Journal of Multivariate Analysis **75**, 219–244 (2000)
doi:10.1006/jmva.2000.1901, available online at http://www.idealibrary.com on

Design Adaptive Nearest Neighbor Regression Estimation

Emmanuel Guerre

LSTA (Université Paris 6) and CREST, Paris, France
E-mail: eguerre@ccr.jussieu.fr

Received December 18, 1997; published online October 18, 2000

This paper deals with nonparametric regression estimation under arbitrary sampling with an unknown distribution. The effect of the distribution of the design, which is a nuisance parameter, can be eliminated by conditioning. An upper bound for the conditional mean squared error of $k-NN$ estimates leads us to consider an optimal number of neighbors, which is a random function of the sampling. The

and random designs with vanishing or diverging densities. The proposed estimate is rate optimal for standard designs. © 2000 Academic Press

AMS 1991 subject classifications: 62G07, 62G20, 62G15.

Key words and phrases: conditional nonparametric inference, design adaptation, $k-NN$ nonparametric regression, nonasymptotic inference, nonparametric rates of convergence.

1. INTRODUCTION

Consider n observations (X_i, Y_i) following the regression model

$$Y_i = m(X_i) + \sigma\varepsilon_i, \tag{1.1}$$

where the Y_i's are real variable and the X_i's are in a space $\mathscr{X}$ with a distance d and are independent from the independent and identically distributed (i.i.d.) real variables ε_i's, with $\mathbb{E}\varepsilon_i = 0$, $\operatorname{Var}(\varepsilon_i) = 1$. This paper deals with nonparametric estimation of the regression function $m(\cdot)$ under arbitrary sampling with unknown distribution and investigates the effect of the design on a nearest neighbor $(k-NN)$ estimate. We assume that the regression function from $\mathscr{X}$ to $\mathbb{R}$ is Lipschitz with constant $L > 0$, i.e., in the set of functions

$$C(L) = \{m(\cdot) : \mathscr{X} \mapsto \mathbb{R};\ |m(x) - m(x')| \leqslant L d(x, x'),\ \forall x, x' \in \mathscr{X}\}. \tag{1.2}$$

Most existing work generally consider the case where the X_i's are i.i.d. variables. Györfi (1981) and Devroye (1982) studied $k-NN$ estimates

0047-259X/00 $35.00

under an arbitrary marginal distribution for these variables. When the i.i.d. X_i's have a common continuous density $f_X(\cdot)$, Fan (1993) established that the optimal minimax rate to estimate $m(x)$ is

$$\left(\frac{f_X(x)}{L\sigma^2}\right)^{1/3} n^{1/3}, \qquad f_X(x)>0. \tag{1.3}$$

The dependence of this rate with respect to the density illustrates the effect of sampling in estimation: higher rates can be expected if the density is large, which is intuitively clear because it should be easier to estimate $m(x)$ when many X_i's are close to x. Motivated by nonparametric dimension reduction methods, Hall *et al.* (1997) relaxed the condition $f_X(x)>0$ and studied the case of low designs $f_X(x)=0$.

But such assumptions on the process $\{X_n\}_{n\geqslant 1}$ are difficult to check in practice and do not hold in many applications. The X_i's may not be identically distributed or the independence assumption may not hold, as in the case of an estimated percentage. The rise of the unit root in econometrics suggests that many variables $\{X_n\}_{n\geqslant 1}$ of practical interest should be nonstationary or have seasonalities which do not fit in the standard i.i.d. framework. More generally, design-free nonparametric methods are interesting because the distribution $\mathbb{P}_X$ of the sampling $\{X_n\}_{n\geqslant 1}$ is a nuisance parameter in the regression model (1.1) which usually is not specified. Important steps in this direction can be found in Kulkarni and Posner (1995), who introduced a $k-NN$ estimate for arbitrary design. When the X_i's are valued in a compact subset of $\mathbb{R}$, they showed that the order of the time average risk of their estimate is $n^{-1/3}$, independent of the design distribution $\mathbb{P}_X$, the order corresponding to the i.i.d. case. However, their approach is somehow unsatisfactory because their results depend upon the support of the unknown sampling distribution.

We propose instead to eliminate the design distribution by conditioning with respect to the ancillary statistic $(X_1, ..., X_n)$. This leads us, in difference to the aforementioned authors, to consider a random choice of the number of nearest neighbors which we introduce now. Assume for the moment that the X_i's are deterministic and that there is not tie among the $d(X_i, x)$'s. Denote by $(X_j^x, Y_j^x, \varepsilon_j^x)$ $(1\leqslant j\leqslant n)$ the ordering of $(X_i, Y_i, \varepsilon_i)$ according to the increasing values of $d(X_i, x)$. The $k-NN$ estimate is

$$m_{k,n}(x)=\frac{1}{k}\sum_{j=1}^{k} Y_j^x.$$

Because the X_i's are deterministic variables, $\{\varepsilon_i^x\}_{1\leqslant i\leqslant n}$ and $\{\varepsilon_i\}_{1\leqslant i\leqslant n}$ have the same distribution. Then the mean squared error of $m_{k,n}(x)$ admits the simple upper bound

$$\begin{aligned}\mathbb{E}_m[m_{k,n}(x)-m(x)]^2 &= \mathbb{E}_m\left[\frac{1}{k}\sum_{j=1}^{k}(m(X_j^x)-m(x))+\frac{\sigma}{k}\sum_{j=1}^{k}\varepsilon_j^x\right]^2\\ &= \left[\frac{1}{k}\sum_{j=1}^{k}(m(X_j^x)-m(x))\right]^2+\frac{\sigma}{k}\\ &\leqslant 2\max(L^2 d(X_k^x, x)^2, \sigma^2/k),\end{aligned}$$

for any $m(\cdot)$ in $C(L)$. This suggests that we should consider the following number of neighbors:

$$K_n(x)=\arg\min_{1\leqslant k\leqslant n}\max(L^2 d(X_k^x, x)^2, \sigma^2/k).$$

Define

$$\hat{m}_n(x)=m_{K_n(x),n}(x), \qquad \hat{R}_n(x)=\min\left(\frac{1}{L^2 d(X_k^x, x)^2}, \frac{K_n(x)^{1/2}}{\sigma}\right).$$

This gives the bound for the mean squared error of $\hat{m}_n(x)$,

$$\mathbb{E}_m[\hat{R}_n(x)^2\,(\hat{m}_n(x)-m(x))^2]\leqslant 2, \tag{1.4}$$

which holds independent of the deterministic X_i's and is therefore design-free. In this approach, the weight function $\hat{R}_n(\cdot)$ takes into account the repartition of $X_1, ..., X_n$ over $\mathcal{X}$. Indeed, $\hat{R}_n(x)$ is large if x is close to many X_i's and small if x is far from $X_1, ..., X_n$.

The remainder of the paper is organized as follows. In Section 2 we introduce an extension of the regression model (1.1) and precisely define the $k-NN$ estimates $m_{k,n}(\cdot)$. We provide some design-free nonasymptotic bounds for the mean integrated squared error of $\hat{m}_n(\cdot)$ weighted by $\hat{R}_n(\cdot)$. As is shown by (1.4), the random variable $\hat{R}_n(x)$ is an upper bound for the convergence rate of $\hat{m}_n(x)$, which converges to $m(x)$ if $\hat{R}_n(x)$ diverges with the sample size. We show in Section 3 that $\hat{m}_n(x)$ converges to $m(x)$ if and only if consistent estimation is possible with the design at hand. In Section 4 we compare our conditional approach with standard ones by studying the behavior of $\hat{R}_n(x)$ under some examples of designs. For some i.i.d. X_i's with density $f_X(\cdot)$ our estimate $\hat{m}_n(x)$ converges to $m(x)$ with the optimal rate (1.3) given by Fan (1993). This suggests that our estimation procedure automatically adapts to the design at hand. We also explain how to improve $k-NN$ estimates by selecting a linear smoother via our conditional approach. Final comments and our conclusion are given in Section 5, and the proofs are gathered in the appendixes.

2. NONASYMPTOTIC BOUNDS FOR CONDITIONAL MEAN SQUARED ERRORS

We describe now a general nonparametric regression model which allows for both arbitrary design and general disturbance terms. Denote by $\mathbb{P}$ the distribution of $\{(X_n, Y_n)\}_{n\geqslant 1}$ and by $\mathbb{E}_{\mathbb{P}}$ the associated expectation. We consider $n\geqslant 2$ observations (X_i, Y_i) with, for any $1\leqslant i\neq j\leqslant n$,

$$\begin{aligned}
&\text{(a)} \qquad \mathbb{E}_{\mathbb{P}}[Y_i \mid X_1, ..., X_n]=\mathbb{E}_{\mathbb{P}}[Y_i \mid X_i]=m_{\mathbb{P}}(X_i)\\
&\text{(b)} \qquad m_{\mathbb{P}}(\cdot)\in C(L),\\
&\text{(c)} \qquad \operatorname{Var}_{\mathbb{P}}[Y_i \mid X_1, ..., X_n]\leqslant \sigma^2, \qquad \sigma\geqslant 0,\\
&\text{(d)} \qquad \operatorname{Cov}_{\mathbb{P}}[Y_i, Y_j \mid X_1, ..., X_n]=0,
\end{aligned}$$

where the set of Lipschitz regression functions $C(L)$ is defined in (1.2). The conditional noncorrelation condition (2.1.d) slightly weakens Kulkarni and Posner (1995), who assumed that the distribution of Y_i given $X_1, ..., X_n$, $Y_1, ..., Y_{i-1}, Y_{i+1}, ..., Y_n$ is the distribution of Y_i given X_i. From now on, we denote $\mathscr{C}_n(L, \sigma)$ the family of distributions $\mathbb{P}$ such that the four conditions in (2.1) hold and by

$$\mathscr{C}(L, \sigma)=\bigcap_{n\geqslant 2} \mathscr{C}_n(L, \sigma)$$

the class of $\mathbb{P}$ such that (2.1) holds for all $n\geqslant 2$.

We now introduce some suitable definitions related to the $k-NN$ estimates. For any x in $\mathscr{X}$ and $h>0$, let $N_n(x, h)$ be the number of X_i's with $d(x, X_i)\leqslant h$, i.e.,

$$N_n(x, h)=\sum_{i=1}^{n} \mathbb{1}(d(x, X_i)\leqslant h).$$

Definition 1. For each x in $\mathscr{X}$ and each integer number k, $1\leqslant k\leqslant n$, define

$$H_k(x)=H_{k:n}(x)=\min\{h; N_n(x, h)\geqslant k\}.$$

Consider the subset of integers in $[0, n]$,

$$\mathscr{K}_n(x)=\{N_n(x, h); h\geqslant 0\}.$$

For $k \geqslant 1$ in $\mathscr{K}_n(x)$, the k-nearest neighbor estimate of $m(x)$ is

$$m_{k,n}(x) = \frac{1}{k} \sum_{i=1}^{n} Y_1 \mathbb{1}(d(x, X_i) \leqslant H_k(x)) = \frac{1}{k} \sum_{j=1}^{n} Y_j^x,$$

with, for the second equality, an appropriate ordering in the case of ties.

The set $\mathscr{K}_n(x)$ has been introduced to average over all groups of ties among the X_i's, when necessary, without introducing any ordering. The $H_k(x)$'s are the successive ordered distances of $\{X_1, ..., X_n\}$ to x, that is,

$$H_k(x) \in \{d(x, X_i), 1 \leqslant i \leqslant n\}, \qquad H_1(x) \leqslant H_2(x) \leqslant \cdots \leqslant H_n(x).$$

Definition 1 shows that the $k-NN$ estimate may be viewed as a kernel nonparametric regression estimate with a random bandwidth $H_k(x)$.

As explained above, our approach is based on a simple bound for the conditional means squared error of $m_{k,n}(x)$ under the regression model (2.1) we introduce now. From now on, denote

$$\mathscr{X}_n = (X_1, ..., X_n).$$

Under (2.1.a) and because the $H_k(\cdot)$'s only depend upon $\mathscr{X}_n$, the conditional mean squared error of the $k-NN$ estimate admits the standard bias variance decomposition

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}}[(m_{k,n}(x) - m_{\mathbb{P}}(x))^2 \mid \mathscr{X}_n] \\
&\quad = \mathbb{E}_{\mathbb{P}}\left[\left\{\frac{1}{k} \sum_{i=1}^{n} (Y_i - m_{\mathbb{P}}(X_i) + m_{\mathbb{P}}(X_i) - m_{\mathbb{P}}(x))\right.\right. \\
&\qquad \left.\left. \times \mathbb{1}[d(x, X_i) \leqslant H_k(x)]\right\}^2 \,\middle|\, \mathscr{X}_n\right] \\
&\quad = \left(\frac{1}{k} \sum_{i=1}^{n} (m_{\mathbb{P}}(X_i) - m_{\mathbb{P}}(x))\, \mathbb{1}[d(x, X_i) \leqslant H_k(x)]\right)^2 \qquad (2.2) \\
&\qquad + \mathrm{Var}_{\mathbb{P}}\left[\frac{1}{k} \sum_{i=1}^{n} Y_i \mathbb{1}[d(x, X_i) \leqslant H_k(x)] \,\middle|\, \mathscr{X}_n\right]. \qquad (2.3)
\end{aligned}$$

The first term (2.2) has the interpretation of a (conditional) squared bias term, and the second (2.3) corresponds to a variance one. Note that $N_n(x, \cdot)$ is a càdlàg function, and then

$$\sum_{i=1}^{n} \mathbb{1}[d(x, X_i) \leqslant H_k(x)] = N_n[x, H_k(x)] = k, \qquad k \in \mathscr{K}_n(x).$$

Because $m(\cdot)$ is in the set of Lipschitz functions $C(L)$ defined in (1.2) under (2.1.b), the bias term (2.2) is bounded by

$$\left(\frac{1}{k}\sum_{i=1}^{n} Ld(x, X_i)\,\mathbb{1}[d(x, X_i) \leqslant H_k(x)]\right)^2 \leqslant L^2 H_k^2(x)\frac{N_n[H_k(x)]}{k}$$
$$= L^2 H_k^2(x).$$

For the variance term (2.3), note that the Y_i's are uncorrelated given by X_i's under (2.1.d). This gives, since $H_k(x)$ only depends upon $\mathscr{X}_n$,

$$\mathrm{Var}_{\mathbb{P}}\left[\frac{1}{k}\sum_{i=1}^{n} Y_i \mathbb{1}[d(x, X_i) \leqslant H_k(x)] \,\middle|\, \mathscr{X}_n\right]$$
$$= \frac{1}{k^2}\sum_{1 \leqslant i,\, j \leqslant n} \mathbb{1}[d(x, X_i) \leqslant H_k(x)]\,\mathbb{1}[d(x, X_j) \leqslant H_k(x)]$$
$$\times \mathrm{Cov}[Y_i, y_j \mid \mathscr{X}_n]$$
$$= \frac{1}{k^2}\sum_{i=1}^{n} \mathbb{1}[d(x, X_i) \leqslant H_k(x)]\,\mathrm{Var}_{\mathbb{P}}[Y_i \mid \mathscr{X}_n].$$

Condition (2.1.c), the fact that $N_n[x, H_k(x)] = k$ for k in $\mathscr{K}_n(x)$, yields that the variance (2.3) term is smaller than

$$\frac{\sigma^2}{k^2}\sum_{i=1}^{n} \mathbb{1}[d(x, X_i) \leqslant H_k(x)] = \frac{\sigma^2 N_n[x, H_k(x)]}{k^2} = \frac{\sigma^2}{k}.$$

Therefore, under (2.1), we have for the conditional mean squared error of the $k - NN$ estimate

$$\mathbb{E}_{\mathbb{P}}[(m_{k,n}(x) - m_{\mathbb{P}}(x))^2 \mid \mathscr{X}_n] \leqslant L^2 H_k^2(x) + \frac{\sigma^2}{k}$$
$$\leqslant 2\max\left(L^2 H_k^2(x), \frac{\sigma^2}{k}\right). \tag{2.4}$$

This bound is the basis of our random choice of the number of neighbors k. Note that $1/k$ is strictly decreasing and $H_k^2(x)$ is strictly increasing with respect to k in $\mathscr{K}_n(x)$. The optimal number of neighbors $K_n(x)$ considered here is the smallest k in $\mathscr{K}_n(x)$ with

$$\max\left(L^2H^2_{K_n(x)}(x), \frac{\sigma^2}{K_n(x)}\right) = \min_{k \in \mathscr{K}_n(x)} \max\left(L^2H^2_k(x), \frac{\sigma^2}{k}\right)$$
$$= \min_{1 \leqslant k \leqslant n} \max\left(L^2H^2_k(x), \frac{\sigma^2}{k}\right). \tag{2.5}$$

The set $\mathscr{K}_n(x)$ can be replaced by $\{1, ..., n\}$ in the equation above because, for any $1 \leqslant k' \leqslant n$ which is not in $\mathscr{K}_n(x)$, there is a k in $\mathscr{K}_n(x)$ with $k \geqslant k'$ and $H_{k'}(x) = H_k(x)$.

This leads us to introduce the following optimal $k-NN$ estimate:

$$\hat{m}_n(x) = m_{K_n(x),\, n}(x). \tag{2.6}$$

We denote by $1/\hat{R}^2_n(x)$ the bound (2.4) of the conditional mean squared error associated with the optimal number of neighbors $K_n(x)$, that is,

$$\hat{R}_n(x) = \min\left(\frac{1}{LH_{K_n(x)}(x)}, \frac{K_n^{1/2}(x)}{\sigma}\right)$$
$$= \max_{1 \leqslant k \leqslant n} \min\left(\frac{1}{LH_k(x)}, \frac{k^{1/2}}{\sigma}\right). \tag{2.7}$$

The variables $K_n(x)$, $H_{K_n(x)}$, $\hat{R}_n(x)$, x in $\mathscr{X}$ depend only upon $X_1, ..., X_n$ and then the upper bound (2.4) yields

$$\mathbb{E}_{\mathbb{P}}[\hat{R}_n(x)^2\,(\hat{m}_n(x) - m_{\mathbb{P}}(x))^2 \mid \mathscr{X}_n] \leqslant 2, \tag{2.8}$$

for any x in $\mathscr{X}$. The next theorem extends this result to the integrated mean squared error.

THEOREM 1. *Consider the regression model* (2.1), *and let* μ *be any probability measure over* $\mathscr{X}$ (*possibly depending upon* $X_1, ..., X_n$). *Then we have*

$$\mathbb{E}_{\mathbb{P}}\left\{\int [\hat{R}_n(x)(\hat{m}_n(x) - m_{\mathbb{P}}(x))]^2\,\mu(dx) \,\middle|\, \mathscr{X}_n\right\} \leqslant 2 \tag{2.9}$$

and

$$\mathbb{E}_{\mathbb{P}}\left\{\int [\hat{R}_n(x)(\hat{m}_n(x) - m_{\mathbb{P}}(x))]^2\,\mu(dx)\right\} \leqslant 2, \tag{2.10}$$

for any $\mathbb{P}$ *in* $\mathscr{C}_n(L, \sigma)$.

Proof of Theorem 1. We have, since $\hat{R}_n(\cdot)$ is a function of $\mathscr{X}_n$ and by the Fubini theorem,

$$\mathbb{E}_{\mathbb{P}}\left\{\int [\hat{R}_n(x)(\hat{m}_n(x)-m_{\mathbb{P}}(x))]^2\,\mu(dx)\right\}$$

$$=\mathbb{E}\int \mathbb{E}_{\mathbb{P}}\{[\hat{R}_n(x)(\hat{m}_n(x)-m_{\mathbb{P}}(x))]^2\mid\mathscr{X}_n\}\,\mu(dx).$$

The bound (2.8) for the conditional mean squared error at x yields

$$\int \mathbb{E}_{\mathbb{P}}[\hat{R}_n(x)(\hat{m}_n(x)-m_{\mathbb{P}}(x))^2\mid\mathscr{X}_n]\,\mu(dx)\leqslant 2\int \mu(dx)=2,$$

which shows that (2.9) holds. Taking expectation with respect to $\mathscr{X}_n$ shows that the inequality (2.10) is proved. ∎

Kulkarni and Posner (1995) suggested various choices for the distribution μ. They considered the case of independent X_i's drawn according to a fixed known distribution μ. If the distribution of the process $\{X_i\}_{i\leqslant 1}$ is unknown, it is possible to consider the empirical measure associated with $X_1, ..., X_n$. In this case, Theorem 1 gives bounds for expectations of the averaged errors

$$\frac{1}{n}\sum_{i=1}^{n}\hat{R}_n(X_i)^2\,[\hat{m}_n(X_i)-m_{\mathbb{P}}(X_i)]^2.$$

Bounds can also be obtained for the time-average risk considered in Kulkarni and Posner (1995).Taking a Dirac mass for μ in Theorem 1 yields the upper bound for the mean squared error,

$$\mathbb{E}_{\mathbb{P}}[\hat{R}_n(x)^2\,(\hat{m}_n(x)-m_{\mathbb{P}}(x))^2]\leqslant 2, \tag{2.11}$$

for any x in $\mathscr{X}$.

Theorem 1 deals with sampling process $\{X_n\}_{n\geqslant 1}$ with arbitrary distribution, as in Kulkarni and Posner (1995). The bounds given by these authors depend upon the Lipschitz constant L, the variance bound σ^2, and the support of the unknown distribution of $\{X_n\}_{n\geqslant 1}$. Our conditional approach avoids the introduction of such unknown sets, but the price to be paid is the introduction of the weight function $\hat{R}_n(\cdot)$ when computing the integrated error with respect to μ. However, doing this takes into account the fact that the repartition of the sample $X_1, ..., X_n$ can be nonhomogeneous over the space $\mathscr{X}$, leading to erratic behavior of $\hat{m}_n(\cdot)-m_{\mathbb{P}}(\cdot)$. If x is far from $X_1, ..., X_n$, all the estimates $m_{k,n}(x)$ suffer from a large bias, as $\hat{m}_n(x)$. On the other side, the bias and variance of the optimal

$k-NN$ estimate are small if x is close to many X_i's: if, for instance, $X_i=x$ for all i then the optimal $k-NN$ estimate averages over all the X_i's and $\hat{R}_n(x)=\sqrt{n}/\sigma$.

A byproduct of this conditional approach is that Theorem 1 gives some nonasymptotic bound for the risk of the optimal $k-NN$ estimate. This agrees with the general message in Barndorff-Nielsen and Cox (1994) which argues that ancillary statistics like $\mathscr{X}_n$ are useful for computing exact or accurate approximations for distributions of estimation errors. For instance, the mean squared error bound (2.11) can be used to propose a nonasymptotic confidence interval when the constants L and σ are available. The Tschebyscheff inequality gives

$$\mathbb{P}(|\hat{R}_n(x)(\hat{m}_n(x)-m_{\mathbb{P}}(x))|\geqslant t)\leqslant\frac{2}{t^2},\qquad t>0,$$

and

$$I_n=\left[\hat{m}_n(x)-\frac{1}{\hat{R}_n(x)}\left(\frac{2}{\alpha}\right)^{1/2},\hat{m}_n(x)+\frac{1}{\hat{R}_n(x)}\left(\frac{2}{\alpha}\right)^{1/2}\right],\qquad \alpha\in\,]0,1[,$$

is then a conservative confidence interval of level $1-\alpha$ for $m_{\mathbb{P}}(x)$.

3. CONSISTENT ESTIMATION

The preceding section dealt with a nonasymptotic point of view. We study now the consistency of the optimal $k-NN$ estimate $\hat{m}_n(\cdot)$ when the distribution of the process $\{X_n\}_{n\geqslant 1}$ is arbitrary. For the sake of simplicity, we limit ourselves to the estimation of $m_{\mathbb{P}}(x)$ for a given x. This leads to introduce the following definition.

Definition 2. Let $\mathscr{P}$ be a family of distribution for the process $\{(X_n, Y_n)\}_{n\geqslant 1}$ such that, for any n,

$$\mathbb{E}_{\mathbb{P}}[Y_n\,|\,X_n=x]=m_{\mathbb{P}}(x).$$

The estimate $\tilde{m}_n(x)$ is $\mathscr{P}$-consistent if and only if

$$\mathbb{P}-\lim_{n\to+\infty}\tilde{m}_n(x)=m_{\mathbb{P}}(x),$$

for any distribution $\mathbb{P}$ in $\mathscr{P}$.

Let us now briefly recall some results when the (X_n, Y_n)'s are independent and identically drawn random variables. Devroye (1982) has obtained a necessary and sufficient condition on the number of neighbors k_n ensuring that the estimate $m_{k_n, n}(x)$ is consistent for $\mathbb{P}_X$ almost all x: the deterministic sequence k_n must diverge with k_n/n going to 0.

The case of arbitrary sampling differs considerably. Indeed, for any given deterministic sequence $\{k_n\}_{n \geqslant 1}$ of number of neighbors as above, it may now be possible to find a sampling $\{X_n\}_{n \geqslant 1}$ such that $m_{k_n, n}(x)$ does not converge to $m_{\mathbb{P}}(x)$. Let us now shortly illustrate this fact, with $\mathscr{X} = \{0, +\infty\}$ and $m_{\mathbb{P}}(x) = Lx$, under the condition of the regression model (2.1). We want to estimate $m_{\mathbb{P}}(0)$ with $m_{k_n, n}(0)$, k_n diverging with n and k_n/n going to 0. For such k_n, it is easily shown that it is possible to find a deterministic sequence $\{X_n\}_{n \geqslant 1}$, with, for an infinite number of sample sizes n

$$X_j^0 = 0, \quad 1 \leqslant j \leqslant k_n - 1, \qquad X_j^0 = +\infty, \quad k_n \leqslant j \leqslant n,$$

where the X_j^0's are the X_i's ordered with respect to $d(0, X_i)$. The gives $m_{k_n, n}(0) = +\infty$ and $m_{k_n, n}(0)$ cannot converge to $m_{\mathbb{P}}(0) = 0$. For estimation in stochastic processes, similar examples, where a deterministic choice of smoothing parameters fails to give consistent estimates, can be found in Györfi and Lugosi (1992), Morvai *et al.* (1996), and Adams and Nobel (1998).

However, it is easily seen that our optimal $k - NN$ estimate $\hat{m}_n(0)$ converges in probability to $m_{\mathbb{P}}(0)$ for the example o sampling above, due to the design-dependent choice of the optimal number of neighbors $K_n(0)$. More generally, the bound (2.11) of the mean squared error of $\hat{m}_n(x)$ yields that

$$\hat{m}_n(x) - m_{\mathbb{P}}(x) = O_{\mathbb{P}}\left(\frac{1}{\hat{R}_n(x)}\right), \tag{3.1}$$

and $\hat{m}_n(x)$ converges in probability to $m_{\mathbb{P}}(x)$ if $\hat{R}_n(x)$ goes to infinity. The next lemma implies that $\hat{R}_n(x)$ diverges or stays bounded away from infinity.

Lemma 1. *For any distribution of the process $\{X_n\}_{n \geqslant 1}$ and any x in $\mathscr{X}$, the random sequence $\{\hat{R}_n(x)\}_{n \geqslant 1}$ increases with the sample size n.*

Proof of Lemma 1. Note that $H_{k:n}(x) \geqslant H_{k:n+1}(x)$, $1 \leqslant k \leqslant n$. The definition (2.7) of $\hat{R}_n(x)$ gives

$$\hat{R}_n(x) = \max_{1 \leqslant k \leqslant n} \min\left(\frac{1}{LH_{k:n}(x)}, \frac{k^{1/2}}{\sigma}\right) \leqslant \max_{1 \leqslant k \leqslant n} \min\left(\frac{1}{LH_{k:n+1}(x)}, \frac{k^{1/2}}{\sigma}\right)$$

$$\leqslant \max_{1 \leqslant k \leqslant n+1} \min\left(\frac{1}{LH_{k:n+1}(x)}, \frac{k^{1/2}}{\sigma}\right) = \hat{R}_{n+1}(x). \quad \blacksquare$$

Lemma 1 and (3.1) show that a natural question deals with the estimation of $m_{\mathbb{P}}(x)$ when $\hat{R}_n(x)$ is bounded. The next theorem shows that it is impossible to find a consistent estimate of $m_{\mathbb{P}}(x)$ when $\hat{R}_n(x)$ does not go to infinity. Therefore, the estimate $\hat{m}_n(x)$ converges to $m_{\mathbb{P}}(x)$ if and only if consistent estimation is feasible with the design at hand and is then sampling adaptive in this sense. This illustrates the superiority of our design-dependent choice of the number of nearest neighbors compared to deterministic ones in the context of arbitrary sampling.

THEOREM 2. *Let $\mathscr{P}_X$ be a family of distributions $\mathbb{P}_X$ for $\{X_i\}_{i\geqslant 1}$. Denote by*

$$\mathscr{P} = \{\mathbb{P} \in \mathscr{C}(\sigma, L), \mathbb{P}_X \in \mathscr{P}_X\}$$

the regression model (2.1) *with design distribution in $\mathscr{P}_X$. Consider a given x in $\mathscr{X}$. Then the following propositions are equivalent*:

1. *For any distribution $\mathbb{P}$ in $\mathscr{P}$, there exists, $\mathbb{P}_X$-almost surely, a subsequence of $\{X_i\}_{i\geqslant 1}$ which converges to x.*
2. *For any distribution $\mathbb{P}$ in $\mathscr{P}$, $\lim_{n\to+\infty}\hat{R}_n(x) = +\infty$, $\mathbb{P}_X$-almost surely.*
3. *The estimate $\hat{m}_n(x)$ is a $\mathscr{P}$-consistent estimate of $m_{\mathbb{P}}(x)$.*
4. *There exists a $\mathscr{P}$-consistent estimate of $m_{\mathbb{P}}(x)$.*

Proof of Theorem 2. See Appendix A.

Theorem 2 gives two necessary and sufficient conditions under which consistent estimation of $m_{\mathbb{P}}(x)$ is feasible or, equivalently, such that $\hat{m}_n(x)$ converges to $m_{\mathbb{P}}(x)$ in probability. The most interesting condition is the recurrence condition (1). Condition (2) involves the variable $\hat{R}_n(x)$ which is specific to our Lipschitz regularity assumption in model (2.1), an assumption that can be weakened to continuity or to the approximation hypothesis of Györfi (1981); see Section 4.2 below. Our optimal $k-NN$ estimate is still consistent under such weaker regularity assumptions.

The recurrence condition (1) in Theorem 2 is intuitively clear because it should not be possible to estimate $m_{\mathbb{P}}(x)$ if there are not enough X_i's close to x. Theorem 2 shows the limits of the nonparametric approach under arbitrary sampling. For instance, it is impossible to find a consistent estimate of $m_{\mathbb{P}}(x)$ if the X_i's are trended variables. Stronger versions of the recurrence condition (1) have been used previously in the literature. For the regression model with real deterministic designs Li (1984) has used a stronger recurrence condition, assuming that the number of X_i, $1\leqslant i\leqslant n$, in the intervals $[x-h, x+h]$ is asymptotically larger than κns, $\kappa>0$.

A similar condition also holds for i.i.d. designs with continuous density bounded away from 0 and infinity, as in Fan (1993).

The importance of recurrence assumptions was also noted in nonparametric inference for stochastic processes. For Markov autoregression models

$$Y_i = m(Y_{i-1}) + \sigma\varepsilon_i,$$

Yakowitz (1993) used a Harris null recurrence condition and studied the consistency of a $k-NN$ estimate. For deterministic dynamical systems

$$Y_i = m(Y_{i-1}), \qquad m(\cdot)\colon [0, 1] \mapsto [0, 1],$$

Guerre and Maës (1999) showed the rate optimality of the closest neighbor estimate with a rate function similar to $\hat{R}_n(x)$, assuming that the sequence $\{Y_n\}_{n\geqslant 1}$ is dense in $[0, 1]$.

4. EXAMPLES OF DESIGNS AND BETTER ADAPTATION

Theorem 2 illustrates the design adaptation of our conditional nonparametric estimation procedure from a consistency point of view. This section is devoted to the rate adaptation of $\hat{m}_n(\cdot)$, showing that its convergence rate is optimal for two standard examples of designs. As shown by Theorem 1, (2.11), and (3.1), the weight function $\hat{R}_n(\cdot)$ is an upper bound for the rate of convergence of $\hat{m}_n(\cdot)$ to $m_{\mathbb{P}}(\cdot)$. In this section the order of $\hat{R}_n(\cdot)$ is given for two simple families of deterministic and i.i.d. designs on $[0, 1]$. In each case, we consider nonhomogeneous sampling over $[0, 1]$ and investigate the effect on $\hat{R}_n(\cdot)$ of the repartitioning of the X_i's. The examples studied here include as a special case some low designs considered in Hall *et al.*, but also designs with diverging density. The simple definition (2.7) of $\hat{R}_n(x)$ yields that the order of this variable is easily derived from the study of $N_n(x, \cdot)$; see Appendix B. It is easily seen that Propositions 1 and 2 stated below extend to a large class of deterministic and random samplings. After the examples, we briefly explain how to improve the $k-NN$ estimate $\hat{m}_n(\cdot)$ by implementing our conditional approach to a local linear smoother.

4.1. *Deterministic Designs*

Define X_i in $[0, \theta]$ as

$$X_i = X_{i,n} = \theta\left(\frac{i-1}{n}\right)^{\alpha}, \qquad 1 \leqslant i \leqslant n, \quad \alpha, \theta > 0. \tag{4.1}$$

The parameters θ and α describe the concentration of the X_i's in the vicinity of 0. For simple designs like (4.1), it is easily shown that $\hat{R}_n(x)$ is the rate of convergence of $\hat{m}_n(x)$ to $m_{\mathbb{P}}(x)$. The next proposition gives the asymptotic behavior of $\hat{R}_n(x)$, for each x in $[0, \theta]$, as a function of the sample size n, the regression model parameters L, σ, and the design parameters α, θ.

Proposition 1. *Assume* $\sigma > 0$, *and let* $\{X_i\}_{1 \leqslant i \leqslant n}$ *be as in* (4.1). *Define*

$$r_n(x) = \begin{cases} \dfrac{1}{L}\left(\dfrac{L^2}{\sigma^2}\dfrac{1}{\theta^{1/\alpha}}\right)^{\alpha/(2\alpha+1)} n^{\alpha/(2\alpha+1)}, & x = 0, \\ \dfrac{1}{L}\left\{\dfrac{2L^2}{\sigma^2}\left(\dfrac{x}{\theta}\right)^{1/\alpha - 1}\dfrac{1}{\alpha\theta}\right\}^{1/3} n^{1/3}, & 0 < x < \theta, \\ \dfrac{1}{L}\left(\dfrac{L^2}{\sigma^2}\dfrac{1}{\alpha\theta}\right)^{1/2} n^{1/3}, & x = \theta. \end{cases}$$

Then,

$$\lim_{n \to +\infty} \frac{\hat{R}_n(x)}{r_n(x)} = 1,$$

for any x *in* $[0, \theta]$.

Proof of Proposition 1. See Appendix B.

Korostelev and Tsybakov (1993) have shown, under arbitrary deterministic designs on $[0, \theta]$, that the minimax optimal rate for estimating $m_{\mathbb{P}}(\cdot)$ with respect to the mean integrated squared error is $n^{1/3}$, and Kulkarni and Posner (1995) gave a $k-NN$ estimate which achieves this optimal rate. Proposition 1 described the nonhomogeneous asymptotic behavior of $\hat{R}_n(\cdot)$. Theorem 1 suggests that the use of a rate independent of x is misleading: the rate function $r_n(x)$ is not constant with respect to x. Moreover, Proposition 1 shows that $\hat{R}_n(0)$ is of order $n^{\alpha/(2\alpha+1)}$ which diverges faster than $n^{1/3}$ if $\alpha > 1$ and slower if $\alpha < 1$. Taking $\alpha = +\infty$ in (4.1) gives $X_i = 0$, $1 \leqslant i \leqslant n$. The $k-NN$ estimate $\hat{m}_n(0)$ averages over all the Y_i's and then achieves the parametric optimal rate $\sqrt{n}$. It is worth mentioning that the effect of the concentration parameter α can be compared with the effect of the smoothness of the regression function. Indeed, if $m_{\mathbb{P}}(\cdot)$ has α bounded derivatives, the optimal minimax rate to estimate $m_{\mathbb{P}}(0)$ under the regular design $X_i = (i-1)/n$ is $n^{\alpha/(2\alpha+1)}$ (see Korostelev and Tsybakov, 1993), which is also the order of $\hat{R}_n(0)$ given by Proposition 1.

When $\alpha = 1$, Theorem 1 and Proposition 1 yield that the optimal $k-NN$ estimate $\hat{m}_n(\cdot)$ adapts to the design in the sense that the optimal rate $n^{1/3}$

is achieved. But the estimate $\hat{m}_n(\cdot)$ suffers from a standard side effect (see e.g. Fan and Gijbels, 1996), that is,

$$r_n(0) = r_n(\theta) < r_n(x), \qquad 0 < x < \theta,$$

and estimation is slower at the extremities 0 and θ.

4.2. *The Case of Independent and Identically Distributed Random Variables*

Consider now the case of independent and identically distributed X_i's with a common density $f_X(\cdot)$. Recently attention has been paid to the effect of the design density on nonparametric regression estimation. Györfi (1981) studied $k-NN$ estimates under an approximation condition of the regression function, under arbitrary distribution of the i.i.d. X_i's. Hall *et al.* (1997) considered low designs, that is, vanishing densities $f_X(\cdot)$ at some x, say 0, with $f_X(x) \sim_0 cx^a$, $c > 0$, for some known $a > 0$. Fan (1993) gave bounds for the minimax mean squared error depending upon $f_X(\cdot)$; see also Fan and Gijbels (1996). In this section we consider random i.i.d. designs and compare our results with Györfi (1981), Hall *et al.* (1997), and Fan (1993).

Define now

$$X_i = \theta U_i^{\alpha}, \qquad \alpha \geqslant 0, \qquad \theta \geqslant 0, \tag{4.2}$$

where the U_i's are i.i.d. uniform random variables over $[0, 1]$, which is the random counterpart of the deterministic designs (4.1). The density of the X_i's is

$$f(x) = \frac{1}{\alpha\theta}\left(\frac{x}{\theta}\right)^{1/\alpha - 1} \mathbb{1}_{[0,\theta]}(x).$$

This example of density slightly enlarges the framework of Hall *et al.* (1997). Indeed, taking $\alpha < 1$ gives the low design case with $f_X(0) = 0$ studied by these authors, but considering $\alpha > 1$ yields a diverging density at 0. The next proposition shows that the asymptotic behavior of $\hat{R}_n(\cdot)$ for random designs (4.2) is similar to the one given in Proposition 1 for deterministic designs (4.1).

PROPOSITION 2. *Assume $\sigma > 0$ and let $\{X_i\}_{i \geqslant 1}$ be as in* (4.2). *Then, for the rate function $r_n(\cdot)$ introduced in Proposition* 1, *we have*

$$\mathbb{P} - \lim_{n \to +\infty} \frac{\hat{R}_n(x)}{r_n(x)} = 1,$$

for any x in $[0, \theta]$.

The comments following Proposition 1 also apply to Proposition 2, and we first compare our results with Györfi (1981). This author considered n i.i.d. observations (X_i, Y_i) and a $k-NN$ estimate of the regression function, assuming that the variables X_i's have a common unknown distribution $\mathbb{P}_X$. Arguing that local $\mathbb{P}_X$-means of $m_{\mathbb{P}}(\cdot)$ should be close to the regression function, he introduced an approximation assumption

$$\frac{1}{\mathbb{P}(X \in [x-h, x+h])} \int_{x-h}^{x+h} m_{\mathbb{P}}(u)\, \mathbb{P}_X(du) - m_{\mathbb{P}}(x) \\ = O(h^{\beta}), \qquad \beta > 0, \qquad h > 0. \tag{4.3}$$

for x in the support of the distribution $\mathbb{P}_X$, i.e., such that

$$\liminf_{h \to 0} \frac{\mathbb{P}(X \in [x-h, x+h])}{h} > 0 \tag{4.4}$$

(see also Eq. (8) in Györfi, 1981). Assuming that β is known, he proposed a $k-NN$ estimate achieving at least the rate $n^{\beta/(2\beta+1)}$. Consider now $\mathscr{X} = \mathbb{R}$, $d(x, y) = |x-y|$, and the simple regression function $m_{\mathbb{P}}(x) = Lx$ which is in $C(L)$. For the particular X_i distribution defined in (4.2), we have, for $x = 0$ and $h \leqslant \theta$,

$$\begin{aligned} \frac{1}{\mathbb{P}(X \in [-h, h])} \int_{-h}^{h} m_{\mathbb{P}}(u)\, f_X(u)\, du &= \frac{1}{(h/\theta)^{1/\alpha}} \int_0^h Lu \frac{1}{\alpha\theta} \left(\frac{u}{\theta}\right)^{1/\alpha - 1} du \\ &= \frac{1}{(h/\theta)^{1/\alpha}} \frac{L}{\alpha\theta^{1/\alpha}} \frac{1}{1+1/\alpha} h^{1+1/\alpha} \\ &= O(h), \end{aligned}$$

and we get $\beta = 1$ for $x = 0$ in (4.3), as for any x in $[0, \theta]$. If $x = 0$, Condition (4.4) holds for designs (4.2) with a strictly positive density at 0, that is, $\alpha \geqslant 1$. The convergence rate for $\beta = 1$ in Györfi (1981) is $n^{1/3}$. Proposition 2 gives the rate $n^{\alpha/(2\alpha+1)}$ with $\alpha/(2\alpha+1) > 1/3$ when $\alpha > 1$.

Hall *et al.* (1997) studied the case of vanishing densities $f_X(x) \sim_0 cx^a$, $c > 0$ for some known $a > 0$. For random designs (4.2), Proposition 2 gives, for $\hat{R}_n(0)$, the order

$$n^{1/(3+a)}, \qquad a = \frac{1}{\alpha} - 1 > 0.$$

This shows that there is a loss in the convergence rate for low designs, the exponent 1/3 of uniform designs being replaced by $1/(3+a)$. Our results cannot be compared directly to those of Hall *et al.* (1997), who studied estimation of twice continuously differentiable regression functions. However, they reached a similar conclusion, obtaining the order $n^{2/(5+a)}$, which is slower than the usual rate $n^{2/5}$ for estimation of these smoother regression functions under regular sampling.

Fan (1993) studied asymptotic minimax efficiency for designs with continuous density bounded away from 0 and infinity. He showed that, for any estimate $\tilde{m}_n(x)$,

$$3^{1/3}\left(\frac{f_X(x)}{L\sigma}\right)^{2/3} n^{2/3} \sup_{\mathbb{P}} \mathbb{E}_{\mathbb{P}}[\tilde{m}_n(x)-m_{\mathbb{P}}(x)]^2 \geqslant 0.92^2+o(1), \qquad (4.5)$$

where the supremum is taken over the distribution $\mathbb{P}$ in $\mathscr{C}_n(L,\sigma)$ corresponding to i.i.d. (X_i, Y_i)'s, $f_X(\cdot)$ being the common marginal continuous density of the X_i's. Fan (1993) gives an asymptotic optimal kernel estimate

$$\bar{m}_n(x)=\frac{\sum_{i=1}^{n}(1-|(x-X_i)/h_n|)_+\, Y_i}{1+\sum_{i=1}^{n}(1-|(x-X_i)/h_n|)_+},$$

where $(\cdot)_+$ is the positive part and h_n is an appropriate bandwidth such that

$$3^{1/3}\left(\frac{f_X(x)}{L\sigma^2}\right)^{2/3} n^{2/3} \sup_{\mathbb{P}} \mathbb{E}_{\mathbb{P}}[\bar{m}_n(x)-m_{\mathbb{P}}(x)]^2 = 1+o(1). \qquad (4.6)$$

From (2.8) and Proposition 2 it is expected that the mean squared error of our optimal $k-NN$ estimate is such that

$$\tfrac{1}{2} r_n(x)^2 \sup_{\mathscr{P}} \mathbb{E}_{\mathbb{P}}[\hat{m}_n(x)-m_{\mathbb{P}}(x)]^2 \leqslant 1+o(1), \qquad (4.7)$$

for the design (4.2). For x in $]0,\theta[$, we have

$$\frac{1}{2} r_n(x)^2 = 2^{-1/3}\frac{1}{L^{2/3}\sigma^{4/3}}\left[\frac{1}{\alpha\theta}\left(\frac{x}{\theta}\right)^{1/\alpha-1}\right]^{2/3} n^{2/3}$$

$$=2^{-1/3}\left(\frac{f(x)}{L\sigma^2}\right)^{2/3} n^{2/3},$$

where $f(\cdot)$ is the design density associated with (4.2). As a consequence, (4.5) and (4.7) yield that, for designs (4.2), our optimal $k-NN$ estimate recovers the optimal rate $n^{1/3}$, and the dependence of the rate $r_n(\cdot)$ with respect to $f(\cdot)$, L, and σ^2 is the expected one. The relative asymptotic minimax efficiency of $\hat{m}_n(x)$ with respect to the optimal kernel estimate $\bar{m}_n(x)$ in (4.6) is larger than

$$\frac{r_n(x)/\sqrt{2}}{3^{1/6}(f(x)/(L\sigma^2))^{1/3}\, n^{1/3}} = 6^{-1/6} \approx 0.74,$$

and there is possibly some loss of asymptotic efficiency from using the optimal $k-NN$ estimate $\hat{m}_n(x)$ instead of $\bar{m}_n(x)$ for the design (4.2). We explain in the next section how to improve our optimal $k-NN$ estimate independent of the sampling.

4.3. *Better Adaptation*

The optimal estimate $\hat{m}_n(\cdot)$ has some interesting theoretical features because the order of the variables $\hat{R}_n(\cdot)$ can be derived easily for the standard sampling examples considered above. This was the basis for the comparison of our conditional approach with previous results. Better (but less explicit) bounds can be obtained for the conditional mean squared error by considering a larger family of estimates. Following Stone (1977), $k-NN$ estimates $m_{k,n}(\cdot)$ are particular case of the linear smoothers

$$m_{\pi,n}(x) = \sum_{i=1}^{n} \pi_i Y_i,$$

where $\pi = (\pi_1, ..., \pi_n)$ are some weights depending upon x and $X_1, ..., X_n$.

From now on we denote by Π the class of weights with $\sum_{i=1}^n \pi_i = 1$. It is possible to choose an optimal $\pi_n(x)$ in Π as we derived an optimal number of neighbors $K_n(x)$, by finding a simple bound for the conditional mean squared error of $m_{\pi,n}(x)$. We have, for any $\mathbb{P}$ in $\mathscr{C}_n(L, \sigma)$,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}}[(m_{\pi,n}(x) - m_{\mathbb{P}}(x))^2 \mid \mathscr{X}_n] \\
&\quad = \mathbb{E}_{\mathbb{P}}\left[\left(\sum_{i=1}^n \pi_i\{m_{\mathbb{P}}(X_i) - m_{\mathbb{P}}(x)\} + \sum_{i=1}^n \pi_i\{Y_i - m_{\mathbb{P}}(X_i)\}\right)^2 \middle| \mathscr{X}_n\right] \\
&\quad = \left(\sum_{i=1}^n \pi_i\{m_{\mathbb{P}}(X_i) - m_{\mathbb{P}}(x)\}\right)^2 + \sum_{i=1}^n \pi_i^2 \operatorname{Var}[Y_i \mid \mathscr{X}_n] \\
&\quad \leqslant \left(L \sum_{i=1}^n |\pi_i|\, d(X_i, x)\right)^2 + \sigma^2 \sum_{i=1}^n \pi_i^2. \qquad (4.8)
\end{aligned}$$

The standard Lagrange multiplier method shows that the optimal weight vector $\pi_n(x)$ is one of the $\pi(x, h)$ with

$$\pi_i(x, h) = \left(1 - \frac{d(X_i, x)}{h}\right)_+.$$

For real X_i's and $d(x, y) = |x - y|$, such weights give for some h_n the optimal kernel estimate $\bar{m}_n(\cdot)$ given in Fan (1993) up to a negligible term 1 in the denominator. An optimal (random) bandwidth $\eta_n(x)$ can be found by taking h such that the upper bound (4.8) is the smallest possible. The random rate corresponding to this optimal linear smoother $m_n(\cdot)$ is

$$R_n(x) = \max\left[\frac{1}{L\sum_{i=1}^n \pi_i(x, \eta_n(x))\, d(X_i, x)}, \frac{(\sum_{i=1}^n \pi_i^2(x, \eta_n(x)))^{-1/2}}{\sigma}\right].$$

Because the upper bound (4.8) for the conditional mean squared error is always smaller than (2.4), we have

$$\hat{R}_n(x) \leqslant R_n(x)$$

for any x in $\mathscr{X}$, which suggests that $m_n(x)$ is more efficient than $\hat{m}_n(x)$. However, note that finding deterministic equivalents of $R_n(\cdot)$ may be more difficult than for $\hat{R}_n(\cdot)$.

5. CONCLUSION AND FURTHER COMMENTS

A conditional approach for selecting an optimal $k - NN$ estimate $\hat{m}_n(\cdot)$ has been proposed. The weighted mean-integrated squared error of $\hat{m}_n(\cdot)$ is bounded by 2 for any sample size n. The role of the weight function $\hat{R}_n(\cdot)$ is to capture the impact of the repartition of the X_i's without any a priori information on the distribution of the design. The random choice (2.5) of the number of neighbors $K_n(x)$ defining the estimate $\hat{m}_n(\cdot)$ is also useful when the sample size n grows. Our $k - NN$ estimate is design adaptive, meaning that $\hat{m}_n(x)$ is consistent if and only if consistent estimation is possible for the design at hand. Moreover, $\hat{m}_n(x)$ is rate optimal for some standard examples of designs without using this addition a priori information on the sampling distribution. For some nonstandard designs clustering at x it is also shown that $\hat{m}_n(x)$ improves on the $k - NN$ estimate of Györfi (1981). An improved design-adaptive estimate is also derived via conditioning. Furthermore, it should be possible to derive, equivalent to the random variable $\hat{R}_n(x)$, an upper bound of the convergence rate of $\hat{m}_n(x)$ for new examples of designs such that the behavior of $N_n(x, .)$ is suitable, as for random walks and ARIMA processes; see Akonom (1993) and Appendix B.

Lots of work remain to be done. In many applications, the Lipschitz constant L and the variance bound σ^2 are unknown. It is possible to choose an arbitrary random choice of the number $K'_n(x)$, taking for instance $L = \sigma = 1$ in (2.5),

$$K'_n(x) = \arg \max_{k \in \mathscr{K}_n(x)} \min\left(\frac{1}{H_k(x)}, k^{1/2}\right).$$

If

$$R'_n(x) = \min\left(\frac{1}{H_{K'_n(x)}(x)}, K'_n(x)^{1/2}\right),$$

a standard bias-variance decomposition similar to (2.8) gives

$$\frac{2R'_n(x)^2}{L^2 + \sigma^2} \mathbb{E}_{\mathbb{P}}[(m_{K'_n(x), n}(x) - m_{\mathbb{P}}(x))^2 \mid \mathscr{X}_n] \leqslant 2.$$

Nonasymptotic bounds for weighted mean-integrated squared errors will follow from this inequality as in Theorem 1. The recurrence condition (1) in Theorem 2 implies that $m_{K'_n(x), n}(x)$ is consistent if and only if there exists a consistent estimate for the design at hand. This $k - NN$ estimate also achieves the rates given in Propositions 1 and 2. However, such estimates are somehow arbitrary and may perform poorly in practice. A better choice of the number of neighbors can be derived from the cross-validation procedure given in Li (1984), who studied deterministic sampling fulfilling a recurrence condition. Extensions of the empirical bandwidth choice in Hall *et al.* (1997) for low designs can also be considered. Spokoiny (1998) proposed, for change-point analysis, a selection method based on residual analysis which may apply in our conditional framework.

APPENDIX A

Proof of Theorem 2. Clearly, Statement 2 implies Statement 3 by (3.1), which gives Statement 4. Theorem 2 is true if we show that Statements 1 and 2 and Statements 1 and 4 are equivalent.

Statements 1 *and* 2 *are equivalent*. Let $X_1, ..., X_n, ...$ be a fixed sequence such that $\hat{R}_n(x)$ diverges. Then $K_n(x)$ diverges and $H_{K_n(x)}(x)$ goes to 0; see the definition (2.7) of $\hat{R}_n(x)$. Because $H_{K_n(x)}(x) = d(x, X_{\tilde{K}_n(x)})$ for some nonconstant sequence $\{\tilde{K}_n(x)\}_{n \geqslant 1}$, $X_{\tilde{K}_n(x)}$ converges to x, and Statement 2 implies Statement 1.

Assume that Statement 2 holds and consider $\{X_{n(p)}\}_{p \geqslant 1}$ such that $n(p)$ increases to infinity with p, and $d(x, X_{n(p)})$ decreases to 0. Let q be a fixed integer, strictly smaller than p. Because $d(x, X_{n(p)}) \leqslant d(x, X_{n(p-1)}) \leqslant \cdots \leqslant d(x, X_{n(p-q)})$, the number of X_i, $1 \leqslant i \leqslant n(p)$, with $d(x, X_i) \leqslant d(x, X_{n(p-q)})$

is larger than or equal to q. Then, by the definition of $H_{q:n(p)}$, we get that $H_{q:n(p)} \leqslant d(x, X_{n(p-q)})$. This gives

$$\hat{R}_{n(p)}(x) = \max_{1 \leqslant k \leqslant n(p)} \min\left(\frac{1}{LH_{k:n(p)}}, \frac{\sqrt{k}}{\sigma}\right) \geqslant \min\left(\frac{1}{LH_{q:n(p)}}, \frac{\sqrt{q}}{\sigma}\right)$$
$$\geqslant \min\left(\frac{1}{Ld(x, X_{n(p-q)})}, \frac{\sqrt{q}}{\sigma}\right),$$

and the lower bound can be made as large as is wanted by taking q large. Therefore, $\hat{R}_{n(p)}(x)$ diverges. Because Lemma 1 shows that $\hat{R}_n(x)$ is increasing, $\hat{R}_n(x)$ diverges.

Statements 1 *and* 4 *are equivalent*. Due to the preceding step, Statement 1 implies 3, which yields Statement 4. Let us now prove that Statement 4 implies Statement 1.

Consider the Gaussian regression submodel $\mathscr{P}_g$ of $\mathscr{P}$ given by

$$Y_i = m(X_i) + \sigma\varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\rightsquigarrow} \mathcal{N}(0, 1),$$

where the sampling $(X_1, ..., X_n, ...) \rightsquigarrow \mathbb{P}_X$, $\mathbb{P}_X$ in $\mathscr{P}_X$, is independent of the regression disturbance terms. If (4) is true, there is a $\mathscr{P}_g$-consistent estimate $\tilde{m}_n(x)$. The proof works by contradiction, assuming that such an estimate $\tilde{m}_n(x)$ exists without Statement 1.

Let $v_n = N_n(x, 0) = \sum_{i=1}^{n} \mathbb{1}(X_i = x)$. If Statement 1 does not hold, v_n is bounded away from infinity and the $d(x, X_i)$'s with $X_i \neq x$ must stay bounded away from 0. Because $X_i \neq x$ gives that $d(x, X_i) \geqslant H_{v_n+1:n}$, $1 \leqslant i \leqslant n$, this is equivalent to

$$v_n = v \qquad \text{for } n \text{ large enough} \quad \text{and} \quad H = \lim_{n \to +\infty} H_{v_n+1:n}(x) > 0, \quad \mathbb{P}_X\text{-a.s.},$$

$H_{v_n+1:n}$ being decreasing with n as soon as $v_n = v$.

Let $h > 0$ be such that $\mathbb{P}(H > h) > 0$. Let $\psi(\cdot)$ be a real C^∞ function supported by $[-1, 1]$ with $\psi(0) > 0$, $\sup_{t \in \mathbb{R}} |\psi'(t)| \leqslant L$. Define

$$\phi(x') = \frac{h}{2}\psi\left(\frac{2d(x, x')}{h}\right),$$

for x' in $\mathscr{X}$. For any x_1, x_2 in $\mathscr{X}$, we have by the triangular inequality

$$|\phi(x_1) - \phi(x_2)| \leqslant \sup_{t \in \mathbb{R}} |\psi'(t)| \, |d(x, x_1) - d(x, x_2)| \leqslant Ld(x_1, x_2),$$

and $\phi(\cdot)$ is in $C(L)$, with $\phi(x) = \psi(0) > 0$. Denote by $\mathbb{P}_\phi$ and $\mathbb{P}_0$ the distribution associated with $m(\cdot) = \phi(\cdot)$ and $m(\cdot) = 0$, respectively.

We introduce the two simple hypotheses $\mathscr{H}_0: m_{\mathbb{P}}(\cdot)=0$ against $\mathscr{H}_1$: $m_{\mathbb{P}}(\cdot)=\phi(\cdot)$. Assume that $\tilde{m}_n(x)$ is $\mathscr{P}$, and then $\mathscr{P}_g$-consistent. Because $\psi(0)>0$, this implies that

$$\lim_{n\to+\infty} \max[\mathbb{P}_0(|\tilde{m}_n(x)|\geqslant\psi(0)/2), \mathbb{P}_\phi(|\tilde{m}_n(x)-\phi(x)|\geqslant\psi(0)/2)]=0.$$

Define now the test

$$T_n=0 \qquad \text{if and only if} \quad |\tilde{m}_n(x)|<\psi(0)/2.$$

This gives

$$\begin{aligned}\mathbb{P}_\phi(T_n=0)&=\mathbb{P}(|\tilde{m}_n(x)-\phi(x)+\phi(x)|\leqslant\psi(0)/2)\\ &\leqslant\mathbb{P}_\phi(\psi(0)-|\tilde{m}_n(x)-\phi(x)|\leqslant\psi(0)/2)\\ &=\mathbb{P}_\phi(|\tilde{m}_n(x)-\phi(x)|\geqslant\psi(0)/2).\end{aligned}$$

Then the limit above yields that the sum of the testing errors goes to 0, i.e.,

$$\mathbb{P}_0(T_n=1)+\mathbb{P}_\phi(T_n=0)\to 0. \tag{5.1}$$

Denote by $\mathbb{P}_{0,n}$ and $\mathbb{P}_{\phi,n}$ the distributions of the n first (X_i, Y_i). Le Cam and Yang (1990) showed that the sums of the two type errors of any test of $\mathscr{H}_0$ against $\mathscr{H}_\phi$ are bounded from below as follows:

$$\begin{aligned}\mathbb{P}_0(T_n=\phi)+\mathbb{P}_\phi(T_n=0)&\geqslant 1-\tfrac{1}{2}\int|d\mathbb{P}_{0,n}-d\mathbb{P}_{\phi,n}|\\ &\geqslant 1-\left[1-\left(\int\sqrt{d\mathbb{P}_{0,n}\,d\mathbb{P}_{\phi,n}}\right)^2\right]^{1/2}.\end{aligned} \tag{5.2}$$

We have, under Gaussian,

$$\begin{aligned}\int\sqrt{d\mathbb{P}_{0,n}\,d\mathbb{P}_{\phi,n}}&=\int\sqrt{d\mathbb{P}_{0,n}(\cdot\mid\mathscr{X}_n)\,d\mathbb{P}_{\phi,n}(\cdot\mid\mathscr{X}_n)}\,d\mathbb{P}_X\\ &=\mathbb{E}\left[\exp\left(-\frac{1}{8\sigma^2}\sum_{i=1}^n\phi^2(X_i)\right)\right]\\ &\geqslant\mathbb{E}\left[\exp\left(-\frac{1}{8\sigma^2}\sum_{i=1}^n\phi^2(X_i)\right)\mathbb{1}(H_{v_n+1:n}>h)\right]\\ &=\mathbb{E}\left[\exp\left(-\frac{v_n\psi(0)}{8\sigma^2}\right)\mathbb{1}(H_{v_n+1:n}>h)\right]\\ &\to\mathbb{E}\left[\exp\left(-\frac{v\psi(0)}{8\sigma^2}\right)\mathbb{1}(H_{v+1}>h)\right]=\ell\leqslant 1,\end{aligned}$$

because, for $1 \leqslant i \leqslant n$, $\phi(X_i) = 0$ if $X_i \neq x$ and $H_{v_n+1:n} > h$, and by the Lebesgue dominated convergence theorem. Therefore inequality (5.2) yields

$$\mathbb{P}_0(T_n = 1) + \mathbb{P}_\phi(T_n = 0) \geqslant 1 - [1 - \ell^2]^{1/2} + o(1),$$

with $1 - [1 - \ell^2]^{1/2} > 0$ because $\ell > 0$ by the definitions of h and v. This contradicts the limit (5.1). Then Statement 4 implies Statement 1, and Theorem 2 is proved. ▮

APPENDIX B

We first give a convenient expression of $\hat{R}_n(x)$. Note that $L^2 h^2 N_n(x, h)$ is a càdlàg function with respect to h, and define

$$B_n(x) = \min\{h \geqslant 0;\ L^2 h^2 N_n(x, h) \geqslant \sigma^2\}. \tag{5.3}$$

Lemma 2. *Let $\hat{R}_n(x)$ be as in* (2.7) *and $B_n(x)$ be as above. Then*

$$\hat{R}_n(x) = \sup_{h \geqslant 0} \min\left(\frac{1}{Lh}, \frac{N_n^{1/2}(x, h)}{\sigma}\right) = \frac{1}{LB_n(x)}.$$

Proof of Lemma 2. For k in $\mathscr{K}_n(x)$, let $s(k)$ be the smallest element of $\mathscr{K}_n(x)$ strictly larger than k, with $s(\sup \mathscr{K}_n(x)) = +\infty$, $H_{+\infty}(x) = +\infty$. Note that for any h in $[H_k(x), H_{s(k)}(x)[$, $N_n(x, h) = k$; see Definition 1. This gives, for k in $\mathscr{K}_n(x)$,

$$\begin{aligned}\sup_{h \in [H_k(x), H_{s(k)}(x)[} \min\left(\frac{1}{Lh}, \frac{N_n^{1/2}(x, h)}{\sigma}\right) &= \sup_{h \in [H_k(x), H_{s(k)}(x)[} \min\left(\frac{1}{Lh}, \frac{k^{1/2}}{\sigma}\right) \\ &= \min\left(\frac{1}{LH_k(x)}, \frac{k^{1/2}}{\sigma}\right).\end{aligned}$$

Combining this with the definitions (2.7) and (2.5) yields

$$\begin{aligned}\hat{R}_n(x) &= \max_{k \in \mathscr{K}_n(x)} \min\left(\frac{1}{LH_k(x)}, \frac{k^{1/2}}{\sigma}\right) \\ &= \sup_{h \geqslant H_{\inf \mathscr{K}_n(x)}(x)} \min\left(\frac{1}{Lh}, \frac{N_n^{1/2}(x, h)}{\sigma}\right) \\ &= \sup_{h \geqslant 0} \min\left(\frac{1}{Lh}, \frac{N_n^{1/2}(x, h)}{\sigma}\right).\end{aligned}$$

To prove the second equality, note that

$$B_n(x) = \min\{h \geqslant 0;\, N_n^{1/2}(x, h)/\sigma \geqslant 1/(Lh)\},$$

that $1/(Lh)$ continuously decreases with h, and that $N_n^{1/2}(x, h)/\sigma$ is a càdlàg increasing function with respect to h. This gives for $h \geqslant B_n(x)$,

$$\min\left(\frac{1}{Lh}, \frac{N_n^{1/2}(x, h)}{\sigma}\right) = \frac{1}{Lh} \leqslant \frac{1}{LB_n(x)},$$

the upper bound being achieved for $h = B_n(x)$. For $h < B_n(x)$, we distinguish two cases.

1. The curves $\{(h, 1/(Lh));\, h \geqslant 0\}$ and $\{(h, N_n^{1/2}(x, h)/\sigma);\, h \geqslant 0\}$ do not cross each other. In this case we have for $h < B_n(x)$ that $N_n^{1/2}(x, h)/\sigma < 1/(Lh)$, by the definition of $B_n(x)$, and then, by the continuity of $1/(Lh)$, that

$$\min\left(\frac{1}{Lh}, \frac{N_n^{1/2}(x, h)}{\sigma}\right) = \frac{N_n^{1/2}(x, h)}{\sigma} \leqslant \frac{1}{LB_n(x)}.$$

This gives that $\hat{R}_n(x) = 1/(LB_n(x))$.

2. The curves $\{(h, 1/(Lh));\, h \geqslant 0\}$ and $\{(h, N_n^{1/2}(x, h)/\sigma);\, h \geqslant 0\}$ have a unique intersection. In this case $B_n(x)$ is such that $1/(LB_n(x)) = N_n^{1/2}(x, B_n(x))/\sigma$, and the variations of the two curves imply $\hat{R}_n(x) = 1/(LB_n(x))$. ▮

Proof of Proposition 1. Let $[\,\cdot\,]$ be the integer part of a real number. We have, for the design (4.1),

$$N_n(x, h) = \sum_{i=1}^{n} \mathbb{1}(x - h \leqslant X_i \leqslant x + h)$$
$$= \begin{cases} \left[n \left(\dfrac{h}{\theta}\right)^{1/\alpha} + 1 \right] & x = 0, \\ \left[n \left(\left(\dfrac{x+h}{\theta}\right)^{1/\alpha} - \left(\dfrac{x-h}{\theta}\right)^{1/\alpha} \right) 1 + \right] & 0 < x < \theta, \\ \left[n \left(1 - \left(1 - \dfrac{h}{\theta}\right)^{1/\alpha} \right) + 1 \right] & x = \theta, \end{cases} \tag{5.4}$$

for n large enough. Define now

$$b_n(x) = \frac{1}{Lr_n(x)}.$$

This gives, for any $\lambda > 0$,

$$
\begin{aligned}
&L^2(\lambda b_n(x))^2\, N_n(x, \lambda b_n(x)) \\
&= L^2\lambda^2 \begin{cases}
\left(\dfrac{\sigma^2}{L^2}\right)^{2\alpha/(2\alpha+1)} \theta^{2/(2\alpha+1)} n^{-2\alpha/(2\alpha+1)} \\
\qquad \times \left[\lambda^{1/\alpha} n^{2\alpha/(2\alpha+1)} \theta^{-2/(2\alpha+1)} \left(\dfrac{\sigma^2}{L^2}\right)^{1/(2\alpha+1)} + 4 \right], & x = 0, \\
\left(\dfrac{\sigma^2\alpha\theta}{2L^2 n}\left(\dfrac{x}{\theta}\right)^{(\alpha-1)/\alpha}\right)^{2/3} \left[2n\lambda \dfrac{1}{\alpha\theta}\left(\dfrac{x}{\theta}\right)^{(1-\alpha)/\alpha} \right. \\
\qquad \left. \times \left(\left(\dfrac{\sigma^2\alpha\theta}{2L^2 n}\left(\dfrac{x}{\theta}\right)^{(\alpha-1)/\alpha}\right)^{1/3} + o(n^{-1/3}) \right) + 1 \right], & 0 < x < \theta, \\
\left(\dfrac{\sigma^2\alpha\theta}{L^2 n}\right)^{2/3} \left[n\lambda \dfrac{1}{\alpha\theta} \left(\left(\dfrac{\sigma^2\alpha\theta}{L^2 n}\right)^{1/3} + o(n^{-1/2}) \right) + 1 \right], & x = \theta.
\end{cases} \\
&\to \sigma^2 \begin{cases} \lambda^{2+1/\alpha} & x = 0, \\ \lambda^3 & 0 < x \leqslant \theta, \end{cases}
\end{aligned}
$$

by the mean value for $0 < x \leqslant \theta$. This implies that, for $0 < \eta < 1$ and for some $0 < \eta' = \eta'(\eta) < 1$,

$$
\begin{aligned}
L^2((1+\eta)\, b_n(x))^2\, N_n[x, (1+\eta)\, b_n(x)] &\geqslant \sigma^2(1 + \eta' + o(1)), \\
L^2((1-\eta)\, b_n(x))^2\, N_n[x, (1-\eta)\, b_n(x)] &\leqslant \sigma^2(1 - \eta' + o(1)).
\end{aligned}
$$

Because $L^2 h^2 N_n(x, h)$ increases in h, $B_n(x)$ as defined in (5.3) is in $[(1-\eta)\, b_n(x),\ (1+\eta)\, b_n(x)]$ for n large enough. As a consequence, $B_n(x)/b_n(x)$ goes to 1, and Lemma 2 shows that Proposition 1 is proved. ▮

Proof of Proposition 2. The proof follows the arguments used to establish Proposition 1 as soon as it is shown that $N_n(x, \lambda b_n(x))$ divided by (5.4) taken at $h = \lambda b_n(x)$ converges to 1, in probability.

Recall that

$$
\begin{aligned}
N_n(x, h) &= \sum_{i=1}^{n} \mathbb{1}(x - h \leqslant X_i \leqslant x + h) \\
&= \sum_{i=1}^{n} \mathbb{1} \left\{ \left(\frac{x-h}{\theta}\right)^{1/\alpha} \leqslant U_i \leqslant \left(\frac{x+h}{\theta}\right)^{1/\alpha} \right\}.
\end{aligned}
$$

Denote by $np_n(x, \lambda)$ the expectation of $N_n(x, \lambda b(x))$, which is the expression in (5.4) taken at $h = \lambda b_n(x)$, with

$$
p_n(x, \lambda) = \mathbb{P} \left\{ \left(\frac{x-h}{\theta}\right)^{1/\alpha} \leqslant U_i \leqslant \left(\frac{x+h}{\theta}\right)^{1/\alpha} \right\}.
$$

We have

$$\operatorname{Var}\left\{\frac{N_n(x, \lambda b_n(x))}{np_n(x, \lambda)}\right\} = \frac{1-p_n(x, \lambda)}{np_n(x, \lambda)} \leqslant \frac{1}{np_n(x, \lambda)}.$$

Recall that $np_n(x)$ is the expression (5.4) taken at $h = \lambda b_n(x)$. Then it has been shown in the proof of Proposition 4.1 that $np_n(x) = O(1/b_n^2(x))$, and then the variance of $N_n(x, \lambda b_n(x))/(np_n(x, \lambda))$ goes to 0, and then $N_n(x, \lambda p_n(x))/(np_n(x, \lambda))$ converges to 1 in probability, for any given x. ▮

ACKNOWLEDGMENTS

The author thanks two anonymous reviewers whose comments and suggestions greatly helped to improve the presentation of the paper.

REFERENCES

1. T. M. Adams and A. B. Nobel, On density estimation from ergodic processes, *Ann. Probab.* **26** (1998), 794–804.
2. J. Akonom, Comportement asymptotique du temps d'occupation du processus des sommes partielles, *Ann. Inst. H. Poincaré* **29** (1993), 57–81.
3. O. E. Barndorff-Nielsen and D. R. Cox, "Inference and Asymptotics," Chapman & Hall, London, 1994.
4. L. Devroye, Necessary and sufficient conditions for the pointwise convergence of the nearest neighbor regression function estimates, *Z. Wahrsch. verw. Gebiete* **61** (1982), 467–481.
5. J. Fan, Local linear regression smoothers and their minimax efficiencies, *Ann. Statist.* **21** (1993), 196–216.
6. J. Fan and I. Gijbels, "Local Polynomial Modelling and Its Applications," Chapman and Hall, London, 1996.
7. E. Guerre and J. Maës, Optimal rate for nonparametric estimation in deterministic dynamical systems, *Statist. Inference Stochast. Process.*, to appear.
8. L. Györfi, The rate of convergence of k_n-NN regression estimates and classification rules, *IEEE Trans. Inform. Theory* **27** (1981), 362–364.
9. L. Györfi and G. Lugosi, Kernel density estimation from ergodic sample is not universally consistent, *Comput. Statist. Data Anal.* **14** (1992), 437–442.
10. P. Hall, J. S. Marron, M. H. Neumann, and D. M. Titterington, Curve estimation when the design density is low, *Ann. Statist.* **25** (1997), 756–770.
11. A. P. Korostelev and A. B. Tsybakov, "Minimax Theory of Image Reconstruction," Lecture Notes in Statistics, Vol. 82, Springer-Verlag, New York/Berlin, 1993.
12. S. R. Kulkarni and S. E. Posner, Rates of convergence of nearest neighbor estimation under arbitrary sampling, *IEEE Trans. Inform. Theory* **41** (1995), 1028–1039.
13. L. Le Cam and G. L. Yang, "Asymptotics in Statistics: Some Basic Concepts," Springer-Verlag, New York/Berlin, 1990.
14. K. C. Li, Consistency for cross-validated nearest neighbor estimates in nonparametric regression, *Ann. Statist.* **12** (1984), 230–240.

15. G. Morvai, S. Yakowitz, and L. Györfi, Nonparametric inference for ergodic stationary time series, *Ann. Statist.* **24** (1996), 370–379.
16. V. G. Spokoiny, Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice, *Ann. Statist.* **26** (1998), 1356–1378.
17. C. J. Stone, Consistent nonparametric regression (with discussion), *Ann. Statist.* **5** (1977), 595–645.
18. S. Yakowitz, Nearest neighbor regression estimation for null-recurrent Markov time series, *Stochastic Process. Appl.* **48** (1993), 311–318.