Error Bounds for Approximation with Neural Networks

MARTIN BURGER*AND ANDREAS NEUBAUER

Institut für Industriemathematik, Johannes-Kepler-Universität,
A-4040 Linz, Austria

Abstract. 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 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.

Key Words: Neural networks, error bounds, nonlinear function approximation.

AMS Subject Classifications: 41A30, 41A25, 92B20, 68T05

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

$$f_n(x) = \sum_{j=1}^{n} c_j \phi(x, t_j)\,, \tag{1.1}$$

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 (1.1) are so-called Ridge-constructions, i.e.,

$$f_n(x) = \sum_{j=1}^{n} c_j \sigma(a_j^T x + b_j)\,. \tag{1.2}$$

The interest in such networks grew, since Hornik et al. [5] showed that functions of the form (1.2) 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. [7]).

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

$$\inf_{g \in X_n} \|f - g\|_X\,, \tag{1.3}$$

*Supported by the Austrian Fonds zur Förderung der wissenschaftlichen Forschung under grant SFB F013/1308

where X_n denotes the set of all functions of form (1.1), i.e.,

$$X_n = \{g = \sum_{j=1}^{n} c_j \phi(x, t_j) \ : \ t_j \in P \subset \mathbb{R}^p, c_j \in \mathbb{R}\}\,. \tag{1.4}$$

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

Usually, the convergence of solutions of (1.3) 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 (1.1) converge is just the closure of the range of the integral operator

$$h \mapsto \int_P h(t)\phi(\cdot, t)\, dt\,.$$

Rates are usually only obtained under additional conditions on f (cf., e.g., [4]). A natural condition seems to be that f is in the range of the above operator, i.e.,

$$f(x) = \int_P h(t)\phi(x, t)\, dt \tag{1.5}$$

It was shown in [6] that under this condition the rate

$$\inf_{g \in X_n} \|f - g\|_{L^2(\Omega)} = \mathcal{O}(n^{-\frac{1}{2}}) \tag{1.6}$$

is obtained if ϕ is a continuous function. We improve this result under additional smoothness assumptions on the basis function ϕ in the next section with estimates also in $H^m(\Omega)$. Moreover, we will give error bounds in $W^{m,r}(\Omega)$ that depend on the dimension p of P, 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 (1.5) to hold.

2. Error Bounds

An inspection of the proof of (1.6) in [6] 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* (1.4) *with $P \subset R^p$ bounded and ϕ such that*

$$\|\phi(\cdot, t) - \phi(\cdot, s)\|_{H^m(\Omega)} \le c\|t - s\|^\rho, \quad \rho \in (0, 1]\,, c > 0\,, m \in \mathbb{N}_0\,. \tag{2.1}$$

Moreover, let $f \in H^m(\Omega)$ satisfy (1.5) *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 $\overline{P} := \{t \in P \ : \ h(t) \ge 0\}$ (note that $\overline{P}$ 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

$$\begin{gathered} \overline{P} = \bigcup_{j=1}^{\bar{n}} P_j\,, \qquad P \backslash \overline{P} = \bigcup_{j=\bar{n}+1}^{n} P_j\,, \\ P_i \cap P_j = \{\}\,, i \ne j\,, \quad \mathrm{diam}(P_j) = \mathcal{O}(n^{-\frac{1}{p}})\,, \quad |P_j| = \mathcal{O}(\tfrac{1}{n})\,, \quad i, j = 1, \ldots, n\,. \end{gathered} \tag{2.2}$$

We now define coefficients

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

and probability measures

$$\mu_j(t) := \left\{ \begin{array}{ll} \frac{1}{c_j} h(t)\,, & t \in P_j \\ 0\,, & \text{otherwise} \end{array} \right\} \quad \text{for } c_j > 0, \text{ and } \mu_j \text{ is arbitrary for } c_j = 0\,.$$

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

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

With c_j and μ_j as above and f as in (1.5) we obtain that

$$\begin{aligned}
& E[\|f - \sum_{j=1}^n c_j \phi(\cdot, t_j)\|^2_{H^m(\Omega)}] \\
= \quad & \|f\|^2_{H^m(\Omega)} - 2 \sum_{j=1}^n c_j \langle\, f(x), \int_P \mu_j(t_j) \phi(\cdot, t_j)\, dt_j\, \rangle_{H^m(\Omega)} \\
& + \sum_{i \neq j = 1}^n c_i c_j \langle \int_P \mu_i(t_i) \phi(\cdot, t_i)\, dt_i, \int_P \mu_j(t_j) \phi(\cdot, t_j)\, dt_j\, \rangle_{H^m(\Omega)} \\
& + \sum_{j=1}^n c_j^2 \int_P \mu_j(t_j)\, \|\phi(\cdot, t_j)\|^2_{H^m(\Omega)} dt_j \\
= \quad & \| \int_P [h(t) - \sum_{j=1}^n c_j \mu_j(t)] \phi(\cdot, t)\, dt \|^2_{H^m(\Omega)} \\
& + \sum_{j=1}^n c_j^2 \Big[\int_P \mu_j(t)\, \|\phi(\cdot, t)\|^2_{H^m(\Omega)} dt - \| \int_P \mu_j(t) \phi(\cdot, t)\, dt \|^2_{H^m(\Omega)} \Big] \\
= \quad & \sum_{j=1}^n c_j^2 \sum_{|\alpha| \le m} \int_\Omega \Big[\int_P \mu_j(t) (\tfrac{\partial^{|\alpha|}}{\partial x^\alpha} \phi(x, t))^2 dt - \Big(\int_P \mu_j(t) \tfrac{\partial^{|\alpha|}}{\partial x^\alpha} \phi(x, t)\, dt \Big)^2 \Big] dx \\
= \quad & \sum_{j=1}^n c_j^2 \sum_{|\alpha| \le m} \int_\Omega \Big[\int_{P_j} \mu_j(t) \Big(\int_{P_j} \mu_j(s) (\tfrac{\partial^{|\alpha|}}{\partial x^\alpha} \phi(x, t) - \tfrac{\partial^{|\alpha|}}{\partial x^\alpha} \phi(x, s))\, ds \Big)^2 dt \Big] dx
\end{aligned}$$

Noting that $h \in L^\infty(P)$ and (2.2) imply that $c_j = \mathcal{O}(\frac{1}{n})$, we now obtain together with (2.1) and (2.2) that

$$\begin{aligned}
& E[\|f - \sum_{j=1}^n c_j \phi(\cdot, t_j)\|^2_{H^m(\Omega)}] \\
\le \quad & \sum_{j=1}^n c_j^2 \sum_{|\alpha| \le m} \int_\Omega \Big[\int_{P_j} \mu_j(t) \int_{P_j} \mu_j(s) \Big(\tfrac{\partial^{|\alpha|}}{\partial x^\alpha} \phi(x, t) - \tfrac{\partial^{|\alpha|}}{\partial x^\alpha} \phi(x, s) \Big)^2 ds\, dt \Big] dx \\
= \quad & \sum_{j=1}^n c_j^2 \int_{P_j} \mu_j(t) \int_{P_j} \mu_j(s)\, \|\phi(\cdot, t) - \phi(\cdot, s)\|^2_{H^m(\Omega)}\, ds\, dt \\
= \quad & \mathcal{O}(n \cdot n^{-2} \cdot n^{-\frac{2\rho}{p}}) = \mathcal{O}(n^{-1-\frac{2\rho}{p}})
\end{aligned}$$

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

$$\begin{aligned}\inf_{g\in X_n} \|f-g\|_{H^m(\Omega)} \le \|f-\sum_{j=1}^n c_j\phi(\cdot,\bar{t}_j)\|_{H^m(\Omega)} &\le \Big(E[\|f-\sum_{j=1}^n c_j\phi(\cdot,t_j)\|^2_{H^m(\Omega)}]\Big)^{\frac{1}{2}} \\ &= \mathcal{O}(n^{-\frac{1}{2}-\frac{\varrho}{p}})\end{aligned}$$

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 (2.2) 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 $P \subset R^p$. The proofs are based on the following results from finite-element theory:

Let

$$P := \bigtimes_{i=1}^p [\underline{p}_i, \overline{p}_i] \quad \text{and} \quad P_{l_1\dots l_p} := \bigtimes_{i=1}^p [\underline{p}_i + \tfrac{\overline{p}_i-\underline{p}_i}{\tau} l_i, \underline{p}_i + \tfrac{\overline{p}_i-\underline{p}_i}{\tau}(l_i+1)]\,, \quad \tau \in \mathbb{N}\,.$$

Then, obviously

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

Moreover, we define

$$t_{j_1\dots j_p} := (t_{j_1\dots j_p;1},\dots,t_{j_1\dots j_p;p}) \in \mathbb{R}^p\,, \qquad t_{j_1\dots j_p;i} := \underline{p}_i + \tfrac{\overline{p}_i-\underline{p}_i}{k\tau} j_i\,, \quad j_i = 0,\dots,k\tau\,. \tag{2.3}$$

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

$$\begin{aligned} q_{\nu_1\dots\nu_p} \in Q_{k,l_1\dots l_p} := \{q(t) = \sum c_{j_1\dots j_p} t_1^{j_1}\cdots t_p^{j_p} \;\; : \;\; & 0 \le j_i \le k,\; 1 \le i \le p, \\ & t = (t_1,\dots,t_p) \in P_{l_1\dots l_p}\} \end{aligned} \tag{2.4}$$

satisfying

$$q_{\nu_1\dots\nu_p}(t_{j_1\dots j_p}) = \prod_{i=1}^p \delta_{\nu_i j_i}\,, \quad kl_i \le \nu_i,\; j_i \le k(l_i+1)\,. \tag{2.5}$$

The function u_I, defined by

$$u_I\Big|_{P_{l_1\dots l_p}} := \sum_{kl_i \le j_i \le k(l_i+1)} u(t_{j_1\dots j_p}) q_{j_1\dots j_p}\,, \tag{2.6}$$

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

Proposition 2.2. *Let $P \subset \mathbb{R}^p$ be rectangular. If $u \in H^k(P)$ with $k > \frac{p}{2}$, then there is a constant $c_\kappa > 0$ such that for all multiindices β with $|\beta| = \kappa < k$ and for all $l_i \in \{0,\dots,\tau-1\}$, $i = 1,\dots,p$, it holds that*

$$\|D^\beta(u-u_I)\|_{L^2(P_{l_1\dots l_p})} \le c_\kappa \tau^{-(k-\kappa)} |u|_{H^k(P_{l_1\dots l_p})}\,. \tag{2.7}$$

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

$$\|D^\beta(u-u_I)\|_{L^\infty(P_{l_1\dots l_p})} \le \overline{c}_\kappa \tau^{-(k-\kappa)} \max_{|\gamma|=k} \|D^\gamma u\|_{L^\infty(P_{l_1\dots l_p})}\,. \tag{2.8}$$

Proof. The proof follows with Theorem 3.1 and Theorem 3.3 in [8]. ∎

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

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

Theorem 2.3. *Let X_n be defined as in (1.4) with $P \subset R^p$ 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 (1.5) 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 (1.5), then we obtain that

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

Let us define $\tau := \frac{[n^{\frac{1}{p}}]-1}{k}$ and $\bar n := (k\tau+1)^p \le n$. Then we choose t_j and γ_j as follows: For $j = \bar n+1, \ldots, n$ let t_j be arbitrary and $\gamma_j \equiv 0$. For $j = 1, \ldots, \bar n$ 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 (2.3) – (2.6)), i.e.,

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

Note that this interpolating property also holds for all derivatives of ϕ with respect to x. Applying (2.7) ($\beta = 0$) for $Y = H^k(P)$ and (2.8) ($\beta = 0$) for $Y = C^k(P)$ we obtain the estimates

$$\|f - \sum_{j=1}^n c_j\phi(\cdot,t_j)\|_{W^{m,r}(\Omega)} \le c_0\tau^{-k}\|h\|_{L^2(P)}\|\phi\|_{W^{m,r}(\Omega,H^k(P))} \tag{2.9}$$

and

$$\|f - \sum_{j=1}^n c_j\phi(\cdot,t_j)\|_{W^{m,r}(\Omega)} \le \bar c_0\tau^{-k}\|h\|_{L^1(P)}\|\phi\|_{W^{m,r}(\Omega,C^k(P))} \tag{2.10}$$

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

Remark 2.4. The idea of choosing c_j, t_j and γ_j as in the prove above was found in a paper by Whaba [9] for one-dimensional P. This idea was extended to higher dimensions, i.e., $P \subset \mathbb{R}^p$.

The following extensions of Theorem 2.3 are obvious from the proof:

- If P is not rectangular but $\text{supp}(h) \subset \overline{P} \subset P$ with $\overline{P}$ rectangular, then the results are still valid.
- If $Y = C^k(P)$, the condition (1.5) 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

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

- Condition (1.5) may be generalized to

$$f(x) = \sum_{|\alpha| \leq \kappa} \int_P h_\beta(t) \tfrac{\partial^{|\beta|}}{\partial t^\beta} \phi(x, t)\, dt\,, \quad \kappa < k\,. \tag{2.11}$$

  If the functions γ_j are chosen such that for each β they coincide with the appropriate derivative of the basis functions $q_{j_1 \dots j_p}$ in $P_{l_1 \dots l_p}$, we obtain together with Proposition 2.2 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 2.3 decrease with increasing dimension p. There is no dimensionless term like $n^{-\frac{1}{2}}$ in (1.6) or Theorem 2.1. Since the estimates in the proof of Theorem 2.3 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 2.2 is valid also for many other non-uniform choices of knots t_j, the rates in Theorem 2.3 are valid for many choices t_j (also non-optimal ones) if at least c_j is chosen optimally.

3. Application to perceptrons

We now apply the results of the previous section to perceptrons with a single hidden layer, namely Ridge-constructions (cf. (1.2)) where σ is a function of sigmoidal form, i.e.,

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

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

$$\sigma(t) := \begin{cases} 1\,, & t > 1\,, \\ p(t)\,, & -1 \leq t \leq 1\,, \\ 0\,, & t < -1\,, \end{cases} \tag{3.1}$$

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

$$p(-1) = 0\,,\ p(1) = 1\,,\ \text{and } p^{(l)}(-1) = 0 = p^{(l)}(1)\,,\ 1 \leq l \leq k\,, \tag{3.2}$$

then $\sigma \in C^{k,1}$ and $\sigma \in W^{k+1,\infty}$ (see Figure 3.1).

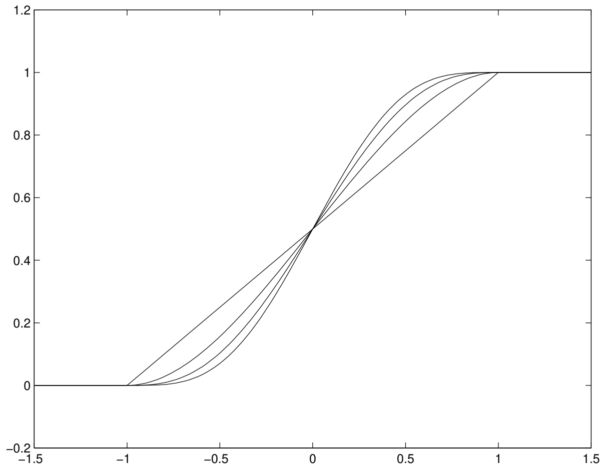

Figure 3.1: Function σ from (3.1) and (3.2) for $k = 0, 1, 2, 3$

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

$$\sigma(t) := \begin{cases} 1\,, & t > 1\,, \\ \frac{t+1}{2}\,, & -1 \le t \le 1\,, \\ 0\,, & t < -1\,, \end{cases} \tag{3.3}$$

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

$$\forall a \in A \;\; \forall x \in \Omega : \; |a^T x| \le \overline{b} - 1\,.$$

Since $\phi(x, a, b) := \sigma(a^T x + b)$ satisfies (2.1) 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

$$\begin{aligned} f(x) &= \int_A \int_{-\overline{b}}^{\overline{b}} h(a, b) \sigma(a^T x + b)\, db\, da \\ &= \int_A \Big[\int_{-1-a^T x}^{1-a^T x} h(a, b) \tfrac{1 + a^T x + b}{2}\, db + \int_{1-a^T x}^{\overline{b}} h(a, b)\, db \Big] da \end{aligned} \tag{3.4}$$

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

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

Since $\phi(x, a, b) := \sigma(a^T x + b)$ satisfies that $\phi \in W^{m,\infty}(\Omega, C^{k-m}(A \times B))$ $(m \le k)$ and $\phi \in W^{m,\infty}(\Omega, H^{k+1-m}(A \times B))$ $(m \le k + 1)$, we may apply Theorem 2.3 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

$$f(x) = \int_A \Big[\int_{-1-a^T x}^{1-a^T x} h(a, b) p(a^T x + b)\, db + \int_{1-a^T x}^{\overline{b}} h(a, b)\, db \Big] da \tag{3.5}$$

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 (3.5) 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 (3.4) and (3.5) can be only satisfied if f is several times differentiable. We will now give a sufficient condition on f that guarantees (3.4):

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 $\rho_n := \frac{\varepsilon_n}{\varepsilon_{n+1}}$. We define the function h as follows:

$$h(a,b) = \sum_{n=1}^{\infty} (\kappa_n(a) \cos(b\varepsilon_n) + \lambda_n(a) \sin(b\varepsilon_n)) \tag{3.6}$$

where

$$\begin{aligned} \kappa_n(a) &:= \begin{cases} -(2\pi)^{-\frac{d}{2}} \varepsilon_n^3 \Im \hat{f}(a\varepsilon_n)\,, & \text{if } a \in A \backslash \rho_{n-1} A\,, \\ 0\,, & \text{else}\,, \end{cases} \\ \lambda_n(a) &:= \begin{cases} (2\pi)^{-\frac{d}{2}} \varepsilon_n^3 \Re \hat{f}(a\varepsilon_n)\,, & \text{if } a \in A \backslash \rho_{n-1} A\,, \\ 0\,, & \text{else}\,. \end{cases} \end{aligned} \tag{3.7}$$

Note that, due to the definition of κ_n and λ_n, the sum in (3.6) will be almost always finite. $\Im$ and $\Re$ denote the imaginary and real part, respectively. With $\hat{f}$ we denote the Fourier transform of any function $\tilde{f}$ satisfying $\tilde{f} = f$ in Ω.

Lemma 3.3. *Let f be such that $(1 + |\cdot|^{3+\alpha-\frac{1}{p}})\hat{f}(\cdot) \in L^p(\mathbb{R}^d)$, where $\hat{f}$ is as above and $\alpha = 0$ for $p = 1$ and $\alpha > 0$ for $1 < p \leq \infty$, and let A and B be as in Example* 3.1. *Then it holds for h defined by* (3.6) *and* (3.7) *with $j \in \mathbb{N}$ sufficiently large (see the definition of ε_n above) that*

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

Proof. Let $p < \infty$. Then we obtain with (3.6), (3.7), and ε_n, ρ_k defined as above that

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

This together with the estimate

$$\Big(\sum_{n=1}^{k} \varepsilon_n^3 |\hat{f}(a\varepsilon_n)| \Big)^p \leq \Big(\sum_{n=1}^{k} \varepsilon_n^{(3+\alpha)p} |\hat{f}(a\varepsilon_n)|^p \Big) \Big(\sum_{n=1}^{k} \varepsilon_n^{-\frac{\alpha p}{p-1}} \Big)^{p-1}$$

and the fact that

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

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

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

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

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

we finally obtain that

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

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 $a \in \rho_k A\backslash\rho_{k-1}A$ that

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

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

Proposition 3.4. *Let f, A, and B satisfy the conditions in Lemma 3.3. Moreover, let f be such that $(1 + |\cdot|)\hat f(\cdot) \in L^1(\mathbb{R}^d)$. Then f has an integral representation (3.4) for some $h \in L^p(A \times B)$.*

Proof. With the special choice of h as in (3.6) and (3.7) we know from Lemma 3.3 that $h \in L^p(A \times B)$. We will now show that

$$\begin{aligned}g(x) &:= \int_A \Big[\int_{-1-a^Tx}^{1-a^Tx} h(a,b)\tfrac{1+a^Tx+b}{2}\, db + \int_{1-a^Tx}^{\bar b} h(a,b)\, db\Big] da\\ &= \sum_{k=1}^{\infty} \int_{\rho_k A\backslash\rho_{k-1}A} \sum_{n=1}^{k} \Big[\kappa_n(a)\Big(\int_{-1-a^Tx}^{1-a^Tx} \cos(b\varepsilon_n)\tfrac{1+a^Tx+b}{2}\, db + \int_{1-a^Tx}^{\bar b} \cos(b\varepsilon_n)\, db\Big)\\ &\qquad + \lambda_n(a)\Big(\int_{-1-a^Tx}^{1-a^Tx} \sin(b\varepsilon_n)\tfrac{1+a^Tx+b}{2}\, db + \int_{1-a^Tx}^{\bar b} \sin(b\varepsilon_n)\, db\Big)\Big] da\end{aligned}$$

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

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

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 3.3.) 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 (3.4) for some $\overline{h} \in L^\infty(A \times B)$. Let $\overline{h}(a,b) := \frac{C}{\overline{b}|A|}$. Then we obtain that

$$
\int_A \Big[\int_{-1-a^T x}^{1-a^T x} \overline{h}(a,b) \tfrac{1+a^T x + b}{2}\, db + \int_{1-a^T x}^{\overline{b}} \overline{h}(a,b)\, db \Big] da = \tfrac{C}{\overline{b}|A|} \int_A (\overline{b} + a^T x)\, da = C\,,
$$

where we used the fact that

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

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

Remark 3.5. For the case $p = 1$, the condition $(1 + |\cdot|)\hat{f}(\cdot) \in L^1(\mathbb{R}^d)$ in Proposition 3.4 is superfluous, since it is implied by condition $(1 + |\cdot|^2)\hat{f}(\cdot) \in L^1(\mathbb{R}^d)$ in Lemma 3.3. This sufficient condition for (3.4) 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 (3.4) to hold it is necessary that f is two-times weakly differentiable.

For the case $p = 2$, the conditions in Proposition 3.4 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 3.4 by constructing the function h in Lemma 3.3 similarly to (3.6) and (3.7). The sufficient conditions for (3.5) to hold are:

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

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

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

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 2.3 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 2.3 are also valid for non-optimally chosen $\{t_j\}$ (compare Remark 2.4).

References

[1] A. R. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Trans. Inf. 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. Comp. Math. (2000), to appear.

[4] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht, 1996.

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

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

[7] R. Schaback, *Approximation by radial basis functions with finitely many centers*, Const. Appr. 12 (1996), 331–340.

[8] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, 1973.

[9] G. Wahba, *Convergence rates of certain approximate solutions to Fredholm integral equations of the first kind*, J. Approx. Theory 7 (1973), 167–185.