Regularized Optimization of Fuzzy Controllers

Martin Burger
SFB F 013 Numerical and Symbolic Scientific Computing and
Industrial Mathematics Institute, Johannes Kepler Universität
Linz, Altenbergerstr.69, A-4040 Linz, Austria.

Josef Haslinger Ulrich Bodenhofer
Software Competence Center Hagenberg GmbH
Hauptstrasse 99, A-4232 Hagenberg, Austria

Abstract

This paper is devoted to the mathematical analysis and the numerical solution of the problem of designing fuzzy controllers. We show that for a special class of controllers (so-called *Sugeno controllers*), the design problem is equivalent to a nonlinear least squares problem, which turns out to be *ill-posed*. Therefore we investigate the use of regularization methods in order to obtain stable approximations of the solution. We analyze a smoothing method, which is common in spline approximation, as well as Tikhonov regularization with respect to stability and convergence.

In addition, we develop an iterative method for the regularized problems, which uses the special structure of the problem and test it in some typical numerical examples. We also compare the behavior of the iterations for the original and the regularized least squares problems. It turns out that the regularized problem is not only more robust but also favors solutions that are interpretable easily, which is an important criterion for fuzzy systems.

Keywords: fuzzy control, regularization, stability, nonlinear least squares, optimization, ill-posed problems.

Subject Classification (AMS 2000): 65F22, 93C42, 49N45, 65K10.

1 Introduction

Fundamentally, the idea of fuzzy sets and systems, dated back to Zadeh [31, 32], is to provide a mathematical model that can present and process vague, imprecise and uncertain knowledge. It has been modeled on human thinking and the ability of humans to perform approximate reasoning, so that precise and yet significant statements can be made on the behavior of a complex system. Successful applications of fuzzy logic control include

automatic train operation systems, elevator control, temperature control, power plant control, fuzzy refrigerators, washing machines, etc. The main advantage of fuzzy controllers in comparison with other adaptive systems like neural networks is the linguistic interpretability of the controller's function.

1.1 Fuzzy Control

Basically, a fuzzy logic controller consists of three components [1, 7, 16]:

1. The rules, i.e. a verbal description of the relationships usually of a form as the following (n is the number of rules):

   if x is A_i then u is B_i $\qquad (i = 1, \ldots, n)$

2. The fuzzy sets (membership functions), i.e. the semantics of the vague expressions A_i, B_i used in the rules. More precisely (cf. [2]): Given a universe of discourse X a fuzzy subset A of X is characterized by its membership function

   $$\mu_A : X \to [0, 1] \tag{1.1}$$

   where for $x \in X$ the number $\mu_A(x)$ is interpreted as the degree of membership of x in the fuzzy set A.

3. An inference machine, i.e. a mathematical methodology for processing a given input through the rule base. The general inference process proceeds in three (or four) steps.

   (a) Under *Fuzzification*, the membership functions defined on the input variables are applied to their actual values, to determine the degree of truth for each rule premise.

   (b) Under *Inference*, the truth value for the premise of each rule is computed, and applied to the conclusion part of each rule. This results in one fuzzy subset to be assigned to each output variable for each rule. Usually only minimum or product are used as inference rules as special cases of a triangular norm (t-norm, [2]).

   (c) Under *Composition*, all of the fuzzy subsets assigned to each output variable are combined together to form a single fuzzy subset for each output variable. Usually maximum or summation are used.

   (d) Finally is the (optional) *defuzzification*, which is used to convert the fuzzy output set to a crisp number. Two of the more common defuzzification methods are the centroid (center of gravity) and the maximum method.

In the following we assume that a reasonable inference scheme—a Sugeno controller [27], where the output membership functions are crisp values—is given. For a complete definition of a Sugeno controller, see Section 2.

There are still two components left which have to be specified in order to design a fuzzy controller—the rules and the fuzzy sets. Recent effort has been concentrated on developing new techniques which may be able to design the membership functions and rule base automatically. Genetic algorithms have played a special role in fuzzy control design as well as methods treating fuzzy systems as artificial neural networks to adjust membership functions using back propagation. For references see the article of Tan and Hu [28]. Also classical optimization algorithms, such as the method of steepest descent have been applied in tuning small and medium sized controllers.

Under the quite natural assumptions that product is used as fuzzy inference rule, summation as the composition scheme, and center of gravity as the defuzzification method, the tuning of a Sugeno controller reduces to fitting a set of data $\{(x_i, y_i)\}_{i=1,\dots,m}$ by a linear combination of membership functions in the least squares sense, i.e. seeking a solution of the minimization problem

$$\sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} \alpha_j b_j(x_i; \mathbf{t}) \right)^2 = \min_{(\alpha, \mathbf{t})}, \tag{1.2}$$

where b_j represents the membership functions and $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)^T$ the coefficients. The concrete shape of the membership functions depends on the knot sequence $\mathbf{t}$, which is also included in the optimization procedure. Therefore, the minimization problem (1.2) is nonlinear.

Among the wide range of possible membership functions for Sugeno controllers, we will concentrate on two different kinds: trapezoidal and B-spline membership functions, firstly for the one-dimensional case (see Section 2). The more general class of B-spline membership functions for Sugeno controllers, including the often used triangular membership functions, were proposed in Zhang and Knoll [33].

1.2 Ill-posedness and Regularization

Assuming for the moment that the knot sequence $\mathbf{t}$ is fixed, we end up with a linear least squares problem

$$\frac{1}{2}\|\mathbf{y} - B(\mathbf{t})\alpha\|^2 = \min_{\alpha}, \tag{1.3}$$

where $B(\mathbf{t}) := (b_j(x_i, \mathbf{t}))_{i=1,\dots,m;\, j=1\dots,n}$ is the so-called observation matrix. (1.3) has a unique solution, if and only if the observation matrix B has full rank which is equivalent to the—in approximation theory well-known—Schoenberg-Whitney condition [6]. In case of a rank-deficient observation

matrix B (i.e., $r :=\text{rank}(B) < n$), the least squares problem (1.3) is no longer uniquely solvable. The set of solutions consists of the linear manifold

$$x^\dagger + N(B).$$

$x^\dagger$ denotes the unique least squares solution of minimal (Euclidean) norm, given by $x^\dagger = B^\dagger y$ ($B^\dagger$ the Moore-Penrose inverse or pseudo inverse) and $N(B)$ denotes the nullspace of B with dimension $n - r$. Because of roundoff errors any numerical scheme for computing the Moore-Penrose inverse of a matrix B will, at best, produce the Moore-Penrose inverse of a perturbed matrix $B + E$. However, it is well known that if a matrix $B + E$ is close to B, but is of different rank than B, then its Moore-Penrose inverse $(B + E)^\dagger$ will be different from $B^\dagger$, and the smaller E is, the greater the difference will be. Generally, problems the solution of which does not depend continuously on the data belong to the class of so-called *ill-posed problems*. In our case, we also have to take into account data errors. Usually, the data $\mathbf{y}$ is the result of measurements contaminated by noise. Often, the exact position x_i of the measurement is only known approximately, i.e. we get a set of noisy data $(\mathbf{x}^\gamma, \mathbf{y}^\delta)$ with error bounds γ and δ. Hence, we have to use so-called regularization techniques to obtain a stable solution to our problem.

We note that an analogous ill-posed problem arises in the problem of function approximation with neural networks. In this case the problem is also given by (1.2), the basis functions are usually of the form

$$b_j(x; \mathbf{a}, \mathbf{b}) = \sigma(a_j^T x + b_j), \tag{1.4}$$

with $a_j \in \mathbf{R}^N$ and $b_j \in \mathbf{R}$. The so-called activation function σ is usually chosen to be a *sigmoidal function*, i.e., a monotone and piecewise continuous function on $\mathbf{R}$, which satisfies

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

Similar to our problem in fuzzy control, the minimization is performed with respect to the weights and also with respect to the parameters a_j and b_j on which the output depends in a nonlinear way. The main difference is that in the approximation with neural networks one is not interested in the behaviour of the parameters a_j and b_j, since they do not have a particular meaning, but one rather wants to achieve convergence of the approximating output $f_n := \sum_{j=1}^n \alpha_j b_j(x; \mathbf{a}, \mathbf{b})$ to the function from which the samples y_i are taken. For this reason the results obtained in the sequel cannot be transferred directly to neural networks, but there are several techniques that could be carried over to that field in future work. For further details we refer the reader to the monograph by Bishop [4] and also to [5, 10, 25].

In the case of linear ill-posed problems, the regularization theory is very well developed [8]. The ill-posed problem is replaced with a family of similar

well-posed problems through the introduction of a regularization operator and a regularization parameter. For a problem to be well-posed, it must satisfy the requirements of existence, uniqueness, and the solution must depend continuously on the data. The regularization parameter is chosen dependent on the noise level and possibly on the data. The regularized solution approaches the true solution as the noiselevels tends to zero only if certain conditions upon the choice of the regularization parameter are satisfied [8].

It is shown by a simple example in Section 2, that the full nonlinear minimization problem (1.2) is indeed ill-posed in the sense that solutions do not necessarily depend on the data in a continuous way. Generally, the theory for nonlinear ill-posed problems (cf. [8], Chapter 10) involves more technical problems as the linear case. The case of an ill-posed nonlinear least squares problem, where no "attainability assumption" is fulfilled, is even more complicated and by far not so well developed [3].

Consider the nonlinear ill-posed problem

$$F(x) = y_0, \tag{1.5}$$

where F is an operator from a (subset of) a Hilbert space to another Hilbert space. We assume that a-priori information about a suitable solution of (1.5) has been incorporated into a vector x^*. The choice of x^* is very crucial; in the case of multiple solutions x^* plays the role of a selection criterion. We are searching for a x^* minimum-norm least squares solution, that is a least squares solution of (1.5) which minimizes the distance to x^* over all least squares solutions.

Among several regularization methods for obtaining a stable approximation to a x^* minimum-norm least squares solution, Tikhonov regularization is one of them. Minimizing the Tikhonov functional

$$J_{TIK}(x) := \|F(x) - y_0\|^2 + \beta\|x - x^*\|^2 \tag{1.6}$$

where $\beta > 0$ is the regularization parameter, is a trade-off between matching the data and stabilizing the solution. A large value of β produces a stable solution; however it may not adequately satisfy the original data. For a small value of β, we could expect to approximate the minimum of (1.5) well; however, the problem is then approaching the original ill-posed problem, and becomes unstable. In Section 3 we discuss the problem of choosing β appropriately.

Another possibility to stabilize an ill-posed problem is regularization with differential operators, i.e., we minimize the functional

$$J^L_{TIK}(x) := \|F(x) - y_0\|^2 + \beta\|Lx\|^2 \tag{1.7}$$

or even

$$J^L_{STATE}(x) := \|F(x) - y_0\|^2 + \beta\|L(F(x))\|^2, \tag{1.8}$$

where L represents a differential operator. Especially in classical approximation theory, instead of the spline approximation problem, a spline smoothing problem is often considered, where the smoothing term $L(F(x))$ characterizes the smoothness of the spline (cf. [6] for fixed knots). It should be mentioned that in practical applications, the smoothness of the controller output is one of the most important design requirements.

1.3 Approximation Properties of Sugeno Controllers

It has been shown by several authors ([14], [15], [29]), that fuzzy controllers are universal approximators in the sense that it is possible to construct such rule bases that approximate uniformly any continuous function defined on a compact subset of $\mathbf{R}^m$ with arbitrary accuracy. Proofs are based upon the Stone-Weierstrass Theorem and purely existential in nature. From a practical—fuzzy control oriented—point of view, these theorems suffer from the fact that the number of rules in the base is not bounded, in addition to that even the supports of the terms in the rules are not bounded (e.g. Gaussian membership functions).

As already mentioned, the tuning of a Sugeno controller reduces to a data fitting problem by a linear combination of membership functions. From a purely mathematical point of view, we now let both the number of membership functions and data points tend to infinity and examine the approximation power. We consider the case of B-spline membership functions, where a wide range of convergence results ([6], [24]) exists.

It is well known, that sequences of polynomials interpolating a predetermined sequence of points in an interval $[a, b]$ may not converge (Theorem of Faber, [24]). If polynomials are forced to follow points in an interval, they may respond by oscillating wildly. This tendency to oscillate becomes increasingly pronounced as the order of the polynomial is increased. The situation is completely different for spline approximation. Here, we are not interested in large order polynomials. To get accurate approximations we would prefer to keep the order of piecewise polynomials between two knots fixed at a rather low value, and increase the number of knots. Not surprisingly, convergence results were obtained for wide classes of functions relating the approximation error to the number of knots or, in our case, to the number of membership functions. Especially, we can approximate a large class of functions arbitrarily well by splines of a fixed order if we are willing to use many knots. The order of approximation attainable will increase with the smoothness of the class of functions being approximated. Additionally, it will turn out, that substantial gains in the rate of convergence can be achieved when using the knots as free parameters that can be adjusted to the particular function being approximated (cf. [24], Chapters 6-7).

Let $W_p^k[a, b] := \{u : u^{(r)} \in L_p[a, b]; \forall\, r \leq k\}$ denote a Sobolev space of order k, p on $[a, b]$, $L_p[a, b]$ the Lebesgue space of order p, $\|\cdot\|_{W_p^k}$ the usual

Sobolev norm, $d(f, S_{m,\ell})_X := \inf_{s \in S_{m,\ell}} \|f - s\|_X$ the distance of a function f to the spline space $S_{m,\ell}$ defined by ℓ knots and piecewise polynomials of order m. Assuming a fixed knot sequence (equally spaced knots, for example), Shumaker [24] cites the following convergence rate result, where $1 \leq p \leq q \leq \infty$:

$$\forall\ (1 \leq s \leq m) \quad \forall\ f \in W_p^s[a,b] \quad d(f, S_{m,\ell})_{L_q[a,b]} = \mathcal{O}(\ell^{-(s+1/q-1/p)}). \tag{1.9}$$

When using free knots

$$\forall\ (1 \leq s \leq m) \quad \forall\ f \in W_1^s[a,b] \quad d(f, S_{m,\ell})_{L_q[a,b]} = \mathcal{O}(\ell^{-s}). \tag{1.10}$$

Indeed, it was shown that order m convergence is the maximum that can be obtained for smooth functions. Consider the uniform norm ($q = \infty$). Then, using splines with equally spaced knots, for example, the maximal order of convergence is $m - 1/p$. On the other hand, a free knot sequence leads to maximal convergence rate m.

1.4 Organization of the paper

The paper is organized as follows: In Section 2, we give mathematically precise definitions of a Sugeno controller and membership functions. The optimization of a Sugeno controller is treated as a nonlinear least squares problem, as not only the coefficients but also the position of knots defining the shape of membership functions is sought for. It is also shown that solving the minimization problem is indeed ill-posed by a rather simple finite dimensional example. In Section 3, two different approaches of regularizing the least squares problem are investigated: the first one is spline smoothing—commonly used in the area of spline approximation—where additional constraints are introduced to avoid coalescing knots; the other one is the classical Tikhonov regularization. We develop existence, stability, and convergence results. Finally, in Section 4 we give a short description of the numerical optimization algorithm—a generalized Gauss-Newton like algorithm—and prove that the results of reconstructing a-priori given functions from noisy data are in agreement with theoretical results obtained in Section 3.

2 Optimization of Sugeno Controllers

2.1 Basic definitions of Sugeno controller and membership functions

If we look at a Sugeno controller from the point of view of mappings which assign to each crisp observation a crisp value (vector) in the output space,

i.e., there is a function $F_s : X \to \mathbf{R}^{d_o}$ associating to each input x its corresponding output y, it is possible to construct an explicit formula substituting the fuzzy control system completely.

Definition 2.1. Let X be an input space, let $A_1, A_2, \ldots, A_n$ be normalized fuzzy subsets of X with $\sum \mu_{A_i}(x) > 0$ for all $x \in X$, and $f_1, f_2, \ldots, f_n$ be functions from X to $\mathbf{R}^{d_o}$, and consider the rulebase ($i = 1, 2, \ldots, n$)

$$\text{if } x \text{ is } A_i \text{ then } u = f_i(x).$$

Then the Sugeno controller defines the following input-output function $F_s : X \to \mathbf{R}^{d_o}$

$$F_s(x) = \frac{\sum \mu_{A_i}(x)\, f_i(x)}{\sum \mu_{A_i}(x)}. \tag{2.1}$$

In the following we consider the special case, that for $i = 1, 2, \ldots, n$ the functions f_i are constant, that is $f_i(x) \equiv \alpha_i$. In a first step, we restrict ourselves to the one-dimensional case, i.e., a single input-single output controller. However, for the output variable this is no restriction. If the number of output variables is higher than one, it can easily be shown that in every case it is possible to decompose the controller into as many independent controllers as many output variables we have [14].

Among the class of membership functions, we consider first the classical trapezoidal ones. Let the knot sequence $\mathbf{t} = \{t_i\}$, where

$$a = t_1 \le t_2 \le \ldots \le t_{2n-1} \le t_{2n} = b \tag{2.2}$$

be a partition of the universe of an input variable defined over $[a, b]$, corresponding to n linguistic terms. Then the mathematical formulation of the trapezoidal membership functions b_j ($j \in \{1, \ldots, n\}$) is as follows:

$$b_1(x, \mathbf{t}) := \begin{cases} 1 & \text{if } x \in [t_1, t_2] \\ \frac{-x+t_3}{t_3-t_2} & \text{if } x \in (t_2, t_3) \\ 0 & \text{otherwise} \end{cases}$$

$$b_j(x, \mathbf{t}) := \begin{cases} \frac{x-t_{2j-2}}{t_{2j-1}-t_{2j-2}} & \text{if } x \in (t_{2j-2}, t_{2j-1}) \\ 1 & \text{if } x \in [t_{2j-1}, t_{2j}] \\ \frac{-x+t_{2j+1}}{t_{2j+1}-t_{2j}} & \text{if } x \in (t_{2j}, t_{2j+1}) \\ 0 & \text{otherwise} \end{cases}$$

$$b_n(x, \mathbf{t}) := \begin{cases} \frac{x-t_{2n-2}}{t_{2n-1}-t_{2n-2}} & \text{if } x \in (t_{2n-2}, t_{2n-1}) \\ 1 & \text{if } x \in [t_{2n-1}, t_{2n}] \\ 0 & \text{otherwise} \end{cases}$$

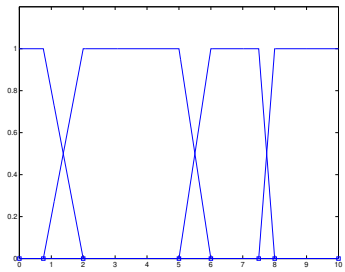

Figure 1: Trapezoidal membership functions

Figure 1 shows a typical example.

Now we turn to the more general class of B-spline membership functions for Sugeno controllers.

Assume that x is an input variable of a Sugeno controller that is defined on the interval $[a, b]$. Given a sequence of ordered knots $\mathbf{t} = \{t_i\}$, where

$$t_1 = \ldots = t_k = a < t_{k+1} \leq ... \leq t_n < b = t_{n+1} = \ldots = t_{n+k} \tag{2.3}$$

the j-th normalized B-spline basis function $B_{j,k}$ of order k for the knot sequence $\mathbf{t}$ is recursively defined as

$$B_{j,1}(x, \mathbf{t}) := \begin{cases} 1 & \text{if } t_j \leq x < t_{j+1}, \\ 0 & \text{otherwise} \end{cases}$$

$$B_{j,k}(x, \mathbf{t}) := \omega_{j,k}(x)\, B_{j,k-1}(x, \mathbf{t}) + (1 - \omega_{j+1,k}(x))\, B_{j+1,k-1}(x, \mathbf{t}) \qquad \text{for } k > 1$$

where

$$\omega_{j,k}(x) := \begin{cases} \frac{x - t_j}{t_{j+k-1} - t_j} & \text{if } t_j < t_{j+k-1}, \\ 0 & \text{otherwise} \end{cases}$$

The complete knots consist of two parts, the interior knots that lie within the universe of discourse, and extended knots that are generated at both ends of the universe for a unified definition of B-splines (leading to the so-called marginal linguistic terms in [33]).

In the following, we summarize some properties of B-splines, where especially the positivity, local support, and partition of unity qualify them as membership functions.

- Positivity: $B_{j,k}(x, \mathbf{t}) \geq 0 \qquad \text{for all} \qquad x \in [a, b]$
- Local support: $B_{j,k}(x, \mathbf{t}) = 0 \qquad \text{if} \qquad x \notin [t_j, t_{j+k})$
- C^{k-2} continuity: if the knots $t_k, \ldots t_{n+1}$ are pairwise different from each other, then $B_{j,k}(x, \mathbf{t})$ is $(k-2)$ times continuously differentiable.

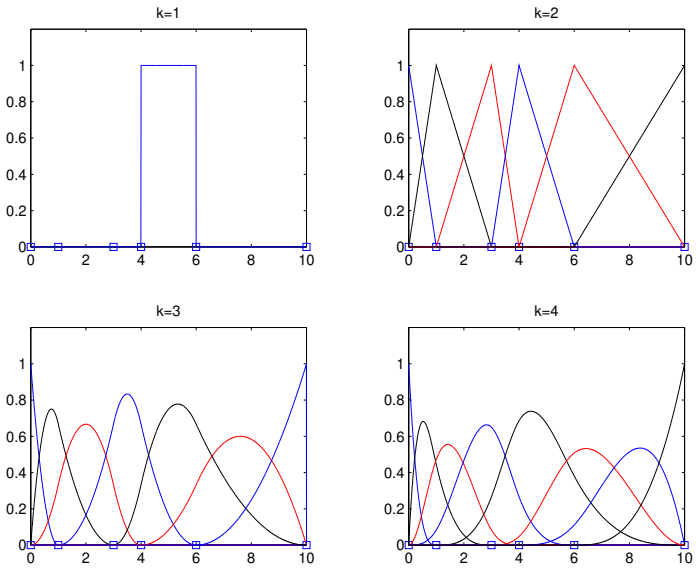

Figure 2: B-spline basis functions of order 1 - 4 for a non-uniform knot sequence

- Partition of unity:

$$\sum_{j=1}^{n} B_{j,k}(x, \mathbf{t}) = 1 \tag{2.4}$$

From the point of view of fuzzy control, B-spline membership functions suffer from the drawback that they are not—except for orders less than three—normalized membership functions, i.e. the largest membership grade is not necessarily one, but a smaller value in the interval $[0, 1)$. Additionally for higher order B-splines, the linguistic interpretation of membership degree is rather complicated.

2.2 Tuning of Sugeno controllers as an ill-posed least squares problem

Assume $(\mathbf{x}, \mathbf{y})$ is a set of so-called training data, where $\mathbf{x} = (x_1, x_2, \ldots, x_m)^T$ is the training data vector, and $\mathbf{y} = (y_1, y_2, \ldots, y_m)^T$ the desired output for $\mathbf{x}$. It follows immediately from (2.1), and the partition of unity (2.4), that designing a Sugeno controller from training data, is then equivalent to the least squares problem

$$\sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} \alpha_j b_j(x_i; \mathbf{t}) \right)^2 = \min_{((\alpha_1, \alpha_2, \ldots, \alpha_n), \mathbf{t}) \in \mathbf{R}^n \times [a,b]^\ell}, \tag{2.5}$$

where $(b_j)_{j=1,\dots,n}$ is one of the membership functions introduced above. The concrete shape of the membership functions is determined by the ℓ-dimensional knot vector $\mathbf{t}$. ℓ represents the number of free knots.

As already mentioned, we have to consider data errors in $\mathbf{y}$ and $\mathbf{x}$, i.e., instead

$$\begin{aligned} \|\mathbf{x} - \mathbf{x}^\gamma\|_{\ell^2} &\leq \gamma & (2.6)\\ \|\mathbf{y} - \mathbf{y}^\delta\|_{\ell^2} &\leq \delta, & (2.7) \end{aligned}$$

where $\|\mathbf{x}\|_{\ell^2} := \sqrt{\sum_{i=1}^m x_i^2}$ denotes the usual ℓ_2 norm.

The following example shows that the problem of finding a minimum to (2.5) is ill-posed, even if we have complete information about the function f, from which the samples y are taken.

Example 2.2. Let $n = 2$, $k \in \mathbf{N}$, $k \geq 2$, $a = t_1^k = 0$, $t_2^k = k^{-3}$ and $t_3^k = 2k^{-3}$, $t_4^k = b = 1$, and choose $\alpha_1^k = k$, $\alpha_2^k = 0$. The fuzzy membership functions b_1 and b_2 are defined by

$$\begin{aligned} b_1(x;\mathbf{t}) &= \begin{cases} 1 & \text{if } x \leq t_2 \\ \frac{t_3 - x}{t_3 - t_2} & \text{if } t_2 < x < t_3 \\ 0 & \text{if } t_3 \leq x \end{cases} & (2.8)\\ b_2(x;\mathbf{t}) &= 1 - b_1(x;\mathbf{t}). & (2.9) \end{aligned}$$

Then $f^k = \alpha_1^k b_1(x;\mathbf{t}^k) + \alpha_2^k b_2(x;\mathbf{t}^k)$ converges to zero in $L_2([0,1])$, but α^k has no bounded subsequence. Hence, the optimization problem is unstable with respect to perturbations in the data.

2.3 Tuning of Sugeno controllers - The multiple input single output case

Under the assumptions of the previous sections, the input-output function F_s of a Sugeno controller with d-dimensional input variable is given by

$$\begin{aligned} &F_s(x_1, x_2, \dots, x_d) = \\ &\alpha_{j_1,j_2,\dots,j_d} \sum_{j_1=1}^{n_1} \sum_{j_2=1}^{n_2} \dots \sum_{j_d=1}^{n_d} b_{j_1}(x_1, \mathbf{t}_1) \cdot b_{j_2}(x_2, \mathbf{t}_2) \cdot \dots \cdot b_{j_d}(x_d, \mathbf{t}_d). \quad (2.10) \end{aligned}$$

F_s represents a $d-$dimensional tensor product spline. Only, if the data is given on a regular grid (e.g. a rectangular grid in the 2D case), then the $d-$dimensional tuning problem splits up into d one-dimensional problems. For irregular data, it is hard to define a Schoenberg-Whitney like condition; practical examples show that the observation matrix B is very often rank-deficient. Hence, regularization is strongly recommendable or even a must.

3 Regularization

In the following we want to investigate two different approaches to the regularization of the least squares problem (2.5). The first one is a common method for spline approximation (cf e.g. [21]) and consists of replacing (2.5) by

$$\sum_{i=1}^{m}\left(y_i-\sum_{j=1}^{n}\alpha_j b_j(x_i;\mathbf{t})\right)^2+\beta|\sum_{j=1}^{n}\alpha_j b_j(.;\mathbf{t})|^2_{H^k(\Omega)}=\min_{(\alpha,\mathbf{t})}, \tag{3.1}$$

where $|.|_{H^k(\Omega)}$ denotes the norm or seminorm in the Sobolev space $H^k(\Omega) := W_2^k(\Omega)$. In addition we impose the constraints

$$t_{j+1}-t_j\geq\epsilon, \qquad j=1,\ldots,\ell-1, \tag{3.2}$$

which are necessary to remove the possible instabilities caused by two equal or almost equal knots. For notational simplicity, we do not bother with multiple knots at the end of the intervals (cf. the definition of the knot sequences (2.2), (2.3)). In the subsequent Section 3.1 we will see that (3.1) subject to (3.2) is a well-posed problem and its solution will converge to a minimizer of the original problem with the additional constraint (3.2) for fixed ϵ and appropriately chosen $\beta\to 0$ as $\gamma,\delta\to 0$. However, we cannot show convergence as $\epsilon\to 0$, which is a serious disadvantage.

The second approach under investigation is classical Tikhonov regularization in the parameter space $\mathbf{R}^n\times\mathbf{R}^\ell$, it consists of minimizing the functional

$$\sum_{i=1}^{m}\left(y_i^\delta-\sum_{j=1}^{n}\alpha_j b_j(x_i^\gamma;\mathbf{t})\right)^2+\beta_1\sum_{j=1}^{n}\alpha_j^2+\beta_2\sum_{j=1}^{\ell}(t_j-t_j^*)^2=\min_{(\alpha,\mathbf{t})} \tag{3.3}$$

for appropriately chosen β_1 and β_2 (in dependence of δ and $\mathbf{y}^\delta$), where $\mathbf{t}^*$ is a prior for $\mathbf{t}$, e.g. the uniform grid points. In this case we can show convergence for appropriate choice of $\beta_1\to 0$ as the noise level tends to zero even for $\beta_2=0$.

In both cases we will assume that the functions b_j satisfy the Lipschitz-estimate

$$|b_j(x,\mathbf{t})-b_j(\tilde{x},\mathbf{t})|\leq L|x-\tilde{x}|, \qquad \forall\ x,\tilde{x}, \quad \forall\mathbf{t}\in[a,b]^\ell$$

with some nonnegative real constant L.

3.1 Smoothing

Now we turn our attention to the stabilized problem (3.1) supplemented by (3.2). For the sake of simplicity we restrict our analysis to the case of

$\Omega = (0,1)$, trapezoidal functions b_j and the H^1-norm defined by

$$\|u\|^2_{H^1(\Omega)} = \int_\Omega \left(|u|^2 + |\nabla u|^2\right)\ dx$$

as the stabilizer. Obviously, the number of inner grid points must be even in this case to ensure that the output equals one in the intervals $(0, t_1)$ and $(t_\ell, 1)$. The number of basis functions is then given by $n = \frac{\ell}{2} + 1$. We note that a similar but technically much more complicated reasoning is possible for other spline basis functions, but the technical details would shadow the basic concepts. Therefore, they are omitted here.

The stabilizing term can be transformed to a bilinear expression in terms of the variable α via

$$\|\sum_{j=1}^{n} \alpha_j b_j(.;\mathbf{t})\|^2_{H^1(\Omega)} = \alpha^T A(\mathbf{t})\alpha + \alpha^T B(\mathbf{t})\alpha, \tag{3.4}$$

where the symmetric, positive definite matrices $A(\mathbf{t})$ and $B(\mathbf{t})$ are defined by

$$A(\mathbf{t}) = \left(\int_0^1 b_i(x;\mathbf{t})b_j(x;\mathbf{t})\ dx\right)_{i,j=1,\dots,n} \tag{3.5}$$

$$B(\mathbf{t}) = \left(\int_0^1 b_i'(x;\mathbf{t})b_j'(x;\mathbf{t})\ dx\right)_{i,j=1,\dots,n}. \tag{3.6}$$

Now we define a new grid s_j, which does not include the intervals (t_{2j}, t_{2j+1}), on which $b_i' = 0$ for all i, more precisely,

$$s_1 = t_1, \quad s_{j+1} = s_j + t_{2j} - t_{2j-1}, j = 1, \dots, \frac{\ell}{2}. \tag{3.7}$$

This allows us to find an equivalent definition for the matrix $B(\mathbf{t})$:

Lemma 3.1. *Let $\{\phi_j\}_{j=1,\dots,\frac{\ell}{2}}$ denote the usual piecewise affinely linear finite elements on the grid $\{s_j\}_{j=1,\dots,\frac{\ell}{2}}$, i.e.,*

$$\phi_j|_{(s_i,s_{i+1})} \text{ is affinely linear }, \quad \phi_j(s_i) = \delta_{ij}, \forall\ i, j,$$

where δ_{ij} denotes the Kronecker delta symbol. Then the matrix $\tilde{B}(\mathbf{t})$ defined by

$$\tilde{B}(\mathbf{t}) = \left(\int \phi_i'(s)\phi_j'(s)\ ds\right)_{i,j=1,\dots,\frac{\ell}{2}}$$

equals $B(\mathbf{t})$ defined by (3.6). Furthermore, the matrix $A(\mathbf{t})$ can be represented in the form

$$A(\mathbf{t}) = \left(\int \phi_i(s)\phi_j(s)\ ds\right)_{i,j=1,\dots,\frac{\ell}{2}} + A_0(\mathbf{t}), \tag{3.8}$$

where $A_0(\mathbf{t})$ is a positive semidefinite matrix.

Proof. Since $b_j' = 0$ on (t_{2i}, t_{2i+1}) for all i, j and $b_j'(x;\mathbf{t}) = \phi_j'(S_i(x))$ on (t_{2i-1}, t_{2i}), where S_i is the unique transformation of the form $S_i(x) = x + \sigma_i$ that maps (t_{2i-1}, t_{2i}) onto (s_i, s_{i+1}) we obtain

$$\int_0^1 b_i'(x;\mathbf{t}) b_j'(x;\mathbf{t})\, dx = \int \phi_i'(s)\phi_j'(s)\, ds$$

and consequently $\tilde{B}(\mathbf{t}) = B(\mathbf{t})$.

An analogous argument yields the decomposition

$$A(\mathbf{t}) = \left(\int \phi_i(s)\phi_j(s)\, ds \right)_{i,j=1,\dots,n} + \left(\int_{(0,1)-S} b_i(s) b_j(s)\, ds \right)_{i,j=1,\dots,n},$$

where $S = \bigcup (t_{2i-1}, t_{2i})$. We now define $A_0(\mathbf{t})$ as the second term in the previous identity and since

$$b_i(s)b_j(s) = \begin{cases} 1 & \text{if } i = j,\ s \in (t_{2i-2}, t_{2i-1}) \\ 0 & \text{else for } s \in (0,1) - S \end{cases},$$

A_0 is a diagonal matrix with nonnegative entries and therefore positive semidefinite. □

To carry out the stability analysis we will use the following result adapted from stability estimates in finite element theory:

Lemma 3.2. *For each $c_0 > 0$ there exists a positive real number c_1 such that for all $\mathbf{t}$ satisfying*

$$\inf_{j \in \{1,\dots,\ell-1\}} \{t_{j+1} - t_j\} \geq \frac{c_0}{\ell}$$

the estimate

$$\sum_{j=1}^n \alpha_j^2 \leq c_1 \ell \| \sum_{j=1}^n \alpha_j b_j(x;\mathbf{t}) \|^2_{H^1(\Omega)} \tag{3.9}$$

holds.

Proof. Lemma 3.1 and (3.4) yield the identity

$$\| \sum_{j=1}^n \alpha_j b_j(x;\mathbf{t}) \|^2_{H^1(\Omega)} = \alpha^T \Phi \alpha + \alpha^T A_0(\mathbf{t}) \alpha \geq \alpha^T \Phi \alpha,$$

where

$$\Phi = \left(\int \left[\phi_i(s)\phi_j(s) + \phi_i'(s)\phi_j'(s) \right] ds \right)_{i,j=1,\dots,n}.$$

A standard argument from finite-element theory (cf. [26]) implies that the minimal eigenvalue of the symmetric matrix Φ is bounded below by $c_1\ell$, where c_1 depends only on $\frac{c_0}{2}$, which is a lower bound for the length of the interval (s_1, s_n). □

Now we are able to show that the stabilized problem 3.1 is well-posed, i.e., a minimizer exists and the dependence of the minimizers on the data is stable (in a set-valued way), which is expressed in the following propositions:

Proposition 3.3 (Existence of a minimizer). *For all* $\mathbf{y} \in \mathbf{R}^m$ *and* $\mathbf{x} \in [0,1]^m$ *there exists a minimizer of* (3.1)*, if* $\epsilon > 0$ *and* $\beta > 0$*.*

Proof. Since a minimizer must yield an output less or equal than the one from $\alpha = 0$, we may add the additional constraint (using Lemma 3.2, $\epsilon = \frac{c_0}{\ell}$ and the notation $C = \frac{c_1}{c_0}$)

$$\sum_{j=1}^{n} \alpha_j^2 \leq \frac{C}{\beta\epsilon} \sum_{i=1}^{m} y_i^2.$$

The resulting set of admissible points is compact in $\mathbf{R}^n \times \mathbf{R}^\ell$ and since the objective functional is continuous, the existence of a minimizer follows from a standard principle in optimization. □

Proposition 3.4 (Stability). *Let* $\beta > 0$*,* $\epsilon > 0$*,* $\mathbf{y}^k \to \mathbf{y}$ *and* $\mathbf{x}^k \to \mathbf{x}$*. Then the according sequence of minimizers of* (3.1) *has a convergent subsequence and the limit of every convergent subsequence is a minimizer of* (3.1)*.*

Proof. As in the proof of Proposition 3.3 we obtain the estimate

$$\sum_{j=1}^{n} |\alpha_j^k|^2 \leq \frac{C}{\beta\epsilon} \sum_{i=1}^{m} |y_i^k|^2.$$

Consequently, the sequence $(\alpha^k, \mathbf{t}^k)$ is bounded, which implies the existence of a convergent subsequence. Let $(\alpha^{k_\ell}, \mathbf{t}^{k_\ell})$ be a convergent subsequence with limit $(\alpha, \mathbf{t})$, then the continuity of the objective functional together with the definition of $(\alpha^k, \mathbf{t}^k)$ implies that $(\alpha, \mathbf{t})$ is a minimizer of (3.1). □

Finally, we want to investigate the question of convergence of minimizers of the regularized problem as the noise level (γ, δ) and the regularization parameter β tend to zero. Of course, it would be of interest to let ϵ tend to zero, too, but in this case one cannot guarantee that the minimizers are uniformly bounded.

Theorem 3.5 (Convergence under Constraints). *Let* $\epsilon > 0$ *be fixed, let* (γ^k, δ^k) *be a monotone sequence convergent to* $(0,0)$ *and let* $(x^{\gamma^k}, y^{\delta^k})$ *be a corresponding data sequence satisfying* (2.6)*,* (2.7) *with* $(\gamma, \delta) = (\gamma^k, \delta^k)$*. Moreover, let the regularization parameter* β^k *be chosen such that*

$$\beta^k \to 0, \qquad \frac{\max\{\gamma^k, \delta^k\}}{\beta^k} \to 0.$$

If a minimizer of (2.5) *with exact data exists, then each sequence of minimizers* $(\alpha^k, \mathbf{t}^k)$ *of* (3.1), (3.2) *with noisy data* $(x^{\gamma^k}, y^{\delta^k})$ *and* $\beta = \beta^k$ *has a convergent subsequence and the limit of each convergent subsequence is a minimizer of the least squares problem* (2.5) *subject to* (3.2).

Proof. Let $(\hat{\alpha}, \hat{\mathbf{t}})$ be a minimizer of the problem with exact data, then the definition of $(\alpha^k, \mathbf{t}^k)$ implies

$$\begin{aligned}
&\sum_{i=1}^{m} \left(y_i^{\delta^k} - \sum_{j=1}^{n} \alpha_j^k b_j(x_i^{\gamma^k}, \mathbf{t}^k) \right)^2 + \beta^k \frac{\epsilon}{C} \sum_{j=1}^{n} (\alpha_j^k)^2 \\
&\quad \le \sum_{i=1}^{m} \left(y_i^{\delta^k} - \sum_{j=1}^{n} \alpha_j^k b_j(x_i^{\gamma^k}, \mathbf{t}^k) \right)^2 + \beta^k (\alpha^k)^T [A(\mathbf{t}^k) + B(\mathbf{t}^k)] \alpha^k \\
&\quad \le \sum_{i=1}^{m} \left(y_i^{\delta^k} - \sum_{j=1}^{n} \hat{\alpha}_j b_j(x_i^{\gamma^k}, \hat{\mathbf{t}}) \right)^2 + \beta^k \hat{\alpha}^T [A(\hat{\mathbf{t}}) + B(\hat{\mathbf{t}})] \hat{\alpha} \\
&\quad \le \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} \hat{\alpha}_j b_j(x_i, \hat{\mathbf{t}}) \right)^2 + c_1(\delta^k + L\|\hat{\alpha}\|_{\ell^1} \gamma^k) + c_2 \beta^k \sum_{j=1}^{n} \hat{\alpha}_j^2
\end{aligned}$$

for some constants c_1, c_2. The noisy residual can be estimated by

$$\begin{aligned}
&\sum_{i=1}^{m} \left(y_i^{\delta^k} - \sum_{j=1}^{n} \alpha_j^k b_j(x_i^{\gamma^k}, \mathbf{t}^k) \right)^2 \\
&\quad \ge \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} \alpha_j^k b_j(x_i, \mathbf{t}^k) \right)^2 - c_1(\delta^k + L\|\alpha^k\|_{\ell^1} \gamma^k) \\
&\quad \ge \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} \hat{\alpha}_j b_j(x_i, \hat{\mathbf{t}}) \right)^2 - c_1(\delta^k + L\|\alpha^k\|_{\ell^1} \gamma^k),
\end{aligned}$$

and hence,

$$\sum (\alpha_j^k)^2 \le \frac{c_1 C}{\epsilon \beta_k} (2\delta^k + L[\|\hat{\alpha}\|_{\ell^1} + \|\alpha^k\|_{\ell^1}]\gamma^k) + \frac{c_2 C}{\epsilon} \sum \hat{\alpha}_j^2.$$

Finally, with the standard estimate $\|\alpha^k\|_{\ell^1} \le \sqrt{n}\|\alpha^k\|_{\ell^2}$ we conclude that

$$\sum (\alpha_j^k)^2 \le \frac{c_1 C}{\epsilon} \left(4\frac{\delta^k}{\beta_k} + 2L\|\hat{\alpha}\|_{\ell^1} \frac{\gamma^k}{\beta^k} \right) + L^2 \frac{c_1^2 C^2}{\epsilon^2} \left(\frac{\gamma^k}{\beta^k} \right)^2 + 2\frac{c_2 C}{\epsilon} \sum \hat{\alpha}_j^2,$$

which implies

$$\limsup \sum (\alpha_j^k)^2 \le 2\frac{c_2 C}{\epsilon} \sum \hat{\alpha}_j^2.$$

Thus, the sequence $(\alpha_j^k, \mathbf{t}^k)$ is bounded and therefore there exists a convergent subsequence. The fact that the limit of a convergent subsequence is a mininizer of (2.5) follows from

$$\limsup \sum_{i=1}^{m} \left(y_i^{\delta^k} - \sum_{j=1}^{n} \alpha_j^k b_j(x_i^{\gamma^k}, \mathbf{t}^k) \right)^2 \leq \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} \hat{\alpha}_j b_j(x_i, \hat{\mathbf{t}}) \right)^2 .$$

□

3.2 Tikhonov Regularization

In this section we investigate the Tikhonov regularization applied to (2.5), i.e., the minimization problem (3.3). We restrict our attention again to the case $\Omega = (0,1)$, but we note that the method and all proofs can be carried out in the same way (but with vectors t_j). In the general theory (cf. e.g. [3, 8, 9]), the existence of a minimizer of problem (3.3) can be shown if $\beta_1 > 0$ and $\beta_2 > 0$. In our special case, the positivity of the second regularization parameter β_2 is not necessary to guarantee the existence as we will show in the following proposition:

Proposition 3.6 (Existence of a minimizer). *For all $\mathbf{y} \in \mathbf{R}^m$ and $\mathbf{x} \in [0,1]^m$ there exists a minimizer of (3.3), if $\beta_1 > 0$.*

Proof. As in the proof of Proposition 3.3 it suffices to show that the set of admissible α can be restricted to a compact set by an a-priori estimate. Again a comparison with the output functional at the point $\alpha = 0$, we may conclude that a minimizer $(\alpha, \mathbf{t})$ of (3.3) must satisfy

$$\sum_{j=1}^{n} \alpha_j^2 \leq \frac{1}{\beta_1} \sum_{i=1}^{m} |y_i^\delta|^2.$$

□

We note that the stability and convergence analysis of Tikhonov regularization with respect to the perturbation in the output $\mathbf{y}$ can be carried over directly from [8, 22]. Since we are also interested in perturbations in the positions $\mathbf{x}$, we need some modifications, which we will prove in the following:

Proposition 3.7 (Stability). *Let $\beta_1 > 0$, $\mathbf{y}^k \to \mathbf{y}$ and $\mathbf{x}^k \to \mathbf{x}$. Then the according sequence of minimizers $(\alpha^k, \mathbf{t}^k)$ of (3.3) has a convergent subsequence and the limit of every convergent subsequence is a minimizer of (3.3).*

Proof. Again we compare the value of the objective functional achieved at $(\alpha^k, \mathbf{t}^k)$ for the data $\mathbf{x}^k$ and $\mathbf{y}^k$ with the one achieved with $(\mathbf{0}, \mathbf{t}^k)$ and obtain the a-priori estimate

$$\sum_{j=1}^{n} \alpha_j^2 \leq \frac{1}{\beta_1} \sum_{i=1}^{m} |y_i^k|^2.$$

Since $\mathbf{y}^k \to \mathbf{y}$, the right-hand side is uniformly bounded as $k \to \infty$ and therefore the set of minimizers is bounded, which implies the existence of a weakly convergent subsequence.

A convergent subsequence (without restriction of generality $(\alpha^k, \mathbf{t}^k)$ itself and limit $(\overline{\alpha}, \overline{\mathbf{t}})$) satisfies

$$\begin{aligned}
&\sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} \overline{\alpha}_j b_j(x_i; \overline{\mathbf{t}}) \right)^2 + \beta_1 \sum_{j=1}^{n} \overline{\alpha}_j^2 + \beta_2 \sum_{j=1}^{\ell} (\overline{t}_j - t_j^*)^2 \\
&\quad \leq \liminf \sum_{i=1}^{m} \left(y_i^k - \sum_{j=1}^{n} \alpha_j^k b_j(x_i^k; \mathbf{t}^k) \right)^2 + \beta_1 \sum_{j=1}^{n} |\alpha_j^k|^2 + \beta_2 \sum_{j=1}^{\ell} (t_j^k - t_j^*)^2 \\
&\quad \leq \liminf \sum_{i=1}^{m} \left(y_i^k - \sum_{j=1}^{n} \alpha_j b_j(x_i^k; \mathbf{t}) \right)^2 + \beta_1 \sum_{j=1}^{n} |\alpha_j|^2 + \beta_2 \sum_{j=1}^{\ell} (t_j - t_j^*)^2 \\
&\quad = \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} \alpha_j b_j(x_i; \mathbf{t}) \right)^2 + \beta_1 \sum_{j=1}^{n} |\alpha_j|^2 + \beta_2 \sum_{j=1}^{\ell} (t_j - t_j^*)^2
\end{aligned}$$

for all admissible $(\alpha, \mathbf{t})$ and thus, the limit is again a minimizer of (3.3). □

The convergence result in this case holds for the full problem (2.5), not only for a constrained version:

Theorem 3.8 (Convergence). *Assume that a minimizer of problem* (3.3) *exists. Moreover, let* (γ^k, δ^k) *be a sequence converging to* $(0, 0)$ *and denote by* $(\alpha^k, \mathbf{t}^k)$ *the according sequence of minimizers of* (3.3) *with data* $(\mathbf{x}^\gamma, \mathbf{y}^\delta)$*, satisfying* (2.6), (2.7)*. Then* $(\alpha^k, \mathbf{t}^k)$ *has a convergent subsequence and the limit of every convergent subsequence is a minimizer of* (3.3) *with exact data* $(\mathbf{x}, \mathbf{y})$ *if the regularization parameters satisfy*

$$\beta_1^k \to 0, \qquad \beta_2^k \to 0 \tag{3.10}$$

$$\frac{\max\{\gamma^k, \delta^k\}}{\beta_1^k} \to 0 \tag{3.11}$$

$$\exists \epsilon > 0 : \frac{\beta_1^k}{\beta_2^k} \geq \epsilon. \tag{3.12}$$

Proof. By similar reasoning to the proof of Theorem 3.5 we can deduce that

$$\limsup \sum (\alpha_j^k)^2 \leq \sum \hat{\alpha}_j^2 + \limsup \frac{\beta_2^k}{\beta_1^k} \sum (\hat{t}_j - t_j^*)^2$$

for a minimizer $(\hat{\alpha}, \hat{\mathbf{t}})$ of (2.5). The remaining steps of the proof are the same as for Theorem 3.5. □

Finally, we want to investigate the rate of convergence of the regularized solutions as $\delta \to 0$. For this sake we need additional smoothness of the *parameter-to-output map*, which we will define and analyze in the following Lemma:

Lemma 3.9. *Let $b_j \in C([0,1]^{\ell+1})$ for all $j \in \{1, \dots, n\}$, then the nonlinear parameter-to-output operator F defined by*

$$\begin{array}{rccc} F: & \mathbf{R}^n \times [0,1]^\ell & \to & \mathbf{R}^m \\ & (\alpha, \mathbf{t}) & \mapsto & \left(\sum_{j=1}^n \alpha_j b_j(x_i; \mathbf{t}) \right)_{i=1,\dots,m} \end{array} \tag{3.13}$$

is continuous. Moreover, if the partial derivatives $\frac{\partial b_j}{\partial t_k}$ exist and are continuous functions for all $j \in \{1, \dots, n\}$, $k \in \{1, \dots, \ell\}$, then F is continuously Frèchet-differentiable with partial derivatives

$$\frac{\partial}{\partial \alpha_k} F(\alpha, \mathbf{t}) = (b_k(x_i; \mathbf{t}))_{i=1,\dots,m} \tag{3.14}$$

$$\frac{\partial}{\partial t_k} F(\alpha, \mathbf{t}) = \left(\sum_{j=1}^n \alpha_j \frac{\partial b_j}{\partial t_k}(x_i; \mathbf{t}) \right)_{i=1,\dots,m}. \tag{3.15}$$

If the partial derivatives above are all Lipschitz-continuous, then F' is Lipschitz-continuous, too.

For convergence rates, we restrict our attention to the case of $\gamma = 0$, which enables the application of the standard theory of Tikhonov regularization. As usual for ill-posed problems, the convergence can be arbitrarily slow in general (cf. e.g. [19]), rates can only be achieved under additional conditions on the solution. A standard condition of this kind is the *source condition*

$$\exists\ \mathbf{w} \in \mathbf{R}^m : \ (\overline{\alpha}, \overline{\mathbf{t}} - \mathbf{t}^*) = F'(\overline{\alpha}, \overline{\mathbf{t}})^* \mathbf{w}, \tag{3.16}$$

which is an abstract smoothness condition. The adjoint of the operator F' defined in (3.13) is given by

$$F'(\alpha, \mathbf{t})^*(u, v) = \begin{pmatrix} \sum_{i=1}^n b_k(x_i; \mathbf{t}) u_i \\ \sum_{i=1}^n \sum_{j=1}^m \alpha_j \frac{\partial b_j}{\partial t_k}(x_i; \mathbf{t}) v_i \end{pmatrix} \tag{3.17}$$

Theorem 3.10 (Rate of Convergence). *Let $\mathbf{y}^\delta \in \mathbf{R}^m$ satisfy (2.7) and let $\alpha_0, \mathbf{t}_0$ be a solution of minimal distance (in the product space $\ell^2 \times \ell^2$) to the prior $(0, \mathbf{t}^*)$. Furthermore, let the metric projection of the exact data y onto $\mathcal{R}(F)$ be unique and equal the projection of $\mathcal{R}(F) \cap B_\epsilon(F(\alpha_0, \mathbf{t}_0))$. Finally, let $b_j \in C^{1,1}([0,1])$ and denote by L_F the resulting Lipschitz-constant of F' in $B_r(\alpha_0, \mathbf{t}_0)$ due to Lemma 3.9. If (3.16) holds with*

$$L_F \|w\|_{\ell^2} < 1, \tag{3.18}$$

then the choice $\beta_1 = \beta_2 \sim \sqrt{\delta}$ yields

$$\|(\alpha^\delta - \alpha_0, \mathbf{t}^\delta - \mathbf{t}_0)\|_{\ell^2 \times \ell^2} = \mathcal{O}\left(\sqrt{\delta}\right), \tag{3.19}$$

where $(\alpha^\delta, \mathbf{t}^\delta)$ denotes the solution of (3.3) with noisy data $\mathbf{y}^\delta$.

Proof. The assertion follows by an application of Theorem 3.7 in [3]. □

Remark 3.11. It is clear that the source condition is a severe restriction if $m < n + \ell$, since the set of parameters that can fulfill the source condition is a lower-dimensional manifold. However, the case of $m \gg n + \ell$ usually arises in practical applications and thus, the source condition is mainly an assumption on the regularity of the distribution of the parameters t_k with respect to the grid points. To illustrate this, we consider the case of cubic B-splines on the unit interval, where the free knots are given by $t_2, ..., \dots, t_{n-1}$ and we have $t_1 = 0$ and $t_n = 1$. Suppose that the following condition is fulfilled:

$$\forall\, k \in \{1, \dots, n-1\} \quad \exists\, i_1(k), i_2(k) \quad x_{i_1(k)}, x_{i_2(k)} \in (t_k, t_{k+1}),$$

then we can set $w_i = 0$ for all $i \notin \{i_1(k), i_2(k)\}_{k \in \{1,\dots,n-1\}}$ and write the source condition as a system for $(w_{i_1(1)}, w_{i_2(1)}, \dots, w_{i_1(n-1)}, w_{i_2(n-1)})$, which is an upper-diagonal system of size $2n-2 \times 2n-2$. Since the diagonal entries are all nonzero (note that $x_{i_1(k)}$ and $x_{i_2(k)}$ are in the interior of the interval (t_k, t_{k+1}), there exists a unique solution. Hence, the source condition (3.16) is satisfied and (3.18) holds if in addition $\|\alpha\|$ and $\|\mathbf{t} - \mathbf{t}^*\|$ are sufficiently small.

4 Numerical Solution of the Regularized Problem

In this section we want to verify theoretical results obtained above by numerical experiments. The description of the optimization algorithm—a generalized Gauss-Newton like method—follows Schütze [20, 21].

4.1 Description of Optimization Algorithm

The common characteristic of both the primal nonlinear least squares problem (2.5) as well as the regularized problems (3.1) and (3.3) is that they are linear in one set of variables (the coefficients α) but nonlinear in the set of free knots $\mathbf{t}$. In the unconstrained case such semi-linear separable problems were first analyzed in detail by Golub/Pereyra [11]. Later Parks [18] treated general constrained nonlinear problems of this type.

Consider the following semi-linear least squares problem with linear inequality constraints:

$$\min_{\alpha,\mathbf{t}}\{\|G(\mathbf{t})\alpha - \mathbf{y}(\mathbf{t})\|^2 \ : C\mathbf{t} \geq \mathbf{h},\ \mathbf{t} \in [0,1]^\ell, \alpha \in \mathbf{R}^n\} \tag{4.1}$$

representing (3.1), (3.2) with appropriately chosen regularized observation matrix $G \in \mathbf{R}^{m+p,n}$, ($p = n - k$ in case of (3.1), $p = n + \ell$ for (3.3)) vector of coefficients $\alpha \in \mathbf{R}^n$ and data vector $\mathbf{y} \in \mathbf{R}^m$. The constraints (3.2) on the knot positions are expressed equivalently in matrix formulation. In the case of (3.3) we do not include the inequality constraints.

The linear subproblem

$$\min_{\alpha}\{\|G(\mathbf{t})\alpha - \mathbf{y}(\mathbf{t})\|^2 \ : \ \alpha \in \mathbf{R}^n\} \tag{4.2}$$

can be solved easily for fixed $\mathbf{t}$, e.g. by reducing G to upper triangular form by a series of Givens rotations, leading to the minimum norm solution

$$\alpha(\mathbf{t}) = G^\dagger(\mathbf{t})\mathbf{y}(\mathbf{t}). \tag{4.3}$$

where $G^\dagger(\mathbf{t})$ is the pseudoinverse of $G(\mathbf{t})$. It follows that the original separable problem can be written

$$\min_{\mathbf{t}}\{\|G(\mathbf{t})\, G^\dagger(\mathbf{t})\, \mathbf{y}(\mathbf{t}) - \mathbf{y}(\mathbf{t})\|^2 \ : \ \mathbf{t} \in [0,1]^\ell\} \tag{4.4}$$

which is now a nonlinear least squares problem in the free knots $\mathbf{t}$ only.

Golub and Pereyra [11] showed that under natural assumptions which guarantee the continuity of the pseudoinverse, the reduction is feasible in the sense that the change from minimizing the full problem to minimizing the reduced problem does not add any critical points and does not exclude the solution of the original problem. Such a natural assumption is that the rank of the matrix $G(\mathbf{t})$ is constant on an open neighborhood which contains the solution. The constant rank assumption, even the full rank assumption on $G(\mathbf{t})$ is satisfied in the case of the regularized problems (3.1), (3.2) and (3.3). Unfortunately for arbitrary data the matrix $G(\mathbf{t})$ of the original problem (2.5) does not satisfy this full rank assumption (cf. Schoenberg-Whitney condition [6]).

Since $G(\mathbf{t})\, G^\dagger(\mathbf{t})$ is the orthogonal projector on the range of $G(\mathbf{t})$, algorithms based on (4.4) are often called variable projection algorithms. A

variable projection algorithm using a Gauss-Newton method applied to the reduced problem (4.4) was used to solve the original least squares problem. The Gauss-Newton method is based on a sequence of linear approximations of the residuum. If $\mathbf{t}^\nu$ denotes the current approximation, then a correction $\mathbf{p}^\nu$ is computed as a solution to the quadratic problem

$$\min_{\mathbf{p}}\{\|[I - G(\mathbf{t}^\nu)\, G^\dagger(\mathbf{t}^\nu)]\, \mathbf{y}(\mathbf{t}^\nu) + J(\mathbf{t}^\nu)\mathbf{p}\|^2 \,:\, \mathbf{p} \in \mathbf{R}^\ell\}. \tag{4.5}$$

with J the Jacobi matrix of $R(\mathbf{t}) := [I - G(\mathbf{t})\, G^\dagger(\mathbf{t})]\, \mathbf{y}(\mathbf{t})$ evaluated at $\mathbf{t}^\nu$. If the Jacobian has full rank then (4.5) has a unique solution $\mathbf{p}^\nu$ which defines the new approximate

$$\mathbf{t}^{\nu+1} = \mathbf{t}^\nu + \mathbf{p}^\nu. \tag{4.6}$$

The Gauss-Newton method can be generalized to constrained problems. A search direction $\mathbf{p}^\nu$ is then computed as a solution to

$$\min_{\mathbf{p}}\{\|R(\mathbf{t}^\nu) + J(\mathbf{t}^\nu)\mathbf{p}\|^2 \,:\, C(\mathbf{t}+\mathbf{p}) \geq \mathbf{h},\, \mathbf{p} \in \mathbf{R}^\ell\} \tag{4.7}$$

by first transforming (4.7) by Householder reflections into a least distance problem and finally using an active set strategy for solving the resulting nonnegative least squares problem [17].

For evaluating J, the derivative of R has to be computed. Expressions for the derivatives of B-splines with respect to its knots can be found, e.g. in [20], the formulas for the Frechet derivative of an orthogonal projector in [11]. Alternatively the derivatives can be approximated by finite differences. Then l additional least squares problems have to be solved in each computation of the derivative. However, in our case, the (regularized) observation matrix G is banded, so that the costs of realizing the linear algebra involved are relatively cheap.

The undamped generalized Gauss-Newton method converges only locally and for small residual problems. In order to globalize the method, a Armijo-Goldstein line search has been implemented. To be robust the algorithm must employ stabilizing techniques for the Gauss-Newton steps when the Jacobian J is nearly rank deficient. This is done by applying a Levenberg-Marquardt method.

Jupp [13] referred to the potentially high number of local extrema for free knot least squares problems. For illustration Figure 3 shows the residuals of least squares approximation of the function $8\sin(10x^2 + 5x + 1)$ on $[-1,\, 1]$ depending on the position of the two free knots t_1 and t_2 (triangular membership functions). Not surprisingly, the local minimum to which the optimization algorithm converges heavily depends on the starting knot sequence $\mathbf{t^0}$. Hence, the generalized Gauss-Newton method is rerun several times with equally distributed random starting values to obtain the global minimum.

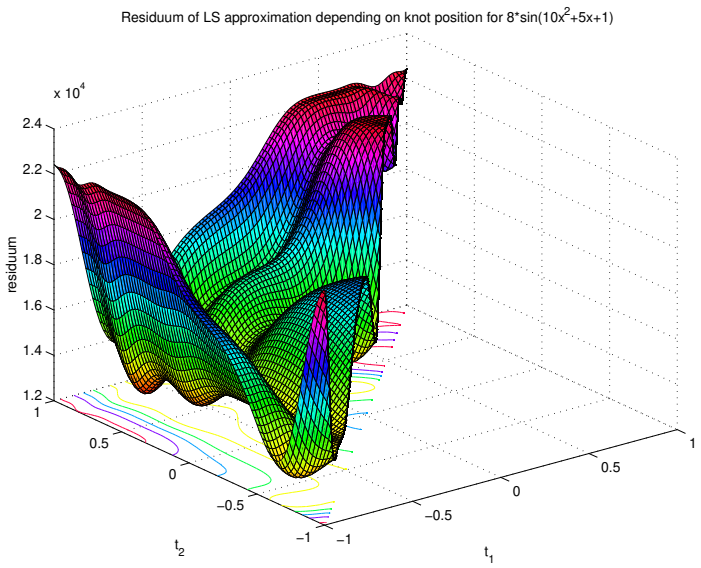

Figure 3: Residuals of least squares approximation with two free knots (triangular membership functions)

4.2 Results for fixed error levels

In the following we compare the results of reconstructing an a-priori given function from noisy measurements taking into account spline approximation, smoothing and Tikhonov regularization. The exact data values are perturbed with uniformly distributed random noise. In the first two examples we take the emphasis on approximation properties, in the third example we take a more careful look onto constructing an interpretable fuzzy controller.

In the figures, the starting knot sequences for the reduced free knot optimization problem are marked with $*$ whereas the locations of the resulting (local) optimal knots are labeled with $\square$. The noisy data are represented by dots, the solid line represents the 'optimal' spline approximation.

In the tables, we compare the residuals $r_{0,0}$ and $r_{\gamma,\delta}$ for exact and noisy data, i.e.

$$r_{\gamma,\delta} := \sqrt{\sum_{i=1}^{m} \left(y_i^{\delta} - \sum_{j=1}^{n} \alpha_j^{\gamma,\delta} \; b_j(x_i^{\gamma}; \mathbf{t}^{\gamma,\delta}) \right)^2} \tag{4.8}$$

where $\alpha^{\gamma,\delta}$ and $\mathbf{t}^{\gamma,\delta}$ denote the solutions to the (regularized) optimization problems with noisy data.

4.2.1 Example 1

As a first example, we consider the reconstruction of the function

$$f_1(x) := sin(4\pi x) \qquad x \in [0,\, 1] \tag{4.9}$$

from noisy data. The 30 x-values, which are chosen randomly distributed in [0, 1], are perturbed with equally distributed random noise up to 1%, the data values up to 25% resulting in $\gamma = 0.0042$ and $\delta = 0.9012$. 15 B-splines of order 5 are used as membership functions. Starting from an equidistant knot sequence the search for the optimal location of the 10 free knots is carried out by the optimization algorithm described above. The results are presented in Figure 4 - Figure 6.

Figure 4 and Table 1 show the results for approximating the data set where the minimal distance between knots is set to 0.002. The minimal distance constraints on the knots of the optimized sequence are active for most of the knots clustering them into three groups. However, the Schoenberg-Whitney condition is not violated, since the support of B-splines of order 5 is relatively large such that there is at least one data point within each support interval. The approximation follows closely the noisy data points leading to abrupt changes in the approximating function and a large discrepancy from f_1.

Smoothing and Tikhonov regularization find approximations much closer to the original function. In the smoothing term the second derivative of the spline is included combined with a relatively small smoothing parameter β. Not surprisingly, the shape of the solution depends strongly on the choice of the regularization parameters. This is of special importance for standard Tikhonov regularization, see Figure 6, where the results are plotted for two different sets of parameters. Additionally the solution depends strongly on the choice of the a-priori knot sequence $\mathbf{t}^*$ which is, in this case, the equidistant knot sequence.

When comparing the results (Table 1), smoothing and Tikhonov regularization are much better than approximation without any regularization. In this example, smoothing gives a slightly smaller residual for exact data than Tikhonov regularization (case 1), mainly because the underlying function f_1 is very smooth and we are using second derivatives in the smoothing term. If somebody wants well separated knots or has a good initial guess of the solution, Tikhonov regularization is preferable.

4.2.2 Example 2

The second example deals with the reconstruction of the function

$$f_2(x) := \frac{10x}{1 + 100x^2} \qquad x \in [-2,\, 2] \tag{4.10}$$

	Approx	Smoothing	Tik 1	Tik 2
$r_{0,0}$	0.7079	0.5763	0.5983	1.4189
$r_{\gamma,\delta}$	0.5710	0.6712	0.6617	1.5739
$\|\alpha\|_{\ell^2}$	4.6017	3.0649	2.9534	1.8563
$\|\mathbf{t} - \mathbf{t}^*\|_{\ell^2}$	0.3613	0.2479	0.0386	0.1840

Table 1: Ex.1: Results for different solution strategies.

$\mathbf{t}^0 = \mathbf{t}^*$	0.091	0.182	0.273	0.364	0.455	0.545	0.636	0.727	0.818	0.909
Approximation	0.056	0.058	0.060	0.492	0.494	0.496	0.816	0.818	0.820	0.822
Smoothing	0.224	0.234	0.252	0.503	0.517	0.522	0.676	0.744	0.746	0.810
Tikhonov 1	0.091	0.160	0.296	0.369	0.425	0.578	0.625	0.717	0.842	0.887
Tikhonov 2	0.082	0.131	0.348	0.350	0.352	0.642	0.644	0.682	0.842	0.856

Table 2: Ex.1: Starting and optimized knot sequences.

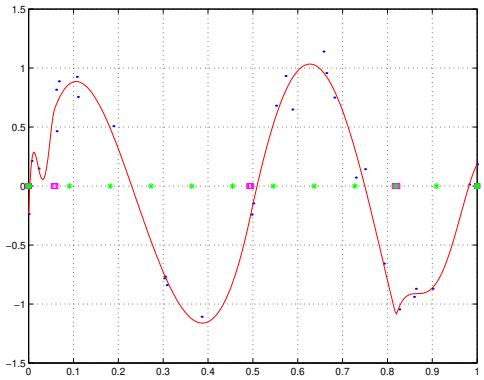

Figure 4: Ex.1: Approximation with 15 B-splines of order 5 ($\varepsilon = 0.002$)

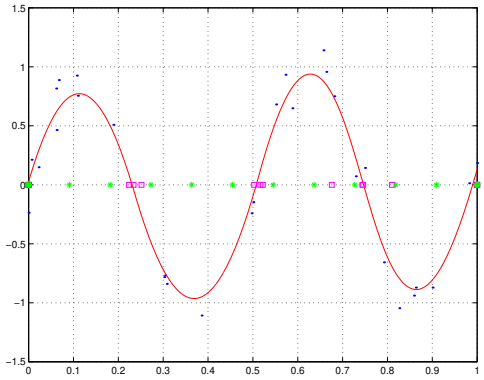

Figure 5: Ex.1: Smoothing with 15 B-splines of order 5 ($k = 2, \beta = 0.0001, \varepsilon = 0.002$)

(see Figure 7), a function already considered in [12] and [20] in the context of spline approximation and smoothing.

Both the abscissa as well as the data values are perturbed with uniformly distributed random noise. The perturbations of the 40 data samples are within a level of 5% and 12%, respectively leading to noiselevels $\gamma = 0.364$ and $\delta = 0.199$. For the reconstruction 11 B-splines of order 3 (quadratic splines) are used. The optimization algorithm for the 8 free knots is started with an equidistant knot sequence.

When approximating f_2 without including any smoothing terms, the resulting function is rather arbitrary (cf. Figure 8); in most cases the optimization procedure breaks down. The Schoenberg-Whitney condition is not satisfied for the knot sequences in the iterative optimization process, the system matrix becomes singular. In smoothing some positions of the optimized knot sequence nearly coincide. However, the smoothing term stabilizes the calculations. In Tikhonov regularization knots are quite separated due to the choice of $\mathbf{t}^*$.

When comparing residuals for exact data, Tikhonov regularization gives better results than regularization via smoothing, and of course, much better results than approximation without applying any regularization technique. But Tikhonov regularization also gives better results with regard to the linguistic interpretability of the resulting fuzzy controller, as we will se in the next example.

	Approximation	Smoothing	Tikhonov reg.
$r_{0,0}$	4.99659	0.55853	0.52923
$r_{\gamma,\delta}$	0.16833	0.59593	0.59760
$\|\alpha\|_{\ell^2}$	8.19528	0.56305	0.60057
$\|\mathbf{t} - \mathbf{t}^*\|_{\ell^2}$	1.11397	0.83290	0.41486

Table 3: Ex.2: Results for different solution strategies.

$\mathbf{t}^0 = \mathbf{t}^*$	-1.566	-1.111	-0.667	-0.222	0.222	0.667	1.111	1.556
Appr.	-1.803	-1.012	-0.192	-0.032	-0.030	0.069	1.405	1.705
Smoothing	-1.872	-0.788	-0.226	-0.224	0.186	0.190	1.037	1.802
Tikhonov	-1.563	-1.120	-0.711	-0.237	0.062	0.288	1.126	1.558

Table 4: Ex.2: Starting and optimized knot sequences.

4.2.3 Example 3

Similar to the paper of Setnes et.al. [23] we want to construct a transparent rule-based model from noisy data measurements considering the spectral data function

$$f_3(x) := 12\,\mathrm{e}^{\frac{-(x-4.8)(x-5.8)}{0.7}} - 12\mathrm{e}^{-(x+3.5)^2} + 0.8x \qquad x \in [-10,\ 10] \quad (4.11)$$

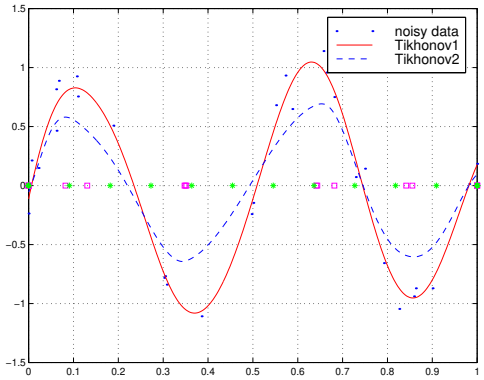

Figure 6: Ex.1: Tikhonov regularization with 15 B-splines of order 5 (Tikhonov 1: $\beta_1 = 0.1$, $\beta_2 = 0.95$, Tikhonov 2: $\beta_1 = 0.95$, $\beta_2 = 0.95$)

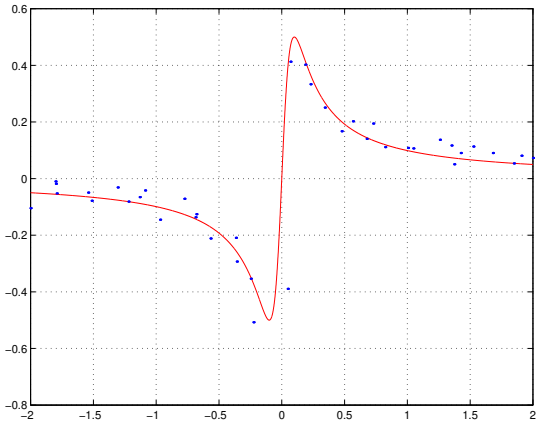

Figure 7: Ex.2: The function $\frac{10x}{1+100x^2}$ and noisy data.

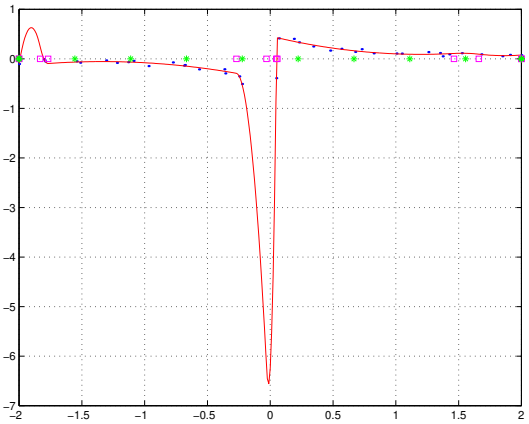

Figure 8: Ex.2: Approximation with 11 quadratic B-splines ($\varepsilon = 0.001$)

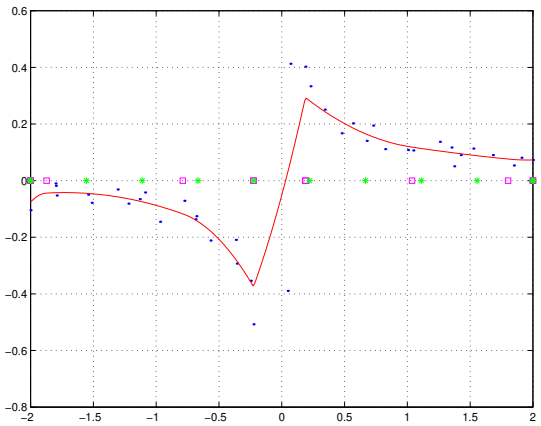

Figure 9: Ex.2: Reconstruction by smoothing with 11 quadratic B-splines ($k = 1, \beta = 0.06, \varepsilon = 0.001$)

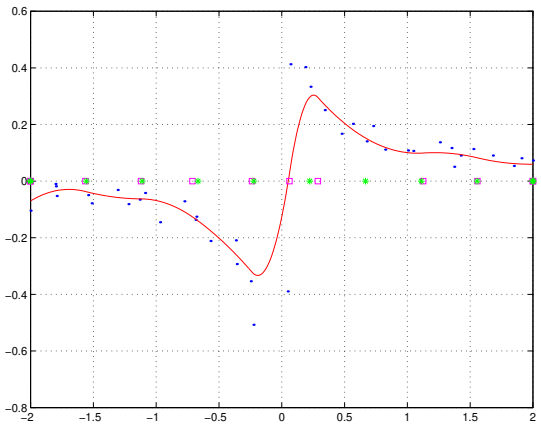

Figure 10: Ex.2: Tikhonov regularization with 11 quadratic B-splines ($\beta_1 = 0.4,\ \beta_2 = 0.4$)

(cf. Figure 11). By using inputs x uniformly distributed in $[-10, 10]$ 50 samples of $f_3(x)$ were obtained and then disturbed with uniformly distributed noise within a noiselevel of 10% (δ =9.5804, maximal error =2.0398).

When constructing a Sugeno controller from measurements, the question on the optimal number of rules or equivalently knots arises. In the context of spline approximation and smoothing, Schütze [20] proposes a knot removal strategy leading to a nearly optimal number of knots. However, we just fix the number of rules to be equal to eight. Accordingly, the universe of discourse is split into eight fuzzy sets interpretable linguistically as negative big, negative medium, negative small, negative very small, positive very small, positive small, positive medium and positive big. To be interpretable easily, the shape of the membership functions is chosen to be triangular.

Figure 12 - Figure 14 show the results for approximation, smoothing and Tikhonov regularization of the noisy data problem. Although the residuum is smaller for approximation than for smoothing and Tikhonov regularization (Table 5), only the later succeeds in constructing an interpretable fuzzy controller since knots are separated appropriately. In approximation and smoothing knots of the optimized sequence nearly coincide (Table 6) leading to questionable and not linguistically interpretable membership functions (Figure 12 - Figure 14, lower part). For Tikhonov regularization $\mathbf{t}^*$ is chosen to be equidistant within the underlying interval.

The linguistic fuzzy model constructed from Tikhonov regularization is given in Table 7.

	Approx	Smoothing	Tikhonov
$r_{0,0}$	12.0859	14.6041	14.5291
$r_{\gamma,\delta}$	12.3130	14.7508	16.9215
$\|\alpha\|_{\ell^2}$	32.6938	25.8770	22.7836
$\|\mathbf{t} - \mathbf{t}^*\|_{\ell^2}$	4.1379	4.4046	1.8854

Table 5: Ex.3: Results for different solution strategies.

$\mathbf{t}^0 = \mathbf{t}^*$	-7.143	-4.286	-1.429	1.429	4.286	7.143
Appr.	-5.585	-2.661	-2.608	3.999	5.658	5.668
Smoothing	-5.372	-3.119	-2.399	4.215	4.346	4.655
Tikhonov	-7.124	-3.718	-0.989	2.790	5.354	7.354

Table 6: Ex.3: Starting and optimized knot sequences.

4.3 Results for error level tending to zero

Again, we consider the reconstruction of the function f_2 (cf. (4.10), Figure 7) and try to validate the convergence properties stated in Theorem 3.10.

Rule: Antecedent		Consequent singleton	Consequent label
R1 : If x is *Negative Big*	then	y= -7.605	*Negative Medium*
R2 : If x is *Negative Medium*	then	y= -5.025	*Negative Medium*
R3 : If x is *Negative Small*	then	y=-11.063	*Negative Big*
R4 : If x is *Negative very Small*	then	y= -0.460	*Negative very Small*
R5 : If x is *Positive very Small*	then	y= 1.367	*Positive very Small*
R6 : If x is *Positive Small*	then	y= 15.095	*Positive Big*
R7 : If x is *Positive Medium*	then	y= 4.968	*Positive Medium*
R8 : If x is *Positive Big*	then	y= 7.682	*Positive Medium*

Table 7: Sugeno controller identified from noisy data.

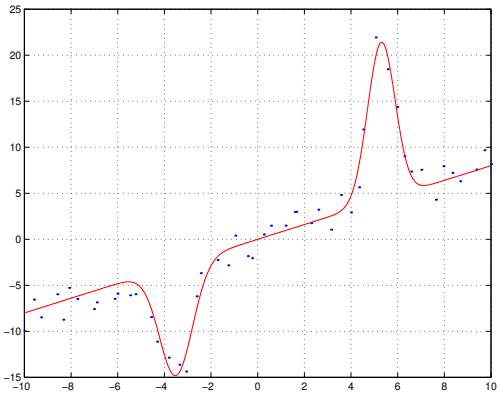

Figure 11: Ex.3: Spectral data function f_3 and noisy measurements

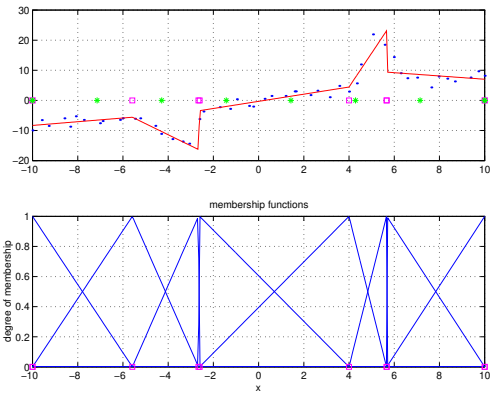

Figure 12: Ex.3: Approximation with 8 triangular membership functions ($\varepsilon = 0.01$)

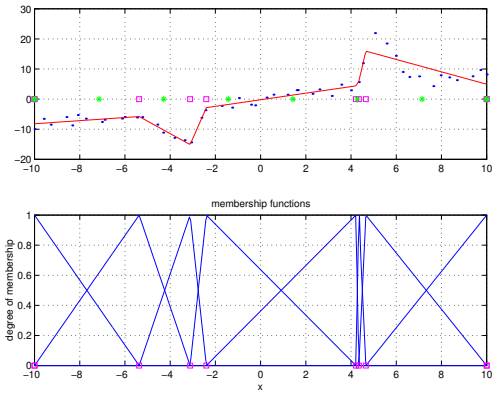

Figure 13: Ex.3: Smoothing with 8 triangular membership functions ($k = 1, \beta = 0.01$, $\varepsilon = 0.01$)

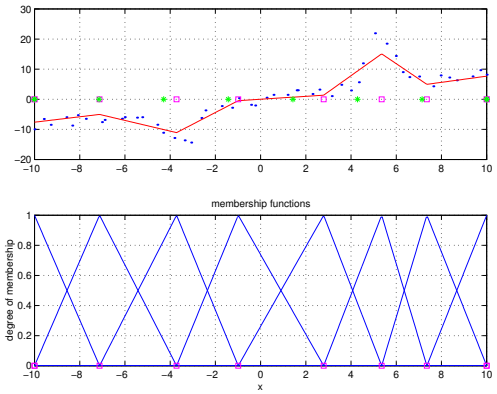

Figure 14: Ex.3: Tikhonov regularization with 8 triangular membership functions ($\beta_1 = \beta_2 = 0.5$)

We take 90 data samples equidistant in $[-2, 2]$ and perturb the y-values with uniformly distributed random noise up to a noiselevel of 20 % (maximal error $= 0.0986$, maximal $\delta = 0.5226$). 15 B-splines of order 5 act as membership functions in Tikhonov regularization. It is easily shown that the assumptions of Theorem 3.10 are satisfied.

The residuum for the least squares approximation of the exact data is equal to 0.004322. The resulting knot sequence

$$\begin{aligned}\overline{\mathbf{t}} \ = \{&-1.2665,\ -0.7356,\ -0.1896,\ -0.0351,\ -0.0350,\\ &0.0349,\ 0.0350,\ 0.1896,\ 0.7355,\ 1.2656\}\end{aligned}$$

is taken as the prior $\mathbf{t}^*$.The regularization parameters are chosen according to the theory ($\beta_1 = \beta_2 = \sqrt{\delta}$). Figure 15 shows the ℓ_2 difference of the coefficients and knots obtained from exact data vs. noisy data. It is noticeable that the difference between the knot sequences is nearly constant or even declines with increasing δ, which could be explained by the increased weighting of the β_2 term in the objective functional. The ℓ_2 difference of the coefficients is quite well in agreement with the theory.

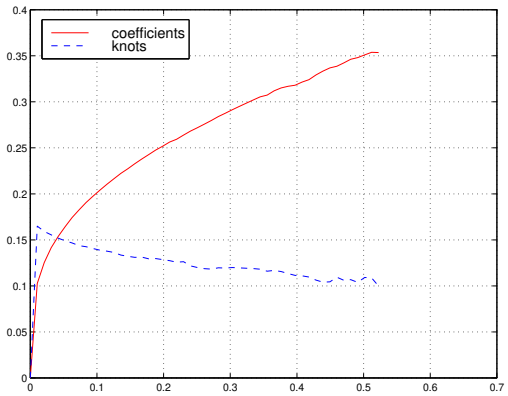

Figure 15: $\|\alpha^\delta - \alpha\|_{\ell_2}$ and $\|\mathbf{t}^\delta - \mathbf{t}^*\|_{\ell_2}$ vs. δ

Finally, in Figure 16 the residuum of the Tikhonov regularized approximation to noisy data is plotted against the error level δ.

5 Extensions and Open Problems

We have seen in the preceding sections that regularization leads to stable approximations of the minimizers and, in addition, improves the interpretability of the arising fuzzy systems, because grid points are separated. So far, we have restricted our analysis to a one-dimensional situations, but multidimensional problems arise in many applications. However, the results on

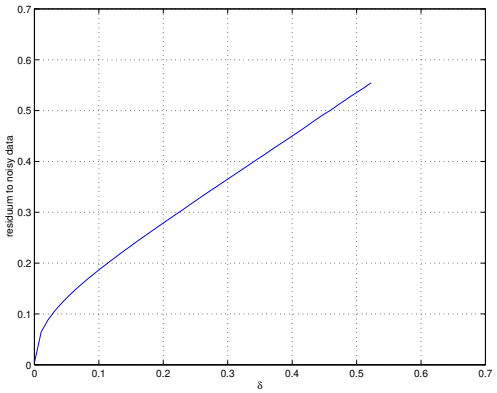

Figure 16: Residuum $\|F(\alpha^\delta, \mathbf{t}^\delta) - \mathbf{y}^\delta\|_{\ell_2}$ of Tikhonov regularized approximation to noisy data vs. noise level

Tikhonov regularization can be carried over to a multi-dimensional situation without many modifications (except with respect to notation). In the case of smoothing the change to higher dimensions is more difficult, since it is not obvious how the singular values of the system matrix can be estimated for arbitrary parameters $\mathbf{t}$.

Since our analysis seems to be a novel approach to the optimization of fuzzy systems, there are still open problems connected to it, which might be of importance for application. In particular we want to mention the so-called *generalization error* (cf. [30]), which means the error of the approximator at points $x \notin \{x_i\}_{i=1,\dots,m}$. A desirable property of the approximators would be convergence to the function from which the samples are taken, as the number of grid points m tends to infinity. However, such a convergence result can be obtained only if also $n \to \infty$, which is often not desirable for fuzzy systems. Nevertheless a meaningful approximation should yield boundedness (and smallness) of the error as $m \to \infty$.

If the grid is regular enough one could consider the case $h \to 0$, where h is a real number such that $|x_i - x_{i-1}| < h$ for all i, which allows a rather standard deterministic analysis. For more irregular distributions of sampling points, one should use different concepts such as stochastic models for the locations. This will be one of our main items for future research.

Acknowledgements

The work of Martin Burger is supported by the *Austrian National Science Foundation FWF* under grant SFB F013/08.

Josef Haslinger and Ulrich Bodenhofer are working in the framework of the *Kplus Competence Center Program* which is funded by the Austrian Government, the Province of Upper Austria, and the Chamber of Commerce

of Upper Austria.

We thank Prof. Heinz W. Engl (University of Linz) for the idea and motivation of this cooperation and for stimulating and fruitful discussions.

References

[1] R. Babuška. *Fuzzy Modeling for Control.* Kluwer Academic Publishers, Boston, 1998.

[2] P. Bauer, E. P. Klement, B. Moser, and A. Leikermoser. Modeling of control functions by fuzzy controllers. In H. T. Nguyen, M. Sugeno, R. M. Tong, and R. R. Yager, editors, *Theoretical Aspects of Fuzzy Control*, chapter 5, pages 91–116. John Wiley & Sons, New York, 1995.

[3] A. Binder, H.W. Engl, C.W. Groetsch, A. Neubauer, and O. Scherzer. Weakly closed nonlinear operators and parameter identification in parabolic equations by Tikhonov regularization. *Appl. Anal.*, 55:215–234, 1994.

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

[5] M. Burger and H.W. Engl. Training neural networks with noisy data as an ill-posed problem. *Adv. Comp. Math.*, 2000. to appear.

[6] C. de Boor. *A practical guide to splines.* Springer-Verlag, New York, Heidelberg, Berlin, 1978.

[7] D. Driankov, H. Hellendoorn, and M. Reinfrank. *An Introduction to Fuzzy Control.* Springer, Heidelberg, 1993.

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

[9] H.W. Engl, K. Kunisch, and A. Neubauer. Convergence rates for Tikhonov regularization of nonlinear ill-posed problems. *Inverse Problems*, 5:523–540, 1989.

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

[11] G.H. Golub and V. Pereyra. The differentiation of pseudoinverses and nonlinear least squares problems whose variables seperate. *SIAM J. Numer. Anal.*, 10:413–432, 1973.

[12] Y. Hu. An algorithm for data reduction using splines with free knots. *IMA J. Numer. Anal.*, 13(3):365–381, 1993.

[13] D.L.B. Jupp. Approximation to data by splines with free knots. *SIAM J. Numer. Anal*, 15(2):328–343, 1978.

[14] E.P. Klement, L.T. Koczy, and B. Moser. Are fuzzy systems universal approximators? *Int. J. General Systems*, 28(2-3):259–282, 1999.

[15] B. Kosko. Fuzzy systems as universal approximators. In *IEEE Int. Conf. on Fuzzy Systems*, pages 1153–1162, San Diego, 1992.

[16] R. Kruse, J. Gebhardt, and F. Klawonn. *Foundations of Fuzzy Systems.* John Wiley & Sons, New York, 1994.

[17] C.L. Lawson and R.J. Hanson. *Solving Least Squares Problems.* SIAM Publications, Philadelphia, 1995.

[18] T.A. Parks. *Reducible Nonlinear Programming Problems.* PhD thesis, Houston Univ., Department of Mathematics, 1985.

[19] E. Schock. Approximate solution of ill-posed equations: arbitrarily slow convergence vs. superconvergence. In G. Hämmerlin and K.H. Hoffmann, editors, *Constructive Methods for the Practical Treatment of Integral Equations*, pages 234–243. Birkhäuser, Basel, 1985.

[20] T. Schütze. *Diskrete Quadratmittelapproximation durch Splines mit freien Knoten.* PhD thesis, Technical University Dresden, Department of Mathematics, 1997.

[21] H. Schwetlick and T. Schütze. Least squares approximation by splines with free knots. *BIT*, 35(3):361–384, 1995.

[22] T.I. Seidman and C.R. Vogel. Well-posedness and convergence of some regularization methods for nonlinear ill-posed problems. *Inverse Problems*, 5:227–238, 1989.

[23] M. Setnes, R. Babuška, and H.B. Verbruggen. Rule-based modeling: Precision and transparency. *IEEE Transactions on Systems, Man, and Cybernetics - Part C*, 28(1):165–169, 1998.

[24] L.L. Shumaker. *Spline Functions: Basic Theory.* John Wiley and Sons, New York, 1981.

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

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

[27] T. Takagi and M. Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE Trans. Syst. Man Cybern.*, 15(1):116–132, 1985.

[28] G. V. Tan and X. Hu. On designing fuzzy controllers using genetic algorithms. In *Proc. FUZZ-IEEE'96*, volume II, pages 905–911, 1996.

[29] L.X. Wang. Fuzzy systems are universal approximators. In *IEEE Int. Conf. on Fuzzy Systems*, pages 1163–1169, San Diego, 1992.

[30] D.H. Wolpert, editor. *The mathematics of generalization*, volume 20 of *SFI Studies in the Sciences of Complexity*, Sante Fe, 1995. Santa Fe Institute, Addison-Wesley. The proceedings of the SFI/CNLS workshop on formal approaches to supervised learning, Summer 1992.

[31] L.A. Zadeh. Fuzzy sets. *Inform. Control*, 8:338–353, 1965.

[32] L.A. Zadeh. Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Trans. Syst. Man Cybern.*, 3(1):28–44, 1973.

[33] J. Zhang and A. Knoll. Constructing fuzzy controllers with B-spline models. In *Proc. FUZZ-IEEE'96*, volume I, pages 416–421, 1996.