Estimation of a density in a simulation model [*]

Ann-Kathrin Bott, Tina Felber and Michael Kohler[†]
Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: abott@mathematik.tu-darmstadt.de, tfelber@mathematik.tu-darmstadt.de, kohler@mathematik.tu-darmstadt.de

June 17, 2013

Abstract
The problem of estimating a density in a simulation model is considered, where given a value of an $\mathbb{R}^d$-valued random input parameter X the value of a real-valued random variable $Y = m(X)$ is computed. Here $m : \mathbb{R}^d \to \mathbb{R}$ is a function which measures the quality $m(X)$ of a technical system with input X. It is assumed that both X and Y have densities. Given a sample of (X, Y) the task is to estimate the density of Y. We estimate in a first step m and the density of X, and by using these estimates we compute in a second step an estimate of the density of Y. Results concerning L_1-consistency and rate of convergence of the estimates are proven and the estimates are illustrated by applying them to simulated and real data.

AMS classification: Primary 62G07; secondary 62G20.

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

1 Introduction

We consider a simulation model of a technical system, which computes for an $\mathbb{R}^d$-valued random variable X the quality $Y = m(X)$ of a corresponding technical system. The simulation model is described by the (measurable) function $m : \mathbb{R}^d \to \mathbb{R}$. We assume that we can observe a sample of the input parameter X and the corresponding values of Y, and we are interested in the distribution of the (random) quality $Y = m(X)$. This distribution is described by its density g, which we assume to exist. By controlling the L_1-error of an estimate of this density we can bound via the Lemma of Scheffé (cf., e.g., Devroye and Györfi (1985)) the total variation error of the corresponding estimate of the distribution. So given a sample $X_1, \ldots, X_n$ of X, we compute $Y_1 = m(X_1), \ldots, Y_n = m(X_n)$ and our aim is to use the data

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

[*] Running title: *Density estimation in a simulation model*
[†] Corresponding author: Tel. +49-6151-16-6846, Fax. +49-6151-16-6822

to construct an estimate $g_n(\cdot) = g_n(\cdot, \mathcal{D}_n) : \mathbb{R} \to \mathbb{R}$ of g such that its L_1-error

$$\int |g_n(y) - g(y)| \, dy$$

is "small".

The easiest way of doing this is to ignore the x-values of the data completely and to estimate the density g of Y using only $Y_1, \ldots, Y_n$. Here we can use, e.g., the famous kernel density estimate (Rosenblatt (1956), Parzen (1962)) defined by

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

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

However, the interesting question is whether we can get better estimates (e.g., estimates with better performance for simulated data or with a better rate of convergence of the L_1-error under appropriate smoothness assumptions on g) by using the values of X, too. In case that additional independent observations of X are available a positive answer to this question was given in Devroye, Felber and Kohler (2013) and Kohler and Krzyżak (2012). The estimates there can be applied in our current situation, too, provided we ignore a part of the values of Y, i.e., we choose $1 < n_l < n$ (e.g., $n_l \approx n/2$), define an estimate m_{n_l} of m depending only on the first n_l data points $\mathcal{D}_{n_l}$ and estimate g by

$$\bar{g}_n(y) = \frac{1}{(n - n_l) \cdot \hat{h}_n} \cdot \sum_{i=n_l+1}^{n} K\left(\frac{y - m_{n_l}(X_i)}{\hat{h}_n}\right).$$

The estimate m_{n_l} of m can be chosen as a suitable regression estimate of m in a regression problem without noise in the dependent variable. For instance, we can use kernel regression estimates (cf., e.g., Nadaraya (1964, 1970), Watson (1964), Devroye and Wagner (1980), Stone (1977, 1982) or Devroye and Krzyżak (1989)), partitioning regression estimates (cf., e.g., Györfi (1981) or Beirlant and Györfi (1998)), nearest neighbor regression estimates (cf., e.g., Devroye (1982) or Devroye et al. (1994)), least squares estimates (cf., e.g., Lugosi and Zeger (1995) or Kohler (2000)) or smoothing spline estimates (cf., e.g., Wahba (1990) or Kohler and Krzyżak (2001)). Due to the fact that the dependent variable is observable without noise, here suitable interpolation estimates can achieve even better rates, especially if the distribution of X satisfies regularity assumptions, cf. Kohler and Krzyżak (2013).

In this paper we use a slightly different approach in the sense that we assume that a density f of X exists, too. By the Lebesgue density theorem (cf, e.g., Devroye and Györfi (1985), Chapter 2, Theorem 1) we know that

$$g(y) \approx \frac{\int K\left(\frac{y-z}{h}\right) \cdot g(z) \, dz}{h} = \frac{1}{h} \cdot \mathbf{E}\left\{K\left(\frac{y - Y}{h}\right)\right\} = \frac{1}{h} \cdot \int K\left(\frac{y - m(x)}{h}\right) \cdot f(x) \, dx$$

for h small. We estimate the latter term by plugging in an estimate m_n of m (based on $\mathcal{D}_n$) and an estimate f_n of the density f of X (based on the values of X in $\mathcal{D}_n$) and use

$$g_n(y) = \frac{1}{h_n} \cdot \int K\left(\frac{y - m_n(x)}{h_n}\right) \cdot f_n(x)\, dx. \tag{2}$$

Our main result states that this estimate is L_1-consistent in case that the bandwidth h_n tends to zero slower than the L_1-error of the estimate m_n of m and in case that the estimate f_n of f is L_1-consistent. Furthermore, we analyze the rate of convergence of our newly proposed estimate and identify situations where it is better than the rate of convergence of the estimate (1) which ignores the values of X completely. Finally, we illustrate our estimate by applying it to simulated and real data.

The outline of this paper is as follows: In Section 2 we give the precise definition of our estimate and present the main results concerning consistency and rate of convergence. Application of the estimate to simulated and real data is contained in Section 3. The proofs are given in Section 4.

2 Main results

In this section the estimate

$$g_n(y) = \frac{1}{h_n} \cdot \int K\left(\frac{y - m_n(x)}{h_n}\right) \cdot f_n(x)\, dx$$

is analyzed, where $h_n > 0$ is the bandwidth and the kernel function K is a symmetric and bounded density which is monotonically decreasing on $\mathbb{R}_+$ (e.g., the naive kernel $K(u) = \frac{1}{2} \cdot \mathbb{1}_{[-1,1]}(u)$), $m_n(\cdot) = m_n(\cdot; (X_1, Y_1), \dots, (X_n, Y_n)) : \mathbb{R}^d \to \mathbb{R}$ is a suitable estimate of m based on the sample $\mathcal{D}_n$ of (X, Y) and

$$f_n(x) = \frac{1}{n \cdot \hat{h}_n^d} \cdot \sum_{i=1}^{n} L\left(\frac{x - X_i}{\hat{h}_n}\right)$$

is the standard kernel density estimate of the density f of X based on the x-values X_1, …, X_n of the sample $\mathcal{D}_n$ with kernel $L : \mathbb{R}^d \to \mathbb{R}_+$ and bandwidth $\hat{h}_n > 0$.
In our first theorem we present sufficient conditions for consistency of g_n.

Theorem 1 *Let (X, Y), (X_1, Y_1), … be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$-valued random variables. Assume that densities (with respect to the Lebesgue-Borel-measure) f and g of X and Y exist. Let the kernel function K be a symmetric and bounded density which is monotonically decreasing on $\mathbb{R}_+$, let the kernel function L : $\mathbb{R}^d \to \mathbb{R}_+$ be a a density, let m_n be an estimate of m based on $\mathcal{D}_n$ and let the estimate g_n of g be defined as above. Assume that*

$$\hat{h}_n \to 0 \quad (n \to \infty), \quad n \cdot \hat{h}_n^d \to \infty \quad (n \to \infty), \tag{3}$$

$$h_n \to 0 \quad (n \to \infty) \tag{4}$$

and

$$\frac{\mathbf{E}\{|m_n(X) - m(X)|\}}{h_n} \to 0 \quad (n \to \infty). \tag{5}$$

Then g_n is weakly consistent, i.e.,

$$\mathbf{E} \int |g_n(y) - g(y)| \, dy \to 0 \quad (n \to \infty). \tag{6}$$

If (3), (4) and, in addition,

$$\frac{\int |m_n(x) - m(x)| \mathbf{P}_X(dx)}{h_n} \to 0 \quad a.s.$$

hold, then g_n is strongly consistent, i.e.,

$$\int |g_n(y) - g(y)| \, dy \to 0 \quad a.s. \tag{7}$$

Remark 1. Condition (5) requires that the expected L_1-error of the estimate m_n converges to zero faster than the bandwidth h_n of g_n. Under appropriate smoothness assumption on m, rate of convergence results for appropriate estimates m_n have been derived in Kohler and Krzyżak (2013). E.g., let m_n be the 1-nearest-neighbor estimate defined by $m_n(x) = Y_{(1)}(x)$ where $(X_{(1)}(x), Y_{(1)}(x)), \dots, (X_{(n)}(x), Y_{(n)}(x))$ is a permutation of $\mathcal{D}_n$ satisfying

$$\|X_{(1)}(x) - x\| \leq \dots \leq \|X_{(n)}(x) - x\|.$$

Then Theorem 1 and Remark 2 in Kohler and Krzyżak (2013) imply that in case $supp(X)$ bounded and m Hölder continuous with exponent $p \leq 1$ condition (5) is satisfied if

$$n^p \cdot h_n^d \to \infty \quad (n \to \infty).$$

Next, we study the rate of convergence of our estimate. Here it is well-known (cf., e.g., Devroye and Györfi (1985)) that smoothness assumptions on g are necessary to derive nontrivial results.

For $p \in (0, 1]$ and $C > 0$ we call a function $h : \mathbb{R}^d \to \mathbb{R}$ (p, C)-smooth if

$$|h(x) - h(z)| \leq C \cdot \|x - z\|^p \quad \text{for all } x, z \in \mathbb{R}^d,$$

i.e., if h is Hölder-continuous with exponent p and Hölder-constant C.

Theorem 2 *Let (X, Y), (X_1, Y_1), ... be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$-valued random variables. Assume that densities (with respect to the Lebesgue-Borel-measure) f and g of X and Y exist. Let the kernel function K be a symmetric and bounded density with bounded support which is monotonically decreasing on $\mathbb{R}_+$, let $L : \mathbb{R}^d \to \mathbb{R}_+$ be a kernel function which is a density, let m_n be an estimate of m based on $\mathcal{D}_n$ and let the estimate g_n of g be defined as above.*
If g is (r, C)-smooth with bounded support, then

$$\mathbf{E} \int |g_n(y) - g(y)| \, dy \leq \mathbf{E} \int |f_n(x) - f(x)| \, dx + 2 \cdot K(0) \cdot \frac{\mathbf{E}\left\{|m_n(X) - m(X)|\right\}}{h_n} + c \cdot h_n^r.$$

From Theorem 2 we get rate of convergence results for g_n provided we make assumptions on the smoothness of m and f and choose appropriate estimates m_n and f_n.

Corollary 1 *Assume that the assumptions of Theorem 2 hold. Let $p, r, s \in (0, 1]$, $C_1, C_2, C_3 > 0$ and assume that f is (p, C_1)-smooth with bounded support, g is (r, C_2)-smooth with bounded support and m is (s, C_3)-smooth. Let $c_1, c_2 > 0$, let f_n be the kernel density estimate with a density with compact support as kernel and with bandwidth*

$$\hat{h}_n = c_1 \cdot n^{-1/(2p+d)},$$

let m_n be the 1-nearest-neighbor estimate and define the estimate g_n as above with bandwidth

$$h_n = c_2 \cdot n^{-s/(d\cdot(r+1))}.$$

Then

$$\mathbf{E} \int |g_n(y) - g(y)| \, dy = O\left(n^{-\min\left\{ \frac{p}{2p+d}, \frac{r \cdot s}{d \cdot (1+r)} \right\}} \right).$$

Proof. By Theorem 1 and Remark 1 in Kohler and Krzyżak (2013) we have

$$\mathbf{E} \left\{ |m_n(X) - m(X)| \right\} \leq n^{-\frac{s}{d}},$$

furthermore

$$\begin{aligned} \mathbf{E} \int |f_n(x) - f(x)| \, dx &\leq \int \mathbf{E}\{|f_n(x) - \mathbf{E} f_n(x)|\} \, dx + \int |\mathbf{E} f_n(x) - f(x)| \, dx \\ &\leq c_3 \cdot \frac{1}{n \cdot h_n^d} + c_4 \cdot h_n^p \qquad (8) \end{aligned}$$

(cf., e.g., proof of Theorem 9.5 in Devroye and Lugosi (2001) and proof of Theorem 2 below), which implies

$$\mathbf{E} \int |f_n(x) - f(x)| \, dx \leq c_5 \cdot n^{-p/(2p+d)}.$$

Application of Theorem 2 yields the assertion. □

Remark 2. If g is Hölder continuous with exponent r and g has compact support then the estimate (1) with bandwidth $h_n = c_6 \cdot n^{-1/(2r+1)}$ satisfies

$$\mathbf{E} \int |\hat{g}_n(y) - g(y)| \, dy \leq c_7 \cdot n^{-r/(2r+1)}$$

(cf., e.g., proof of (8)). Comparing this with the above results we see that the estimate g_n achieves a better rate of convergence e.g. if

$$d = 1, \quad p > r \quad \text{and} \quad s > \frac{r+1}{2r+1}.$$

Remark 3. Any application of the estimate to simulated or real data requires a data dependent choice of the smoothing parameters (here, bandwidths h_n and $\hat{h}_n$ and smoothing parameter of m_n) of the estimate. For choosing the parameter of f_n we can use the combinatorial method described in Devroye and Lugosi (2001), and the parameter of m_n can be chosen by cross-validation (cf., e.g., Chapter 8 in Györfi et al. (2002)). For the choice of the bandwidth h_n in the definition of g_n we can use an adaptation of the combinatorial method of Devroye and Lugosi (2001) similar to the one described in Kohler and Krzyżak (2012).

3 Application to simulated and real data

In this section we illustrate the finite sample size performance of our density estimate by applying it to simulated and real data.

In our first example we set $X = (X_1, X_2)$ with independent standard normally distributed random variables X_1 and X_2 and choose $m(x_1, x_2) = 2 \cdot x_1 + x_2 + 2$. In this case $Y = m(X)$ 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 using 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 estimate f_n we use the standard kernel density estimate of Rosenblatt and Parzen with Gaussian kernel L and bandwidth $\hat{h}_n$ chosen by cross-validation. For the kernel function K we use the naive kernel. The bandwidth h_n of g_n is chosen by minimizing the L_1-errors of the estimate via comparing the estimated density with the underlying density. So we assume that we have available an oracle which chooses the optimal bandwidth, such that we can ignore effects occuring due to inproper choice of bandwidths. We set the sample size $n = 200$ and compare the estimate with the standard kernel density estimate of Rosenblatt and Parzen. Figure 1 shows both estimates and the underlying density.

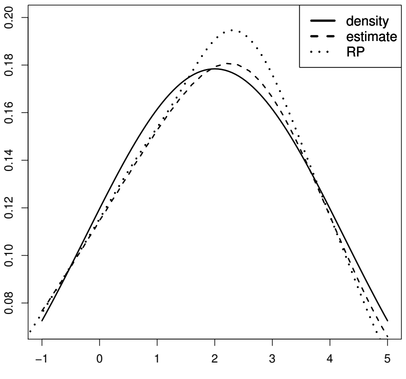

Figure 1: Typical simulation in the first model

Since the result of our simulation depends on the randomly occuring data points, we repeat the whole procedure 100 times with independent realizations of the occuring random variables and report boxplots of the L_1-errors in Figure 2. The mean of the L_1-errors of the proposed estimate (0.0651) is less than the mean L_1-error of the Rosenblatt-Parzen density estimate (0.0723).

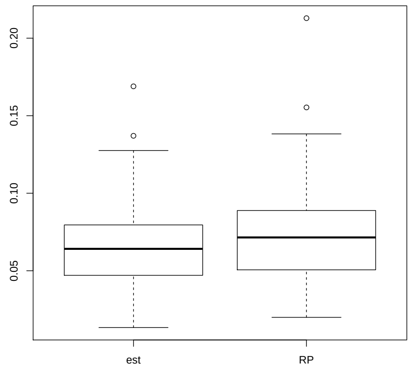

Figure 2: Boxplots of the occuring L_1-errors in the first model

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 $m(x_1, x_2) = x_1^2 + x_2^2$. Then $Y = m(X)$ is chi-squared distributed with two degrees of freedom. We define the estimate as in the first example and choose again $n = 200$. Figure 3 shows the estimate g_n, the Rosenblatt-Parzen density estimate and the underlying density in a typical simulation.

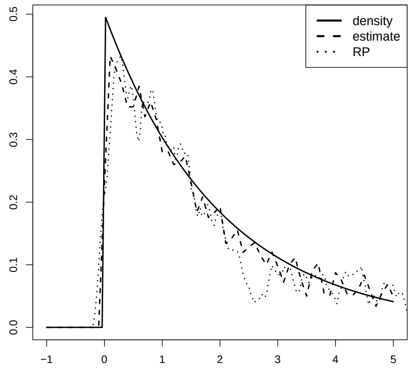

Figure 3: Typical simulation in the second model

In Figure 4 we compare boxplots of the occuring L_1-errors of the two estimates. The mean L_1-error of our estimate (0.1200) is much lower than the one of Rosenblatt and Parzen (0.2042).

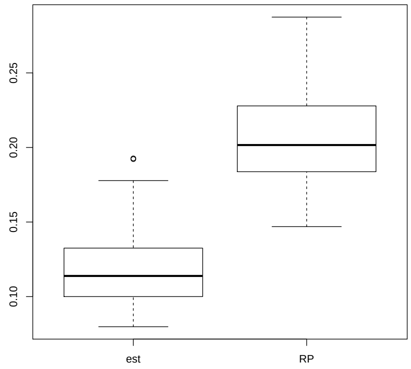

Figure 4: Boxplots of the occuring L_1-errors in the second model

In Figure 5 and Figure 6 we repeat the same simulation with $n = 200$ choosing X as a standard normally distributed random variable and $m(x) = exp(x)$. In this case $Y = m(X)$ is log-normally distributed. The mean of the L_1-errors of the estimate g_n is again much lower (0.0706) than the mean error of the Rosenblatt-Parzen estimate (0.1019).

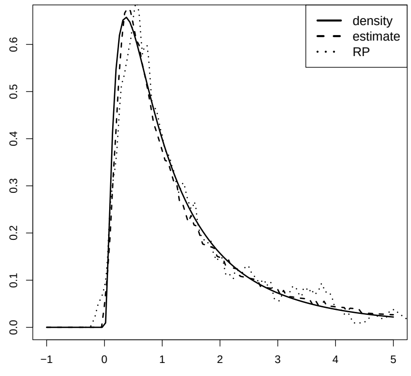

Figure 5: Typical simulation in the third model

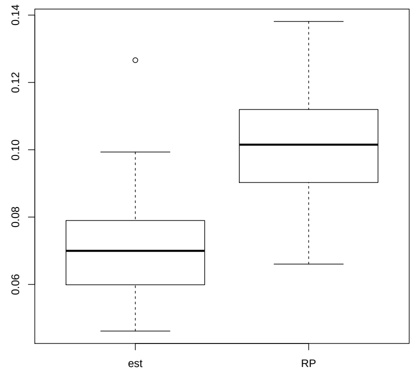

Figure 6: Boxplot of the occuring L_1-errors in the third model

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 $\alpha = 5°$ 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 independent standard normally distributed random variable with expectation 15 and standard deviation 0.5.

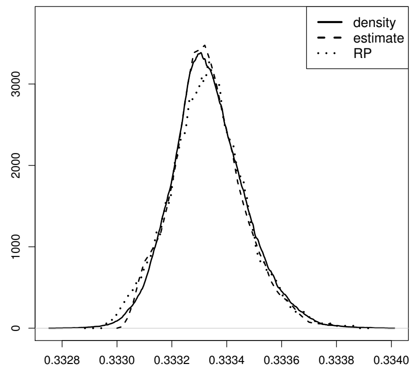

Figure 7: Density estimation in a simulation model

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. For simplicity, we consider only one leg of the tripod. Since in this case the real density is unknown, we repeat the simulation 10.000 times to generate a high sample of relative loads. Application of the routine *density* in the statistics package *R* to these 10.000 observed values leads to the solid line in Figure 7. We calculate our estimates as described before assuming that $n = 200$ measurements are available. Again, the newly proposed estimate is printed by the dashed line, and the dotted line represents the estimate of Rosenblatt and Parzen. Similary as before, the run of the curve of our estimate lies much closer to the solid line than the one of Rosenblatt and Parzen.

4 Proofs

4.1 Auxiliary results

In this subsection we prove two auxiliary results, from which we will conclude Theorem 1 and Theorem 2 easily.

Lemma 1 *Let the kernel function K be a symmetric, bounded density which is monotonically decreasing on $\mathbb{R}_+$. Then it holds*

$$\int \left| K\left(\frac{y-z_1}{h_n}\right) - K\left(\frac{y-z_2}{h_n}\right) \right| dy \leq 2 \cdot K(0) \cdot |z_1 - z_2|$$

for arbitrary $z_1, z_2 \in \mathbb{R}$.

Proof of Lemma 1. Without loss of generality we assume $z_1 \leq z_2$ and set

$$\bar{K}(y) := \left| K\left(\frac{y-z_1}{h_n}\right) - K\left(\frac{y-z_2}{h_n}\right) \right|.$$

This function is axial symmetric to $x = \frac{z_1+z_2}{2}$, because for all $t \in \mathbb{R}$ we can conclude from $K(u) = K(-u)$ $(u \in \mathbb{R})$

$$\begin{aligned}
\bar{K}\left(\frac{z_1+z_2}{2} + t\right) &= \left| K\left(\frac{z_2 - z_1 + 2t}{2h_n}\right) - K\left(\frac{z_1 - z_2 + 2t}{2h_n}\right) \right| \\
&= \left| K\left(\frac{z_1 - z_2 - 2t}{2h_n}\right) - K\left(\frac{z_2 - z_1 - 2t}{2h_n}\right) \right| \\
&= \left| K\left(\frac{z_2 - z_1 - 2t}{2h_n}\right) - K\left(\frac{z_1 - z_2 - 2t}{2h_n}\right) \right| \\
&= \bar{K}\left(\frac{z_1+z_2}{2} - t\right).
\end{aligned}$$

With the assumption $z_1 \leq z_2$ we can conclude that

$$|y - z_1| \leq |y - z_2| \quad \text{for all} \quad y \leq \frac{z_1 + z_2}{2}, \tag{9}$$

because in case of $y \leq z_1$ we have

$$|y - z_1| \;=\; z_1 - y \;\leq\; z_2 - y \;=\; |y - z_2|,$$

and in case of $z_1 < y \leq \frac{z_1+z_2}{2}$ we get

$$|y - z_1| \;=\; y - z_1 \;\leq\; \frac{z_1 + z_2}{2} - z_1 \;=\; z_2 - \frac{z_1 + z_2}{2} \;\leq\; z_2 - y = |y - z_2|.$$

Since the symmetric kernel K is monotonically decreasing on $\mathbb{R}_+$, we obtain by (9)

$$K\left(\frac{y - z_1}{h_n}\right) \geq K\left(\frac{y - z_2}{h_n}\right) \quad \text{for all } y \leq \frac{z_1 + z_2}{2}. \tag{10}$$

Applying the observation about the symmetry, we can conclude

$$\begin{aligned}
& \int \left| K\left(\frac{y - z_1}{h_n}\right) - K\left(\frac{y - z_2}{h_n}\right) \right| dy \\
= \; & \int \bar{K}(y)\, dy \\
= \; & 2 \int_{-\infty}^{\frac{z_1+z_2}{2}} \bar{K}(y)\, dy \\
\overset{(10)}{=} \; & 2 \left[\int_{-\infty}^{\frac{z_1+z_2}{2}} K\left(\frac{y - z_1}{h_n}\right) dy - \int_{-\infty}^{\frac{z_1+z_2}{2}} K\left(\frac{y - z_2}{h_n}\right) \right] dy \\
= \; & 2 \cdot h_n \left[\int_{-\infty}^{\frac{z_2-z_1}{2h_n}} K\left(u\right) du - \int_{-\infty}^{\frac{z_1-z_2}{2h_n}} K\left(v\right) dv \right] \\
= \; & 2 \cdot h_n \int_{\frac{z_1-z_2}{2h_n}}^{\frac{z_2-z_1}{2h_n}} K\left(y\right) dy
\end{aligned}$$

Because K is symmetric and monotonically decreasing on $\mathbb{R}_+$, K(0) is the maximum value of K and thus we get

$$\begin{aligned}
\int \left| K\left(\frac{y - z_1}{h_n}\right) - K\left(\frac{y - z_2}{h_n}\right) \right| dy \;&\leq\; 2 \cdot h_n \cdot K(0) \cdot \left(\frac{z_2 - z_1}{2h_n} - \frac{z_1 - z_2}{2h_n} \right) \\
&=\; 2 \cdot K(0) \cdot |z_2 - z_1|
\end{aligned}$$

The proof is complete. □

Lemma 2 *Let (X, Y), (X_1, Y_1), ... be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$-valued random variables. Assume that densities (with respect ot the Lebesgue-Borel-measure) f and g of X and Y exist. Let K be a symmetric, bounded density, which is monotonically decreasing on $\mathbb{R}_+$, let f_n be an estimate of f, let m_n be an estimate of m and let the estimate g_n of g be defined as in Section 2. Then*

$$\int |g_n(y) - g(y)|\, dy$$

$$\leq \int |f_n(x) - f(x)| \, dx + 2 \cdot K(0) \cdot \frac{\int |m_n(x) - m(x)| \, \mathbf{P}_X(dx)}{h_n}$$
$$+ \int \left| \int \frac{1}{h_n} K\left(\frac{y-z}{h_n}\right) \cdot g(z) \, dz - g(y) \right| \, dy.$$

Proof of Lemma 2. We use the error decomposition

$$\begin{aligned}
& \int |g_n(y) - g(y)| \, dy \\
\leq \; & \int \left| \frac{1}{h_n} \cdot \int K\left(\frac{y - m_n(x)}{h_n}\right) \cdot f_n(x) \, dx - \frac{1}{h_n} \cdot \int K\left(\frac{y - m_n(x)}{h_n}\right) \cdot f(x) \, dx \right| \, dy \\
& + \int \left| \frac{1}{h_n} \cdot \int K\left(\frac{y - m_n(x)}{h_n}\right) \cdot f(x) \, dx - \frac{1}{h_n} \cdot \int K\left(\frac{y - m(x)}{h_n}\right) \cdot f(x) \, dx \right| \, dy \\
& + \int \left| \frac{1}{h_n} \cdot \int K\left(\frac{y - m(x)}{h_n}\right) \cdot f(x) \, dx - g(y) \right| \, dy \\
=: \; & T_{1,n} + T_{2,n} + T_{3,n}.
\end{aligned}$$

Application of the theorem of Fubini yields

$$\begin{aligned}
T_{1,n} & \leq \int \int \frac{1}{h_n} \cdot K\left(\frac{y - m_n(x)}{h_n}\right) \cdot |f_n(x) - f(x)| \, dx \, dy \\
& = \int |f_n(x) - f(x)| \cdot \int \frac{1}{h_n} \cdot K\left(\frac{y - m_n(x)}{h_n}\right) \, dy \, dx \\
& = \int |f_n(x) - f(x)| \, dx,
\end{aligned}$$

where the last equality follows from the fact that K is a density.
Next we bound $T_{2,n}$. Using again the theorem of Fubini and Lemma 1 we conclude

$$\begin{aligned}
T_{2,n} & \leq \frac{1}{h_n} \cdot \int \int \left| K\left(\frac{y - m_n(x)}{h_n}\right) - K\left(\frac{y - m(x)}{h_n}\right) \right| \cdot f(x) \, dx \, dy \\
& = \frac{1}{h_n} \cdot \int \int \left| K\left(\frac{y - m_n(x)}{h_n}\right) - K\left(\frac{y - m(x)}{h_n}\right) \right| \, dy \cdot f(x) \, dx \\
& \leq \frac{1}{h_n} \cdot \int 2 \cdot K(0) \cdot |m_n(x) - m(x)| \cdot f(x) \, dx \\
& = 2 \cdot K(0) \cdot \frac{\int |m_n(x) - m(x)| \, \mathbf{P}_X(dx)}{h_n}.
\end{aligned}$$

Finally, we bound $T_{3,n}$. Because X has density f and $Y = m(X)$ has density g we have

$$\begin{aligned}
T_{3,n} & = \int \left| \mathbf{E} \left\{ \frac{1}{h_n} K\left(\frac{y - m(X)}{h_n}\right) \right\} - g(y) \right| \, dy \\
& = \int \left| \int \frac{1}{h_n} K\left(\frac{y-z}{h_n}\right) \cdot g(z) \, dz - g(y) \right| \, dy.
\end{aligned}$$

The proof is complete. □

4.2 Proof of Theorem 1

By Lemma 2 we have

$$
\begin{aligned}
\mathbf{E} \int |g_n(y) - g(y)| \, dy \quad &\leq \quad \mathbf{E} \int |f_n(x) - f(x)| \, dx + 2 \cdot K(0) \cdot \frac{\mathbf{E}\left\{|m_n(X) - m(X)|\right\}}{h_n} \\
&\quad + \int \left| \int \frac{1}{h_n} K\left(\frac{y-z}{h_n}\right) \cdot g(z) \, dz - g(y) \right| dy \\
&=: \quad T_{4,n} + T_{5,n} + T_{6,n}. \qquad (11)
\end{aligned}
$$

Because of (3) we can apply Theorem 1, Chapter 3, in Devroye and Györfi (1985) which implies

$$T_{4,n} \to 0 \quad (n \to \infty).$$

By (5) we get

$$T_{5,n} \to 0 \quad (n \to \infty).$$

Finally, using (4) and Theorem 1, Chapter 2, in Devroye and Györfi (1985) we get

$$T_{6,n} \to 0 \quad (n \to \infty).$$

This proves (6), assertion (7) follows in the same way. □

4.3 Proof of Theorem 2

Since $\int |g_n(y) - g(y)| \, dy \leq 2$ we can assume w.l.o.g. that h_n is bounded. Since g and K have compact support we can choose a bounded set S such that

$$g(y) = 0 \quad \text{and} \quad \int K\left(\frac{y-z}{h_n}\right) g(z) \, dz = 0$$

for $y \notin S$. Since g is Hölder continuous with exponent r we have

$$
\begin{aligned}
&\int \left| \int \frac{1}{h_n} K\left(\frac{y-z}{h_n}\right) \cdot g(z) \, dz - g(y) \right| dy \\
&= \int_S \left| \int \frac{1}{h_n} K\left(\frac{y-z}{h_n}\right) \cdot (g(z) - g(y)) \, dz \right| dy \\
&\leq \int_S \int \frac{1}{h_n} K\left(\frac{y-z}{h_n}\right) \cdot C \cdot |z-y|^r \, dz \, dy \\
&= C \cdot h_n^r \cdot \int |u|^r K(u) \, du \cdot \int_S 1 \, dy \\
&\leq c \cdot h_n^r.
\end{aligned}
$$

This together with (11) implies the assertion. □

5 Acknowledgment

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

References

[1] 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.

[2] 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.

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

[4] Devroye, L., Felber, T., and Kohler, M. (2013). Estimation of a density using real and artificial data. *IEEE Transactions on Information Theory*, **59**, pp. 1917–1928.

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

[6] 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.

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

[8] 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.

[9] 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.

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

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

[12] 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.

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

[14] Kohler, M. and Krzyżak, A. (2012). Adaptive density estimation based on real and artificial data. Submitted for publication.

[15] Kohler, M. and Krzyżak, A. (2013). Optimal global rates of convergence for interpolation problems with random design. *Statistics and Probability Letters*, **83**, pp. 1871–1879.

[16] Lugosi, G. and Zeger, K. (1995). Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory*, **41**, pp. 677–687.

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

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

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

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

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

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

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

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