mean L_1–error (0.1341) is smaller than if we use only artificial data (0.1402).

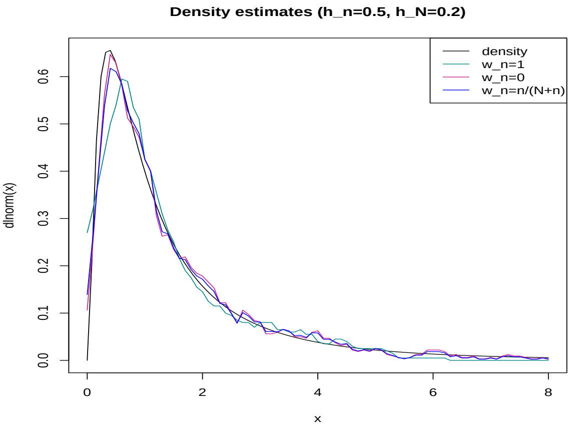

Figure 5: Density estimates in the third model

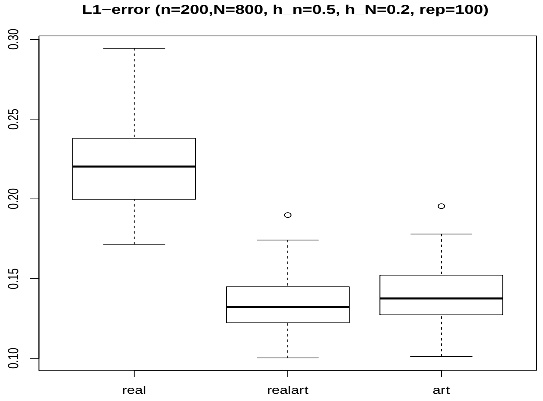

Figure 6: Boxplots in the third model

Finally we illustrate the usefulness of our estimation procedure by applying it to a density estimation problem in a simulation model. Here we consider the load distribution in the three legs of a simple tripod. More precisely, a static force is applied on the symmetric tripod to induce mechanical loading equivalent to the weight of 4,5 kg in its three legs. On the bottom side of the legs, force sensors are mounted to measure the leg's axial force. For a safe and stable standing of the tripod, the legs are angled with $\alpha = 5^\circ$ from the middle axis of the connecting devise. Engineers expect that if the holes where the legs are plugged in have a diameter of 15 mm, a third of the general load should be measured in each leg. Unfortunately, a gouching of exactly 15 mm is not possible in the manufactering process. In the simulation we assume that the diameters behave like a standard normally distributed random variable with expectation 15 and standard deviation 0.5. Based on the physical model of the tripod we are able to calculate the resulting load distribution in dependence of the three values of the diameter. Since in this case the real density is unknown, we repeat the simulation 10.000 times to generate a high sample of relative loads. For simplicity, we consider only one leg of the tripod. Application of the routine *density* in the statistics package R to these 10.000 observed values leads to the black line in figure 7. We calculate our estimates as described before using 200 real and 800 artificial data. Again the estimates which use artificial data achieve better results than the estimate with $w_n = 1$. The difference between the blue and the red line is not visible.

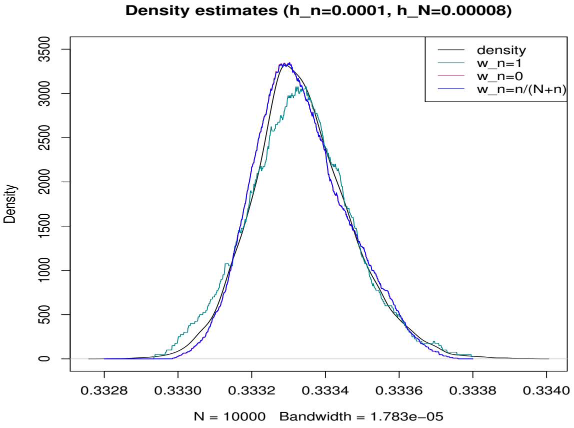

Figure 7: Density estimates in the fourth model

6 An outlook

The consistency result gives us only general guidelines for the joint choice of $\widehat{h}_N$ and m_n. The rate of convergence of $\int |g_n(y) - f(y)|\,dy$ needs to be studied in detail. This can only be attempted if we specify the data-dependent weight function w_n. It is clear that the individual rates of

$$\int |f_n(y) - f(y)|\,dy \quad \text{and} \quad \int |\widehat{f}_N(y) - f(y)|\,dy$$

will form the basis of such a study.

Help can come from $\widehat{f}_N$ only if it performs well. Our examples outlined a few possible situations. The discrete analog of the previous section shows even more dramatically the improvements one can expect if N is very large.

There is also a need to develop a suitable minimax theory for our estimation problem for appropriate subclasses of distributions of X and regression functions m. In fact, even for $N = 0$, there is no known minimax theory when m is restricted in some sense.

Even further afield, one might consider vector-valued regression functions m.

Finally, our generic estimate itself was kernel-based. It is of interest to explore a nearest-neighbor or space partitioning version for both m_n and the definition of $\widehat{f}_N$.

7 Appendix: Proof of Theorem 1

We prove only the almost sure version of the Theorem 1. The other part can be derived in the same way. Since

$$\begin{aligned}
&\int_{\mathbb{R}} |g_n(y) - f(y)|\, dy \\
&\leq w_n \cdot \int_{\mathbb{R}} |f_n(y) - f(y)|\, dy + (1 - w_n) \cdot \int_{\mathbb{R}} |\widehat{f}_N(y) - f(y)|\, dy \\
&\leq \int_{\mathbb{R}} |f_n(y) - f(y)|\, dy + \int_{\mathbb{R}} |\widehat{f}_N(y) - f(y)|\, dy,
\end{aligned}$$

it suffices to show

$$\int_{\mathbb{R}} |f_n(y) - f(y)|\, dy \to 0 \quad \text{almost surely} \tag{12}$$

and

$$\int_{\mathbb{R}} |\widehat{f}_N(y) - f(y)|\, dy \to 0 \quad \text{almost surely.} \tag{13}$$

(12) follows from Devroye (1983), so it suffices to show (13).

As in the proof of Theorem 1 in Devroye et al. (2012) we conclude from McDiarmid's inequality (cf., McDiarmid (1989)) that (13) is implied by

$$\mathbf{E}\left\{ \int_{\mathbb{R}} |\widehat{f}_N(y) - f(y)|\, dy \big| \mathcal{D}_n \right\} \to 0 \quad \text{almost surely,} \tag{14}$$

which we show in the sequel.

Set $(a)_+ = \max\{a, 0\}$ for $a \in \mathbb{R}$. By Scheffé's Lemma we know

$$\begin{aligned}
\int_{\mathbb{R}} |\widehat{f}_N(y) - f(y)|\, dy &= 2 \cdot \int_{\mathbb{R}} (f(y) - \widehat{f}_N(y))_+\, dy \\
&\leq 2 \cdot \int_B (f(y) - \widehat{f}_N(y))_+\, dy + 2 \cdot \int_{B^c} f(y)\, dy
\end{aligned}$$

for any $B \subseteq \mathbb{R}$, hence it suffices to show that we have for any compact set $B \subseteq \mathbb{R}$

$$\mathbf{E}\left\{ \int_B (f(y) - \widehat{f}_N(y))_+\, dy \big| \mathcal{D}_n \right\} \to 0 \quad \text{almost surely.} \tag{15}$$

Since

$$(a)_+ \leq |b| + (a - b)_+ \text{ for } a, b \in \mathbb{R}, \tag{16}$$

this in turn is implied by

$$\mathbf{E}\left\{\int_B \left|\widehat{f}_N(y) - \mathbf{E}\left\{\widehat{f}_N(y)|\mathcal{D}_n\right\}\right| \, dy|\mathcal{D}_n\right\} \to 0 \quad \text{almost surely} \tag{17}$$

and

$$\int_B \left(f(y) - \mathbf{E}\left\{\widehat{f}_N(y)|\mathcal{D}_n\right\}\right)_+ dy \to 0 \quad \text{almost surely.} \tag{18}$$

In the first step of the proof we show (17). By Cauchy-Schwarz and the inequality of Jensen we have

$$\begin{aligned}
&\mathbf{E}\left\{\int_B \left|\widehat{f}_N(y) - \mathbf{E}\left\{\widehat{f}_N(y)|\mathcal{D}_n\right\}\right| \, dy|\mathcal{D}_n\right\} \\
&\leq \sqrt{\int_B 1\, dy} \cdot \mathbf{E}\left\{\sqrt{\int_B \left|\widehat{f}_N(y) - \mathbf{E}\left\{\widehat{f}_N(y)|\mathcal{D}_n\right\}\right|^2 dy}\middle|\mathcal{D}_n\right\} \\
&\leq \sqrt{\int_B 1\, dy} \cdot \sqrt{\mathbf{E}\left\{\int_B \left|\widehat{f}_N(y) - \mathbf{E}\left\{\widehat{f}_N(y)|\mathcal{D}_n\right\}\right|^2 dy\middle|\mathcal{D}_n\right\}}.
\end{aligned}$$

Using the theorem of Fubini and the conditional independence of $\widehat{Y}_1, \ldots, \widehat{Y}_N$ we get

$$\begin{aligned}
&\mathbf{E}\left\{\int_B \left|\widehat{f}_N(y) - \mathbf{E}\left\{\widehat{f}_N(y)|\mathcal{D}_n\right\}\right|^2 dy|\mathcal{D}_n\right\} \\
&= \int_B \mathbf{E}\left\{\left|\widehat{f}_N(y) - \mathbf{E}\left\{\widehat{f}_N(y)|\mathcal{D}_n\right\}\right|^2 |\mathcal{D}_n\right\} dy \\
&\leq \int_B \frac{1}{N^2 \cdot \widehat{h}_N^2} \cdot \sum_{i=1}^N \mathbf{E}\left\{K^2\left(\frac{y - m_n(X_{n+i})}{\widehat{h}_N}\right)\middle|\mathcal{D}_n\right\} dy \\
&= \frac{1}{N \cdot \widehat{h}_N^2} \cdot \int_B \int K^2\left(\frac{y - m_n(z)}{\widehat{h}_N}\right) \mu(dz)\, dy \\
&= \frac{1}{N \cdot \widehat{h}_N^2} \cdot \int \int_B K^2\left(\frac{y - m_n(z)}{\widehat{h}_N}\right) dy\, \mu(dz) \\
&\leq \frac{1}{N \cdot \widehat{h}_N} \cdot \int \int_{\mathbb{R}} K^2\left(y\right) \, dy\, \mu(dz) \\
&= \frac{1}{N \cdot \widehat{h}_N} \cdot \int_{\mathbb{R}} K^2\left(y\right) \, dy \to 0 \quad (n \to \infty),
\end{aligned}$$

from which we conclude (17) via (1).

In the second step of the proof we show (18). Let $B \subseteq \mathbb{R}$ be an arbitrary compact set, let $\epsilon > 0$ be arbitrary and let $A_{n,\epsilon}$ be defined as in the theorem. Then

$$\begin{aligned}
&\int_B \left(f(y) - \mathbf{E}\left\{\widehat{f}_N(y)|\mathcal{D}_n\right\}\right)_+ dy \\
&= \int_B \left(f(y) - \int \frac{1}{\widehat{h}_N} K\left(\frac{y - m_n(x)}{\widehat{h}_N}\right) \mu(dx)\right)_+ dy
\end{aligned}$$

$$\leq \int_B \left(f(y) - \int \frac{1}{\widehat{h}_N} K\left(\frac{y - m_n(x)}{\widehat{h}_N}\right) \cdot 1_{A_{n,\epsilon}}(x)\, \mu(dx) \right)_+ dy.$$

Since K is the naive kernel we have for any $x \in A_{n,\epsilon}$ in case that $\widehat{h}_N > \|m_n - m\|_{\infty, A_{n,\epsilon}}$

$$\begin{aligned}
& K\left(\frac{y - m(x)}{\widehat{h}_N - \|m_n - m\|_{\infty,A_{n,\epsilon}}}\right) = \frac{1}{2} \\
\Leftrightarrow \quad & m(x) - \widehat{h}_N + \|m_n - m\|_{\infty,A_{n,\epsilon}} \leq y \leq m(x) + \widehat{h}_N - \|m_n - m\|_{\infty,A_{n,\epsilon}} \\
\Rightarrow \quad & m(x) - \widehat{h}_N + (m_n(x) - m(x)) \leq y \leq m(x) + \widehat{h}_N - (m(x) - m_n(x)) \\
\Leftrightarrow \quad & m_n(x) - \widehat{h}_N \leq y \leq m_n(x) + \widehat{h}_N \\
\Leftrightarrow \quad & K\left(\frac{y - m_n(x)}{\widehat{h}_N}\right) = \frac{1}{2}
\end{aligned}$$

which implies

$$K\left(\frac{y - m_n(x)}{\widehat{h}_N}\right) \quad \geq \quad K\left(\frac{y - m(x)}{\widehat{h}_N - \|m_n - m\|_{\infty,A_{n,\epsilon}}}\right).$$

Using this and (16) we conclude that we have in case $\widehat{h}_N > \|m_n - m\|_{\infty,A_{n,\epsilon}}$ (which happens for n sufficiently large with probability one by assumption (5))

$$\begin{aligned}
& \int_B \left(f(y) - \int \frac{1}{\widehat{h}_N} K\left(\frac{y - m_n(x)}{\widehat{h}_N}\right) \cdot 1_{A_{n,\epsilon}}(x)\, \mu(dx) \right)_+ dy \\
& \leq \int_B \left(f(y) - \int \frac{1}{\widehat{h}_N} K\left(\frac{y - m(x)}{\widehat{h}_N - \|m_n - m\|_{\infty,A_{n,\epsilon}}}\right) \cdot 1_{A_{n,\epsilon}}(x)\, \mu(dx) \right)_+ dy \\
& \leq \int_B \left(f(y) - \int \frac{1}{\widehat{h}_N} K\left(\frac{y - m(x)}{\widehat{h}_N - \|m_n - m\|_{\infty,A_{n,\epsilon}}}\right) \mu(dx) \right)_+ dy \\
& \quad + \int_B \int \frac{1}{\widehat{h}_N} K\left(\frac{y - m(x)}{\widehat{h}_N - \|m_n - m\|_{\infty,A_{n,\epsilon}}}\right) \cdot 1_{A^c_{n,\epsilon}}(x)\, \mu(dx)\, dy \\
& \leq \int_B \left(f(y) - \int_{\mathbb{R}} \frac{1}{\widehat{h}_N} K\left(\frac{y - z}{\widehat{h}_N - \|m_n - m\|_{\infty,A_{n,\epsilon}}}\right) \cdot f(z)\, dz \right)_+ dy \\
& \quad + \frac{\widehat{h}_N - \|m_n - m\|_{\infty,A_{n,\epsilon}}}{\widehat{h}_N} \cdot \int_{\mathbb{R}} K(y)\ dy \int 1_{A^c_{n,\epsilon}}(x)\, \mu(dx).
\end{aligned}$$

By Lebesgue's density theorem (cf., e.g., Theorem 2 or Theorem 3 in Devroye and Györfi (1985)) and (1) and (5) we know that

$$\int_{\mathbb{R}} \frac{1}{\widehat{h}_N} K\left(\frac{y - z}{\widehat{h}_N - \|m_n - m\|_{\infty,A_{n,\epsilon}}}\right) \cdot f(z)\, dz$$

$$
\begin{aligned}
&= \frac{\widehat{h}_N - \|m_n - m\|_{\infty,A_{n,\epsilon}}}{\widehat{h}_N} \cdot \int_{\mathbb{R}} \frac{1}{\widehat{h}_N - \|m_n - m\|_{\infty,A_{n,\epsilon}}} K\left(\frac{y-z}{\widehat{h}_N - \|m_n - m\|_{\infty,A_{n,\epsilon}}}\right) \cdot f(z)\, dz \\
&\to 1 \cdot f(y) = f(y) \quad (n \to \infty)
\end{aligned}
$$

for almost all $y \in \mathbb{R}$ with probability one, which implies (via the dominated convergence theorem)

$$
\int_B \left(f(y) - \int_{\mathbb{R}} \frac{1}{\widehat{h}_N} K\left(\frac{y-z}{\widehat{h}_N - \|m_n - m\|_{\infty,A_{n,\epsilon}}}\right) \cdot f(z)\, dz \right)_+ dy \to 0 \quad \text{almost surely.}
$$

Furthermore,

$$
\frac{\widehat{h}_N - \|m_n - m\|_{\infty,A_{n,\epsilon}}}{\widehat{h}_N} \cdot \int_{\mathbb{R}} K(y)\ dy \int 1_{A^c_{n,\epsilon}}(x)\, \mu(dx) \leq 1 \cdot 1 \cdot \mu(A^c_{n,\epsilon}).
$$

Summarizing the above result we get

$$
\limsup_{n\to\infty} \int_B \left(f(y) - \mathbf{E}\left\{ \widehat{f}_N(y) | \mathcal{D}_n \right\} \right)_+ dy \leq \epsilon \quad \text{almost surely,}
$$

and with $\epsilon \to 0$ this implies (18). The proof is complete. □

8 Acknowledgment

The authors would like to thank the German Research Foundation (DFG) for founding this project within the Collaborative Research Center 805.

References

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

[2] Beirlant, J. and Györfi, L. (1998). On the asymptotic L_2-error in partitioning regression estimation. *Journal of Statistical Planning and Inference*, **71**, pp. 93–107.

[3] Castelli, V. and Cover, T. (1996). The relative value of labeled and unlabeld samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on Information Theory*, **42**, pp. 2101–2117.

[4] Chapelle, O., Schölkopf, B. and Zien, A. (2006). Semi-Supervised Learning. *MIT Press,* Cambridge.

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

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

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

[8] Devroye, L. (1987). A Course in Density Estimation. *Birkhäuser*, Basel.

[9] Devroye, L., Felber, T., Kohler, M., and Krzyżak, A. (2012). L_1-consistent estimation of the density of residuals in random design regression models. *Statistics and Probability Letters*, **82**, pp. 173-179.

[10] Devroye, L. and Györfi, L. (1985). Nonparametric Density Estimation. The L1 view. *Wiley Series in Probability and Mathematical Statistics: Tracts on Probability and Statistics. JohnWiley and Sons*, New York.

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

[12] Devroye, L. and Lugosi, G. (2001). Combinatorial Methods in Density Estimation. *Springer-Verlag*, New York.

[13] Devroye, L. and Krzyżak, A. (1989). An equivalence theorem for L_1 convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference*, **23**, pp. 71–82.

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

[15] Efromovich, S. (2005). Estimation of the density of regression errors. *Annals of Statistics*, **33**, pp. 2194–2227.

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

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

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

[19] Kohler, M. (2000). Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. *Journal of Statistical Planning and Inference* **89**, pp. 1-23.

[20] Kohler, M. and Krzyżak, A. (2001). Nonparametric regression estimation using penalized least squares. *IEEE Transactions on Information Theory*, **47**, pp. 3054–3058.

[21] Lugosi, G. and Zeger, K. (1995). Nonparametric estimation via empirical risk minimization. *IEEE Trans. Inform. Theory* **41** 677-687.

[22] Mnatsakanov, R. M., and Khmaladze, E. V. (1981). On L_1-convergence of statistical kernel estimators of distribution densities. *Soviet Mathematics Doklady* **23**, pp. 633-636.

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

[24] Nadaraya, E. A. (1964) On estimating regression. *Theory of Probability and its Applications*, **9**, 141–142.

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

[26] Parzen, E. (1962). On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, **33**, pp. 1065–1076.

[27] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, pp. 832–837.

[28] Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statististics*, **5**, pp. 595–645.

[29] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, **10**, pp. 1040–1053.

[30] Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

[31] Watson, G. S. (1964). Smooth regression analysis. *Sankhya Series A*, **26**, pp. 359–372.