APPLICATIONES MATHEMATICAE
28,3 (2001), pp. 257–270

Waldemar Popiński (Warszawa)

ON ORTHOGONAL SERIES ESTIMATION OF BOUNDED REGRESSION FUNCTIONS

Abstract. The problem of nonparametric estimation of a bounded regression function $f \in L^2([a,b]^d)$, $[a,b] \subset \mathbb{R}$, $d \ge 1$, using an orthonormal system of functions e_k, $k = 1,2,\ldots,$ is considered in the case when the observations follow the model $Y_i = f(X_i) + \eta_i$, $i = 1,\ldots,n$, where X_i and η_i are i.i.d. copies of independent random variables X and η, respectively, the distribution of X has density ϱ, and η has mean zero and finite variance. The estimators are constructed by proper truncation of the function $\widehat{f}_n(x) = \sum_{k=1}^{N(n)} \widehat{c}_k e_k(x)$, where the coefficients $\widehat{c}_1,\ldots,\widehat{c}_{N(n)}$ are determined by minimizing the empirical risk $n^{-1}\sum_{i=1}^n (Y_i - \sum_{k=1}^{N(n)} c_k e_k(X_i))^2$. Sufficient conditions for convergence rates of the generalization error $E_X|f(X) - \widehat{f}_n(X)|^2$ are obtained.

1. Introduction. Let observations Y_i, $i = 1,\ldots,n$, follow the model $Y_i = f(X_i) + \eta_i$, where $f \in L^2([a,b]^d)$, $[a,b] \subset \mathbb{R}$, $d \ge 1$, is an unknown regression function, the errors η_i, $i = 1,\ldots,n$, are i.i.d. copies of a random variable with zero mean value and finite variance σ_η^2, and X_i, $i = 1,\ldots,n$, form a sample from the distribution of a random variable X ranging over a compact subset $[a,b]^d$ of some euclidean space $\mathbb{R}^d$, $d \ge 1$. We further assume that the distribution of X is absolutely continuous with density ϱ and that the vector random variables $X_1^n = (X_1,\ldots,X_n)$ and $\eta_1^n = (\eta_1,\ldots,\eta_n)$ are independent.

Let functions f_k, $k = 1,2,\ldots,$ be analytic in (a,b) and constitute a complete orthonormal system in $L^2[a,b]$. Then the functions $e_{k_1\ldots k_d}(x_1,\ldots,x_d) = f_{k_1}(x_1)\ldots f_{k_d}(x_d)$, where $x_l \in [a,b]$, $l = 1,\ldots,d$, $k_1,\ldots,k_d = 1,2,\ldots,$ are a complete orthonormal system in $L^2([a,b]^d)$. We assume that e_k, $k =$

2000 *Mathematics Subject Classification*: 62G08, 62G20.

Key words and phrases: orthonormal system, least squares method, nonparametric regression, weak and strong consistency, convergence rate.

$1, 2, \ldots$, is a complete orthonormal system in $L^2([a,b]^d)$ constructed in this way. The well known system of trigonometric functions in $L^2([0,2\pi]^d)$ is an example of an orthonormal system satisfying the above requirements, and another one can be constructed in $L^2([-1,1]^d)$ using tensor products of Legendre polynomials.

In this work we assume that the regression function f is bounded (i.e. $|f| \le L < \infty$) and examine asymptotic properties of the *generalization error*

$$E_X(f(X) - \overline{f}_n(X))^2 = \int\limits_{[a,b]^d} (f(x) - \overline{f}_n(x))^2 \varrho(x)\, dx$$

for estimators $\overline{f}_n$ which are defined in the following two steps:

In the first step we determine, for a fixed N, the vector of coefficients $\widehat{c}^N = (\widehat{c}_1, \ldots, \widehat{c}_N)^T$ by minimizing the empirical risk:

$$\widehat{c}^N = \arg \min_{c \in \mathbb{R}^N} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle c, e^N(X_i) \rangle)^2,$$

where $e^N(x) = (e_1(x), \ldots, e_N(x))^T$.

If the functions $e_k,\ k = 1, 2, \ldots$, are constructed as tensor products of functions $f_l,\ l = 1, 2, \ldots$, orthogonal in $L^2[a,b]$ and analytic in (a,b), as assumed, then for $N \le n$ the vector $\widehat{c}^N$ can be uniquely determined with probability one as the solution of the normal equations

$$\widehat{c}^N = G_n^{-1} g_n, \tag{1}$$

where

$$G_n = \frac{1}{n} \sum_{i=1}^{n} e^N(X_i) e^N(X_i)^T, \qquad g_n = \frac{1}{n} \sum_{i=1}^{n} Y_i e^N(X_i).$$

This follows from Lemma 2.2 of [11] which assures that the matrices G_n are almost surely positive definite for $N \le n$ when $X_i,\ i = 1, \ldots, n$, form a random sample from a distribution with density $\varrho \in L^1([a,b]^d)$.

In the second step we construct the estimator $\overline{f}_n$ by truncating the function $\widehat{f}_n(x) = \sum_{k=1}^{N} \widehat{c}_k e_k(x)$ at B_n and $-B_n$, i.e.

$$\overline{f}_n(x) = (\widehat{f}_n(x) \vee (-B_n)) \wedge B_n \quad \text{for } x \in [a,b]^d, \tag{2}$$

where we use the notation $a \vee b := \max\{a,b\},\ a \wedge b := \min\{a,b\}$ for $a, b \in \mathbb{R}$, with constant $B_n \in \mathbb{R}_+$ depending only on $n,\ B_n \to \infty$ as $n \to \infty$.

In author's earlier works, consistency in the sense of the generalization error [12], [13] and its convergence rates [14] for orthogonal series estimators were investigated only in the case when the density ϱ satisfies the condition $0 < c \le \varrho$. Similarly, Huang [6] obtained convergence rates for such estimators only for $0 < c \le \varrho \le D < \infty$. In the present work we prove that the truncated orthogonal series estimators are consistent in the sense of the generalization error and give their convergence rates for the observation model with bounded density $\varrho \le D < \infty$ and bounded regression functions. In order to obtain the convergence rates for an arbitrary density ϱ we have to assume that the regression function can be approximated in the supremum norm by linear combinations of the functions $e_k,\ k = 1, 2, \ldots$

Results concerning weak and strong universal consistency of series type regression estimators were obtained by Lugosi and Zeger [8] for a more general observation model, where i.i.d. realizations of a pair of random variables (X, Y) are given, even without the assumption of absolute continuity of the predictor variable distribution. However, they consider estimators of the form $\widehat{g}_n(x) = \sum_{k=1}^{N} \widehat{a}_k e_k(x)$, where the coefficients $\widehat{a}_1, \ldots, \widehat{a}_N$ are determined by minimizing the empirical risk $n^{-1} \sum_{i=1}^{n} (Y_i - \sum_{k=1}^{N} a_k e_k(X_i))^2$ under the constraint $\sum_{k=1}^{N} |a_k| \le \beta_n,\ \beta_n \to \infty$.

As remarked by Györfi and Walk [4], obtaining the empirically optimal estimator $\widehat{g}_n$ is difficult if the minimum is not unique. The same remark holds for radial basis function estimators investigated in Niyogi and Girosi [9], obtained by similar constrained empirical risk minimization, for which convergence rates of the generalization error were given.

Our estimators are almost surely uniquely determined and can be constructed by solving a system of linear equations, which may also reduce computation time in comparison to constrained empirical risk minimization. Thus, the approach applied in this work, similarly to [4], aims at solving the numerical difficulties which appear in obtaining the estimators using constrained empirical risk minimization.

Results obtained for other approaches to nonparametric regression function estimation giving weakly and universally consistent estimators are discussed in [4], [7].

In Section 2 we give an overview of results of the Vapnik–Chervonenkis theory which are necessary to prove the results of the present work. Our results are formulated and proved in Sections 3 and 4 concerning, respectively, the weak and strong consistency of the relevant estimators in the sense of the generalization error.

2. Some results of the Vapnik–Chervonenkis theory. In this section we list the definitions and basic results of the Vapnik–Chervonenkis theory which we use in the following sections to prove our results.

Let us start with the definition of covering numbers of function classes.

Definition 2.1. Let F be a class of functions $f : \mathbb{R}^d \to \mathbb{R}$. The *covering number* $\aleph(\varepsilon, F, z_1^n)$ is defined for any $\varepsilon > 0$ and $z_1^n = (z_1, \ldots, z_n) \in \mathbb{R}^{dn}$ as the smallest integer k such that there exist functions $g_1, \ldots, g_k : \mathbb{R}^d \to \mathbb{R}$

with

$$\min_{1\le j\le k}\frac{1}{n}\sum_{i=1}^{n}|f(z_i)-g_j(z_i)|\le\varepsilon$$

for each $f\in F$.

If $Z_1^n=(Z_1,\dots,Z_n)$ is a sequence of $\mathbb{R}^d$-valued random variables, then $\aleph(\varepsilon,F,Z_1^n)$ is a random variable with expected value $E\aleph(\varepsilon,F,Z_1^n)$. The next result due to Pollard is the main tool in the proof of our results.

Lemma 2.1 (Pollard [10], Section II.5, Theorem 24). *Let F be a class of functions $f:\mathbb{R}^d\to[0,B]$, and let $Z_1^n=(Z_1,\dots,Z_n)$ be $\mathbb{R}^d$-valued i.i.d. random variables. Then for any $\varepsilon>0$,*

$$P\Big\{\sup_{f\in F}\Big|\frac{1}{n}\sum_{i=1}^{n}f(Z_i)-Ef(Z_1)\Big|>\varepsilon\Big\}\le 8E\Big(\aleph\Big(\frac{\varepsilon}{8},F,Z_1^n\Big)\Big)\exp\Big(-\frac{n\varepsilon^2}{128B^2}\Big).$$

In order to bound covering numbers we need the notion of the VC-dimension which is defined as follows.

Definition 2.2. Let Λ be a class of subsets of $\mathbb{R}^d$ and let $S\subseteq\mathbb{R}^d$. One says that Λ *shatters* S if each subset of S has the form $\lambda\cap S$ for some $\lambda\in\Lambda$. The *VC-dimension* V_Λ of Λ is defined as the largest integer k for which there exists a set of cardinality k shattered by Λ.

A connection between covering numbers and the VC-dimension is given by the following lemma, which uses the notation V_{F^+} for the VC-dimension of the set class

$$F^+:=\{\{(x,t)\in\mathbb{R}^d\times\mathbb{R}:t\le f(x)\}:f\in F\}.$$

Lemma 2.2 (Haussler [5], Theorem 6). *Let F be a class of functions $f:\mathbb{R}^d\to[-B,B]$. Then for any $z_1^n=(z_1,\dots,z_n)\in\mathbb{R}^{dn}$ and any $\varepsilon>0$,*

$$\aleph(\varepsilon,F,z_1^n)\le 2\Big(\frac{4eB}{\varepsilon}\ln\Big(\frac{4eB}{\varepsilon}\Big)\Big)^{V_{F^+}}.$$

For $B>0$ define $T_B:\mathbb{R}\to\mathbb{R}$ by the formula

$$T_B(t)=\begin{cases}B & \text{if } t>B,\\ t & \text{if } |t|\le B,\\ -B & \text{if } t<-B,\end{cases}$$

and let $T_BF=\{T_B\circ f:f\in F\}$ be the set of truncated functions from F. The following lemma gives a bound on the VC-dimension of a class of truncated functions.

Lemma 2.3. *Let F be a class of functions $f:\mathbb{R}^d\to\mathbb{R}$ and $B>0$. Then for the class of functions $T_BF=\{T_B\circ f:f\in F\}$ we have $V_{T_BF^+}\le V_{F^+}$.*

Proof. It is enough to show that every set shattered by T_BF^+ is also shattered by F^+. Suppose there exists a sequence $((x_1,t_1),\ldots,(x_m,t_m))$ which is shattered by T_BF^+. By definition, this means that for every boolean vector $b\in\{0,1\}^m$ there is some function $g_b=T_B\circ f_b$ ($f_b\in F$) satisfying $g_b(x_i)\ge t_i$ if and only if $b_i=1$ for $i=1,\ldots,m$. We now show that the set $((x_1,t_1),\ldots,(x_m,t_m))$ is shattered by F^+. Set

$$a_i=\min_b\{g_b(x_i)=T_B(f_b(x_i))\mid b_i=1\},$$
$$A_i=\max_b\{g_b(x_i)=T_B(f_b(x_i))\mid b_i=0\},$$

for $i=1,\ldots,m$. Since T_BF^+ shatters $((x_1,t_1),\ldots,(x_m,t_m))$, we have $-B\le A_i<t_i\le a_i\le B$. Now, by construction of T_B,

$$f_b(x_i)\ge t_i\equiv T_B(f_b(x_i))\ge t_i\equiv b_i=1\qquad\text{for } i=1,\ldots,m.$$

Since each $f_b\in F$, we see that $((x_1,t_1),\ldots,(x_m,t_m))$ is shattered by F^+. ■

The following result is often useful for bounding the VC-dimension.

Lemma 2.4 (Dudley [2]). *Let F be a k-dimensional vector space of functions $f:\mathbb{R}^d\to\mathbb{R}$. Then the class of sets of the form $\{x\in\mathbb{R}^d: f(x)\ge 0\}$, where $f\in F$, has the VC-dimension less than or equal to k.*

For simplicity of notation let T_nF denote the class of truncated functions $T_{B_n}F$. Assume now that we are given a random sample $(X_1,Y_1),\ldots,(X_n,Y_n)$ from the distribution of (X,Y), where X is the predictor variable from the observation model and Y is a real-valued bounded random variable.

Lemma 2.5. *Let $|Y|\le L<\infty$ and $F_N=\operatorname{span}\{e_1,\ldots,e_N\}$. Then for $\varepsilon>0$ and sufficiently large n,*

$$P\bigg\{\sup_{g\in T_nF_N}\bigg|\frac{1}{n}\sum_{i=1}^n(Y_i-g(X_i))^2-E(Y-g(X))^2\bigg|>\varepsilon\bigg\}$$
$$\le 2^4\bigg(\frac{2^7eB_n^2}{\varepsilon}\ln\bigg(\frac{2^7eB_n^2}{\varepsilon}\bigg)\bigg)^{N+1}\exp\bigg(-\frac{n\varepsilon^2}{2^{11}B_n^4}\bigg).$$

Proof. Consider the class of functions $H_n=\{h(x,y)=(y-g(x))^2: [a,b]^d\times\mathbb{R}\to\mathbb{R}\mid g\in T_nF_N\}$. Since $B_n\to\infty$ we have $|Y|\le B_n$ for sufficiently large n. Consequently, $(y-g(x))^2\le 4B_n^2$ for $g\in T_nF_N$ and $x\in[a,b]^d$. Applying Lemma 2.1 to the function class H_n and random variables $Z_i=(X_i,Y_i),\ i=1,\ldots,n$, we obtain the bound

$$P\bigg\{\sup_{g\in T_nF_N}\bigg|\frac{1}{n}\sum_{i=1}^n(Y_i-g(X_i))^2-E(Y-g(X))^2\bigg|>\varepsilon\bigg\}$$
$$\le 8E\bigg(\aleph\bigg(\frac{\varepsilon}{8},H_n,Z_1^n\bigg)\bigg)\exp\bigg(-\frac{n\varepsilon^2}{2048B_n^4}\bigg).$$

Now, we will bound the covering number in the above inequality. Observe first that if $h_j(x,y)=(y-g_j(x))^2$, $g_j\in T_nF_N$ for $j=1,2$, then

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}&|h_1(X_i,Y_i)-h_2(X_i,Y_i)|\\ &=\frac{1}{n}\sum_{i=1}^{n}|2Y_i-g_1(X_i)-g_2(X_i)|\cdot|g_1(X_i)-g_2(X_i)|\\ &\le 4B_n\frac{1}{n}\sum_{i=1}^{n}|g_1(X_i)-g_2(X_i)|.\end{aligned}$$

In consequence,

$$\aleph\left(\frac{\varepsilon}{8},H_n,Z_1^n\right)\le\aleph\left(\frac{\varepsilon}{32B_n},T_nF_N,X_1^n\right),$$

so using the notion of the VC-dimension and Lemmas 2.2 and 2.3 we further obtain

$$\begin{aligned}\aleph\left(\frac{\varepsilon}{32B_n},T_nF_N,X_1^n\right)&\le 2\left(\frac{128eB_n^2}{\varepsilon}\ln\left(\frac{128eB_n^2}{\varepsilon}\right)\right)^{V_{T_nF_N^+}}\\ &\le 2\left(\frac{128eB_n^2}{\varepsilon}\ln\left(\frac{128eB_n^2}{\varepsilon}\right)\right)^{V_{F_N^+}}.\end{aligned}$$

Since F_N is a linear function space of dimension N we also have (see Lemma 2.4) $V_{F_N^+}\le N+1$, which completes the proof. ■

Remark 2.1. To ensure measurability of the supremum in the above lemma it is necessary to impose regularity conditions on uncountable collections of functions F. For the finite-dimensional spaces $F_N=\operatorname{span}\{e_1,\dots,e_N\}$ one can use the fact that every linear combination of functions $e_1,e_2,\dots$ is a pointwise limit of linear combinations with rational coefficients (see Pollard [10], p. 38).

3. Weak convergence of the generalization error.

In this section we examine convergence rates in probability of the generalization error for the truncated orthogonal series estimators $\bar{f}_n$ defined by (1) and (2) for $N\le n$. It follows from the construction that $\bar{f}_n\in T_nF_N$, where $F_N=\operatorname{span}\{e_1,\dots,e_N\}$. We start by proving the following lemma.

Lemma 3.1. *Assume that the regression function f is bounded, $N\le n$, and $\varepsilon>0$. Then for $h_N(x)=\langle c^N,e^N(x)\rangle$, $c^N\in\mathbb{R}^N$, and sufficiently large n,*

$$\begin{aligned}P\Big\{E_\eta E_X(f(X)-\bar{f}_n(X))^2>\frac{1}{n}\sum_{i=1}^{n}(f(X_i)-h_N(X_i))^2+\sigma_\eta^2\frac{N}{n}+\varepsilon\Big\}&\\ \le 2^4\left(\frac{2^7eB_n^2}{\varepsilon}\ln\left(\frac{2^7eB_n^2}{\varepsilon}\right)\right)^{N+1}\exp\left(-\frac{n\varepsilon^2}{2^{11}B_n^4}\right).&\end{aligned}$$

Proof. Putting

$$D_n = 2^4\biggl(\frac{2^7eB_n^2}{\varepsilon}\ln\biggl(\frac{2^7eB_n^2}{\varepsilon}\biggr)\biggr)^{N+1}\exp\biggl(-\frac{n\varepsilon^2}{2^{11}B_n^4}\biggr)$$

and applying Lemma 2.5 to the random variables $(X_i, f(X_i))$, $i = 1,\dots,n$, we see that for $\varepsilon > 0$ and sufficiently large n,

$$P\biggl\{\sup_{g\in T_nF_N}\biggl|\frac{1}{n}\sum_{i=1}^n(f(X_i)-g(X_i))^2 - E_X(f(X)-g(X))^2\biggr| > \varepsilon\biggr\} \le D_n.$$

Since $\overline{f}_n(x) = T_{B_n}(\widehat{f}_n(x))$, where $\widehat{f}_n(x) = \sum_{k=1}^N \widehat{c}_k e_k(x)$, by the inequality

$$\biggl|\frac{1}{n}\sum_{i=1}^n E_\eta(f(X_i)-\overline{f}_n(X_i))^2 - E_\eta E_X(f(X)-\overline{f}_n(X))^2\biggr| \le E_\eta\biggl|\frac{1}{n}\sum_{i=1}^n(f(X_i)-\overline{f}_n(X_i))^2 - E_X(f(X)-\overline{f}_n(X))^2\biggr|$$

we also have

$$P\biggl\{\biggl|\frac{1}{n}\sum_{i=1}^n E_\eta(f(X_i)-\overline{f}_n(X_i))^2 - E_\eta E_X(f(X)-\overline{f}_n(X))^2\biggr| > \varepsilon\biggr\} \le D_n$$

for n such that $|f| \le B_n$ (see Lemma 2.5). Furthermore, the obvious inequality

$$(f(X_i)-\overline{f}_n(X_i))^2 \le (f(X_i)-\widehat{f}_n(X_i))^2$$

implies

$$P\biggl\{E_\eta E_X(f(X)-\overline{f}_n(X))^2 > \frac{1}{n}\sum_{i=1}^n E_\eta(f(X_i)-\widehat{f}_n(X_i))^2 + \varepsilon\biggr\} \le D_n. \tag{3}$$

The standard squared bias plus variance decomposition with respect to η variable yields

$$\begin{aligned}R_{nN} &= \frac{1}{n}\sum_{i=1}^n E_\eta(f(X_i)-\widehat{f}_n(X_i))^2\\ &= \frac{1}{n}\sum_{i=1}^n(f(X_i)-E_\eta\widehat{f}_n(X_i))^2 + \frac{1}{n}\sum_{i=1}^n E_\eta(\widehat{f}_n(X_i)-E_\eta\widehat{f}_n(X_i))^2.\end{aligned}$$

Taking into account (1) we obtain for $N \le n$,

$$\frac{1}{n}\sum_{i=1}^n E_\eta(\widehat{f}_n(X_i)-E_\eta\widehat{f}_n(X_i))^2 = \frac{1}{n}\sum_{i=1}^n E_\eta\biggl\langle e_N(X_i), G_n^{-1}\frac{1}{n}\sum_{j=1}^n \eta_j e^N(X_j)\biggr\rangle^2$$

$$= \frac{\sigma_\eta^2}{n^3}\sum_{i=1}^n\sum_{j=1}^n \langle e_N(X_i), G_n^{-1}e_N(X_j)\rangle^2 = \frac{\sigma_\eta^2}{n^2}\sum_{i=1}^n \langle e_N(X_i), G_n^{-1}e_N(X_i)\rangle$$

$$= \frac{\sigma_\eta^2}{n}\operatorname{Tr} G_n G_n^{-1} = \sigma_\eta^2\frac{N}{n},$$

which implies the equality

$$R_{nN} = \frac{1}{n}\sum_{i=1}^n (f(X_i) - E_\eta \widehat{f}_n(X_i))^2 + \sigma_\eta^2\frac{N}{n}.$$

Now, since for fixed observation points $X_i,\ i = 1,\dots,n$, we have

$$\frac{1}{n}\sum_{i=1}^n (f(X_i) - E_\eta \widehat{f}_n(X_i))^2 \le \frac{1}{n}\sum_{i=1}^n (f(X_i) - h_N(X_i))^2$$

for any linear combination $h_N = \sum_{k=1}^N c_k e_k$, $(c_1,\dots,c_N)^T \in \mathbb{R}^N$, we immediately obtain

$$R_{nN} \le \frac{1}{n}\sum_{i=1}^n (f(X_i) - h_N(X_i))^2 + \sigma_\eta^2\frac{N}{n}.$$

The above inequality together with (3) proves the lemma. ■

If the function f can be approximated in the supremum norm by linear combinations of the functions $e_k,\ k = 1,2,\dots$, the following theorem holds.

Theorem 3.1. *Assume that the regression function $f \in L^2([a,b]^d)$ is bounded and for $N = 1,2,\dots$, there exist $g_N \in \operatorname{span}\{e_1,\dots,e_N\}$ such that $\|f - g_N\|_\infty \to 0$ as $N \to \infty$. If the sequences of reals B_n and integers $N(n),\ n = 1,2,\dots$, satisfy*

$$N(n) \to \infty, \quad B_n \to \infty, \quad \frac{N(n)}{n} \to 0, \quad \frac{N(n)B_n^4\ln(n)}{n^{1-2\beta}} \to 0,$$

where $0 < \beta < 1/2$, then the estimator $\bar{f}_n$ constructed according to (1) *and* (2) *satisfies*

$$E_X(f(X) - \bar{f}_n(X))^2 = O_p\bigg(\|f - g_{N(n)}\|_\infty^2 + \sigma_\eta^2\frac{N(n)}{n} + n^{-\beta}\bigg).$$

Proof. Putting $\varepsilon = n^{-\beta}$ and applying Lemma 3.1 we have

$$P\bigg\{E_\eta E_X(f(X) - \bar{f}_n(X))^2 > \frac{1}{n}\sum_{i=1}^n (f(X_i) - g_N(X_i))^2 + \sigma_\eta^2\frac{N}{n} + n^{-\beta}\bigg\}$$

$$\le 2^4(2^7 e B_n^2 n^\beta \ln(2^7 e B_n^2 n^\beta))^{N+1}\exp\bigg(-\frac{n^{1-2\beta}}{2^{11}B_n^4}\bigg)$$

for $N \le n$ and sufficiently large n. As one can easily verify, the conditions imposed on the sequences B_n and $N(n)$ assure that the right hand side tends to zero as $n \to \infty$. In consequence,

$$P\left\{E_\eta E_X(f(X) - \bar{f}_n(X))^2 > \|f - g_{N(n)}\|_\infty^2 + \sigma_\eta^2 \frac{N(n)}{n} + n^{-\beta}\right\} \to 0,$$

and hence

$$E_X(f(X) - \bar{f}_n(X))^2 = O_p\left(\|f - g_{N(n)}\|_\infty^2 + \sigma_\eta^2 \frac{N(n)}{n} + n^{-\beta}\right)$$

which proves the theorem. ∎

In the case when the convergence rate of the uniform approximation error $\|f - g_{N(n)}\|_\infty$ is known, the following corollary is valid.

COROLLARY 3.1. *Assume that the regression function* $f \in L^2([a,b]^d)$ *is bounded and for* $N = 1, 2, \ldots,$ *there exist* $g_N \in \operatorname{span}\{e_1, \ldots, e_N\}$ *such that* $\|f - g_N\|_\infty = O(N^{-\alpha})$, *where* $\alpha > 0$. *If* $N(n) \sim n^{1/(1+2\alpha)}$ (*i.e.* $r_1 \geq N(n)n^{-1/(1+2\alpha)} \geq r_2$, $r_1, r_2 > 0$), *and the sequence of reals* B_n, $n = 1, 2, \ldots,$ *satisfies*

$$B_n \to \infty, \qquad \frac{B_n^4 \ln(n)}{n^\delta} \to 0,$$

where $0 < \delta < 2\alpha/(1+2\alpha)$, *then the estimator* $\bar{f}_n$ *constructed according to* (1) *and* (2) *satisfies*

$$E_X(f(X) - \bar{f}_n(X))^2 = O_p(n^{-\alpha/(1+2\alpha)+\delta/2}).$$

Proof. It follows from the assumptions that $\|f - g_{N(n)}\|_\infty^2 + \sigma_\eta^2 N(n)/n = O(n^{-2\alpha/(1+2\alpha)})$. Putting

$$\beta = \frac{1}{2}\left(\frac{2\alpha}{1+2\alpha} - \delta\right)$$

we have $1 - 2\beta = 1/(1+2\alpha) + \delta > 0$. The condition $B_n^4 \ln(n)/n^\delta \to 0$ implies $N(n)B_n^4 \ln(n)/n^{1-2\beta} \to 0$ and by Theorem 3.1,

$$E_X(f(X) - \bar{f}_n(X))^2 = O_p(n^{-2\alpha/(1+2\alpha)} + n^{-\beta}) = O_p(n^{-\beta}),$$

which completes the proof. ∎

The next theorem holds for bounded regression functions $f \in L^2([a,b]^d)$ which can be approximated only in the mean-square sense.

THEOREM 3.2. *Assume that the regression function* $f \in L^2([a,b]^d)$ *and density* ϱ *are bounded and for* $N = 1, 2, \ldots,$ *let* f_N *be the orthogonal projection of* f *on the subspace* $\operatorname{span}\{e_1, \ldots, e_N\}$. *If the sequences of reals* B_n *and integers* $N(n)$, $n = 1, 2, \ldots,$ *satisfy*

$$N(n) \to \infty, \quad B_n \to \infty, \quad \frac{N(n)}{n} \to 0, \quad \frac{N(n)B_n^4 \ln(n)}{n^{1-2\beta}} \to 0,$$

where $0<\beta<1/2$, *then the estimator* $\bar{f}_n$ *constructed according to* (1) *and* (2) *satisfies*

$$E_X(f(X)-\bar{f}_n(X))^2=O_p\biggl(\|f-f_{N(n)}\|+\sigma_\eta^2\frac{N(n)}{n}+n^{-\beta}\biggr).$$

Proof. Putting $\varepsilon=n^{-\beta}$ and following the proof of Theorem 3.1 we see that

$$P\biggl\{E_\eta E_X(f(X)-\bar{f}_n(X))^2>\frac{1}{n}\sum_{i=1}^n(f(X_i)-f_{N(n)}(X_i))^2+\sigma_\eta^2\frac{N(n)}{n}+n^{-\beta}\biggr\}$$

tends to zero as $n\to\infty$. Furthermore, since $\varrho\le D<\infty$ we have

$$E_X\frac{1}{n}\sum_{i=1}^n(f(X_i)-f_{N(n)}(X_i))^2=E_X(f(X)-f_{N(n)}(X))^2\le D\|f-f_{N(n)}\|^2,$$

which further yields

$$P\biggl\{\frac{1}{n}\sum_{i=1}^n(f(X_i)-f_{N(n)}(X_i))^2>\|f-f_{N(n)}\|\biggr\}\le D\|f-f_{N(n)}\|\to 0.$$

Thus,

$$P\biggl\{E_\eta E_X(f(X)-\bar{f}_n(X))^2>\|f-f_{N(n)}\|+\sigma_\eta^2\frac{N(n)}{n}+n^{-\beta}\biggr\}\to 0$$

as $n\to\infty$ and consequently

$$E_X(f(X)-\bar{f}_n(X))^2=O_p\biggl(\|f-f_{N(n)}\|+\sigma_\eta^2\frac{N(n)}{n}+n^{-\beta}\biggr). \blacksquare$$

For $f\in L^2([a,b]^d)$ we have $\|f-f_N\|\to 0$ as $N\to\infty$, so under the assumptions of Theorem 3.2, $E_X(f(X)-\widehat{f}_{N(n)}(X))^2=o_p(1)$, i.e. the relevant estimator is consistent in the sense of the generalization error. Inspection of the proof of Theorem 3.2 reveals that the same conclusion holds for an arbitrary density ϱ if the set of finite linear combinations of e_k, $k=1,2,\ldots,$ is dense in $L^2([a,b]^d,\mu)$ for any probability measure μ. Examples of orthogonal systems which satisfy this condition are given in the following section.

In the case when the convergence rate of the mean-square approximation of f by its Fourier series is known, the following corollary holds.

Corollary 3.2. *Assume that the regression function* $f\in L^2([a,b]^d)$ *and density* ϱ *are bounded and the* N*-term Fourier series* f_N *approximates* f *in* $L^2([a,b]^d)$ *at the rate* $\|f-f_N\|=O(N^{-\alpha})$, *where* $\alpha>0$. *If* $N(n)\sim n^{1/(1+\alpha)}$ (*i.e.* $r_1\ge N(n)n^{-1/(1+\alpha)}\ge r_2$, $r_1,r_2>0$), *and the sequence of reals* B_n, $n=1,2,\ldots,$ *satisfies*

$$B_n\to\infty,\qquad \frac{B_n^4\ln(n)}{n^\delta}\to 0,$$

where $0 < \delta < \alpha/(1+\alpha)$, *then the estimator* $\bar{f}_n$ *constructed according to* (1) *and* (2) *satisfies*

$$E_X(f(X) - \bar{f}_n(X))^2 = O_p(n^{-\alpha/(2(1+\alpha))+\delta/2}).$$

Proof. It follows from the assumptions that $\|f - f_{N(n)}\| + \sigma_\eta^2 N(n)/n = O(n^{-\alpha/(1+\alpha)})$. Putting

$$\beta = \frac{1}{2}\left(\frac{\alpha}{1+\alpha} - \delta\right)$$

we have $1 - 2\beta = 1/(1+\alpha) + \delta > 0$. The condition $B_n^4 \ln(n)/n^\delta \to 0$ implies $N(n)B_n^4 \ln(n)/n^{1-2\beta} \to 0$ and by Theorem 3.2,

$$E_X(f(X) - \bar{f}_n(X))^2 = O_p(n^{-\alpha/(1+\alpha)} + n^{-\beta}) = O_p(n^{-\beta}),$$

which is our claim. ■

As one can clearly see from Corollary 3.1, our convergence rates do not attain Stone's [15] bound on the best obtainable rate $n^{-2\alpha/(1+2\alpha)}$.

4. Strong convergence of the generalization error. Let us now consider the observation model where i.i.d. realizations (X_i, Y_i), $i = 1, \ldots, n$, of a pair of random variables (X, Y) are given, where X is the previously defined predictor variable and Y is real-valued and satisfies $|Y| \le L < \infty$ a.s. For such a model, estimators defined by (1) and (2) estimate the regression function $f(x) = E(Y \mid X = x)$. In this section we assume that the estimators are constructed using the system of trigonometric functions in $L^2([0,2\pi]^d)$ or algebraic polynomials in $L^2([-1,1]^d)$, respectively.

Let $B_nF_N := \{g \in F_N \mid \forall x \in [a,b]^d : |g(x)| \le B_n\}$ be the set of functions in $F_N = \operatorname{span}\{e_1, \ldots, e_N\}$ which are bounded in absolute value by B_n. The next lemma allows us to obtain sufficient conditions for strong consistency of the estimators considered.

LEMMA 4.1. *Assume that* $N(n) \to \infty$, $B_n \to \infty$ *as* $n \to \infty$,

$$\text{(i)} \qquad \sup_{g \in T_n F_{N(n)}} \left| \frac{1}{n} \sum_{i=1}^{n} (g(X_i) - Y_i)^2 - E(g(X) - Y)^2 \right| \to 0 \qquad \textit{a.s.}$$

for every distribution of (X, Y) *with* $|Y| \le L$ *for some* $L \in \mathbb{R}$ *and*

$$\text{(ii)} \qquad \inf_{g \in B_n F_{N(n)}} \int_{[a,b]^d} |g(x) - f(x)|^2 \, \mu(dx) \to 0$$

for every distribution of (X, Y) *with* $EY^2 < \infty$, *where* μ *denotes the marginal distribution of* X. *Then the sequence of estimators* $\bar{f}_n$, $n = 1, 2, \ldots$, *defined by* (1) *and* (2) *is strongly consistent.*

Proof. The proof is analogous to the proof of Lemma 2 (Section 4) in Kohler [7]. ■

Now we can prove the following theorem on strong consistency of trigonometric and polynomial regression estimators.

THEOREM 4.1. *If* $|Y| \le L < \infty$, X *has absolutely continuous distribution with density* $\varrho \in L^1([0,2\pi]^d)$ (*resp.* $\varrho \in L^1([-1,1]^d)$), *and the sequences of reals* B_n *and integers* $N(n)$, $n = 1, 2, \ldots$, *satisfy*

$$N(n) \to \infty, \qquad B_n \to \infty, \qquad \frac{N(n)B_n^4 \ln(n)}{n} \to 0, \qquad \frac{B_n^4}{n^{1-\delta}} \to 0$$

for some $\delta > 0$ *and* $n \to \infty$, *then the trigonometric* (*resp. polynomial*) *estimator* $\bar{f}_n$ *constructed according to* (1) *and* (2) *is strongly consistent in the sense of the generalization error*, *i.e.*

$$\lim_{n\to\infty} E_X(f(X) - \bar{f}_n(X))^2 = 0 \qquad a.s.$$

Proof. The assumptions on the sequences B_n and $N(n)$, $n = 1, 2, \ldots$, imply by Lemma 2.5 and the Borel–Cantelli lemma that condition (i) of Lemma 4.1 is satisfied (see [7], [8] for details). Moreover, for any probability measure μ the set of continuous functions of compact support is dense in $L^2([0,2\pi]^d,\mu)$ (resp. $L^2([-1,1]^d,\mu)$) [4]. Since such functions can be uniformly approximated by trigonometric (resp. algebraic) polynomials condition (ii) of Lemma 4.1 is also satisfied. Hence, the assertion follows. ■

5. Conclusions. Results concerning convergence rates of the generalization error for orthogonal series estimators were earlier obtained in [6], [14] under the assumption that the density ϱ satisfies $0 < c \le \varrho$. In the case of bounded regression functions, applying the results from empirical process theory and the notion of the VC-dimension of function classes enabled obtaining the convergence rates without this restrictive assumption. However, the rates given in the present work are not optimal. Another approach aiming at relaxing the condition $0 < c \le \varrho$ is presented in [1].

As already remarked in [12], [13] series type regression function estimators constructed using the system of multivariate trigonometric functions from $L^2([0,2\pi]^d)$ can be represented as single-layer neural network estimators with an appropriate activation function called the cosine squasher. Since the estimators considered in the present work can be obtained by proper truncation of least squares trigonometric estimators, they can also be represented as neural network estimators with slightly more complex structure, i.e. as two-layer neural network estimators. Thus, our results contribute to understanding the asymptotic properties of neural network estimators once again.

Other approaches to nonparametric regression function estimation, based on empirical process theory and properties of the VC-dimension, are presented in [3]. They include Regularization Networks and Support Vector

Machines of which Radial Basis Functions are a special case. The regression problem considered is formulated as a variational problem of finding a function f that minimizes the functional

$$\min_{f\in H} Q[f] = \frac{1}{n}\sum_{i=1}^{n} U(Y_i, f(X_i)) + \lambda\|f\|_K^2,$$

where U is a loss function, $\|f\|_K$ is a norm in a Reproducing Kernel Hilbert Space H defined by the positive definite kernel function K and λ is a regularization parameter. Under rather general conditions the solution of the above variational problem is $\widetilde{f}_n(x) = \sum_{i=1}^{n} c_i K(x, X_i)$.

It is worth mentioning that the kernel function in the RKHS space H can be given by the formula $K(x,y) = \sum_{k=1}^{\infty} \lambda_k e_k(x) e_k(y)$, where λ_k is a sequence of positive numbers and $e_k,\ k = 1, 2, \ldots,$ is an orthogonal system of functions. Thus, the estimators investigated in [3] can be constructed using orthogonal functions, although they are not series type estimators.

References

[1] M. Delecroix and C. Protopopescu, *Consistency of a least squares orthonormal series estimator for a regression function*, Discussion Paper No. 7 (2000), National Research Center on Quantification and Simulation of Economic Processes, Interdisciplinary Research Project 373, Humboldt Universität zu Berlin.

[2] R. Dudley, *Central limit theorems for empirical measures*, Ann. Probab. 6 (1978), 899–929.

[3] T. Evgeniou, M. Pontil and T. Poggio, *Regularization networks and support vector machines*, Adv. Comput. Math. 13 (2000), 1–50.

[4] L. Györfi and H. Walk, *On the strong universal consistency of a series type regression estimate*, Math. Methods Statist. 5 (1996), 332–342.

[5] D. Haussler, *Decision theoretic generalization of the PAC model for neural net and other learning applications*, Inform. Comput. 100 (1992), 78–150.

[6] J. Z. Huang, *Projection estimation in multiple regression with application to functional ANOVA models*, Ann. Statist. 26 (1998), 242–272.

[7] M. Kohler, *Universally consistent regression function estimation using hierarchical B-splines*, J. Multivariate Anal. 68 (1999), 138–164.

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

[9] P. Niyogi and F. Girosi, *Generalization bounds for function approximation from scattered noisy data*, Adv. Comput. Math. 10 (1999), 51–80.

[10] D. Pollard, *Convergence of Stochastic Processes*, Springer, New York, 1984.

[11] W. Popiński, *On least squares estimation of Fourier coefficients and of the regression function*, Appl. Math. (Warsaw) 22 (1993), 91–102.

[12] —, *Consistency of trigonometric and polynomial regression estimators*, ibid. 25 (1998), 73–83.

[13] —, *A note on orthogonal series regression function estimators*, ibid. 26 (1999), 281–291.

[14] W. Popiński, *Convergence rates of orthogonal series regression estimators*, ibid. 27 (2000), 445–454.

[15] C. J. Stone, *Optimal global rates of convergence for nonparametric regression*, Ann. Statist. 10 (1982), 1040–1053.

Department of Survey Design
Central Statistical Office
Al. Niepodległości 208
00-925 Warszawa, Poland
E-mail: w.popinski@stat.gov.pl

Received on 5.7.2000;
revised version on 13.2.2001 (1543)