An approximation result for nets in functional estimation

Sebastian Döhler and Ludger Rüschendorf*

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

16. June 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, Stinchcombe, and White (1995) with probabilistic based approximation results of Makovoz (1996) for the optimal approximation of functions by convex combination of n basis elements.

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, ...) 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, Györfi, and Lugosi (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 results is closely connected with generalizations of Wiener's closure theorem

* Corresponding author. Tel.: +49-761-2035665
E-mail address: ruschen@stochastik.uni-freiburg.de

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 $O\left(\frac{1}{\sqrt{n}}\right)$ for neural net approximation was obtained in Barron (1993), Girosi and Anzellotti (1993), Jones (1992) and others. Barron's approximation result was extended to sup-norm metrics in Yukich, Stinchcombe, and White (1995). The method of proof in that paper is based on empirical process theory. It gives a $\frac{1}{\sqrt{n}}$-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 $\frac{1}{\sqrt{n}}$ 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, Stinchcombe, and White (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 $\mathcal{T} = \{a_\vartheta : U \to \mathbb{R}^p;\ \vartheta \in \Theta\}$ for $U \subset \mathbb{R}^d$ be a parametric class of functions, with $\Theta \subset \mathbb{R}^k$ a measurable set, which generate a class $\mathcal{F}_0$ of 'basis'-functions by some transformation $\Psi : \mathbb{R}^p \to [0,1]$

$$\mathcal{F}_0 = \{\Psi \circ a_\vartheta;\ \vartheta \in \Theta\}. \tag{1}$$

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

$$\begin{aligned} \mathcal{Z}(\Psi,\mathcal{T}) &= \mathcal{Z}(\mathcal{F}_0) \\ &= \{f : U \to \mathbb{R};\ \exists \nu \in \mathcal{M}_k, f(x) = \int_\Theta \Psi(a_\vartheta(x))d\nu(\vartheta)\} \end{aligned} \tag{2}$$

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

Functions of this type were introduced in Yukich, Stinchcombe, and White (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, Stinchcombe, and White (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

$$\mathcal{Z}_K(\mathcal{F}_0) = \{f_K;\ \ f_K = \sum_{i=1}^{K} \beta_i f_i, f_i \in \mathcal{F}_0, \beta_i \in \mathbb{R}\} \tag{3}$$

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

$$f(x_1,\dots,x_n) = \sum_{i=1}^{2n+1} \Psi\left(\sum_{j=1}^{n} \alpha_{ij}(x_j)\right) \tag{4}$$

where α_{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 Kolmogorv'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 μ 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 α_{ij} and Ψ ; the α_{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 μ 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 Θ (for the notions of covering numbers and P-Donsker class see e.g. van der Vaart and Wellner (1996) and Pollard (1990)). Denote

$$\mathcal{V} := \{b_z : \Theta \to \mathbb{R}^p, b_z(\vartheta) = a_\vartheta(z);\ \ z \in U\}, \tag{5}$$

let $M^1(U), M^1(\Theta), \dots$ denote the set of probability measures on $U, \Theta, \dots$ and let $q^* = \lceil q \rceil$ 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. $$N(\delta, \mathcal{F}_0, d_{L^2(\mu)}) = O\left(\left(\frac{1}{\delta}\right)^{2(D-1)}\right) \tag{6}$$

   for some $D > 1$ *as* $\delta \to 0$.

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

Then for any $f = \int_\Theta \Psi \circ a_\vartheta \, d\nu(\vartheta) \in \mathcal{Z}(\mathcal{F}_0)$ *and* $q \geq 1$ *there exists a finite net* $g_K = \sum_{i=1}^K \beta_i f_i \in \mathcal{Z}_K(\mathcal{F}_0), K \in \mathbb{N}$, *with* $\sum_{i=1}^K |\beta_i| \leq 2|\nu|$ *such that*

$$\left\| f - \sum_{i=1}^K \beta_i f_i \right\|_{L^q(\mu)} = O\left(K^{-\frac{1}{2}\left(1+\frac{2}{q^*(D-1)}\right)} \right). \tag{7}$$

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 $\mathcal{F}_0 = \Psi \circ \mathcal{V}$ is a P-Donsker class for all $P \in M^1(\Theta)$. Therefore, we can follow the empirical process argument in Yukich, Stinchcombe, and White (1995) for the proof of Theorem 2.1 to obtain the following *sup*-norm approximation result: For all $n \in \mathbb{N}$ there exist $\beta_1, \ldots, \beta_{2n} \in [-|\nu|, |\nu|]$ and $\vartheta_1, \ldots, \vartheta_{2n} \in \Theta$ such that:

$$\sup_{z \in U} \left| f(z) - \frac{1}{n} \sum_{i=1}^{2n} \beta_i \Psi(a_{\vartheta_i}(z)) \right| \leq C|\nu| \frac{1}{\sqrt{n}} \tag{8}$$

where $C = C(\nu, d)$ is a constant depending on ν and the dimension d.

Denote for $M > 0$

$$V(M) = c\ell \left\{ v : U \to \mathbb{R};\;\; v(z) = \sum_i \beta_i \Psi(a_{\vartheta_i}(z)), \sum_i |\beta_i| \leq M, \vartheta \in \Theta \right\} \tag{9}$$

where the closure is taken in $L^2(\mu)$.

Then from the uniform approximation result in (8) we obtain that $f \in V(2|\nu|)$. Since $V(1) = c\ell\, \mathrm{conv}(\mathcal{F}_0 \cup (-\mathcal{F}_0))$, conv denoting the convex hull, we obtain from the approximation result in Makovoz (1996, Theorem 2) for $1 \leq q < \infty$

$$\begin{aligned} &\sup_{f \in V(1)} \inf \left\{ \|f - g\|_{L^q(\mu)};\;\; g = \sum_{i=1}^K \beta_i f_i \in \mathcal{Z}_K(\mathcal{F}_0), \sum_{i=1}^K |\beta_i| \leq 1 \right\} \\ &\leq \;\; C_q \varepsilon_K(\mathcal{F}_0)^{\frac{2}{q^*}} \frac{1}{\sqrt{K}} \end{aligned} \tag{10}$$

where $\varepsilon_K(\mathcal{F}_0) = \inf\{\varepsilon > 0;\;\; N(\varepsilon, \mathcal{F}_0, d_{L^2(\mu)}) \leq K\}$.

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

$$\begin{aligned} &\sup_{f \in V(1)} \inf \left\{ \|f - g\|_{L^q(\mu)};\;\; g = \sum_{i=1}^K \beta_i f_i \in \mathcal{Z}_K(\mathcal{F}_0), \sum_{i=1}^K |\beta_i| \leq 1 \right\} \\ &= \;\; O\left(K^{-\frac{1}{2}\left(1+\frac{2}{q^*(D-1)}\right)} \right) \end{aligned}$$

and the result follows. □

Remark 2.2

a) *The assumption on the integral representation* $f \in \mathcal{Z}(\mathcal{F}_0)$ *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:*

$$\textit{There exists } M > 0 \textit{ such that } f \in V(M) \tag{11}$$

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

$$N\left(\varepsilon \|F\|_{L^q(\mu)}, \mathcal{F}, d_{L^q(\mu)}\right) \leq C d (16e)^d \left(\frac{1}{\varepsilon}\right)^{q(d-1)} \tag{12}$$

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'

$$f(x) = \int_{\mathbb{R}^d} \int_{\mathbb{R}} \Psi(\gamma \cdot x + \delta) d\nu(\gamma, \delta) \tag{13}$$

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*

$$f_K(x) = \sum_{i=1}^{K} \beta_i \Psi(\gamma_i \cdot x + \delta_i) \in \mathcal{Z}_K(\mathcal{F}_0)$$

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

$$\|f - f_K\|_{L^q(\mu)} = O\left(K^{-\frac{1}{2}\left(1+\frac{2}{q^*(d+2)}\right)}\right) \tag{14}$$

Proof: Note that the neural net is a special case of the net in section 2 where $p = 1$, $\mathcal{T} = \{a_{(\gamma,\delta)} : U \to \mathbb{R};\ a_{(\gamma,\delta)}(x) = \gamma \cdot x + \delta, \gamma \in \mathbb{R}^d, \delta \in \mathbb{R}\}$, $\Theta = \mathbb{R}^{d+1}, \vartheta = (\gamma, \delta)$. Since $\mathcal{T}$ 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 $d+3$, $\dim_{VC}(\mathcal{T}) \le d+3$. Since $\dim_{VC}(\Phi \circ \mathcal{F}) \le \dim_{VC}(\mathcal{F})$ for monotone transformations Φ and as Ψ is of bounded variation, we conclude that $\dim_{VC} \mathcal{F}_0 = \dim_{VC} \Psi \circ \mathcal{T} \le d+3$. Therefore, by Pollards inequality (12) assumption 1 of Theorem 2.1 is fulfilled with $D = d+3$.

Also $\mathcal{V} = \{b_z : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}, (\gamma, \delta) \to \gamma \cdot z + \delta;\ z \in U\}$ is a subset of a finite dimensional vector space and, therefore, $\mathcal{V}$ is a VC-class. This implies that $\Psi \circ \mathcal{V}$ is a VC-class and, therefore, is a P-Donsker class for any $P \in M^1(\Theta)$ (cf. van der Vaart and Wellner (1996, pg. 141)). From Theorem 2.1 we obtain the approximation result in Theorem 3.1. □

Remark 3.2 *In special cases (e.g. if $U = [0,1]^d, \mu = \lambda^d$) the convergence rate can be improved to $\frac{1}{2}\left(1 + \frac{2}{q^* d}\right)$ 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.g. that $U = [0,1]^d$. Let $\Psi : \mathbb{R} \to [0,1]$ be monotonically non-decreasing, $\lim_{x \to \infty} \Psi(x) = 1, \lim_{x \to -\infty} \Psi(x) = 0$ and let $f : [0,1]^d \to \mathbb{R}$ be a bounded measurable function with representation

$$f(x) = \int_{\mathbb{R}^d} \exp(i\omega \cdot x) \tilde{f}(\omega) d\omega \tag{15}$$

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

Theorem 3.3 *Assume that*

$$C_f := \int |\omega|_1 |\tilde{f}(\omega)| d\omega < \infty, \tag{16}$$

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

$$\|f - f_K\|_{L^q(\mu)} = O\left(K^{-\frac{1}{2}\left(1 + \frac{2}{q^*(d+2)}\right)}\right) \tag{17}$$

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 $\Psi(x) = 0$ for $x \le c_1$ and $\Psi(x) = 1$ for $x \ge c_2$, then from Remark 5 and the proof of Theorem 2.2 in Yukich, Stinchcombe, and White (1995) we obtain that $f \in \mathcal{Z}(\mathcal{F}_0)$. 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 $\varrho : \mathbb{R}^+ \to [0,1]$ be monotonically nonincreasing or monotonically nondecreasing, $U \subset \mathbb{R}^d$ and let $f : U \to \mathbb{R}$ have a continuous radial basis function representation

$$f(x) = \int_{\mathbb{R}^d} \int_{\mathbb{R}} \varrho(\|\delta(x-\gamma)\|) d\nu(\gamma, \delta) \tag{18}$$

for some $\nu \in \mathcal{M}_{d+1}$. As remarked in Yukich, Stinchcombe, and White (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 $f_K(x) = \sum_{i=1}^K \beta_i \varrho(\|\delta_i(x-\gamma_i)\|) \in \mathcal{Z}_K(\mathcal{F}_0)$ of f with $\sum_{i=1}^K |\beta_i| \leq 2|\nu|$ such that*

$$\|f - f_K\|_{L^q(\mu)} = O\left(K^{-\frac{1}{2}\left(1+\frac{2}{q^*(d+3)}\right)}\right). \tag{19}$$

Proof: Note that with

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

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

3.3 Wavelet nets

Wavelet nets are alternatives to neural networks in approximating nonlinear functions. Yukich, Stinchcombe, and White (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

$$f(x) = \int_{\mathbb{R}^d} \int_{\mathbb{R}} \Psi(\delta(x-\gamma)) d\nu(\gamma, \delta) \tag{20}$$

for some $\nu \in \mathcal{M}_{d+1}$.

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 $f_K(x) = \sum_{i=1}^K \beta_i \Psi(\delta_i(x-\gamma_i)) \in \mathcal{Z}_K(\mathcal{F}_0)$ of f with $\sum_{i=1}^K |\beta_i| \leq 2|\nu|$ and*

$$\|f - f_K\|_{L^q(\mu)} = O\left(K^{-\frac{1}{2}\left(1+\frac{2}{q^*3d}\right)}\right) \tag{21}$$

Proof: Let $\mathrm{supp}(\Psi) \subset [-M, M]^d$,

$$\begin{aligned}\mathcal{T} &= \{a_{(\gamma,\delta} : U \to \mathbb{R}^d, x \to \delta(x-\gamma);\;\; \gamma \in \mathbb{R}^d, \delta \in \mathbb{R}\} \\ \mathcal{V} &= \{b_z : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d, (\gamma,\delta) \to \delta(z-\gamma);\;\; z \in U\}\end{aligned}$$

and

$$\mathcal{V}_i = \{b_z : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}, (\gamma,\delta) \to \delta(z_i-\gamma_i);\;\; z \in U\}.$$

Let $T_M : \mathbb{R} \to [-M, M]$ denote the truncation operator

$$T_M(x) = \begin{cases} M & \text{for } x > M \\ x & \text{for } -M \leq x \leq \quad M \\ -M & \text{for } x < -M. \end{cases} \tag{22}$$

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 Ψ 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 $\Psi \circ \mathcal{V}$ 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 2-dimensional vector space $\{x \to ax_i + b; a, b \in \mathbb{R}\}$ and, therefore, $\dim_{VC}(T_M \circ \mathcal{T}_i) \leq \dim_{VC}(\mathcal{T}_i) \leq 4$.

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

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

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

For $\varepsilon := \dfrac{\delta}{\sqrt{d}\,\mathrm{Lip}\,\Psi}$ and $i_1, \dots, i_d \le N(\varepsilon)$ define $g_{i_1,\dots,i_d} = \Psi \circ \left(f^1_{i_1}, \dots, f^d_{i_d}\right)$. Then for $g = \Psi \circ (f^1, \dots, f^d) \in \Psi \circ (T_M \circ \mathcal{T}_1, \dots, T_m \circ \mathcal{T}_d)$ and $f^i_{j_i} \in T_M \circ \mathcal{T}_i$ with $\|f^i - f^i_{j_i}\|_{L^2(\mu)} \le \varepsilon$ we obtain

$$\begin{aligned}
\left\|g - g_{i_1,\dots,i_d}\right\|^2_{L^2(\mu)} &= \int |\Psi \circ (f^1, \dots, f^d) - \Psi \circ (f^1_{j_1}, \dots, f^d_{j_d})|^2 d\mu \\
&\le (\mathrm{Lip}\,\Psi) \sum_{i=1}^d \|f^i - f^i_{j_i}\|^2_{L^2(\mu)} \le d(\mathrm{Lip}\,\Psi)\varepsilon^2.
\end{aligned}$$

This estimate implies (23).

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

References

Barron, A. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory 39*(3), 930–945.

Barron, A. (1994). Approximation and estimation bound for artificial neural networks. *Machine Learning 14*, 115–133.

Cybenko, G. (1989). Approximation by superposition of a sigmoidal function. *Math. Control Systems Signals 2*(4), 303–314.

Devroye, L., L. Györfi, and G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.

Döhler, S. (2000). *Empirische Risiko-Minimierung bei zensierten Daten*. Ph. D. thesis, Universität Freiburg.

Girosi, F. and G. Anzellotti (1993). Rates of convergence for radial basis functions and neural networks. In R. Mammone (Ed.), *Artificial Neural Networks for Speech and Vision*, pp. 97–113. Chapman & Hall.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks 4*, 251–257.

Hornik, K., M. Stinchcombe, and H. White (1989). Multilayer feedforward networks are universal approximators. *Neural Networks 2*, 359–366.

Isenbeck, M. and L. Rüschendorf (1992). Completeness in location families. *Prob. and Math. Stat. 13*, 321–343.

Jones, L. (1992). A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann. Stat. 20*, 608–613.

Kolmogorov, A. (1957). On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *Dokl. Akad. Nauk SSSR 114*, 953–953.

Lugosi, G. and K. Zeger (1995). Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory 41*(3), 677–687.

Makovoz, Y. (1996). Random approximants and neural networks. *Journal of Approximation Theory 85*, 98–109.

Mhaskar, H. (1996). Neural networks for optimal approximation of smooth and analytical functions. *Neural Computation 8*, 164–177.

Park, J. and I. Sandberg (1991). Universal approximation using radial-basis function networks. *Neural Computation 3*, 246–257.

Park, J. and I. Sandberg (1993). Approximation and radial-basis-function networks. *Neural Computation 5*, 305–316.

Petrushev, P. (1999). Approximation by ridge functions and neural networks. *SIAM J. Math. Anal. 30*(1), 155–189.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*. Regional Converence Series in Probability and Statistics. Institute of Mathematical Statistics, Hayward, ASA.

Ripley, B. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Rüschendorf, L. and W. Thomsen (1997). Closedness of sum spaces and the generalized Schrödinger problem. *Theory Prob. Appl. 42*, 576–590.

van der Vaart, A. and J. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer.

Yukich, J., M. Stinchcombe, and H. White (1995). Sup-norm approximation bounds for networks through probabilistic methods. *IEEE Transactions on Information Theory 41*(4), 1021–1027.