Analysis of Tikhonov Regularization for Function Approximation by Neural Networks

Martin Burger*and Andreas Neubauer

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

Abstract. This paper is devoted to the convergence and stability analysis of Tikhonov regularization for function approximation by a class of feed-forward neural networks with one hidden layer and linear output layer. We investigate two frequently used approaches, namely *regularization by output smoothing* and *regularization by weight decay*, as well as a combination of both methods to combine their advantages. We show that in all cases stable approximations are obtained converging to the approximated function in a desired Sobolev space as the noise in the data tends to zero (in the weaker L^2-norm) if the regularization parameter and the number of units in the network are chosen appropriately. Under additional smoothness assumptions we are able to show convergence rates results in terms of the noise level and the number of units in the network.

In addition, we show how the theoretical results can be applied to the important classes of *perceptrons* with one hidden layer and to *translation networks*. Finally, the performance of the different approaches is compared in some numerical examples.

Key Words: Ill-posed problems, neural networks, Tikhonov regularization, output smoothing, weight decay, function approximation.

AMS Subject Classifications: 65J20, 92B20, 41A30.

1. Introduction

In this paper we deal with the problem of approximating a function $f \in H^m(\Omega)$ for which only noisy measurements $f^\delta \in L^2(\Omega)$ with an error bound

$$\|f - f^\delta\|_{L^2(\Omega)} \leq \delta \tag{1.1}$$

are known. The class of approximating functions under consideration are *feed-forward neural networks with one hidden layer and linear output layer*, i.e., the set of functions of the form

$$X_n := \Big\{ \sum_{i=1}^{n} c_i \phi(x; t_i) \ : \ c_i \in \mathbb{R}, t_i \in P \subset \mathbb{R}^p \Big\}, \tag{1.2}$$

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

where P is a compact subset of $\mathbb{R}^p$ and ϕ is a given activation function. The above network architecture is frequently used for approximation problems because of its good approximation properties (cf. e.g. [5, 13, 15, 19]), especially in the case of Ridge-constructions (cf. e.g. [1, 7, 14]) where ϕ is of the form

$$\phi(x; a, b) = \sigma(a^T x + b), \qquad a \in A \subset \mathbb{R}^d,\ b \in B \subset \mathbb{R}. \tag{1.3}$$

Hornik et al. [13] showed that the union of the sets X_n defined in (1.2) with ϕ given by (1.3) are dense in $C(\Omega)$ ($\Omega \subset \mathbb{R}^d$), if σ is a continuous function of sigmoidal form, i.e., σ is monotone and

$$\lim_{s \to -\infty} \sigma(s) = 0\,, \qquad \lim_{s \to +\infty} \sigma(s) = 1\,.$$

In subsequent papers, the approximation capabilities of several network constructions with linear output layers have been investigated (cf. e.g. [1, 15, 19] and the references therein). A result of particular interest is the dimension-independent convergence rate

$$\inf_{f_n \in X_n} \|f - f_n\|_{L^2(\Omega)} = \mathcal{O}(n^{-\frac{1}{2}})\,, \tag{1.4}$$

which can be achieved under the additional conditions (cf. [19])

$$\sup_{t \in P} \|\phi(*; t)\|_{L^2(\Omega)} < \infty \qquad \text{and} \qquad f = \int_P h(t)\phi(*; t)\, dt \quad \text{for some } h \in L^1(P)\,.$$

Under stronger conditions on ϕ, this rate result can be even improved (cf. [5]).

It is well-known that the approximation problem in $H^m(\Omega)$ is asymptotically ill-posed if the observation error is bounded in the weaker L^2-norm (cf. e.g. [4]), i.e., an arbitrarily small data error may lead to arbitrarily high errors in the solution as n tends to infinity. A reasonably small choice of n would be an inherent regularization, but it is a difficult task to find such a parameter choice $n = n(\delta, f^\delta)$ that yields convergence as the data error δ decreases to zero (cf. [4]). Therefore, in practice other regularization methods are used that allow larger values of n, namely either iterative techniques (often called *early stopping*, cf. e.g. [2, 20]) or Tikhonov-type methods (cf. e.g. [2, 3, 11, 12, 16, 17]). In this paper we will concentrate on the latter, for which we prove stability and develop a convergence analysis as the noise level δ tends to zero.

Tikhonov regularization of an operator equation of the form

$$F(x) = y\,, \tag{1.5}$$

with noisy data y^δ, means to replace (1.5) by the minimization problem

$$\min_{x \in X} \|F(x) - y^\delta\|_Y^2 + \alpha \|x - x_*\|_X^2\,, \tag{1.6}$$

where X and Y are function spaces such that F maps $\mathcal{D}(F) \subset X$ to Y and x_* is an initial guess for a solution of (1.5) (see [8] for a general overview of linear and nonlinear Tikhonov regularization). As for any regularization method, the minimizers of (1.6) converge to a solution of (1.5) only if the regularization parameter α is chosen appropriately in dependence on the noise level δ and possibly the noisy data y^δ. Therefore, the mathematical theory is important also for practical computations, since it yields rules for the optimal choice of the regularization parameter. Such a convergence analysis does not yet exist for the regularized training of neural networks.

There are mainly two different approaches to regularization with Tikhonov-type stabilizers in this area, namely *regularization by output smoothing* (cf. e.g. [2, 3, 11, 12]) and *regularization by weight decay* (cf. e.g. [2, 16, 17]).

Regularization by output smoothing means solving the minimization problem

$$\min_{f_n \in X_n} \|f_n - f^\delta\|^2_{L^2(\Omega)} + \alpha \|f_n - f_*\|^2_{H^m(\Omega)} \,. \tag{1.7}$$

Since X_n is usually not weakly closed, this problem might not have a solution. In the next section, we show existence, stability and convergence in $H^m(\Omega)$ of solutions for a slight modification of (1.7) (see (2.1) below). Under additional smoothness assumptions on f we can even guarantee convergence rates.

The minimization of a functional like (1.7) is also a common method for standard classes of approximating functions like splines (cf. e.g. [21]) and seems to be a good choice if one is interested in the output f_n only, but not in the behaviour of the parameters c_i and t_i. More emphasis on these parameters is put in the so-called *regularization by weight decay* where the following minimization problem with respect to the parameters $\{(c_i, t_i)\}$ is solved:

$$\min_{\{(c_i,t_i)\} \in (\mathbb{R}\times P)^n} \|\sum_{i=1}^n c_i \phi(*; t_i) - f^\delta\|^2_{L^2(\Omega)} + \beta n \sum_{i=1}^n c_i^2 \tag{1.8}$$

Sometimes an additional penalty term for the parameters $\{(t_i)\}$ is used, e.g. $\tilde{\beta} \sum_{i=1}^n t_i^2$. However, since the parameters $\{(t_i)\}$ are restricted to a compact set P in our considerations, this term is not necessary and will therefore be omitted. In Section 3 we shall prove existence, stability and weak convergence of solutions of (1.8) in the Sobolev space $H^m(\Omega)$. Strong convergence and convergence rates can be derived in weaker norms, i.e., in $H^s(\Omega)$ with $s < m$.

As a consequence of the analysis in Sections 2 and 3 we combine both methods, output smoothing and weigt decay, in Section 4, and investigate the properties of the resulting method. In Section 5, the theoretical results are applied to *perceptrons* with one hidden layer and to *translation networks*. Finally, a comparison of the methods and numerical results will be presented in Section 6.

For our analysis we need the following three basic assumptions:

(A1) The set of parameters $P \subset \mathbb{R}^p$ is compact.

(A2) The activation function ϕ is in the space $C(P; H^m(\Omega))$.

(A3) The function $f \in H^m(\Omega)$, $m \in \mathbb{N}$, to be approximated by functions in X_n, satisfies the representation

$$f = \int_P h(t)\phi(*; t)\, dt \quad \text{for some } h \in L^1(P)\,.$$

Assumption (A2) guarantees that the evaluation of $\phi(*; t)$ (and its derivatives with respect to x up to order m) is well-defined and that $X_n \subset H^m(\Omega)$. Assumption (A3) is a smoothness condition for f. For special cases of ϕ, e.g. for some perceptrons, (A3) is implied by a certain rate of decay of the Fourier transform $\hat{f}$ (cf. [5]). Sometimes we will need slightly stronger versions of (A2) and (A3) namely:

(A2') The activation function ϕ satisfies:

$$\|\phi(*;t) - \phi(*;s)\|_{H^m(\Omega)} \leq c|t-s|^\rho\,, \quad \rho \in (0,1]\,, c \in \mathbb{R}^+\,.$$

(A3') $f \in H^m(\Omega)$ satisfies (A3) where h is even in $L^2(P)$.

In addition to these assumption the following approximation result will be fundamental for our convergence analysis:

Theorem 1.1. *Let X_n be defined by* (1.2) *and let assumptions* (A1) – (A3) *are fulfilled. Then there exists an element*

$$f_n = \sum_{i=1}^n c_i^n \phi(*;t_i^n) \in X_n \qquad \textit{with} \quad \sum_{i=1}^n (c_i^n)^2 \leq \gamma n^{-1}\,, \tag{1.9}$$

for some $\gamma > 0$, such that

$$\|f - f_n\|_{H^m(\Omega)} = \mathcal{O}(n^{-\frac{1}{2}})\,. \tag{1.10}$$

If in addition the stronger assumptions (A2') *and* (A3') *are fulfilled, then an element f_n as in* (1.9) *exists with*

$$\|f - f_n\|_{H^m(\Omega)} = \mathcal{O}(n^{-\frac{1}{2}-\frac{\rho}{p}})\,. \tag{1.11}$$

Proof. The first rate was shown in [19] for the case $m = 0$, i.e., in $L^2(\Omega)$. It follows from the proof there that the result is valid for $m > 0$, too, if (A1) – (A3) are satisfied. The second rate result was shown in [5, Theorem 2.1] under a slightly stronger assumption, namely that $h \in L^\infty(P)$. In the proof we used this assumption to show that

$$\sum_{i=1}^n (c_i^n)^2 = \mathcal{O}(n^{-1}) \tag{1.12}$$

noting that $c_i^n = \mathcal{O}(n^{-1})$ for $h \in L^\infty(P)$, where c_i^n is defined by

$$c_i^n := \int_{P_i} h(t)\,dt$$

and the sets P_i are such that

$$P = \bigcup_{i=1}^n P_i\,, \qquad P_i \cap P_j = \{\}\,, i \neq j\,, \qquad |P_i| = \mathcal{O}(n^{-1})\,.$$

Using the Cauchy-Schwarz inequality we show that (1.12) even holds under the weaker assumption $h \in L^2(P)$:

$$\begin{aligned}\sum_{i=1}^n (c_i^n)^2 &= \sum_{i=1}^n \Big(\int_{P_i} h(t)\,dt\Big)^2 \leq \sum_{i=1}^n \int_{P_i} 1\,dt \int_{P_i} h^2(t)\,dt \\ &= \sum_{i=1}^n |P_i| \int_{P_i} h^2(t)\,dt = \mathcal{O}(n^{-1}) \sum_{i=1}^n \int_{P_i} h^2(t)\,dt = \mathcal{O}(n^{-1})\,\|h\|^2_{L^2(P)}\end{aligned}$$

∎

2. Regularization by output smoothing

In this section we investigate stability and convergence properties of regularization by output smoothing. As mentioned in the introduction, the minimization problem (1.7) might not have a solution since the set X_n in (1.2) is, in general, not weakly closed. Therefore, we consider the following modified problem: for $\alpha > 0$ and $\mu > 0$ we look for a solution $f_{n,\alpha}^{\delta,\mu}$ satisfying

$$\begin{aligned} \|f_{n,\alpha}^{\delta,\mu} - f^\delta\|^2_{L^2(\Omega)} &+ \alpha\|f_{n,\alpha}^{\delta,\mu} - f_*\|^2_{H^m(\Omega)} \\ &\leq \|f_n - f^\delta\|^2_{L^2(\Omega)} + \alpha\|f_n - f_*\|^2_{H^m(\Omega)} + \mu \qquad \text{for all } f_n \in X_n\,. \end{aligned} \tag{2.1}$$

This problem reflects the fact that a minimizer of (1.7), even if it exists, can not be calculated exactly in practice. It is obvious that problem (2.1) always has several solutions that are stable in the following sense:

Proposition 2.1. *Let $\{f^k\}$ be a sequence converging towards f^δ in $L^2(\Omega)$ and let $f_{n,\alpha}^{k,\mu}$ be solutions of* (2.1) *with f^δ replaced by f_k. Then for every $\varepsilon > 0$, $f_{n,\alpha}^{k,\mu}$ is also a solution of* (2.1) *with μ replaced by $\mu + \varepsilon$ if k is sufficiently large, i.e.,*
$\forall \varepsilon > 0\, \exists \overline{k} \in \mathbb{N}\, \forall k \geq \overline{k}$:

$$\begin{aligned} \|f_{n,\alpha}^{k,\mu} - f^\delta\|^2_{L^2(\Omega)} &+ \alpha\|f_{n,\alpha}^{k,\mu} - f_*\|^2_{H^m(\Omega)} \\ &\leq \|f_n - f^\delta\|^2_{L^2(\Omega)} + \alpha\|f_n - f_*\|^2_{H^m(\Omega)} + \mu + \varepsilon \qquad \textit{for all } f_n \in X_n\,. \end{aligned}$$

The convergence analysis follows the lines of [8, Theorem 10.3].

Theorem 2.2. *Let $f^\delta \in L^2(\Omega)$ satisfy* (1.1) *and let assumptions* (A1) – (A3) *be fulfilled. Moreover, let $\alpha = \alpha(n, \delta, \mu)$ be chosen such that*

$$\alpha \to 0\,, \qquad \delta^2\alpha^{-1} \to 0\,, \qquad \mu\alpha^{-1} \to 0\,, \qquad \textit{and} \qquad n\alpha \to \infty\,,$$

as $n \to \infty$, $\delta \to 0$, and $\mu \to 0$. Then $f_{n,\alpha}^{\delta,\mu} \to f$ in $H^m(\Omega)$.

If in addition the stronger assumptions (A2') *and* (A3') *are fulfilled, then the condition $n\alpha \to \infty$ may be weakened to*

$$n^{1+\frac{2\rho}{p}}\alpha \to \infty\,.$$

Proof. Let f_n be an approximating function as in (1.9). Then, (1.1), (1.10), (2.1), and

$$\|g\|_{L^2(\Omega)} \leq \|g\|_{H^m(\Omega)} \quad \text{for all } g \in H^m(\Omega)\,, \tag{2.2}$$

yield the estimate

$$\begin{aligned} \|f_{n,\alpha}^{\delta,\mu} - f^\delta\|^2_{L^2(\Omega)} + \alpha\|f_{n,\alpha}^{\delta,\mu} - f_*\|^2_{H^m(\Omega)} &\leq \|f_n - f^\delta\|^2_{L^2(\Omega)} + \alpha\|f_n - f_*\|^2_{H^m(\Omega)} + \mu \\ &\leq \mathcal{O}(n^{-1} + \delta^2) + \mu \\ &\qquad + \alpha(\mathcal{O}(n^{-\frac{1}{2}}) + \|f - f_*\|_{H^m(\Omega)})^2\,. \end{aligned}$$

Hence,

$$f_{n,\alpha}^{\delta,\mu} \to f \text{ in } L^2(\Omega) \quad \text{as } n \to \infty\,,\ \delta \to 0\,,\ \mu \to 0\,,$$

and

$$\limsup_{n,\delta,\mu} \|f_{n,\alpha}^{\delta,\mu} - f^*\|_{H^m(\Omega)} \leq \|f - f_*\|_{H^m(\Omega)} \,.$$

Since $H^m(\Omega)$ is compactly embedded in $L^2(\Omega)$, strong convergence of $f_{n,\alpha}^{\delta,\mu}$ towards f in $H^m(\Omega)$ can now be shown with the same technique as in [8, Theorem 10.3].

The proof for the weaker condition $n^{1+\frac{2\rho}{p}}\alpha \to \infty$ under the assumptions (A2') and (A3') follows similarly using (1.11) instead of (1.10). ∎

It is obvious from the proof above that depending on the choice of α one always gets rates for $(f_{n,\alpha}^{\delta,\mu} - f)$ in the norm of $L^2(\Omega)$. However, as usual for ill-posed problems, the convergence might be arbitrarily slow in $H^m(\Omega)$ and rates can be proven only under additional smoothness assumptions on f. Usually, such smoothness assumptions are source conditions of the form

$$f - f_* \in \mathcal{R}((E^*E)^\nu)\,, \quad \text{for some } 0 < \nu \leq \tfrac{1}{2}\,, \tag{2.3}$$

where, in our case, E denotes the embedding operator from $H^m(\Omega)$ into $L^2(\Omega)$.

Theorem 2.3. *Let f satisfy* (2.3) *and f^δ be such that* (1.1) *holds. Moreover, let assumptions* (A1) – (A3) *be fulfilled. If*

$$\mu = \mathcal{O}(n^{-1} + \delta^2) \qquad \text{and} \qquad \alpha \sim (n^{-\frac{1}{2}} + \delta)^{\frac{2}{1+2\nu}}\,,$$

then

$$\|f_{n,\alpha}^{\delta,\mu} - f\|_{H^s(\Omega)} = \mathcal{O}((n^{-\frac{1}{2}} + \delta)^{1-\frac{s}{m(1+2\nu)}}) \quad \text{for any } 0 \leq s \leq m\,. \tag{2.4}$$

If in addition the stronger assumptions (A2') *and* (A3') *are fulfilled and if*

$$\mu = \mathcal{O}(n^{-1-\frac{2\rho}{p}} + \delta^2) \qquad \text{and} \qquad \alpha \sim (n^{-\frac{1}{2}-\frac{\rho}{p}} + \delta)^{\frac{2}{1+2\nu}}\,,$$

then

$$\|f_{n,\alpha}^{\delta,\mu} - f\|_{H^s(\Omega)} = \mathcal{O}((n^{-\frac{1}{2}-\frac{\rho}{p}} + \delta)^{1-\frac{s}{m(1+2\nu)}}) \quad \text{for any } 0 \leq s \leq m\,. \tag{2.5}$$

Proof. Let f_n be as in (1.9) and $f - f_* = (E^*E)^\nu w$. Then it follows with (1.1), (1.10), (2.1), (2.2), and $\mu = \mathcal{O}(n^{-1} + \delta^2)$ that

$$\begin{aligned}
&\|f_{n,\alpha}^{\delta,\mu} - f^\delta\|_{L^2(\Omega)}^2 + \alpha\|f_{n,\alpha}^{\delta,\mu} - f\|_{H^m(\Omega)}^2 \\
&\quad\leq \|f_n - f^\delta\|_{L^2(\Omega)}^2 + \mu \\
&\qquad + \alpha\Big(\|f_n - f\|_{H^m(\Omega)}^2 + 2\langle f_n - f, f - f_*\rangle_{H^m(\Omega)} + 2\langle f_{n,\alpha}^{\delta,\mu} - f, f_* - f\rangle_{H^m(\Omega)}\Big) \\
&\quad= \mathcal{O}(n^{-1} + \delta^2 + \alpha n^{-\frac{1}{2}} + \alpha\|(E^*E)^\nu(f_{n,\alpha}^{\delta,\mu} - f)\|_{H^m(\Omega)})\,.
\end{aligned}$$

Together with the interpolation inequality (cf. (2.49) in [8]) and the fact that

$$\|(E^*E)^{\frac{1}{2}}g\|_{H^m(\Omega)} = \|Eg\|_{L^2(\Omega)} = \|g\|_{L^2(\Omega)} \qquad \text{for all } g \in H^m(\Omega)\,,$$

we now obtain the estimate

$$\begin{aligned}
&\|f_{n,\alpha}^{\delta,\mu} - f\|_{L^2(\Omega)}^2 + \alpha\|f_{n,\alpha}^{\delta,\mu} - f\|_{H^m(\Omega)}^2 \\
&\quad= \mathcal{O}(n^{-1} + \delta^2 + \alpha n^{-\frac{1}{2}} + \delta\|f_{n,\alpha}^{\delta,\mu} - f\|_{L^2(\Omega)} + \alpha\|f_{n,\alpha}^{\delta,\mu} - f\|_{L^2(\Omega)}^{2\nu}\|f_{n,\alpha}^{\delta,\mu} - f\|_{H^m(\Omega)}^{1-2\nu})\,,
\end{aligned}$$

which immediately implies

$$\begin{aligned}
&\max\{\|f_{n,\alpha}^{\delta,\mu}-f\|_{L^2(\Omega)}^2, \alpha\|f_{n,\alpha}^{\delta,\mu}-f\|_{H^m(\Omega)}^2\}\\
&\quad = \mathcal{O}\Big(n^{-1}+\delta^2+\alpha n^{-\frac{1}{2}}+(\delta+\alpha^{\frac{1}{2}+\nu})\max\{\|f_{n,\alpha}^{\delta,\mu}-f\|_{L^2(\Omega)}, \alpha^{\frac{1}{2}}\|f_{n,\alpha}^{\delta,\mu}-f\|_{H^m(\Omega)}\}\Big)\,.
\end{aligned}$$

Hence, the order estimate

$$\max\{\|f_{n,\alpha}^{\delta,\mu}-f\|_{L^2(\Omega)}, \alpha^{\frac{1}{2}}\|f_{n,\alpha}^{\delta,\mu}-f\|_{H^m(\Omega)}\} = \mathcal{O}(n^{-\frac{1}{2}}+\delta+\alpha^{\frac{1}{2}}n^{-\frac{1}{4}}+\alpha^{\frac{1}{2}+\nu})$$

holds, and together with $\alpha\sim(n^{-\frac{1}{2}}+\delta)^{\frac{2}{1+2\nu}}$ we now obtain

$$\begin{aligned}
\|f_{n,\alpha}^{\delta,\mu}-f\|_{L^2(\Omega)} &= \mathcal{O}(n^{-\frac{1}{2}}+\delta)\,,\\
\|f_{n,\alpha}^{\delta,\mu}-f\|_{H^m(\Omega)} &= \mathcal{O}((n^{-\frac{1}{2}}+\delta)^{\frac{2\nu}{1+2\nu}})\,.
\end{aligned}$$

Finally, (2.4) follows with the interpolation inequality.

The proof of the rates under the assumptions (A2') and (A3') follows similarly using (1.11) instead of (1.10). ∎

Remark 2.4. The convergence rate in (2.4) suggests to choose $n\sim\delta^{-2}$. If n grows faster than δ^{-2}, we do not gain anything in the rate, but make the dimension of problem (2.1) larger than necessary.

The assumption (2.3) is a smoothness condition, which means that, in addition to condition (A3), $f-f_*$ has to be an element of $H^{(1+2\nu)m}(\Omega)$ satisfying some boundary conditions. E.g., for the case $m=1$ and $\nu=\frac{1}{2}$, we have (cf. [18])

$$\mathcal{R}((E^*E)^{\frac{1}{2}}) = \mathcal{R}(E^*) = \{z\in H^2(\Omega)\,:\,\tfrac{\partial z}{\partial\nu}=0 \text{ on } \partial\Omega\}\,, \tag{2.6}$$

where $\frac{\partial z}{\partial\nu}$ denotes the normal derivative at the boundary.

3. Regularization by weight decay

In this section we consider the nonlinear minimization problem (1.8) in the parameter set $(\mathbb{R}\times P)^n$. Since the nonlinear operator $F:(\mathbb{R}\times P)^n\to H^m(\Omega)$ defined by

$$F((c_i,t_i)_{i=1}^n) := \sum_{i=1}^n c_i\phi(*;t_i)$$

is obviously continuous and weakly sequentially closed, the existence and stability of regularized solutions

$$f_{n,\beta}^\delta := \sum_{i=1}^n c_{i,\beta}^\delta\phi(*;t_{i,\beta}^\delta) \tag{3.1}$$

follow from [8, Theorem 10.2].

In the next theorem we prove weak convergence of the regularized solutions in $H^m(\Omega)$ as well as strong convergence and convergence rates in weaker norms.

Theorem 3.1. *Let $f^\delta\in L^2(\Omega)$ satisfy* (1.1) *and let assumptions* (A1) – (A3) *be fulfilled. Moreover, let $\beta=\beta(n,\delta)$ be such that*

$$\beta\to 0\,,\qquad \delta^2\beta^{-1}\le\gamma_1\,,\qquad \textit{and}\qquad n\beta\ge\gamma_2$$

for some positive constants γ_1, γ_2, as $n \to \infty$ and $\delta \to 0$. Then $f^\delta_{n,\beta} \rightharpoonup f$ in $H^m(\Omega)$.

If $\beta \sim (n^{-1} + \delta^2)$, then we even obtain convergence rates in weaker norms given by

$$\|f^\delta_{n,\beta} - f\|_{H^s(\Omega)} = \mathcal{O}((n^{-\frac{1}{2}} + \delta)^{1-\frac{s}{m}}) \quad \textit{for any } 0 \le s < m\,.$$

If in addition the stronger assumptions (A2') *and* (A3') *are fulfilled, then condition $n\beta \ge c_2$ may be weakened to*

$$n^{1+\frac{2\rho}{p}}\beta \ge c_2\,.$$

Moreover, for the choice $\beta \sim (n^{-1-\frac{2\rho}{p}} + \delta^2)$ we obtain the rates

$$\|f^\delta_{n,\beta} - f\|_{H^s(\Omega)} = \mathcal{O}((n^{-\frac{1}{2}-\frac{\rho}{p}} + \delta)^{1-\frac{s}{m}}) \quad \textit{for any } 0 \le s < m\,.$$

Proof. Let f_n be an approximating function as in (1.9). Then, with (1.1), (1.8), (1.10), (2.2), and (3.1), we conclude that

$$\begin{aligned}\|f^\delta_{n,\beta} - f^\delta\|^2_{L^2(\Omega)} + \beta n \sum_{i=1}^n (c^\delta_{i,\beta})^2 &\le \|f_n - f^\delta\|^2_{L^2(\Omega)} + \beta n \sum_{i=1}^n (c^n_i)^2 \\ &= \mathcal{O}(n^{-1} + \delta^2 + \beta) \end{aligned} \tag{3.2}$$

Hence,

$$f^\delta_{n,\beta} \to f \text{ in } L^2(\Omega) \quad \text{as } n \to \infty\,,\ \delta \to 0\,,$$

and

$$\limsup_{n,\delta}\, n \sum_{i=1}^n (c^\delta_{i,\beta})^2 = \mathcal{O}(1)\,.$$

Therefore,

$$\begin{aligned}\limsup_{n,\delta} \|f^\delta_{n,\beta}\|^2_{H^m(\Omega)} &= \limsup_{n,\delta} \sum_{i,j=1}^n c^\delta_{i,\beta} c^\delta_{j,\beta} \langle\, \phi(*; t^\delta_{i,\beta}), \phi(*;, t^\delta_{j,\beta})\,\rangle_{H^m(\Omega)} \\ &= \mathcal{O}\Big(\limsup_{n,\delta} n \sum_{i=1}^n (c^\delta_{i,\beta})^2\Big) = \mathcal{O}(1)\,.\end{aligned}$$

Note that $\sup_{t\in P} \|\phi(*;t)\|_{H^m(\Omega)} < \infty$, due to (A2). Now the compactness of the embedding of $H^m(\Omega)$ into $L^2(\Omega)$ implies that $f^\delta_{n,\beta} \rightharpoonup f$ in $H^m(\Omega)$. Moreover, it follows from the estimate (3.2) and the interpolation inequality that for the choice $\beta \sim (n^{-1} + \delta^2)$ we obtain the asserted convergence rates.

The proof for the weaker condition $n^{1+\frac{2\rho}{p}}\beta \to \infty$ under the assumptions (A2') and (A3') follows similarly using (1.11) instead of (1.10). ∎

Remark 3.2. From the proof of Theorem 3.1 one observes that for the second part condition (A2') could be even weakened to: ϕ has to be in the space $C(P; H^m(\Omega))$ and

$$\|\phi(*;t) - \phi(*;s)\|_{L^2(\Omega)} \le c|t-s|^\rho\,, \quad \rho \in (0,1]\,, c \in \mathbb{R}^+\,,$$

i.e., Hölder continuity of ϕ is only needed in $L^2(\Omega)$ and not in $H^m(\Omega)$, since the convergence rate result (1.11) is only needed in $L^2(\Omega)$.

4. A combination of output smoothing and weight decay

In this section we want to combine both approaches, output smoothing and weight decay, i.e., we look for a minimizer

$$f_{n,\alpha,\beta}^{\delta} := \sum_{i=1}^{n} c_{i,\alpha,\beta}^{\delta}\phi(*; t_{i,\alpha,\beta}^{\delta}) \tag{4.1}$$

of the problem

$$\min_{\{(c_i,t_i)\}\in(\mathbb{R}\times P)^n} \|\sum_{i=1}^{n} c_i\phi(*;t_i) - f^{\delta}\|_{L^2(\Omega)}^2 + \alpha\|\sum_{i=1}^{n} c_i\phi(*;t_i) - f_*\|_{H^m(\Omega)}^2 + \beta n \sum_{i=1}^{n} c_i^2 \,. \tag{4.2}$$

It follows in an analogous way to Section 3 that for all $\alpha, \beta > 0$ solutions $f_{n,\alpha,\beta}^{\delta}$ exist and are stable with respect to data noise, which is due to the second penalty term. As we will see below the first penalty term will guarantee convergence and convergence rates as in Section 2.

Theorem 4.1. *Let $f^{\delta} \in L^2(\Omega)$ satisfy* (1.1) *and let assumptions* (A1) – (A3) *be fulfilled. Moreover, let $\alpha = \alpha(n,\delta)$ and $\beta = \beta(n,\delta)$ be such that*

$$\alpha \to 0\,, \qquad \delta^2\alpha^{-1} \to 0\,, \qquad n\alpha \to \infty\,, \qquad \beta \to 0\,, \qquad \text{and} \qquad \beta\alpha^{-1} \to 0$$

as $n \to \infty$ and $\delta \to 0$. Then $f_{n,\alpha,\beta}^{\delta} \to f$ in $H^m(\Omega)$.

If in addition the stronger assumptions (A2') *and* (A3') *are fulfilled, then condition $n\alpha \to \infty$ may be weakened to*

$$n^{1+\frac{2\rho}{p}}\alpha \to \infty\,.$$

Proof. Let f_n be as in (1.9), then it holds for the corresponding weights $\{c_i^n\}$ that

$$n\sum_{i=1}^{n}(c_i^n)^2 = \mathcal{O}(1)\,. \tag{4.3}$$

Now we obtain similar to the proof of Theorem 2.2 the estimate

$$\begin{aligned} &\|f_{n,\alpha,\beta}^{\delta} - f^{\delta}\|_{L^2(\Omega)}^2 + \alpha\|f_{n,\alpha,\beta}^{\delta} - f_*\|_{H^m(\Omega)}^2 + \beta n\sum_{i=1}^{n}(c_{i,\alpha,\beta}^{\delta})^2 \\ &\quad\leq\ \mathcal{O}(n^{-1} + \delta^2 + \beta) + \alpha\Big(\mathcal{O}(n^{-\frac{1}{2}}) + \|f - f_*\|_{H^m(\Omega)}\Big)^2\,. \end{aligned}$$

The proof now follows as in Theorem 2.2 noting that β now plays the role of μ. ∎

Theorem 4.2. *Let f satisfy* (2.3) *and f^{δ} be such that* (1.1) *holds. Moreover, let assumptions* (A1) – (A3) *be fulfilled. If*

$$\alpha \sim (n^{-\frac{1}{2}} + \delta)^{\frac{2}{1+2\nu}} \qquad \text{and} \qquad \beta = \mathcal{O}(n^{-1} + \delta^2)\,,$$

then

$$\|f_{n,\alpha,\beta}^{\delta} - f\|_{H^s(\Omega)} = \mathcal{O}((n^{-\frac{1}{2}} + \delta)^{1-\frac{s}{m(1+2\nu)}}) \quad \text{for any } 0 \leq s \leq m\,.$$

If in addition the stronger assumptions (A2') *and* (A3') *are fulfilled and if*

$$\alpha \sim (n^{-\frac{1}{2}-\frac{\rho}{p}} + \delta)^{\frac{2}{1+2\nu}} \qquad \text{and} \qquad \beta = \mathcal{O}(n^{-1-\frac{2\rho}{p}} + \delta^2)\,,$$

then

$$\|f^{\delta}_{n,\alpha,\beta} - f\|_{H^s(\Omega)} = \mathcal{O}((n^{-\frac{1}{2}-\frac{\rho}{p}} + \delta)^{1-\frac{s}{m(1+2\nu)}}) \quad \text{for any } 0 \le s \le m\,.$$

Proof. The proof is similar to the one of Theorem 2.3. Note that, as in the proof of Theorem 4.1, (4.3) holds and β plays the role of μ. ∎

5. Applications

In this section we apply the above results to some typical constructions for neural networks. The two classes we consider are perceptrons with one hidden layer and translation networks, whose use for approximation problems (cf. [10]) and deconvolution (cf. [6]) has been investigated recently.

Perceptrons are a classical construction for neural networks. They consist of an input layer of ridge-type and an activation function σ as in (1.3). The activation function σ is usually chosen as a Heaviside function or a smoothed version like, e.g.,

$$\sigma(t) = \frac{1}{1+e^{-t}}\,.$$

A multi-layer perceptron with one hidden layer and linear output layer is then of the form

$$f_n(x) = \sum_{i=1}^{n} c_i \sigma(a_i^T x + b_i)\,. \tag{5.1}$$

The assumptions (A1) – (A3) on ϕ and f can be easily interpreted in this case if Ω is a bounded domain. A canonical choice for the set of parameters $P = A \times B$ is $A := [-\overline{a}, \overline{a}]^d$ and $B := [-\overline{b}, \overline{b}]$, where $\overline{b}$ has to be chosen sufficiently large with respect to $\overline{a}$ (cf. [5]). This choice of P also allows a simple numerical implementation of the training process.

If σ is such that

$$\sigma(t) := \begin{cases} 1\,, & t > 1\,, \\ p_k(t)\,, & -1 \le t \le 1\,, \\ 0\,, & t < -1\,, \end{cases} \tag{5.2}$$

where p_k is the unique polynomial of degree $2k+1$, $k \in \mathbb{N}_0$, satisfying

$$p_k(-1) = 0\,,\ p_k(1) = 1\,,\ \text{and } p_k^{(l)}(-1) = 0 = p_k^{(l)}(1)\,,\ 1 \le l \le k\,,$$

then $\sigma \in C^{k,1}$ and $\sigma \in W^{k+1,\infty}$. We will prove in the next lemma that ϕ satisfies (A2') for $m \le k+1$.

Lemma 5.1. *Let Ω be a bounded domain and let σ be as in* (5.2)*. Then it holds that the activation function $\phi(x;a,b) := \sigma(a^T x + b)$ satisfies* (A2') *with $\rho = 1$ for $m \le k$ and $\rho = \frac{1}{2}$ for $m = k+1$.*

Proof. By definition of σ and ϕ, $\phi(*;a,b)$ is obviously in $H^m(\Omega)$ for $m \le k+1$ and we obtain

$$\|\phi(*;a,b)-\phi(*;\overline{a},\overline{b})\|^2_{H^m(\Omega)} = \sum_{|\kappa|\le m}\int_\Omega \left(\sigma^{(|\kappa|)}(a^Tx+b)a^\kappa - \sigma^{(|\kappa|)}(\overline{a}^Tx+\overline{b})\overline{a}^\kappa\right)^2 dx\,, \quad (5.3)$$

where $\kappa = (\kappa_1,\ldots,\kappa_d)$ is a multiindex and $a^\kappa := a_1^{\kappa_1}\cdots a_d^{\kappa_d}$.

If $m \le k$, then $\sigma^{(|\kappa|)}$ is Lipschitz continuous. Hence, (A2') obviously holds with $\rho = 1$.

Let us now consider the case $m = k+1$: Noting that $\sigma^{(k+1)}$ is a polynomial of degree k in $[-1,1]$ and 0 outside, we obtain together with (5.3) that

$$\|\phi(*;a,b)-\phi(*;\overline{a},\overline{b})\|^2_{H^m(\Omega)} \le \gamma_1(|a-\overline{a}|^2+|b-\overline{b}|^2)+\gamma_2\eta|a|^2\,, \quad (5.4)$$

where γ_1,γ_2 are positive constants and

$$\eta := \operatorname{meas}\{x\in\Omega\,:\,(|a^Tx+b|\le 1 \wedge |\overline{a}^T+\overline{b}|>1)\vee(|a^Tx+b|>1\wedge|\overline{a}^T+\overline{b}|\le 1)\}\,. \quad (5.5)$$

Let us now estimate $\eta|a|^2$. First we consider the case where $a^T\overline{a} \le \frac{1}{4}|a|^2$: Since we then have that

$$|a-\overline{a}|^2 = |a|^2+|\overline{a}|^2-2a^T\overline{a} \ge \frac{1}{2}|a|^2+|\overline{a}|^2$$

and since $\eta \le |\Omega|$, we obtain the estimate

$$\eta|a|^2 \le 2|\Omega||a-\overline{a}|^2\,. \quad (5.6)$$

Let us now consider the case where $a^T\overline{a} > \frac{1}{4}|a|^2$: note that then $a \ne 0$ and $a^T\overline{a} \ne 0$. Moreover, it is obvious that η is bounded by a constant times the maximal distance between the hyperplanes $a^Tx+b=1$ and $\overline{a}^Tx+\overline{b}=1$ (as well as with 1 replaced by -1) with $x\in\Omega$. Let x be such that $a^Tx+b=1$ and $\overline{x}$ be such that $\overline{a}^T\overline{x}+\overline{b}=1$ and $\overline{x} = x+\lambda a$. Then

$$|x-\overline{x}| = \frac{|(a-\overline{a})^Tx+(b-\overline{b})||a|}{|a^T\overline{a}|} < 4\frac{|a-\overline{a})||x|+|b-\overline{b}|}{|a|}$$

and hence

$$\eta|a|^2 \le \gamma_3(|a-\overline{a})|+|b-\overline{b}|) \quad (5.7)$$

for some positive constant γ_3. Now (5.4) – (5.7) imply that (A2') holds with $\rho = \frac{1}{2}$. ∎

Assumption (A3) may be interpreted as a smoothness condition upon f (cf. [5]), which becomes more and more restrictive with increasing smoothness of ϕ. Therefore, it is advantageous to choose σ not much smoother than needed to satisfy assumption (A2), i.e., $m = k+1$ seems to be an optimal choice.

For the special case $m=1$, $k=0$ a sufficient condition for (A3) to hold is that the Fourier transform $\hat{f}$ of f is such that (cf. [5, Proposition 3.4, Remark 3.5])

$$(1+|\cdot|^2)\hat{f}(\cdot)\in L^1(\mathbb{R}^d)\,,$$

while $f\in W^{2,1}(\Omega)$ is obviously a necessary condition. Slightly stronger conditions are necessary for (A3') to be satisfied.

Another popular construction are *translation networks*, which are of the form

$$f_n(x) = \sum_{i=1}^n c_i\psi(x-t_i)\,,$$

i.e., they fit into the form (1.2) with

$$\phi(x;t) = \psi(x-t)\,.$$

The obvious choice for the parameters in this case is $P = \overline{\Omega}$ for a bounded domain Ω.

Translation networks cover the important class of radial basis functions, which are also used in many other applications such as density estimation. A particular example are so-called *regularization networks*, where the activation function ψ is the fundamental solution of a symmetric elliptic differential operator D of order $2k$ with constant coefficients, such that

$$N_D(g) := -\int_{\mathbb{R}^d} (Dg)(x)g(x)\,dx$$

is an equivalent norm on $H^k(\mathbb{R}^d)$. For such networks, assumption (A2) is satisfied at least if $k > m + \frac{d}{2}$, since all derivatives up to order m are continuous in this case, which can be seen from a standard embedding theorem (cf. [9, p. 270]). A sufficient condition for (A3) to be fulfilled is $f \in H^{2k}(\Omega)$, since then

$$f(x) = \langle\, \delta(x-\cdot), f\,\rangle = \langle\, -D\psi(x-\cdot), f\,\rangle = \langle\, \psi(x-\cdot), -Df\,\rangle = \int_\Omega \psi(x-t)h(t)\,dt$$

with $h = -Df \in L^2(\Omega) \subset L^1(\Omega)$.

6. Numerical results

In order to test our theoretical results in numerical examples, we consider the approximation of functions in $H^1(\Omega)$ with a multilayer perceptron of the form (5.1) in two examples: in the first example $\Omega = [0,1]$, i.e., the spatial dimension $d=1$, and in the second example $\Omega = [0,1]^2$, i.e., the spatial dimension $d=2$.

The noise in our examples is an artificial high-frequency perturbation added to the exact data. The resulting noisy data are then sampled on a uniform grid $\mathcal{G}$ with step size $h = 10^{-2}$. The integral of a function over Ω is numerically approximated by a trapezoidal rule for each cell in the grid $\mathcal{G}$. Hence, the discretized optimization problems arising from (1.7), (1.8), and (4.2), respectively, are of the form

$$\min_{(\{a_j,b_j,c_j\})\in(A\times B\times\mathbb{R})^n} \sum_{x\in\mathcal{G}} w(x)\Big(f^\delta(x) - \sum_{j=1}^n c_j\sigma(a_j^T x + b_j)\Big)^2 + S_{\mathcal{G}}(\{a_j,b_j,c_j\})\,, \tag{6.1}$$

where $w(x)$ denotes the sum of the quadrature weights at point x and $S_{\mathcal{G}}$ denotes the discretization of the stabilizing term over the grid $\mathcal{G}$. In the case of weight decay (cf. (1.8)), the stabilizing term is independent of the grid and therefore we have

$$S_{\mathcal{G}}(\{a_j,b_j,c_j\}) = \beta n \sum_{j=1}^n c_j^2\,, \tag{6.2}$$

while we need again quadrature rules to discretize the stabilizer in the case of output smoothing in $H^1(\Omega)$ (cf. (1.7) and (2.1)), which yields (with $f^* = 0$)

$$S_{\mathcal{G}}(\{a_j,b_j,c_j\}) = \alpha \sum_{x\in\mathcal{G}} w(x)\Big(\Big(\sum_{j=1}^n c_j\sigma(a_j^T x + b_j)\Big)^2 + \Big|\sum_{j=1}^n c_j a_j\sigma'(a_j^T x + b_j)\Big|^2\Big)\,. \tag{6.3}$$

In the combined approach as presented in Section 4, the stabilizing term $S_{\mathcal{G}}$ is just the sum of the ones in (6.2) and (6.3).

All numerical tests that are presented in the following were performed with the software package MATLAB 6 on an SGI Origin 3800. In both cases, we used routines for constrained minimization from the MATLAB Optimization Toolbox to solve the discretized optimization problem (6.1).

Example 6.1. Our first example is the approximation of

$$f(x) = 1 - \sin(1.8\pi x)\,, \qquad x \in \Omega := [0,1]\,,$$

with a multilayer perceptron of the form (5.1) and activation function

$$\sigma(t) = \frac{1}{1 + e^{-100t}}\,.$$

The set of admissible parameters $(a_j, b_j) \in \mathbb{R}^2$ is given by $P = A \times B = [-1,1]^2$.

In order to investigate the convergence behavior as $\delta \to 0$, we choose a sequence $\delta_k = \delta_0 2^{-k}$ for $k = 1, \dots, 6$ and $\delta_0 = 0.32$. The regularization parameters α, β and the number of units n are chosen according to Theorem 2.2 and Theorem 3.1, respectively, with $\rho = 1$ and $p = 2$, i.e.,

$$\alpha(\delta) = \alpha_0 \delta\,, \qquad \beta(\delta) = \beta_0 \delta^2\,, \qquad n = c\delta^{-1}\,.$$

We want to mention that in our test case, the parameters α_0 and β_0 are tuned such that an optimal approximation with respect to the H^1-norm is achieved. In the combined approach, we choose α as for output smoothing and the parameters β are chosen as $\beta = o(\alpha n^{-\frac{1}{2}})$, i.e., the conditions of Theorem 4.1 are fulfilled.

A general observation in the numerical minimization of (6.1) is that the optimization algorithm must be stopped earlier for output smoothing than for weight decay or the combination of both. We think that this could be due to the possible non-existence of a minimizer of (1.7) and corresponds to the relaxed problem (2.1). Another observable effect is that the iteration numbers in the numerical minimization are usually smaller for weight decay than for the other two methods. In addition, the numerical effort for the evaluation of $S_{\mathcal{G}}$ is obviously much lower for weight decay than for output smoothing.

In Figure 6.1, the sensitivity of the resulting error between the regularized solutions and the exact function f in the H^1-norm is illustrated. The left picture shows the error $\|f^{\delta,\mu}_{n,\alpha} - f\|_{H^1(\Omega)}$ plotted vs. $\log\alpha$ and the right one shows $\|f^{\delta}_{n,\alpha,\beta} - f\|_{H^1(\Omega)}$ plotted vs. $\log\beta$ for fixed number of units $n = 16$ and noise level $\delta = 2\%$. Note that for output smoothing, the minimization procedure did not yield reasonable results for $\alpha \le 10^{-10}$, but was trapped in a stationary point; for weight decay, this effect occured later at around $\beta = 10^{-12}$. As one would expect, both methods do not yield reasonable results if the regularization parameters are extremely small or extremely large, but one observes that the error does not change dramatically in a relatively large scale between 10^{-4} and 10^{-8}, i.e., both methods seem to be robust with respect to over- or underestimating the regularization parameter.

In Figure 6.2, the error between the regularized solutions obtained with output smoothing, weight decay and the combination of both, are plotted vs. the noise level δ. The left picture shows the error in the L^2-norm, which is almost the same for the three methods for most choices of δ; the only visible difference occurs in a region between

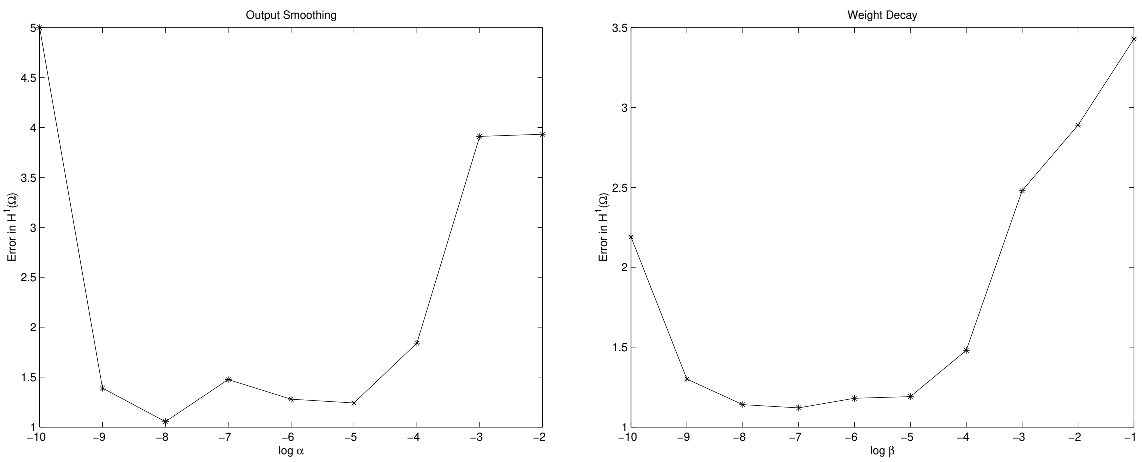

Figure 6.1: Error in the H^1-norm plotted vs. $\log \alpha$ for output smoothing (left) and vs. $\log \beta$ for weight decay (right).

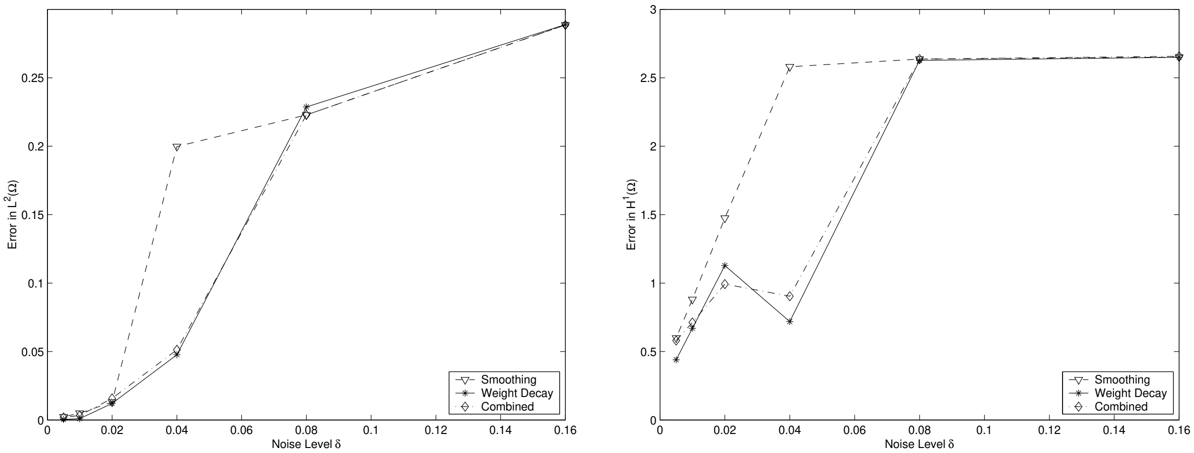

Figure 6.2: Error in the L^2-norm (left) and in the H^1-norm (right) plotted vs. the noise level δ in %.

2% and 8%, where output smoothing yields a significantly larger error. Note that the problem is not ill-posed in the L^2-norm, so also nonregularized solutions are expected to converge to the exact function in this norm, while this is not necessarily true in the H^1-norm. The errors for the latter are plotted vs. the noise level in the right picture, which shows that output smoothing yields the worst results in this case, except for large noise level ($\delta \geq 8\%$), where all three methods behave very similar. However, the numerical results indicate that convergence is obtained with any of the regularization methods, if the regularization parameter and the number of units are chosen appropriately.

Example 6.2. Our second numerical example is the approximation of the function

$$f(x_1, x_2) = \Big(\frac{x_1^2}{2} - \frac{x_1^3}{3}\Big)\Big(\frac{x_2^2}{2} - \frac{x_2^3}{3}\Big), \qquad (x_1, x_2) \in \Omega := [0,1]^2,$$

with a multi-layer perceptron of the form (5.1) with activation function given by (5.2) with $k = 1$. With this choice it was shown in Lemma 5.1 that (A2') is satisfied with $\rho = 1$. One can show that f satisfies (A3') and (2.3) with $\nu = \frac{1}{2}$, which can be seen

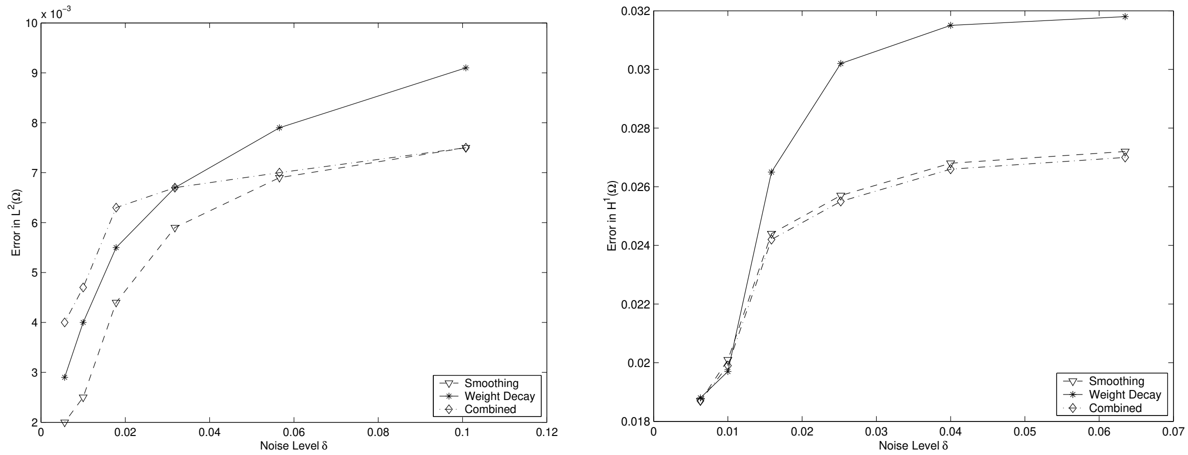

Figure 6.3: Error in the L^2-norm (left) and in the H^1-norm (right) plotted vs. the noise level δ in %.

from (2.6), and f and f_n are in $H^2(\Omega)$. Consequently, we choose the number of units and regularization parameters according to Theorem 2.3 with $\nu = \frac{1}{2}$, $s = m = 1$, $p = 3$, and Theorem 3.1 with $m = 2$, $s = 1$, and $p = 3$, i.e.,

$$n \sim \delta^{-\frac{6}{5}}\,, \qquad \alpha \sim \delta\,, \qquad \beta \sim \delta^2\,.$$

With this choice we may expect the rates

$$\|f_n - f\|_{H^1(\Omega)} = \mathcal{O}(\sqrt{\delta}) = \mathcal{O}(n^{-\frac{5}{3}}) \tag{6.4}$$

for the regularized solutions f_n obtained with output smoothing, weight decay or the combined approach.

From Lemma 5.1 we see that even the choice $k = 0$ in (5.2) would have been possible. However, $k = 1$ guarantees that the objective function in (6.1) is differentiable with respect to a_j, b_j, c_j, which is a desireable property for the minimization algorithm. Moreover, we obtain convergence for weight decay in $H^1(\Omega)$ whereas for $k = 0$ merely weak convergence can be guaranteed.

In the numerical minimization of (6.1) we obtain similar results with respect to the number of iterates and the numerical effort as in the one-dimensional Example 6.1. In fact, the difference in the numerical effort between weight decay and methods involving a stabilizer in the H^1-norm is even larger in two spatial dimensions due to the increase in the number of grid points.

Figure 6.3 shows the error between the regularized solutions and the exact function f plotted vs. the noise level δ. Opposed to Example 6.1, output smoothing and the combined method yield now a smaller error in the H^1-norm than weight decay, and the combined method is now the one with smallest error. The combined approach is now rather close to output smoothing, while it was much closer to weight decay in Example 6.1. This confirms quite well our intuition that the combined method might also combine the advantages of both methods and therefore yield an optimal performance.

Finally, we numerically investigate the rate of convergence, which is predicted by (6.4). Note that, if (6.4) holds, one should obtain that

$$q_k := \frac{\log(\|f_{n_{k+1}} - f\|_{H^1(\Omega)}) - \log(\|f_{n_k} - f\|_{H^1(\Omega)})}{\log \delta_{k+1} - \log \delta_k} \to \frac{1}{2} \tag{6.5}$$

for a sequence $\delta_k \to 0$. The values of q_k for our choices of δ are shown in Table 6.1. One observes that the values of q_k gradually increase to 0.5 for all three methods, which numerically confirms (6.5).

	Smoothing	Weight Decay	Combined
q_1	-0.0660	0.0065	0.0131
q_2	0.0599	-0.0257	0.0332
q_3	0.1662	0.0986	0.1397
q_4	0.4495	0.4187	0.3273

Table 6.1: Numerical estimate of the convergence rate according to (6.5).

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] ——, *Training with noise is equivalent to tihkhonov regularization*, Neural Computation 7 (1995), 108–116.

[4] M. Burger and H. W. Engl, *Training neural networks with noisy data as an ill-posed problem*, Adv. Comp. Math. (2001), to appear.

[5] M. Burger and A. Neubauer, *Error bounds for approximation with neural networks*, SFB-Report 00-17, University of Linz, 2000, submitted.

[6] M. Burger and O. Scherzer, *Regularization methods for blind deconvolution and blind source separation problems*, Math. Cont. Signals & Systems (2001), to apppear.

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

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

[9] L. C. Evans, *Partial Differential Equations*, Vol. 19 of AMS Graduate Studies in Mathematics, AMS, Providence, Rhode Island, 1998.

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

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

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

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

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

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

[16] J. E. Moody, *Note on generalization, regularization, and architecture selection in nonlinear learning systems*, in: Proceedings of the First IEEE-SP Workshop on Neural Networks for Signal Processing, IEEE Computer Society Press, Los Alamitos, 1991, 1–10.

[17] ——, *The effective number of parameters: an analysis of generalization and regularization in nonlinear learning systems*, in: J. E. Moody, S. J. Hansen, and R. P. Lippmann, eds., Advances in Neural Information Processing Systems 4, Morgan Kaufmann, Palo Alto, 1992, 847–854.

[18] A. Neubauer, *When do Sobolev spaces form a Hilbert scale?*, Proc. Amer. Math. Soc. 103 (1988), 557–562.

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

[20] J. Sjöberg and L. Ljung, *Overtraining, regularization and searching for a minimum, with application to neural networks*, Int. J. Control 62 (1995), 1391–1407.

[21] G. Wahba, *Spline Models for Observational Data*, SIAM, Philadelphia, 1990.