Training Neural Networks with Noisy Data as an Ill-Posed Problem

Martin Burger Heinz W. Engl

Industrial Mathematics Institute, Johannes Kepler Universität Linz, Altenbergerstr. 69, A-4040 Linz, Austria.

Abstract

This paper is devoted to the analysis of network approximation in the framework of approximation and regularization theory. It is shown that training neural networks and similar network approximation techniques are equivalent to least-squares collocation for a corresponding integral equation with mollified data.

Results about convergence and convergence rates for exact data are derived based upon well-known convergence results about least-squares collocation. Finally, the stability properties with respect to errors in the data are examined and stability bounds are obtained, which yield rules for the choice of the number of network elements.

Keywords: ill-posed problems, least-squares collocation, neural networks, network training, regularization.

AMS Subject Classification: 41A15, 41A30, 45L10, 65J20, 92B20.

Short Title: Training Neural Networks with Noisy Data.

1 Introduction

In this paper we consider the approximation of a function $f \in H^s(\Omega)$ by a (neural) network of the form

$$f_n(x) = \sum_{j=1}^{n} c_i \phi(x; t_i), \tag{1.1}$$

where Ω is a bounded domain in $\mathbf{R}^d$ and $H^s(\Omega)$ denotes the Sobolev space of order $s \in \mathbf{N}_0$. Recently, there is a growing interest in applying such approximation schemes in practice, a popular special case is the approximation with one-layer feed-forward neural networks, but also radial basis function networks or wavelets fit into this scheme. We will give more details about these specific cases in Section 1.1 - 1.3.

In practice, instead of f only a noisy observation $f^\delta \in L^2(\Omega)$ with

$$\|f - f^\delta\|_{L^2(\Omega)} \leq \delta \tag{1.2}$$

is known; the function f^δ represents inexactly measured and interpolated data. Non-parametric approximation of f is based on choosing that function f_n which satisfies some interpolation or moment conditions and minimizes some additional functional, e.g. the norm. Such approximations have been introduced as abstract splines by Sard [34] and generalized by Groetsch

[17]. The latter proved that the approximation obtained in that way depends continuously upon the data if and only if the set of *attainable* functions is closed. Since we are interested in approximation schemes in $H^s(\Omega)$ and $C(\Omega)$ and these spaces are non-closed subspaces of $L^2(\Omega)$, the approximation problem is *ill-posed*, i.e., the solution does not depend upon the data in a stable way (see e.g. [10, 23, 36] for a general overview of ill-posed problems). We will show below that approximations of the form (1.1) fit into this framework, i.e., satisfy some kind of moment conditions and minimize the H^s-norm under all such functions. This means that we have to expect instabilities as n tends to infinity, if we do not choose n carefully, i.e., the approximations f_n^δ obtained with noisy data will not converge to f as $n \to \infty$ and $\delta \to 0$.

Hence, in order to approximate f in a stable way, regularization methods should be used. In general, regularization methods for the solution of linear equations

$$Ax = y, \tag{1.3}$$

where A is a linear operator acting between two Hilbert spaces, replace the generalized inverse $A^\dagger$ (or the inverse A^{-1}, if it exists) by a family of continuous operators R_α, which converge pointwise to $A^\dagger$. The *regularization parameter* α is chosen dependent upon the noise level and possibly upon the data, i.e., $\alpha = \alpha(\delta, y^\delta)$. The regularized solutions $R_\alpha y^\delta$ converge to $A^\dagger y$ as $\delta \to 0$ only if certain conditions upon the choice of α are satisfied (cf. e.g. [10]).

Approximations such as (1.1) can be interpreted as regularization methods, where the regularization parameter is related to the number n of basis functions. It turns out that the condition number of the linear system obtained for the computation of the coefficients c_i will tend to infinity with $n \to \infty$. The error between the exact solution f and the approximation obtained for noisy data (denoted by f_n^δ), can be estimated by

$$\|f - f_n^\delta\| \leq \|f - f_n\| + \|f_n - f_n^\delta\|, \tag{1.4}$$

where f_n denotes the approximation of the form (1.1) with exact data. Thus, the first term on the right-hand side of (1.4) is an approximation error, whereas the second term is a stability bound on the finite-dimensional subspace $\text{span}\{\phi(.;t_i)\}$. We note that (1.4) is sharp in the sense that for any δ one can find a perturbation of noise level δ which yields equality in (1.4). Although there exists a large variety of results about the approximation by neural networks (c.f. [3, 7, 13, 21, 26, 29] and the references quoted there), the stability aspect has been largely neglected so far, only few authors give a rigourous treatment of regularization methods for the approximation problem (cf. [14, 15, 16, 33, 40]).

In Section 2 we will show that network approximation in Sobolev spaces is equivalent to *least-squares collocation* for a corresponding integral equation of the first kind. Based on the well-known results about least-squares collocation we will derive results about the convergence in the case of exact data in Section 3. It turns out, that the rate of convergence will increase with the smoothness of the network function, but the set of functions on which the approximation converges is smaller for smooth basis functions.

In Section 4 we will use these results to obtain convergence results for perturbed data and to construct strategies for the optimal choice of the number of knots in the network.

1.1 Artificial Neural Networks

Artificial Neural Networks are a popular method of nonlinear system identification and function approximation. In their simplest form, namely as *one-layer feed-forward networks*, they can be written in the form (1.1) with

$$\phi(x;t) = \sigma(a^T x - b), \tag{1.5}$$

where $t_i = (a,b) \in \mathbf{R}^{d+1}$ or $t_i \in S^1(\mathbf{R}^d) \times \mathbf{R}$ and σ is of sigmoidal shape, in most applications either the Heaviside function (cf. e.g. [21, 26])

$$\sigma(x) = \begin{cases} 0 & \text{if x} < 0 \\ 1 & \text{if x} \geq 0 \end{cases} \tag{1.6}$$

or the smooth sigmoidal function (cf. e.g. [7, 35])

$$\sigma(z) = \frac{1}{1+e^{-z}}. \tag{1.7}$$

So-called *multi-layer neural networks* are constructed in a similar way. The parameter (a_i^m, b_i^m) in the $m-th$ layer are calculated via (1.1), (1.5) from a network with the parameter $\{(a_j^{m-1}, b_j^{m-1})\}_{j=1,\ldots,n}$ and so on. Obviously this procedure is expected to produce a network which is more flexible than a one-layer feed-forward neural net. Nevertheless the class of functions which can be approximated with a single layer is usually large enough for many practical applications. Our analysis will include only one-layer networks, a similar analysis for multi-layer networks should be based on regularization methods for nonlinear problems (c.f. e.g. [10, Section 10]) and is left to future work.

In most cases of approximation by neural networks, the parameter (a_i, b_i) are not chosen a-priori, but determined by an optimization procedure, too. In the language of neural networks, the estimation of the parameter (a_i, b_i, c_i) is called *training*. Due to the specific structure of the network, efficient methods like *back-propagation* (cf. [4, 35]) can be employed for the calculation of gradients, which is necessary for the numerical solution of the minimization problem. We will give an analysis for arbitrary choice of the parameter t, obviously all approximation results hold for the optimal choice of t, too (which is the case so far mainly studied in the literature, cf. [3, 7, 13, 26, 33]). The stability results deduced in Section 4 cannot be applied to the case of optimized t in a simple way, since the dependence of t upon the data f^δ has to be examined additionally. An extension of the stability results to this nonlinear problem will be one of our main future projects.

So far, different kinds of regularization methods have been used for the training of neural networks with the aim of reducing the complexity of the network (cf. [5]). It turns out that even for exact data, regularized networks are able to produce better results if the function to be approximated is only sampled at a finite number of points (cf. e.g. [35]). The error resulting from this lack of information is usually called *generalization error*, obviously it will be an important task for the future to investigate the effects of regularization on all kinds of errors, i.e., approximation and generalization error and stability bounds, together.

1.2 Radial Basis Function Networks

Another important class of networks covered by (1.1) are so-called *radial basis function networks* (cf. e.g. [4, 35]), i.e.,

$$\phi(x;t) = \psi(\|x-t\|), \tag{1.8}$$

with $t \in \mathbf{R}^d$ or $t \in \Omega$. The function ψ is usually chosen to be some kind of peak function, a frequently used basis function is the gaussian

$$\psi(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{z^2}{2\sigma^2}}, \tag{1.9}$$

another common choice is the multiquadric function

$$\psi(z) = \frac{\sqrt{r^2 + z^2}}{r}. \tag{1.10}$$

Contrary to the philosophy of neural networks, the parameters t_i are not optimized in most applications, which yields even more efficient algorithms because of the simple linear convolutive structure (cf. [30]). Obviously, a disadvantage of a model with fixed parameters is the loss of flexibility; nevertheless the parameters can be chosen a-priori in a meaningful way if additional information about the structure of the solution is available.

1.3 Wavelets and Gabor Transforms

In the context of image and signal processing, approximations based on wavelets (cf. e.g. [8, 27]) and Gabor analysis (cf. e.g. [11]) are heavily used. With an extension to complex-valued basis functions and coefficients, they fit into the form (1.1), with

$$\phi(x; t_{jk}) = a_j^{-\frac{1}{2}} \psi\left(\frac{x - b_{jk}}{a_j}\right) \tag{1.11}$$

for wavelets, where usually $a_j = a_0^j$ and $b_{jk} = a_0^j k b_0$; respectively

$$\phi(x; t_{jk}) = g(x - b_k) e^{2\pi i a_j} \tag{1.12}$$

for Gabor transforms, where usually $a_j = j a_0$ and $b_k = k b_0$. The parameter t_{jk} are not trained (i.e., optimized) in such applications, but chosen a-priori. Connections between wavelets and regularization methods have been examined recently (cf. [1, 12]).

2 Network Approximation and Least-squares Collocation

We assume that $\phi \in C^{2s}(\Omega \times P)$, where $\Omega \subset \mathbf{R}^d$ and the set of parameters $P \subset \mathbf{R}^p$ are bounded domains. Let $L_s : H^s(\Omega) \to L^2(\Omega)$ denote the linear operator with

$$\|g\|_{H^s(\Omega)} = \|L_s g\|_{L^2(\Omega)}, \qquad \forall\, g \in H^s(\Omega). \tag{2.1}$$

Under the above smoothness assumptions we may define a continuous function $k \in C(\Omega \times P)$ by

$$k(x, t) := L_s^* L_s \phi(x, t), \tag{2.2}$$

where $L_s^* : L^2(\Omega) \to H^s(\Omega)$ is the adjoint of L_s and the elliptic differential operator $L_s^* L_s$ is evaluated with respect to x. Furthermore, we define an integral operator by

$$\begin{array}{rccc} \mathcal{F}_k : & H^s(\Omega) & \to & L^2(P) \\ & g & \mapsto & \int_\Omega k(x, .) g(x)\; dx. \end{array} \tag{2.3}$$

The operator $\mathcal{F}_k$ has the property that

$$\begin{aligned}
\langle \phi(.,t), g\rangle_{H^s(\Omega)} &= \langle L_s\phi(.,t), L_s g\rangle_{L^2(\Omega)} \\
&= \langle L_s^* L_s \phi(.,t), g\rangle_{L^2(\Omega)} \\
&= \langle k(.,t), g\rangle_{L^2(\Omega)} \\
&= (\mathcal{F}_k g)(t).
\end{aligned} \tag{2.4}$$

Lemma 2.1. *The linear operator $\mathcal{F}_k$ is continuous from $L^2(\Omega)$ (and $H^s(\Omega)$) to $L^2(P)$ and $C(P)$. Its adjoint $\mathcal{F}_k^* : L^2(P) \to H^s(\Omega)$ is given by*

$$\mathcal{F}_k^* h = \int_P \phi(.,t)h(t)\ dt. \tag{2.5}$$

Proof. The fact that $\mathcal{F}_k$ maps $L^2(\Omega)$ to $C(P) \subset L^2(P)$ and the continuity on these spaces is a standard result (cf. [18]). Because of the continuous embedding in $L^2(\Omega)$ it is also continuous from $H^s(\Omega)$ to $L^2(P)$ and $C(P)$.

Using the definition of k and Fubini's theorem we obtain

$$\begin{aligned}
\langle \mathcal{F}_k g, h\rangle_{L^2(P)} &= \int_P h(t) \int_\Omega k(x,t) g(x)\ dx\ dt \\
&= \int_P h(t) \int_\Omega L_s\phi(x,t) L_s g(x)\ dx\ dt \\
&= \int_\Omega L_s g(x) L_s \left(\int_P \phi(x,t) h(t)\ dt \right)\ dx \\
&= \langle g, \mathcal{F}_k^* h\rangle_{H^s(\Omega)},
\end{aligned}$$

and hence,

$$\mathcal{F}_k^* h = \int_P \phi(.,t)h(t)\ dt.$$

□

Now we can characterize the best approximating function of the form (1.1) as the solution of a discretized integral equation involving the operator $\mathcal{F}_k$:

Lemma 2.2. *Let f_n be the unique minimizer of the approimation problem, i.e.,*

$$\|f - f_n\|_{H^s(\Omega)} = \min_{f_n \in\ span\{\phi(.,t_i)\}} \|f - g_n\|_{H^s(\Omega)} \tag{2.6}$$

with fixed t_j, $1 \le j \le n$. Then f_n is the minimum-norm solution of

$$(\mathcal{F}_k f_n)(t_j) = y_j, \qquad j = 1, \dots, n, \tag{2.7}$$

with the right-hand side defined by

$$y_j := \int_\Omega f(x) k(x, t_j)\ dx. \tag{2.8}$$

Proof. Since the minimization problem (2.6) is continuous, strictly convex and quadratic, its solution is unique and characterized by the condition

$$\begin{aligned} 0 &= \langle f - f_n, \phi(., t_j)\rangle_{H^s(\Omega)} \\ &= \langle f - f_n, L_s^* L_s \phi(., t_j)\rangle_{L^2(\Omega)} \\ &= \langle f - f_n, k(., t_j)\rangle_{L^2(\Omega)} \\ &= y_j - (\mathcal{F}_k f_n)(t_j). \end{aligned}$$

This shows that f_n is a solution of the discretized integral equation; it remains to verify that it is the one with minimal norm in $H^s(\Omega)$. Let g be any other solution of (2.7), then

$$\begin{aligned} \langle g - f_n, f_n\rangle_{H^s(\Omega)} &= \langle g - f_n, L_s^* L_s f_n\rangle_{L^2(\Omega)} \\ &= \sum_{j=1}^{n} c_j \langle g - f_n, L_s^* L_s \phi(., t_j)\rangle_{L^2(\Omega)} \\ &= \sum_{j=1}^{n} c_j \langle g - f_n, k(., t_j)\rangle_{L^2(\Omega)} \\ &= \sum_{j=1}^{n} c_j \big((\mathcal{F}_k g)(t_j) - (\mathcal{F}_k f_n)(t_j)\big) = 0. \end{aligned}$$

Using this equality, we obtain

$$\begin{aligned} \|g\|^2_{H^s(\Omega)} &= \|f_n\|^2_{H^s(\Omega)} + \|g - f_n\|^2_{H^s(\Omega)} + 2\langle g - f_n, f_n\rangle_{H^s(\Omega)} \\ &= \|f_n\|^2_{H^s(\Omega)} + \|g - f_n\|^2_{H^s(\Omega)} \geq \|f_n\|^2_{H^s(\Omega)}. \end{aligned}$$

Thus, f_n is the minimum-norm solution. □

We note that the statement of Lemma 2.2 still holds if f is replaced by $f^\delta \in L^2(\Omega)$, since due to Lemma 2.1, $\mathcal{F}_k$ is a continuous operator from $L^2(\Omega)$ onto $C(P)$.

We can now interpret the network approximation procedure as a two-step algorithm:

1. The mollified data function

$$y := \mathcal{F}_k f \ \in C(P) \tag{2.9}$$

   is computed.

2. The integral equation

$$y(t) = (\mathcal{F}_k g)(t) = \int_\Omega k(x,t) g(x)\ dx, \qquad t \in P, \tag{2.10}$$

   is solved approximately via least-squares collocation in (2.7).

The second step is a well-known regularization method for Fredholm integral equations of the first kind (cf. [10, 31, 37]) and other ill-posed linear operator equations (cf. [9, 38]), called *least-squares collocation*.

The coefficients c_j can be computed explicitely by solving the following $n \times n$ system (cf. e.g. [9]):

$$\Phi_n C_n = Y_n, \tag{2.11}$$

where Φ_n is the matrix

$$\begin{aligned} \Phi_n &= (\langle \phi(.,t_i), \phi(.,t_j) \rangle_{H^s(\Omega)})_{i,j=1,\dots,n} \\ &= (\langle \phi(.,t_i), k(.,t_j) \rangle_{L^2(\Omega)})_{i,j=1,\dots,n}, \end{aligned} \tag{2.12}$$

and the vectors C_n and Y_n are defined by

$$C_n := (c_1, \dots, c_n)^T \tag{2.13}$$

$$Y_n := (y_1, \dots, y_n)^T. \tag{2.14}$$

The matrix Φ_n is positive definite, hence the discretized equation can be solved in a stable way. Since the integral operator $\mathcal{F}_k$ is compact under the above assumptions (cf. [18]), the minimal eigenvalue of Φ_n must tend to zero, i.e., the condition number of Φ_n is increasing and the constants in the stability estimates will diverge. We will investigate this problem in Section 4.

Mollification is a well-known concept in regularization theory, too. The basic idea is to replace the data f^δ by $M_\gamma f^\delta$, where $\{M_\gamma\}_{\gamma>0}$ is a sequence of smoothing operators such that $M_\gamma f \to f$ as $\gamma \to 0$ (cf. [20, 24, 25, 28]). This means that if the mollifier is an integral operator on $L^2(\Omega)$, its kernel should approximate the Dirac delta-distribution in an appropriate space, e.g. in $H^{-s}(\Omega)$. As we have seen before (cf. (2.9),(2.10)), network approximation combines mollification with an approximate inversion of M_γ, thus it is not necessary to let γ tend to zero. Nonetheless, also in the case of network approximation it is a desireable property that the kernel k approximates the Dirac delta, which means that less information is lost in the mollification step and that the reconstruction step works better. The limiting case of

$$k(x,y) = L_s^* L_s \phi(x,y) = \delta(x-y) \tag{2.15}$$

yields a network whose basis function is the Green's function of the differential operator $L_s^* L_s$. These so-called *regularization networks* have been introduced by Girosi and Poggio (cf. [14, 15, 16]) and analyzed with respect to their approximation properties. The class of functions generated by a regularization network is dense in $H^s(\Omega)$. A disadvantage of these networks is that the 'mollifier' (the Dirac delta) is not a continuous operator on $L^2(\Omega)$, and hence stability results similar to Section 4 cannot be expected in this framework.

A similar problem arises when the network approximation is interpreted as a projection method. Formally one may look at (2.7), (2.8) as finding $f_n \in \text{span}\,\{\phi(.,t_i)\}_{i=1,\dots,n}$ such that

$$\langle g, f - f_n \rangle_{H^s(\Omega)} = 0, \qquad \forall\, g \in \text{ span } \{\phi(.,t_i)\}_{i=1,\dots,n}, \tag{2.16}$$

i.e., f_n is the projection of f onto span $\{\phi(.,t_i)\}$. For $f \in L^2$, equation (2.16) makes sense if $g \in H^{2s}(\Omega)$, but the classical analysis of projection methods (cf. [10, 32] would need estimates of f^δ in the H^s-norm, which is not possible for perturbations in $L^2(\Omega)$.

We recall that the solution of linear equations with compact operators whose range is not finite dimensional, is an ill-posed problem (cf. [10, Proposition 2.7]). Of course, compactness and the resulting ill-posedness also depend on the spaces on which the operator is considered.

On the spaces $L^2(\Omega)$ or $C(\Omega)$ an integral operator may be non-compact if Ω is not bounded and / or the kernel k is strongly singular (cf. [22]), so that an equation of the first kind involving such an operator may be well-posed. For example, the inversion of the Fourier transform from $L^2(\mathbf{R}^p)$ to $L^2(\mathbf{C}^p)$ is well-posed. Another example is the solution of first kind equations for logarithmic single-layer potentials on arcs, which is well-posed between $C(\Gamma)$ and $C^1(\Gamma)$ and appropriate collocation methods converge (cf. [2, 22, 39]). Here, the well-posedness is a consequence of the choice of spaces.

In general, linear integral equations are well-posed if and only if the range of the operator is closed. For equations of the second kind, this follows from the Riesz-Schauder theory (cf. [22]). Closedness of the range can always be achieved by appropriate choice of spaces. Nevertheless, in our case we cannot choose the output space Y arbitrarily, because we must guarantee that

$$y^\delta := \int_\Omega k(x,.) f^\delta(x)\, dx \in Y \tag{2.17}$$

for all perturbed data $f^\delta \in L^2(\Omega)$. Furthermore we need k to be a smooth kernel ($k \in L^2(\Omega \times p)$ or $k \in C(P; L^2(\Omega))$ for (2.17) to make sense. Thus, if we want to solve the problem in $C(\Omega)$ or $H^s(\Omega)$ with positive s, we know that the range of $\mathcal{F}_k$ cannot be closed in Y if k is non-degenerate, and consequently (2.10) is ill-posed. The only possibility for the choice of k that would make (2.10) well-posed is the form of a degenerate kernel

$$k(x,t) = \sum_{j=1}^{m} a_j(x) b_j(t), \tag{2.18}$$

with fixed $m \in N$. However, the form (2.18) is not of interest for neural networks, since it neither matches the typical choices of ϕ nor allows the approximation of functions with arbitrary precision (since f_n is always in the finite-dimensional subspace spanned by the functions $\{(L_s^* L_s)^{-1} a_j\}_{j=1,\ldots,m}$).

3 Convergence Results

Convergence analysis of least-squares collocation is based on the reproducing kernel Q, which is defined by

$$Q(t,s) = \langle q_t, q_s \rangle_{H^s(\Omega)}, \tag{3.1}$$

where q_t is the unique element in $H^s(\Omega)$ such that

$$\langle q_t, g \rangle_{H^s(\Omega)} = \mathcal{F}_k g, \tag{3.2}$$

for all $g \in H^s(\Omega)$ (cf. [31, p.78]). The identity (2.4) immediately implies

$$q_t = \phi(.,t), \qquad \forall\, t \in P, \tag{3.3}$$

and using the definition of k we may write

$$\begin{aligned} Q(t,s) &= \langle \phi(.,t), \phi(.,s) \rangle_{H^s(\Omega)} \\ &= \langle \phi(.,t), k(.,s) \rangle_{L^2(\Omega)}. \end{aligned} \tag{3.4}$$

We first give a standard convergence result for least-squares collocation (see Theorem 3.1 in [31] and its extension from $L^2(\Omega)$ to the case of general Hilbert-Spaces [31, p.78])

Theorem 3.1. *Let $y \in \mathcal{R}(\mathcal{F}_k)$, then,*

1. *$f_n \to f^\dagger$ as $\Delta_n \to 0$, if $Q \in C(P \times P)$.*
2. *if $P \subset \mathbf{R}^1$, $f^\dagger \in \mathcal{R}(\mathcal{F}_k^*)$ and furthermore the derivatives $\frac{\partial^j}{\partial t^j} Q(t,s)$ exist and are continuous for $j = 1, \dots, 2m$, then*

$$\|f_n - f^\dagger\|_{H^s(\Omega)} \le \gamma \Delta_n^m \tag{3.5}$$

   for some constant $\gamma > 0$ as $\Delta_n \to 0$,

where $f^\dagger$ is the minimum-norm solution of (2.10) *and*

$$\Delta_n := \sup_{t \in D} \inf_{i=1,\dots,n} |t - t_i| \tag{3.6}$$

Some of the conditions of Theorem 3.1 are always satisfied in the above setup: first $y \in \mathcal{R}(\mathcal{F}_k)$ because of its definition (2.9). In addition, if $\phi \in C^{2s}(\Omega \times P)$, we obtain $k \in C(\Omega \times P)$ and thus $Q \in C(P \times P)$, which follows immediately from (3.4). Therefore, we can apply Theorem 3.1 easily to the special structure of the approximation problem (2.6). The results we are able to deduce are similar to existing results in literature (cf. e.g. [13, 15, 33]), but now with the advantage, that the parameters t_i need not to be optimized (i.e., trained), which is relevant for many applications. Below we will give a corollary which directly connects our results to the case of optimized parameters. Also, we give a connection to the theory of integral equations of the first kind.

Proposition 3.2. *Let $f \in H^s(\Omega)$ and f_n as above. Then $f_n \to f$ if $\Delta_n \to 0$ and if $f \in \mathcal{N}(\mathcal{F}_k)^\perp$, i.e.,*

$$\langle f, g \rangle = 0, \tag{3.7}$$

for all $g \in \mathcal{N}(\mathcal{F}_k)$. If $f \notin \mathcal{N}(\mathcal{F}_k)^\perp$, then $f_n \to \mathcal{P}f$, where $\mathcal{P}$ denotes the orthogonal projector onto $\mathcal{N}(\mathcal{F}_k)^\perp$

Proof. The convergence of f_n to $f^\dagger$ follows from Theorem 3.1. Because of the definition of y, $\mathcal{P}f$ satisfies

$$\mathcal{F}_k \mathcal{P} f = \mathcal{F}_k f = y,$$

hence it is a solution and obviously it is the one of minimal norm. Thus, if $f \in \mathcal{N}(\mathcal{F}_k)^\perp$, then

$$f_n \to f^\dagger = \mathcal{P} f = f,$$

otherwise f_n converges to $\mathcal{P}f \neq f$. □

A direct consequence is the convergence for those elements in the range of the adjoint operator $\mathcal{F}_k^*$:

Corollary 3.3. *Let $f \in H^s(\Omega)$ such that there exists a function $h \in L^2(P)$ with*

$$f(x) = \int_P \phi(x,t) h(t) \, dt, \qquad \forall\, x \in \Omega. \tag{3.8}$$

Then $f_n \to f$ if $\Delta_n \to 0$.

Proof. Because of $\mathcal{R}(\mathcal{F}_k^*) \subset \mathcal{N}(\mathcal{F}_k)^\perp$, we obtain convergence of f_n to f from Proposition 3.2. $\square$

Proposition 3.4. *Let $P \subset \mathbf{R}^1$ and $f \in H^s(\Omega)$ such that (3.8) holds for some $h \in L^2(P)$. Moreover, let $\phi \in C^{2s+2m}(\Omega \times P)$, $m \geq 0$. Then $f_n \to f$ as $\Delta_n \to 0$ and the estimate*

$$\|f_n - f\|_{H^s(\Omega)} \leq c_1 \Delta_n^{s+m}, \tag{3.9}$$

holds for some positive contant c_1.

Proof. The assertion follows directly from Theorem 3.1, since for $\phi \in C^{2s+2m}$ we have

$$\frac{\partial^j}{\partial t^j} Q(t,s) = \int_\Omega \frac{\partial^j}{\partial t^j} \phi(x,t) k(x,s)\ dx \in C(P \times P), \qquad j = 1, \ldots, 2s+2m,$$

and the condition $f \in \mathcal{R}(\mathcal{F}_k^*)$ is just (3.8). $\square$

Proposition 3.4 shows that the rate of convergence increases with the smoothness of the basis function ϕ. Vice versa, the set of functions on which the approximation scheme converges or on which this rate can be achieved, decreases with the smoothness of ϕ, since the condition (3.8) gets stronger. This result confirms a central point in the philosophy of network approximation, namely that the architecture should incorporate a-priori knowledge and qualitative ideas about the solution. The basis function should be chosen such that the expected solution satisfies the condition (3.8), which is an abstract smoothness condition.

If the parameters t_i are not fixed, but chosen in an optimal way, the approximation must be at least as good as for the special choice of uniformly distributed t_i (i.e., $\Delta_n = Cn^{-1}$). Together with Proposition 3.4 this immediately implies

Corollary 3.5 (Optimal Choice of all Parameters). *Let, in addition to the assumptions of Proposition 3.4, $\{(c_i, t_i)\}_{i=1,\ldots,n}$ be such that*

$$\|f - f_n\|_{H^s(\Omega)} = \min_{\{(c_i,t_i)\}} \|f - \sum_{i=1}^n c_i \phi(x, t_i)\|_{H^s(\Omega)}. \tag{3.10}$$

Then, the asymptotic estimate

$$\|f - f_n\|_{H^s(\Omega)} \quad = \quad \mathcal{O}\left(n^{-(m+s)}\right) \tag{3.11}$$

holds.

In general, we need $n \sim \ell^p$ points to cover a domain $P \subset \mathbf{R}^p$ such that $\Delta_n = \mathcal{O}(\frac{1}{\ell})$. Hence, we have

$$\Delta_n = \mathcal{O}\left(n^{-\frac{1}{p}}\right) \tag{3.12}$$

and an error estimate similar to the one of Corollary 3.4 would take the form

$$\|f_n - f\| = \mathcal{O}\left(n^{-\frac{m+s}{p}}\right), \tag{3.13}$$

if Proposition 3.4 also holds for $P \subset \mathbf{R}^d$. Using similar ideas as Nashed and Wahba [31] in the proof of Theorem (3.1) one can even show the following result [6]:

Theorem 3.6 (Optimal Choice of all Parameters in general Dimensions). *Let $P \subset \mathbf{R}^p$ be a rectangle, $\phi \in C^{m+s}(\Omega \times P)$ and $f \in H^s(\Omega)$ be a function such that* (3.8) *holds for some $h \in L^2(P)$. If $\{(c_i, t_i)\}_{i=1,\dots,n}$ are such that*

$$\|f - f_n\|_{H^s(\Omega)} = \min_{\{(c_i,t_i)\}} \|f - \sum_{i=1}^{n} c_i \phi(x, t_i)\|_{H^s(\Omega)}, \tag{3.14}$$

the asymptotic estimate

$$\|f - f_n\|_{H^s(\Omega)} = \mathcal{O}\left(n^{-\frac{m}{p}}\right). \tag{3.15}$$

holds.

We note that our approximation result is very similar to the approximation results in [13] and [33], where under the condition (3.8) with $h \in L^1$ a convergence rate of $\frac{1}{\sqrt{n}}$ has been proven for optimal choice of the parameters t_i:

Example 3.7. Now we apply Corollary 3.6 to neural networks with a single layer, i.e., the approximation scheme (1.1), (1.5). In this case we have $P = S^1(\mathbf{R}^d) \times (b_-, b_+)$, where $S^1(\mathbf{R}^d)$ denotes the unit sphere in $\mathbf{R}^d$ and (b_-, b_+) is some finite interval. In order to obtain a rectangle $\overline{P} = [-\pi, \pi)^{d-1} \times (b_-, b_+)$ we may use a trigonometric parametrisation of the unit sphere (which does not change the continuity properties). Hence, the dimension of the set of parameters is $p = d$, which yields the estimate

$$\|f - f_n\|_{H^s(\Omega)} \quad = \quad \mathcal{O}\left(n^{-\frac{s}{d}}\right)$$

if $\sigma \in C^s(\mathbf{R})$ and f satisfies (3.8). In the setup of [13], [33], i.e., with $2s > d$, which guarantees the continuous embedding into $C(\overline{\Omega})$ this yields

$$\|f - f_n\|_{H^s(\Omega)} \quad = \quad o\left(\frac{1}{\sqrt{n}}\right), \tag{3.16}$$

i.e., the rate $\mathcal{O}(\frac{1}{\sqrt{n}})$ obtained in [33] is even improved.

Example 3.8. As a simple example we consider the approximation with gaussian kernels, i.e.,

$$\phi(x,t) = e^{-\frac{|x-t|^2}{2\sigma}}, \tag{3.17}$$

and $P = \Omega$. Since ϕ is an analytic function, a convergence rate faster than any power of n is obtained from Proposition 3.4, but only for those functions which can be represented as a convolution with a gaussian kernel. However, all band-limited functions satisfy this condition (cf. [13]), which is of interest for many approximation problems.

We consider the approximation of the function

$$f(x) := x + \sin 4\pi x \tag{3.18}$$

in the interval $\Omega = (0,1)$ with $n = 11, 21, 31, 41$ knots. An obvious choice for the norm is $s = 1$, since $H^1(\Omega)$ can be embedded continuously into $C(\Omega)$ for dimension 1. The operator L_1 is just the first derivative and the corresponding elliptic operator $L_1^* L_1$ is given by

$$\begin{array}{rccc} L_1^* L_1 : & H^1(\Omega) & \to & H^{-1}(\Omega) \\ & g & \mapsto & -g'' \end{array}.$$

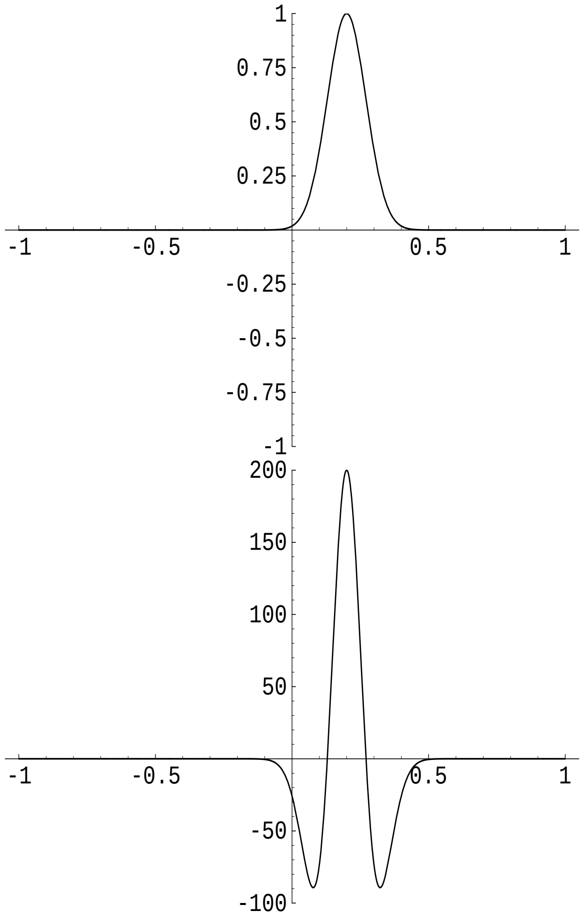

Figure 1: The gaussian kernel and its second derivative, $t = 0.2$, $\sigma = 0.005$.

n	Δ_n	$\|f - f_n\|_\infty$	$\|f - f_n\|_{H^1}$
11	0.1	0.5981	0.3828
21	0.05	0.3699	0.1945
31	0.033	0.0549	0.0289
41	0.025	$2.15\ 10^{-3}$	$6.87\ 10^{-4}$

Table 1: Numerical results with a translation network on a uniform grid using gaussian basis-functions.

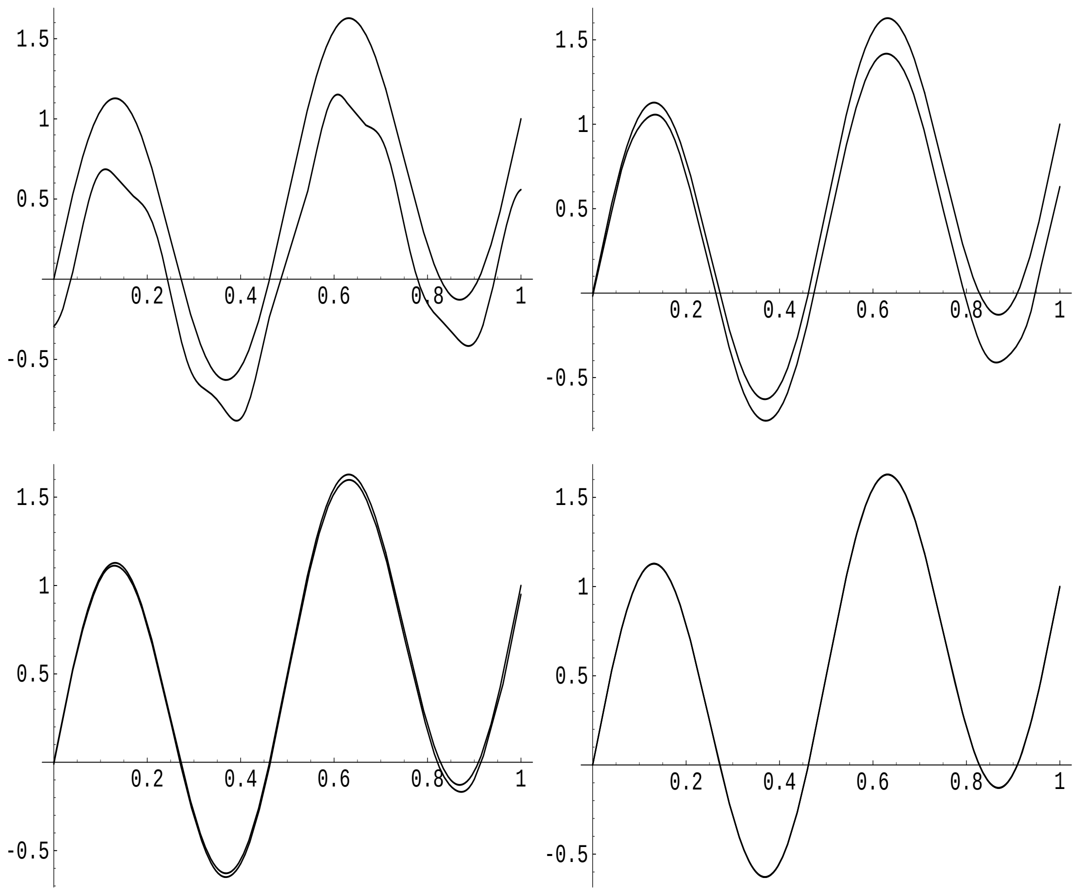

Figure 2: Exact and approximated data for $n = 11, 21, 31, 41$, $\sigma = 0.002$.

Figure 1 shows the gaussian kernel (ϕ) and its second derivative ($\phi'' = -k$) for small σ in one spatial dimension. One observes that the kernel is concentrated around t, which illustrates that it approximates the Dirac delta distribution well. Hence, following the remarks at the end of Section 2, an approximation scheme of the form (1.1), (3.17) seems to be a good choice if it can be assumed that the function to be approximated is sufficiently smooth, which is true for our choice of f.

The convergence as $n \to \infty$ is illustrated in Figure 2, where the results for $n = 11, 21, 31, 41$ are given. Table 1 shows the corresponding errors in the H^1-norm and in the supremum-norm, which decrease fast as expected from (3.9).

4 Stability

We now turn to the problem of stability with respect to data errors.

Proposition 4.1. *Let Φ_n be as in* (2.12) *and let σ_n denote its minimal nonzero eigenvalue. Then*

$$\|f_n - f_n^\delta\| \leq c_2 \frac{\sqrt{n}\delta}{\sqrt{\sigma_n}}. \tag{4.1}$$

Proof. We may estimate

$$\begin{aligned} \|f_n - f_n^\delta\|^2 &= \sum_{i,j=1}^{n} (c_i - c_i^\delta)(c_j - c_j^\delta)\langle \phi(.,t_i), \phi(.,t_j)\rangle_{H^k(\Omega)} \\ &= (c - c^\delta)^T \Phi_n (c - c^\delta) \\ &= (y - y^\delta)^T \Phi_n^\dagger (y - y^\delta) \\ &\leq \frac{\|y - y^\delta\|^2}{\sigma_n}. \end{aligned}$$

The discrete data error can be estimated by

$$\begin{aligned} \|y - y^\delta\|^2 &= \sum_{j=1}^{n} |y_j - y_j^\delta|^2 \\ &\leq \sum_{j=1}^{n} \|k(.,t_j)\|_{L^2(\Omega)}^2 \|f - f^\delta\|_{L^2(\Omega)}^2 \\ &= n\delta^2 c_2^2. \end{aligned}$$

Combining these estimates we obtain (4.1) □

The stability estimate (4.1) is sharp in the worst-case sense, since one can always construct a perturbation f^δ such that $y - y^\delta$ lies in the direction of the minimal eigenvector of Φ_n and $\|y - y^\delta\|$ is of order δ. This shows again the ill-posedness of the approximation problem, because for $n \to \infty$ this produces arbitrarily high errors.

Now we are able to give a general convergence result for f_n^δ:

Theorem 4.2. *Let $\phi \in C^{2s}(\Omega \times P)$ and $f \in \mathcal{N}(\mathcal{F}_k)^\perp$. If $n = n(\delta)$ is chosen such that $\frac{n\delta^2}{\sigma_n} \to 0$ and if $\Delta_{n(\delta)} \to 0$, then*

$$f_{n(\delta)}^\delta \to f \quad \text{in } H^s(\Omega). \tag{4.2}$$

Proof. The result is a direct consequence of Propositions 3.2 and 4.1 and the triangle inequality,

$$\|f - f_n^\delta\|_{H^s(\Omega)} \quad \le \quad \|f - f_n\|_{H^s(\Omega)} + \|f_n - f_n^\delta\|_{H^s(\Omega)}.$$

□

In a similar way we deduce the following result about convergence rates:

Theorem 4.3. *Let $\phi \in C^{2s+2m}(\Omega \times P)$ and let f satisfy (3.8), then the estimate*

$$\|f_n - f_n^\delta\|_{H^s(\Omega)} \le c_1 \Delta_n^{s+m} + c_2 \frac{\sqrt{n}\delta}{\sqrt{\sigma_n}} \tag{4.3}$$

holds.

Theorem 4.2 shows that the number of knots in the network must depend upon the data error to obtain convergence. Together with the remark after Proposition 4.1 it follows that there exists a sequence of perturbations f^{δ_k} such that

$$\|f_{n(\delta_k)}^{\delta_k} - f\| \to \infty \qquad \text{if } \frac{n\delta^2}{\sigma_n} \to \infty.$$

From Theorem 4.3 we can deduce a method for the choice of n. If we fix δ in (4.3), then the first term converges to zero as $n \to \infty$, while $\frac{n}{\sigma_n} \to \infty$. Hence, a heuristic for choosing the number of knots in the network dependent upon δ is to balance both terms, i.e.,

$$\Delta_n^{2s+2m} \frac{\sigma_n}{n} \sim \delta^2. \tag{4.4}$$

Based on continuous embeddings, we can now use our results in sobolev spaces to derive an analogous result in the space of continuous functions:

Corollary 4.4. *Let $2s > d$, then under the conditions of Theorem 4.2,*

$$f_{n(\delta)}^\delta \to f \text{ in } C(\overline{\Omega}), \tag{4.5}$$

and under the conditions of Theorem 4.3,

$$\|f_n - f_n^\delta\|_{C(\overline{\Omega})} \le \widetilde{c}_1 \Delta_n^{s+m} + \widetilde{c}_2 \frac{\sqrt{n}\delta}{\sqrt{\sigma_n}} \tag{4.6}$$

hold.

Proof. If $2s > d$, the Sobolev space $H^s(\Omega)$ may be embedded continuously into $C(\overline{\Omega})$, which implies convergence and rates of the supremum norm, too. □

Example 4.5. To illustrate the arising instabilities, we revisit Example 3.8, but now we use a perturbed observation f^δ with noise level $\delta = 5\%$. Figure 3 shows the development of the error $\|f - f_n\|$ with increasing n (and decreasing δ_n, respectively) for random noise. One observes that the errors in the H^1-norm and supremum norm first decrease until the reach an optimal value of n and then increase again, i.e., Proposition 4.1 is confirmed well numerically.

Finally, Table 2 gives the minimal and maximal eigenvalue of Φ_n (denoted by σ_n and λ_n). While the maximal eigenvalues increase linearly, the minimal eigenvalues tend to zero extremly fast, the matrix Φ_n is ill-conditioned already for small values of n.

n	λ_n	σ_n
11	34.14	0.4288
21	44.83	0.2976
31	67.21	$3.22\ 10^{-3}$
41	89.57	$1.05\ 10^{-8}$
51	111.94	$9.16\ 10^{-11}$.

Table 2: Maximal and minimal eigenvalues of Φ_n.

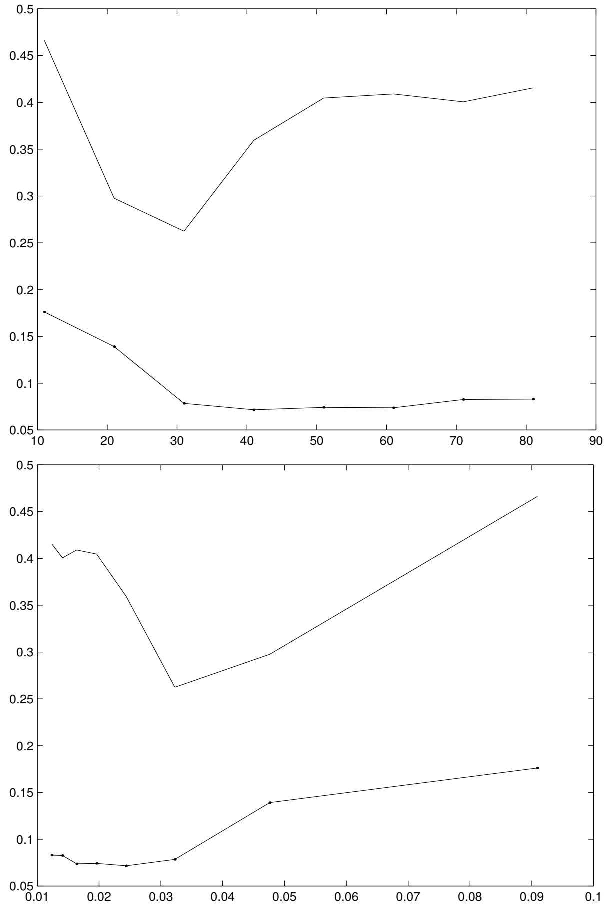

Figure 3: Error $\|f - f_n\|$ vs. n (upper) and vs. Δ_n (lower). The solid line with dots shows the error in the H^1-norm, the second line shows the error in the supremum-norm.

5 Extensions

So far we have only treated the case when the complete function f^δ is known. In all practical applications only scattered data $\{f^\delta(x_j)\}_{j=1,\ldots,N}$ are available, with an error estimate of the form

$$\sum_{j=1}^{N} w_j(f(x_j) - f^\delta(x_j))^2 \leq \delta^2, \tag{5.1}$$

for appropriate weights w_j. If one cannot assume that interpolation with an L^2-error δ is possible, a third type of error will appear, the so-called *generalization error*, which arises because of the lack of information. A special case, namely the approximation in the H^1-norm for $d = 1$, which is just the well-known problem of numerical differentiation, has been investigated by Hanke and Scherzer [19] in this setup. The expected value of the generalization error has been estimated in a recent publication by Niyogi and Girosi (cf. [33]). It should be an important task for the future to investigate the stability properties of network approximation in this more general case. The basic idea for this generalization is the approximation of the integrals needed for the operator $\mathcal{F}_k$ by quadrature at the sampling points $\{x_j\}$.

Another possible extension is to investigate different regularization methods than just collocation in the integral equation (2.10). In literature, many applications of Tikhonov regularization to networks can be found, but so far only with the aim of reducing the generalization error. Some effects of iterative regularization (under the key word *overtraining*) have been studied in that context, too. Thus, it is a natural next step to utilize the machinery of linear and nonlinear regularization theory (cf. [10]) for the stability analysis of these methods combined with network approximation.

Acknowledgements

This work has been supported by the Austrian Fonds zur Förderung der wissenschaftlichen Forschung, project F 1308.

References

[1] U.Amato, D.T.Vuza, *Besov regularization, thresholding and wavelets for smoothing data*, Numer. Funct. Anal. Optimiz. **18**(1997), 461-493.

[2] K.E.Atkinson, I.H.Sloan, *The numerical solution of first-kind logarithmic-kernel integral equations on smooth open arcs*, Math. Comp. **56** (1991), 119-139.

[3] A.R.Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Trans. Inf. Theory **39** (1993), 930-945.

[4] C.M.Bishop, *Neural Networks for Pattern Recognition* (Clarendon Press, Oxford, 1995).

[5] C.M.Bishop, *Regularization and complexity control in feed-forward networks*, EC2 & Cie (1995), 141-148.

[6] M.Burger, A.Neubauer, *Error bounds for approximation with neural networks*, submitted.

[7] C.K.Chui, X.Li, *Approximation by ridge functions and neural networks with one hidden layer*, J. Approx. Theory **70** (1992), 131-141.

[8] I.Daubechies, *Ten Lectures on Wavelets* (SIAM, Philadelphia, 1992).

[9] H.W.Engl, *Regularization by least-squares collocation*, in P.Deuflhard, E.Hairer (eds.), *Numerical Treatment of Inverse Problems in Differential and Integral Equations* (Birkhäuser, Boston, Basel, Stuttgart, 1983), 345-354.

[10] H.W.Engl, M.Hanke, A.Neubauer, *Regularization of Inverse Problems* (Kluwer, Dordrecht, 1996).

[11] H.G.Feichtinger, T.Strohmer, eds., *Gabor Analysis and Applications* (Birkhäuser, Boston, Basel, Berlin, 1998).

[12] W.Freeden, F.Schneider, *Regularization wavelets and multiresolution*, Inverse Problems **14**(1998), 225-243.

[13] F.Girosi, G.Anzellotti, *Convergence rates of approximation by translates*, AI Memo 1288 (AI Laboratory, MIT, Cambridge, Massachusetts, 1995).

[14] F.Girosi, T.Poggio, *A theory of networks for approximation and learning*, AI Memo 1140 (AI Laboratory, MIT, Cambridge, Massachusetts, 1989).

[15] F.Girosi, T.Poggio, *Networks and the best approximation property*, Biol.Cybern. **63** (1990), 169-176.

[16] F.Girosi, M.Jones, T.Poggio, *Regularization theory and neural networks architectures*, Neural Computation **7** (1995), 219-269.

[17] C.W. Groetsch, *Generalized inverses and generalized splines*, Numer. Funct. Anal. Optimiz. **2** (1980), 93-97.

[18] W.Hackbusch, *Integral Equations* (Birkhäuser, Basel, 1995).

[19] M.Hanke, O.Scherzer, *Numerical differentiation as an example for inverse problems*, Preprint 98/16 (University Karlsruhe, 1998).

[20] M.Hegland, R.S.Anderssen, *A mollification framework for improperly posed problems*, Numer. Math. **78** (1998), 549-575.

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

[22] R.Kress, *Linear Integral Equations* (2nd Edition, Springer, New York, 1999).

[23] A.K.Louis, *Inverse und schlecht gestellte Probleme* (Teubner, Stuttgart, 1989).

[24] A.K.Louis, P.Maaß, *Smoothed projection methods for the moment problem*, Numer. Math. **59** (1991), 277-294.

[25] A.K.Louis, *Approximate inverse for linear and some nonlinear problems*, Inverse Problems **12** (1996), 175-190.

[26] Y.Makovoz, *Uniform approximation by neural networks*, J. Approx. Theory **95** (1998), 215-228.

[27] Y.Meyer, *Wavelets. Algorithms and Applications* (SIAM, Philadelphia, 1993).

[28] D.A.Murio, *The Mollification Method and the Numerical Solution of Ill-posed Problems* (Wiley, New York, 1993).

[29] H.N.Mhaskar, C.A.Micchelli, *Degree of approximation by neural and translation networks with a single hidden layer*, Adv. Appl. Math. **16** (1995), 151-183.

[30] I.T.Nabney, *Efficient training of RBF networks for classification*, NCRG Tech Report 99/02 (1999).

[31] M.Z.Nashed, G.Wahba, *Convergence rates of approximate least squares solutions to linear integral and operator equations of the first kind*, Math. Comp. **28** (1974), 69-80.

[32] F.Natterer, *Regularisierung schlecht gestellter Probleme durch Projektionsverfahren*, Numer. Math. **28** (1977), 329-341.

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

[34] A.Sard, Approximations based on nonscalar observations, J. Approx. Theory **8** (1973), 315-334.

[35] J.Sjöberg, Q.Zhang, L.Ljung, A.Benveniste, B.Deylon, P.Y.Glorennec, H.Hjalmarsson, A.Juditsky, *Nonlinear black-box modeling in system identification: A unified overview*, Automatica **31** (1995), 1691-1724.

[36] A.N.Tikhonov, V.Y.Arsenin, *Solutions of Inverse Problems* (Wiley, New York, 1977).

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

[38] G.Wahba, *A class of approximate solutions to linear operator equations*, J. Approx. Theory, **9** (1973), 61-77.

[39] Y.Xu, Y.Zhao, *Quadratures for boundary integral equations of the first kind with logarithmic kernels*, Journal of Integral Equations and Applications **8** (1996), 239-268.

[40] P.Yee, S.Haykin, *Pattern classification as an ill-posed, inverse problem: a regularization approach*, Proc. ICASSP-93 (Minneapolis, 1993).