An approximation result for nets in functional estimation

Sebastian Döhler*, Ludger Rüschendorf

Institute for Mathematical Stochastics, University of Freiburg, Eckerstr. 1, D-79104 Freiburg, Germany

Received July 2000; received in revised form September 2000

Abstract

In this paper a quantitative approximation result is obtained for a general class of function nets which is of interest in functional estimation. Specific applications are given to approximation by neural nets, radial basis function nets, and wavelet nets. For the proof we combine the empirical process based results of a paper of Yukich et al. (IEEE Trans. Inform. Theory 41 (4) (1995) 1021) with probabilistic based approximation results of Makovoz (J. Approx. Theory 85 (1996) 98) for the optimal approximation of functions by convex combination of $n$ basis elements. © 2001 Elsevier Science B.V. All rights reserved

Keywords: Neural nets; Radial basis functions; Wavelet nets; Functional estimation

1. Introduction

A basic principle in nonparametric functional estimation based on empirical risk minimization is to decompose the estimation error into two parts, the systematic approximation error induced by the functional class (like splines, polynomial nets, neural nets, etc.) and the stochastic estimation error. Both types of errors have to be estimated and balanced in order to obtain consistency results for the proposed estimators. For general reference on this method we refer to Lugosi and Zeger (1995) and Devroye et al. (1996).

In this paper, we concentrate on a result for the approximation error under general conditions on the net used in functional estimation. Several approximation results for nets are available in the literature. Denseness results for neural nets, and radial basis function nets were obtained by different methods in Cybenko (1989), Hornik (1991), Park and Sandberg (1991, 1993), Petrushev (1999) (see also additional references in Ripley (1996)). This type of result is closely connected with generalizations of Wiener’s closure theorem as noted in Cybenko (1989), and Park and Sandberg (1991, 1993)). For a detailed exposition see Isenbeck and Rüschendorf (1992). An approximation rate of order $O(1/\sqrt{n})$ for neural net approximation was obtained in Barron (1993), Giorosi and Anzellotti (1993), Jones (1992) and others. Barron’s approximation result was extended to sup-norm metrics in Yukich et al. (1995). The method of proof in that paper is based on empirical process theory.

* Corresponding author. Tel.: +49-0761-203-5661.
E-mail address: ruschen@stochastic.uni-freiburg.de (S. Döhler).
It gives a $1/\sqrt{n}$-approximation rate independent of the dimension under the assumption that the underlying function has an integral representation in terms of the neural net kernel. In comparison Barron (1993) used a condition on the Fourier transform related to some smoothness condition.

Makovoz (1996) introduced a probabilistic approximation technique for elements in the closure of a convex symmetric set in $L_q$ by linear combination of $n$ elements. His result leads to an improvement of the $1/\sqrt{n}$ rate of approximation for certain sigmoidal neural nets with respect to $L_q$ norm. We combine in this paper the result of Makovoz (1996) with the development in Yukich et al. (1995) on sup-norm approximation to obtain improved $L_q$-approximation rates for general net types like neural nets, radial basis functions and wavelet nets. In a subsequent paper, we will apply these results to obtain nonparametric estimation rates for net estimators.

For some details on the arguments and the connection to functional estimation we refer to the dissertation of Döhler (2000).

2. Approximation rates by finite nets

Let $\mathcal{F} = \{a_\theta : U \rightarrow \mathbb{R}^p; \; \theta \in \Theta\}$ for $U \subset \mathbb{R}^d$ be a parametric class of functions, with $\Theta \subset \mathbb{R}^k$ a measurable set, which generates a class $\mathcal{F}_0$ of ‘basis’-functions by some transformation $\Psi : \mathbb{R}^p \rightarrow [0, 1]$

$$\mathcal{F}_0 = \{\Psi \circ a_\theta; \; \theta \in \Theta\}.$$  

In this section, we study approximation properties of functions $f$ which allow an integral representation over basis-functions. Define the class of ‘continuous net functions’ by

$$\mathcal{Z}(\Psi, \mathcal{F}) = \mathcal{Z}(\mathcal{F}_0)$$

$$= \left\{ f : U \rightarrow \mathbb{R}; \; \exists v \in \mathcal{M}_k, f(x) = \int_{\mathcal{F}_0} \Psi(a_\theta(x)) \, dv(\theta) \right\},$$

where $\mathcal{M}_k$ is the class of signed measures $v$ on $\mathbb{R}^k$ with finite total variation $|v|$.

Functions of this type were introduced in Yukich et al. (1995). For basis functions of the form $\Psi(\gamma \cdot x + \delta)$ they are related to neural nets, for the form $\varrho(\|\delta(x - \gamma)\|)$ to radial basis nets and for $\Psi(\delta(x - \gamma))$ to wavelet nets. Several denseness results for the class of functions $\mathcal{Z}(\mathcal{F}_0)$ are known and for some cases sufficient smoothness conditions to imply a representation of this form are known (see the papers of Barron (1993) and Yukich et al. (1995)).

The problem considered in this paper is to find ‘good’ $L_q$-approximations for $f \in \mathcal{Z}(\mathcal{F}_0)$ in the class of finite nets with $K$ components:

$$\mathcal{Z}_K(\mathcal{F}_0) = \left\{ f_K; \; f_K = \sum_{i=1}^K \beta_i f_i, f_i \in \mathcal{F}_0, \beta_i \in \mathbb{R} \right\},$$

or in some subclass of $\mathcal{Z}_K(\mathcal{F}_0)$ with additional boundedness restrictions on $\beta_i$. A famous theorem of Kolmogorov (1957) states for any continuous function $f$ of $n$ variables an exact representation of the form

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{2n+1} \Psi \left( \sum_{j=1}^n x_{ij} x_j \right),$$

where $x_{ij}$ are certain Lipschitz functions independent of $f$ and $\Psi = \Psi_f$ is a continuous function of one variable. (In fact this is a variant of Kolmogorov’s original theorem.) So with a small number $K = 2n + 1$ an exact representation holds. A similar representation holds true also for any measurable function $f$ if exact equality is replaced by a.s. equality w.r.t. any measure $\mu$ on $\mathbb{R}^n$ (see Rüschendorf and Thomsen (1997)). For functional estimation one wants to find good approximations of $f$ by function of the form as in (4) with simple functions $x_{ij}$ and $\Psi$; the $x_{ij}$ typically coming from a class of affine functions.
In the following theorem, we impose a growth condition on the basis $\mathcal{F}_0$ in terms of the covering numbers $N(\delta, \mathcal{F}_0, d_{L^2(\mu)})$ w.r.t. some probability measure $\mu$ on $U$ and secondly a condition that the class of functions $\vartheta \mapsto \Psi \circ a_\vartheta(z), z \in U$ is not too ‘big’, in technical terms is a $P$-Donsker class for any probability measure $P$ on $\Theta$ (for the notions of covering numbers and $P$-Donsker class see e.g. van der Vaart and Wellner (1996) and Pollard (1990)). Denote
\[
\vartheta^* := \{ b_\vartheta : \Theta \to \mathbb{R}^p, b_\vartheta(\vartheta) = a_\vartheta(z); z \in U \}, \tag{5}
\]
let $M^1(U), M^1(\Theta), \ldots,$ denote the set of probability measures on $U, \Theta, \ldots$, and let $q^* = \lfloor q \rfloor$ be the smallest even number greater than or equal to $q$ for $q > 0$.

**Theorem 2.1.** Assume that $\mu \in M^1(U)$ and
\[
1.
\]
\[
N(\delta, \mathcal{F}_0, d_{L^2(\mu)}) = O((1/\delta)^{2(D-1)})
\tag{6}
\]
for some $D > 1$ as $\delta \to 0$.

2. $\Psi \circ \vartheta^*$ is a $P$-Donsker class for all $P \in M^1(\Theta)$.

Then for any $f = \int_\Theta \Psi \circ a_\vartheta d\nu(\vartheta) \in \mathcal{F}(\mathcal{F}_0)$ and $q \geq 1$ there exists a finite net $g_K = \sum_{i=1}^K \beta_i f_i \in \mathcal{F}_K(\mathcal{F}_0), K \in \mathbb{N}$, with $\sum_{i=1}^K |\beta_i| \leq 2|v|$ such that
\[
\left\| f - \sum_{i=1}^K \beta_i f_i \right\|_{L^q(\mu)} = O(K^{-1/2(1+(2q^*(D-1)))}). \tag{7}
\]

**Proof.** Define $\Psi_z : \Theta \to [0, 1], \Psi_z(\vartheta) = \Psi(a_\vartheta(z))$ i.e. $\Psi_z = \Psi \circ b_z$ for $z \in U$. By our assumption $\mathcal{F}_0 = \Psi \circ \vartheta^*$ is a $P$-Donsker class for all $P \in M^1(\Theta)$. Therefore, we can follow the empirical process argument in Yukich et al. (1995) for the proof of Theorem 2.1 to obtain the following sup-norm approximation result: For all $n \in \mathbb{N}$ there exist $\beta_1, \ldots, \beta_{2n} \in [-|v|, |v|]$ and $\vartheta_1, \ldots, \vartheta_{2n} \in \Theta$ such that:
\[
\sup_{z \in U} \left| f(z) - \frac{1}{n} \sum_{i=1}^{2n} \beta_i \Psi(a_{\vartheta_i}(z)) \right| \leq C|v| \frac{1}{\sqrt{n}}, \tag{8}
\]
where $C = C(v, d)$ is a constant depending on $v$ and the dimension $d$.

Denote for $M > 0$
\[
V(M) = c \bigg\{ v : U \to \mathbb{R}; v(z) = \sum_i \beta_i \Psi(a_{\vartheta_i}(z)), \sum_i |\beta_i| \leq M, \vartheta_i, \ldots, \vartheta_{2n} \in \Theta \bigg\}, \tag{9}
\]
where the closure is taken in $L^2(\mu)$.

Then from the uniform approximation result in (8) we obtain that $f \in V(2|v|)$. Since $V(1) = c \big\{ v \big\} \text{conv} (\mathcal{F}_0 \cup (-\mathcal{F}_0))$, conv denoting the convex hull, we obtain from the approximation result in Makovoz (1996, Theorem 2) for $1 \leq q < \infty$
\[
\sup_{f \in V(1)} \inf_{g \in V(1)} \left\{ \|f - g\|_{L^q(\mu)}; g = \sum_{i=1}^K \beta_i f_i \in \mathcal{F}_K(\mathcal{F}_0), \sum_{i=1}^K |\beta_i| \leq 1 \right\}
\leq C_q \varepsilon_K(\mathcal{F}_0)^2 q^* \frac{1}{\sqrt{K}}, \tag{10}
\]
where $\varepsilon_K(\mathcal{F}_0) = \inf \{ \varepsilon > 0; N(\varepsilon, \mathcal{F}_0, d_{L^2(\mu)}) \leq K \}$. 


If \( \delta > 0 \) and \( N(\delta, \mathcal{F}_0, d_{L^p}) \leq K \), then \( \delta \geq \varepsilon_K(\mathcal{F}_0) \). Since by assumption (6) \( N(\delta, \mathcal{F}_0, d_{L^p}) \leq C(D)(1/\delta)^{1/(D-1)} \) for some constant \( C(D) \), we conclude that \( \varepsilon_K(\mathcal{F}_0) \leq (C(D)/K)^{1/(2(D-1))} \). This implies by (10)

\[
\sup_{f \in V(1)} \inf_{g \in Z(K(\mathcal{F}_0), \mathcal{F}_0)} \left\{ \| f - g \|_{L^q} \right\} = O(K^{-1/2(1+2/q^*(D-1))})
\]

and the result follows.

\[\square\]

**Remark 2.2.**

(a) The assumption on the integral representation \( f \in Z(\mathcal{F}_0) \) and the Donsker-class condition only serve in the proof of Theorem 2.1 to imply that \( f \in V(M) \) for some \( M > 0 \). Therefore, the approximation result remains valid under the alternative condition:

\[\text{there exists } M > 0 \text{ such that } f \in V(M). \tag{11}\]

(b) For Vapnik Cervonenkis classes \( \mathcal{F} \) with majorant \( F \) holds by Pollard’s inequality (see van der Vaart and Wellner (1996, Theorem 2.6.7)) for \( 0 < \varepsilon < 1, q \geq 1 \)

\[
N(\varepsilon\|F\|_{L^p}, \mathcal{F}, d_{L^p}) \leq Cd(16\varepsilon)^d \left( \frac{1}{\varepsilon} \right)^{q(d-1)}, \tag{12}\]

where \( d = \dim_{VC}(\mathcal{F}) \) is the VC-dimension of \( \mathcal{F} \) and \( C \) is a universal constant. Therefore, condition 1 of Theorem 2.1 holds with \( D = \dim_{VC}(\mathcal{F}_0) \).

### 3. Examples

In this section, we apply the approximation result in Theorem 2.1 to several examples which are relevant in functional estimation and which allow to obtain convergence rates results for empirical risk minimization estimators.

#### 3.1. Neural nets

Let \( \Psi : \mathbb{R} \to [0, 1] \) be of bounded variation, \( U \subset \mathbb{R}^d \) and let \( f : U \to \mathbb{R} \) have a representation as ‘continuous neural net’

\[
f(x) = \int_{\mathbb{R}^d} \int_{\mathbb{R}} \Psi(\gamma \cdot x + \delta) \, d\nu(\gamma, \delta)
\]

for some \( \nu \in \mathcal{M}_{d+1} \) and let \( \mu \in \mathcal{M}^1(U) \).

**Theorem 3.1.** For \( K \in \mathbb{N} \) there exists a finite neural net approximation

\[
f_K(x) = \sum_{i=1}^{K} \beta_i \Psi(\gamma_i \cdot x + \delta_i) \in Z_K(\mathcal{F}_0)
\]

of \( f \) with \( \sum_{i=1}^{K} |\beta_i| \leq 2|\nu| \) such that

\[
\| f - f_K \|_{L^p} = O(K^{-1/2(1+2/q^*(d+2))}). \tag{14}\]
Proof. Note that the neural net is a special case of the net in Section 2 where \( p = 1 \), \( \mathcal{F} = \{ a(\gamma, \delta) : U \to \mathbb{R} ; a(\gamma, \delta)(x) = \gamma \cdot x + \delta, \gamma, \delta \in \mathbb{R}^d, \delta \in \mathbb{R} \} \), \( \Theta = \mathbb{R}^{d+1} \), \( \vartheta = (\gamma, \delta) \). Since \( \mathcal{F} \) is a \( d + 1 \)-dimensional vector space its VC-dimension is by a well-known result of Dudley and Steele (cf. van der Vaart and Wellner (1996, Theorem 2.6.15)) bounded by \( d + 3 \), \( \dim_{\text{VC}}(\mathcal{F}) \leq d + 3 \). Since \( \dim_{\text{VC}}(\Phi \circ \mathcal{F}) = \dim_{\text{VC}}(\mathcal{F}) \) for monotone transformations \( \Phi \) and as \( \Psi \) is of bounded variation, we conclude that \( \dim_{\text{VC}}(\mathcal{F}_0) = \dim_{\text{VC}}(\Psi \circ \mathcal{F}) \leq d + 3 \).

Therefore, by Pollard’s inequality (12) assumption 1 of Theorem 2.1 is fulfilled with \( D = d + 3 \).

Remark 3.2. In special cases (e.g. if \( U = [0, 1]^d, \mu = \lambda^d \)) the convergence rate can be improved to \( \frac{1}{2}(1 + 2/q^*d) \) as is shown in Makovoz (1996, Theorem 3).

For functions defined on a compact set \( U \) the assumption on the integral representation of \( f \) can be replaced by an assumption on the Fourier transform of \( f \) due to Barron (1993). Assume w.l.o.g. that \( U = [0, 1]^d \). Let \( \mathcal{F} : \mathbb{R} \to [0, 1] \) be monotonically non-decreasing, \( \lim_{x \to -\infty} \mathcal{F}(x) = 1 \), \( \lim_{x \to -\infty} \mathcal{F}(x) = 0 \) and let \( f : [0, 1]^d \to \mathbb{R} \) be a bounded measurable function with representation

\[
 f(x) = \int_{\mathbb{R}^d} \exp(\imath \omega \cdot x) \tilde{f}(\omega) \, d\omega, \tag{15}
\]

with Fourier transform \( \tilde{f} \). Finally let \( \mu \in M^1([0, 1]^d) \).

Theorem 3.3. Assume that

\[
 C_f := \int |\omega||\tilde{f}(\omega)| \, d\omega < \infty, \tag{16}
\]

then there exists for \( K \in \mathbb{N} \) some finite neural net approximation \( f_K(x) = \sum_{i=1}^K b_i \mathcal{F}(\gamma_i \cdot x + \delta_i) \in \mathcal{F}_0 \) with \( \sum_{i=1}^K |b_i| \leq C_f + \|f\|_{\infty} \) such that

\[
 \|f - f_K\|_{L^1(\mu)} = O(K^{-1/2(1+2/q^*(d+2))}). \tag{17}
\]

Proof. From Barron (1994) we conclude that \( f \in V(C_f + \|f\|_{\infty}) \). As in the proof of Theorem 3.1 condition 1 of Theorem 2.1 is fulfilled with \( D = d + 3 \). Therefore, we obtain the approximation result from Remark 2.2, a).

Remark 3.4. If \( \mathcal{F}(x) = 0 \) for \( x \leq c_1 \) and \( \mathcal{F}(x) = 1 \) for \( x \geq c_2 \), then from Remark 5 and the proof of Theorem 2.2 in Yukich et al. (1995) we obtain that \( f \in \mathcal{F}(\mathcal{F}_0) \). So in this case the representation condition in Theorem 3.1 is implied by Barron’s condition in Theorem 3.3.

3.2. Radial basis function nets

Let \( g : \mathbb{R}^+ \to [0, 1] \) be monotonically nonincreasing or monotonically nondecreasing, \( U \subseteq \mathbb{R}^d \) and let \( f : U \to \mathbb{R} \) have a continuous radial basis function representation:

\[
 f(x) = \int_{\mathbb{R}^d} \int_{\mathbb{R}} g(\|\delta \cdot x - \gamma\|) \, d\nu(\gamma, \delta) \tag{18}
\]

for some \( \nu \in \mathcal{M}_{d+1} \). As remarked in Yukich et al. (1995) this representation holds true under some smoothness conditions on \( f \).
Theorem 3.5. Let $\mu \in M^1(U)$, $q \geq 1$, then for $K \in \mathbb{N}$ there exists a finite radial basis net approximation $f_K(x) = \sum_{i=1}^{\infty} \beta_i \varphi(||\delta_i(x - \gamma)||) \in \mathcal{F}_K(\mathcal{F}_0)$ of $f$ with $\sum_{i=1}^{\infty} |\beta_i| \leq 2|v|$ such that

$$\|f - f_K\|_{L^q(\mu)} = O(K^{-1/2(1+(2/q^*(d+3)))}).$$

(19)

Proof. Note that with

$$\mathcal{F} = \{a(y,\delta) : U \to \mathbb{R}^+, x \to |\delta||x - \gamma|; \gamma \in \mathbb{R}^d, \delta \in \mathbb{R}, a \in \mathbb{R}, b \in \mathbb{R} \},$$

$\mathcal{F}^2 \{a^2; a \in \mathcal{F}\}$ is contained in the $d + 2$-dimensional vector space $\{x \to a||x||^2 + b \cdot x + c; a, c \in \mathbb{R}, b \in \mathbb{R} \}$. Therefore, dim$_{\mathcal{V}} \mathcal{F}^2 \leq d + 4$ and, as in the proof of Theorem 3.5.1 dim$_{\mathcal{V}} \mathcal{F}_0 = \dim_{\mathcal{V}} \mathcal{F} \circ \mathcal{F} \leq d + 4$, i.e. condition 1 of Theorem 2.1 is fulfilled with $D = d + 4$. Further, with $\mathcal{V} = \{b_\gamma : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^+, (\gamma, \delta) \to |\delta||z - \gamma|; z \in U \}$ we obtain that $\mathcal{V}^2$ is a subset of the finite dimensional vector space $\{(\gamma, \delta) \to y^T \delta^2 - 2\delta \gamma \cdot z; z \in \mathbb{R}^d, y \in \mathbb{R} \}$. Therefore, $\mathcal{V}$ is a VC-class and the approximation result follows as in the proof of Theorem 3.1. \qed

3.3. Wavelet nets

Wavelet nets are alternatives to neural networks in approximating nonlinear functions. Yukich et al. (1995) obtained a uniform $\sqrt{K}$-approximation result for wavelet networks. Again, we obtain improved convergence rates for $L'_q$-approximation.

Let $\Psi : \mathbb{R}^d \to [0,1]$ be Lipschitz-continuous with compact support, let $U \subset \mathbb{R}^d$ and let $f : U \to \mathbb{R}$ have a continuous wavelet representation of the form:

$$f(x) = \int_{\mathbb{R}^d} \int_{\mathbb{R}} \Psi(\delta(x - \gamma)) d\nu(\gamma, \delta)$$

(20)

for some $v \in \mathcal{M}_d$.1.

Theorem 3.6. Let $\mu \in M^1(U)$, $q \geq 1$, then for $K \in \mathbb{N}$ there exists a finite wavelet net approximation $f_K(x) = \sum_{i=1}^{\infty} \beta_i \Psi(\delta_i(x - \gamma_i)) \in \mathcal{F}_K(\mathcal{F}_0)$ of $f$ with $\sum_{i=1}^{\infty} |\beta_i| \leq 2|v|$ and

$$\|f - f_K\|_{L^q(\mu)} = O(K^{-1/2(1+(2/q^*d)))}).$$

(21)

Proof. Let supp($\Psi$) $\subset [-M, M]$, $\mathcal{F} = \{a(y,\delta) : U \to \mathbb{R}^d, x \to \delta(x - \gamma); \gamma \in \mathbb{R}^d, \delta \in \mathbb{R} \},$

$$\mathcal{V} = \{b_\gamma : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^+, (\gamma, \delta) \to \delta(z - \gamma); z \in U \}$$

and

$$\mathcal{V}_i = \{b_\gamma : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}, (\gamma, \delta) \to \delta(z_i - \gamma_i); z \in U \}.$$ Let $T_M : \mathbb{R} \to [-M, M]$ denote the truncation operator

$$T_M(x) = \begin{cases} M & \text{for } x > M, \\ x & \text{for } -M \leq x \leq M, \\ -M & \text{for } x < -M. \end{cases}$$

(22)

Then $\mathcal{V}_i$ is a VC-class. This implies that $T_M \circ \mathcal{V}_i$ is a VC-class and, therefore, also a Donsker-class. As $\Psi$ is Lipschitz continuous and $\Psi \circ (\mathcal{V}_1, \ldots, \mathcal{V}_d) = \Psi \circ (T_M \circ \mathcal{V}_1, \ldots, T_M \circ \mathcal{V}_d)$ we obtain from Theorem 2.10.6 in
van der Vaart and Wellner (1996) that $\Psi \circ \mathcal{F}$ is a Donsker-class and, therefore, condition 2 of Theorem 2.1 is satisfied.

To establish condition 1 of Theorem 2.1 we introduce

$$\mathcal{T}_i = \{ a_{\gamma, \delta} : U \to \mathbb{R}, \ x \to \delta (x_i - \gamma_i); \ \gamma \in \mathbb{R}^d, \delta \in \mathbb{R} \}. $$

$\mathcal{T}_i$ is a subset of the two-dimensional vector space $\{ x \to ax_i + b; a, b \in \mathbb{R} \}$ and, therefore, $\text{dim}_{VC}(T_M \circ \mathcal{T}_i) \leq \text{dim}_{VC}(\mathcal{T}_i) \leq 4$.

In the next step, we prove for $\delta > 0$ the estimate

$$N(\delta, \Psi \circ \mathcal{T}, d_{L^2(\mu)}) \leq N \left( \frac{\delta}{\sqrt{d \text{Lip } \Psi}}, T_m \circ \mathcal{T}_1, d_{L^2(\mu)} \right)^d. \quad (23)$$

For the proof note that

$$N(\delta, \Psi \circ \mathcal{T}, d_{L^2(\mu)}) \leq N(\delta, \Psi \circ (T_M \circ \mathcal{T}_1, \ldots, T_M \circ \mathcal{T}_d), d_{L^2(\mu)}).$$

Let $N(\epsilon) := N(\epsilon, T_M \circ \mathcal{T}_1, d_{L^2(\mu)}) = \cdots = N(\epsilon, T_M \circ \mathcal{T}_d, d_{L^2(\mu)})$, then for $1 \leq i \leq d$ there exist $f_1, \ldots, f_{N(\epsilon)} \in T_M \circ \mathcal{T}_i$ such that $\inf_1 \leq N(\epsilon) \| g - f_i \|_{L^2(\mu)} \leq \epsilon$ for all $g \in T_M \circ \mathcal{T}_i$.

For $\epsilon := \delta/\sqrt{d \text{Lip } \Psi}$ and $i_1, \ldots, i_d \leq N(\epsilon)$ define $g_{i_1, \ldots, i_d} = \Psi \circ (f_{i_1}^1, \ldots, f_{i_d}^d)$. Then for $g = \Psi \circ (f_1^1, \ldots, f_d^d) \in \Psi \circ (T_M \circ \mathcal{T}_1, \ldots, T_M \circ \mathcal{T}_d)$ and $f_{j_i}^i \in T_M \circ \mathcal{T}_i$ with $\| f_i^i - f_{j_i}^i \|_{L^2(\mu)} \leq \epsilon$ we obtain

$$\| g - g_{i_1, \ldots, i_d} \|_{L^2(\mu)}^2 = \int \| \Psi \circ (f_1^1, \ldots, f_d^d) - \Psi \circ (f_{j_1}^1, \ldots, f_{j_d}^d) \|_{\mu}^2 \text{d}\mu$$

$$\leq (\text{Lip } \Psi) \sum_{i=1}^d \| f_i^i - f_{j_i}^i \|_{L^2(\mu)}^2 \leq d(\text{Lip } \Psi) \epsilon^2.$$

This estimate implies (23).

Since $\text{dim}_{VC}(T_M \circ \mathcal{T}_1) \leq 4$ and $T_M \circ \mathcal{T}_1$ is uniformly bounded, we obtain from Pollard’s estimate (12) $N(\delta, T_M \circ \mathcal{T}_1, d_{L^2(\mu)}) = O(1/\delta^6)$ and, therefore, from (23) $N(\delta, \Psi \circ \mathcal{T}, d_{L^2(\mu)}) = O(1/\delta^{6d})$. So condition 1 of Theorem 2.1 is fulfilled with $D = 3d + 1$ and the approximation result follows. \qed

Uncited References

Hornik et al., 1989; Mhaskar, 1996.

References


