JOURNAL OF APPROXIMATION THEORY 51, 193–201 (1987)

On the Asymptotic Mean Square Error of L_1 Kernel Estimates of Smooth Functions

HANS-GEORG MÜLLER

Institut für medizinisch-biologische Statistik und Dokumentation der Philipps-Universität Marburg, 3550 Marburg/Lahn, West Germany

Communicated by Frank Deutsch

Received November 14, 1984; revised January 2, 1986

Rates of convergence of Mean Squared Error of convolution type estimators of regression functions or density functions in $\mathcal{C}^r$ are derived, employing L_1 kernel functions. The idea is to let the order of the kernel (number of vanishing moments) tend to infinity with increasing number of observations. In this setting, the rate n^{-1} is achieved if and only if the function to be estimated has a specific property. For a broad class of functions, the optimal rate is seen to be $O(\alpha_n^{1/2}/n)$, where $(\alpha_n/e)^{\alpha_n} \sim n$.

1. INTRODUCTION

During the last 25 years, estimation of functions has gained considerable interest in statistics. A classical measure of the performance of such estimators at a point is the Mean Square Error (MSE). If $\hat{g}(t)$ is a curve estimate of a function $g(t)$, the MSE can be decomposed into variance and bias squared:

$$E(\hat{g}(t) - g(t))^2 = \operatorname{var}(\hat{g}(t)) + (E\hat{g}(t) - g(t))^2.$$

Both variance and bias of a curve estimator are to be kept small, and usually a compromise has to be found. Keeping the bias small is equivalent to finding a good deterministic approximation to the function g.

The simplest smoothing and approximation methods of analysis are those of convolution type; compare with Shapiro [10] for an overview. The function to be smoothed is convolved with a smooth function called the kernel function. Because of its simplicity, this kernel smoothing method is widely applied in statistical curve estimation. Let us define kernels of

0021-9045/87 $3.00

order k as kernel functions K with $(k-1)$ vanishing moments which are contained in

$$\tilde{\mathcal{M}}_k := \left\{ f \in L^1([-1,1]) : \int f(x)\, x^j\, dx = \begin{cases} 1, & j=0, \\ 0, & 0<j<k, \end{cases} \quad x^k f \in L^1([-1,1]) \right\}$$

and satisfy $B_k(K) = \int K(x)\, x^k\, dx \neq 0$.

Discussing purely deterministic approximation by kernels of order k, assume that g is a bounded measurable function and that $g^{(k)}$ exists. We then obtain by the pointwise saturation theorem [10]

$$\lim_{h \to 0} h^{-k} \left[g(t) - \int \frac{1}{h} K\left(\frac{t-x}{h} \right) g(x)\, dx \right] = g^{(k)}(t)\, B_k(K) \frac{(-1)^k}{k!}, \qquad (1)$$

where h is a scaling factor or bandwidth. Even if $g \in \mathcal{C}^\infty$, where $\mathcal{C}^\infty$ denotes the space of infinitely often differentiable functions, the rate of convergence remains at h^k (saturation). A faster rate can be obtained only by

(A) increasing the order k or

(B) dropping the assumption that the kernel be in L_1.

In statistical curve estimation, approach (B) has been taken in a series of articles concerned with the estimation of smooth densities [2, 3, 6]. There it is shown that with the Fourier integral kernel $K^*(x) = (\pi x)^{-1} (\sin x)$ which is not in L_1, faster rates of convergence of MSE can be obtained than with L_1-kernels if the characteristic function of the curve to be estimated decreases fast enough. The fastest rate, n^{-1}, is obtained iff the characteristic function is compactly supported [3]. This approach, however, suffers from the drawback that the support of the kernels employed is necessarily unbounded and therefore these rates are not attained if the curve to be estimated is of bounded support. This assumption is always made in the statistical curve-fitting problem to be outlined below. The problem lies in the fact that there are always boundary effects which dominate the convergence if kernel and function to be estimated both have unbounded support. This is reflected in the fact that the characteristic function of a compactly supported curve decreases only algebraically of degree $p=1$ and according to Theorem 4.2 of [2], the Fourier integral kernel then is not competitive to any L_1 kernel with $k \geq 1$.

Therefore, we will be concerned here with approach (A). We apply it to the curve-fitting problem in the fixed design regression model

$$Y_i = g(t_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \qquad (2)$$

where n observations Y_i and times of measurement $t_i = i/n$ are given and the unknown smooth regression function $g \in \mathscr{C}^\infty$ ($[0, 1]$) is to be estimated. The errors (ε_i) are assumed to be independently and identically distributed with $E\varepsilon_1 = 0$, $E\varepsilon_1^2 = \sigma^2 < \infty$. The results hold also for other models as well as for kernel density estimation, but for the sake of simplicity we restrict the discussion to the model (2). Given $t \in (0, 1)$, a kth order kernel estimate of $g(t)$ is defined as

$$g_n(t) = \frac{1}{b} \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} K_k\left(\frac{t-u}{b}\right) du \; Y_i, \tag{3}$$

where b is the bandwidth now depending on n, $K_k \in \mathscr{M}_k := \tilde{\mathscr{M}}_k \cap \mathrm{Lip}([-1, 1])$ for some k (where Lip denotes the class of Lipschitz-continuous functions) is a kernel function of order k and $s_i = (2i+1)/(2n)$, $i = 1, \ldots, n-1$, $s_0 = 0$, $s_n = 1$. In order to avoid boundary effects, we always assume that $b \leqslant \min(t, 1-t)$. The estimate (3) can be viewed as a discretization of the convolution integral in (1).

For fixed k, observing (1) and approximating sums by integrals, using the Lipschitz continuity of g and K_k, we obtain for bias and variance of the estimate (3), assuming $b \to 0$ and $nb \to \infty$:

$$Eg_n(t) - g(t) = \frac{(-1)^k}{k!} b^k (g^{(k)}(t)\, B_k(K_k) + o(1)) + O\left(\frac{1}{n}\right) \tag{4}$$

$$\mathrm{var}(g_n(t)) = \frac{\sigma^2}{nb} (V(K_k) + o(1)), \tag{5}$$

where $V(K_k) = \int K_k^2(x)\, dx$ (cf. [1]). By (4), (5) the MSE optimal bandwidth sequence is seen to be $b \sim n^{-1/(2k+1)}$, and this yields the rate of convergence $\mathrm{MSE} \sim n^{-2k/(2k+1)}$ (for a derivation of this rate in density estimation, cf. [1] or [9]). For functions $g \in \mathscr{C}^k([0, 1])$, this rate is optimal [5, 11]. It remains the same if $g \in \mathscr{C}^\infty([0, 1])$ and a kernel of order k is used.

If we assume $g \in \mathscr{C}^\infty([0, 1])$, there is no corresponding kernel of order $k = \infty$ (see Theorem 2.1, below); if we choose any fixed order $k < \infty$, the rate of convergence will again be $n^{-2k/(2k+1)}$. But the rate of convergence can be improved if we let $k \to \infty$, as the number of observations increases (Theorem 3.1, Corollary 3.1). This idea is consistent with practical considerations: Since constants in the leading expressions of MSE depending on the kernel increase with increasing k, the improvement in the rate of convergence will lead to a smaller MSE only in large samples, and the larger k is, the larger the sample has to be. For instance, even if we assume only $g \in \mathscr{C}^{10}([0, 1])$, for usual sample sizes of $n = 25$-100 it is not reasonable to exploit this smoothness fully by choosing a kernel of order 10; compare also simulation results of Gasser *et al.* [4]. Therefore, assum-

ing $g \in \mathscr{C}^{\infty}([0, 1])$, for any n there will be a MSE optimal finite value for the order $k(n)$ as there is an optimal bandwidth b for any given k and n.

Assuming certain growth conditions for $g^{(k)}$ as $k \to \infty$, we will show in Section 3 that by letting $k(n) \to \infty$ $(n \to \infty)$ at a specific rate, we obtain a MSE rate better than any of the rates $n^{-2k/(2k+1)}$, k fixed. In Section 2 we investigate the behavior of two classes of kernels of order k as $k \to \infty$, in order to assess the behavior of kernel-dependent constants determining MSE for large k.

2. Asymptotic Properties of Two Classes of Kernels

We discuss the asymptotic behavior of two classes of kernels. These kernels with support $[-1, 1]$ are defined as solutions of the variational problems

$$\int K^{(\mu)}(x)^2 \, dx = \min! \text{ under } K \in \mathscr{A}_k \cap \mathscr{C}^{\mu}([-1, 1]) \text{ and}$$

$$K^{(j)}(-1) = K^{(j)}(1) = 0, \qquad j = 0, \ldots, \mu - 1. \tag{6}$$

We consider this problem for $\mu = 0, 1$ and denote the solution by 0-optimal (or minimum variance) kernels $\mathscr{S}_k (\mu = 0)$ and by 1-optimal (or just optimal) kernels $\psi_k (\mu = 1)$. In the following, we assume that k is even. The general variational problem was discussed in [7] and further special formulas for the cases $\mu = 0, 1$ were derived in [4]. In order to apply these formulas, we need an asymptotic approximation for binomial coefficients.

Lemma 1. $(\sqrt{k}/2^k)\binom{k}{k/2} \to \sqrt{2/\pi}$ as $k \to \infty$.

Proof. Apply Stirling's formula.

The functionals of any kernel $K \in \mathscr{A}_k$ that have an influence on MSE are, according to (4), (5), $V(K) := \int_{-1}^{1} K^2(x) \, dx$ and $B_k(K) := \int_{-1}^{1} K(x) x^k \, dx$.

The asymptotic behavior of these functionals for the solutions of (6) for $\mu = 0, 1$ is given in the following:

Lemma 2. *As $k \to \infty$, we obtain*

(i) $2^k |B_k(\mathscr{S}_k)| \to \sqrt{2}$;

(ii) $V(\mathscr{S}_k)/k \to 1/\pi$;

(iii) $2^{k+1} |B_k(\psi_k)| \to \sqrt{2}$;

(iv) $V(\psi_k)/k \to 1/\pi$;

(v) $\int_{-1}^{1} \psi_k'(x)^2 \, dx / k^2(k+2) \to 1/(3\pi)$.

Proof. (i) By formula (5), Theorem 1 of [4], we obtain

$$|B_k(\mathscr{S}_k)| = \binom{k}{k/2} \Big/ \binom{2k}{k},$$

and the result follows from Lemma 1.

(ii) By formula (6), Theorem 1 of [4],

$$V(\mathscr{S}_k) = \frac{k^2}{2^{2k+1}} \binom{k}{k/2}^2.$$

Again, the result follows from Lemma 1.

(iii) Follows in the same way as formula (11) of [4]:

$$|B_k(\psi_k)| = \frac{k+1}{2k+1} \binom{k}{k/2} \Big/ \binom{2k}{k}.$$

(iv) Follows from formula (12) of [4]:

$$V(\psi_k) = \frac{k+1}{2k+1} \frac{k^2}{2^{2k}} \binom{k}{k/2}^2.$$

(v) By partial integration, we find that ψ_k' is a kernel function satisfying

$$\int_{-1}^{1} \psi_k'(x)\, x^j\, dx = \begin{cases} 0, & j = 0, 2, \ldots, k \\ -1, & j = 1 \end{cases}$$

and minimizes $\int_{-1}^{1} \psi_k'(x)^2\, dx$ according to the definition of ψ_k. Therefore, ψ_k' corresponds to a minimum variance kernel (with $\nu = 1$, $k' = k+1$, $\mu = 0$ in the notation of [7]). Formula (6) of [4] yields for this kernel

$$\int_{-1}^{1} \psi_k'(x)^2\, dx = \frac{k^2}{3} \left(\frac{k+2}{2}\right)^2 \frac{1}{2^{k+3}} \binom{k+2}{(k+2)/2},$$

and the result follows from Lemma 1.

Next we show that it is not possible to find a kernel corresponding to the smoothness of a $\mathscr{C}^\infty$ function.

THEOREM 1. *There is no kernel of order* $k = \infty$, *i.e.*, $\mathscr{M}_\infty := \bigcap_{i=1}^{\infty} \mathscr{M}_i = \varnothing$.

Proof. Assume that $K_\infty \in \mathscr{M}_\infty$. It follows that for any k, $K_\infty \in \mathscr{M}_k$. According to the definition of the $\mathscr{S}_k$ as solution of (6) for $\mu = 0$, it holds that $\int K_k(x)^2\, dx \geqslant \int \mathscr{S}_k(x)^2\, dx$ for any kernel $K_k \in \mathscr{M}_k$, and therefore

$\int_{-1}^{1} K_\infty(x)^2\,dx \geq \int_{-1}^{1} \mathscr{S}_k(x)^2\,dx$ for any k. But by Lemma 2.2(ii), $V(\mathscr{S}_k) \to \infty$ as $k \to \infty$, which implies $K_\infty \notin L_2([-1,1])$. This is a contradiction since $\mathscr{M}_\infty \subset \mathrm{Lip}([-1,1]) \subset L_2([-1,1])$.

3. Rates of Convergence of Mean Squared Error

According to Theorem 1, the only possibility to exploit the smoothness of $\mathscr{C}^\infty$ functions with L_1 kernels is to let k depend on n such that $k(n) \to \infty$ as $n \to \infty$. In view of the nearly identical asymptotic behavior of kernels $\mathscr{S}_k$ and ψ_k according to Lemma 2, we consider in the following only kernels ψ_k that solve (6) for $\mu = 1$. Considering varying k, we obtain for the bias:

Lemma 3. *Let* $k = k(n) \to \infty$ *and*

$$\sup_{n \in \mathbb{N}} b \leqslant s < \infty, \qquad 0 < t - s < t + s < 1. \tag{7}$$

There exists $\xi_{b,k} \in [t-b, t+b]$ *such that*

$$Eg_n(t) - g(t) = b^k g^{(k)}(\xi_{b,k})\, B_k(\psi_k) + O\left(\frac{\sqrt{k}}{n}\right). \tag{8}$$

Proof.

$$\begin{aligned}
\left| Eg_n(t) - \int_{-1}^{1} \psi_k(x)\, g(t - xb)\, dx \right| &= \left| \frac{1}{b} \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} \psi_k\left(\frac{t-u}{b}\right) du\, g(t_i) \right. \\
&\quad \left. - \frac{1}{b} \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} \psi_k\left(\frac{t-u}{b}\right) g(u)\, du \right| \\
&\leqslant \frac{1}{nb} \int_0^1 \left| \psi_k\left(\frac{t-u}{b}\right) \right| du \\
&\leqslant \frac{\sqrt{2}}{n} \left(\int_{-1}^{1} |\psi_k(x)|^2\, dx \right)^{1/2} \\
&= O(\sqrt{k}/n)
\end{aligned}$$

by Lemma 2(iv). Therefore

$$|Eg_n(t) - g(t)| = O(\sqrt{k}/n) + \left| \int_{-1}^{1} \psi_k(x)\, g(t - xb)\, dx - g(t) \right|$$

and the result follows from a Taylor expansion of g around t with the Lagrange remainder term.

As for the variance, we get the following approximation.

LEMMA 4.

$$\operatorname{var}(g_n(t)) = \frac{\sigma^2}{nb} \int_{-1}^{1} \psi_k(x)^2 \, dx + O(k^2/(n^2b^2)).$$

Proof. With mean values η_i, ζ_i

$$\begin{aligned} &\left| \operatorname{var} g_n(t) - \frac{\sigma^2}{nb} \int_{-1}^{1} \psi_k(x)^2 \, dx \right| \\ &\quad \leqslant \frac{1}{n^2b^2} \sum_{i=1}^{n} \left| \psi_k^2\left(\frac{t-\eta_i}{b}\right) - \psi_k^2\left(\frac{t-\zeta_i}{b}\right) \right| \\ &\quad \leqslant \frac{1}{n^2b^2} T(\psi_k^2), \end{aligned}$$

where T denotes total variation. For any function $f \in \mathcal{C}^1([0, 1])$, $T(f) \leqslant \int |f'(x)| \, dx$. Therefore

$$\begin{aligned} T(\psi_k^2) &\leqslant 2 \left(\int \psi_k(x)^2 \, dx \right)^{1/2} \left(\int \psi_k'(x)^2 \, dx \right)^{1/2} \\ &= O(k^2) \end{aligned}$$

by Lemma 2(iv), (v).

Combining Lemmas 2(iii), (iv), 3, and 4, we obtain for the MSE (observing (8)):

THEOREM 2. *Under the assumptions of Lemma* 3,

$$\begin{aligned} E(g_n(t) - g(t))^2 = g^{(k)}(\xi_{b,k})^2 \frac{b^{2k}}{k!^2 \, 2^{2k+1}} (1 + o(1)) + \frac{k}{nb} \frac{\sigma^2}{\pi} (1 + o(1)) \\ + O\left(\frac{\sqrt{k}}{n} \frac{b^k}{k! \, 2^k} \right) + O\left(\frac{k^2}{n^2b^2} \right). \end{aligned} \tag{9}$$

We observe that $E(g_n(t) - g(t))^2 \to 0$ as $n \to \infty$ implies that the O-terms disappear, since then $k/(nb) \to 0$ and since the first O-term is $o(1/n)$. A condition for achieving the rate n^{-1} (which, e.g., is obtained in parametric regression models, if the model fits the data) is obtained as a consequence of Theorem 2.

COROLLARY 1. *The rate* n^{-1} *for* MSE *of* $g_n(t)$ *is attained iff there exist* k_0 *and* b_0 *such that* $g^{(k_0)}(\xi_{b_0,k_0}) = 0$.

Proof. If the condition is satisfied, we fix $k = k_0$ and $b = b_0$ for all n. The bias squared disappears and the variance decreases as n^{-1}. If MSE $\sim n^{-1}$,

we must have bias squared $\sim n^{-1}$ and var $\sim n^{-1}$. The latter implies that there exist k_1 and b_1 such that $\sup_{n\in\mathbb{N}} k(n)\leqslant k_1$ and $\inf_{n\in\mathbb{N}} b(n)\geqslant b_1$. Therefore there exists k_0 such that infinitely many of the $\xi_{b,k}$ can be written as ξ_{b,k_0}. Further, bias squared $\sim n^{-1}$ requires that $g^{(k_0)}(\xi_{b,k_0})=O(n^{-1})$, where $b=b(n)\in[s,b_1]$. Now $\{b(n)\mid n\in\mathbb{N}\}$ has a subsequence b^* with a limit b_0. Assume that $g^{(k_0)}(\xi_{b_0,k_0})\neq 0$. It follows that $\lim_{n\to\infty}\xi_{b^*,k_0}=\xi_{b_0,k_0}$ and therefore $g^{(k_0)}(\xi_{b_0,k_0})=0$.

The condition for achieving the MSE-rate n^{-1} is satisfied, e.g., if g is a polynomial.

In general, this condition will not be satisfied and the rate n^{-1} will be unattainable. The optimal rate of convergence of MSE then has to be achieved for variance as well as for bias squared. Setting $c_{b,k}:=g^{(k)}(\xi_{b,k})$, this yields the condition

$$c_{b,k}\frac{b^{2k}}{k!^2\,2^{2k}}\sim\frac{k}{nb}. \tag{10}$$

Applying Stirling's formula $k!=k^{k+1/2}e^{-k}(2\pi)^{1/2}e^{\theta/12k}$ for some $0<\theta<1$, we see that this is equivalent to

$$\left(\frac{2k}{e}\right)^{2k}\frac{1}{c_{b,k}}\sim nb^{2k+1}. \tag{11}$$

If we choose, e.g., $b\sim(1/n)^{1/(2k+1)}$ as after (5), we conclude that $\sup_{n\in\mathbb{N}} k\leqslant k_0$ and the rate of convergence of MSE becomes $n^{-2k/(2k+1)}$. Assume now that

$$0<\inf_{n\in\mathbb{N}} c_{b,k}<\sup_{n\in\mathbb{N}} c_{b,k}<\infty. \tag{12}$$

Setting $\lambda=k/b$, we see that (11) is equivalent to

$$\left(\frac{2}{e}\lambda\right)^{2\lambda b}\sim nb \qquad \text{or} \qquad \left(\frac{2}{e}\lambda\right)^{2\lambda}\sim n^{1-b}, \tag{13}$$

and the rate of convergence of MSE is then λ/n. Obviously, this rate is fastest if $\inf_{n\in\mathbb{N}} b>0$. Then we obtain the following result.

COROLLARY 2. *Assume that* (12) *holds. Then the optimal rate of convergence of the* MSE *is obtained if* $\inf_{n\in\mathbb{N}} b>0$ *and* k *is chosen in such a way that* $(2k/e)^{2k}\sim n$. *This rate is then* k/n.

Assumption (12) covers a broad class of functions. Knowledge of the behavior of the $c_{b,k}$ may lead to other optimal rates. The best possible rate n^{-1} can be only achieved under the condition of Corollary 1.

An open question is how to choose the order k in practical situations. An obvious proposal is to employ an estimator of IMSE (k, b) like cross-validation or an estimator proposed by Rice [8], minimize this w.r.t. b for various values of k, i.e., various kernel orders, and choose the k which yields the minimal value.

REFERENCES

1. M. S. BARTLETT, Statistical estimation of density functions, *Sankhyā Ser. A* **25** (1963), 245–254.
2. K. B. DAVIS, Mean square error properties of density estimates, *Ann. Statist.* **3** (1975), 1025–1030.
3. K. B. DAVIS, Mean integrated square error properties of density estimates, *Ann. Statist.* **5** (1977), 530–535.
4. TH. GASSER, H. G. MÜLLER, AND V. MAMMITZSCH, Kernels for nonparametric curve estimation, *J. Roy. Statist. Soc. Ser. B* **47** (1985), 238–252.
5. G. HALASZ, Statistical interpolation *in* "Fourier Analysis and Approximation Theory (Budapest, 1976), Vol. I." (G. Alexits and P. Twan Eds.), pp. 403–410, North-Holland, Amsterdam, 1978.
6. I. A. IBRAGIMOV AND R. Z. KHAS'MINSKI, Estimation of distribution density belonging to a class of entire functions, *Theoret. Probab. Appl.* **27** (1982), 551–562.
7. H. G. MÜLLER, Smooth optimum kernel estimators of regression curves, densities and modes, *Ann. Statist.* **12** (1984), 766–774.
8. J. RICE, Bandwidth choice for nonparametric kernel regression, *Ann. Statist.* **12** (1984), 1215–1231.
9. M. ROSENBLATT, Curve estimates, *Ann. Math. Statist.* **42** (1971), 1815–1842.
10. J. S. SHAPIRO, "Smoothing and Approximation of Functions." Van Nostrand, New York, 1969.
11. C. J. STONE, Optimal rates of convergence for nonparametric estimators, *Ann. Statist.* **8** (1980), 1348–1360.