THE HILBERT KERNEL REGRESSION ESTIMATE

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

Laszlo Györfi
Department of Mathematics and Computer Science
Technical University of Budapest
Stoczek u. 2-4
Budapest, H-1521 Hungary

and

Adam Krzyżak
Department of Computer Science
Concordia University
Montreal, Canada H3G 1M8

Abstract. Let (X, Y) be an $\mathbb{R}^d \times \mathbb{R}$-valued regression pair, where X has a density and Y is bounded. If n i.i.d. samples are drawn from this distribution, the Nadaraya-Watson kernel regression estimate in $\mathbb{R}^d$ with Hilbert kernel $K(x) = 1/\|x\|^d$ is shown to converge weakly for all such regression pairs. We also show that strong convergence cannot be obtained. This is particularly interesting as this regression estimate does not have a smoothing parameter.

Keywords and phrases. Regression function estimation, kernel estimate, convergence, bandwidth selection, Nadaraya-Watson estimate, nonparametric estimation.

1991 Mathematics Subject Classifications: Primary 62G05.

The first author's work was supported by NSERC Grant A3456 and by FCAR Grant 90-ER-0291. The second author's work was supported by the Hungarian Academy of Sciences (MTA IEKCS). The third author was supported by a grant from the Humboldt Foundation, by NSERC Grant OGP0000270 and by FCAR grant 97-ER-1661.

1. Introduction.

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

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

where $h > 0$ is a smoothing factor depending upon n, K is an absolutely integrable function (the kernel), and $K_h(x) = K(x/h)$ (Nadaraya (1964, 1970), Watson (1964)).

We are concerned with the pointwise and L_1 convergence of m_n to m, where the latter is measured by $J_n = \int |m_n(x) - m(x)| \mu(dx)$. This quantity is particularly important in discrimination based on the kernel rule (see Devroye and Wagner (1980) or Stone (1977)). Stone (1977) first pointed out that there exist estimators for which $J_n \to 0$ in probability for all distributions of (X, Y) with $\mathbf{E}|Y| < \infty$. In 1980, Devroye and Wagner, and independently Spiegelman and Sacks, showed that the kernel estimate with smoothing factor h has the same property provided that K is a bounded nonnegative function with compact support such that for a small fixed sphere S centered at the origin, $\inf_{x \in S} K(x) > 0$, and that

$$h \to 0 \ , \ nh^d \to \infty$$

as $n \to \infty$. These results were extended and complemented by Greblicki, Krzyżak and Pawlak (1984) (who allowed bounded but possibly non-integrable kernels), Krzyżak (1986) and Krzyżak and Pawlak (1984). Weak pointwise convergence at almost all x and for all distributions of (X, Y) with $\mathbf{E}|Y| < \infty$ was first obtained by Devroye (1981).

Interestingly, it turns out that the conditions for the "in probability" convergence of J_n are also sufficient for the strong convergence of J_n, thus rendering all modes of convergence equivalent. Assuming that Y is uniformly bounded, the *kernel estimate* is strongly consistent ($J_n \to 0$ almost surely) if the above condition on h holds, K is a Riemann integrable kernel and $K \geq aI_S$, where $a > 0$ is a constant, and S is a ball centered at the origin that has a positive radius (Devroye and Krzyżak, 1989).

In this note, we study the Nadaraya-Watson estimate with Hilbert kernel

$$K(x) = \frac{1}{\|x\|^d} \ .$$

The name refers to the related Hilbert transform in real analysis. This kernel is neither integrable nor bounded, so that none of the papers cited above covers its behavior. Interestingly, the regression function estimate becomes independent of h due to cancelation in numerator and denominator:

$$m_n(x) = \frac{\sum_{i=1}^n Y_i/\|x-X_i\|^d}{\sum_{i=1}^n 1/\|x-X_i\|^d}\ .$$

No parameter is picked in this estimate! Because it is the only kernel with this invariance property, it occupies a special place, and we take the liberty to call the regression estimate the *Hilbert estimate*. Interestingly, this estimate is universally consistent in a sense made precise in the next Theorem.

THEOREM. *Let m_n be the Hilbert regression estimate. Let X have any density f on $\mathbb{R}^d$ and let Y be bounded. Then:*

(i) At almost all x with $f(x) > 0$, $m_n(x) \to m(x)$ in probability as $n \to \infty$.

(ii) $\int |m_n(x) - m(x)|\, f(x)\, dx \to 0$ in probability as $n \to \infty$.

(iii) There exists a distribution of (X, Y) on $[0,1] \times \{-1, 1\}$ such that for all x with $f(x) > 0$, $m_n(x) \not\to m(x)$ almost surely as $n \to \infty$.

We make no claims about the convergence when X does not have a density.

2. Proof of parts (i) and (ii) of the Theorem.

Note that (ii) follows from (i) by a standard argument (Devroye, 1981, p. 1316). Let $S(x,r)$ denote the closed ball in $\mathbb{R}^d$ of radius r centered at x. We will show (i) for all Lebesgue points for f and m, that is, for all x for which $f(x) > 0$ and for which at the same time

$$\lim_{r\downarrow 0} \frac{\int_{S(x,r)} f(y)\,dy}{\int_{S(x,r)} dy} = f(x)$$

and

$$\lim_{r\downarrow 0} \frac{\int_{S(x,r)} m(y)f(y)\,dy}{\int_{S(x,r)} f(y)\,dy} = m(x) \ .$$

As f is a density and $\int |m|f < \infty$, we know that almost all x satisfy the properties given above (Wheeden and Zygmund, 1977, p. 189; see also Devroye, 1981, Lemma 1.1). Let x be such a point.

Fix $\epsilon \in (0,1)$ and find $\delta > 0$ such that

$$\sup_{0<r\leq\delta} \left| \frac{\int_{S(x,r)} f(y)\,dy}{\int_{S(x,r)} dy} - f(x) \right| \leq \epsilon f(x)$$

and

$$\sup_{0<r\leq\delta} \left| \frac{\int_{S(x,r)} m(y)f(y)\,dy}{\int_{S(x,r)} f(y)\,dy} - m(x) \right| \leq \epsilon \ .$$

Define $p = \int_{S(x,\delta)} f$. Let V_d be the volume of the unit ball of $\mathbb{R}^d$, and let F be the univariate distribution function of $W \stackrel{\text{def}}{=} \|x - X\|^d V_d$. Note that F has a density and that if $u \leq V_d\delta^d$,

$$\begin{aligned} F(u) &= \mathbf{P}\left\{V_d\|x-X\|^d \leq u\right\} \\ &= \mathbf{P}\left\{X \in S(x,(u/V_d)^{1/d})\right\} \\ &= \int_{S(x,(u/V_d)^{1/d})} f(y)\,dy \\ &\in [(1-\epsilon)f(x)u, (1+\epsilon)f(x)u] \ . \end{aligned}$$

Define $W_i = V_d\|x - X_i\|^d$, $1 \leq i \leq n$, and let $W_{(1)} < \cdots < W_{(n)}$ be the order statistics for $W_1, \ldots, W_n$. If $U_{(1)} < \cdots < U_{(n)}$ are uniform order statistics, we have in fact the representation

$$U_{(i)} \stackrel{\mathcal{L}}{=} F(W_{(i)}), W_{(i)} \stackrel{\mathcal{L}}{=} F^{\text{inv}}(U_{(i)})$$

jointly for all i. Thus,

$$(1-\epsilon)f(x)W_{(i)} \leq U_{(i)} \leq (1+\epsilon)f(x)W_{(i)}$$

provided $W_{(i)} \leq V_d \delta^d$. Put differently, under the latter condition,

$$\frac{U_{(i)}}{(1+\epsilon)f(x)} \leq W_{(i)} \leq \frac{U_{(i)}}{(1-\epsilon)f(x)} \ .$$

The Hilbert estimate $m_n(x)$ may be written as follows:

$$m_n(x) = \frac{\sum_{i=1}^n Y_i/W_i}{\sum_{i=1}^n 1/W_i} \ .$$

Thus,

$$|m_n(x) - m(x)| \leq \left| \frac{\sum_{i=1}^n (Y_i - m(X_i))/W_i}{\sum_{i=1}^n 1/W_i} \right| + \frac{\sum_{i=1}^n |m(X_i) - m(x)|/W_i}{\sum_{i=1}^n 1/W_i} \stackrel{\text{def}}{=} I + II \ .$$

We show that I and II tend to zero in probability. We will repeatedly use the following special form of the Hajek-Rényi inequality (see Chow and Teicher, 1978):

Lemma 1. *If $X_1, X_2, \ldots$ are i.i.d. zero mean random variables with variance σ^2 and $\epsilon > 0$, then*

$$\mathbf{P}\left\{ \bigcup_{i=n}^{\infty} \left[\left| \frac{\sum_{j=1}^i X_j}{i} \right| \geq \epsilon \right] \right\} \leq \frac{2\sigma^2}{n\epsilon^2} \ .$$

Part I. We may assume without loss of generality that $|Y| \leq 1$ (so that $|m| \leq 1$ as well). Given $X_1, \ldots, X_n$ (and thus $W_1, \ldots, W_n$), we have $Y_1, \ldots, Y_n$ conditionally independent, and thus

$$\begin{aligned}
\mathbf{E}\left\{I^2 | X_1, \ldots, X_n\right\} &\leq \sum_{i=1}^n \left(\frac{1/W_i}{\sum_{j=1}^n 1/W_j} \right)^2 \\
&\leq \max_{1 \leq i \leq n} \frac{1/W_i}{\sum_{j=1}^n 1/W_j} \\
&= \frac{1/W_{(1)}}{\sum_{j=1}^n 1/W_{(j)}} \\
&\leq \frac{(1+\epsilon)f(x)/U_{(1)}}{\sum_{j=1}^k (1-\epsilon)f(x)/U_{(j)}} + I_{W_{(k)} > V_d \delta^d} \\
&\leq \left(\frac{1+\epsilon}{1-\epsilon} \right) \frac{1/U_{(1)}}{\sum_{j=1}^k 1/U_{(j)}} + I_{W_{(k)} > V_d \delta^d} \ .
\end{aligned}$$

Recall that all of this assumes that the U_i's are related to the X_i's by the probability integral transform given above. By Lemma 2 below, if k is the largest integer satisfying $W_{(k)} \leq V_d \delta^d$, $\mathbf{E}\{I^2 | X_1, \ldots, X_n\} \to 0$ in probability. This implies $\mathbf{E} I^2 \to 0$, and thus $I \to 0$ in probability. $\square$

LEMMA 2. *Let $U_{(1)} < \cdots < U_{(n)}$ be uniform order statistics. Let $\delta > 0$ be arbitrary, and let k be the largest integer such that $W_{(k)} \le V_d \delta^d$. Then*

$$Z \stackrel{\text{def}}{=} \frac{1/U_{(1)}}{1/U_{(1)} + \cdots 1/U_{(k)}} \to 0 \text{ in probability}$$

as $n \to \infty$.

PROOF. We may use a well-known connection between uniform samples and Poisson point processes. If $E_1, E_2, \ldots$ are i.i.d. standard exponential random variables, then

$$(U_{(1)}, \ldots, U_{(n)}) \stackrel{\mathcal{L}}{=} \left(\frac{\sum_{i=1}^{1} E_i}{\sum_{i=1}^{n+1} E_i}, \ldots, \frac{\sum_{i=1}^{n} E_i}{\sum_{i=1}^{n+1} E_i} \right)$$

(see, e.g., chapter 8 of Shorack and Wellner, 1986). Thus,

$$Z \stackrel{\mathcal{L}}{=} \frac{1/E_1}{1/E_1 + 1/(E_1+E_2) + \cdots + 1/(E_1 + \cdots + E_k)} .$$

Clearly, it suffices to show that the denominator of this expression tends to ∞ in probability. By the Hajek-Rényi inequality,

$$\mathbf{P}\left\{ \bigcup_{i=\ell}^{\infty} \left[\frac{E_1 + \cdots + E_i}{i} \ge 2 \right] \right\} \le \frac{2\mathrm{Var}\{E_1\}}{\ell} = \frac{2}{\ell} ,$$

and hence

$$\mathbf{P}\left\{ 1/E_1 + 1/(E_1+E_2) + \cdots + 1/(E_1 + \cdots + E_n) \ge \sum_{i=\ell}^{n} \frac{1}{2i} \right\} \ge 1 - \frac{2}{\ell} .$$

Therefore, we are done if we can show that $k \to \infty$ in probability. It is a trivial exercise to show, in fact, that there exists an $\epsilon > 0$ such that $\mathbf{P}\{k < \epsilon n\} \to 0$. This concludes the proof of Lemma 2. □

Part II.

We first show that

$$A_n \stackrel{\text{def}}{=} \frac{\sum_{i \le \theta n} 1/W_{(i)}}{\sum_{i=1}^{n} 1/W_{(i)}} \to 1 \quad \text{in probability}$$

as $n \to \infty$ for all fixed $\theta \in (0,1)$. As $A_n \le 1$, we need only be concerned with a lower bound. If $B = [W_{(\lfloor \theta n \rfloor)} \le V_d \delta^d]$ holds, then

$$\frac{(1-\epsilon) f(x)}{U_{(i)}} \le \frac{1}{W_{(i)}} \le \frac{(1+\epsilon) f(x)}{U_{(i)}}$$

for all $i \leq \theta n$. For larger i, we have in any case $1/W_{(i)} \leq 1/W_{(\lfloor \theta n \rfloor)}$. Therefore, if $\epsilon < 1$, and using the inequality $a/(a+b) \geq (a-b)/a$, valid for $a, b > 0$,

$$\begin{aligned}
A_n I_B &\geq \frac{(1-\epsilon) f(x) \sum_{i \leq \theta n} 1/U_{(i)}}{(1+\epsilon) f(x) \sum_{i \leq \theta n} 1/U_{(i)} + n/W_{(\lfloor \theta n \rfloor)}} \\
&\geq 1 - \frac{2\epsilon f(x) \sum_{i \leq \theta n} 1/U_{(i)}}{(1+\epsilon) f(x) \sum_{i \leq \theta n} 1/U_{(i)}} - \frac{n/W_{(\lfloor \theta n \rfloor)}}{(1+\epsilon) f(x) \sum_{i \leq \theta n} 1/U_{(i)}} \\
&\geq 1 - \frac{2\epsilon}{(1+\epsilon)} - \frac{n/W_{(\lfloor \theta n \rfloor)}}{f(x) \sum_{i \leq \theta n} 1/U_{(i)}} \ .
\end{aligned}$$

The middle term is as small as desired by our choice of ϵ, while the last term will be shown to tend to zero in probability. This follows from Lemma 3 below, and the fact that $\mathbf{P}\{W_{(\lfloor \theta n \rfloor)} \leq q\} \to 0$, where q is the $\theta/2$-quantile of $W_1 = V_d \|x - X\|^d$.

To complete the proof of part II, let $\epsilon > 0$ be fixed. First, we find $\delta > 0$ so small that

$$\sup_{r \leq \delta} \frac{\int_{S_{x,r}} |m(y) - m(x)| f(y)\, dy}{\int_{S_{x,r}} f(y)\, dy} < \epsilon^2 \ .$$

Let $A = \{y : |m(y) - m(x)| > \epsilon\}$. By Markov's inequality,

$$\sup_{r \leq \delta} \frac{\int_{S_{x,r} \cap A} f(y)\, dy}{\int_{S_{x,r}} f(y)\, dy} \leq \sup_{r \leq \delta} \frac{\int_{S_{x,r} \cap A} |m(y) - m(x)| f(y)\, dy}{\epsilon \int_{S_{x,r}} f(y)\, dy} \leq \frac{\epsilon^2}{\epsilon} = \epsilon \ .$$

Choose $\theta \in (0,1)$ small enough so that $\mathbf{P}\{\|X_{(\lfloor \theta n \rfloor)} - x\| > \delta\} \to 0$. We set $Z_i = 1/W_i$ for convenience, and note the following:

$$\begin{aligned}
&\frac{\sum_{i=1}^n |m(X_i) - m(x)| Z_i}{\sum_{i=1}^n Z_i} \\
&\leq \frac{2 \sum_{i: X_i \notin S_{x,\delta}} Z_i}{\sum_{i=1}^n Z_i} + \frac{\sum_{i: X_i \in S_{x,\delta}} |m(X_i) - m(x)| Z_i}{\sum_{i=1}^n Z_i} \\
&= \frac{2 \sum_{i: X_i \notin S_{x,\delta}} Z_i}{\sum_{i=1}^n Z_i} (I_{||X_{(\lfloor \theta n \rfloor)} - x|| \leq \delta} + I_{||X_{(\lfloor \theta n \rfloor)} - x|| > \delta}) + \frac{\sum_{i: X_i \in S_{x,\delta}} |m(X_i) - m(x)| Z_i}{\sum_{i=1}^n Z_i} \\
&\leq \frac{2 \sum_{i: X_i \notin S_{x,\delta}} Z_i}{\sum_{i=1}^n Z_i} I_{||X_{(\lfloor \theta n \rfloor)} - x|| \leq \delta} + 2 I_{||X_{(\lfloor \theta n \rfloor)} - x|| > \delta} \\
&+ \frac{\sum_{i: X_i \in S_{x,\delta} \cap A^c} |m(X_i) - m(x)| Z_i}{\sum_{i=1}^n Z_i} + \frac{\sum_{i: X_i \in S_{x,\delta} \cap A} |m(X_i) - m(x)| Z_i}{\sum_{i=1}^n Z_i} \\
&\leq \frac{2 \sum_{i > \theta n} Z_i}{\sum_{i=1}^n Z_i} + 2 I_{||X_{(\lfloor \theta n \rfloor)} - x|| > \delta} + \epsilon + \frac{\sum_{i: X_i \in S_{x,\delta} \cap A} Z_i}{\sum_{i=1}^n Z_i} \ . \\
&= V_1 + V_2 + V_3 + V_4 \ .
\end{aligned}$$

Clearly, $V_1 \to 0$ in probability as $A_n \to 1$ in probability. As noted above, $V_2 \to 0$ in probability by choice of θ. Also, V_3 can be made as small as desired by choice of ϵ. To show that $V_4 \to 0$ in probability, we note that the Z_i's are decreasing random variables if we reorder the X_i's according to distance from x: $\|X_1 - x\| \leq \|X_2 - x\| \leq \cdots$. Set $\xi_i = I_{X_i \in A^c \cap S_{x,\delta}}$, where X_i is the i-th furthest point from x. Define $D_i = Z_i - Z_{i+1}$, where we formally set $Z_{n+1} = 0$. Thus, $Z_i = \sum_{j=i}^n D_j$. Therefore,

$$\begin{aligned}
V_4 &= \frac{\sum_{i=1}^n \xi_i Z_i}{\sum_{i=1}^n Z_i} \\
&= \frac{\sum_{i=1}^n \xi_i \sum_{j=i}^n D_j}{\sum_{i=1}^n Z_i} \\
&= \frac{\sum_{j=1}^n D_j \sum_{i=1}^j \xi_i}{\sum_{i=1}^n Z_i} \\
&\leq \frac{\sum_{j=1}^n D_j (2\epsilon j + M I_{j \leq M})}{\sum_{i=1}^n Z_i} \\
&\qquad \left(\text{if } \textstyle\sum_{i=1}^j \xi_i \leq 2\epsilon j, \text{ all } j \geq M\right) \\
&\leq 2\epsilon \frac{\sum_{j=1}^n j D_j}{\sum_{i=1}^n j D_j} + \frac{M Z_1}{\sum_{i=1}^n Z_i} \\
&\leq 2\epsilon + \frac{M/W_1}{\sum_{i=1}^n 1/W_i} \ .
\end{aligned}$$

The last term tends to 0 in probability by part I of the proof, while the first term can be made as small as desired by choice of ϵ. Thus, $V_4 \to 0$ in probability if

$$\lim_{M \to \infty} \mathbf{P} \left\{ \bigcup_{j \geq M} \left[\frac{\sum_{i=1}^j \xi_i}{j} \geq 2\epsilon \right] \right\} = 0 \ .$$

Fix j. Then $\sum_{i=1}^j \xi_i$ is stochastically not greater than a binomial (j, p) random variable, where

$$p \leq \sup_{r \leq \delta} \mathbf{P}\{X_1 \in A \cap S_{x,r} | X_1 \in S_{x,r}\}$$

(this follows by removing the order of the X_i's again and conditioning on the j-th furthest point from x). But, as noted earlier, $p \leq \epsilon$. Therefore,

$$\begin{aligned}
\mathbf{P} \left\{ \bigcup_{j \geq M} \left[\frac{\sum_{i=1}^j \xi_i}{j} \geq 2\epsilon \right] \right\} &\leq \sum_{j=M}^{\infty} \mathbf{P} \left\{ \frac{\text{binomial}(j, \epsilon)}{j} \geq 2\epsilon \right\} \\
&\leq \sum_{j=M}^{\infty} e^{-2j\epsilon^2} = o(1)
\end{aligned}$$

as $M \to \infty$, where we used Hoeffding's inequality (Hoeffding, 1963). $\square$

Lemma 3. *Let $U_{(1)} < \cdots < U_{(n)}$ be uniform order statistics, and let $\theta \in (0,1)$ be fixed. Then*

$$Z \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i \le \theta n} \frac{1}{U_{(i)}} \to \infty \text{ in probability}$$

as $n \to \infty$.

Proof. As in the proof of Lemma 2, we use the representation of a uniform sample in terms of exponentials. Thus,

$$\begin{aligned} Z &\stackrel{\mathcal{L}}{=} \frac{1}{n} \sum_{i \le \theta n} \frac{E_1 + \cdots + E_{n+1}}{E_1 + \cdots + E_i} \\ &= \frac{E_1 + \cdots + E_{n+1}}{n} \times \sum_{i \le \theta n} \frac{1}{E_1 + \cdots + E_i} \stackrel{\text{def}}{=} III \times IV \ . \end{aligned}$$

By the law of large numbers, $III \to 1$ in probability. By the strong law of large numbers (see Lemma 1), for every $\delta > 0$, there exists $\ell = \ell(\delta)$ such that with probability $> 1 - \delta$, for all $i \ge \ell$,

$$E_1 + \cdots + E_i \le 2i \ .$$

Thus, with probability $> 1 - \delta$,

$$IV = \sum_{i \le \theta n} \frac{1}{E_1 + \cdots + E_i} \ge \sum_{\ell \le i \le \theta n} \frac{1}{2i}$$

and the lower bound diverges with n. $\square$

We note that in fact $Z / \log n \to 1$ in probability.

3. Proof of part (iii) of the Theorem.

Here we construct a simple example in which strong pointwise convergence occurs nowhere, so that the mode of convergence in the Theorem cannot be improved. Let X be uniform on $[0,1]$, and let Y be independent of X and take the values 1 and -1 with probability $1/2$ each. Clearly, $m \equiv 0$. If we define the event

$$B_n = \left[1/|x-X_n| > 2\sum_{i=1}^{n-1} 1/|x-X_i| \right] ,$$

then it is clear that on B_n

$$\begin{aligned} |m_n(x)| &= \left| \frac{Y_n/|x-X_n| + \sum_{i=1}^{n-1} Y_i/|x-X_i|}{1/|x-X_n| + \sum_{i=1}^{n-1} 1/|x-X_i|} \right| \\ &= \left| \frac{1+\sum_{i=1}^{n-1} Y_iY_n|x-X_n|/|x-X_i|}{1+\sum_{i=1}^{n-1} |x-X_n|/|x-X_i|} \right| \\ &\geq \frac{1-\sum_{i=1}^{n-1} |x-X_n|/|x-X_i|}{1+\sum_{i=1}^{n-1} |x-X_n|/|x-X_i|} \geq 1/3. \end{aligned}$$

Therefore,

$$B_n \subseteq [|m_n(x)| \geq 1/3] .$$

Let $\mathcal{F}_n$ denote the σ-field generated by $(X_1,Y_1),\ldots,(X_n,Y_n)$. Then it is well-known that

$$\mathbf{P}\{B_n \text{ i.o.}\} = 1$$

if

$$\sum_{n=1}^{\infty} \mathbf{P}\{B_n|\mathcal{F}_{n-1}\} = \infty \text{ almost surely}$$

(see, e.g., Chow and Teicher, 1978, p. 245). The last condition is equivalent to

$$\sum_{n=1}^{\infty} \mathbf{P}\left\{ \frac{1}{|x-X_n|} > 2\sum_{i=1}^{n-1} \frac{1}{|x-X_i|} \middle| \mathcal{F}_{n-1} \right\} = \infty \text{ almost surely}$$

and by the uniformity of X, this in turn is equivalent to

$$\sum_{n=1}^{\infty} \frac{1}{\sum_{i=1}^{n} \frac{1}{|x-X_i|}} = \infty \text{ almost surely} .$$

It is easy to see that this is true if for an i.i.d. uniform $[0,1]$ sequence $U_1,U_2,\ldots$, we have with probability one

$$\sum_{n=1}^{\infty} \frac{1}{\sum_{i=1}^{n} \frac{1}{U_i}} = \infty .$$

This is shown in Lemma 5. All this thus implies that at every $x \in [0,1]$, $|m_n(x)| \geq 1/3$ infinitely often with probability one, and thus, m_n does not converge strongly to m at any such x (while it converges weakly to x at all such x).

LEMMA 4. *Let $V_1, V_2, \ldots$ be a sequence of (possibly dependent) positive random variables. Let a_n be a sequence of positive numbers with $\sum_n a_n = \infty$. If $\lim_{n\to\infty} \mathbf{P}\{V_n < a_n\} = 0$, then*

$$\sum_{n=1}^{\infty} V_n = \infty$$

almost surely.

PROOF. Define the event $A_{N,k} = [\sum_{n=N}^{\infty} V_n < 1/k]$ where N and k are integers. We will show that $\mathbf{P}\{A_{N,k}\} = 0$. This implies that

$$\mathbf{P}\{\cup_N \cup_k A_{N,k}\} = 0 \ ,$$

and thus that

$$\mathbf{P}\left\{\sum_{n=1}^{\infty} V_n < \infty\right\} = 0 \ .$$

For fixed k, note that $\mathbf{P}\{A_{N,k}\}$ is nondecreasing in N. Assume $\mathbf{P}\{A_{N,k}\} = p > 0$. Note that if $B_n = [V_n \geq a_n]$ and B_n^c is its complement,

$$\begin{aligned}
\frac{1}{k} &\geq \mathbf{E}\left\{I_{A_{N,k}} \sum_{n=N}^{\infty} V_n\right\} \\
&\geq \mathbf{E}\left\{I_{A_{N,k}} \sum_{n=N}^{\infty} V_n I_{B_n}\right\} \\
&\geq \mathbf{E}\left\{I_{A_{N,k}} \sum_{n=N}^{\infty} a_n I_{B_n}\right\} \\
&\geq \mathbf{E}\left\{\sum_{n=N}^{\infty} a_n (I_{A_{N,k}} - I_{B_n^c})\right\} \\
&= \sum_{n=N}^{\infty} a_n (\mathbf{P}\{A_{N,k}\} - \mathbf{P}\{B_n^c\})\} \\
&\geq \sum_{n=N}^{\infty} a_n \mathbf{P}\{A_{N,k}\}/2 \\
&\qquad \text{(for all } N \text{ large enough)} \\
&= \infty \ ,
\end{aligned}$$

which is a contradiction. Therefore, for all N, k, $\mathbf{P}\{A_{N,k}\} = 0$. $\square$

LEMMA 5. *Let $U_1, U_2, \ldots$ be i.i.d. uniform $[0,1]$ random variables. Then, with probability one, we have*

$$\sum_{n=1}^{\infty} \frac{1}{\sum_{i=1}^{n} \frac{1}{U_i}} = \infty \ .$$

PROOF. Define $V_n = 1/\sum_{i=1}^{n} 1/U_i$ and $a_n = 1/(4n \log n)$. Lemma 5 now follows from Lemma 4 if we can show that $\mathbf{P}\{V_n < a_n\} \to 0$, or equivalently, if

$$\lim_{n\to\infty} \mathbf{P}\left\{\sum_{i=1}^{n} \frac{1}{U_i} > 4n \log n\right\} = 0 \ .$$

By the representation of a uniform sample as a function of independent exponentials $E_1, E_2, \ldots$, we see that

$$\sum_{i=1}^{n} \frac{1}{U_i} \stackrel{\mathcal{L}}{=} \sum_{i=1}^{n+1} E_i \times \sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{i} E_j} \stackrel{\text{def}}{=} I \times II \ .$$

Clearly, by the law of large numbers, $I/n \to 1$ in probability. Thus, it suffices to show that $\mathbf{P}\{II > 3(1 + \log n)\} \to 0$. By the strong law of large numbers, for every $\delta > 0$, there exists $\ell = \ell(\delta)$ such that with probability $> 1 - \delta$, for all $i \geq \ell$,

$$E_1 + \cdots + E_i \geq i/2 \ .$$

Thus, with probability $> 1 - \delta$,

$$II \leq \frac{\ell}{E_1} + \sum_{i=\ell}^{n} \frac{2}{i} \leq \frac{\ell}{E_1} + 2(1 + \log n) \ .$$

Therefore,

$$\mathbf{P}\{II > 3(1 + \log n)\} \leq \delta + \mathbf{P}\{\ell/E_1 > 1 + \log n\} \leq \delta + \frac{\ell}{1 + \log n} \ ,$$

which can be made as small as desired by letting n tend to ∞ and picking δ small enough. $\square$

4. Inconsistent generalizations.

One may consider a generalization of the Hilbert kernel regression estimate as follows: take $a > 0$ and define

$$m_n(x) = \frac{\sum_{i=1}^n Y_i/\|x - X_i\|^{ad}}{\sum_{i=1}^n 1/\|x - X_i\|^{ad}} \ .$$

This estimate is not universally consistent unless $a = 1$. The simple example given in this section should drive home our point. For simplicity, we assume $d = 1$, and draw X uniformly on $[0,1]$. Assume first that $a > 1$. Hints for the case $a < 1$ will be given later. We take Y independent of X and and Bernoulli (p). We have

$$m_n(x) = \frac{\sum_{i=1}^n Y_i/\|x - X_i\|^{a}}{\sum_{i=1}^n 1/\|x - X_i\|^{a}} \ .$$

It really suffices to study the behavior at $x = 0$. We let the $U_{(i)}$'s be as in Lemma 3, and note that

$$m_n(0) \stackrel{\mathcal{L}}{=} Z \stackrel{\text{def}}{=} \frac{\sum_{i=1}^n Y_i/U_{(i)}^a}{\sum_{i=1}^n 1/U_{(i)}^a} \ .$$

Note from the proof of Lemma 3 that jointly for all i,

$$\frac{U_{(1)}}{U_{(i)}} \stackrel{\mathcal{L}}{=} \frac{E_1}{E_1 + \cdots + E_i}$$

and that by the Hajek-Rényi inequality,

$$\mathbf{P}\left\{\cup_{i=\ell}^n \left[E_1 + \cdots + E_i \le i/2\right]\right\} \le \frac{4}{\ell} \ .$$

Thus, with probability larger than $1/2$, we have jointly $E_1 + \cdots + E_9 \le 20$ and $E_1 + \cdots + E_i \ge i/2$ for all $i \ge 10$. On the latter event, we have, if $C = 9 + 40^a/(a-1)9^{a-1}$,

$$\begin{aligned} Z &\ge \frac{Y_1}{\sum_{i=1}^n (U_{(1)}/U_{(i)})^a} \\ &\ge \frac{Y_1}{9 + \sum_{i=10}^n (40/i)^a} \\ &\ge \frac{Y_1}{9 + \int_9^\infty (40/x)^a \, dx} \\ &\ge \frac{Y_1}{9 + 40^a/(a-1)9^{a-1}} \\ &= \frac{Y_1}{C} \ . \end{aligned}$$

Hence, $\mathbf{P}\{Z \ge 1/C\} \ge p > 0$, so that $Z \not\to p$ in probability if $1/C > p$.

The case $a > 1$ leads to inconsistency because the contribution of far-away data pairs is too large. Without formally constructing the counterexample, it is helpful to note that the weights of the (X_i, Y_i) pairs now ordered according to increasing values of $\|X_i - x\|$ are roughly proportional to $1/i$. If $H_n = \sum_{i=1}^n 1/i$, then the Hilbert regression estimate is roughly a weighted nearest neighbor regression function estimate

$$\sum_{i=1}^{n} w_{ni} Y_i$$

where $w_{ni} = 1/(iH_n)$, $1 \leq i \leq n$. Note in particular that the w_{ni}'s form a probability vector, and that $\max_i w_{ni} \to 0$. Furthermore, for any $\epsilon > 0$, $\sum_{i > \epsilon n} w_{ni} \to 0$ as $n \to \infty$. The latter two conditions on general weights were obtained by Devroye in 1982 as necessary and sufficient conditions for weak convergence almost everywhere of nearest neighbor type regression function estimates. If we take weights proportional to $1/i^a$ normalized to one, as in the generalization suggested above, then the maximal weight does not tend to zero when $a > 1$, and the ϵn tail of the sum of the weights does not tend to zero when $a < 1$. The counterexamples are thus identical to those given by Devroye (1982).

5. Universal consistency.

The Hilbert kernel estimate is also not universally consistent, i.e., the condition that X has a density cannot be removed in general. This is easily seen from examples such as the following. The Hilbert kernel regression estimate with $d = 2$, $a = 1$ is considered when X is uniformly distributed on $[0, 1] \times \{0\}$, and Y is Bernoulli (p) and independent of X (as in the previous section), then m_n behaves as in a one-dimensional example in which X is uniform on $[0, 1]$. Indeed, both estimates would be statistically indistinguishable. But the Hilbert kernel for $d = 1$ has $a = 2$ and is thus not consistent, as proved in the previous section.

6. True interpolation.

Remarkably, the estimate provides true interpolation as $m_n(X_i) = Y_i$: the regression estimate passes through all data points (X_i, Y_i). This obviously introduces unnecessary noise, but at the same time, except for immediate regions around the data points, the estimate feels and behaves like a true kernel smoother. One could of course introduce various devices to get rid of the noisy peaks, but that will not be attempted in this short note.

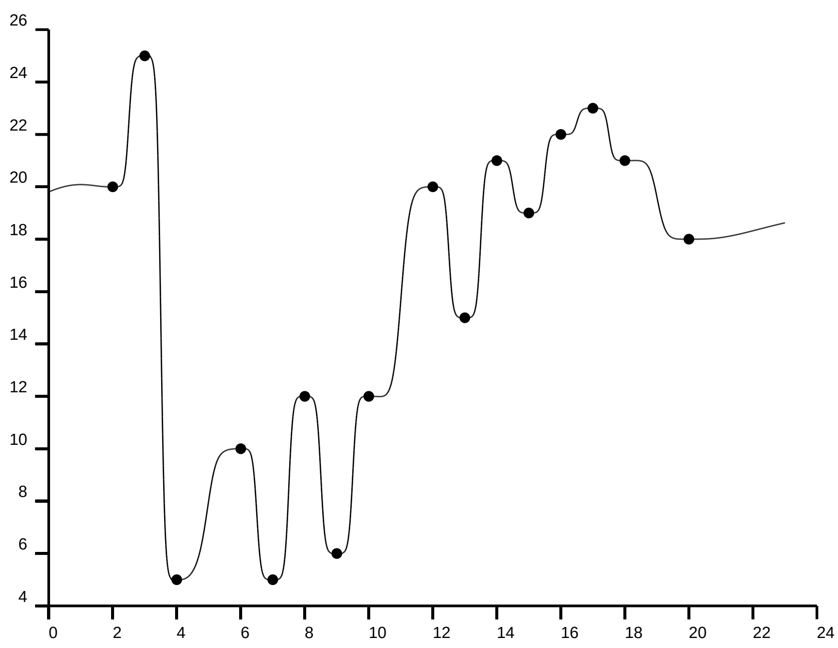

Figure 1. The generalized Hilbert kernel regression estimate with (the inconsistent choice) $a = 4$.

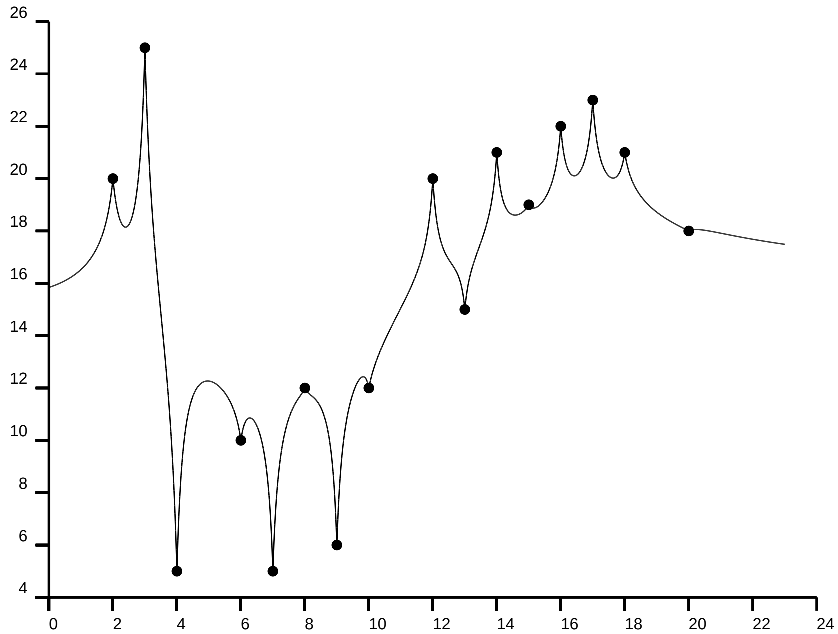

Figure 2. The Hilbert kernel regression estimate with $a = 1$.

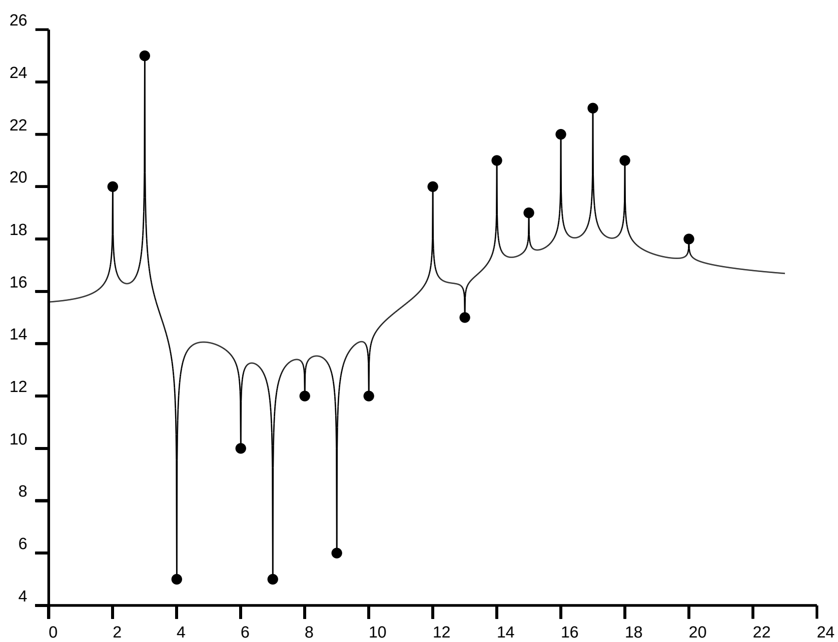

Figure 3. The generalized Hilbert kernel regression estimate with (the inconsistent choice) $a = 1/2$.

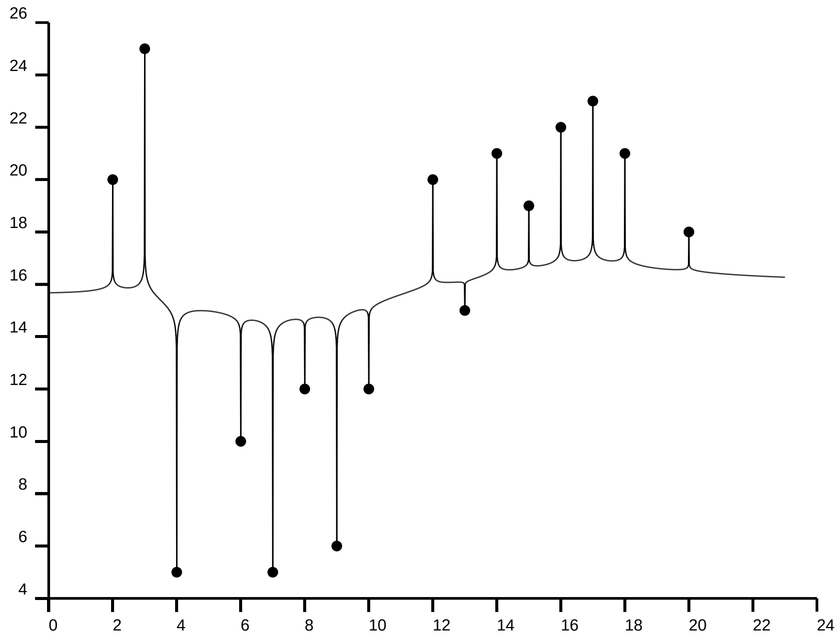

Figure 4. The generalized Hilbert kernel regression estimate with (the inconsistent choice) $a = 1/4$.

7. Application in discrimination.

The early motivation for the Hilbert regression estimate comes from the field of pattern recognition and discrimination, where the data are i.i.d. $\mathbb{R}^d \times \{0,1\}$-valued random vectors (X_i, Y_i), $1 \le i \le n$, and one needs to estimate Y given X, where (X, Y) is distributed as (X_1, Y_1). The kernel discrimination rule $g_n(X)$ is defined as follows:

$$g_n(X) = \begin{cases} 1 & \text{if } \sum_{i=1}^n (2Y_i - 1) K(X - X_i) \ge 0 \\ 0 & \text{otherwise.} \end{cases}$$

This is equivalent to $g_n(X) = I_{m_n(X) \ge 1/2}$. In particular, we have

$$\mathbf{P}\{g_n(X) \ne Y\} \le \inf_{g : \mathbb{R}^d \to \{0,1\}} \mathbf{P}\{g(X) \ne Y\} + 2\mathbf{E}\{|m_n(X) - m(X)|\} ,$$

so that Bayes risk consistency of the discrimination rule follows from L_1 consistency of the corresponding regression estimate (see Devroye, Györfi and Lugosi, 1996, p. 16, for the inequality, and elsewhere in the book for a survey and references). This paper thus solves exercise 10.22 of Devroye, Györfi and Lugosi (1996). Kernels that come close to the Hilbert kernel may be found in the Russian “potential function” literature (Bashkirov, Braverman and Muchnik, 1964) and in early books on learning (Sebestyen, 1962).

8. Extensions.

The results of this note may be repeated for other kernel regression function estimates without smoothing factor. We may for example replace all Hilbert kernels outside a unit ball by zero or a finite constant without affecting the consistency result. Hilbert kernels multiplied with slowly varying functions are also easy to deal with. It is an interesting question to characterize all kernels for which one obtains consistency. So, consider the general estimate

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

The unboundedness of K is essential it seems, as a necessary condition for the estimate to converge in probability at almost all points is that either K is not bounded or $K \notin L_2(\mu)$, where μ is the probability measure of X. To see this, note that

$$|m_n(x) - m(x)| \geq \left| \frac{\sum_{i=1}^n (m(X_i) - m(x)) K(x - X_i)}{\sum_{i=1}^n K(x - X_i)} \right| - \left| \frac{\sum_{i=1}^n (Y_i - m(X_i)) K(x - X_i)}{\sum_{i=1}^n K(x - X_i)} \right| \stackrel{\text{def}}{=} I - II \ .$$

Denote $K_i = K(x - X_i)$. If K is bounded or $K \in L_2(\mu)$ then

$$\mathbf{E}\left\{II^2 | X_1, \ldots, X_n\right\} \leq \sum_{i=1}^n \left(\frac{K_i}{\sum_{j=1}^n K_j} \right)^2 \leq \frac{\sum_{i=1}^n K_i^2}{n\mathbf{E}K^2} \frac{(n\mathbf{E}K)^2}{(\sum_{i=1}^n K_i)^2} \frac{n\mathbf{E}K^2}{n^2\mathbf{E}^2 K} \ .$$

By the law of large numbers the right side converges in probability to to the same limit as $\mathbf{E}K^2/n\mathbf{E}^2K$ which is 0 if K is bounded or $K \in L_2(\mu)$. On the other hand $I \to |\mathbf{E}(m(X) - m(x))K| / |\mathbf{E}K| \neq 0$ in probability (by the law of large numbers), so we do not have consistency.

9. References.

O. Bashkirov, E. M. Braverman, and I. E. Muchnik, "Potential function algorithms for pattern recognition learning machines," *Automation and Remote Control*, vol. 25, pp. 692–695, 1964.

Y. S. Chow and H. Teicher, *Probability Theory*, Springer-Verlag, New York, N.Y., 1978.

G. Collomb, "Estimation non parametrique de la regression:," *International Statistical Review*, vol. 49, pp. 75–93, 1981.

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

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

L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New York, 1996.

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

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

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

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

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

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

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

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

G. Sebestyen, *Decision Making Processes in Pattern Recognition*, Macmillan, New York, N.Y., 1962.

G. R. Shorack and J. A. Wellner, *Empirical Processes with Applications to Statistics*, John Wiley, New York, 1986.

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

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

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

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

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