Journal of Approximation Theory **112**, 235–250 (2001)
doi:10.1006/jath.2001.3613, available online at http://www.idealibrary.com on IDEAL®

Error Bounds for Approximation with Neural Networks

Martin Burger[1] and Andreas Neubauer

Institute for Industrial Mathematics, Johannes-Kepler University, A-4040 Linz, Austria
E-mail: burger@indmath.uni-linz.ac.at, neubauer@indmath.uni-linz.ac.at

Communicated by Robert Schaback

Received May 14, 2000; accepted in revised form June 14, 2001

In this paper we prove convergence rates for the problem of approximating functions f by neural networks and similar constructions. We show that the rates are the better the smoother the activation functions are, provided that f satisfies an integral representation. We give error bounds not only in Hilbert spaces but also in general Sobolev spaces $W^{m,r}(\Omega)$. Finally, we apply our results to a class of perceptrons and present a sufficient smoothness condition on f guaranteeing the integral representation. © 2001 Academic Press

Key Words: neural networks; error bounds; nonlinear function approximation.

1. INTRODUCTION

The aim of this paper is to find error bounds for the approximation of functions by feed-forward networks with a single hidden layer and a linear output layer, which can be written as

$$f_n(x) = \sum_{j=1}^{n} c_j \phi(x, t_j), \tag{1}$$

where $c_j \in \mathbb{R}$ and $t_j \in P \subset \mathbb{R}^p$ are parameters to be determined.

An important special case of (1) are so-called Ridge-constructions, i.e.,

$$f_n(x) = \sum_{j=1}^{n} c_j \sigma(a_j^T x + b_j). \tag{2}$$

The interest in such networks grew, since Hornik *et al.* [6] showed that functions of the form (2) are dense in $C(\Omega)$, if σ is a function of sigmoidal form. An other special case are radial basis function networks, where $\phi(x, t) = \psi(\|x - t\|)$ (cf. [11]).

[1] Supported by the Austrian Science Foundation under Grant SFB F013/1308.

0021-9045/01 $35.00

We consider the problem of approximating a function $f \in W^{m,r}(\Omega)$, where $W^{m,r}(\Omega)$ denote the usual Sobolev spaces and Ω is a (not necessarily bounded) domain in $\mathbb{R}^d$. This problem can be written in the abstract form

$$\inf_{g \in X_n} \|f-g\|_X, \tag{3}$$

where $X = W^{m,r}(\Omega)$ and X_n denotes the set of all functions of form (1), i.e.,

$$X_n = \left\{ g = \sum_{j=1}^{n} c_j \phi(x, t_j) : t_j \in P \subset \mathbb{R}^p, c_j \in \mathbb{R} \right\}. \tag{4}$$

ϕ is assumed smooth enough so that $X_n \subset X$; P is a (usually bounded) domain.

Usually, the convergence of solutions of (3) if they exist (note that X_n is not a finite-dimensional subspace of X) is arbitrarily slow, since the approximation problem is asymptotically ill-posed, i.e., arbitrarily small errors in the observation can lead to arbitrarily large errors in the approximation as $n \to \infty$ (cf., e.g., [2, 3]). It was shown in [3] that the set of functions to which networks of the form (1) converge is just the closure of the range of the integral operator

$$K: L^2(P) \to X, \qquad h \mapsto \int_P h(t)\, \phi(\cdot, t)\, dt.$$

Rates are usually only obtained under additional conditions on f (cf., e.g., [5]). A natural condition seems to be that f is in the range of the above operator, i.e.,

$$f(x) = \int_P h(t)\, \phi(x, t)\, dt, \tag{5}$$

where h is allowed to be in $L^1(P)$ if ϕ is smooth enough. It was shown in [9] that under this condition the rate

$$\inf_{g \in X_n} \|f-g\|_{L^2(\Omega)} = \mathcal{O}(n^{-\frac{1}{2}}) \tag{6}$$

is obtained if ϕ is a continuous function (see also [7, 8]). We improve this result under additional smoothness assumptions on the basis function ϕ in the next section with estimates also in $H^m(\Omega) = W^{m,2}(\Omega)$. Moreover, we will give error bounds in $W^{m,r}(\Omega)$ that depend on the dimension p (cf. (4)), where the analysis is based on finite-element theory. In Section 3, we apply the results to perceptrons and give sufficient conditions on f for condition (5) to hold. Similar results on the unit circle have been obtained in [4, 10].

2. ERROR BOUNDS

An inspection of the proof of (6) in [9] shows that the result can be improved if the activation function ϕ is Hölder continuous. Moreover, rates can be obtained in $H^m(\Omega)$:

THEOREM 2.1. *Let X_n be defined as in* (4) *with $P \subset \mathbb{R}^p$ compact and ϕ such that*

$$\|\phi(\cdot, t) - \phi(\cdot, s)\|_{H^m(\Omega)} \leqslant c\, \|t - s\|^\rho, \qquad \rho \in (0, 1], c > 0, m \in \mathbb{N}_0. \tag{7}$$

Moreover, let $f \in H^m(\Omega)$ satisfy (5) *with $h \in L^\infty(P)$. Then we obtain the rate*

$$\inf_{g \in X_n} \|f - g\|_{H^m(\Omega)} = \mathcal{O}(n^{-\frac{1}{2} - \frac{\rho}{p}}).$$

Proof. Let $\bar{P} = \{t \in P : h(t \geqslant 0\}$ (note that $\bar{P}$ is unique up to a set of measure zero) and $\bar{n} := [\frac{n}{2}]$. Since P is bounded, it is possible to find bounded measurable sets P_j such that

$$\begin{gathered} \bar{P} = \bigcup_{j=1}^{\bar{n}} P_j, \qquad P \backslash \bar{P} = \bigcup_{j=\bar{n}+1}^{n} P_j, \qquad P_i \cap P_j = \{\}, i \neq j, \\ \operatorname{diam}(P_j) = \mathcal{O}(n^{-\frac{1}{p}}), \qquad |P_j| = \mathcal{O}\left(\frac{1}{n}\right). \end{gathered} \tag{8}$$

We now define coefficients

$$c_j := \int_{P_j} h(t)\, dt$$

and probability measures

$$u_j(t) := \begin{cases} \dfrac{1}{c_j} h(t), & t \in P_j, \\ 0, & \text{otherwise}, \end{cases} \qquad \text{for } c_j \neq 0 \text{ and } \mu_j \text{ is arbitrary for } c_j = 0.$$

As a direct consequence of our construction we have that

$$h = \sum_{j=1}^{n} c_j \mu_j.$$

Furthermore, we consider the variables $t_j \in P$ as random variables distributed with probability distribution μ_j. The expected value of $z(t_1, ..., t_n)$ is defined as

$$E[z] := \int_P \cdots \int_P z(t_1, ..., t_n)\, \mu_1(t_1) \cdots \mu_n(t_n)\, dt_1 \cdots dt_n.$$

With c_j and μ_j as above and f as in (5) we obtain using Fubini's theorem that

$$\begin{aligned}
E\left[\left\| f - \sum_{j=1}^n c_j \phi(\cdot, t_j)\right\|^2_{H^m(\Omega)}\right]
&= \|f\|^2_{H^m(\Omega)} - 2 \sum_{j=1}^n c_j \left\langle f, \int_P \mu_j(t_j)\, \phi(\cdot, t_j)\, dt_j \right\rangle_{H^m(\Omega)} \\
&\quad + \sum_{i \neq j = 1}^n c_i c_j \left\langle \int_P \mu_i(t_i)\, \phi(\cdot, t_i)\, dt_i, \int_P \mu_j(t_j)\, \phi(\cdot, t_j)\, dt_j \right\rangle_{H^m(\Omega)} \\
&\quad + \sum_{j=1}^n c_j^2 \int_P \mu_j(t_j)\, \|\phi(\cdot, t_j)\|^2_{H^m(\Omega)}\, dt_j \\
&= \left\| \int_P \left[h(t) - \sum_{j=1}^n c_j \mu_j(t) \right] \phi(\cdot, t)\, dt \right\|^2_{H^m(\Omega)} \\
&\quad + \sum_{j=1}^n c_j^2 \left[\int_P \mu_j(t)\, \|\phi(\cdot, t)\|^2_{H^m(\Omega)}\, dt - \left\| \int_P \mu_j(t)\, \phi(\cdot, t)\, dt \right\|^2_{H^m(\Omega)} \right].
\end{aligned}$$

Since the first term on the right hand side vanishes, we may conclude that

$$\begin{aligned}
E\left[\left\| f - \sum_{j=1}^n c_j \phi(\cdot, t_j)\right\|^2_{H^m(\Omega)}\right]
&= \sum_{j=1}^n c_j^2 \sum_{|\alpha| \leqslant m} \int_\Omega \left[\int_P \mu_j(t) \left(\frac{\partial^{|\alpha|}}{\partial x^\alpha} \phi(x, t) \right)^2 dt \right. \\
&\qquad \left. - \left(\int_P \mu_j(t) \frac{\partial^{|\alpha|}}{\partial x^\alpha} \phi(x, t)\, dt \right)^2 \right] dx \\
&= \sum_{j=1}^n c_j^2 \sum_{|\alpha| \leqslant m} \int_\Omega \left[\int_{P_j} \mu_j(t) \left(\int_{P_j} \mu_j(t) \left(\frac{\partial^{|\alpha|}}{\partial x^\alpha} \phi(x, t) \right.\right.\right. \\
&\qquad \left.\left.\left. - \frac{\partial^{|\alpha|}}{\partial x^\alpha} \phi(x, s) \right) ds \right)^2 dt \right] dx.
\end{aligned}$$

Noting that $h \in L^\infty(P)$ and (8) imply that $c_j = \mathcal{O}(\frac{1}{n})$, we now obtain together with (7), (8), and the Cauchy–Schwarz inequality that

$$E\left[\left\|f - \sum_{j=1}^{n} c_j \phi(\cdot, t_j)\right\|^2_{H^m(\Omega)}\right]$$
$$\leqslant \sum_{j=1}^{n} c_j^2 \sum_{|\alpha| \leqslant m} \int_\Omega \left[\int_{P_j} \mu_j(t) \int_{P_j} \mu_j(s) \left(\frac{\partial^{|\alpha|}}{\partial x^\alpha} \phi(x, t) - \frac{\partial^{|\alpha|}}{\partial x^\alpha} \phi(x, s)\right)^2 ds\, dt\right] dx$$
$$= \sum_{j=1}^{n} c_j^2 \int_{P_j} \mu_j(t) \int_{P_j} \mu_j(s) \, \|\phi(\cdot, t) - \phi(\cdot, s)\|^2_{H^m(\Omega)} \, ds\, dt$$
$$= \mathcal{O}(n \cdot n^{-2} \cdot n^{-\frac{2\rho}{p}}) = \mathcal{O}(n^{-1-\frac{2\rho}{p}}).$$

Therefore, there exists a set of elements $\bar{t}_j \in P$ such that

$$\inf_{g \in X_n} \|f - g\|_{H^m(\Omega)} \leqslant \left\|f - \sum_{j=1}^{n} c_j \phi(\cdot, \bar{t}_j)\right\|_{H^m(\Omega)}$$
$$\leqslant \left(E\left[\left\|f - \sum_{j=1}^{n} c_j \phi(\cdot, t_j)\right\|^2_{H^m(\Omega)}\right]\right)^{1/2}$$
$$= \mathcal{O}(n^{-\frac{1}{2} - \frac{\rho}{p}}),$$

where c_j is as above. ∎

We think that the proposition above is also true if $h \in L^2(P)$. However, the choice of the subsets P_j in (8) has to be more tricky, since $c_j = \mathcal{O}(\frac{1}{n})$ will no longer hold, in general.

We will now turn to other estimates in spaces $W^{m,r}(\Omega)$. The error bounds will depend on the dimension p of $P \subset \mathbb{R}^p$. The proofs are based on the following results from finite-element theory (see [12]):

Let

$$P := \mathop{\times}_{i=1}^{p} [\underline{p}_i, \bar{p}_i] \qquad \text{and}$$

$$P_{l_1 \cdots l_p} := \mathop{\times}_{i=1}^{p} \left[\underline{p}_i + \frac{\bar{p}_i - \underline{p}_i}{\tau} l_i, \underline{p}_i + \frac{\bar{p}_i - \underline{p}_i}{\tau} (l_i + 1)\right], \qquad \tau \in \mathbb{N}.$$

Then, obviously

$$P = \bigcup_{\substack{l_i = 0, \ldots, \tau - 1 \\ i = 1, \ldots, p}} P_{l_1 \cdots l_p}.$$

Moreover, we define for some $k \in \mathbb{N}$

$$t_{j_1 \cdots j_p} := (t_{j_1 \cdots j_p;\,1}, \ldots, j_{j_1 \cdots j_p;\,p}) \in \mathbb{R}^p, \qquad t_{j_1 \cdots j_p;\,i} := \underline{p}_i + \frac{\bar{p}_i - \underline{p}_i}{k\tau}\, j_i, \qquad j_i = 0, \ldots, k\tau. \tag{9}$$

Then for all $kl_i \leqslant v_i \leqslant k(l_i + 1)$ there exists a unique polynomial function

$$q_{v_1 \cdots v_p} \in Q_{k,\,l_1 \cdots l_p} := \Big\{ q(t) = \sum c_{j_1 \cdots j_p} t_1^{j_1} \cdots t_p^{j_p} : 0 \leqslant j_i \leqslant k, \; 1 \leqslant i \leqslant p,\, t = (t_1, \ldots, t_p) \in P_{l_1 \cdots l_p} \Big\} \tag{10}$$

satisfying

$$q_{v_1 \cdots v_p}(t_{j_1 \cdots j_p}) = \prod_{i=1}^{p} \delta_{v_i j_i}, \; kl_i \leqslant v_i,\, j_i \leqslant k(l_i + 1). \tag{11}$$

The function u_I, defined by

$$u_I|_{P_{l_1 \cdots l_p}} := \sum_{kl_i \leqslant j_k \leqslant k(l_i + 1)} u(t_{j_1 \cdots j_p})\, q_{j_1 \cdots j_p}, \tag{12}$$

interpolates $u \in C(P)$ at the knots $t_{j_1 \cdots j_p}$, $0 \leqslant j_i \leqslant k\tau$, $1 \leqslant i \leqslant p$. Note that $u_I \in C(P) \cap H^1(P)$.

PROPOSITION 2.1. *Let $P \subset \mathbb{R}^p$ be rectangular. If $u \in H^k(P)$ with $k > \frac{p}{2}$, then there is a constant $n_\kappa > 0$ such that for all multiindices β with $|\beta| = \kappa < k$ and for all $l_i \in \{0, \ldots, \tau - 1\}$, $i = 1, \ldots, p$, it holds that*

$$\|D^\beta (u - u_I)\|_{L^2(P_{l_1 \cdots l_p})} \leqslant \eta_\kappa \tau^{-(k-\kappa)}\, |u|_{H^k(P_{l_1 \cdots l_p})}. \tag{13}$$

If $u \in C^k(P)$, then there is a constant $\bar{\eta}_\kappa > 0$ such that for all multiindices β with $|\beta| = \kappa < k$ and for all $l_i \in \{0, \ldots, \tau - 1\}$, $i = 1, \ldots, p$, it holds that,

$$\|D^\beta (u - u_I)\|_{L^\infty(P_{l_1 \cdots l_p})} \leqslant \bar{\eta}_\kappa \tau^{-(k-\kappa)} \max_{|\gamma| = k} \|D^\gamma u\|_{L^\infty(P_{l_1 \cdots l_p})}. \tag{14}$$

Proof. The proof follows with Theorem 3.1 and Theorem 3.3 in [12]. ∎

For our main result we need the following types of smoothness of ϕ: $\phi \in W^{m,r}(\Omega, Y)$ with $Y = H^k(P)$ or $Y = C^k(P)$ and norms

$$\|\phi\|_{W^{m,r}(\Omega, Y)} := \begin{cases} \left(\sum_{|\alpha| \leqslant m} \int_\Omega \left\| \frac{\partial^{|\alpha|}}{\partial x^\alpha} \phi(x, \cdot) \right\|_Y^r dx \right)^{\frac{1}{r}}, & \text{if } 1 \leqslant r < \infty, \\ \max_{|\alpha| \leqslant m} \operatorname{ess\,sup}_{x \in \Omega} \left\| \frac{\partial^{|\alpha|}}{\partial x^\alpha} \phi(x, \cdot) \right\|_Y, & \text{if } r = \infty. \end{cases}$$

THEOREM 2.2. *Let X_n be defined as in (4) with $P \subset \mathbb{R}^p$ bounded and rectangular and let $\phi \in W^{m,r}(\Omega, Y)$ with $Y = H^k(P)$, $k > \frac{p}{2}$, or $Y = C^k(P)$. Moreover, let $f \in W^{m,r}(\Omega)$ satisfy (5) with $h \in L^2(P)$ if $Y = H^k(P)$ and $h \in L^1(P)$ if $Y = C^k(P)$. Then we obtain the rate*

$$\inf_{g \in X_n} \|f - g\|_{W^{m,r}(\Omega)} = \mathcal{O}(n^{-\frac{k}{p}}).$$

Proof. If we choose c_j as

$$c_j := \int_P h(t)\, \gamma_j(t)\, dt, \qquad \gamma_j \in L^\infty(P),$$

with h as in (5), then we obtain that

$$\left\| f - \sum_{j=1}^n c_j \phi(\cdot, t_j) \right\|_{W^{m,r}(\Omega)} = \left\| \int_P h(t) \left(\phi(\cdot, t) - \sum_{j=1}^n \gamma_i(t)\, \phi(\cdot, t_j) \right) dt \right\|_{W^{m,r}(\Omega)}.$$

Let us define $\tau := ([n^{1/p}] - 1)/k$ and $\bar{n} := (k\tau + 1)^p \leqslant n$. Then we choose t_j and γ_j as follows: For $j = \bar{n} + 1, ..., n$ let t_j be arbitrary and $\gamma_j \equiv 0$. For $j = 1, ..., \bar{n}$ let t_j and γ_j be the appropriate knots and basis functions such that the sum above equals the interpolating function $\phi_I(\cdot, t)$ (see (9)–(12)), i.e.,

$$\left\| f - \sum_{j=1}^n c_j \phi(\cdot, t_j) \right\|_{W^{m,r}(\Omega)} = \left\| \int_P h(t)(\phi(\cdot, t) - \phi_I(\cdot, t))\, dt \right\|_{W^{m,r}(\Omega)}.$$

Note that this interpolating property also holds for all derivatives of ϕ with respect to x, since the interpolation is done with respect to t only and holds independently of x. Applying (13) ($\beta = 0$) for $Y = H^k(P)$ and (14) ($\beta = 0$) for $Y = C^k(P)$ we obtain the estimates

$$\left\| f - \sum_{j=1}^{n} c_j \phi(\cdot, t_j) \right\|_{W^{m,r}(\Omega)} \leqslant \eta_0 \tau^{-k} \|h\|_{L^2(P)} \|\phi\|_{W^{m,r}(\Omega, H^k(P))} \tag{15}$$

and

$$\left\| f - \sum_{j=1}^{n} c_j \phi(\cdot, t_j) \right\|_{W^{m,r}(\Omega)} \leqslant \bar{\eta}_0 \tau^{-k} \|h\|_{L^2(P)} \|\phi\|_{W^{m,r}(\Omega, C^k(P))} \tag{16}$$

respectively. Now the assertion follows together with the fact that $\tau \sim n^{\frac{1}{p}}$. ∎

Remark 2.1. The idea of choosing c_j, t_j and γ_j as in the prove above was found in a paper by Wahba [13] for one-dimensional P. This idea was extended to higher dimensions, i.e., $P \subset \mathbb{R}^p$.

The following extensions of Theorem 2.2 are obvious from the proof:

- If P is not rectangular but $\operatorname{supp}(h) \subset \bar{P} \subset P$ with $\bar{P}$ rectangular, then the results are still valid.
- If $Y = C^k(P)$, the condition (5) for f with $h \in L^1(P)$ may be replaced by: f is such that there exists a uniformly bounded sequence h_l in $L^1(P)$ with

$$\left\| f - \int_P h_l(t)\, \phi(\cdot, t)\, dt \right\|_{W^{m,r}(\Omega)} \to 0 \qquad \text{as} \quad l \to \infty.$$

- Condition (5) may be generalized to

$$f(x) = \sum_{|\alpha| \leqslant \kappa} \int_P h_\beta(t) \frac{\partial^{|\beta|}}{\partial t^\beta} \phi(x, t)\, dt, \qquad \kappa < k. \tag{17}$$

If the functions γ_j are chosen such that for each β they coincide with the appropriate derivative of the basis functions $q_{j_1 \cdots j_p}$ in $P_{l_1 \cdots l_p}$, we obtain together with Proposition 2.1 the rates

$$\inf_{g \in X_n} \|f - g\|_{W^{m,r}(\Omega)} = \mathcal{O}(n^{-\frac{(k-\kappa)}{p}}).$$

Finally, we want to mention that the rates above and in Theorem 2.2 decrease with increasing dimension p. There is no dimensionless term like $n^{-\frac{1}{2}}$ in (6) or Theorem 2.1. Since the estimates in the proof of Theorem 2.2 are based on a fixed choice of knots t_j this dependence on p is to be expected. We were not able to improve the rates for a possible optimal choice of knots. However, since Proposition 2.1 is valid also for many other non-uniform choices of knots t_j, the rates in Theorem 2.2 are valid for many choices t_j (also non-optimal ones) if at least c_j is chosen optimally.

3. APPLICATIONS TO PERCEPTRONS

We now apply the results of the previous section to perceptrons with a single hidden layer, namely Ridge- constructions (cf. (2)) where σ is a function of sigmoidal form, i.e.,

$$X_n = \left\{ g = \sum_{j=1}^{n} c_j \sigma(a_j^T x + b_j) : a_j \in A \subset \mathbb{R}^d,\, b_j \in B \subset \mathbb{R} \right\}$$

and σ is piecewise continuous, monotonically increasing, and such that

$$\lim_{t \to -\infty} \sigma(t) = 0 \qquad \text{and} \qquad \lim_{t \to +\infty} \sigma(t) = 1.$$

If σ is such that

$$\sigma(t) := \begin{cases} 1, & t > 1, \\ p(t), & -1 \leqslant t \leqslant 1, \\ 0, & t < -1, \end{cases} \tag{18}$$

with p the unique polynomial of degree $2k+1$ satisfying

$$p(-1) = 0,\ p(1) = 1, \text{ and } p^{(l)}(-1) = 0 = p^{(l)}(1), \qquad 1 \leqslant l \leqslant k, \tag{19}$$

then $\sigma \in C^{k,1}$ and $\sigma \in W^{k+1,\sigma}$ (see Fig. 1).

EXAMPLE 3.1. Let us consider the special case of $k = 0$, i.e.,

$$\sigma(t) := \begin{cases} 1, & t > 1, \\ \dfrac{t+1}{2}, & -1 \leqslant t \leqslant 1, \\ 0, & t < -1, \end{cases} \tag{20}$$

and let $A := \bigtimes_{i=1}^{d} [-\bar{a}_i, \bar{a}_i]$ and $B := [-\bar{b}, \bar{b}]$ with $\bar{a}_i > 0$ and $\bar{b} > 0$ such that

$$\forall a \in A \ \ \forall x \in \Omega : |a^T x| \leqslant \bar{b} - 1.$$

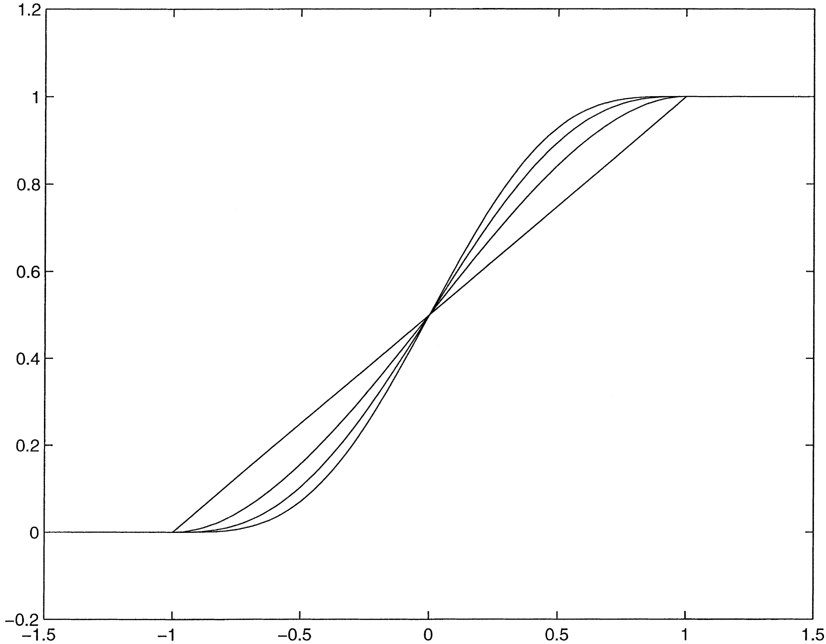

FIG. 1. Function σ from (18) and (19) for $k = 0, 1, 2, 3$.

Since $\phi(x, a, b) := \sigma(a^T x + b)$ satisfies (7) with $m = 0$ and $\rho = 1$, Theorem 2.1 implies that

$$\inf_{g \in X_n} \|f - g\|_{L^2(\Omega)} = \mathcal{O}(n^{-\frac{1}{2} - \frac{1}{d+1}})$$

if

$$\begin{aligned} rcl f(x) &= \int_A \int_{-\bar{b}}^{\bar{b}} h(a, b)\, \sigma(a^T x + b)\, db\, da \\ &= \int_A \left[\int_{-1-a^T x}^{1-a^T x} h(a, b) \frac{1 + a^T x + b}{2}\, db + \int_{1-a^T x}^{\bar{b}} h(a, b)\, db \right] da \qquad (21) \end{aligned}$$

for some $h \in L^\infty(A \times B)$.

Example 3.2. We consider now the general case, where σ is defined by (18), (19), and where A and B are as in Example 3.1.

Since $\phi(x, a, b) := \sigma(a^T x + b)$ satisfies that $\phi \in W^{m,\infty}(\Omega, C^{k-m}(A \times B))$ $(m \leqslant k)$ and $\phi \in W^{m,\infty}(\Omega, H^{k+1-m}(A \times B))$ $(m \leqslant k+1)$, we may apply Theorem 2.2 to obtain

$$\inf_{g \in X_n} \|f - g\|_{W^{m,r}(\Omega)} = \mathcal{O}(n^{-\frac{k-m}{d+1}})$$

if $f \in W^{m,r}(\Omega)$ satisfies

$$f(x) = \int_A \left[\int_{-1-a^T x}^{1-a^T x} h(a, b)\, p(a^T x + b)\, db + \int_{1-a^T x}^{\bar{b}} h(a, b)\, db \right] da \tag{22}$$

for some $h \in L^1(A \times B)$ and

$$\inf_{g \in X_n} \|f - g\|_{W^{m,r}(\Omega)} = \mathcal{O}(n^{-\frac{k+1-m}{d+1}})$$

if $f \in W^{m,r}(\Omega)$ satisfies (22) for some $h \in L^2(A \times B)$ and $k+1-m > \frac{d+1}{2}$. Note that for $m = 0$ and $k > \frac{d+1}{2}$ the rate above is better than the one in Example 3.1.

From both examples, we can see that the conditions (21) and (22) can be only satisfied if f is several times differentiable. We will now give a sufficient condition on f that guarantees (21):

Let $\varepsilon_0 := 0$ and $\varepsilon_n := \frac{\pi}{2}(4n^j - 3)$, $n \in \mathbb{N}$, for some $j \in \mathbb{N}$ to be specified later, and let $\rho_n := \varepsilon_n / \varepsilon_{n+1}$. We define the function h as

$$h(a, b) = \sum_{n=1}^{\infty} (\kappa_n(a) \cos(b\varepsilon_n) + \lambda_n(a) \sin(b\varepsilon_n)), \tag{23}$$

where

$$\begin{aligned}
\kappa_n(a) &:= \begin{cases} -(2\pi)^{-\frac{d}{2}} \varepsilon_n^3 \Im \hat{f}(a\varepsilon_n), & \text{if } \ a \in A \backslash \rho_{n-1} A, \\ 0, & \text{else,} \end{cases} \\
\lambda_n(a) &:= \begin{cases} (2\pi)^{-\frac{d}{2}} \varepsilon_n^3 \Re \hat{f}(a\varepsilon_n), & \text{if } \ a \in A \backslash \rho_{n-1} A, \\ 0, & \text{else.} \end{cases}
\end{aligned} \tag{24}$$

Note that, due to the definition of κ_n and λ_n, the sum in (23) will be almost always finite. $\Im$ and $\Re$ denote the imaginary and real part, respectively. The definition of κ_n and λ_n seem rather technical. It will become clear from the proofs of Lemma 3.1 and Proposition 3.1. With $\hat{f}$ we denote the Fourier transform of any function $\tilde{f}$ satisfying that $\tilde{f} = f$ in Ω.

Lemma 3.1. *Let f be such that $(1+|\cdot|^{3+\alpha-1/p})\,\hat{f}(\cdot)\in L^p(\mathbb{R}^d)$, where $\hat{f}$ is as above and $\alpha=0$ for $p=1$ and $\alpha>0$ for $1<p\leqslant\infty$, and let A and B be as in Example 3.1. Then it holds for h defined by (23) and (24) with $j\in\mathbb{N}$ sufficiently large (see the definition of ε_n) that*

$$h\in L^p(A\times B).$$

Proof. Let $p<\infty$. Then we obtain with (23) and (24) that

$$\begin{aligned}
&\int_A\int_{-\bar{b}}^{\bar{b}}|h(a,b)|^p\,db\,da\\
&\quad=\sum_{k=1}^{\infty}\int_{\rho_k A\setminus\rho_{k-1}A}\int_{-\bar{b}}^{\bar{b}}\left|\sum_{n=1}^{k}(\kappa_n(a)\cos(b\varepsilon_n)+\lambda_n(a)\sin(b\varepsilon_n))\right|^p db\,da\\
&\quad\leqslant 2\bar{b}\sum_{k=1}^{\infty}\int_{\rho_k A\setminus\rho_{k-1}A}\left(\sum_{n=1}^{k}(|\kappa_n(a)|+|\lambda_n(a)|)\right)^p da\\
&\quad=\mathcal{O}\left(\sum_{k=1}^{\infty}\int_{\rho_k A\setminus\rho_{k-1}A}\left(\sum_{n=1}^{k}\varepsilon_n^3\,|\hat{f}(a\varepsilon_n)|\right)^p da\right).
\end{aligned}$$

This together with the estimate

$$\left(\sum_{n=1}^{k}\varepsilon_n^3\,|\hat{f}(a\varepsilon_n)|\right)^p\leqslant\left(\sum_{n=1}^{k}\varepsilon_n^{(3+\alpha)\,p}\,|\hat{f}(a\varepsilon_n)|^p\right)\left(\sum_{n=1}^{k}\varepsilon_n^{-\frac{\alpha p}{p-1}}\right)^{p-1}$$

and the fact that

$$\sum_{n=1}^{\infty}\varepsilon_n^{-\frac{\alpha p}{p-1}}<\infty,$$

if $\alpha>0$, $p>1$, and $1>\frac{p-1}{\alpha p}$, implies that

$$\begin{aligned}
\int_A\int_{-\bar{b}}^{\bar{b}}|h(a,b)|^p\,db\,da&=\mathcal{O}\left(\sum_{n=1}^{\infty}\int_{A\setminus\rho_{n-1}A}\varepsilon_n^{(3+\alpha)\,p}\,|\hat{f}(a\varepsilon_n)|^p\,da\right)\\
&=\mathcal{O}\left(\sum_{n=1}^{\infty}\int_{\varepsilon_n A\setminus\varepsilon_{n-1}A}\varepsilon_n^{(3+\alpha)\,p-1}\,|\hat{f}(z)|^p\,dz\right)
\end{aligned}$$

if j is sufficiently large and $\alpha=0$ for $p=1$ and $\alpha>0$ for $p>1$ which we assume to hold in the following. Since

$$\exists C>0\ \forall z\in\varepsilon_n A\setminus\varepsilon_{n-1}A:\varepsilon_n^{(3+\alpha)\,p-1}\leqslant C(1+|z|^{3+\alpha-\frac{1}{p}})^p,$$

we finally obtain that

$$\int_A \int_{-\bar{b}}^{\bar{b}} |h(a,b)|^p \, db \, da = \mathcal{O}\left(\sum_{n=1}^{\infty} \int_{\varepsilon_n A \setminus \varepsilon_{n-1} A} (1+|z|^{3+\alpha-\frac{1}{p}})^p \, |\hat{f}(z)|^p \, dz \right)$$
$$= \mathcal{O}\left(\int_{\mathbb{R}^d} (1+|z|^{3+\alpha-\frac{1}{p}})^p \, |\hat{f}(z)|^p \, dz \right).$$

This proves the assertion for $p < \infty$.

Let us now consider the case $p = \infty$: We assume that $\alpha > 0$ and that $j > \frac{1}{\alpha}$. Then we obtain for all $a \in \rho_k A \setminus \rho_{k-1} A$ that

$$\begin{aligned}|h(a,b)| &\leqslant \sum_{n=1}^{k} (|\kappa_n(a)| + |\lambda_n(a)|) \\ &= \mathcal{O}\left(\sum_{n=1}^{k} \varepsilon_n^3 \, |\hat{f}(a\varepsilon_n)| \right) \\ &= \mathcal{O}\left(\sum_{n=1}^{k} (1+(|a| \, \varepsilon_n)^{3+\alpha}) \, \varepsilon_n^{-\alpha} \, |\hat{f}(a\varepsilon_n)| \right) \\ &= \mathcal{O}(\|(1+|\cdot|^{3+\alpha}) \, \hat{f}(\cdot)\|_{L^\infty(\mathbb{R}^d)}).\end{aligned}$$

This proves the assertion for $p = \infty$. ∎

PROPOSITION 3.1. *Let f, A, and B satisfy the conditions in Lemma* 3.1. *Moreover, let f be such that $(1+|\cdot|) \, \hat{f}(\cdot) \in L^1(\mathbb{R}^d)$. Then f has an integral representation* (21) *for some $h \in L^p(A \times B)$.*

Proof. With the special choice of h as in (23) and (24) we know from Lemma 3.1 that $h \in L^p(A \times B)$. We will now show that

$$\begin{aligned}g(x) :&= \int_A \left[\int_{-1-a^T x}^{1-a^T x} h(a,b) \, \frac{1+a^T x + b}{2} \, db + \int_{1-a^T x}^{\bar{b}} h(a,b) \, db \right] da \\ &= \sum_{k=1}^{\infty} \int_{\rho_k A \setminus \rho_{k-1} A} \sum_{n=1}^{k} \left[\kappa_n(a) \left(\int_{-1-a^T x}^{1-a^T x} \cos(b\varepsilon_n) \, \frac{1+a^T x + b}{2} \, db \right.\right. \\ &\qquad \left. + \int_{1-a^T x}^{\bar{b}} \cos(b\varepsilon_n) \, db \right) \\ &\qquad + \lambda_n(a) \left(\int_{-1-a^T x}^{1-a^T x} \sin(b\varepsilon_n) \, \frac{1+a^T x + b}{2} \, db \right. \\ &\qquad \left.\left. + \int_{1-a^T x}^{\bar{b}} \sin(b\varepsilon_n) \, dt \right) \right] da\end{aligned}$$

is identical to f up to a constant. The integrals with respect to b may be calculated analytically. Together with $\sin(\varepsilon_n)=1$ this yields that

$$\begin{aligned} g(x) &= \sum_{k=1}^{\infty} \int_{\rho_k A \setminus \rho_{k-1} A} \sum_{n=1}^{k} [\kappa_n(a)(\varepsilon_n^{-1} \sin(\bar{b}\varepsilon_n) + \varepsilon_n^{-2} \sin(a^T x \varepsilon_n)) \\ &\quad + \lambda_n(a)(-\varepsilon_n^{-1} \cos(\bar{b}\varepsilon_n) + \varepsilon_n^{-2} \cos(a^T x \varepsilon_n))]\, da \\ &= (2\pi)^{-\frac{d}{2}} \sum_{n=1}^{\infty} \int_{\varepsilon_n A \setminus \varepsilon_{n-1} A} (\Re \hat{f}(z) \cos(z^T x) - \Im \hat{f}(z) \sin(z^T x))\, dz \\ &\quad - (2\pi)^{-\frac{d}{2}} \sum_{n=1}^{\infty} \int_{\varepsilon_n A \setminus \varepsilon_{n-1} A} \varepsilon_n (\Re \hat{f}(z) \cos(\bar{b}\varepsilon_n) + \Im \hat{f}(z) \sin(\bar{b}\varepsilon_n))\, dz. \end{aligned}$$

The second term above is a constant, since $(1+|\cdot|)\, \hat{f}(\cdot) \in L^1(\mathbb{R}^d)$. (The proof is similar to the one in Lemma 3.1.) We denote this constant by C in the following. Hence, we obtain that

$$\begin{aligned} g(x) &= (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} (\Re \hat{f}(z) \cos(z^T x) - \Im \hat{f}(z) \sin(z^T x))\, dz + C \\ &= (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} \hat{f}(z)\, e^{iz^T x}\, dz + C \\ &= f(x) + C. \end{aligned}$$

It remains to be shown that the constant function satisfies (21) for some $\bar{h} \in L^\infty(A \times B)$. Let $\bar{h}(a, b) := \frac{C}{\bar{b}\,|A|}$. Then we obtain that

$$\begin{aligned} &\int_A \left[\int_{-1-a^T x}^{1-a^T x} \bar{h}(a, b) \frac{1 + a^T x + b}{2}\, db + \int_{1-a^T x}^{\bar{b}} \bar{h}(a, b)\, db \right] da \\ &= \frac{C}{\bar{b}\,|A|} \int_A (\bar{b} + a^T x)\, da = C, \end{aligned}$$

where we used the fact that

$$\int_A a^T x\, da = 0$$

for the special choice of A (see Example 3.1). ∎

Remark 3.1. For the case $p=1$, the condition $(1+|\cdot|)\, \hat{f}(\cdot) \in L^1(\mathbb{R}^d)$ in Proposition 3.1 is superfluous, since it is implied by condition $(1+|\cdot|^2)\, \hat{f}(\cdot) \in L^1(\mathbb{R}^d)$ in Lemma 3.1. This sufficient condition for (21) actually means that f has a C^2-extension into the exterior of Ω. On the other hand, it is easy to see that for condition (21) to hold it is necessary that f is two-times weakly differentiable.

For the case $p=2$, the conditions in Proposition 3.1 mean that f has a C^1-extension into the exterior of Ω and that f may be extended to a function in $H^{\frac{5}{2}+\alpha}(\mathbb{R}^d)$ for some $\alpha>0$.

For the general case of perceptrons ($k\in\mathbb{N}$) in Example 3.2, one can prove a similar result to Proposition 3.1 by constructing the function h in Lemma 3.1 similarly to (23) and (24). The sufficient conditions for (22) to hold are:

$$(1+|\cdot|)\,\hat{f}(\cdot)\in L^1(\mathbb{R}^d) \qquad \text{and} \qquad (1+|\cdot|^{3+k+\alpha-\frac{1}{p}})\,\hat{f}(\cdot)\in L^p(\mathbb{R}^d).$$

It was shown in [1] that $(1+|\cdot|)\,\hat{f}(\cdot)\in L^1(\mathbb{R}^d)$ is sufficient for the rate

$$\inf_{g\in X_n}\|f-g\|_{L^2(\Omega)}=\mathcal{O}(n^{-1/2})$$

if $P=\mathbb{R}^{d+1}$. It is obvious that better rates can only be obtained under stronger conditions on f. Unfortunately, the rates in Theorem 2.2 are only better than $\mathcal{O}(n^{-\frac{1}{2}})$ if k is sufficiently large depending on the dimension d. On the other hand, the rates in Theorem 2.2 are also valid for non-optimally chosen $\{t_j\}$ (compare Remark 2.1).

REFERENCES

1. A. R. Barron, Universal approximation bounds for superpositions of a sigmoidal function, *IEEE Trans. Inform. Theory* **39** (1993), 930–945.
2. C. M. Bishop, "Neural Networks for Pattern Recognition," Clarendon Press, Oxford, 1995.
3. M. Burger and H. W. Engl, Training neural networks with noisy data as an ill-posed problem, *Adv. Comput. Math.* (2000), to appear.
4. R. A. DeVore, K. I. Oskolkov, and P. P. Petrushev, Approximation by feed-forward neural networks, *Ann. Numer. Math.* **4** (1997), 261–287.
5. H. W. Engl, M. Hanke, and A. Neubauer, "Regularization of Inverse Problems," Kluwer, Dordrecht, 1996.
6. K. Hornik, M. Stinchcombe, and H. White, Multilayer feedforward networks are universal approximators, *Neural Networks* **2** (1989), 359–366.
7. H. N. Mhaskar, Approximation properties of a multilayered feed-forward artificial neural network, *Adv. Comput. Math.* **1** (1993), 61–80.
8. H. N. Mhaskar and C. A. Miccheli, Degree of approximation by neural and translation networks with a single hidden layer, *Adv. Appl. Math.* **16** (1995), 151–183.
9. P. Niyogi and F. Girosi, Generalization bounds for function approximation from scattered noisy data, *Adv. Comput. Math.* **10** (1999), 51–80.
10. P. P. Petrushev, Approximation by ridge functions and neural networks, *SIAM J. Math. Anal.* **30** (1999), 155–189.

11. R. Schaback, Approximation by radial basis functions with finitely many centers, *Constr. Appr.* **12** (1996), 331–340.
12. G. Strang and G. J. Fix, "An Analysis of the Finite Element Method," Prentice–Hall, Englewood Cliffs, NJ, 1973.
13. G. Wahba, Convergence rates of certain approximate solutions to Fredholm integral equations of the first kind, *J. Approx. Theory* **7** (1973), 167–185.