

Als Manuskript gedruckt

Technische Universität Dresden  
Herausgeber: Der Rektor

## **Bivariate Free Knot Splines**

Torsten Schütze and Hubert Schwetlick

MATH-NM-13-01

August 2001



# Bivariate Free Knot Splines<sup>\*</sup>

Torsten Schütze<sup>1</sup> and Hubert Schwetlick<sup>2</sup> †

<sup>1</sup>*Siemens AG, Corporate Technology, Information and Communications, Security (CT IC 3),  
D-81730 Munich, Germany.*

[torsten.schuetze@mchp.siemens.de](mailto:torsten.schuetze@mchp.siemens.de), <http://www.math.tu-dresden.de/~schuetze/>

<sup>2</sup>*Department of Mathematics, Dresden University of Technology,  
D-01062 Dresden, Germany.*

[schwetlick@math.tu-dresden.de](mailto:schwetlick@math.tu-dresden.de), <http://www.math.tu-dresden.de/~schwetli/>

## Abstract

We consider the least squares approximation of gridded 2D data by tensor product splines with free knots. The smoothing functional to be minimized—a generalization of the univariate Schoenberg functional—is chosen in such a way that the solution of the bivariate problem separates into the solution of a sequence of univariate problems in case of fixed knots. The resulting optimization problem is a constrained separable least squares problem with tensor product structure. Based on some ideas developed by the authors for the univariate case, an efficient method for solving the specially structured 2D problem is proposed, analyzed and tested on hand of some examples from the literature.

*AMS subject classification:* Primary 65D10, 65D07; Secondary 41A15, 65D17.

*Key words:* Data fitting, bivariate least squares approximation, splines with free knots, constrained separable optimization problems.

## 1 Introduction.

Approximation by tensor product splines with fixed knots is of great importance in theory and applications, mainly due its simple structure and its ease of implementation. However, it is well known that the approximation error is, in general, much smaller if variable knots are allowed instead of fixed ones. This is true for the univariate as well as for the bivariate case.

On the other hand, approximation by splines with free knots leads to difficult but highly structured nonlinear optimization problems. There is a vast amount of literature on Chebyshev approximation by splines with free knots. Up to now mainly univariate problems have been considered. However, a few years ago Nürnberger [21] stated some research problems concerned with bivariate splines with free knots, and in [19] an algorithm was given for bivariate segment approximation which lead to good Chebyshev approximations by splines with free knots .

However, unlike the authors mentioned above, in this paper we address the problem of *least squares approximation* by splines with free knots. We generalize results from least squares approximation by univariate splines with free knots to the case of bivariate tensor product splines with gridded data. While there exist a number of papers for univariate splines with free knots, see [27] for unconstrained free knot splines, [25] and the references cited therein for constrained free knot splines, this idea seems to be new for bivariate tensor product splines.

This paper is organized as follows: In Section 2 we introduce the notation and review some results from the case of univariate spline approximation and smoothing with free knots. In the following section we consider bivariate smoothing splines and formulate the full and reduced approximation and smoothing problem, respectively. The full problems are nonlinear least squares

---

<sup>\*</sup>Date: August 28, 2001.

<sup>†</sup>Research of the first author was partly supported by Deutsche Forschungsgemeinschaft under grant GR 705/4-2

problems with a special structure—so-called separable least squares problems with tensor product structure. In Section 4 we examine general problems of this type and show—in a certain sense—the equivalence of full and reduced problem. We apply these techniques to bivariate tensor product splines with free knots in Section 5. In Section 6 the numerical solution of the reduced problem is considered. Finally, in Section 7 we demonstrate the performance and the advantages of our method by some numerical tests.

## 2 Review of univariate results.

In this section we introduce the notation and shortly review the main results of univariate free knot spline approximation. For details, discussions, and proofs we refer to [27] and [25].

### 2.1 Spline smoothing with fixed knots.

We want to approximate noisy data  $\{x_i, y_i\}$  ( $i = 1, \dots, m$ ) by a function  $s$  from  $\mathcal{S}_{k,\tau}$ , the space of polynomial splines of order  $k \geq 1$  with knot sequence  $\tau \in \mathbb{R}^{n+k}$ , where  $m \geq n$ . The noisy measurements  $y_i = g(x_i) + \varepsilon_i$  ( $i = 1, \dots, m$ ) result from an unknown smooth function  $g \in W_2^q[a, b]$  with  $\varepsilon_i$  being stochastic errors and  $x_i$  monotonously increasing abscissae.

The parameters of the spline  $s$  have to be chosen in such a way that the Schoenberg functional

$$\frac{1}{2} \sum_{i=1}^m [y_i - s(x_i)]^2 + \mu \frac{1}{2} \int_a^b [s^{(r)}(x)]^2 dx$$

with the smoothing parameter  $\mu > 0$  and fixed order  $r \in \{0, \dots, q\}$  in the smoothing term becomes minimal.

Let  $\tau = (\tau_1, \dots, \tau_{n+k})^T$  with

$$\tau_1 = \dots = \tau_k = a < \tau_{k+1} \leq \dots \leq \tau_n < b = \tau_{n+1} = \dots = \tau_{n+k}$$

be a knot sequence and consider a spline  $s \in \mathcal{S}_{k,\tau}$  defined by  $s = \sum_{j=1}^n B_{j,k,\tau} \alpha_j$  where  $B_{j,k,\tau}$  denotes the usual  $j$ -th normalized polynomial B-spline of order  $k$  with knot sequence  $\tau$ . Using the observation matrix  $\mathbf{B}(\tau) := (B_{j,k,\tau}(x_i))_{i=1,\dots,m}^{j=1,\dots,n} \in \mathbb{R}^{m,n}$ , the vector  $\mathbf{y} := (y_1, \dots, y_m)^T \in \mathbb{R}^m$  of data and the coefficients  $\boldsymbol{\alpha} := (\alpha_1, \dots, \alpha_n)^T \in \mathbb{R}^n$ , the *approximation term* can be written as

$$\frac{1}{2} \sum_{i=1}^m [y_i - s(x_i)]^2 = \frac{1}{2} \|\mathbf{y} - \mathbf{B}(\tau)\boldsymbol{\alpha}\|_2^2.$$

The *smoothing term* can be represented in a similar form as

$$\frac{1}{2} \int_a^b [s^{(r)}(x)]^2 dx = \frac{1}{2} \|\mathbf{S}_r(\tau)\boldsymbol{\alpha}\|_2^2$$

with the smoothing matrix  $\mathbf{S}_r(\tau) \in \mathbb{R}^{n-r,n}$  which is either the exact smoothing matrix  $\bar{\mathbf{S}}_r$  defined as above or else a cheaper approximation  $\tilde{\mathbf{S}}_r$  given in [26] and [27].

Now we are able to express the Schoenberg functional as function of spline coefficients  $\boldsymbol{\alpha}$  and knots  $\tau$  as

$$\frac{1}{2} \|\mathbf{y} - \mathbf{B}(\tau)\boldsymbol{\alpha}\|_2^2 + \frac{1}{2} \mu \|\mathbf{S}_r(\tau)\boldsymbol{\alpha}\|_2^2.$$

The system matrix  $\mathbf{B}_\mu(\tau) := \begin{bmatrix} \mathbf{B}(\tau) \\ \sqrt{\mu} \mathbf{S}_r(\tau) \end{bmatrix} \in \mathbb{R}^{m+n-r,n}$  has full rank  $n$  if the regularity condition  $m \geq r$  and  $\mu > 0$  is met.

## 2.2 Spline smoothing with free knots.

We include a subset  $\mathbf{t} = (\tau_{p(1)}, \dots, \tau_{p(l)})^T \in \mathbb{R}^l$  of the inner knots, the so-called *free knots*, into the optimization process whereas the remaining knots have to be given in advance and stay fixed. Hence,  $\mathbf{B}$  and  $\mathbf{S}_r$  become functions of  $\mathbf{t}$  alone, and we write  $\mathbf{B}(\mathbf{t})$  and  $\mathbf{S}_r(\mathbf{t})$  instead of  $\mathbf{B}(\boldsymbol{\tau})$  and  $\mathbf{S}_r(\boldsymbol{\tau})$ , resp. The number of free knots is denoted by  $l$ , and  $\mathbf{p} \in \mathbb{Z}^l$  contains the indices of these free knots.

In approximation by splines with free knots one has to avoid the coalescing of knots. Thus we require  $\tau_{p(j)} \in [\tau_{p(j)-1} + \epsilon h_j, \tau_{p(j)+1} - \epsilon h_j]$ ,  $h_j := t_{p(j)+1} - t_{p(j)-1}$ ,  $j = 1, \dots, l$  with  $0 < \epsilon \ll 1$  which, by using appropriately chosen  $\mathbf{C}$ ,  $\mathbf{h}$ , can equivalently be written as

$$(2.1) \quad \mathbf{C}\mathbf{t} - \mathbf{h} \geq \mathbf{0} \quad \text{with } \mathbf{C} \in \mathbb{R}^{2l, l}, \mathbf{h} \in \mathbb{R}^{2l}.$$

Finally we can formulate the *full smoothing problem*

$$(2.2) \quad \underset{\boldsymbol{\alpha} \in \mathbb{R}^n, \mathbf{t} \in \mathbb{R}^l}{\text{minimize}} \quad f(\boldsymbol{\alpha}, \mathbf{t}) := \frac{1}{2} \left\| \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} - \begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu} \mathbf{S}_r(\mathbf{t}) \end{bmatrix} \boldsymbol{\alpha} \right\|_2^2 \quad \text{s.t. } \mathbf{C}\mathbf{t} - \mathbf{h} \geq \mathbf{0}.$$

Problem (2.2) is a separable nonlinear least squares problem. By inserting the minimum norm solution  $\boldsymbol{\alpha}_{opt}(\mathbf{t}) := \mathbf{B}_\mu(\mathbf{t})^+ \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}$  of (2.2) for fixed knots  $\mathbf{t}$  with respect to  $\boldsymbol{\alpha}$  into the objective function of the full problem we obtain the *reduced smoothing problem*

$$(2.3) \quad \underset{\mathbf{t} \in \mathbb{R}^l}{\text{minimize}} \quad f(\mathbf{t}) := \frac{1}{2} \left\| \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} - \begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu} \mathbf{S}_r(\mathbf{t}) \end{bmatrix} \begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu} \mathbf{S}_r(\mathbf{t}) \end{bmatrix}^+ \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} \right\|_2^2 \quad \text{s.t. } \mathbf{C}\mathbf{t} - \mathbf{h} \geq \mathbf{0}$$

with only the free knots  $\mathbf{t}$  as variables. The investigation of the full and reduced smoothing problem and the efficient numerical solution of the reduced problem was the subject of [27]. There it was shown that the reduced smoothing problem (2.3) has a solution under the conditions

(C1) The knots satisfy  $\tau_j < \tau_{j+k-q}$  ( $j = q + 1, \dots, n$ ).

(C2) The regularity condition  $m \geq r$  and  $\mu > 0$  is met.

If further

(C3) The free knots are simple knots, and it holds  $k \geq 3$ .

is fulfilled, then the change from minimizing the full functional to minimizing the reduced functional does not add any critical points, does not exclude the solution of the original problem, and the critical points are equivalent, see the original papers for details.

The reduced problem was solved by a generalized Gauss-Newton method which requires the Jacobian  $\mathbf{F}'(\mathbf{t})$  of the reduced functional  $f(\mathbf{t}) = \frac{1}{2} \|\mathbf{F}(\mathbf{t})\|_2^2$ . Since the computation of the Jacobian is quite expensive we use instead the so-called *Kaufman-approximation*  $\mathbf{J}_K \approx \mathbf{F}'(\mathbf{t})$ . This results in a very efficient and robust algorithm for the computation of univariate free knot splines.

In the following we will extend these ideas to two dimensions, i. e., we tensorize the univariate results. For simplicity we use the notations from the univariate case and characterize the variables in x- and y-direction by a sub- or superscript 1 and 2, resp., e. g.,  $\boldsymbol{\tau}^1$  denotes the knot sequence in x-direction etc.

## 3 Bivariate smoothing splines.

Let  $\mathbf{Z} = \{z_{i_1, i_2} : i_1 = 1, \dots, m_1; i_2 = 1, \dots, m_2\}$  be noisy measurements of an unknown function  $g \in W_2^{q_1, q_2}[a_1, b_1] \times [a_2, b_2]$ , which are given on a grid  $[x_1, \dots, x_{m_1}] \times [y_1, \dots, y_{m_2}]$ , i. e., it holds

$$z_{i_1, i_2} = g(x_{i_1}, y_{i_2}) + \varepsilon_{i_1, i_2}$$

with the abscissae  $a_1 = x_1 < \dots < x_{m_1} = b_1$ ,  $a_2 = y_1 < \dots < y_{m_2} = b_2$  and the measurement errors  $\varepsilon_{i_1, i_2}$ . The stochastic errors  $\varepsilon_{i_1, i_2}$  are assumed to be independent and identically distributed.

We want to approximate these data by a tensor product spline. Such splines have a simple structure, and its computation allows the separation into a sequence of univariate problems if the data are given on a rectangular grid as in our case.

The extremal properties of univariate splines carry over to both interpolating and smoothing tensor product splines. For example, the natural smoothing bicubic spline is the solution of the variational problem

$$(3.1) \quad \min \left\{ \sum_{i_1=1}^{m_1} \sum_{i_2=1}^{m_2} [z_{i_1, i_2} - s(x_{i_1}, y_{i_2})]^2 + \mu_1 \sum_{i_2=1}^{m_2} \int_{a_1}^{b_1} [D^{2,0} s(x, y_{i_2})]^2 dx + \right. \\ \left. \mu_2 \sum_{i_1=1}^{m_1} \int_{a_2}^{b_2} [D^{0,2} s(x_{i_1}, y)]^2 dy + \mu_1 \mu_2 \int_{a_1}^{b_1} \int_{a_2}^{b_2} [D^{2,2} s(x, y)]^2 dx dy \right\}$$

over  $s \in W_2^{2,2}[a_1, b_1] \times [a_2, b_2]$ , see [15]. Here the operator  $D^{r_1, r_2}$  denotes the partial derivative of order  $r_1$  with respect to  $x$  and of order  $r_2$  with respect to  $y$ , and the parameters  $\mu_1 > 0$  and  $\mu_2 > 0$  are the smoothing parameters.

Splines defined by the above extremal property have the disadvantage that the knots of the spline are identical with the data points, i.e., no data reduction is possible. Therefore we do not use the classical variational approach but follow a so-called direct approach by restricting the approximands  $s$  from  $W_2^{2,2}$  a priori to a fixed tensor product space of B-splines, i.e., the unknown function  $g$  will be approximated by a bivariate spline  $s \in \mathcal{S}_{k_1, \tau^1} \otimes \mathcal{S}_{k_2, \tau^2}$  with knot sequences the cardinality of which may be much smaller than the cardinality of the grid points. The space  $\mathcal{S}_{k_1, \tau^1} \otimes \mathcal{S}_{k_2, \tau^2}$  is the tensor product of the univariate spline spaces  $\mathcal{S}_{k_1, \tau^1}$  and  $\mathcal{S}_{k_2, \tau^2}$  of polynomial splines of order  $k_1$  and  $k_2$  with knot sequence  $\tau^1$  and  $\tau^2$ , resp. The knot sequences are as follows:

$$\begin{aligned} \tau^1 : \tau_1^1 &= \dots = \tau_{k_1}^1 = a_1 < \tau_{k_1+1}^1 \leq \dots \leq \tau_{n_1}^1 < b_1 = \tau_{n_1+1}^1 = \dots = \tau_{n_1+k_1}^1 \\ \tau^2 : \tau_1^2 &= \dots = \tau_{k_2}^2 = a_2 < \tau_{k_2+1}^2 \leq \dots \leq \tau_{n_2}^2 < b_2 = \tau_{n_2+1}^2 = \dots = \tau_{n_2+k_2}^2. \end{aligned}$$

The parameters of the spline  $s$  have to be chosen in such a way that the least squares error defined by the approximation term

$$(3.2a) \quad \varphi(s) := \frac{1}{2} \sum_{i_1=1}^{m_1} \sum_{i_2=1}^{m_2} [z_{i_1, i_2} - s(x_{i_1}, y_{i_2})]^2$$

becomes minimal. It is known that the solution of (3.2a) is unique if the Schoenberg-Whitney condition is fulfilled. In the case of arbitrary data this can not be assured.

By using the thin plate functional

$$\int_{a_1}^{b_1} \int_{a_2}^{b_2} \{ [D^{2,0} s(x, y)]^2 + 2[D^{1,1} s(x, y)]^2 + [D^{0,2} s(x, y)]^2 \} dx dy$$

as smoothing term uniqueness can be achieved independently of the data. This functional has the drawback that it has no tensor product structure and, therefore, the solution process does not separate in the case of gridded data. However, by utilizing a smoothing term which is slightly modified compared to (3.1) and minimizing the functional  $\phi$  defined by

$$(3.2b) \quad \phi(s) := \frac{1}{2} \sum_{i_1=1}^{m_1} \sum_{i_2=1}^{m_2} [z_{i_1, i_2} - s(x_{i_1}, y_{i_2})]^2 + \mu_1 \frac{1}{2} \sum_{i_2=1}^{m_2} \int_{a_1}^{b_1} [D^{r_1, 0} s(x, y_{i_2})]^2 dx \\ + \mu_2 \frac{1}{2} \sum_{i_1=1}^{m_1} \int_{a_2}^{b_2} [D^{0, r_2} s(x_{i_1}, y)]^2 dy + \mu_1 \mu_2 \frac{1}{2} \int_{a_1}^{b_1} \int_{a_2}^{b_2} [D^{r_1, r_2} s(x, y)]^2 dy dx,$$

we again have a separation into a sequence of univariate problems. Unlike the thin plate functional, this smoothing functional—a generalization of the univariate Schoenberg functional—has no nice

physical interpretation. However, it serves quite well as an appropriate regularization term in that it guarantees a unique solution and preserves the tensor product structure.

Bivariate smoothing splines with fixed knots have a long history: Dierckx [6] used first a nonseparable smoothing term before a suitable separable term had been found in [7]. Based on a variational approach, Hu/Schumaker considered natural bicubic smoothing splines [15] and complete smoothing splines [16]. The abstract case of interpolating and smoothing tensor product splines—from which many of the above can be derived as special cases—is investigated in [9]. Finally, in [20], [2] and [22] numerical techniques for minimizing the functional (3.2b) for fixed knots have been considered. Let us point out that, independent of the cited sources, V. Kunert has proposed such a tensor product smoothing term and, unlike the others, developed an efficient and stable solution process along the lines of [26].

The main advantage of the direct approach is that data reduction is possible and that the number and position of knots can be chosen independent of the data.

### 3.1 Representation of the smoothing functional.

We use polynomial B-splines of order  $k_1$  and  $k_2$  with knot sequences  $\boldsymbol{\tau}^1$  and  $\boldsymbol{\tau}^2$  as basis for the univariate spaces  $\mathcal{S}_{k_1, \boldsymbol{\tau}^1}$  and  $\mathcal{S}_{k_2, \boldsymbol{\tau}^2}$ , resp. They are denoted by  $B_{j_1, k_1, \boldsymbol{\tau}^1}$  ( $j_1 = 1, \dots, n_1$ ) and  $B_{j_2, k_2, \boldsymbol{\tau}^2}$  ( $j_2 = 1, \dots, n_2$ ). So one obtains the representation

$$s(x, y) = \sum_{j_1=1}^{n_1} \sum_{j_2=1}^{n_2} B_{j_1, k_1, \boldsymbol{\tau}^1}(x) B_{j_2, k_2, \boldsymbol{\tau}^2}(y) \alpha_{j_1, j_2}$$

with coefficients  $\alpha_{j_1, j_2}$  for a tensor product spline  $s \in \mathcal{S}_{k_1, \boldsymbol{\tau}^1} \otimes \mathcal{S}_{k_2, \boldsymbol{\tau}^2}$ . By using this representation, the functionals (3.2a) and (3.2b) lead to the problems

$$(3.3a) \quad \frac{1}{2} \sum_{i_1=1}^{m_1} \sum_{i_2=1}^{m_2} \left[ z_{i_1, i_2} - \sum_{j_1=1}^{n_1} \sum_{j_2=1}^{n_2} B_{j_1, k_1, \boldsymbol{\tau}^1}(x_{i_1}) B_{j_2, k_2, \boldsymbol{\tau}^2}(y_{i_2}) \alpha_{j_1, j_2} \right]^2 \rightarrow \min_{\alpha_{j_1, j_2}}$$

and

$$(3.3b) \quad \frac{1}{2} \sum_{i_1=1}^{m_1} \sum_{i_2=1}^{m_2} \left[ z_{i_1, i_2} - \sum_{j_1=1}^{n_1} \sum_{j_2=1}^{n_2} B_{j_1, k_1, \boldsymbol{\tau}^1}(x_{i_1}) B_{j_2, k_2, \boldsymbol{\tau}^2}(y_{i_2}) \alpha_{j_1, j_2} \right]^2 \\ + \mu_1 \frac{1}{2} \sum_{i_2=1}^{m_2} \int_{a_1}^{b_1} \left[ \sum_{j_1=1}^{n_1} \sum_{j_2=1}^{n_2} B_{j_1, k_1, \boldsymbol{\tau}^1}^{(r_1)}(x) B_{j_2, k_2, \boldsymbol{\tau}^2}(y_{i_2}) \alpha_{j_1, j_2} \right]^2 dx \\ + \mu_2 \frac{1}{2} \sum_{i_1=1}^{m_1} \int_{a_2}^{b_2} \left[ \sum_{j_1=1}^{n_1} \sum_{j_2=1}^{n_2} B_{j_1, k_1, \boldsymbol{\tau}^1}(x_{i_1}) B_{j_2, k_2, \boldsymbol{\tau}^2}^{(r_2)}(y) \alpha_{j_1, j_2} \right]^2 dy \\ + \mu_1 \mu_2 \frac{1}{2} \int_{a_1}^{b_1} \int_{a_2}^{b_2} \left[ B_{j_1, k_1, \boldsymbol{\tau}^1}^{(r_1)}(x) B_{j_2, k_2, \boldsymbol{\tau}^2}^{(r_2)}(y) \alpha_{j_1, j_2} \right]^2 dy dx \rightarrow \min_{\alpha_{j_1, j_2}}.$$

#### 3.1.1 Matrix formulation.

For notational convenience we will use matrix notation in the following. We define

$$\mathbf{A} := (\alpha_{j_1, j_2})_{\substack{j_2=1, \dots, n_2 \\ j_1=1, \dots, n_1}} \in \mathbb{R}^{n_1, n_2}, \quad \mathbf{Z} := (z_{i_1, i_2})_{\substack{i_2=1, \dots, m_2 \\ i_1=1, \dots, m_1}} \in \mathbb{R}^{m_1, m_2}, \\ \boldsymbol{\beta}^1(x, \boldsymbol{\tau}^1) := (B_{1, k_1, \boldsymbol{\tau}^1}(x), \dots, B_{n_1, k_1, \boldsymbol{\tau}^1}(x))^T \in \mathbb{R}^{n_1}, \\ \boldsymbol{\beta}^2(y, \boldsymbol{\tau}^2) := (B_{1, k_2, \boldsymbol{\tau}^2}(y), \dots, B_{n_2, k_2, \boldsymbol{\tau}^2}(y))^T \in \mathbb{R}^{n_2}, \\ \mathbf{B}_1(\boldsymbol{\tau}^1) := (B_{j_1, k_1, \boldsymbol{\tau}^1}(x_{i_1}))_{\substack{j_1=1, \dots, n_1 \\ i_1=1, \dots, m_1}} = (\boldsymbol{\beta}^1(x_{i_1}, \boldsymbol{\tau}^1)^T)_{i_1=1, \dots, m_1} \in \mathbb{R}^{m_1, n_1}, \\ \mathbf{B}_2(\boldsymbol{\tau}^2) := (B_{j_2, k_2, \boldsymbol{\tau}^2}(y_{i_2}))_{\substack{j_2=1, \dots, n_2 \\ i_2=1, \dots, m_2}} = (\boldsymbol{\beta}^2(y_{i_2}, \boldsymbol{\tau}^2)^T)_{i_2=1, \dots, m_2} \in \mathbb{R}^{m_2, n_2}.$$

Thus we have  $s(x, y) = \beta^1(x, \tau^1)^T \mathbf{A} \beta^2(y, \tau^2)$ , and (3.3a) reads as

$$(3.4a) \quad \frac{1}{2} \|\mathbf{Z} - \mathbf{B}_1(\tau^1) \mathbf{A} \mathbf{B}_2(\tau^2)^T\|_F^2 \rightarrow \min_{\mathbf{A} \in \mathbb{R}^{n_1, n_2}}$$

where  $\|\cdot\|_F$  denotes the Frobenius matrix norm. If we define smoothing matrices

$$\mathbf{S}_{r_1}^1(\tau^1) \in \mathbb{R}^{n_1 - r_1, n_1}, \quad \mathbf{S}_{r_2}^2(\tau^2) \in \mathbb{R}^{n_2 - r_2, n_2}$$

as in the univariate case, (3.3b) becomes

$$\begin{aligned} & \frac{1}{2} \|\mathbf{Z} - \mathbf{B}_1(\tau^1) \mathbf{A} \mathbf{B}_2(\tau^2)^T\|_F^2 + \frac{1}{2} \mu_1 \|\mathbf{S}_{r_1}^1(\tau^1) \mathbf{A} \mathbf{B}_2(\tau^2)^T\|_F^2 \\ & + \frac{1}{2} \mu_2 \|\mathbf{B}_1(\tau^1) \mathbf{A} \mathbf{S}_{r_2}^2(\tau^2)^T\|_F^2 + \frac{1}{2} \mu_1 \mu_2 \|\mathbf{S}_{r_1}^1(\tau^1) \mathbf{A} \mathbf{S}_{r_2}^2(\tau^2)^T\|_F^2 \rightarrow \min_{\mathbf{A} \in \mathbb{R}^{n_1, n_2}} \end{aligned}$$

or, using block notation,

$$(3.4b) \quad \frac{1}{2} \left\| \begin{bmatrix} \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{B}_1(\tau^1) \\ \sqrt{\mu_1} \mathbf{S}_{r_1}^1(\tau^1) \end{bmatrix} \mathbf{A} \begin{bmatrix} \mathbf{B}_2(\tau^2) \\ \sqrt{\mu_2} \mathbf{S}_{r_2}^2(\tau^2) \end{bmatrix}^T \right\|_F^2 \rightarrow \min_{\mathbf{A} \in \mathbb{R}^{n_1, n_2}}.$$

Before we characterize the solutions of (3.4a) and (3.4b) we need some auxiliary results. Recall that the *Kronecker product* of  $\mathbf{A} \in \mathbb{R}^{m, n}$  and  $\mathbf{B} \in \mathbb{R}^{p, q}$  is given by

$$\mathbf{A} \otimes \mathbf{B} := \begin{bmatrix} a_{11} \mathbf{B} & \cdots & a_{1n} \mathbf{B} \\ \vdots & & \vdots \\ a_{m1} \mathbf{B} & \cdots & a_{mn} \mathbf{B} \end{bmatrix} \in \mathbb{R}^{m \cdot p, n \cdot q}.$$

For matrices  $\mathbf{A} = (a_{i,j}) \in \mathbb{R}^{m, n}$  the operation  $\text{vec} : \mathbb{R}^{m, n} \rightarrow \mathbb{R}^{m \cdot n}$  is defined by

$$\text{vec}(\mathbf{A}) := (a_{11}, \dots, a_{m1}, a_{12}, \dots, a_{m2}, \dots, a_{1n}, \dots, a_{mn})^T \in \mathbb{R}^{m \cdot n}.$$

Using elementary facts about Kronecker products and the  $\text{vec}$ -operator, see [1] for details, it can be shown that  $\|\mathbf{A} \mathbf{X} \mathbf{B} - \mathbf{C}\|_F^2 = \|(\mathbf{B}^T \otimes \mathbf{A}) \text{vec}(\mathbf{X}) - \text{vec}(\mathbf{C})\|_2^2$ . Hence, problem (3.4a) is equivalent to

$$(3.5) \quad \frac{1}{2} \|\text{vec}(\mathbf{Z}) - (\mathbf{B}_2(\tau^2) \otimes \mathbf{B}_1(\tau^1)) \text{vec}(\mathbf{A})\|_2^2 \rightarrow \min_{\text{vec}(\mathbf{A}) \in \mathbb{R}^{n_1 n_2}},$$

the minimum norm solution  $\mathbf{A}_{opt}(\tau^1, \tau^2)$  of which is, for fixed  $\tau^1$  and  $\tau^2$ , given by

$$\text{vec}(\mathbf{A}_{opt}) = (\mathbf{B}_2 \otimes \mathbf{B}_1)^+ \text{vec}(\mathbf{Z}) = (\mathbf{B}_2^+ \otimes \mathbf{B}_1^+) \text{vec}(\mathbf{Z}) = \text{vec}(\mathbf{B}_1^+ \mathbf{Z} (\mathbf{B}_2^+)^T), \text{ i.e.,}$$

$$(3.6a) \quad \mathbf{A}_{opt}(\tau^1, \tau^2) := \mathbf{B}_1(\tau^1)^+ \mathbf{Z} (\mathbf{B}_2(\tau^2)^+)^T.$$

It can be computed by successive solution of the following univariate problems:

$$(F) \quad \text{Solve} \quad \frac{1}{2} \|\mathbf{Z} - \mathbf{B}_1(\tau^1) \mathbf{F}\|_F^2 \rightarrow \min_{\mathbf{F} \in \mathbb{R}^{n_1, m_2}} \quad \text{for} \quad \mathbf{F} = \mathbf{B}_1(\tau^1)^+ \mathbf{Z},$$

$$(A) \quad \text{Solve} \quad \frac{1}{2} \|\mathbf{F}^T - \mathbf{B}_2(\tau^2) \mathbf{A}^T\|_F^2 \rightarrow \min_{\mathbf{A} \in \mathbb{R}^{n_1, n_2}} \quad \text{for} \quad \mathbf{A}^T = \mathbf{B}_2(\tau^2)^+ \mathbf{F}^T.$$

Analogously one obtains the minimum norm solution to problem (3.4b)

$$(3.6b) \quad \mathbf{A}_{opt}(\tau^1, \tau^2) := \begin{bmatrix} \mathbf{B}_1(\tau^1) \\ \sqrt{\mu_1} \mathbf{S}_{r_1}^1(\tau^1) \end{bmatrix}^+ \begin{bmatrix} \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \left( \begin{bmatrix} \mathbf{B}_2(\tau^2) \\ \sqrt{\mu_2} \mathbf{S}_{r_2}^2(\tau^2) \end{bmatrix}^+ \right)^T$$

which can successively be computed as follows:

$$(F) \quad \text{Solve} \quad \frac{1}{2} \left\| \begin{bmatrix} \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{B}_1(\boldsymbol{\tau}^1) \\ \sqrt{\mu_1} \mathbf{S}_{r_1}^1(\boldsymbol{\tau}^1) \end{bmatrix} \mathbf{F} \right\|_F^2 \rightarrow \min_{\mathbf{F} \in \mathbb{R}^{n_1, m_2 + n_2 - r_2}} \quad \text{for } \mathbf{F},$$

$$(A) \quad \text{Solve} \quad \frac{1}{2} \left\| \mathbf{F}^T - \begin{bmatrix} \mathbf{B}_2(\boldsymbol{\tau}^2) \\ \sqrt{\mu_2} \mathbf{S}_{r_2}^2(\boldsymbol{\tau}^2) \end{bmatrix} \mathbf{A}^T \right\|_F^2 \rightarrow \min_{\mathbf{A} \in \mathbb{R}^{n_1, n_2}} \quad \text{for } \mathbf{A}^T.$$

For the solution of subproblems (A) and (F) we can use the rowwise Givens factorization developed for the univariate case, see [26] for details. General investigations about the exploitation of structure in large linear least squares problems based on Kronecker products can be found in [10].

### 3.2 Full and reduced approximation problem.

Following our general philosophy, see [27], [25], we now include the knots of the splines into the optimization process. If we consider a subset  $(\mathbf{t}^1, \mathbf{t}^2)$  of the inner knots in problem (3.4a) as variable, we get the *full approximation problem*

$$(3.7) \quad f(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}) := \frac{1}{2} \left\| \mathbf{Z} - \mathbf{B}_1(\mathbf{t}^1) \mathbf{A} \mathbf{B}_2(\mathbf{t}^2)^T \right\|_F^2 \rightarrow \min_{\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}}$$

with linear constraints

$$(3.8) \quad \mathbf{C}_1 \mathbf{t}^1 - \mathbf{h}^1 \geq \mathbf{0}, \quad \mathbf{C}_2 \mathbf{t}^2 - \mathbf{h}^2 \geq \mathbf{0}.$$

which avoid coalescing of knots.

Inserting the minimum norm solution (3.6a) into the functional  $f$  one obtains the *reduced approximation problem*

$$(3.9) \quad f(\mathbf{t}^1, \mathbf{t}^2) := \frac{1}{2} \left\| \mathbf{Z} - \mathbf{B}_1(\mathbf{t}^1) \mathbf{B}_1(\mathbf{t}^1)^+ \mathbf{Z} (\mathbf{B}_2(\mathbf{t}^2)^+)^T \mathbf{B}_2(\mathbf{t}^2)^T \right\|_F^2 \rightarrow \min_{\mathbf{t}^1, \mathbf{t}^2}$$

with the linear inequality constraints (3.8). The objective function can equivalently be written as

$$f(\mathbf{t}^1, \mathbf{t}^2) = \frac{1}{2} \left\| \mathbf{Z} - \mathbf{P}_{B_1} \mathbf{Z} \mathbf{P}_{B_2} \right\|_F^2$$

with the orthogonal projectors  $\mathbf{P}_{B_1} := \mathbf{B}_1(\mathbf{t}^1) \mathbf{B}_1(\mathbf{t}^1)^+$  and  $\mathbf{P}_{B_2} := \mathbf{B}_2(\mathbf{t}^2) \mathbf{B}_2(\mathbf{t}^2)^+$ .

### 3.3 Full and reduced smoothing problem.

For the smoothing problem (3.4b) one obtains analogously the *full smoothing problem*

$$(3.10) \quad f(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}) := \frac{1}{2} \left\| \begin{bmatrix} \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{B}_1(\mathbf{t}^1) \\ \sqrt{\mu_1} \mathbf{S}_{r_1}^1(\mathbf{t}^1) \end{bmatrix} \mathbf{A} \begin{bmatrix} \mathbf{B}_2(\mathbf{t}^2) \\ \sqrt{\mu_2} \mathbf{S}_{r_2}^2(\mathbf{t}^2) \end{bmatrix}^T \right\|_F^2 \rightarrow \min_{\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}}$$

with the linear inequality constraints (3.8).

Again, inserting the minimum norm solution (3.6b) into the functional, one obtains the *reduced smoothing problem*

$$(3.11) \quad f(\mathbf{t}^1, \mathbf{t}^2) := \frac{1}{2} \left\| \begin{bmatrix} \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} - \mathbf{P}_{\left[ \frac{B}{\sqrt{\mu_S}} \right]_1} \begin{bmatrix} \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{P}_{\left[ \frac{B}{\sqrt{\mu_S}} \right]_2} \right\|_F^2 \rightarrow \min_{\mathbf{t}^1, \mathbf{t}^2}$$

with the constraints (3.8). The orthogonal projectors are defined by

$$\mathbf{P}_{\left[ \frac{B}{\sqrt{\mu_S}} \right]_1} := \begin{bmatrix} \mathbf{B}_1(\mathbf{t}^1) \\ \sqrt{\mu_1} \mathbf{S}_{r_1}^1(\mathbf{t}^1) \end{bmatrix} \begin{bmatrix} \mathbf{B}_1(\mathbf{t}^1) \\ \sqrt{\mu_1} \mathbf{S}_{r_1}^1(\mathbf{t}^1) \end{bmatrix}^+$$

$$\mathbf{P}_{\left[ \frac{B}{\sqrt{\mu_S}} \right]_2} := \begin{bmatrix} \mathbf{B}_2(\mathbf{t}^2) \\ \sqrt{\mu_2} \mathbf{S}_{r_2}^2(\mathbf{t}^2) \end{bmatrix} \begin{bmatrix} \mathbf{B}_2(\mathbf{t}^2) \\ \sqrt{\mu_2} \mathbf{S}_{r_2}^2(\mathbf{t}^2) \end{bmatrix}^+.$$

In the rest of this paper we investigate the equivalence of solutions to the full and the reduced problem and develop an algorithm for the solution of the reduced problem. We are especially interested whether the problems are separable and how techniques from the univariate case carry over. Note that a satisfying resolution of the geometric structure expressed by the data is, in principle, impossible if we use tensor product splines. For example, peaks and diagonally placed layers can not be approximated well. Alternatives are splines on triangulations, curved knot lines, or the usage of hierarchical B-Splines. Despite these inherent disadvantages, tensor product splines are widely used in practice since they are simple to handle and inexpensive to compute as long as the knot lines are fixed.

While there are several algorithms available for the direct minimization of the least squares error as function of the free knots in the univariate case, we know of no such algorithms in the bivariate case. If we restrict to a heuristic, adaptive insertion of knot lines we have the algorithm REGRID from [8] at our disposal. For the case of Chebyshev approximation first results were achieved recently in [19] as already mentioned in the introduction.

#### 4 Separable least squares problems with tensor product structure.

In this section we will detach from the problem of spline smoothing and consider general optimization problems of the following form:

*Full problem*

$$(4.1) \quad f(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}) := \frac{1}{2} \|\mathfrak{F}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})\|_F^2 \rightarrow \min_{\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}}$$

with  $\mathfrak{F}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}) := \mathbf{Z} - \mathbf{B}_1(\mathbf{t}^1)\mathbf{A}\mathbf{B}_2(\mathbf{t}^2)^T$  subject to the constraints

$$(4.2) \quad \mathbf{C}_1\mathbf{t}^1 - \mathbf{h}^1 \geq \mathbf{0}, \quad \mathbf{C}_2\mathbf{t}^2 - \mathbf{h}^2 \geq \mathbf{0}.$$

Here  $\mathbf{B}_1$  and  $\mathbf{B}_2$  are arbitrary smooth matrix functions. The remaining quantities  $\mathbf{Z}$ ,  $\mathbf{C}_1$ ,  $\mathbf{C}_2$ ,  $\mathbf{h}^1$  and  $\mathbf{h}^2$  are constant matrices and vectors. The *variable projection solution*, i. e., the optimal solution  $\mathbf{A}_{opt}$  of (4.1) for fixed  $\mathbf{t}^1$  and  $\mathbf{t}^2$ , is defined by

$$(4.3) \quad \mathbf{A}_{opt}(\mathbf{t}^1, \mathbf{t}^2) := \mathbf{B}_1(\mathbf{t}^1)^+ \mathbf{Z} (\mathbf{B}_2(\mathbf{t}^2)^+)^T.$$

*Reduced problem*

$$(4.4) \quad f(\mathbf{t}^1, \mathbf{t}^2) := \frac{1}{2} \|\mathbf{F}(\mathbf{t}^1, \mathbf{t}^2)\|_F^2 \rightarrow \min_{\mathbf{t}^1, \mathbf{t}^2}$$

with  $\mathbf{F}(\mathbf{t}^1, \mathbf{t}^2) := \mathfrak{F}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}_{opt}(\mathbf{t}^1, \mathbf{t}^2)) = \mathbf{P}_{\mathbf{B}_1}^\perp \mathbf{Z} \mathbf{P}_{\mathbf{B}_2}$  under the constraints (4.2). Obviously an equivalent representation of the reduced functional  $f(\mathbf{t}^1, \mathbf{t}^2)$  is given by

$$f(\mathbf{t}^1, \mathbf{t}^2) = \frac{1}{2} \left\| \mathbf{B}_1(\mathbf{t}^1) \mathbf{B}_1(\mathbf{t}^1)^+ \mathbf{Z} (\mathbf{B}_2(\mathbf{t}^2)^+)^T \mathbf{B}_2(\mathbf{t}^2)^T - \mathbf{Z} \right\|_F^2 = \frac{1}{2} \left\| \mathbf{P}_{\mathbf{B}_1} \mathbf{Z} \mathbf{P}_{\mathbf{B}_2}^\perp \right\|_F^2.$$

We will now—in generalization of the results of [14]—investigate to what extent full and reduced problem are equivalent. In the following we will often need the Fréchet derivatives of certain functionals which are defined using Frobenius norms. Let  $\mathfrak{D}$  denote the operator of Fréchet derivative. Then we have

##### Lemma 4.1 (Fréchet derivatives of Frobenius norms).

Let  $\mathbf{A} : \mathbb{R}^l \rightarrow \mathfrak{L}(\mathbb{R}^n, \mathbb{R}^m)$ ,  $\mathbf{x} \in \mathbb{R}^l \rightarrow \mathbf{A}(\mathbf{x}) \in \mathbb{R}^{m,n}$  be a Fréchet differentiable matrix function, and define  $f : \mathbb{R}^l \rightarrow \mathfrak{L}(\mathbb{R})$

$$\mathbf{x} \in \mathbb{R}^l \rightarrow f(\mathbf{x}) := \frac{1}{2} \|\mathbf{A}(\mathbf{x})\|_F^2 = \frac{1}{2} \operatorname{tr} \left\{ \mathbf{A}(\mathbf{x})^T \mathbf{A}(\mathbf{x}) \right\} = \frac{1}{2} \operatorname{tr} \left\{ \mathbf{A}(\mathbf{x}) \mathbf{A}(\mathbf{x})^T \right\} \in \mathbb{R}.$$

Then it holds

$$\begin{aligned}\partial f(\mathbf{x})[\Delta \mathbf{x}] &= \text{tr} \left\{ (\partial \mathbf{A}(\mathbf{x})[\Delta \mathbf{x}])^T \mathbf{A}(\mathbf{x}) \right\} = \text{tr} \left\{ \mathbf{A}(\mathbf{x})^T (\partial \mathbf{A}(\mathbf{x})[\Delta \mathbf{x}]) \right\} \\ &= \text{tr} \left\{ (\partial \mathbf{A}(\mathbf{x})[\Delta \mathbf{x}]) \mathbf{A}(\mathbf{x})^T \right\} = \text{tr} \left\{ \mathbf{A}(\mathbf{x}) (\partial \mathbf{A}(\mathbf{x})[\Delta \mathbf{x}])^T \right\} \quad \text{for all } \Delta \mathbf{x} \in \mathbb{R}^l.\end{aligned}$$

Further, we need the Fréchet derivative of an orthogonal projector. We consider an arbitrary differentiable  $m \times n$  matrix function  $\mathbf{B}(\cdot)$  of locally constant rank and the corresponding orthogonal projector  $\mathbf{P}_B := \mathbf{B}\mathbf{B}^-$ .

**Lemma 4.2 (Golub/Pereyra [14, Lemma 4.1]).**

Let  $\mathbf{B}^-(\cdot)$  be an  $n \times m$  matrix function such that (P1)  $\mathbf{B}\mathbf{B}^-\mathbf{B} = \mathbf{B}$  and (P3)  $(\mathbf{B}\mathbf{B}^-)^T = \mathbf{B}\mathbf{B}^-$ . Then it holds

$$\partial \mathbf{P}_B = \mathbf{P}_B^\perp (\partial \mathbf{B}) \mathbf{B}^- + (\mathbf{P}_B^\perp (\partial \mathbf{B}) \mathbf{B}^-)^T \quad \text{where } \mathbf{P}_B^\perp = \mathbf{I} - \mathbf{P}_B.$$

#### 4.1 Fréchet derivative of the full functional.

Let  $\partial_1$  be the Fréchet derivative with respect to  $\mathbf{t}^1$ . Using Lemma 4.1 we have

$$\partial_1 f(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})[\Delta \mathbf{t}^1] = \text{tr} \left\{ (\partial_1 \mathfrak{F}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})[\Delta \mathbf{t}^1])^T \mathfrak{F}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}) \right\}.$$

For simplicity we will omit the argument of the matrix functions  $\mathbf{B}_1$  and  $\mathbf{B}_2$ .

Obviously it holds  $\partial_1 \mathfrak{F}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})[\Delta \mathbf{t}^1] = -\partial_1 \mathbf{B}_1[\Delta \mathbf{t}^1] \mathbf{A} \mathbf{B}_2^T$ , hence we have

$$(4.5) \quad \partial_1 f(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})[\Delta \mathbf{t}^1] = -\text{tr} \left\{ \mathbf{B}_2 \mathbf{A}^T (\partial_1 \mathbf{B}_1[\Delta \mathbf{t}^1])^T (\mathbf{Z} - \mathbf{B}_1 \mathbf{A} \mathbf{B}_2^T) \right\}.$$

for the Fréchet derivative of the full functional with respect to  $\mathbf{t}^1$ .

Let  $\partial_2$  be the Fréchet derivative with respect to  $\mathbf{t}^2$ . Using the representation

$$f(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}) = \frac{1}{2} \|\mathfrak{F}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})\|_F^2 = \frac{1}{2} \text{tr} \left\{ \mathfrak{F}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}) \mathfrak{F}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})^T \right\}$$

with  $\mathfrak{F}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}) := \mathbf{B}_1 \mathbf{A} \mathbf{B}_2^T - \mathbf{Z}$  we get from Lemma 4.1

$$\partial_2 f(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})[\Delta \mathbf{t}^2] = \text{tr} \left\{ \mathfrak{F}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}) (\partial_2 \mathfrak{F}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})[\Delta \mathbf{t}^2])^T \right\}.$$

With  $\partial_2 \mathfrak{F}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})[\Delta \mathbf{t}^2] = \mathbf{B}_1 \mathbf{A} (\partial_2 \mathbf{B}_2[\Delta \mathbf{t}^2])^T$  we obtain

$$(4.6) \quad \partial_2 f(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})[\Delta \mathbf{t}^2] = \text{tr} \left\{ (\mathbf{B}_1 \mathbf{A} \mathbf{B}_2^T - \mathbf{Z}) (\partial_2 \mathbf{B}_2[\Delta \mathbf{t}^2])^T \mathbf{A}^T \mathbf{B}_1^T \right\}$$

for the Fréchet derivative of the full functional with respect to  $\mathbf{t}^2$ .

#### 4.2 Fréchet derivative of the reduced functional.

We will now compute the Fréchet derivative of the reduced functional  $f$  with respect to  $\mathbf{t}^1$ . First, we have

$$\partial_1 f(\mathbf{t}^1, \mathbf{t}^2)[\Delta \mathbf{t}^1] = \text{tr} \left\{ (\partial_1 \mathbf{F}(\mathbf{t}^1, \mathbf{t}^2)[\Delta \mathbf{t}^1])^T \mathbf{F}(\mathbf{t}^1, \mathbf{t}^2) \right\}$$

because of Lemma 4.1, and with Lemma 4.2

$$\begin{aligned}\partial_1 \mathbf{F}(\mathbf{t}^1, \mathbf{t}^2)[\Delta \mathbf{t}^1] &= (\partial_1 \mathbf{P}_{B_1}^\perp[\Delta \mathbf{t}^1]) \mathbf{Z} \mathbf{P}_{B_2} \\ &= - \left\{ \mathbf{P}_{B_1}^\perp (\partial_1 \mathbf{B}_1[\Delta \mathbf{t}^1]) \mathbf{B}_1^+ + (\mathbf{P}_{B_1}^\perp (\partial_1 \mathbf{B}_1[\Delta \mathbf{t}^1]) \mathbf{B}_1^+)^T \right\} \mathbf{Z} \mathbf{P}_{B_2}.\end{aligned}$$

It follows

$$\begin{aligned}\partial_1 f(\mathbf{t}^1, \mathbf{t}^2)[\Delta \mathbf{t}^1] &= \\ &= -\text{tr} \left\{ \mathbf{P}_{B_2}^T \mathbf{Z}^T \left\{ \mathbf{P}_{B_1}^\perp (\partial_1 \mathbf{B}_1[\Delta \mathbf{t}^1]) \mathbf{B}_1^+ + (\mathbf{P}_{B_1}^\perp (\partial_1 \mathbf{B}_1[\Delta \mathbf{t}^1]) \mathbf{B}_1^+)^T \right\} \mathbf{P}_{B_1}^\perp \mathbf{Z} \mathbf{P}_{B_2} \right\}\end{aligned}$$

and, because of  $\mathbf{B}_1^+ \mathbf{P}_{B_1}^\perp = \mathbf{0}$  and  $(\mathbf{P}_{B_1}^\perp)^T \mathbf{P}_{B_1}^\perp = \mathbf{P}_{B_1}^\perp$ , finally

$$(4.7) \quad \partial_1 f(\mathbf{t}^1, \mathbf{t}^2)[\Delta \mathbf{t}^1] = -\operatorname{tr} \left\{ \mathbf{P}_{B_2} \mathbf{Z}^T ((\partial_1 \mathbf{B}_1)[\Delta \mathbf{t}^1]) \mathbf{B}_1^+ \right\}^T \mathbf{P}_{B_1}^\perp \mathbf{Z} \mathbf{P}_{B_2} \left. \right\}.$$

Analogously we obtain

$$(4.8) \quad \partial_2 f(\mathbf{t}^1, \mathbf{t}^2)[\Delta \mathbf{t}^2] = -\operatorname{tr} \left\{ \mathbf{P}_{B_1} \mathbf{Z} \mathbf{P}_{B_2}^\perp (\partial_2 \mathbf{B}_2[\Delta \mathbf{t}^2]) \mathbf{B}_2^+ \mathbf{Z}^T \mathbf{P}_{B_1} \right\}.$$

using the representation

$$f(\mathbf{t}^1, \mathbf{t}^2) = \frac{1}{2} \|\mathbf{F}(\mathbf{t}^1, \mathbf{t}^2)\|_F^2 = \frac{1}{2} \operatorname{tr} \left\{ \mathbf{F}(\mathbf{t}^1, \mathbf{t}^2) \mathbf{F}(\mathbf{t}^1, \mathbf{t}^2)^T \right\}$$

with  $\mathbf{F}(\mathbf{t}^1, \mathbf{t}^2) := \mathbf{P}_{B_1} \mathbf{Z} \mathbf{P}_{B_2}^\perp$ .

#### 4.3 Relations between Fréchet derivatives.

##### Lemma 4.3.

Let the matrix functions  $\mathbf{B}_1$  and  $\mathbf{B}_2$  have locally constant rank around  $\mathbf{t}^1$  and  $\mathbf{t}^2$ . Let

$$\mathbf{A}_{opt}(\mathbf{t}^1, \mathbf{t}^2) = \mathbf{B}_1(\mathbf{t}^1)^+ \mathbf{Z} \left( \mathbf{B}_2(\mathbf{t}^2)^+ \right)^T$$

be the corresponding variable projection solution. Then

$$\partial_1 f(\mathbf{t}^1, \mathbf{t}^2) = \partial_1 \mathfrak{f}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}_{opt}(\mathbf{t}^1, \mathbf{t}^2)) \quad \text{and} \quad \partial_2 f(\mathbf{t}^1, \mathbf{t}^2) = \partial_2 \mathfrak{f}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}_{opt}(\mathbf{t}^1, \mathbf{t}^2)).$$

*Proof.* (i) First we consider the Fréchet derivative with respect to  $\mathbf{t}^1$ . By inserting the variable projection solution into (4.5) we get

$$\begin{aligned} \partial_1 \mathfrak{f}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}_{opt}(\mathbf{t}^1, \mathbf{t}^2))[\Delta \mathbf{t}^1] &= \operatorname{tr} \left\{ -\mathbf{B}_2 \mathbf{B}_2^+ \mathbf{Z}^T (\mathbf{B}_1^+)^T (\partial_1 \mathbf{B}_1[\Delta \mathbf{t}^1])^T (\mathbf{Z} - \mathbf{B}_1 \mathbf{B}_1^+ \mathbf{Z} (\mathbf{B}_2^+)^T \mathbf{B}_2^T) \right\} \\ &= \operatorname{tr} \left\{ -\mathbf{P}_{B_2} \mathbf{Z}^T ((\partial_1 \mathbf{B}_1[\Delta \mathbf{t}^1]) \mathbf{B}_1^+)^T \mathbf{P}_{B_1}^\perp \mathbf{Z} \mathbf{P}_{B_2} \right\} \\ &= \partial_1 f(\mathbf{t}^1, \mathbf{t}^2)[\Delta \mathbf{t}^1] \quad (\text{see (4.7)}). \end{aligned}$$

(ii) Analogously we get the Fréchet derivative with respect to  $\mathbf{t}^2$  as

$$\begin{aligned} \partial_2 \mathfrak{f}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}_{opt}(\mathbf{t}^1, \mathbf{t}^2))[\Delta \mathbf{t}^2] &= \operatorname{tr} \left\{ (\mathbf{B}_1 \mathbf{B}_1^+ \mathbf{Z} (\mathbf{B}_2^+)^T \mathbf{B}_2^T - \mathbf{Z}) (\partial_2 \mathbf{B}_2[\Delta \mathbf{t}^2]) \mathbf{B}_2^+ \mathbf{Z}^T (\mathbf{B}_1^+)^T \mathbf{B}_1^T \right\} \\ &= \operatorname{tr} \left\{ -\mathbf{P}_{B_1} \mathbf{Z} \mathbf{P}_{B_2}^\perp (\partial_2 \mathbf{B}_2[\Delta \mathbf{t}^2]) \mathbf{B}_2^+ \mathbf{Z}^T \mathbf{P}_{B_1} \right\} \\ &= \partial_2 f(\mathbf{t}^1, \mathbf{t}^2)[\Delta \mathbf{t}^2] \quad (\text{see (4.8)}). \end{aligned}$$

□

The importance of the above lemma and the following theorem lies not only in their statements itself—which are to be expected—but in the formulae for gradient and Jacobian of the reduced functional derived there.

#### 4.4 Correspondence between full and reduced problem.

The following theorem is a natural generalization of [14, Theorem 2.1] to the tensor product case. Note, however, that we have allowed linear inequality constraints with respect to  $\mathbf{t}^1$  and  $\mathbf{t}^2$ .

##### Theorem 4.1 (Correspondence between full and reduced problem).

Let full and reduced problem be defined as above. Further assume that the matrix functions  $\mathbf{B}_1(\mathbf{t}^1)$  (resp.  $\mathbf{B}_2(\mathbf{t}^2)$ ) have constant rank on the open sets  $\Omega_1 \subset \mathbb{R}^{l_1}$  (resp.  $\Omega_2 \subset \mathbb{R}^{l_2}$ ).

- (i) If  $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$  is a critical point (or a global minimizer on  $\Omega_1 \times \Omega_2$ ) of the reduced problem then  $(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*}))$  is a critical point (or a global minimizer for  $(\mathbf{t}^1, \mathbf{t}^2) \in \Omega_1 \times \Omega_2$ ) of the full problem and it holds

$$\mathfrak{f}(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*})) = f(\mathbf{t}^{1*}, \mathbf{t}^{2*})$$

where  $\mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*}) = \mathbf{B}_1(\mathbf{t}^{1*})^+ \mathbf{Z} \left( \mathbf{B}_2(\mathbf{t}^{2*})^+ \right)^T$ .

(ii) Let  $(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$  be a global minimizer of the full problem for  $(\mathbf{t}^1, \mathbf{t}^2) \in \Omega_1 \times \Omega_2$ . Then  $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$  is a global minimizer of the reduced problem on  $\Omega_1 \times \Omega_2$  and it holds

$$f(\mathbf{t}^{1*}, \mathbf{t}^{2*}) = \mathfrak{f}(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*).$$

If there is an unique  $\mathbf{A}^*$  under all minimizing triplets  $\mathfrak{f}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})$ , then it must hold

$$\mathbf{A}^* = \mathbf{B}_1(\mathbf{t}^{1*})^+ \mathbf{Z} \left( \mathbf{B}_2(\mathbf{t}^{2*})^+ \right)^T.$$

*Proof.* We define the Lagrangian of the full and reduced problems

$$\begin{aligned} L_I(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}, \mathbf{w}^1, \mathbf{w}^2) &:= \mathfrak{f}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}) - \sum_{i=1}^{ncstr_1} \mathbf{w}_i^1 r_i^1(\mathbf{t}^1) - \sum_{i=1}^{ncstr_2} \mathbf{w}_i^2 r_i^2(\mathbf{t}^2) \\ L_{II}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{w}^1, \mathbf{w}^2) &:= f(\mathbf{t}^1, \mathbf{t}^2) - \sum_{i=1}^{ncstr_1} w_i^1 r_i^1(\mathbf{t}^1) - \sum_{i=1}^{ncstr_2} w_i^2 r_i^2(\mathbf{t}^2) \end{aligned}$$

with nonnegative multipliers  $\mathbf{w}_i^1, w_i^1$  ( $i = 1, \dots, ncstr_1$ ) and  $\mathbf{w}_i^2, w_i^2$  ( $i = 1, \dots, ncstr_2$ ) and the constraints  $\mathbf{r}^1(\mathbf{t}^1) := \mathbf{C}_1 \mathbf{t}^1 - \mathbf{h}^1 \geq \mathbf{0}$  and  $\mathbf{r}^2(\mathbf{t}^2) := \mathbf{C}_2 \mathbf{t}^2 - \mathbf{h}^2 \geq \mathbf{0}$ .

(i) Let  $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$  be a critical point of the reduced problem. The first order optimality conditions state the existence of multipliers  $\mathbf{w}^{1*}$  and  $\mathbf{w}^{2*}$  so that

$$\begin{aligned} \nabla_{\mathbf{t}^1} L_{II}(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{w}^{1*}, \mathbf{w}^{2*}) &= \mathbf{0} & \nabla_{\mathbf{t}^2} L_{II}(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{w}^{1*}, \mathbf{w}^{2*}) &= \mathbf{0} \\ r_i^1(\mathbf{t}^{1*}) \geq 0 \quad (i = 1, \dots, ncstr_1) & & r_i^2(\mathbf{t}^{2*}) \geq 0 \quad (i = 1, \dots, ncstr_2) & \\ w_i^{1*} r_i^1(\mathbf{t}^{1*}) = 0 \quad (i = 1, \dots, ncstr_1) & & w_i^{2*} r_i^2(\mathbf{t}^{2*}) = 0 \quad (i = 1, \dots, ncstr_2) & \\ w_i^{1*} \geq 0 \quad (i = 1, \dots, ncstr_1) & & w_i^{2*} \geq 0 \quad (i = 1, \dots, ncstr_2) & \end{aligned}$$

Further we assume that appropriate constraint qualifications at  $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$  are satisfied.

Now it holds

$$\begin{aligned} \mathbf{0} &= \nabla_{\mathbf{t}^1} L_{II}(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{w}^{1*}, \mathbf{w}^{2*}) \\ &= \nabla_{\mathbf{t}^1} \mathfrak{f}(\mathbf{t}^{1*}, \mathbf{t}^{2*}) - \sum_{i=1}^{ncstr_1} w_i^{1*} \nabla_{\mathbf{t}^1} r_i^1(\mathbf{t}^{1*}) - \sum_{i=1}^{ncstr_2} w_i^{2*} \nabla_{\mathbf{t}^1} r_i^2(\mathbf{t}^{2*}). \end{aligned}$$

Using Lemma 4.3 and identifying corresponding multipliers we obtain

$$\begin{aligned} &= \nabla_{\mathbf{t}^1} \mathfrak{f}(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*})) - \sum_{i=1}^{ncstr_1} \mathbf{w}_i^{1*} \nabla_{\mathbf{t}^1} r_i^1(\mathbf{t}^{1*}) - \sum_{i=1}^{ncstr_2} \mathbf{w}_i^{2*} \nabla_{\mathbf{t}^1} r_i^2(\mathbf{t}^{2*}) \\ &= \nabla_{\mathbf{t}^1} L_I(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*}), \mathbf{w}^{1*}, \mathbf{w}^{2*}). \end{aligned}$$

Analogously,  $\mathbf{0} = \nabla_{\mathbf{t}^2} L_I(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*}), \mathbf{w}^{1*}, \mathbf{w}^{2*})$ . Together with the feasibility of the constraints and the complementarity (after identification of corresponding multipliers) this yields the first order optimality conditions of the full problem. The constraint qualifications carry over from the full problem since the constraints are unaltered.

Note that  $\nabla_{\mathbf{A}} L_I(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*}), \mathbf{w}^{1*}, \mathbf{w}^{2*}) = \mathbf{0}$  due to the definition of  $\mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*})$ .

Hence  $(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*}), \mathbf{w}^{1*}, \mathbf{w}^{2*})$  is a critical point of the full problem. From the definition of the reduced problem we have

$$\mathfrak{f}(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*})) = f(\mathbf{t}^{1*}, \mathbf{t}^{2*}).$$

The rest of the proof follows the ideas of Golub/Pereyra. For the sake of completeness we state the remaining steps.

Let  $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$  be a global minimizer of the reduced problem on  $\Omega_1 \times \Omega_2$  and let  $\mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*}) = \mathbf{B}_1(\mathbf{t}^{1*})^\dagger \mathbf{Z} \left( \mathbf{B}_2(\mathbf{t}^{2*})^\dagger \right)^T$ . Then  $f(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*})) = f(\mathbf{t}^{1*}, \mathbf{t}^{2*})$ . Assume there exists  $(\mathbf{t}^{1\dagger}, \mathbf{t}^{2\dagger}, \mathbf{A}^\dagger)$  with  $\mathbf{t}^{1\dagger} \in \Omega_1$ ,  $\mathbf{t}^{2\dagger} \in \Omega_2$ , such that  $f(\mathbf{t}^{1\dagger}, \mathbf{t}^{2\dagger}, \mathbf{A}^\dagger) < f(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*}))$ . For all  $(\mathbf{t}^1, \mathbf{t}^2)$  we have  $f(\mathbf{t}^1, \mathbf{t}^2) \leq f(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})$ , hence

$$f(\mathbf{t}^{1\dagger}, \mathbf{t}^{2\dagger}) \leq f(\mathbf{t}^{1\dagger}, \mathbf{t}^{2\dagger}, \mathbf{A}^\dagger) < f(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*})) = f(\mathbf{t}^{1*}, \mathbf{t}^{2*})$$

in contrast to the assumption. Thus  $(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*}))$  is a global minimizer of the full problem in  $\Omega_1 \times \Omega_2$ .

(ii) Let  $(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$  be a global minimizer of the full problem for  $(\mathbf{t}^1, \mathbf{t}^2) \in \Omega_1 \times \Omega_2$  and let  $\mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*}) = \mathbf{B}_1(\mathbf{t}^{1*})^\dagger \mathbf{Z} \left( \mathbf{B}_2(\mathbf{t}^{2*})^\dagger \right)^T$ . It holds  $f(\mathbf{t}^{1*}, \mathbf{t}^{2*}) \leq f(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$ . From the definition of the reduced functional we have  $f(\mathbf{t}^{1*}, \mathbf{t}^{2*}) = f(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*})) \leq f(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$ . Since  $(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$  is a global minimizer equality holds, i. e.,

$$f(\mathbf{t}^{1*}, \mathbf{t}^{2*}) = f(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*).$$

If there is an unique  $\mathbf{A}^*$  among all minimizing triplets of  $f(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})$ , it holds  $\mathbf{A}^* = \mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*})$ .

Let us assume that  $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$  is *no* global minimizer of the reduced problem on  $\Omega_1 \times \Omega_2$ , i. e., there exist  $(\mathbf{t}^{1\dagger}, \mathbf{t}^{2\dagger}) \in \Omega_1 \times \Omega_2$  with  $f(\mathbf{t}^{1\dagger}, \mathbf{t}^{2\dagger}) < f(\mathbf{t}^{1*}, \mathbf{t}^{2*})$ .

With  $\mathbf{A}^\dagger = \mathbf{B}_1(\mathbf{t}^{1\dagger})^\dagger \mathbf{Z} \left( \mathbf{B}_2(\mathbf{t}^{2\dagger})^\dagger \right)^T$  we have

$$f(\mathbf{t}^{1\dagger}, \mathbf{t}^{2\dagger}) = f(\mathbf{t}^{1\dagger}, \mathbf{t}^{2\dagger}, \mathbf{A}^\dagger) < f(\mathbf{t}^{1*}, \mathbf{t}^{2*}) = f(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$$

in contrast to the assumption that  $(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$  is a global minimizer of the full problem.  $\square$

From the argumentation one sees immediately that the statements carry over to the case of nonlinear equality constraints  $\mathbf{s}^1(\mathbf{t}^1) = \mathbf{0}$  and  $\mathbf{s}^2(\mathbf{t}^2) = \mathbf{0}$ .

## 5 Bivariate tensor product splines with free knots.

After these preparations we can now apply the reduction technique to the full smoothing problem (3.10), (3.8). By continuity arguments analogously to the univariate case, see [27], [25, Theorem 4.2], we obtain

### Theorem 5.1 (Existence of solutions to the reduced smoothing problem).

Let the set of feasible knots  $\{(\mathbf{t}^1, \mathbf{t}^2) \in \mathbb{R}^{l_1} \times \mathbb{R}^{l_2} : \mathbf{C}_1 \mathbf{t}^1 - \mathbf{h}^1 \geq \mathbf{0}, \mathbf{C}_2 \mathbf{t}^2 - \mathbf{h}^2 \geq \mathbf{0}\}$  be nonempty, and let the following conditions be fulfilled for fixed  $r_1 \in \{0, \dots, q_1\}$ ,  $0 \leq q_1 < k_1$  and  $r_2 \in \{0, \dots, q_2\}$ ,  $0 \leq q_2 < k_2$ :

(C1) The knots satisfy  $\tau_{j_1}^1 < \tau_{j_1+k_1-q_1}^1$  ( $j_1 = q_1 + 1, \dots, n_1$ ) and  $\tau_{j_2}^2 < \tau_{j_2+k_2-q_2}^2$  ( $j_2 = q_2 + 1, \dots, n_2$ ).

(C2) The regularity conditions  $m_1 \geq r_1$ ,  $\mu_1 > 0$  and  $m_2 \geq r_2$ ,  $\mu_2 > 0$  are met.

Then the reduced smoothing problem (3.11), (3.8) has a solution  $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$ .

We further have the smoothness of the reduced functional and, by applying Theorem 4.1, the equivalence of full and reduced smoothing problem in the following sense:

### Theorem 5.2 (Correspondence between full and reduced smoothing problem).

Let  $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$  be a feasible knot sequence, i. e.,

$$(\mathbf{t}^{1*}, \mathbf{t}^{2*}) \in \{(\mathbf{t}^1, \mathbf{t}^2) \in \mathbb{R}^{l_1} \times \mathbb{R}^{l_2} : \mathbf{C}_1 \mathbf{t}^1 - \mathbf{h}^1 \geq \mathbf{0}, \mathbf{C}_2 \mathbf{t}^2 - \mathbf{h}^2 \geq \mathbf{0}\},$$

and let the following conditions be fulfilled for fixed  $r_1 \in \{0, \dots, q_1\}$ ,  $0 \leq q_1 < k_1$  and  $r_2 \in \{0, \dots, q_2\}$ ,  $0 \leq q_2 < k_2$ :

(C1) The knots satisfy  $\tau_{j_1}^1 < \tau_{j_1+k_1-q_1}^1$  ( $j_1 = q_1 + 1, \dots, n_1$ ) and  $\tau_{j_2}^2 < \tau_{j_2+k_2-q_2}^2$  ( $j_2 = q_2 + 1, \dots, n_2$ ).

(C2) The regularity conditions  $m_1 \geq r_1$ ,  $\mu_1 > 0$  and  $m_2 \geq r_2$ ,  $\mu_2 > 0$  are met.

(C3) The free knots  $(\mathbf{t}^*, \mathbf{t}^{2*})$  are simple knots. It holds  $k_1 \geq 3$  and  $k_2 \geq 3$ .

Then the following relations hold for the full smoothing problem (3.10), (3.8) and the reduced smoothing problem (3.11), (3.8): The reduced function  $\mathbf{F}$  is smooth on the feasible set  $\{(\mathbf{t}^1, \mathbf{t}^2) \in \mathbb{R}^{l_1} \times \mathbb{R}^{l_2} : \mathbf{C}_1 \mathbf{t}^1 - \mathbf{h}^1 \geq \mathbf{0}, \mathbf{C}_2 \mathbf{t}^2 - \mathbf{h}^2 \geq \mathbf{0}\}$ .

- (a) If  $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$  is a critical point (or a global minimizer) of (3.11), (3.8) and  $\mathbf{A}^*$  satisfies (3.6b) at the point  $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$ , then  $(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$  is a critical point (or a global minimizer) of (3.10), (3.8) and it holds  $f(\mathbf{t}^{1*}, \mathbf{t}^{2*}) = f(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$ .
- (b) If  $(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$  is a global minimizer of (3.10), (3.8), then  $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$  is a global minimizer of (3.11), (3.8). It holds  $f(\mathbf{t}^{1*}, \mathbf{t}^{2*}) = f(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$  and (3.6b).

The last two Theorems are straightforward generalizations of the univariate results. Note, however, that the expressions from Section 4 derived for the gradient and Jacobian are crucial for efficient numerical methods.

## 6 Numerical solution of the reduced problem.

After having shown the equivalence of full and reduced problem in the sense of Theorem 5.2 we will now consider the numerical solution of the reduced problem.

The reduced problem is a nonlinear least squares problem in the variables  $\mathbf{t}^1$  and  $\mathbf{t}^2$  with linear inequality constraints. Formally the full problem resembles a separable least squares problem with multiple right hand sides. Such a problem can be expressed as  $\frac{1}{2} \|\mathbf{Z} - \mathbf{B}_1(\mathbf{t}^1) \mathbf{A}\|_F^2 \rightarrow \min_{\mathbf{t}^1, \mathbf{A}}$ , see [13] and [17].

A naive realization transforms the problem  $\frac{1}{2} \|\mathbf{Z} - \mathbf{B}_1(\mathbf{t}^1) \mathbf{A} \mathbf{B}_2^T(\mathbf{t}^2)\|_F^2 \rightarrow \min$  into standard matrix-vector form  $\frac{1}{2} \|\text{vec}(\mathbf{Z}) - (\mathbf{B}_2(\mathbf{t}^2) \otimes \mathbf{B}_1(\mathbf{t}^1)) \text{vec}(\mathbf{A})\|_2^2 \rightarrow \min$  and then applies the reduction technique. In this case we have to factorize the large matrix  $\mathbf{B}_2 \otimes \mathbf{B}_1 \in \mathbb{R}^{m_1 m_2, n_1 n_2}$  to compute the Jacobian of the reduced functional. However, a closer investigation of the structure shows that in every block of the Jacobian the same Fréchet derivative occurs. This structure has to be exploited for the solution process, in particular when solving large scale problems. In the case of separable least squares problems with multiple right hand sides this has first been done in [13].

The method of choice for solving the reduced problem is a generalized Gauss-Newton method. In every step one has to solve the quadratic model problem

$$\mu_{GP}(\mathbf{t}^1 + \Delta \mathbf{t}^1, \mathbf{t}^2 + \Delta \mathbf{t}^2) := \frac{1}{2} \|\mathbf{F}(\mathbf{t}^1, \mathbf{t}^2) + \partial_1 \mathbf{F}(\mathbf{t}^1, \mathbf{t}^2) [\Delta \mathbf{t}^1] + \partial_2 \mathbf{F}(\mathbf{t}^1, \mathbf{t}^2) [\Delta \mathbf{t}^2]\|_F^2 \rightarrow \min_{\Delta \mathbf{t}^1 \in \mathbb{R}^{l_1}, \Delta \mathbf{t}^2 \in \mathbb{R}^{l_2}}$$

subject to the constraints

$$\mathbf{C}_1 \mathbf{t}^1 + \mathbf{C}_1 \Delta \mathbf{t}^1 - \mathbf{h}^1 \geq \mathbf{0}, \quad \mathbf{C}_2 \mathbf{t}^2 + \mathbf{C}_2 \Delta \mathbf{t}^2 - \mathbf{h}^2 \geq \mathbf{0}.$$

This yields

$$\begin{aligned}
\mu_{GP} &= \frac{1}{2} \|\mathbf{F} + \partial_1 \mathbf{F} \Delta \mathbf{t}^1 + \partial_2 \mathbf{F} \Delta \mathbf{t}^2\|_F^2 \\
&= \frac{1}{2} \left\| \mathbf{F} + \sum_{\kappa=1}^{l_1} \partial_1 \mathbf{F}[\mathbf{e}^\kappa] \Delta \mathbf{t}_\kappa^1 + \sum_{\kappa=1}^{l_2} \partial_2 \mathbf{F}[\mathbf{e}^\kappa] \Delta \mathbf{t}_\kappa^2 \right\|_F^2 \\
&= \frac{1}{2} \left\| \text{vec}(\mathbf{F}) + \sum_{\kappa=1}^{l_1} \text{vec}(\partial_1 \mathbf{F}[\mathbf{e}^\kappa]) \Delta \mathbf{t}_\kappa^1 + \sum_{\kappa=1}^{l_2} \text{vec}(\partial_2 \mathbf{F}[\mathbf{e}^\kappa]) \Delta \mathbf{t}_\kappa^2 \right\|_2^2 \\
&= \frac{1}{2} \left\| \text{vec}(\mathbf{F}) + \mathbf{J} \begin{pmatrix} \Delta \mathbf{t}^1 \\ \Delta \mathbf{t}^2 \end{pmatrix} \right\|_2^2
\end{aligned}$$

with

$$\mathbf{J} := \begin{bmatrix} | & & | & & | & & | \\ \text{vec}(\partial_1 \mathbf{F}[\mathbf{e}^1]) & \cdots & \text{vec}(\partial_1 \mathbf{F}[\mathbf{e}^{l_1}]) & \text{vec}(\partial_2 \mathbf{F}[\mathbf{e}^1]) & \cdots & \text{vec}(\partial_2 \mathbf{F}[\mathbf{e}^{l_2}]) \\ | & & | & & | & & | \end{bmatrix}.$$

The Jacobian  $\mathbf{J} \in \mathbb{R}^{(m_1+n_1-r_1)(m_2+n_2-r_2), l_1+l_2}$  in the smoothing case (resp.  $\mathbf{J} \in \mathbb{R}^{m_1 m_2, l_1+l_2}$  in the approximation case), which is in general full, can be computed column-wise.

When computing the Jacobian in the above way, the required QR-factorizations of  $\mathbf{B}_1$  and  $\mathbf{B}_2$  for  $\partial_1 \mathbf{F}$  and  $\partial_2 \mathbf{F}$  have to be computed only once and can then be applied to different right hand sides. All algorithms for the computation of the Jacobian—including the use of the Kaufman approximation and the exploitation of sparsity—carry over from the univariate case. For example, in the case of spline approximation we obtain the Jacobian

$$\partial_1 \mathbf{F}(\mathbf{t}^1, \mathbf{t}^2)[\Delta \mathbf{t}^1] = - \left\{ \mathbf{P}_{B_1}^\perp (\partial_1 \mathbf{B}_1[\Delta \mathbf{t}^1]) \mathbf{B}_1^+ + (\mathbf{P}_{B_1}^\perp (\partial_1 \mathbf{B}_1[\Delta \mathbf{t}^1]) \mathbf{B}_1^+)^T \right\} \mathbf{Z} \mathbf{P}_{B_2}$$

and the Kaufman approximation

$$\mathbf{J}_K[\Delta \mathbf{t}^1] = -\mathbf{P}_{B_1}^\perp (\partial_1 \mathbf{B}_1[\Delta \mathbf{t}^1]) \mathbf{B}_1^+ \mathbf{Z} \mathbf{P}_{B_2}.$$

This procedure bears great resemblance to the computation of the Jacobian for separable least squares problems with multiple right hand sides. If we set  $\mathbf{B}_2(\mathbf{t}^2) = \mathbf{I}$  we arrive straightforward at the results of Golub/LeVeque as a special case. The exploitation of structure for these problems has been intensively investigated during the last years, see [17], [18], [11], [28]. Kaufman/Sylvester report on a drastical reduction in computing time for real world problems with thousands of parameters and millions of observations. While in this case the linear least squares problems were full, Soo/Bates [28] investigate the additional sparsity of the observation matrix. As one example they consider self-modeling free-knot splines, but only in the univariate and not in the tensor product case.

In conclusion, it can be stated that the algorithms for the computation of the Jacobian and the generalized Gauss-Newton method developed in [27], [25] for the univariate case carry over to bivariate free knot splines, in principle.

## 7 Numerical tests.

For the computation of bivariate tensor product splines with free knots we have implemented a method for the solution of the reduced problem in MATLAB. We used the code FMINCON from the MATLAB Optimization Toolbox [4] and the code NPSOL [12] on a PENTIUM 1 GHz processor. Both methods are SQP-methods which use a BFGS-update of the Hessian of the Lagrangian. Gradients have been computed via finite differences, but exact gradients are also available using the formulae in [25].

### 7.1 Bivariate Titanium Heat Data.

In our first example we consider the well-known Titanium Heat Data [5]. Building the tensor product of the univariate data, i.e.,  $z_{i_1, i_2} = y_{i_1} \times y_{i_2}$ , we get  $(m_1 = 49) \times (m_2 = 49)$  points in  $[595, 1075] \times [595, 1075]$ , see Figure 7.1. We want to approximate these 2401 data points by  $n_1 = 11$  cubic B-splines in x-direction and  $n_2 = 9$  cubic B-splines in y-direction. Considering all inner knots as free we achieve at  $l_1 = 7$  and  $l_2 = 5$ .

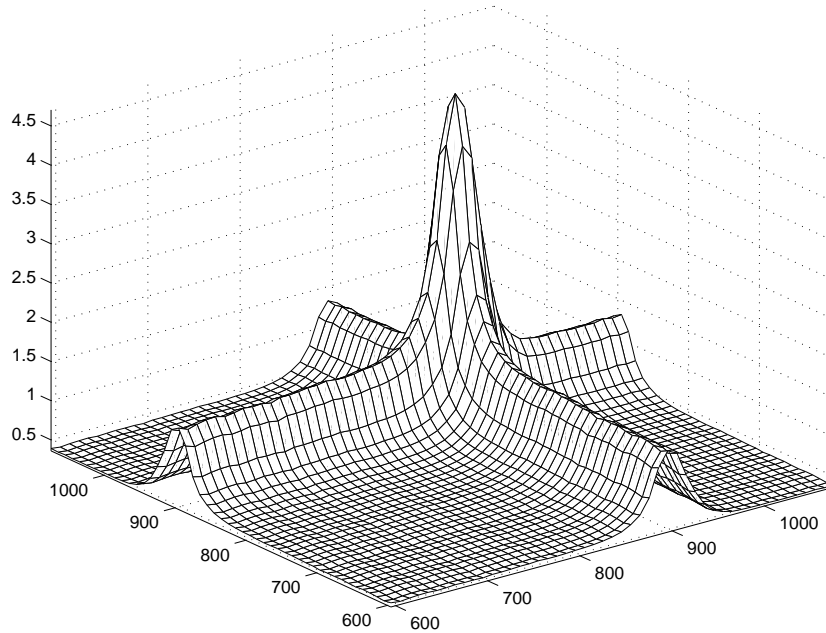


Figure 7.1: Bivariate Titanium Heat Data: data points

If we choose equidistant interior knots we obtain the approximation shown in Figure 7.2. Besides the oscillations in the flat part one observes that, in particular, the approximation near the peak is very poor. Optimizing the location of the knots the oscillations disappear and the residual decreases to 17%. The results are summarized in Table 7.1. Note that the MATLAB procedure cannot achieve the desired accuracy. The resulting spline after the optimization with NPSOL is shown in Figure 7.3. In Figure 7.4 we plot the location of the knots and the corresponding contour before and after the optimization.

Table 7.1: Bivariate Titanium Heat Data: comparison of FMINCON and NPSOL

	initial knot sequence	FMINCON	NPSOL
$\ \mathbf{F}\ $	9.049841 E+00	1.622400 E+00	1.560459 E+00
steps		86	58
func. calls		1209	1336
time [sec]		7.72	8.20
return code		max. no. iterations	successful

Above example clarifies the importance of a good knot sequence. Although the residuals of both approximations do not differ by several magnitudes the approximant with equidistant knots is, in fact, unusable due to its high oscillations. By minimizing the least square error with respect

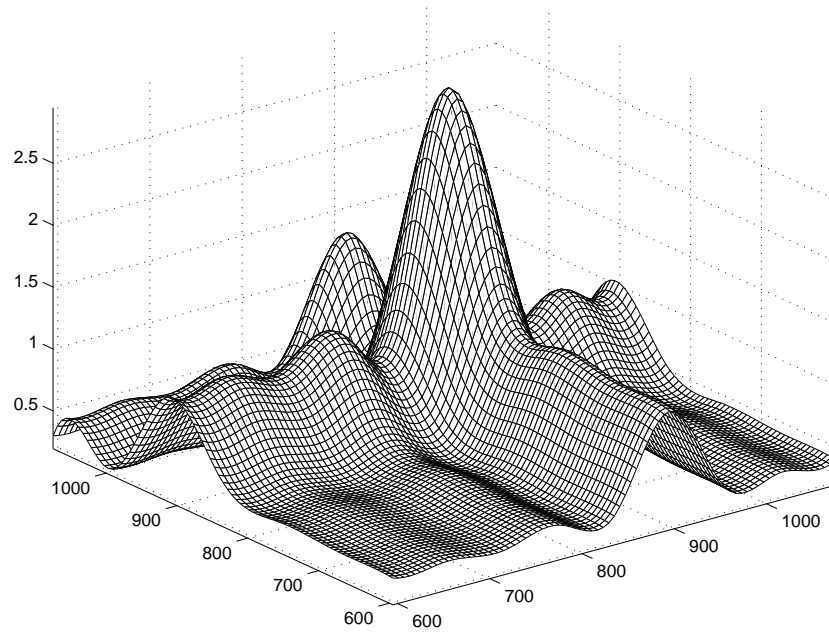
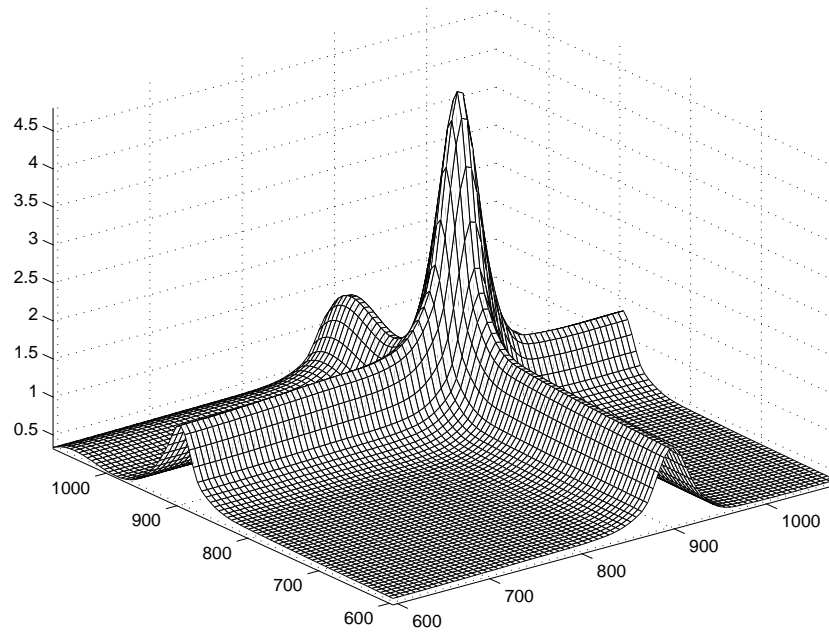
Figure 7.2: Bivariate Titanium Heat Data: spline  $s$ , initial knot sequence

Figure 7.3: Bivariate Titanium Heat Data: optimized location of knots, NPSOL

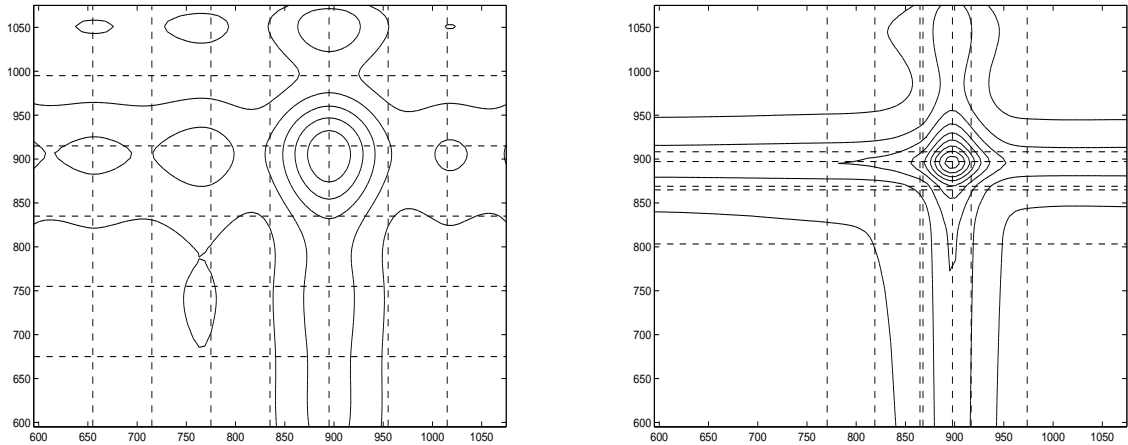


Figure 7.4: Bivariate Titanium Heat Data: contour plot and knots before and after the optimization

to the free knots not only the approximation error gets smaller, even the approximant becomes visually more pleasant.

### 7.2 EOS Aluminium Data.

In a second example we use a standard data set for bivariate *constrained* approximation, see e.g. [3]. The  $(m_1 = 10) \times (m_2 = 6)$  data points in  $[-0.07, 1.13] \times [-2.3, 0]$  describe an equation of state (EOS) for aluminium. Represented is pressure as a function of density and temperature on a log-log scale. The data are in monotone position, see Figure 7.5.

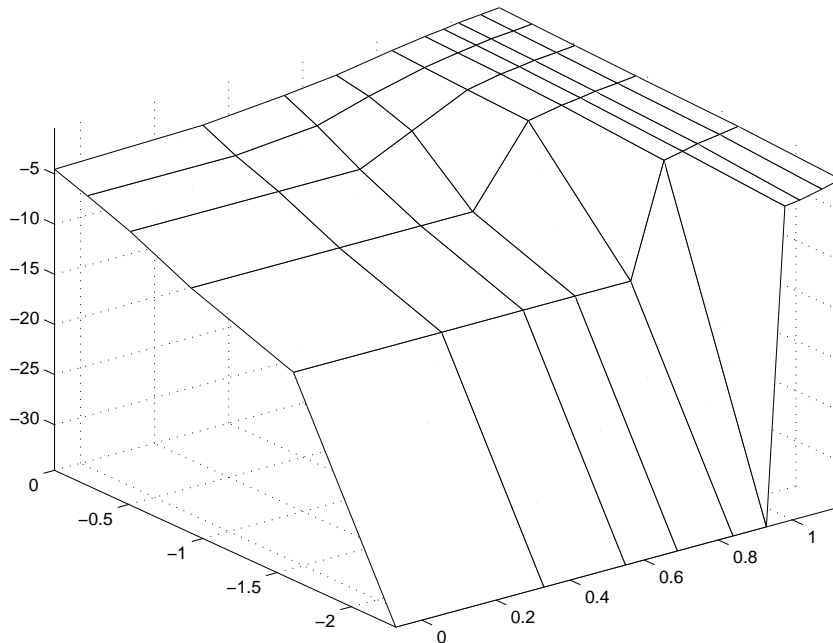


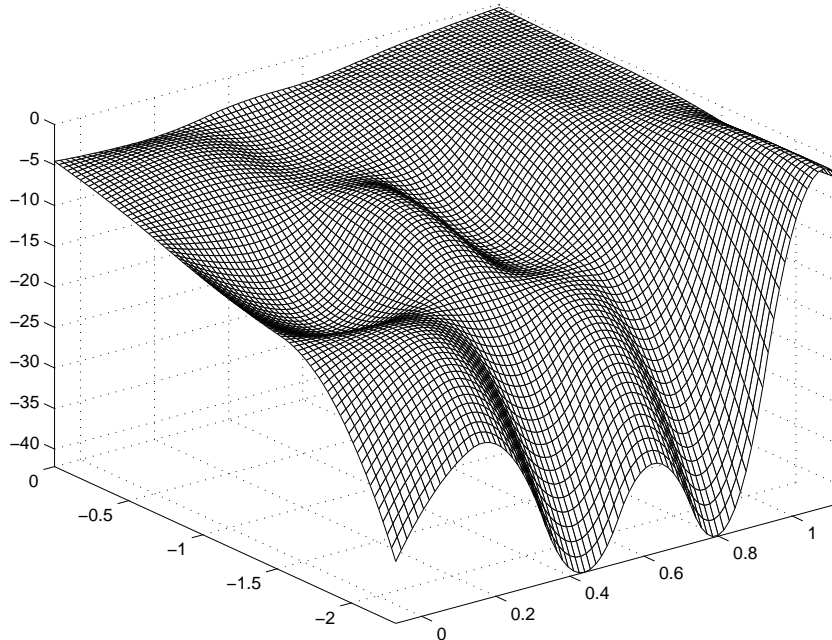
Figure 7.5: EOS Aluminium Data: data set

Although this data set is relatively small it is difficult to approximate. We use  $n_1 = 8$  quadratic B-splines in x-direction,  $n_2 = 5$  quadratic B-splines in y-direction and the smoothing parameters

Table 7.2: EOS Aluminum Data: comparison of FMINCON and NPSOL

	initial knot sequence	FMINCON	NPSOL
$\ \mathbf{F}\ $	1.587922 E+01	1.228879 E+00	1.027007 E+00
steps		65	36
func. calls		603	691
time [sec]		2.03	2.71
return code		successfully	successfully

$\mu_1 = \mu_2 = 1.0 \text{ E-}08$ ,  $r_1 = r_2 = 2$ . For equidistant inner knots we have the unsatisfying approximation shown in Figure 7.6.

Figure 7.6: EOS Aluminum Data: spline  $s$ , initial knot sequence

Optimizing the location of knots one obtains, for example with FMINCON, the spline shown in Figure 7.7. The essential oscillations have disappeared. Table 7.2 and Figure 7.8 summarize the results. The resulting spline approximant is *almost* monotone, it holds  $\min s_y \approx -12$ ,  $\max s_y \approx 86$ ! Note that in this example both methods abort in the case of pure spline approximation without smoothing ( $\mu_1 = \mu_2 = 0$ ) since rank deficient observation matrices—hence loss of differentiability—do occur in the optimization process.

The almost-monotonicity suggests a further area of application for bivariate free knot splines: In so-called *fit-and-modify* methods for constrained interpolation one has to provide good estimates for derivatives. The parameters of the constrained spline are then computed so that the resulting spline deviates as small as possible from the given spline, but satisfies the shape constraints. Since the derivatives of a bivariate free knot spline are of good quality, in general, they qualify as starting parameters for *fit-and-modify* methods. Figure 7.9 represents a monotone interpolating spline that has been computed by methods from [23]. Table 7.3 compares both methods using the EOS Aluminium Data. Combining both methods we expect substantial advantages since the unconstrained splines are *almost optimal* and will serve as a good starting point for the subsequent

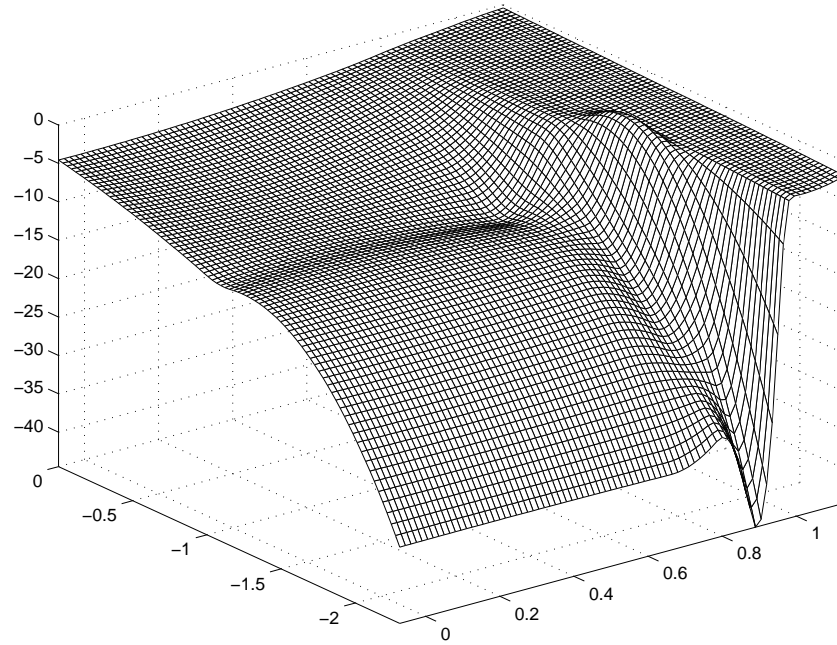


Figure 7.7: EOS Aluminium Data: optimized knots, FMINCON

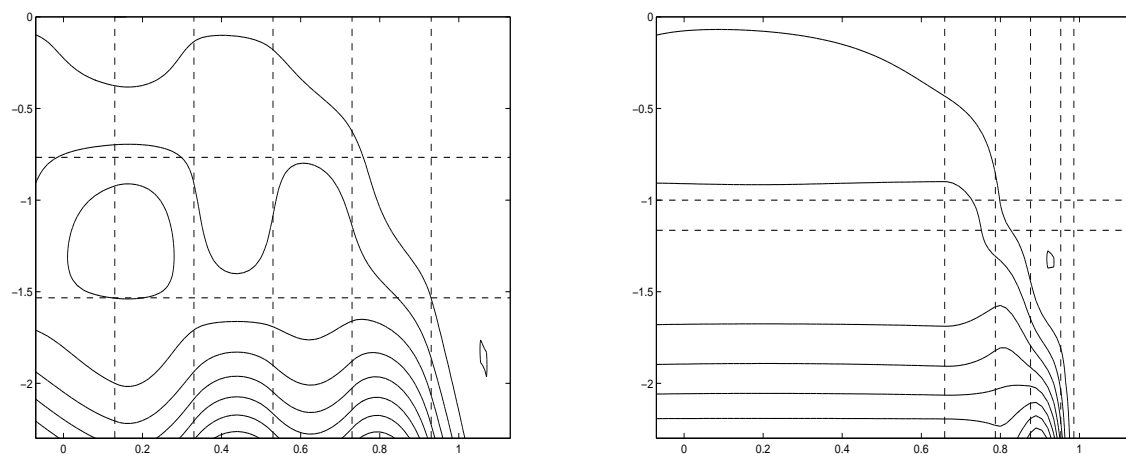


Figure 7.8: EOS Aluminium Data: contour plot and knots before and after the optimization

Table 7.3: Comparison of our method with fit-and-modify method by Schmidt/Bastian-Walther

approximation type	approximation	interpolation
type of optimization problem	nonlinear	quadratic
constraints on derivatives	unconstrained	monotonicity
time [sec]	2.03	45.62
min $s_y$	-12	0
max $s_y$	86	905

iteration method so a reduction of computing time can be expected.

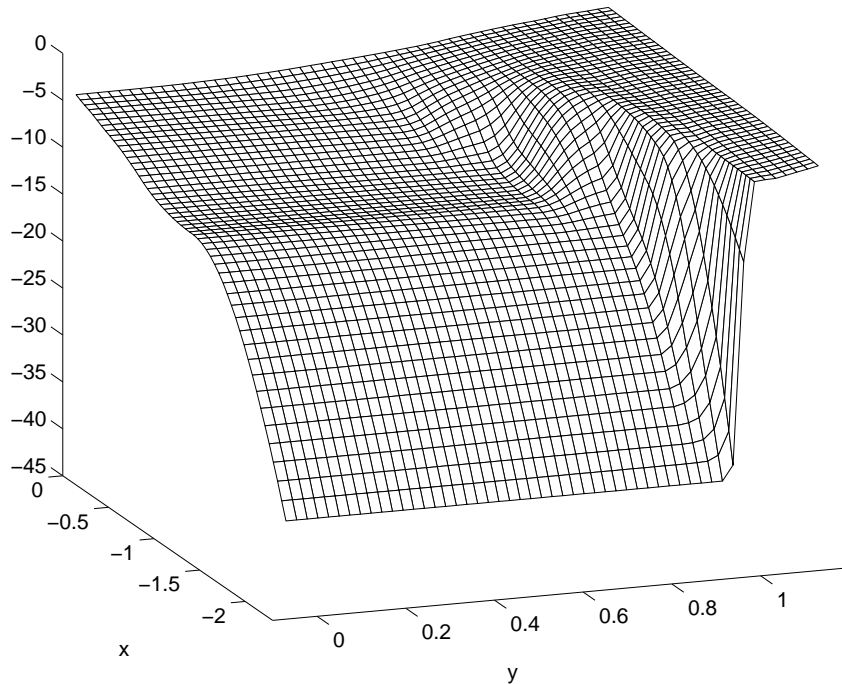


Figure 7.9: EOS Aluminium Data: fit-and-modify method [23], monotone spline

## References

- [1] Å. BJÖRCK, *Numerical methods for least squares problems*, SIAM, Philadelphia, 1996.
- [2] K. BRENNER, *Bivariate Quadratmittelapproximation mittels glättender Tensorprodukt-B-Splines*, Diplomarbeit, Martin-Luther-Universität Halle/Wittenberg, 1990.
- [3] R. E. CARLSON AND F. N. FRITSCH, *Monotone piecewise bicubic interpolation*, SIAM J. Numer. Anal., 22 (1985), pp. 386–400.
- [4] T. COLEMAN, M. A. BRANCH, AND A. GRACE, *Optimization Toolbox User's Guide, Version 2*, The MathWorks, Inc., 1999.
- [5] C. DE BOOR, *A practical guide to splines*, Springer-Verlag, New York, Heidelberg, Berlin, 1978.
- [6] P. DIERCKX, *An algorithm for surface-fitting with spline functions*, IMA J. Numer. Anal., 1 (1981), pp. 267–283.
- [7] ———, *A fast algorithm for smoothing data on a rectangular grid while using spline functions*, SIAM J. Numer. Anal., 19 (1982), pp. 1286–1305.
- [8] ———, *FITPACK user guide, part 2: Surface fitting routines*, Tech. Rep. TW Report 122, Department of Computer Science, Katholieke Universiteit Leuven, Belgium, 1989.

- [9] S. EWALD, H. MÜHLIG, AND B. MULANSKY, *Bivariate interpolating and smoothing tensor product splines*, in *Splines in Numerical Analysis*, Proceed. ISAM-89, J. W. Schmidt and H. Späth, eds., vol. 52 of *Mathematical Research*, Akademie-Verlag, Berlin, 1989, pp. 55–68.
- [10] D. W. FAUSETT AND C. T. FULTON, *Large least squares problems involving Kronecker products*, *SIAM J. Matrix Anal. Appl.*, 15 (1994), pp. 219–227.
- [11] D. M. GAY AND L. KAUFMAN, *Tradeoffs in algorithms for separable nonlinear least squares*, in *Computational and Applied Mathematics, I. Algorithms and Theory*, Selected Papers IMACS 13th World Congress, Dublin, Ireland, 1991, 1992, pp. 179–183.
- [12] P. E. GILL, W. MURRAY, M. A. SAUNDERS, AND M. H. WRIGHT, *User's guide for NPSOL (version 4.0): A Fortran package for nonlinear programming*, Tech. Report SOL 86-2, Department of Operations Research, Stanford University, 1986.
- [13] G. H. GOLUB AND R. J. LEVEQUE, *Extensions and uses of the variable projection algorithm for solving nonlinear least squares problems*, in *Proc. 1979 Army Numerical Analysis and Computer Science Conference*, Army Research Office, 1979.
- [14] G. H. GOLUB AND V. PEREYRA, *The differentiation of pseudoinverses and nonlinear least squares problems whose variables separate*, *SIAM J. Numer. Anal.*, 10 (1973), pp. 413–432.
- [15] C. L. HU AND L. L. SCHUMAKER, *Bivariate natural spline smoothing*, in *Delay Equations, Approximation and Applications*, Mannheim 1994, G. Meinardus and G. Nürnberger, eds., vol. 74 of *ISNM*, Birkhäuser, 1985, pp. 165–179.
- [16] ———, *Complete spline smoothing*, *Numer. Math.*, 49 (1986), pp. 1–10.
- [17] L. KAUFMAN AND G. S. SYLVESTER, *Separable nonlinear least squares with multiple right-hand sides*, *SIAM J. Matrix Anal. Appl.*, 13 (1992), pp. 68–89.
- [18] L. KAUFMAN, G. S. SYLVESTER, AND M. H. WRIGHT, *Structured linear least-squares problems in system identification and separable nonlinear data fitting*, *SIAM J. Optimization*, 4 (1994), pp. 847–871.
- [19] G. MEINARDUS, G. NÜRNBERGER, AND G. WALZ, *Bivariate segment approximation and splines*, *Adv. in Comput. Math.*, 6 (1996), pp. 25–45.
- [20] B. MULANSKY, *Glättung mittels zweidimensionaler Tensorprodukt-Splinefunktionen*, *Wiss. Z. Tech. Univ. Dresden*, 39 (1990), pp. 187–190.
- [21] G. NÜRNBERGER, *Bivariate segment approximation and free knot splines: Research problems 96-4*, *Constr. Approx.*, 12 (1996), pp. 555–558.
- [22] T. PIGORSCH, *Bivariate Quadratmittelapproximation unter Verwendung von Tensorprodukt-B-Splines im Falle von Rechteckgitterdaten*, Diplomarbeit, Martin-Luther-Universität Halle-Wittenberg, 1991.
- [23] J. W. SCHMIDT AND M. WALTHER, *Gridded data interpolation with restrictions on the first order derivatives*, in *Multivariate Approximation and Splines*, G. Nürnberger, J. W. Schmidt, and G. Walz, eds., *ISNM*, Birkhäuser, Basel, 1997, pp. 291–307.
- [24] T. SCHÜTZE, *Diskrete Quadratmittelapproximation durch Splines mit freien Knoten*, Dissertation, Technische Universität Dresden, 1997.
- [25] T. SCHÜTZE AND H. SCHWETLICK, *Constrained approximation by splines with free knots*, *BIT*, 37 (1997), pp. 105–137.
- [26] H. SCHWETLICK AND V. KUNERT, *Spline smoothing under constraints on derivatives*, *BIT*, 33 (1993), pp. 512–528.
- [27] H. SCHWETLICK AND T. SCHÜTZE, *Least squares approximation by splines with free knots*, *BIT*, 35 (1995), pp. 361–384.
- [28] Y. W. SOO AND D. M. BATES, *Loosely coupled nonlinear least squares*, *Computational Statistics and Data Analysis*, 14 (1992), pp. 249–259.