Nonlinear Function Approximation: Computing Smooth Solutions with an Adaptive Greedy Algorithm.[*]

Andreas Hofinger[†]

Abstract

Opposed to linear schemes, nonlinear function approximation allows to obtain a dimension independent rate of convergence. Unfortunately, in the presence of data noise typical algorithms (like e. g., backpropagation) are inherently unstable, whereas greedy algorithms, which are in principle stable, can not be implemented in their original form, since they require unavailable information about the data.

In this work we present a modified greedy algorithm, which does not need this information, but rather recovers it iteratively from the given data. We show that the generated approximations are always at least as smooth as the original function and that the algorithm also remains stable, when it is applied to noisy data. Finally the applicability of this algorithm is demonstrated by numerical experiments.

Keywords: Greedy Algorithm, Nonlinear Function Approximation, Data Noise, Regularization Theory
AMS Subject Classification: 41A46, 41A65, 93C41

1 Introduction

In many *black-box models* the goal is to approximate a function f using a simple representation f_k of the form

$$f_k = \sum_{i=1}^{k} c_i \Phi(\cdot, t_i) \tag{1.1}$$

[*] This work has been supported by the Austrian Science Foundation FWF through project SFB F 013 / 08.

[†] Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Linz, Austria.

(cf. e. g., [12]). If the parameters t_i are chosen a priori, this results in a *linear* problem, which can be solved easily, but only yields a convergence rate that heavily depends on the dimension of the parameter-space (cf. e. g. [11, 10]).

Therefore, typically the parameters t_i are chosen via an optimization process in dependence of the function f. For instance the "learning" of neural networks can be interpreted as special case of *nonlinear* function approximation, also radial basis functions or fuzzy control fall into this scheme (cf. [1, 8, 4, 2]). In this setting one can obtain—of course at higher computational cost—the dimension independent rate

$$\|f - f_k\| = \mathcal{O}\left(k^{-1/2}\right) .$$

Unfortunately, if all t_i are determined at the same time this not only results in a high-dimensional optimization problem with lots of local minima, but also in instabilities if noise is present (see [2, 9]). For instance it is possible that some of the parameters c_i tend to infinity, or that f_k tends to f in L_2 but in no space H^s with $s > 0$.[1]

An astonishingly simple solution to these two problems is a *greedy algorithm* ([7, 6, 5, 13, 3]). In such an algorithm the optimization problem above is not solved at once, but via a sequence of low-dimensional ones; all parameters t_i are determined one after the other. The functions f_k are then defined inductively as convex-combinations of f_{k-1} and the current element $g_k := c_k\Phi(\cdot, t_k)$.

More precisely, let us assume that the parameters t_i are restricted to some compact set P, and define

$$G = \{\Phi(\cdot, t) \mid t \in P\}$$

(in the following we assume $\|\Phi(\cdot, t)\| \leq 1$ for $t \in P$). Furthermore we assume that f is contained in the closed convex hull of the set $G_b := b \cdot G$, which we denote as $f \in \overline{\text{co}}(G_b)$. In the greedy algorithm elements $g_k \in G_b$ are chosen one after the other, and the approximating functions f_k are built iteratively as convex-combination of f_{k-1} and g_k, as shown in Algorithm 1.1[2]. The main purpose of this work will be, to transfer the *conceptual* Algorithm 1.1 into a *realisable* form.

[1] The common reason for these effects is that the nonlinear scheme (1.1) allows constructions of the form $\psi_\varepsilon = c\,(\Phi(\cdot, t+\varepsilon) - \Phi(\cdot, t))$. Clearly, if f_k is a good approximation to f, then also $f_k + \psi_\varepsilon$ is one, no matter how large c is chosen, provided ε is sufficiently small. Furthermore, the fact that—by a similar construction—f_k may (almost) resemble the kth derivative of Φ, results in the second type of instability.

[2] In the following we consider a setup proposed by Dingankar and Sandberg [7]; a slightly different method with the same spirit has extensively been studied by Temlyakov et. al., see [6, 13] and the references therein. The influence of noise and the unavailability of b have not been considered in these workings.

Algorithm 1.1: Abstract greedy approximation of noise free data with a-priori known smoothness.

```
Set $f_0 = 0$.
Choose a constant $M$, such that $M > b^2 - \|f\|^2$.
Choose a positive sequence $\varepsilon_k$ that fulfills
    $\varepsilon_k \leq \frac{M - (b^2 - \|f\|^2)}{k^2} \qquad \text{for} \quad k = 1, 2, \ldots \qquad (1.2)$
for $k := 1$ to maxit do
    Find an element $g_k \in G_b$ such that
        $\left\| f - \frac{k-1}{k} f_{k-1} - \frac{1}{k} g_k \right\|^2 \leq \inf_{g \in G_b} \left\| f - \frac{k-1}{k} f_{k-1} - \frac{1}{k} g \right\|^2 + \varepsilon_k \qquad (1.3)$
    is fulfilled and define $f_k$ as
        $f_k = \frac{k-1}{k} f_{k-1} + \frac{1}{k} g_k \,.$
end for
```

Condition (1.3) in Algorithm 1.1 shows that it is not allowed to take arbitrary elements g_k in the kth step, but only such, which are almost optimal approximations to the function $kf - (k-1)f_{k-1}$. This local (almost-) optimality is sufficient to maintain the dimension-independent convergence rate, as the next theorem shows (cf. [7]).

Theorem 1.1. *Let $f \in \overline{\text{co}}(G_b)$, then the approximating functions f_k generated by Algorithm 1.1 fulfill the error estimate*

$$\|f - f_k\|^2 \leq \frac{M}{k} \,. \tag{1.4}$$

Thus, in principle Algorithm 1.1 yields the optimal convergence rate $\|f - f_k\| = \mathcal{O}\left(k^{-1/2}\right)$; but as already indicated it is only conceptual and has several disadvantages:

1. We need the smoothness parameter[3] b in order to compute the iteration bound M.

2. We need the sequence ε_k and have to estimate infima to verify if g_k is a sufficiently good approximation.

3. The algorithm is only defined for noise-free data f, also Theorem 1.1 does not provide information about the behavior of Algorithm 1.1 when applied to noisy data f^δ.

It turns out (cf. [3, 9]) that the second point does not pose a problem, since the corresponding step in the algorithm may be replaced by:

$$\text{“Find an element } g_k \in G_b \text{ such that } \left\| f - \frac{k-1}{k} f_{k-1} - \frac{1}{k} g_k \right\|^2 \leq \frac{M}{k} \,. \text{”}$$

Nevertheless, still the parameter M and consequently the smoothness b have to be known. The main purpose of this work is, to develop an algorithm, which can be implemented without knowledge of this smoothness parameter b, but which rather adaptively reconstructs the value of b.

This is important, because usually no information about the size of b will be available, even if—e. g., due to physical considerations—it is known that $f \in \overline{\text{co}}(G_b)$ for *some* b.

To obtain the final *adaptive* Algorithm 4.1 we have to start with an apparently independent step, the investigation of the influence of noise. The reason for this is that a (wrongly) estimated parameter b has the same influence on the algorithm, as noisy data—the function f does not fulfill $f \in \overline{\text{co}}(G_b)$.

The outline of this paper is as follows. In Section 2 we give some results on convex approximation, which are used in Section 3 to derive estimates for noisy data. These two sections will build the basis for Section 4 where we present the adaptive greedy algorithm. Finally the applicability of Algorithm 4.1 is demonstrated by numerical examples in Section 5.

2 Convex Approximation of Noisy Data

First we present two basic results about approximation in the convex hull of a set G (see also [5, Chapter 25]).

Lemma 2.1. *Let H be a Hilbert-space and $G \subset H$ a bounded set. Then for all $h \in \text{co}(G)$ and for all $v \in H$ there exists $g \in G$ such that*

$$\langle h - g, v \rangle \leq 0 \,.$$

[3] To construct the instability effects mentioned above we needed unboundedness of b, vice versa a small value of b ensures that $\overline{\text{co}}(G_b)$ is a set of smooth functions.

This result can also be transferred to elements in $\overline{\mathrm{co}}(G)$, the closure of the convex hull of G.

Corollary 2.2. *Let H be a Hilbert-space and $G \subset H$ a bounded set. Then for all $f \in \overline{\mathrm{co}}(G)$ and for all $v \in H$ the estimate*

$$\inf_{g\in G} \langle f - g, v\rangle \leq 0$$

holds.

Using Corollary 2.2 we can now construct a sharp estimate for the error of convex approximations to noisy data. For the case of noise-free data, i. e., $\delta = 0$, the result simplifies to the estimate given in [7, Lemma 2].

Theorem 2.3. *Let $f \in \overline{\mathrm{co}}(G)$ and f^δ such that $\left\|f - f^\delta\right\| \leq \delta$. Furthermore let $h \in H$ and $\lambda \in [0, 1]$. Then, using the setting $b := \sup_{q\in G} \|q\|$, we have*

$$\begin{aligned}\inf_{g\in G} &\left\|f^\delta - \lambda h - (1-\lambda)g\right\|^2 \\ &\leq \lambda^2 \left\|f^\delta - h\right\|^2 + (1-\lambda)^2\big(b^2 - \|f^\delta\|^2\big) + 2\delta(1-\lambda)\|f^\delta - \lambda h\| \,.\end{aligned} \tag{2.1}$$

Proof. First of all we transfer estimate (2.1) to an equivalent form, by splitting the norm on the left hand side such that it cancels the first term on the right. Remaining we have

$$\begin{aligned}\inf_{g\in G}(1-\lambda)^2 \left\|f^\delta - g\right\|^2 + 2\lambda(1-\lambda)\left\langle f^\delta - g, f^\delta - h\right\rangle& \\ \leq (1-\lambda)^2\big(b^2 - \|f^\delta\|^2\big) + 2\delta(1-\lambda)\|f^\delta - \lambda h\| \,.&\end{aligned}$$

For $\lambda = 1$ this is a trivial result, for $\lambda \neq 1$ we may transfer the relation to

$$\begin{aligned}\inf_{g\in G}(1-\lambda)\left(\|f^\delta - g\|^2 + \|f^\delta\|^2\right) + 2\lambda\left\langle f^\delta - g, f^\delta - h\right\rangle& \\ \leq (1-\lambda)b^2 + 2\delta\|f^\delta - \lambda h\| \,.&\end{aligned}$$

Using the identity $\|f^\delta - g\|^2 + \|f^\delta\|^2 = \|g\|^2 + 2\left\langle f^\delta - g, f^\delta\right\rangle$, we can combine the two scalar products on the left into one. The term $\|g\|^2$ is bounded by b^2. Therefore it suffices to show that

$$\inf_{g\in G} 2\left\langle f^\delta - g, f^\delta - \lambda h\right\rangle \leq 2\delta\|f^\delta - \lambda h\|$$

is fulfilled, which is the direct consequence of the identity

$$\left\langle f^\delta - g, f^\delta - \lambda h\right\rangle = \left\langle f^\delta - f, f^\delta - \lambda h\right\rangle + \left\langle f - g, f^\delta - \lambda h\right\rangle \,,$$

the estimate $\|f^\delta - f\| \leq \delta$, the Cauchy-Schwarz-inequality and Corollary 2.2 for the setting $v = f^\delta - \lambda h$. □

Under the assumptions above, the error-estimate (2.1) can not be improved:

Remark 2.4. The estimate in the Theorem above is sharp, as can be seen for the choice

$$\begin{aligned} &g_0 = 0, \quad g_1 = g \in H \quad \text{with} \quad \|g\| = 1\,. \\ &G = \{g_0, g_1\}, \quad f = h = g, \quad f^\delta = (1+\delta)g \quad \text{with some} \quad \delta > 0\,. \end{aligned}$$

With this choice of G, f and f^δ we obtain equality in Theorem 2.3, independent of the value of λ.

When the greedy algorithm is applied to noisy data $f^\delta \notin \overline{\mathrm{co}}(G_b)$, Theorem 1.1 cannot hold, since in this case even the optimal approximation yields a residual greater than 0. Nevertheless, it turns out that the rate M/k can at least be obtained up to a certain iteration index k_*. In the next section we will derive a sharp estimate for this iteration index and the corresponding residual.

3 Optimal Greedy Iteration for Noisy Data

In this section we consider the case that instead of $f \in \overline{\mathrm{co}}(G_b)$ only a noisy version f^δ with $\left\|f - f^\delta\right\| \leq \delta$ is available.

For the case of noise-free data we had to pick $M > (b^2 - \|f\|^2)$ in Algorithm 1.1, now it turns out that we need at least $M > M_0$ with

$$M_0 := \left(b^2 - \|f^\delta\|^2 + 2\delta\|f^\delta\|\right)\,. \tag{3.1}$$

Furthermore, we find (cf. Theorem 3.1 and Remark 3.3) that we cannot guarantee the existence of proper updates g_k as soon as $k > k_*$, where

$$k_* := \left\lceil \frac{\eta^2 M_0}{4\delta^2(1+\eta)} \right\rceil\,, \tag{3.2}$$

and we assumed that $M = (1+\eta)M_0$. Both values will appear in a natural way in Theorems 3.1 and 3.2, but first we have a look at the modified greedy algorithm shown on the following page.

The crucial step in Algorithm 3.1 is to find elements g_k^δ that are a sufficiently good approximation to $kf^\delta - (k-1)f_{k-1}^\delta$. Based on Theorem 2.3 we are able to show, that such elements indeed exist for indices $k \leq k_*$.

Algorithm 3.1: Greedy approximation of noisy data with given smoothness parameter b.

```
Set $f_0^\delta = 0$.
Choose $M > M_0$ with $M_0$ as in (3.1).
Compute $k_*$ via (3.2).
for $k := 1$ to $\min(k_*, \mathrm{maxit})$ do
    Find $g_k^\delta \in G$ (see Theorem 3.1) such that
        $\left\| f^\delta - \frac{k-1}{k} f_{k-1}^\delta - \frac{1}{k} g_k^\delta \right\|^2 \le \frac{M}{k}$
    is fulfilled and define $f_k^\delta$ as
        $f_k^\delta = \frac{k-1}{k} f_{k-1}^\delta + \frac{1}{k} g_k^\delta \,.$
end for
```

Theorem 3.1. *For indices $1 \le k \le k_*$ Algorithm 3.1 is feasible, i. e., in each step suitable elements g_k^δ can be found. The corresponding approximations f_k^δ satisfy the error estimate*

$$\left\| f^\delta - f_k^\delta \right\|^2 \le \frac{M}{k} \quad \textit{for} \quad 1 \le k \le k_* \,. \tag{3.3}$$

Proof. The proof uses an induction argument, based on Theorem 2.3. We consider Algorithm 3.1 with a similar inf-condition as Algorithm 1.1. Therefore we define a sequence ε_k as

$$\varepsilon_k := \frac{1}{k^2} \Big(M - (b^2 - \|f^\delta\|^2 + 2\delta \|f^\delta\|) - 2\delta \sqrt{k-1} \sqrt{M} \Big) \,. \tag{3.4}$$

Since the right hand side of (3.4) becomes negative for $k \to \infty$, for given M, b, δ and f^δ there exists a unique index k_* with

$$\varepsilon_{k_*} > 0 \quad \text{and} \quad \varepsilon_{k_*+1} \le 0 \,.$$

To compute k_* we solve the equation $\varepsilon(k) = 0$ which is equivalent to

$$M - M_0 - 2\delta \sqrt{k-1} \sqrt{M} = 0 \,,$$

the solution for k is given as

$$\tilde{k} = \frac{\eta^2 M_0}{4\delta^2(1+\eta)} + 1\ . \tag{3.5}$$

Since this value is related to the integer value k_* via $\tilde{k} > k_* \geq \tilde{k} - 1$ we obtain (3.2). We will now show that up to this index k_*, the rate M/k can be maintained.

- For the step $k = 1$ we obtain in the modified algorithm

  $$\left\|f^\delta - g_1^\delta\right\|^2 \leq \inf_{g\in G} \left\|f^\delta - g\right\|^2 + \varepsilon_1 \tag{3.6}$$

  which we can estimate using Theorem 2.3 for $\lambda = 0$ via

  $$\begin{aligned} &\leq (b^2 - \left\|f^\delta\right\|^2) + 2\delta\left\|f^\delta\right\| + \varepsilon_1 \\ &\leq M\ , \end{aligned}$$

  since ε_1 was chosen according to (3.4).

- Now we inspect the case $1 < k \leq k_*$. We assume that the convergence rate was preserved up to this step of the iteration, this means that the estimate $\left\|f^\delta - f_{k-1}^\delta\right\| < \frac{\sqrt{M}}{\sqrt{k-1}}$ holds. In the kth step we have

  $$\begin{aligned} &\left\|f^\delta - \frac{k-1}{k} f_{k-1}^\delta - \frac{1}{k} g_k^\delta\right\|^2 \\ &\qquad \leq \inf_{g\in G} \left\|f^\delta - \frac{k-1}{k} f_{k-1}^\delta - \frac{1}{k} g\right\|^2 + \varepsilon_k\ , \end{aligned} \tag{3.7}$$

  which can again be estimated using Theorem 2.3 via

  $$\begin{aligned} &\leq \left(\frac{k-1}{k}\right)^2 \left\|f^\delta - f_{k-1}^\delta\right\|^2 + \frac{1}{k^2}(b^2 - \left\|f^\delta\right\|^2) \\ &\qquad + 2\delta\frac{1}{k}\left\|f^\delta - \frac{k-1}{k} f_{k-1}^\delta\right\| + \varepsilon_k \\ &\leq \frac{k-1}{k^2} M + \frac{1}{k^2}(b^2 - \left\|f^\delta\right\|^2) \\ &\qquad + 2\delta\frac{1}{k}\left(\frac{k-1}{k}\left\|f^\delta - f_{k-1}^\delta\right\| + \frac{1}{k}\left\|f^\delta\right\|\right) + \varepsilon_k \end{aligned}$$

We can now insert the estimate for $\left\|f^\delta - f^\delta_{k-1}\right\|$ a second time and obtain further

$$\begin{aligned}
&\leq \frac{1}{k^2}\left((k-1)M + b^2 - \left\|f^\delta\right\|^2 + 2\delta\left(\sqrt{k-1}\sqrt{M} + \left\|f^\delta\right\|\right)\right) + \varepsilon_k \\
&\leq \frac{M}{k}\,,
\end{aligned}$$

since ε_k was chosen according to (3.4).

Elements g_1^δ and g_k^δ in (3.6) and (3.7) can always be found, since ε_1 and ε_k are positive. These elements yield the rate $\frac{M}{k}$ and thus the algorithm is feasible. □

Since the rate $\mathcal{O}\left(k^{-1/2}\right)$ only holds up to the index k_*, which depends on M, f^δ, δ and b, it is a natural next step to look for parameters $M = M(f^\delta, \delta, b)$, such that the residual at the end of the iteration is minimized. The result of this optimization step is given in the next theorem.

Theorem 3.2. *Let M be chosen as $M = (1+\eta)M_0$, with M_0 as in* (3.1) *and $\eta > 0$. Then for the index k_* defined via* (3.2) *the approximations $f^\delta_{k_*}$ in the greedy algorithm fulfill the estimate*

$$\left\|f^\delta - f^\delta_{k_*}\right\| \leq 2\,\frac{1+\eta}{\eta}\,\delta = \mathcal{O}\left(\delta\right)\,. \tag{3.8}$$

Proof. According to Theorem 3.1 the residual at the end of the iteration is given by $\frac{M}{k_*}$ where k_* is defined via (3.2). Since $k_* \geq \tilde{k} - 1$, with $\tilde{k}$ defined in (3.5) we can estimate the residual as

$$\begin{aligned}
\left\|f^\delta - f^\delta_{k_*}\right\|^2 \leq \frac{M}{k_*} &\leq \frac{M}{\tilde{k} - 1} \\
&= (1+\eta)M_0 \frac{4\delta^2(1+\eta)}{\eta^2 M_0} \\
&= 4\delta^2 \frac{(1+\eta)^2}{\eta^2}
\end{aligned}$$

which completes the proof. □

We will now show that the index k_* is optimal, i.e. that is is in general not possible to find proper updates g_k^δ in the greedy algorithm for indices $k > k_*$. Therefore we demonstrate that the error estimate in the theorem above is a sharp bound for the minimal residual for countably many values of η, in particular for a sequence $\eta_i \to \infty$.

Remark 3.3. To show that estimate (3.8) is a sharp bound for the minimal residual, we choose G as the one-dimensional interval $[0, b]$. The exact data is chosen as $f = b$, and we assume that instead of f we are only given a noisy version $f^\delta = b + \delta$, i. e., the noise level is δ.

We now fix μ and η with $1 \leq \mu < 2\frac{1+\eta}{\eta}$ and $(1 + \eta)/\mu^2 =: k_*$ integer, and construct an approximating sequence, for which the greedy algorithm terminates with residual $\left\| f^\delta - f^\delta_{k_*} \right\| = \mu\delta$.

With this choice of parameters we have for all $k \leq k_*$ that $f^\delta_k := b-(\mu-1)\delta$ is a sufficiently good approximation. Indeed, we have

$$\left\| f^\delta - f^\delta_k \right\| = \mu\delta \leq \sqrt{\frac{(1+\eta)\delta^2}{k}} \quad \text{for} \quad k \leq k_* \,.$$

We now show that the greedy algorithm terminates in the next step of the iteration, which proves that estimate (3.8) is sharp: The optimal element g_{k_*+1} is given as $g_{k_*+1} = b$, hence $f^\delta_{k_*+1} := \frac{k_*}{k_*+1} f^\delta_{k_*} + \frac{1}{k_*+1} b$, but this approximation is not sufficiently good since

$$\left\| f^\delta - \frac{k_*}{k_*+1} f^\delta_{k_*} - \frac{1}{k_*+1} b \right\| = \delta \left(1 + \frac{k_*(\mu-1)}{k_*+1} \right) > \sqrt{\frac{(1+\eta)\delta^2}{k_*+1}} \,,$$

as a straight-forward computation shows. Hence for $\mu < 2\frac{1+\eta}{\eta}$ and appropriate η we obtain that the estimate is sharp.

The reason why we cannot get this result for arbitrary values of η is that in the proof of Theorem 3.2 we had to distinguish between the real value $\tilde{k}$ and the integer k_*. Ideally these values are almost equal, in the worst case their ratio is $\eta^2/(2+\eta)^2$. In this case estimate (3.8) is only sharp up to the factor $\eta/(2+\eta)$.

In principle the estimate could be made sharp for all values of η by introducing the factor

$$\left\lceil \frac{\eta^2}{4(1+\eta)} \right\rceil \Big/ \left(\frac{\eta^2}{4(1+\eta)} + 1 \right) \,,$$

where $\lceil a \rceil$ denotes the ceiling of a. Nevertheless, we omit this factor for the sake of readability.

It should be mentioned that a different estimate is available in the case that within the greedy algorithm also the weighting in the convex-combination is optimized (see [5, Chap. 25]).

4 An Adaptive Greedy Algorithm for Data with Unknown Smoothness

In this section we develop the adaptive greedy algorithm, which will be applicable also if the smoothness of the (noisy) data is not known a-priori.

The motivation for this algorithm is as follows: Assume that we are given data $f \in \overline{\text{co}}(G_B)$, where we do not know the actual value B, but we have the additional knowledge that $f \in \overline{\text{co}}(G_b)$ for *some* b. The natural approach would be to guess $b \lesssim B$ and—if the algorithm does not converge "properly"—increase b by a certain amount. The results of the section above will help us to provide a theoretical basis for this heuristic method.

The main idea is that an incorrect, i. e., too small choice of b has the same effect as noise—the given data f does not fulfill $f \in \overline{\text{co}}(G_b)$. In the previous section we have developed sharp estimates for the corresponding termination index k_*, now we will use these estimates to develop an update rule for the parameter b. As a first step, we have to transfer the results from the previous section to the case of "artificial noise", i. e., noise that is caused by a wrong choice of b.

Corollary 4.1. *Let $f \in \overline{\text{co}}(G_B)$ and $M = (1+\eta)(b^2 - \|f\|^2 + 2\frac{B-b}{B}\|f\|^2)$ with $b \leq B$. Then the approximations of Algorithm 3.1 fulfill*

$$\|f - f_{k_*}\| \leq 2\,\frac{1+\eta}{\eta}\,\frac{B-b}{B}\,\|f\| \tag{4.1}$$

Proof. Since $\frac{b}{B}f \in \overline{\text{co}}(G_b)$, we can interpret f as a noisy version of $\frac{b}{B}f$, where the noise level δ can be estimated as $\delta \leq \frac{B-b}{B}\|f\|$. The proof now follows with Theorem 3.2. □

In practice neither η nor B are known, in the following lemma we express η in terms of B, b, f and τ.

Lemma 4.2. *Let $f \in \overline{\text{co}}(G_B)$ and $M = (1+\tau)(b^2 + \|f\|^2)$ with $0 < b \leq B$. Then the approximations of Algorithm 3.1 fulfill*

$$\|f - f_{k_*}\| \leq 2\,\frac{(1+\tau)(b^2 + \|f\|^2)}{B\tau(b^2+\|f\|^2) + 2b\,\|f\|^2}\,(B-b)\,\|f\| \tag{4.2}$$

Proof. Follows immediately from Corollary 4.1 using the relation $\frac{1+\eta}{\eta} = \frac{M}{M-M_0}$, where $M_0 = (b^2 - \|f\|^2 + 2\frac{B-b}{B}\|f\|^2)$ and $M = (1+\tau)(b^2+\|f\|^2)$. □

With the estimate of this lemma, we can now construct a lower bound for the true, unknown parameter B, which we will use as update-rule in Algorithm 4.1.

Algorithm 4.1: Adaptive greedy algorithm for approximation of data with unknown smoothness parameter B.

1. Choose $b_0 < B$.[a]

   Set $k = 1$ and $f_0 = 0$.

2. Perform iterations in Algorithm 3.1 as follows

   - Take $M = (1+\tau)(b_i^2 + \|f^{(\delta)}\|^2)$ with some $\tau \geq 0$ in the noise-free case and $\tau \geq 4\xi / \left\|f^\delta\right\|$ for noisy data.
   - Perform iterations as long as valid updates g_k can be found.[b]

3. If the discrepancy principle (4.4) is fulfilled, then stop the iteration. Otherwise use the residual in the greedy-algorithm to obtain a better estimate b_{i+1} for B (see (4.3) and (4.5)), and continue with step 2 at the index $k = k_{*,i}$.

[a] Choices that guarantee this are $b_0 = \|f\|/2$ and $b_0 = (\left\|f^\delta\right\| - \xi)/2$ respectively. In general severe underestimation is not a problem, b_0 may trouble-free be 10^5 times smaller than B (cf. the discussion of Figure 5.4).

[b] Since we try to approximate data $f \in \overline{\mathrm{co}}(G_B)$, using elements $f_k \in \overline{\mathrm{co}}(G_{b_i}) \subsetneq \overline{\mathrm{co}}(G_B)$, the greedy-algorithm will fail to find a sufficiently good update after a certain number $k_{*,i}$ of iterations (see also Remark 4.4).

Theorem 4.3. *Let $f \in \overline{\mathrm{co}}(G_B)$ and $M = (1+\tau)(b^2 + \|f\|^2)$ with $0 < b \leq B$. Then the residual at the end of the iteration of Algorithm 3.1 provides a lower bound for B via*

$$B \geq \tilde{b}\left(b, \tau, f, f_{k_*}\right) := b \frac{(1+\tau)\|f\| + \|f - f_{k_*}\| \frac{\|f\|^2}{b^2 + \|f\|^2}}{(1+\tau)\|f\| - \frac{\tau}{2}\|f - f_{k_*}\|} \geq b \tag{4.3}$$

Proof. Follows from Lemma 4.2 under the observation that $\tau \|f - f_{k_*}\| < 2(1+\tau)\|f\|$ for $b > 0$. □

With this update rule, we are now able to construct the adaptive Algorithm 4.1 (given on top of this page). The estimates b_i that are generated within Algorithm 4.1 fulfill $\lim b_i \leq B$, this means that throughout the iteration the generated approximations f_k remain at least as smooth as f.

Nevertheless, Theorem 4.3 is still not a complete result, since there are also numerical effects, that have to be taken care of.

Remark 4.4. In practice there are two effects that may require to adjust estimate (4.3).

Noisy data: Besides the missing information about the value of B we might even only be given a noisy version f^δ with $\left\|f - f^\delta\right\| \leq \xi$. If a bound on the noise level is known, we can derive similar results as in Theorem 4.3 (see Theorem 4.6).

Furthermore, since for noisy data the optimal residual is larger than zero, we have to incorporate a discrepancy-type stopping criterion into the algorithm (see Remark 4.5).

Numerical minimization: The estimates in this section are based on the fact, that sufficiently good approximations g_k^δ exist for indices $k \leq k_*$. The index $\hat{k}$, for which no such approximation $g_{\hat{k}}^\delta$ exists at all is an upper bound for k_*, i. e., $k_* \leq \hat{k}$. Numerically we try to estimate $\hat{k}$ by observing when the algorithm fails to find a sufficiently good update g_k^δ within reasonable time.

If we terminate the algorithm too early, we underestimate $\hat{k}$ and consequently k_*. Fortunately, Lemma 4.7 shows that this does not pose a problem as long as the search for the (almost) optimal element is performed as thorough as in the original algorithm. This lemma also gives a bound for the amount of underestimation.

As mentioned above, in the case of noisy data we cannot obtain an arbitrarily small residual even if the parameter b would be chosen correctly. Therefore we have to use an additional stopping rule.

Remark 4.5 (Discrepancy principle). For noisy data with noise level $\left\|f - f^\delta\right\| \leq \xi$, Algorithm 4.1 should be stopped at the index k for which for the first time the estimate

$$\left\|f^\delta - f_k^\delta\right\| \leq 2 \frac{(1+\tau)(b^2 + \left\|f^\delta\right\|^2)}{\tau(b^2 + \|f^\delta\|^2) + 2\,\|f^\delta\|\,(\|f^\delta\| - \xi)}\,\xi \tag{4.4}$$

is fulfilled. This follows from the fact that with correct choice of b (i. e., $b = B$), this is—according to Theorem 3.2—the minimal residual that we can expect with noisy data.

In practice we do not have to check (4.4) for every k, but only in step 3 of Algorithm 4.1, i. e., when we have to check whether we should update b_j or stop the algorithm.

Using this discrepancy rule we can now give the main result of this work: the update-rule for b for the case of noisy data. This rule was used to generate the numerical examples in Section 5.

Theorem 4.6. *Let $f \in \overline{\mathrm{co}}(G_B)$, f^δ such that $\left\|f - f^\delta\right\| \leq \xi$ and $M = (1 + \tau)(b^2 + \left\|f^\delta\right\|^2)$ with $0 < b \leq B$ and $\tau \geq 4\xi / \left\|f^\delta\right\|$. Then the residual at the end of the iteration in Algorithm 3.1 provides a lower bound for B via*

$$B \geq \tilde{b}\left(b, \tau, f^\delta, f^\delta_{k_*}, \xi\right) := \\ b \, \frac{2\left(\xi + \|f^\delta\|\right)\left(\|f^\delta - f^\delta_{k_*}\| \, \|f^\delta\| + M\right)}{2M\left(2\xi + \|f^\delta\|\right) - \|f^\delta - f^\delta_{k_*}\| \left(\tau(b^2 + \|f^\delta\|^2) - 4\xi \, \|f^\delta\|\right)} \, . \tag{4.5}$$

If furthermore the discrepancy rule (4.4) *is used, we obtain in addition*

$$\tilde{b}\left(b, \tau, f^\delta, f^\delta_{k_*}, \xi\right) \geq b \, , \tag{4.6}$$

i. e., Algorithm 4.1 generates a monotonically increasing sequence b_j with $\lim b_j \leq B$. The discrepancy rule is a necessary condition for monotonicity.

Proof. The proof follows with similar arguments as the proofs of Corollary 4.1 to Theorem 4.3. Again we start with Theorem 3.2, but now with the total noise level, which can be bounded via $\delta \leq \xi(2 - b/B) + (B - b)/B \left\|f^\delta\right\|$. Using the relation $(1 + \eta)/\eta = M/(M - M_0)$ we obtain the estimate

$$\left\|f - f^\delta_{k_*}\right\| \leq \frac{2M\left(B(2\xi + \left\|f^\delta\right\|) - b(\xi + \left\|f^\delta\right\|)\right)}{B(\tau(b^2 + \|f^\delta\|^2) - 4\xi \left\|f^\delta\right\|) + 2b \left\|f^\delta\right\| (\xi + \left\|f^\delta\right\|)} \, ,$$

where we needed that $\tau \geq 4\xi / \left\|f^\delta\right\|$. This result now immediately yields the estimate for B. Under the additional assumption that the discrepancy principle of Remark 4.5 was used, we have a lower bound for the residual and can therefore derive the monotonicity result (4.6). Vice versa, assuming the monotonicity, one obtains (4.6). □

Observe that Theorem 4.5 contains the result of Theorem 4.3, since for the case of noise-free data, estimate (4.5) simplifies to (4.3).

Finally we briefly discuss the second point of Remark 4.4. In the original greedy algorithm we have to look for almost optimal elements, where the distance to the optimum in the kth step is bounded by $\tau(b^2 + \|f\|^2)/k^2$ (cf. the definition of ε_k in (1.2)). If we perform the algorithm for unknown smoothness with the slightly better precision $\lambda\varepsilon_k$ with $\lambda < 1$, we can estimate the ratio of k_* and the actual stopping index $\hat{k}$. In both cases the precision has to tend to zero as $\mathcal{O}\left(k^{-2}\right)$.

Lemma 4.7. *Let Algorithm 1.1 be performed with $f \in \overline{\mathrm{co}}(G_B)$ where $B > b$, and precision $\lambda\tau(b^2 + \|f\|^2)/k^2$ where $\lambda < 1$. Then the algorithm is feasible up to an index $\hat{k}$, where we have the estimate*

$$\frac{\sqrt{k_* - 1}}{\sqrt{\hat{k} - 1}} \leq 1 + \frac{\lambda}{1 - \lambda + 2\frac{b\|f\|^2}{\tau B(b^2 + \|f\|^2)}} \leq \frac{1}{1 - \lambda} \tag{4.7}$$

Proof. The algorithm will terminate at the index $\hat{k}$ for which the working precision $\lambda\tau(b^2 + \|f\|^2)/k^2$ is larger than the required precision ε_k, given in (3.4). For this index $\hat{k}$ we have the equation

$$\frac{\lambda\tau(b^2 + \|f\|^2)}{\hat{k}^2} \geq \frac{1}{\hat{k}^2}\left(M - \left(b^2 - \|f\|^2 + 2\frac{B-b}{B}\|f\|^2\right) - 2\delta\sqrt{\hat{k}-1}\sqrt{M}\right),$$

with δ as in the proof of Corollary 4.1. This yields the estimate

$$2\delta\sqrt{\hat{k}-1}\sqrt{M} \geq (\tau - \lambda\tau)\left(b^2 + \|f\|^2\right) + 2\frac{b}{B}\|f\|^2$$

Since for k_* we have the relation

$$2\delta\sqrt{k_* - 1}\sqrt{M} \geq \tau\left(b^2 + \|f\|^2\right) + 2\frac{b}{B}\|f\|^2$$

we obtain the first estimate in (4.7), the second one follows by $b \geq 0$. □

The following remark shows, how the assumptions of Lemma 4.7 can be fulfilled in a simple manner.

Remark 4.8. Let the quadratic minimization functional $L(t)$ be defined as

$$L(t) := \|F - \Phi(\cdot, t)\|^2 = \int (F(x) - \Phi(x,t))^2\, dx\,.$$

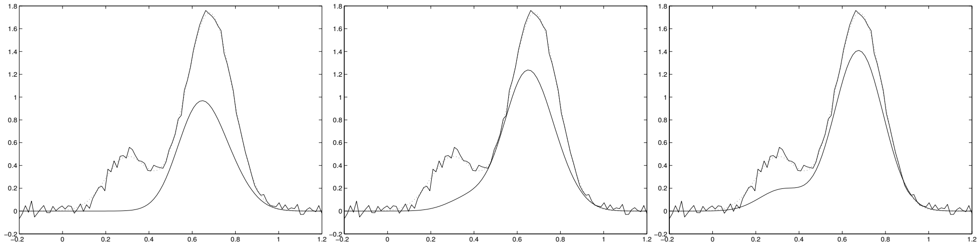
In the setting of Section 5 $\Phi(\cdot,\cdot)$ is a radial basis function, i. e., it can be written as $\Phi(x,t) = \Xi(\|x-t\|^2)$. Therefore the second derivative of $L(t)$ can be estimated as $|L''(t)| \leq 2\,\|\Phi_{t,t}(\cdot,t)\|\,\|F\|$. Close to a local minimum t_0 we now obtain

$$|L(t) - L(t_0)| \simeq |(t-t_0)L'(t_0) + \frac{(t-t_0)^2}{2}L''(t_0)| \leq (t-t_0)^2\,\|\Phi_{t,t}\|\,\|F\|\,.$$

If we insert the functions $F = f^\delta - \frac{k-1}{k}f^\delta_{k-1}$ and $\Phi = \frac{1}{k}g^\delta_k$ we obtain further

$$|L(t) - L(t_0)| \lesssim \frac{(t-t_0)^2}{k^2}\left(\|f^\delta\| + \sqrt{k-1}\sqrt{M}\right)\|g_{t,t}\|\,.$$

Thus, in order to obtain the accuracy $\mathcal{O}\left(1/k^2\right)$ in the kth step of the greedy algorithm, it is sufficient to choose $\mathcal{O}(\sqrt[4]{k})$ evenly distributed values for t. There is no need for additional optimization steps.



$k = k_{*,12} + 1 = 9$ $k = 20$ $k = k_{*,13} = 44$

Figure 5.1: Development of f_k^δ within the greedy algorithm with 5% noise, $B = 1$ and $b = b_{13} = 0.8570$. The three graphs correspond to $k = k_{*,12} + 1 = 9$, $k = 20$ and the $k = k_{*,13} = 44$.

5 Numerical Experiments

To test the results of the preceding sections numerically, we implement a greedy algorithm for a simple, but still infinite-dimensional setting:

The set G_b is generated by Gaussian functions with fixed diameter and variable center, where the centers are taken from the interval $[-0.2, 1.2]$. More precisely we define

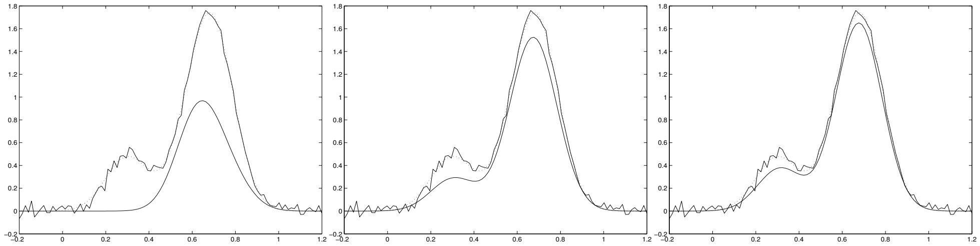
$$G_b := b \cdot G \quad \text{with} \quad G := \left\{ \frac{e^{-50(x-t)^2}}{\sqrt[4]{\pi/100}} \mid x, t \in [-0.2, 1.2] \right\} \tag{5.1}$$

(We do not have $\|g(t)\| = b$ for all t, since part of the function $g(t)$ may lie outside the interval. Nevertheless all theorems only require that $\|g\| \leq b$ for elements $g \in G_b$). The function f to be approximated is given as $0.2g(0.6) + 0.2g(0.3) + 0.6g(0.7)$, i. e., $B = 1$. This function is discretized and afterwards contaminated with Gaussian white noise; as initial guess for B we set $b_0 = 0.001$.

The second step of Algorithm 4.1 is implemented in a very simple way: To find suitable elements g_k we take $t_r \in [-0.2, 1.2]$ randomly, and take $g_k := \pm g(t_r)$, where also the sign is determined at random (see Remark 4.8). If with this element the residual is sufficiently small, the convex combination $f_{k+1} = k/(k+1)f_k + 1/(k+1)g_k$ is computed, otherwise a new element g_k is generated. If this procedure fails to find an update within a given number of trials[4], the algorithm breaks. Figure 5.1 shows the development of the iterates in this procedure for $b = b_{12}$.

If the computed residual at the end of this approach is already sufficiently small, i. e., the discrepancy rule (4.4) is fulfilled, then Algorithm 4.1 is ter-

[4] In the given examples the number of trials in step k was restricted to $25\sqrt[4]{k}$.



$k_{*,12} = 8$, $b_{12} = 0.71$ $k_{*,14} = 97$, $b_{14} = 0.91$ $k_{*,18} = 295$, $b_{18} = 0.96$

Figure 5.2: Development of f_k^δ within the greedy algorithm with 5% noise and different values of b_j. The algorithm was started with $b_0 = 10^{-3}$, the discrepancy-rule was fulfilled with $b_{18} = 0.9575 < B$.

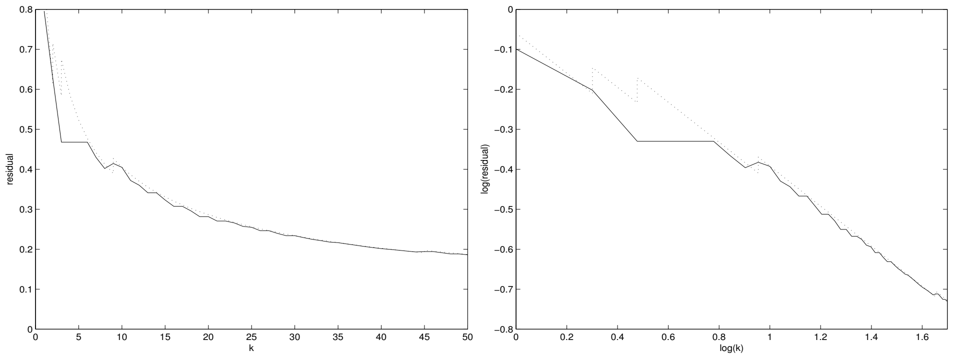


Figure 5.3: Development of the residual within the adaptive greedy algorithm. The solid line represents the residual, the dotted one corresponds to the iteration bound $\sqrt{M/k}$. The updates for b_j lead to the typical saw-tooth structure.

minated. Otherwise the result of Theorem 4.6 is used, in order to generate a better estimate for B. While b_j increases, also the iterates become better approximations to the (noisy) data (see Figure 5.2).

Due to (4.6) we can be assured to obtain an increasing sequence b_j with $\lim b_j \leq B$.

Since $b_{j+1} \geq b_j$ we have $f_{k_*,j}^\delta \in \overline{\text{co}}(G_{b_{j+1}})$, therefore we are allowed to continue the iteration at the current index k, there is no need to restart the whole algorithm with the index $k = 1$. This procedure yields the typical saw-tooth shape in Figure 5.3.

Figure 5.4 shows the development of b_j during the algorithm. As can be seen, the estimates immediately ($k \leq 3$) increase up to the correct order of magnitude. After a few more updates the discrepancy rule is ful-

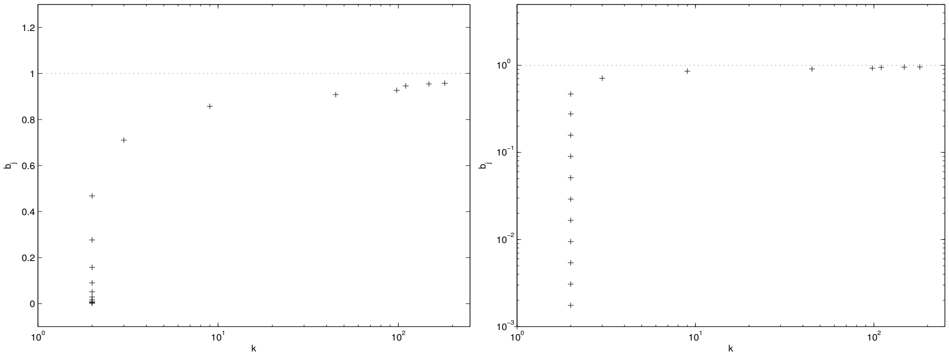


Figure 5.4: Development of the estimates b_j for noise level $\xi = 5\%$. The estimates immediately approach the correct order of magnitude, already for $k \leq 3$ the parameter b is increased from $b_0 = 10^{-3}$ to $b_{12} = 0.7105$.

filled and the algorithm terminates with $b = b_{18} = 0.9757$. The residual is $\left\|f^\delta - f^\delta_k\right\| / \left\|f^\delta\right\| = 9.98\% \approx 2 \cdot \xi$.

Finally, in Figure 5.5 we investigate the influence of the noise level on the quality of the results. Clearly, the residual results. Clearly at the end of the iteration will be larger for higher noise levels, the left plot shows that the ratio between residual and noise level is approximately constant. The right graph demonstrates the influence of noise on the recovery of B. For high noise levels, B is underestimated due to the discrepancy rule—very small values of b (typical: $b \approx 0.5B$) already yield sufficient approximations. For low noise levels, B is estimated correctly or even overestimated. The overestimation is due to the numerical effects described in Lemma 4.7, and could in principle be avoided. Nevertheless, this is not necessary, since typically b stays less than B, and even in the worst case we only observed $b \lesssim 1.2B$. Furthermore, after the first step of overestimation the algorithm will usually terminate due to the discrepancy rule, so there is no danger of substantial overestimation. The algorithm always produces smooth solutions.

Acknowledgements

The author thanks Martin Burger for useful and stimulating discussions. Financial support is acknowledged to the Austrian Science Foundation FWF through project SFB F 013 / 08.

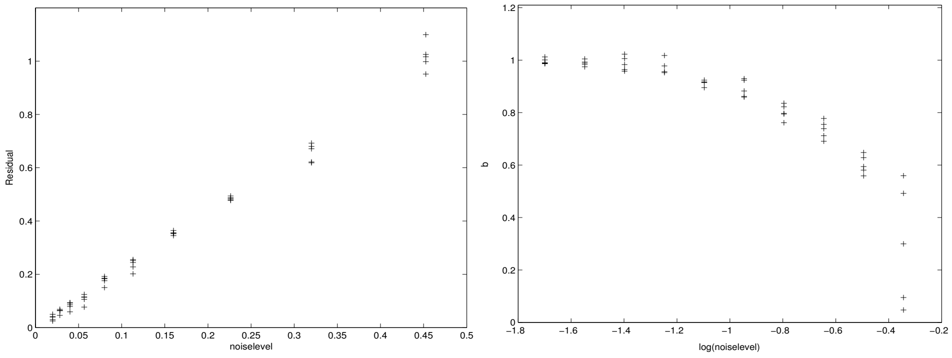


Figure 5.5: The left plot shows the dependence of the final residual on the noise level, the right one demonstrates the influence on the estimates for B. For every noise level the algorithm was run 5 times, the noise ranges from 2% to 45%.

References

[1] A. R. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Trans. Inform. Theory, 39 (1993), pp. 930–945.

[2] U. Bodenhofer, M. Burger, H. W. Engl, and J. Haslinger, *Regularized data-driven construction of fuzzy controllers*, J. Inverse Ill-Posed Probl., 10 (2002), pp. 319–344.

[3] M. Burger and A. Hofinger, *Regularized greedy algorithms for network training with data noise*, Computing, (to appear).

[4] M. Burger and A. Neubauer, *Error bounds for approximation with neural networks*, J. Approx. Theory, 112 (2001), pp. 235–250.

[5] W. Cheney and W. Light, *A Course in Approximation Theory*, Brooks/Cole Publishing Company, 2000.

[6] R. A. DeVore and A. N. Temlyakov, *Some remarks on greedy algorithms*, Adv. Comput. Math., 5 (1996), pp. 173–187.

[7] A. T. Dingankar and I. W. Sandberg, *A note on error bounds for approximation in inner product spaces*, Circuits Syst. Signal Process., 15 (1996), pp. 519–522.

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

[9] A. Hofinger, *Iterative regularization and training of neural networks*, Diplomarbeit, University of Linz, 2003.

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

[11] A. Pinkus, *n-Widths in Approximation Theory*, Springer, Berlin, Heidelberg, 1985.

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

[13] A. N. Temlyakov, *Weak greedy algorithms*, Adv. Comput. Math., 12 (2000), pp. 213–227.