Diskrete Quadratmittelapproximation durch Splines mit freien Knoten

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

vorgelegt

der Fakultät Mathematik und Naturwissenschaften
der Technischen Universität Dresden

von

Dipl.-Math. Torsten Schütze

geboren am 24. September 1969 in Nordhausen

Gutachter:	Prof. Dr. rer. nat. habil. Hubert Schwetlick
	Prof. Dr. rer. nat. habil. Jochen W. Schmidt
	Prof. Dr. rer. nat. habil. Gerhard Opfer
Eingereicht am:	3. September 1997
Tag der Verteidigung:	16. Januar 1998

Mein Dank gilt an erster Stelle Herrn Prof. Dr. H. Schwetlick, der mich vor langer Zeit für die Numerische Mathematik begeisterte und der mich mit dem Thema der Dissertation vertraut machte. Durch die intensive Betreuung und ständige Gesprächsbereitschaft über den gesamten Zeitraum meiner bisherigen wissenschaftlichen Tätigkeit habe ich nicht nur in wissenschaftlichen Dingen von ihm profitiert. Ebenso bin ich für die stets freundliche Unterstützung durch Herrn Prof. Dr. J. W. Schmidt dankbar. Sein Engagement, seine Hinweise und nicht zuletzt die finanzielle Unterstützung im Rahmen des DFG-Projektes „Shape Preserving Spline Approximation“ waren mir eine große Hilfe.

Außerdem möchte ich mich bei allen Mitarbeitern und Angehörigen des Institutes für Numerische Mathematik der Technischen Universität Dresden, die mir hilfreich zur Seite standen, bedanken.

Inhaltsverzeichnis

Abbildungsverzeichnis **iii**

Tabellenverzeichnis **v**

Bezeichnungen **vii**

1 Einleitung **1**

2 Univariate Splines **7**
2.1 Einleitung und historischer Überblick 7
2.2 Problemformulierung 10
2.2.1 Bezeichnungen und Grundlagen aus der Splinetheorie 10
2.2.2 Darstellung des Schoenberg-Funktionals 13
2.2.3 Die freien Knoten 17
2.2.4 Anordnungsbedingung für die freien Knoten 18
2.2.5 Formulierung des vollständigen Optimierungsproblems 20
2.3 Separable Quadratmittelprobleme 20
2.3.1 Gauß-Newton-ähnliche Verfahren für separable Quadratmittelprobleme 21
2.3.2 Der Übergang zum reduzierten Funktional 23
2.3.3 Die Bestimmung der Jacobi-Matrix 24
2.3.4 Konvergenzraten und Aufwand 26
2.3.5 Separable Quadratmittelprobleme mit Nebenbedingungen 27
2.4 Splineglättung mit freien Knoten 27
2.4.1 Existenz von Lösungen des reduzierten Glättungsproblems 28
2.4.2 Äquivalenz von vollständigem und reduziertem Problem 29
2.5 Numerische Lösung des reduzierten Problems 33
2.5.1 Ein verallgemeinertes Gauß-Newton-Verfahren 34
2.5.2 Die Berechnung der Residuumsfunktion 37
2.5.3 Die Berechnung der Jacobi-Matrix 38
2.5.4 FREE – Ein Programm zur Berechnung von Splines mit freien Knoten 42
2.6 Numerische Tests 42
2.6.1 Titanium Heat Data 43
2.6.2 Ein Algorithmus zur Datenreduktion 44
2.6.3 Ein Beispiel von Hu 47

3 Univariate Splines mit Nebenbedingungen 51
3.1 Einleitung . . . 51
3.2 Problemformulierung . . . 53
3.2.1 Glättungsfunktional und Anordnungsnebenbedingungen . . . 53
3.2.2 Nebenbedingungen an Ableitungen . . . 53
3.2.3 Konsistenz der Nebenbedingungen . . . 55
3.2.4 Vollständiges restringiertes Glättungsproblem . . . 58
3.3 Restringierte semi-lineare Quadratmittelprobleme . . . 58
3.3.1 Vollständiges und reduziertes Problem . . . 58
3.3.2 Äquivalenz von vollständigem und reduziertem Problem . . . 59
3.3.3 Quantitative Analyse von Subproblem (A) und reduziertem Problem 61
3.3.4 Struktur der Jacobi-Matrix, Kaufman-Approximation . . . 63
3.4 Splineglättung mit Nebenbedingungen . . . 64
3.5 Numerische Lösung des reduzierten Problems . . . 69
3.5.1 Die Lösung von Subproblem (A) . . . 70
3.5.2 Die Berechnung der Jacobi-Matrix . . . 70
3.6 Numerische Tests . . . 75
3.6.1 Titanium Heat Data . . . 75
3.6.2 Arctan-Daten . . . 78
3.6.3 Volumetric moisture content data . . . 80

4 Bivariate Tensorprodukt-Splines 83
4.1 Einleitung und Problemstellung . . . 83
4.1.1 Darstellung des Zielfunktionals . . . 85
4.1.2 Vollständiges und reduziertes Approximationsproblem . . . 88
4.1.3 Vollständiges und reduziertes Glättungsproblem . . . 88
4.2 Separable Quadratmittelprobleme mit Tensorprodukt-Struktur . . . 89
4.2.1 Die Fréchet-Ableitung des vollständigen Funktionals . . . 91
4.2.2 Die Fréchet-Ableitung des reduzierten Funktionals . . . 91
4.2.3 Beziehungen zwischen den Fréchet-Ableitungen . . . 92
4.2.4 Äquivalenz von vollständigem und reduziertem Problem . . . 93
4.3 Bivariate Tensorprodukt-Splines mit freien Knoten . . . 95
4.4 Numerische Lösung des reduzierten Problems . . . 96
4.5 Numerische Tests . . . 98
4.5.1 Bivariate Titanium Heat Data . . . 99
4.5.2 EOS Aluminium Daten . . . 101

5 Zusammenfassung und Ausblick 105

Literaturverzeichnis 108

Abbildungsverzeichnis

2.1 Titanium Heat Data: Startknotenfolge $\mathbf{t}_2$ 44
2.2 Titanium Heat Data: Optimalstelle $\mathbf{t}^*$ 46
2.3 Beispiel von Hu: Vorgeschalteter Optimierungsschritt ausgehend von äquidistanten inneren Knoten, $n = 20$, $l = 15$ 48
2.4 Beispiel von Hu: Knotenreduktionsalgorithmus, Ergebnis von Stufe II, $n = 10$, $l = 5$ 48

3.1 Äquivalenz von vollständigem restringierten Glättungsproblem FCSP und reduziertem restringierten Glättungsproblem RCSP 69
3.2 Titanium Heat Data: Spline s, Startknotenfolge 76
3.3 Titanium Heat Data: Spline s, Optimierte Knotenfolge, RCSP-Ka-ED 76
3.4 Titanium Heat Data: s'' zur Startknotenfolge und zur optimierten Knotenfolge (RCSP-Ka-ED) 77
3.5 Titanium Heat Data: CONCON (—), RCAP-Ka-ED (- - -), $x \in [925, 1075]$ . . 78
3.6 Arctan-Daten: Spline s (—) und Funktion g (- - -), Startknotenfolge 80
3.7 Arctan-Daten: Erste Ableitungen s' (—) und g' (- - -), Startknotenfolge . . 81
3.8 Arctan-Daten: Spline s (—) und Funktion g (- - -), optimierte Knoten, RCAP-GP-OD 81
3.9 Volumetric Moisture Content Data: CONCON (—), RCAP-Ka-ED (- - -) . . . 82

4.1 Bivariate Titanium Heat Data: Datenpunkte 99
4.2 Bivariate Titanium Heat Data: Spline s, Startknotenfolge 100
4.3 Bivariate Titanium Heat Data: Optimierte Knoten, NPSOL 100
4.4 Bivariate Titanium Heat Data: Contour-Linien und Knoten vor und nach der Optimierung 101
4.5 EOS Aluminium Daten: Datenpunkte 102
4.6 EOS Aluminium Daten: Spline s, Startknotenfolge 102
4.7 EOS Aluminium Daten: Optimierte Knoten, CONSTR 103
4.8 EOS Aluminium Daten: Contour-Linien und Knoten vor und nach der Optimierung 103
4.9 EOS Aluminium Daten: Verfahren von Walther, monotoner Spline (fit-and-modify) 104

Tabellenverzeichnis

2.1 Rückgabewerte und Abbruchtests 43
2.2 Titanium Heat Data: Stationäre Punkte 44
2.3 Titanium Heat Data: Vergleich des Algorithmus mit MATLAB-Routine und drei verschiedenen Startpunkten 45
2.4 Beispiel von Hu: Knotenreduktionsalgorithmus 47

3.1 Titanium Heat Data: Glättung mit Nebenbedingungen ($\mu = 1.0$) 75
3.2 Titanium Heat Data: CONCON — RCAP-Ka-ED 77
3.3 Arctan-Daten: Splineglättung ($\mu = 1.0$ E-03) 79
3.4 Arctan-Daten: Splineapproximation ($\mu = 0$) 79
3.5 Volumetric Moisture Content Data 82

4.1 Bivariate Titanium Heat Data: Vergleich von CONSTR und NPSOL 99
4.2 EOS Aluminium Daten: Vergleich von CONSTR und NPSOL 101
4.3 Vergleich der Verfahren von Schütze und Walther 104

Bezeichnungen

Univariate Problemstellung:

$\|.\|$	Euklidische Vektornorm
$[a,b]$	$[a,b] \subset \mathbb{R}$, betrachtetes Grundintervall
$L_2[a,b]$	Raum der über $[a,b]$ quadratisch integrierbaren Funktionen
$W_2^q[a,b]$	Sobolew-Raum der Funktionen, deren verallgemeinerte Ableitungen bis zur q-ten Ordnung zum $L_2[a,b]$ gehören
g, s	unbekannte Funktion $g \in W_2^q[a,b]$, Spline $s \in \mathcal{S}_{k,\boldsymbol{\tau}}$
φ, ρ	Approximationsterm, Glättungsterm
μ	Glättungsparameter
ϕ	Schoenberg-Funktional $\phi(s) = \varphi(s) + \mu\rho(s)$
$(\mathbf{x}, \mathbf{y})$	Meßdaten $\{(x_i, y_i) : i = 1, \ldots, m\}$
k, n	Ordnung und Anzahl der B-Splines
$\boldsymbol{\tau}$	Splineknoten $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{n+k})^T \in \mathbb{R}^{n+k}$
$\#\tau_j$	Vielfachheit des Knoten τ_j
$B_{j,k,\boldsymbol{\tau}}$	j-ter polynomialer B-Spline der Ordnung k zu den Knoten $\boldsymbol{\tau}$
$\mathcal{S}_{k,\boldsymbol{\tau}}$	Splineraum
$\boldsymbol{\alpha}$	Splinekoeffizienten $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T \in \mathbb{R}^n$
$\mathbf{B}$	Beobachtungsmatrix $\mathbf{B} = (B_{j,k,\boldsymbol{\tau}}(x_i))_{i=1,\ldots,m;j=1,\ldots,n} \in \mathbb{R}^{m,n}$
$\mathbf{D}_q$	Ableitungsmatrix $\mathbf{D}_q \in \mathbb{R}^{n-q,n}$
r	Ableitungsordnung im Glättungsterm
$\bar{\mathbf{S}}_r, \tilde{\mathbf{S}}_r, \mathbf{S}_r$	(exakte, approximierte) Glättungsmatrix
$\mathbf{B}_\mu$	Systemmatrix $\mathbf{B}_\mu := \begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix} \in \mathbb{R}^{m+n-r,n}$
l, $\mathbf{p}$	Anzahl und Indexvektor der freien Knoten
$\mathbf{t}$	freie Knoten $\mathbf{t} = (\tau_{p(1)}, \ldots, \tau_{p(l)})^T \in \mathbb{R}^l$
$\mathbf{Ct} - \mathbf{h} \geq \mathbf{0}$	Anordnungsnebenbedingungen, $\mathbf{C} \in \mathbb{R}^{ncstr,l}$, $\mathbf{h} \in \mathbb{R}^{ncstr}$
ϵ	(absolutes, relatives) Distanzmaß
$\mathbf{A}^+$	Moore-Penrose-Inverse
$\mathbf{P}_A$	orthogonaler Projektor $\mathbf{P}_A = \mathbf{A}\mathbf{A}^+$
rank, cond	Rang und Konditionszahl einer Matrix
$\mathcal{R}(\mathbf{A})$, $\mathcal{N}(\mathbf{A})$	Spaltenraum, Nullraum der Matrix $\mathbf{A}$
im, ker	Bild, Kern einer Abbildung
$\mathfrak{f}, \mathfrak{F}$	vollständiges Problem: Funktional und Residuumsfunktion
$f, \mathbf{F}$	reduziertes Problem: Funktional und Residuumsfunktion

$\nabla_{\boldsymbol{\alpha}}, \nabla_{\mathbf{t}}$	Gradienten bez. $\boldsymbol{\alpha}$ und $\mathbf{t}$
$\boldsymbol{\partial}$	Operator der Fréchet-Ableitung bez. der freien Knoten, $\boldsymbol{\partial} = \nabla_{\mathbf{t}}^T$
$\mathbf{J}_K$	Kaufman-Approximation an die Jacobi-Matrix $\mathbf{F}'$
μ_{GP}, μ_K	Golub/Pereyra-Modell, Kaufman-Modell
p	Ableitungsordnung in Nebenbedingungen
$l_i^{(p)} \leq s^{(p)}(x) \leq u_i^{(p)}$	Nebenbedingungen an Ableitungen
$\mathbf{L} \leq \mathbf{D}_p(\mathbf{t})\boldsymbol{\alpha} \leq \mathbf{U}$	hinreichende Nebenbedingungen an Ableitungen, $\mathbf{L}, \mathbf{U} \in \mathbb{R}^{n-p}$, $\mathbf{g}(\boldsymbol{\alpha}, \mathbf{t}) := \begin{bmatrix} \mathbf{D}_p(\mathbf{t}) \\ -\mathbf{D}_p(\mathbf{t}) \end{bmatrix} \boldsymbol{\alpha} - \begin{bmatrix} \mathbf{L} \\ -\mathbf{U} \end{bmatrix} \geq \mathbf{0}$
$\mathcal{I}, nact$	Indexmenge und Anzahl aktiver Restriktionen
$\mathbf{R}, \bar{\mathbf{R}}$	$\mathbf{R} := -(\nabla_{\boldsymbol{\alpha}}\mathbf{g})^T \in \mathbb{R}^{2(n-p),n}$, $\bar{\mathbf{R}} := -(\nabla_{\boldsymbol{\alpha}}\mathbf{g}_i)_{i\in\mathcal{I}}^T \in \mathbb{R}^{nact,n}$
$\boldsymbol{\Gamma}, \bar{\boldsymbol{\Gamma}}$	$\boldsymbol{\Gamma} := -(\nabla_{\mathbf{t}}\mathbf{g})^T \in \mathbb{R}^{2(n-p),l}$, $\bar{\boldsymbol{\Gamma}} := -(\nabla_{\mathbf{t}}\mathbf{g}_i)_{i\in\mathcal{I}}^T \in \mathbb{R}^{nact,l}$
$\mathbf{N}$	$\mathbf{N} \in \mathbb{R}^{n,n-nact}$ Nullraumbasis von $\bar{\mathbf{R}}$
macheps	Maschinenepsilon

Bivariate Problemstellung:

$[a_1, b_1] \times [a_2, b_2]$	$[a_1, b_1] \times [a_2, b_2] \subset \mathbb{R}^2$, betrachteter Bereich
$W_2^{q_1,q_2}[a_1, b_1] \times [a_2, b_2]$	Sobolew-Raum der Funktionen, deren gemischte verallgemeinerte partiellen Ableitungen bis zur Ordnung q_1 bzw. q_2 in x und y aus L_2 sind
D^{r_1,r_2}	Ableitungsoperator $\frac{\partial^{r_1+r_2}}{\partial x^{r_1} \partial y^{r_2}}$
$(\mathbf{x}, \mathbf{y}, \mathbf{Z})$	Meßdaten $\{(x_{i_1}, y_{i_2}, z_{i_1,i_2}) : i_1 = 1, \ldots, m_1; i_2 = 1, \ldots, m_2\}$
$\mathbf{A}$	Splinekoeffizienten $\mathbf{A} = (\alpha_{j_1,j_2}) \in \mathbb{R}^{n_1,n_2}$
vec ()	vec ()-Operator, spaltenweise Anordnung einer Matrix
$\|.\|_F$	Frobeniusnorm einer Matrix
tr	Spur einer Matrix
$s^1 \otimes s^2$, $\mathcal{S}_1 \otimes \mathcal{S}_2$	Tensorprodukt zweier Funktionen, Funktionenräume
$\mathbf{A} \otimes \mathbf{B}$	Kronecker-Produkt zweier Matrizen

Alle anderen Bezeichnungen werden aus dem univariaten Teil übernommen und durch entsprechende Indizes gekennzeichnet.

Kapitel 1

Einleitung

Although this may seem a paradox, all exact science is dominated by the idea of approximation.

Bertrand Russell

Die Approximation von Daten und Funktionen ist von großer praktischer Bedeutung in der Mathematik und den Naturwissenschaften. In den letzten Jahren rückte dabei die formerhaltende (shape preserving) Approximation besonders in den Mittelpunkt. Diese Approximation ist nicht nur aus innermathematischer Sicht heraus interessant: In vielen Anwendungen ist die Einhaltung bestimmter Nebenbedingungen wie Nichtnegativität, Monotonie oder Konvexität wesentlich für ein physikalisch oder technisch sinnvolles Ergebnis.

Gegeben seien Daten $\{(x_i, y_i) : i = 1, \ldots, m\}$, wobei $\{x_i\}$ streng monoton steigende Abszissen mit $a \le x_1 < \cdots < x_i < \cdots < x_m \le b$ und $\{y_i\}$ fehlerbehaftete Meßwerte einer unbekannten glatten Funktion $g \in W_2^q[a, b]$ sind, d. h. $y_i = g(x_i) + \epsilon_i$, $i = 1, \ldots, m$. Die zufälligen Fehler ϵ_i seien stochastisch unabhängig und identisch verteilt.

Die Funktion g bzw. die Daten $\{x_i, y_i\}$ sollen durch eine Funktion s aus einem geeignet gewählten Teilraum $\mathcal{S} \subset W_2^q[a, b]$ approximiert werden. An diese Funktion s stellen wir folgende Forderungen:

(i) Die Funktion und ihre Ableitungen können leicht ausgewertet werden.

(ii) Die Funktion kann durch wenige Parameter beschrieben werden (Datenreduktion!).

(iii) Die Funktion ist glatt und repräsentiert die Funktion g innerhalb des Fehlerniveaus der Daten.

(iv) Die Funktion ist eine formerhaltende Approximation, z. B. $s^{(p)}(x) \ge 0$ für alle $x \in [a, b]$, falls diese Eigenschaft von g bekannt ist.

Unter den Voraussetzungen $\mathrm{E}\,\epsilon_i = 0$ und $\mathrm{Var}\,\epsilon_i = \sigma^2 > 0$, $i = 1, \ldots, m$, an die Fehler stellt der *Approximationsterm*

$$\varphi(s) := \frac{1}{2} \sum_{i=1}^{m} [y_i - s(x_i)]^2 ,$$

d. h. die Fehlerquadratsumme, ein natürliches Maß für die Approximationsgüte dar. Eine Funktion s ist als zulässig anzusehen, wenn der Approximationsterm $\varphi(s)$ eine bestimmte Toleranz φ_{tol} der Größenordnung $m\sigma^2$ nicht überschreitet, d. h. es gelte

$$\varphi(s) \leq \varphi_{tol}. \tag{1.1}$$

Bei der praktischen Anwendung von Approximationsschemata ist oftmals ein Kompromiß zwischen der Güte der Approximation und der Glattheit erforderlich. Wir betrachten daher zusätzlich den *Glättungsterm*

$$\rho(s) := \frac{1}{2}\left\|s^{(r)}\right\|_{L_2}^2 = \frac{1}{2}\int_a^b \left[s^{(r)}(x)\right]^2 dx$$

mit festem $r \in \{0, \ldots, q\}$. Die Größe des Glättungsterms wird als Kriterium zur Auswahl einer geeigneten Approximation benutzt.

Bei der Behandlung obiger Approximationsprobleme haben Splines eine weite Verbreitung gefunden. Zur Definition eines Splines werden sog. Splineknoten benötigt. Der Einfachheit halber werden diese Splineknoten oftmals äquidistant im Approximationsintervall verteilt oder interaktiv durch den Anwender vorgegeben. Es ist wohlbekannt, daß man in solchen Fällen die Approximationsgüte wesentlich verbessern kann, wenn man die Lage der Knoten als freie Parameter betrachtet und zusätzlich in den Optimierungsprozeß einbezieht. Gegenstand dieser Arbeit ist daher die numerische Berechnung von Splines mit freien Knoten unter besonderer Berücksichtigung der formerhaltenden Approximation.

Wir wollen im folgenden zur Motivation unseres direkten Zugangs den sog. Variationszugang zur Splinetheorie kurz darstellen:

Im Fall der unrestringierten Approximation betrachtet man das Problem

$$\min\left\{\rho(s) \,:\, \varphi(s) \leq \varphi_{tol}, s \in W_2^r[a,b]\right\}. \tag{1.2}$$

Es ist bekannt (Schoenberg [Sch64], Reinsch [Rei67, Rei71]), daß das Variationsproblem (1.2) für jedes $\varphi_{tol} \geq 0$ Lösungen besitzt und daß alle Lösungen natürliche Splines der Ordnung $k = 2r$ mit Knoten an den Datenstellen $\{x_i\}$ sind. Falls das Regressionspolynom der Ordnung r die Bedingung (1.1) nicht erfüllt, so ist überdies die Lösung von (1.2) eindeutig und löst das unrestringierte Problem

$$\min\left\{\phi(s) := \varphi(s) + \mu\rho(s) \,:\, s \in W_2^r[a,b]\right\}. \tag{1.3}$$

Der zugehörige Lagrange-Parameter oder *Glättungsparameter* $\mu > 0$ ist eindeutig festgelegt durch die Gleichung $\varphi(s_\mu) = \varphi_{tol}$, s_μ eindeutige Lösung von (1.3) für $\mu > 0$. Das Funktional $\phi : s \in W_2^r[a,b] \to \mathbb{R}$ mit

$$\phi(s) := \varphi(s) + \mu\rho(s) = \frac{1}{2}\sum_{i=1}^{m}\left[y_i - s(x_i)\right]^2 + \mu\frac{1}{2}\int_a^b \left[s^{(r)}(x)\right]^2 dx \tag{1.4}$$

wird als *Schoenberg-Funktional* bezeichnet.

Im Fall der Interpolation mit Nebenbedingungen betrachtet man

$$\min\left\{\rho(s) \,:\, \varphi(s) = 0, s^{(p)}(x) \geq 0 \text{ für alle } x \in [a,b], s \in W_2^r[a,b]\right\}.$$

Existieren Lösungen dieses Problems, so erhält man erneut Splines der Ordnung $k = 2r$, allerdings mit zusätzlichen Knoten zwischen den Datenstellen $\{x_i\}$, siehe etwa [OO88], [FOP91] und [AE95] für nichtnegative Splines und Splines mit Hindernissen ($p = 0$), [Hor78] und [AE91] für monotone Splines ($p = 1$) sowie [AE87] für den konvexen Fall ($p = 2$). Allgemeine Charakterisierungsaussagen finden sich in [MU88], [MU91] und [MSSW85]. Im Fall der Approximation mit Nebenbedingungen betrachten [EA88] das Variationsproblem

$$\min \left\{ \varphi(s) + \mu\rho(s) \,:\, s^{(p)}(x) \geq 0 \text{ für alle } x \in [a,b], s \in W_2^r[a,b] \right\}.$$

Die oben beschriebenen Variationszugänge haben Gemeinsamkeiten, die sich aber im Rahmen unserer Zielstellungen als Nachteile erweisen:

- Die Ordnung $k = 2r$ des Splines ist durch die Ordnung r im Glättungsterm fest vorgegeben.
- In Regionen, in denen keine Nebenbedingungen aktiv sind, werden alle Datenstellen $\{x_i\}$ als Knoten des Splines benötigt, d. h. es ist keine Datenreduktion möglich. Im praktisch relevanten Fall großer Datenmengen kann man aber oft schon mit wenigen Parametern die Funktion innerhalb des Fehlerniveaus der Daten beschreiben.
- Im restringierten Fall hat die Lösung im allgemeinen zusätzliche Knoten zwischen den Datenstellen $\{x_i\}$, deren Anzahl und Lage a priori nicht bekannt ist.

Daher favorisieren wir den sog. direkten Zugang zur Splinetheorie und beschränken uns von vornherein auf den endlich-dimensionalen Raum $\mathcal{S}_{k,\boldsymbol{\tau}}$ der polynomialen Splines der Ordnung k zu einer gegebenen Knotenfolge $\boldsymbol{\tau} = \{\tau_j\}$ mit

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

und $m \geq n$. Im allgemeinen ist die Dimension n des Raums $\mathcal{S}_{k,\boldsymbol{\tau}}$ sehr viel kleiner als die Anzahl m der Datenpunkte, d. h. man erreicht durch den direkten Zugang eine Datenreduktion. Außerdem ist damit ein gewisser Regularisierungseffekt verbunden. Die Parameter des Splines werden so bestimmt, daß das Schoenberg-Funktional (1.4) minimiert wird. Als Basis für den Raum $\mathcal{S}_{k,\boldsymbol{\tau}}$ verwenden wir die bekannten polynomialen B-Splines. Ein Spline $s \in \mathcal{S}_{k,\boldsymbol{\tau}}$ ist dann durch Splineknoten $\boldsymbol{\tau}$ und Koeffizienten $\boldsymbol{\alpha}$ eindeutig bestimmt.

Die Lage der Knoten $\boldsymbol{\tau}$ ist für die Approximationsgüte von großer Bedeutung. Es ist wohlbekannt, daß man die Approximationsgüte, insbesondere bei nichtglatten Funktionen, verbessern kann, wenn man die Knoten als freie Parameter betrachtet, siehe [Sch81, Kapitel 7].

Gegenstand dieser Arbeit ist daher die Minimierung des Schoenberg-Funktionals (1.4) bei vorgegebener Dimension n bezüglich der Koeffizienten $\boldsymbol{\alpha}$ und einer Teilmenge $\mathbf{t}$ der Knoten $\boldsymbol{\tau}$, den sog. *freien Knoten*, unter Beachtung eventueller Nebenbedingungen an Ableitungen des Splines. Das resultierende Optimierungsproblem ist ein nichtlineares Quadratmittelproblem in den freien Knoten und den Koeffizienten. Bei der Verwendung von B-Splines haben die auftretenden Matrizen Bandstruktur. Es zeigt sich, daß Standardprogramme der nichtlinearen Optimierung für die Berechnung von Splines mit freien Knoten ohne weitere Modifikationen versagen oder sehr ineffizient sind. Unser Ziel ist die Entwicklung spezieller Verfahren für diese Problemstellung.

Bei der Minimierung des Schoenberg-Funktionals bei *festen* Knoten entsteht ein lineares Quadratmittelproblem in den Koeffizienten $\boldsymbol{\alpha}$, dessen Lösung effizient berechnet werden kann. Für die Berechnung solcher Splines mit festen Knoten existieren eine Reihe ausgefeilter Algorithmen, siehe [dB78], [Die87] sowie [SK93].

Das Ausgangsproblem der Minimierung des Schoenberg-Funktionals bezüglich der Koeffizienten *und* der freien Knoten ist ein *separables Quadratmittelproblem*, falls keine Nebenbedingungen an Ableitungen auftreten, bzw. ein Spezialfall sog. *restringierter semi-linearer Quadratmittelprobleme* bei Nebenbedingungen an Ableitungen. Grundlage für Lösungsverfahren solcher spezieller nichtlinearer Optimierungsprobleme ist die Tatsache, daß man die optimalen Koeffizienten $\boldsymbol{\alpha}$ bei festen Parametern $\mathbf{t}$ explizit darstellen bzw. sehr effizient berechnen kann. Setzt man jetzt diese optimalen Koeffizienten $\boldsymbol{\alpha}$ in das Ausgangsproblem ein, so erhält man ein *reduziertes Problem*, in welchem nur noch die freien Knoten $\mathbf{t}$ als Variable auftreten.

Nachdem wir die Existenz von Lösungen des reduzierten Problems gezeigt haben, weisen wir unter Benutzung von Ergebnissen der Theorie dieser speziellen Optimierungsprobleme die Äquivalenz von Ausgangsproblem und reduziertem Problem im Fall der Splineapproximation nach. Ein Schwerpunkt der Arbeit liegt dann in der numerischen Lösung des reduzierten Problems, welches ein nichtlineares Quadratmittelproblem mit linearen Ungleichheitsnebenbedingungen ist. Dazu setzen wir ein verallgemeinertes Gauß-Newton-Verfahren ein. Da die benötigte Jacobi-Matrix eine sehr komplizierte Struktur hat, verwenden wir eine billiger zu berechnende Approximation. Bei der Berechnung der Residuumsfunktion und der Jacobi-Matrix bzw. deren Approximation wird die Bandstruktur der Matrizen ausgenutzt. Dies und die Verwendung stabiler Orthogonalisierungstechniken führt schließlich zu einem robusten und effizienten Verfahren zur Berechnung von Splines mit freien Knoten, welches auch bei großen Datenmengen praktikabel ist.

Typische Anwendungsgebiete für unsere Verfahren sind Aufgaben, bei denen der Spline als Ausgangspunkt für weitere kompliziertere oder häufig wiederkehrende Rechnungen dient, etwa Inverse Probleme, Optimierung und Optimale Steuerung sowie Implementierung in Hardware. Als Beispiele seien genannt Parameterschätzung in Differentialgleichungen [Var82], Verfahren zur Datenreduktion, z. B. bei der Kennlinienspeicherung in Mikroprozessoren [Wev89], sowie „self-modeling free-knot splines“ zur statistischen Analyse biologischer Modelle [SB92]. Weitere technische Anwendungen findet man in [Wev89].

Es sei betont, daß die Methoden nicht für das „bloße Visualisieren“ von glatten Funktionen, welche an äquidistanten Punkten ohne stochastische Fehler berechnet werden, gedacht sind. Der Aufwand zur Lösung der nichtlinearen Quadratmittelprobleme ist dann im allgemeinen zu hoch im Vergleich zum Qualitätszuwachs des Graphen des Splines. Ziel ist stattdessen eine möglichst gute Approximation von fehlerbehafteten Daten, die durch *wenige* Parameter beschrieben wird. Durch die Verwendung des regularisierenden Glättungsterms sind dabei auch Daten mit Lücken zugelassen, d. h. bestimmte Bereiche des Intervalls $[a, b]$, an denen – etwa aus technischen Gründen – keine Messung erfolgen kann. Da wir an einer Approximation mit möglichst wenig Knoten interessiert sind, benutzen wir auch nicht die Aussagen zur asymptotisch optimalen Knotenverteilung, siehe [BS78] im univariaten Fall und [DYS91] im Fall von bivariaten Tensorprodukt-Splines, sondern minimieren direkt das Schoenberg-Funktional.

Die Arbeit ist wie folgt aufgebaut: In Kapitel 2 betrachten wir zunächst die Minimierung des univariaten Schoenberg-Funktionals (1.4) ohne Nebenbedingungen an Ableitungen. Zu diesem Gebiet liegen bereits von anderen Autoren Arbeiten vor. Durch die Verwendung

des Glättungsfunktionals an Stelle des Quadratmittelfehlers konnte jedoch erstmalig die Äquivalenz von vollständigem und reduziertem Problem unabhängig von den Knoten gezeigt werden. Unter Benutzung eines verallgemeinerten Gauß-Newton-Verfahrens werden sowohl die linearen Nebenbedingungen, welche das Zusammenfallen von Knoten verhindern, direkt behandelt als auch die Quadratmittelstruktur ausgenutzt. In der ganzen Arbeit haben wir großen Wert auf die Verwendung numerisch stabiler Orthogonalisierungstechniken gelegt, welche eine weitestgehende Ausnutzung der Schwachbesetztheit gestatten. Gegenüber den in [SS95] veröffentlichten Ergebnissen aus Kapitel 2 wurden noch wesentliche algorithmische Verbesserungen (exakte Ableitungen) und eine durchgängige Darstellung erreicht.

Kapitel 3 beschäftigt sich mit der Minimierung von (1.4) unter Nebenbedingungen der Form

$$l_i^{(p)} \le s^{(p)}(x) \le u_i^{(p)} \qquad \text{für alle } x \in [\tau_i, \tau_{i+1}),\ i = k, \ldots, n$$

mit festem $p \in \{0, \ldots, q\}$. In diesem Kapitel sehen wir den Hauptbeitrag der Dissertation. Uns sind bisher keine Arbeiten bekannt, die sich mit der direkten Minimierung des Schoenberg-Funktionals (oder des Quadratmittelfehlers) als Funktion der freien Knoten unter Nebenbedingungen an Ableitungen befassen. Die aus den separablen Quadratmittelproblemen bekannte Kaufman-Approximation wird auf den restringierten Fall verallgemeinert. Ein Verfahren für restringierte semi-lineare Quadratmittelprobleme mit dieser Approximation wird entwickelt und numerisch getestet. Die Hauptergebnisse aus Kapitel 3 wurden in [SS97] veröffentlicht, die Aussagen zur Strukturausnutzung bei der Berechnung der Jacobi-Matrix werden hier vertieft und ergänzt.

In Kapitel 4 wird schließlich die Problemstellung auf die bivariate Approximation von Daten auf Rechteckgittern durch Tensorprodukt-Splines verallgemeinert, zunächst ohne die Betrachtung von Nebenbedingungen an Ableitungen. Diese Verallgemeinerung ebenso wie die auf den Fall von Tensorprodukt-Splines mit unregelmäßig verteilten Daten kann ohne Probleme durchgeführt werden, sofern man einen separablen Glättungsterm benutzt.

Eine naheliegende weitere Aufgabenstellung wäre nun die Quadratmittelapproximation durch Tensorprodukt-Splines mit freien Knoten und Ungleichheitsnebenbedingungen an Ableitungen. Im Gegensatz zu den bisher beschriebenen Fällen sind hier selbst Untersuchungen zu Splines mit festen Knoten noch nicht in der Literatur vorhanden. Kapitel 5 enthält Anregungen zur Behandlung dieses Problems und faßt die erreichten Ergebnisse der Dissertation zusammen. Wir möchten anmerken, daß selbst im Fall einer erfolgreichen theoretischen Behandlung der bivariaten restringierten Approximation mit freien Knoten eine erfolgreiche numerische Behandlung des Problems schwierig und teuer ist, da die auftretenden linearen Probleme nicht mehr zerfallen.

Jedes der Kapitel 2 bis 4 hat die folgende Struktur: Nachdem wir zur Einordnung unserer Methode einen Überblick über existierende bzw. ähnliche Verfahren in der Literatur gegeben haben, widmen wir uns der Formulierung des vollständigen Problems in Abhängigkeit von den Knoten und Koeffizienten. Danach stellen wir die benötigten Resultate für allgemeine Quadratmittelprobleme dieser speziellen Struktur zusammen, d. h. unabhängig vom Kontext der Splineapproximation. Anschließend wenden wir diese Ergebnisse auf den Fall der Splineapproximation an. Die Entwicklung eines Algorithmus zur numerischen Lösung des reduzierten Problems unter Ausnutzung der Schwachbesetztheitsstruktur stellt jeweils einen Schwerpunkt der Kapitel dar. Abschließend wird unser Verfahren ausgiebig an Beispielen aus der Literatur und selbstkonstruierten Beispielen getestet.

Kapitel 2

Univariate Splines

2.1 Einleitung und historischer Überblick

Es gibt eine Fülle von Arbeiten zur Bestapproximation von Funktionen durch Splines mit freien Knoten, insbesondere zur Chebyshev-Approximation. Eine vollständige Charakterisierung von (eindeutigen) Bestapproximationen ist jedoch auch im Fall der Chebyshev-Approximation durch Splines mit freien Knoten bisher nicht bekannt [Nür96]. Zur numerischen Berechnung einer *guten* Splineapproximation mit freien Knoten wird ein zweistufiges Verfahren vorgeschlagen, siehe [MNSS89]. Im ersten Schritt werden freie Knoten bei stückweisen Polynomen (ohne Glattheitsforderungen) bestimmt. Die als Ergebnis dieses Segmentapproximationsproblems erhaltenen guten Knoten verwendet man nun zur Konstruktion eines bestapproximierenden Splines mit festen Knoten. Es sei bemerkt, daß die Chebyshev-Approximation im Rahmen unserer statistischen Grundvoraussetzungen ungeeignet ist, da sie sehr empfindlich auf Ausreißer reagiert.

Einfache Beispiele zeigen, daß das Problem der Bestapproximation durch Splines mit freien Knoten in der Menge der Splines mit einfachen Knoten nicht immer lösbar ist. Dagegen existiert stets eine Lösung, wenn man mehrfache Knoten zuläßt. Das erste Existenztheorem geht auf [Ric69] zurück und besagt, daß für jede Funktion $g \in L_p[a,b]$, $1 \le p \le \infty$, eine beste L_p-Approximation im Raum der Splines mit freien (eventuell mehrfachen) Knoten existiert.

Charakteristisch für die Approximation durch Splines mit freien Knoten – oder allgemeiner die Approximation mit γ-Polynomen – ist das „Lethargie-Syndrom“ [Jup75]. Die Konsequenzen dieser Eigenschaft sind

- die Existenz vieler stationärer Punkte von φ auf dem Rand des zulässigen Bereichs $\left\{\boldsymbol{\tau} \in \mathbb{R}^{n+k} : \tau_1 = \cdots = \tau_k = a < \tau_{k+1} \le \cdots \le \tau_n < b = \tau_{n+1} = \cdots = \tau_{n+k}\right\}$,
- die Nichtkonvexität von φ als Funktion der Knoten sowie
- das schlechte Konvergenzverhalten von Algorithmen in der Nähe des Randes des zulässigen Bereichs.

Im Gegensatz zur Approximation von Funktionen durch Splines mit freien Knoten wollen wir uns in dieser Arbeit mit der Approximation von Daten beschäftigen. In dem Vorliegen von lediglich diskreter Information liegt eine zusätzliche Schwierigkeit begründet: Beim Fehlen von Information, d. h. fehlende Datenpunkte $\{x_i, y_i\}$ innerhalb von einigen Intervallen,

kann eine gewisse Regularitätsbedingung (Schoenberg-Whitney-Bedingung, siehe (2.3)) und damit die Eindeutigkeit des Splines zu festen Knoten nicht garantiert werden. Arbeiten, die sich mit der numerischen Berechnung freier Knoten bei der diskreten Quadratmittelapproximation von Daten beschäftigen, gibt es daher vergleichsweise wenig.

Einige Eigenschaften der Approximation durch Splines mit freien Knoten – wie das Lethargie-Syndrom – übertragen sich jedoch vom kontinuierlichen auf den diskreten Fall. Die Vermeidung von (fast) zusammenfallenden Knoten ist deshalb eine Gemeinsamkeit vieler Algorithmen. Man kann derartige Algorithmen zur Quadratmittelapproximation durch Splines mit freien Knoten und zur Datenreduktion grob in verschiedene Klassen einteilen:

Einige Verfahren beginnen mit wenigen Knoten und fügen iterativ Knoten nach geeigneten Regeln ein, bis der resultierende Spline eine hinreichend gute Qualität hat. Zu dieser Klasse gehört de Boor's Algorithmus NEWKNOT [dB78, S. 184ff], bei welchem die Knoten so gewählt werden, daß eine von der k-ten Ableitung von g abhängige Indikatorfunktion gleichverteilt wird. Während de Boor eine stückweise konstante Approximation an die unbekannte Funktion $g^{(k)}$ benutzt, wird in der Arbeit [Hu93] diese Funktion in einem vorbereitenden Schritt durch einen Spline höherer Ordnung mit vielen Knoten bestimmt. Algorithmen dieser Art beruhen auf lokalen Fehlerabschätzungen bzw. auf Aussagen zur asymptotisch optimalen Knotenanordnung.

In einer zweiten Klasse von Algorithmen startet man mit einer großen Anzahl von Knoten, welche meist aus der Menge der Datenstellen $\{x_i\}$ gewählt werden. Dann wird mittels einer geeigneten Wichtung entschieden, welche Knoten weniger bedeutsam für den Approximationsfehler sind. Diese werden dann iterativ entfernt. Vertreter dieser Klasse sind die Verfahren von Lyche/Mørken [LM88] und das neue Verfahren von Schumaker/Stanley [SS96], siehe auch [ADLM90] und [LM87].

Während die obigen Algorithmen keine *optimale* Plazierung der Knoten bezüglich des Quadratmittelfehlers φ liefern, minimieren die folgenden Methoden diesen Fehler direkt für eine fest vorgegebene Anzahl von Knoten. In der Einbeziehung der Nebenbedingung $\tau_j < \tau_{j+1}$, $j = k, \ldots, n$, welche das Zusammenfallen von Knoten verhindern soll, unterscheiden sich die Verfahren.

In einer ersten Arbeit [dBR68] minimieren die Autoren zyklisch den L_2-Fehler im Intervall

$$[\tau_{j-1} + \epsilon(\tau_{j+1} - \tau_{j-1}), \tau_{j+1} - \epsilon(\tau_{j+1} - \tau_{j-1})], \qquad \epsilon = 0.0625, j = k+1, \ldots, n,$$

als Funktion des Knoten τ_j. Zur Lösung dieser eindimensionalen Optimierungsaufgabe wird das Newton-Verfahren eingesetzt, als Basis für den Splineraum verwendet man stückweise Polynome. Jupp [Jup78] optimiert die Knoten simultan und benutzt die Technik der variablen Projektion aus [GP73] zur Elimination der linearen Variablen. Mittels einer nichtlinearen Transformation der Knoten

$$\mu_j := \log \frac{\tau_{j+1} - \tau_j}{\tau_j - \tau_{j-1}}, \qquad j = k+1, \ldots, n,$$

wird der Rand des zulässigen Bereichs ins Unendliche transformiert. Jupp zeigt, daß diese Transformation in gewisser Weise die Schwierigkeiten der originalen Formulierung abschwächt. Bei der Lösung des unrestringierten Problems werden Quasi-Newton-Verfahren und Gauß-Newton-Techniken verglichen. Dierckx beschreibt in seinem Buch [Die93] ein Verfahren, welches ursprünglich in [Die79] entwickelt wurde und ebenfalls die linearen Variablen

eliminiert. Er verwendet einen Barriereterm der Form

$$P(\boldsymbol{\tau}) := \sum_{j=k}^{n} \frac{1}{\tau_{j+1} - \tau_j} \qquad \text{(inverse Barrierefunktion)}$$

zur Transformation in ein unrestringiertes Optimierungsproblem, d. h. es wird

$$\xi(\boldsymbol{\tau}) := \varphi(\boldsymbol{\tau}) + pP(\boldsymbol{\tau}) = \underbrace{\frac{1}{2}\sum_{i=1}^{m}[y_i - s(x_i)]^2}_{=:\varphi(\boldsymbol{\tau})} + p\underbrace{\sum_{j=k}^{n}\frac{1}{\tau_{j+1}-\tau_j}}_{=:P(\boldsymbol{\tau})} \to \min$$

betrachtet. Der Strafparameter p wird heuristisch gewählt

$$p = \epsilon_1 \frac{\varphi(\boldsymbol{\tau}^0)}{P(\boldsymbol{\tau}_{equi})}, \qquad \begin{array}{lll} \epsilon_1 & - & \text{relative Genauigkeit,} \\ \boldsymbol{\tau}^0 & - & \text{Startpunkt für die Knoten,} \\ \boldsymbol{\tau}_{equi} & - & \text{äquidistante Knoten.} \end{array}$$

Zur Minimierung von ξ wird das Fletcher/Reeves CG-Verfahren verwendet. Man beachte, daß obige Wahl des Strafparameters keine Konvergenz zu einer Lösung des restringierten Problems im Sinne der Optimierungstheorie sichert. Außerdem wird die ursprüngliche Quadratmittelstruktur nicht ausgenutzt. Im Rahmen der Untersuchung nichtlinearer Quadratmittelprobleme mit speziell strukturierten Nebenbedingungen (Schranken an die Variablen und Anordnungsnebenbedingungen) betrachten Holt/Fletcher [HF79] ebenfalls Splines mit freien Knoten, allerdings ohne eine Separation von linearen und nichtlinearen Variablen vorzunehmen.

Bei den obigen Verfahren zur direkten Minimierung des Quadratmittelfehlers φ werden stets alle inneren Splineknoten in den Optimierungsprozeß einbezogen. Außerdem ist die erwähnte Reduktion auf ein Problem, in welchem nur die nichtlinearen Parameter $\boldsymbol{\tau}$ auftreten, nur zulässig, falls die Schoenberg-Whitney-Bedingung erfüllt ist. In den Verfahren wird explizit – manchmal auch stillschweigend – angenommen, daß diese Regularitätsbedingung für alle während des Optimierungsprozesses auftretenden Splineknoten erfüllt ist. Suchomski [Suc91] formuliert dagegen die Schoenberg-Whitney-Bedingung direkt als zusätzliche Nebenbedingung an die Knoten. Er betrachtet allerdings eine Art simultaner Interpolation und Approximation, d. h. von den m Daten $\{x_i, y_i\}$ werden n ausgewählte Daten interpoliert. Man beachte, daß diese Problemformulierung unsachgemäß ist, falls alle Datenwerte $\{y_i\}$ fehlerbehaftet sind und keine durch einen kleineren stochastischen Fehler ausgezeichnet sind. Mittels einer nichtlinearen Transformation ähnlich der von Jupp wird das Problem in ein unrestringiertes Quadratmittelproblem überführt.

Wir werden in diesem Kapitel ein Verfahren entwickeln, welches

- keine Schoenberg-Whitney-Bedingung erfordert, sondern durch den Übergang von

$$\varphi(s) := \frac{1}{2}\sum_{i=1}^{m}[y_i - s(x_i)]^2 \to \min$$

zu

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

unabhängig von den Knoten stets die Zulässigkeit der Reduktionstechnik sichert.

- die linear restringierten nichtlinearen Quadratmittelprobleme direkt behandelt, d. h. ohne eine künstliche Transformation in ein unrestringiertes Problem wie die logarithmische Transformation von Jupp. Für die letztere ist bekannt, daß die Umkehrfunktion in sehr hoher Genauigkeit ausgeführt werden muß, da sonst ein drastischer Verlust an numerischer Genauigkeit auftritt. Dies wird durch die numerischen Tests des Autors belegt.

- nur die Lage einer Teilmenge der inneren Knoten optimiert. Diese Option ist hilfreich bei der Approximation von verschiedenen Datensätzen, welche sich nur in wenigen Intervallen voneinander unterscheiden.

2.2 Problemformulierung

Aus den in Kapitel 1 erwähnten Gründen wählen wir für den Ansatzraum $\mathcal{S}$ a priori den Raum $\mathcal{S}_{k,\boldsymbol{\tau}}$ der polynomialen Splines der Ordnung k zur Knotenfolge $\boldsymbol{\tau}$. Als Basis für diesen Raum verwenden wir B-Splines. Zunächst wollen wir einige benötigte Bezeichnungen aus der Splinetheorie bereitstellen.

2.2.1 Bezeichnungen und Grundlagen aus der Splinetheorie

Definition 2.1 (Knotenfolge)**.** Eine Folge $\boldsymbol{\tau} = (\dots, \tau_{-1}, \tau_0, \tau_1, \dots)$ mit $\tau_j \in \mathbb{R}$, $\tau_j \le \tau_{j+1}$ für alle $j \in \mathbb{Z}$ und $\lim\limits_{j\to\pm\infty} \tau_j = \pm\infty$ heißt biinfinite Knotenfolge.

Definition 2.2 (Normalisierte polynomiale B-Splines)**.** Der j-te normalisierte polynomiale B-Spline der Ordnung k für die Knotenfolge $\boldsymbol{\tau}$ wird mit $B_{j,k,\boldsymbol{\tau}}$ bezeichnet und definiert durch:

$$B_{j,1,\boldsymbol{\tau}}(x) := \begin{cases} 1, & \text{falls } \tau_j \le x < \tau_{j+1}; \\ 0, & \text{sonst;} \end{cases}$$

$$B_{j,k,\boldsymbol{\tau}}(x) := \omega_{j,k}(x) \cdot B_{j,k-1,\boldsymbol{\tau}}(x) + (1 - \omega_{j+1,k}(x)) \cdot B_{j+1,k-1,\boldsymbol{\tau}}(x) \quad \text{für } k > 1$$

mit

$$\omega_{j,k}(x) := \begin{cases} \dfrac{x - \tau_j}{\tau_{j+k-1} - \tau_j}, & \text{falls } \tau_j < \tau_{j+k-1}; \\ 0, & \text{sonst.} \end{cases}$$

Die Bezeichnung $B_{j,k,\boldsymbol{\tau}}$ deutet dabei an, daß die B-Splines von der Knotenfolge $\boldsymbol{\tau}$ abhängig sind. Diese Abhängigkeit ist nichtlinear. Wenn aus dem Zusammenhang deutlich wird, auf welche Knotenfolge sich der B-Spline bezieht, verzichten wir gegebenenfalls auf die gesonderte Kennzeichnung der Knotenfolge.

Für $\tau_j = \dots = \tau_{j+k}$ gilt offensichtlich $B_{j,k,\boldsymbol{\tau}} \equiv 0$. Es ist deshalb vernünftig, $\tau_j < \tau_{j+k}$ für alle $j \in \mathbb{Z}$ zu fordern, um die trivialen, identisch verschwindenden B-Splines von vornherein von der Betrachtung auszuschließen. Der Knoten τ_j hat die Vielfachheit $\nu_j = \#\tau_j$, wenn $\exists i : \tau_i < \tau_j = \tau_{i+1} = \dots = \tau_{i+\nu_j} < \tau_{i+\nu_j+1}$. Die Bedingung $\tau_j < \tau_{j+k}$ ist damit äquivalent zu $\#\tau_j < k$ für alle $j \in \mathbb{Z}$. Nun sind wir in der Lage, den Splineraum $\mathcal{S}_{k,\boldsymbol{\tau}}$ zu definieren.

Definition 2.3 (Spline, Splineraum)**.** Die Menge $\mathcal{S}_{k,\boldsymbol{\tau}} := \left\{ \sum_j B_{j,k,\boldsymbol{\tau}} \alpha_j \, , \, \alpha_j \in \mathbb{R} \right\}$ heißt Splineraum der Ordnung k bezüglich der Knotenfolge $\boldsymbol{\tau}$. Die Elemente $s \in \mathcal{S}_{k,\boldsymbol{\tau}}$ heißen Splines.

Der folgende klassische Satz der Splinetheorie zeigt den Zusammenhang zwischen Knotenvielfachheit und Glattheit sowie die lineare Unabhängigkeit der B-Splines.

Satz 2.1 (Curry, Schoenberg, zitiert nach [dBH87])**.**
Wenn $\tau_j < \tau_{j+k}$ für alle j gilt, dann sind die B-Splines $B_{j,k,\boldsymbol{\tau}}$ linear unabhängig und bilden eine Basis für den Raum $\tilde{\mathcal{S}}$ aller stückweisen Polynome vom Grade kleiner als k mit Bruchstellen τ_j, welche in den Bruchstellen $(k-1-\#\tau_j)$-mal stetig differenzierbar sind. Es gilt:

Anzahl der Glattheitsbedingungen in τ_j + Knotenvielfachheit $\#\tau_j$ = Ordnung k.

Spezialisierung auf ein endliches Intervall

Da jeder B-Spline wegen $\operatorname{supp} B_{j,k,\boldsymbol{\tau}} = [\tau_j, \tau_{j+k}]$ nur einen endlichen Träger besitzt, genügt es, sich auf ein endliches Intervall zu beschränken. Wir betrachten daher im folgenden ausschließlich die (finite) Knotenfolge $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{n+k})^T \in \mathbb{R}^{n+k}$. Dabei sei n eine fest vorgegebene Anzahl von B-Splines.

Später werden wir B-Splines benutzen, um gegebene Daten $\{x_i, y_i\}$, $i = 1, \ldots, m$, zu approximieren. Die Meßstellen x_i befinden sich innerhalb des gegebenen Intervalls $[a, b] \subset \mathbb{R}$. Wir beschränken uns daher auf die Knoten

$$\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{n+k})^T \in \mathbb{R}^{n+k} \tag{2.1a}$$

mit

$$\tau_1 \leq \cdots \leq \tau_k \leq a \quad \text{und} \quad b \leq \tau_{n+1} \leq \cdots \leq \tau_{n+k} \tag{2.1b}$$

sowie

$$a < \tau_{k+1} \leq \cdots \leq \tau_n < b. \tag{2.1c}$$

Die Knoten $\tau_{k+1}, \ldots, \tau_n$ werden als *innere Knoten* bezeichnet. Es gelte stets $\tau_j < \tau_{j+k}$, $j = 1, \ldots, n$, d. h. $\#\tau_j < k$. Stärkere Glattheitsforderungen, etwa einfache innere Knoten $\#\tau_j = 1$, $j = k+1, \ldots, n$, werden an späterer Stelle diskutiert.

Für die *Randknoten* $\tau_1, \ldots, \tau_k$ und $\tau_{n+1}, \ldots, \tau_{n+k}$ gelte speziell

$$\tau_1 = \cdots = \tau_k = a \qquad \text{und} \qquad b = \tau_{n+1} = \cdots = \tau_{n+k}. \tag{2.1d}$$

Diese Bedingung ist nach [Cox82] günstig für die Kondition der auftretenden Matrizen. Da bei dieser Wahl $s(a) = \alpha_1$ und $s(b) = \alpha_n$ gilt, eignet sie sich insbesondere für die Berücksichtigung von Randbedingungen. Die Bedingung (2.1d) ist für die weiteren Ausführungen nicht wesentlich.

Wir betrachten jetzt die Einschränkung von $\mathcal{S}_{k,\boldsymbol{\tau}}$ auf $[a, b]$

$$\mathcal{S}_{k,\boldsymbol{\tau}_{|[a,b]}} := \left\{ s \, : \, s(x) = \sum_{j=1}^{n} B_{j,k,\boldsymbol{\tau}}(x) \alpha_j, x \in [a, b], \alpha_j \in \mathbb{R} \right\}.$$

Statt $\mathcal{S}_{k,\boldsymbol{\tau}_{|[a,b]}}$ werden wir im folgenden kurz $\mathcal{S}_{k,\boldsymbol{\tau}}$ schreiben. Für einen Spline $s \in \mathcal{S}_{k,\boldsymbol{\tau}}$ gibt es eine Darstellung $s(x) = \sum_{j=1}^{n} B_{j,k,\boldsymbol{\tau}}(x)\alpha_j$ bzw. $s(x) = \boldsymbol{\beta}(x,\boldsymbol{\tau})^T\boldsymbol{\alpha}$ mit $\boldsymbol{\beta}(x,\boldsymbol{\tau}) := (B_{1,k,\boldsymbol{\tau}}(x),\ldots,B_{n,k,\boldsymbol{\tau}}(x))^T \in \mathbb{R}^n$ und Splinekoeffizienten $\boldsymbol{\alpha} := (\alpha_1,\ldots,\alpha_n)^T \in \mathbb{R}^n$. Unter der Voraussetzung $\tau_j < \tau_{j+k}$, $j = 1,\ldots,n$, ist die Dimension von $\mathcal{S}_{k,\boldsymbol{\tau}}$ gleich n und die obige Darstellung ist eindeutig. Nach Definition 2.2 sind die B-Splines rechtsstetig. Am rechten Intervallende definieren wir daher $B_{j,k,\boldsymbol{\tau}}(b) := B_{j,k,\boldsymbol{\tau}}(b-)$.

Bei der Quadratmittelapproximation mit B-Splines spielt die Struktur folgender Matrix, der sog. *Beobachtungsmatrix*, eine Rolle

$$\mathbf{B} = \mathbf{B}(\boldsymbol{\tau}) := (B_{j,k,\boldsymbol{\tau}}(x_i))_{i=1,\ldots,m;\, j=1,\ldots,n} \in \mathbb{R}^{m,n}.$$

Satz 2.2 (Schoenberg, Whitney)**.**
Die Matrix $\mathbf{B} \in \mathbb{R}^{m,n}$ hat Vollrang n genau dann, wenn es eine geordnete Teilfolge der Datenabszissen (x_{i_j}) mit $1 \le i_1 < \cdots < i_n \le m$ gibt, so daß

$$B_{j,k,\boldsymbol{\tau}}\left(x_{i_j}\right) \neq 0, \qquad j = 1,\ldots,n. \tag{2.2}$$

Eine hinreichende (und bei $\tau_j < \tau_{j+k}$ auch notwendige) Bedingung für (2.2) ist die *Schoenberg-Whitney-Bedingung*

$$\tau_j < x_{i_j} < \tau_{j+k}, \qquad j = 1,\ldots,n. \tag{2.3}$$

In der Originalarbeit [SW53] wird der Satz für den Fall der Splineinterpolation bei Benutzung abgebrochener Potenzfunktionen bewiesen. Die zitierte Version stammt aus [dB78, Theorem XIII.1, Lemma XIV.2]. Die Bedingung (2.3) wird auch als „interlacing property“ bezeichnet. Die Matrix $\mathbf{B}$ hat eine Zeilenbandbreite kleiner oder gleich k.

Ableitung eines Splines nach seinem Argument

Zur Darstellung des Glättungsterms $\rho(s) = \frac{1}{2}\int_a^b [s^{(r)}(x)]^2\,dx$ benötigen wir die Ableitung eines Splines bezüglich seines Arguments. Wir geben die Ableitung der Ordnung q nach [SK93] in Matrixnotation an.

Lemma 2.3 (Ableitung eines Splines nach seinem Argument)**.**
Sei $s \in \mathcal{S}_{k,\boldsymbol{\tau}}$ und gelte $\tau_j < \tau_{j+k-q}$, $j = q+1,\ldots,n$. Dann existiert die Ableitung der Ordnung q von s bez. des Arguments und ist ein Spline der Ordnung $k-q$ zu denselben Knoten. Falls $s(x) = \sum_{j=1}^{n} B_{j,k,\boldsymbol{\tau}}(x)\alpha_j$, so besitzt $s^{(q)}$ die Darstellung

$$s^{(q)}(x) = \sum_{j=q+1}^{n} B_{j,k-q,\boldsymbol{\tau}}(x)\alpha_j^{(q)} = \boldsymbol{\beta}_q(x,\boldsymbol{\tau})^T\boldsymbol{\alpha}^{(q)}$$

mit

$$\boldsymbol{\beta}_q(x,\boldsymbol{\tau}) := (B_{q+1,k-q,\boldsymbol{\tau}}(x),\ldots,B_{n,k-q,\boldsymbol{\tau}}(x))^T \in \mathbb{R}^{n-q},$$
$$\boldsymbol{\alpha}^{(q)} := \left(\alpha_{q+1}^{(q)},\ldots,\alpha_n^{(q)}\right)^T \in \mathbb{R}^{n-q}.$$

Die Koeffizienten $\boldsymbol{\alpha}^{(q)}$ sind durch folgende Beziehung mit den Koeffizienten $\boldsymbol{\alpha}$ verbunden:

$$\boldsymbol{\alpha}^{(q)} := \mathbf{D}_q\boldsymbol{\alpha} \quad \textit{mit} \quad \mathbf{D}_0 := \mathbf{I} \in \mathbb{R}^{n,n} \quad \textit{und} \quad \mathbf{D}_q := \mathbf{H}_q\mathbf{L}_q\ldots\mathbf{H}_1\mathbf{L}_1 \in \mathbb{R}^{n-q,n} \quad q \ge 1.$$

Die Matrizen $\mathbf{H}_\nu$ *und* $\mathbf{L}_\nu$ *sind definiert durch*

$$\mathbf{H}_\nu := (k-\nu)\,\mathrm{diag}\left(\frac{1}{\tau_{k+j}-\tau_{\nu+j}}\right)_{j=1,\dots,n-\nu} \in \mathbb{R}^{n-\nu,n-\nu}$$

$$\mathbf{L}_\nu := \begin{bmatrix} -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & \ddots & \ddots & & \\ & & & -1 & 1 & \\ & & & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{n-\nu,n-\nu+1} \quad \textit{für } \nu = 1,\dots,q.$$

Die Bedingung $\tau_j < \tau_{j+k-q}$, $j = q+1,\dots,n$ impliziert $q < k$ und $q+1 \le n$. Die Matrix $\mathbf{D}_q = \mathbf{D}_q(\boldsymbol{\tau})$ hängt nichtlinear von den Knoten ab, sie ist eine obere Dreiecksmatrix der Bandbreite $q+1$. Mittels vollständiger Induktion zeigt man leicht

$$(\mathbf{D}_q)_{ii} = (-1)^q \prod_{\nu=1}^{q} (k-\nu)\frac{1}{\tau_{k+i}-\tau_{i+\nu}} \quad \text{für} \quad i = 1,\dots,n-q \quad \text{und} \quad q \ge 1. \tag{2.4}$$

Daraus erhält man die Zeilenregularität von $\mathbf{D}_q$, d. h. $\operatorname{rank}\mathbf{D}_q = n-q$.

2.2.2 Darstellung des Schoenberg-Funktionals

Nach Einführung dieser grundlegenden Bezeichnungen kommen wir nun zur Darstellung des Schoenberg-Funktionals in Abhängigkeit von den Splinekoeffizienten und Splineknoten: Gegeben seien Meßdaten $\{(x_i, y_i) : i = 1,\dots,m\}$ einer unbekannten glatten Funktion $g \in W_2^q[a,b]$ mit $a \le x_1 < \cdots < x_m \le b$ und $y_i = g(x_i) + \epsilon_i$, $i = 1,\dots,m$. Die stochastischen Fehler ϵ_i seien unabhängig und identisch verteilt. Es gelte

$$\mathrm{E}\,\epsilon_i = 0 \qquad \text{und} \qquad \mathrm{Var}\,\epsilon_i = \sigma^2 > 0, \qquad i = 1,\dots,m. \tag{2.5}$$

Diese Daten wollen wir durch einen Spline s der Ordnung k zur Knotenfolge $\boldsymbol{\tau}$ approximieren, d. h. $s \in \mathcal{S}_{k,\boldsymbol{\tau}}$. Unter der Voraussetzung (2.5) stellt nach dem Gauß-Markov-Theorem der *Approximationsterm* $\varphi(s)$ ein geeignetes Maß für die Approximationsgüte dar.

Als Basis des Raums $\mathcal{S}_{k,\boldsymbol{\tau}}$ verwenden wir B-Splines der Ordnung k zur Knotenfolge $\boldsymbol{\tau}$ mit

$$\tau_1 = \cdots = \tau_k = a < \tau_{k+1} \le \cdots \le \tau_n < b = \tau_{n+1} = \cdots = \tau_{n+k}.$$

Wir definieren die Ansatzfunktion $s(x) := \sum_{j=1}^{n} B_{j,k,\boldsymbol{\tau}}(x)\alpha_j$ bzw. in Vektorform $s(x) := \boldsymbol{\beta}(x,\boldsymbol{\tau})^T\boldsymbol{\alpha}$ mit den Splinekoeffizienten $\boldsymbol{\alpha} := (\alpha_1,\dots,\alpha_n)^T \in \mathbb{R}^n$ und den B-Spline-Werten $\boldsymbol{\beta}(x,\boldsymbol{\tau}) := (B_{1,k,\boldsymbol{\tau}}(x),\dots,B_{n,k,\boldsymbol{\tau}}(x))^T \in \mathbb{R}^n$. Die Knotenfolge $\boldsymbol{\tau}$ erfülle die Bedingung

$$\tau_j < \tau_{j+k}, \qquad j = 1,\dots,n, \tag{2.6}$$

bzw.

$$\tau_j < \tau_{j+k-q}, \qquad j = q+1,\dots,n \tag{2.7}$$

mit einer Zahl $q \in \mathbb{Z}$ mit $0 \le q < k$ sowie $q \le n-1$.

Damit erhalten wir für den Approximationsterm die Darstellung

$$\varphi(s) = \frac{1}{2}\sum_{i=1}^{m}\left[y_i - \sum_{j=1}^{n} B_{j,k,\boldsymbol{\tau}}(x_i)\alpha_j\right]^2.$$

Mit dem Vektor der Beobachtungen $\mathbf{y} := (y_1, \ldots, y_m)^T \in \mathbb{R}^m$ und der Beobachtungsmatrix

$$\mathbf{B}(\boldsymbol{\tau}) := \left(B_{j,k,\boldsymbol{\tau}}(x_i)\right)_{i=1,\ldots,m;\, j=1,\ldots,n} \in \mathbb{R}^{m,n}$$

ergibt sich die äquivalente Matrixformulierung

$$\varphi(s) = \frac{1}{2}\left\|\mathbf{y} - \mathbf{B}(\boldsymbol{\tau})\boldsymbol{\alpha}\right\|^2.$$

Korollar 2.4.
Für die Knotenfolge $\boldsymbol{\tau}$ gelte $\tau_j < \tau_{j+k}$, $j = 1, \ldots, n$. Das Approximationsproblem

$$\frac{1}{2}\left\|\mathbf{y} - \mathbf{B}(\boldsymbol{\tau})\boldsymbol{\alpha}\right\|^2 \to \min_{\boldsymbol{\alpha}\in\mathbb{R}^n}$$

hat für jede feste Knotenfolge $\boldsymbol{\tau}$ eine eindeutige Lösung $\boldsymbol{\alpha}_{opt}(\boldsymbol{\tau})$ genau dann, wenn die Schoenberg-Whitney-Bedingung $\tau_j < x_{i_j} < \tau_{j+k}$, $j = 1, \ldots, n$, erfüllt ist. Die Optimallösung ist gegeben durch

$$\boldsymbol{\alpha}_{opt}(\boldsymbol{\tau}) := \mathbf{B}(\boldsymbol{\tau})^{+}\mathbf{y} = \left(\mathbf{B}(\boldsymbol{\tau})^T\mathbf{B}(\boldsymbol{\tau})\right)^{-1}\mathbf{B}(\boldsymbol{\tau})^T\mathbf{y}.$$

Beweis. Unter Verwendung von Satz 2.2 und Bedingung (2.3) erhalten wir, daß unter den angegebenen Voraussetzungen die Matrix $\mathbf{B} \in \mathbb{R}^{m,n}$ Vollrang n hat. Aus der Theorie linearer Quadratmittelprobleme ist klar, daß für spaltenreguläre Matrizen $\mathbf{B}$ die Minimum-Norm-Lösung $\mathbf{B}^{+}\mathbf{y}$ die einzige Lösung des Problems $\mathbf{B}\boldsymbol{\alpha} \cong \mathbf{y}$ ist. Die Moore-Penrose-Inverse $\mathbf{B}^{+} \in \mathbb{R}^{n,m}$ ist durch die Penrose-Bedingungen

$$\begin{array}{ll} (P1)\quad \mathbf{B}\mathbf{B}^{+}\mathbf{B} = \mathbf{B} & (P2)\quad \mathbf{B}^{+}\mathbf{B}\mathbf{B}^{+} = \mathbf{B}^{+} \\ (P3)\quad (\mathbf{B}\mathbf{B}^{+})^T = \mathbf{B}\mathbf{B}^{+} & (P4)\quad (\mathbf{B}^{+}\mathbf{B})^T = \mathbf{B}^{+}\mathbf{B} \end{array}$$

eindeutig festgelegt. Im Fall $m \geq n$, $\operatorname{rank}\mathbf{B} = n$ gilt speziell $\mathbf{B}^{+} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$. □

Bei der praktischen Anwendung von Approximationsschemata ist oftmals ein Kompromiß zwischen der Güte der Approximation und der Glattheit gefragt. Wir betrachten daher den *exakten Glättungsterm*

$$\rho(s) = \bar{\rho}(s) := \frac{1}{2}\left\|s^{(r)}\right\|_{L_2}^2 = \frac{1}{2}\int_a^b \left[s^{(r)}(x)\right]^2 dx$$

mit festem $r \in \{0, \ldots, q\}$ und das zugehörige Minimierungsproblem

$$\phi(s) := \varphi(s) + \mu\rho(s) = \frac{1}{2}\sum_{i=1}^{m}\left[y_i - s(x_i)\right]^2 + \mu\frac{1}{2}\int_a^b \left[s^{(r)}(x)\right]^2 dx \to \min$$

mit dem *Glättungsparameter* $\mu > 0$ und dem *Schoenberg-Funktional* ϕ. Falls die Knoten die Bedingung (2.7) erfüllen, so existiert $s^{(r)}$ und besitzt die Darstellung

$$s^{(r)}(x) = \sum_{j=r+1}^{n} B_{j,k-r,\boldsymbol{\tau}}(x)\alpha_j^{(r)}$$

mit $\boldsymbol{\alpha}^{(r)} := \mathbf{D}_r(\boldsymbol{\tau})\boldsymbol{\alpha}$ und $\mathbf{D}_r(\boldsymbol{\tau}) \in \mathbb{R}^{n-r,n}$ wie in Lemma 2.3. Setzt man dies in den Glättungsterm $\bar{\rho}$ ein, so erhält man

$$\begin{aligned}\bar{\rho} &= \frac{1}{2}\int_a^b \left[\sum_{i=r+1}^{n} B_{i,k-r,\boldsymbol{\tau}}(x)\alpha_i^{(r)}\right]\left[\sum_{j=r+1}^{n} B_{j,k-r,\boldsymbol{\tau}}(x)\alpha_j^{(r)}\right] dx \\ &= \frac{1}{2}\sum_{i,j=r+1}^{n} \alpha_i^{(r)}\alpha_j^{(r)} \int_a^b B_{i,k-r,\boldsymbol{\tau}}(x)B_{j,k-r,\boldsymbol{\tau}}(x)\,dx \\ &= \frac{1}{2}\boldsymbol{\alpha}^{(r)^T}\bar{\mathbf{M}}_r(\boldsymbol{\tau})\boldsymbol{\alpha}^{(r)}\end{aligned}$$

mit der, von den Knoten $\boldsymbol{\tau}$ abhängigen, Matrix $\bar{\mathbf{M}}_r(\boldsymbol{\tau}) = \left(\bar{m}_{ij}^{(r)}\right) \in \mathbb{R}^{n-r,n-r}$

$$\bar{m}_{ij}^{(r)} := \int_a^b B_{i,k-r,\boldsymbol{\tau}}(x)B_{j,k-r,\boldsymbol{\tau}}(x)\,dx \quad i,j = r+1,\ldots,n.$$

Die Matrix $\bar{\mathbf{M}}_r(\boldsymbol{\tau})$ ist eine symmetrische Matrix der Bandbreite $2(k-r)-1$ und als Gramsche Matrix linear unabhängiger Funktionen positiv definit. Sei $\bar{\mathbf{M}}_r(\boldsymbol{\tau}) = \bar{\mathbf{F}}_r(\boldsymbol{\tau})^T\bar{\mathbf{F}}_r(\boldsymbol{\tau})$ die Cholesky-Faktorisierung. Die reguläre obere Dreiecksmatrix $\bar{\mathbf{F}}_r(\boldsymbol{\tau}) \in \mathbb{R}^{n-r,n-r}$ besitzt die Bandbreite $k-r$. Damit erhält man für den Glättungsterm

$$\bar{\rho} = \frac{1}{2}\boldsymbol{\alpha}^{(r)^T}\bar{\mathbf{M}}_r(\boldsymbol{\tau})\boldsymbol{\alpha}^{(r)} = \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{D}_r(\boldsymbol{\tau})^T\bar{\mathbf{F}}_r(\boldsymbol{\tau})^T\bar{\mathbf{F}}_r(\boldsymbol{\tau})\mathbf{D}_r(\boldsymbol{\tau})\boldsymbol{\alpha} = \frac{1}{2}\left\|\bar{\mathbf{S}}_r(\boldsymbol{\tau})\boldsymbol{\alpha}\right\|^2$$

mit der *Glättungsmatrix*

$$\bar{\mathbf{S}}_r(\boldsymbol{\tau}) := \bar{\mathbf{F}}_r(\boldsymbol{\tau})\mathbf{D}_r(\boldsymbol{\tau}) \in \mathbb{R}^{n-r,n}. \tag{2.8}$$

Die Glättungsmatrix $\bar{\mathbf{S}}_r(\boldsymbol{\tau})$ ist eine obere Dreiecksmatrix mit Bandbreite k. Sie hat wegen der Zeilenregularität von $\mathbf{D}_r(\boldsymbol{\tau})$ und der Regularität von $\bar{\mathbf{F}}_r(\boldsymbol{\tau})$ Vollrang $n-r$.

Die Berechnung der Elemente $\bar{m}_{ij}^{(r)}$ der Gramschen Matrix ist für $k-r > 2$ teuer. Bei sorgfältiger Implementierung kann sie mittels dividierter Differenzen in $\mathcal{O}\left(n(k-r)^2\right)$ Operationen erfolgen. Dieser Prozeß ist jedoch für große Ordnung k und sehr ungleichmäßig verteilte Knoten numerisch instabil. In [dBLS76] wird deshalb eine stabile Rekursionsformel mit $\mathcal{O}\left(n(k-r)^3\right)$ Operationen angegeben. Der „naive“ Zugang, nämlich die Anwendung der Gauß-Quadratur auf den einzelnen Teilintervallen, benötigt ebenfalls $\mathcal{O}\left(n(k-r)^3\right)$ Operationen und ist numerisch stabil. Er wird in [Sch81] empfohlen. Eine Übersicht und Wertung verschiedener Verfahren zur Berechnung der Gramschen Matrix findet man in [VBH92].

Eine Alternative zum exakten Glättungsterm $\bar{\rho}$ besteht in der Verwendung einer billigeren Approximation $\tilde{\rho}$. Wir ersetzen die L_2-Semi-Norm durch ihr diskretes Analogon im Splineraum $\mathcal{S}_{k-r,\boldsymbol{\tau}}$ und definieren für $s(x) = \sum_{j=1}^{n} B_{j,k,\boldsymbol{\tau}}(x)\alpha_j$ die diskreten Normen

$$\|s\|_{l_p} := \begin{cases} \sum_{j=1}^{n} |\alpha_j|^p \left(\frac{\tau_{j+k}-\tau_j}{k}\right)^{1/p}, & \text{für } 1 \le p < \infty, \\ \max|\alpha_j|, & \text{für } p = \infty. \end{cases}$$

Die Normen sind in folgendem Sinne äquivalent:

Es existiert eine Konstante $d_{k,p}$, welche nicht von den Knoten $\boldsymbol{\tau}$ abhängt, so daß gilt

$$d_{k,p}^{-1} \|s\|_{l_p} \leq \|s\|_{L_p} \leq \|s\|_{l_p} \qquad \forall s \in \mathcal{S}_{k,\boldsymbol{\tau}}.$$

Die Existenz der Konstanten $d_{k,p}$ für beliebige L_p-Normen und ihr diskretes Äquivalent l_p geht auf de Boor [dB76] zurück. Für $k = 2, \ldots, 10$ findet man die exakten Werte für $d_{k,\infty}$ etwa in [dB78, S. 155], allgemein gilt $d_{k,\infty} \sim 2^k$.

Die Ersetzung der L_p-Norm durch die diskrete l_p-Norm wird im Zusammenhang mit der Approximation von Daten bereits in [LM88] und [ADLM90] vorgeschlagen. Nehmen wir die Ersetzung innerhalb des exakten Glättungsterms $\bar{\rho}$ vor, so erhalten wir den *approximierten Glättungsterm*

$$\rho(s) = \tilde{\rho}(s) := \frac{1}{2} \left\| s^{(r)} \right\|_{l_2}^2 := \frac{1}{2} \sum_{j=r+1}^{n} \left(\alpha_j^{(r)} \right)^2 \frac{\tau_{j+k-r} - \tau_j}{k-r}.$$

Für diese Approximation erhalten wir

$$\tilde{\rho} = \frac{1}{2} \boldsymbol{\alpha}^{(r)^T} \tilde{\mathbf{F}}_r(\boldsymbol{\tau})^T \tilde{\mathbf{F}}_r(\boldsymbol{\tau}) \boldsymbol{\alpha}^{(r)} = \frac{1}{2} \boldsymbol{\alpha}^T \mathbf{D}_r(\boldsymbol{\tau})^T \tilde{\mathbf{F}}_r(\boldsymbol{\tau})^T \tilde{\mathbf{F}}_r(\boldsymbol{\tau}) \mathbf{D}_r(\boldsymbol{\tau}) \boldsymbol{\alpha}$$

mit der Diagonalmatrix

$$\tilde{\mathbf{F}}_r(\boldsymbol{\tau}) := \operatorname{diag} \left(\sqrt{\frac{\tau_{k+j} - \tau_{r+j}}{k-r}} \right)_{j=1,\ldots,n-r} \in \mathbb{R}^{n-r,n-r} \tag{2.9}$$

und schließlich $\tilde{\rho} = \frac{1}{2} \|\tilde{\mathbf{S}}_r(\boldsymbol{\tau}) \boldsymbol{\alpha}\|^2$ mit der *approximierten Glättungsmatrix*

$$\tilde{\mathbf{S}}_r(\boldsymbol{\tau}) := \tilde{\mathbf{F}}_r(\boldsymbol{\tau}) \mathbf{D}_r(\boldsymbol{\tau}) \in \mathbb{R}^{n-r,n}. \tag{2.10}$$

Die approximierte Glättunsgmatrix $\tilde{\mathbf{S}}_r(\boldsymbol{\tau})$ ist eine obere Dreicksmatrix der Bandbreite $r+1$, im Gegensatz zur Bandbreite k bei der exakten Glättungsmatrix $\bar{\mathbf{S}}_r$.

Wegen der Äquivalenz der Normen besitzen die Matrizen $\bar{\mathbf{S}}_r(\boldsymbol{\tau})$ und $\tilde{\mathbf{S}}_r(\boldsymbol{\tau})$ beide vollen Rang $n - r$ und haben denselben Nullraum

$$\ker \left(\tilde{\mathbf{S}}_r(\boldsymbol{\tau}) \right) = \ker \left(\bar{\mathbf{S}}_r(\boldsymbol{\tau}) \right) = \left\{ \boldsymbol{\alpha} \in \mathbb{R}^n \,:\, \alpha_j^{(r)} = 0 \,,\, j = r+1, \ldots, n \right\},$$

welcher durch alle Polynome der Ordnung r aus $\mathcal{S}_{k,\boldsymbol{\tau}}$ gekennzeichnet ist.

Nun können wir das zu minimierende *Glättungsfunktional* $\phi = \varphi + \mu\rho$ in Abhängigkeit von den Knoten und Koeffizienten definieren.

Definition 2.4 (Glättungsfunktional)**.** Das Funktional $\mathfrak{f} : \mathbb{R}^n \times \mathbb{R}^{n+k} \to \mathbb{R}$ mit

$$\mathfrak{f}(\boldsymbol{\alpha}, \boldsymbol{\tau}) := \frac{1}{2} \|\mathbf{y} - \mathbf{B}(\boldsymbol{\tau}) \boldsymbol{\alpha}\|^2 + \mu \frac{1}{2} \|\mathbf{S}_r(\boldsymbol{\tau}) \boldsymbol{\alpha}\|^2 = \frac{1}{2} \|\mathfrak{F}(\boldsymbol{\alpha}, \boldsymbol{\tau})\|^2$$

und der Residuumsfunktion $\mathfrak{F} : \mathbb{R}^n \times \mathbb{R}^{n+k} \to \mathbb{R}^{m+n-r}$ mit

$$\mathfrak{F}(\boldsymbol{\alpha}, \boldsymbol{\tau}) := \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} - \begin{bmatrix} \mathbf{B}(\boldsymbol{\tau}) \\ \sqrt{\mu} \mathbf{S}_r(\boldsymbol{\tau}) \end{bmatrix} \boldsymbol{\alpha},$$

sowie $\mu > 0$ heißt *Glättungsfunktional*, wobei entweder $\rho = \bar{\rho} = \frac{1}{2} \left\| \bar{\mathbf{S}}_r(\boldsymbol{\tau}) \boldsymbol{\alpha} \right\|^2$, $\mathbf{S}_r(\boldsymbol{\tau}) = \bar{\mathbf{S}}_r(\boldsymbol{\tau})$ (exakter Glättungsterm) oder $\rho = \tilde{\rho} = \frac{1}{2} \left\| \tilde{\mathbf{S}}_r(\boldsymbol{\tau}) \boldsymbol{\alpha} \right\|^2$, $\mathbf{S}_r(\boldsymbol{\tau}) = \tilde{\mathbf{S}}_r(\boldsymbol{\tau})$ (approximierter Glättungsterm).

Die weiteren Ausführungen gelten für beide Glättungsterme gleichermaßen. Bei der Implementierung ergeben sich jedoch auf Grund der unterschiedlichen Bandbreite minimale Unterschiede.

Lemma 2.5 (Vollrangeigenschaft der Systemmatrix, Eindeutigkeit)**.**
Für die Knotenfolge $\boldsymbol{\tau}$ *gelte* $\tau_j < \tau_{j+k-q}$, $j = q+1, \ldots, n$ *und* $r \in \{0, \ldots, q\}$. *Falls die* Regularitätsbedingung

$$m \geq r \qquad \textit{und} \qquad \mu > 0 \tag{2.11}$$

erfüllt ist, so hat die Systemmatrix

$$\mathbf{B}_\mu(\boldsymbol{\tau}) := \begin{bmatrix} \mathbf{B}(\boldsymbol{\tau}) \\ \sqrt{\mu}\mathbf{S}_r(\boldsymbol{\tau}) \end{bmatrix} \in \mathbb{R}^{m+n-r,n}$$

Vollrang n *und das Glättungsproblem*

$$\frac{1}{2}\left\| \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} - \begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}) \end{bmatrix} \boldsymbol{\alpha} \right\|^2 \to \min_{\boldsymbol{\alpha} \in \mathbb{R}^n}$$

für jede feste Knotenfolge $\boldsymbol{\tau}$ *eine eindeutige Lösung* $\boldsymbol{\alpha}_{opt}(\boldsymbol{\tau})$, *welche gegeben ist durch*

$$\boldsymbol{\alpha}_{opt}(\boldsymbol{\tau}) := \left(\begin{bmatrix} \mathbf{B}(\boldsymbol{\tau}) \\ \sqrt{\mu}\mathbf{S}_r(\boldsymbol{\tau}) \end{bmatrix}^T \begin{bmatrix} \mathbf{B}(\boldsymbol{\tau}) \\ \sqrt{\mu}\mathbf{S}_r(\boldsymbol{\tau}) \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{B}(\boldsymbol{\tau}) \\ \sqrt{\mu}\mathbf{S}_r(\boldsymbol{\tau}) \end{bmatrix}^T \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}.$$

Beweis. Die Systemmatrix $\mathbf{B}_\mu$ ist genau dann spaltenregulär, wenn $\ker(\mathbf{B}_\mu) = \ker(\mathbf{B}) \cap \ker(\sqrt{\mu}\mathbf{S}_r) = \mathbf{0}$. Sei $\boldsymbol{\alpha} \in \ker(\mathbf{B})$, d. h. $s(x_i) = 0$ $(i = 1, \ldots, m)$. Sei weiterhin $\boldsymbol{\alpha} \in \ker(\sqrt{\mu}\mathbf{S}_r)$, d. h. $\sqrt{\mu}\mathbf{S}_r\boldsymbol{\alpha} = \mathbf{0}$. Dies gilt genau dann, wenn $\mathbf{S}_r\boldsymbol{\alpha} = \mathbf{0}$ bzw. $s^{(r)} \equiv \mathbf{0}$, da $\mu > 0$ vorausgesetzt war. Ein Element $\boldsymbol{\alpha} \in \ker(\mathbf{B}) \cap \ker(\sqrt{\mu}\mathbf{S}_r)$ ist also dadurch charakterisiert, daß ein Polynom r-ter Ordnung die Daten $\{x_i, 0\}$ $(i = 1, \ldots, m)$ interpoliert. Daraus folgt $\boldsymbol{\alpha} = \mathbf{0}$ im Fall $m \geq r$. □

Die Bedingung (2.11) kann unabhängig von der Lage der Daten gesichert werden und ersetzt gewissermaßen die Schoenberg-Whitney-Bedingung (2.3). Bei praktischen Aufgabenstellungen ist sie in natürlicher Weise erfüllt (im allgemeinen gilt: Anzahl m der Meßwerte $\gg$ Ordnung r im Glättungsterm). Die Schoenberg-Whitney-Bedingung läßt sich dagegen im Laufe des Minimierungsprozesses nur schwer sichern.

2.2.3 Die freien Knoten

Die Splinekoeffizienten $\boldsymbol{\alpha}$ *und* die Splineknoten $\boldsymbol{\tau}$ sollen jetzt so bestimmt werden, daß das Glättungsfunktional $\mathfrak{f}$ mit $\mathfrak{f}(\boldsymbol{\alpha}, \boldsymbol{\tau}) = \frac{1}{2}\|\mathfrak{F}(\boldsymbol{\alpha}, \boldsymbol{\tau})\|^2$ minimiert wird. Um eine größere Flexibilität zu erreichen, beschränken wir uns bei der Optimierung auf eine Teilmenge $\mathbf{t}$ der Knoten $\boldsymbol{\tau}$, die sog. *freien Knoten*. Diese Option ist hilfreich bei der Approximation mehrerer Datensätze, welche sich nur in einem kleinen Bereich des Ausgangsintervalls $[a, b]$ voneinander unterscheiden. In solchen Fällen sollte man nur die Lage der Knoten in diesem kleinen Bereich optimieren. Außerdem werden wir in den weiteren Ausführungen fordern, daß die freien Knoten einfach sind. Durch die Betrachtung einer Teilmenge der Knoten kann man dann a priori feste mehrfache Knoten festlegen, um eine schwächere Glattheitsforderung an diesen Stellen zu erfüllen.

Bezeichne also l die Anzahl der freien Knoten und sei

$$\mathbf{t} = \left(\tau_{p(1)}, \ldots, \tau_{p(l)}\right)^T \in \mathbb{R}^l$$

der Vektor der freien Knoten, wobei

$$\mathbf{p} = (p(1), \ldots, p(l))^T \in \mathbb{Z}^l$$

die Indizes der freien Knoten enthält. Der Indexvektor $\mathbf{p}$ erfülle die Bedingung

$$k < p(1) < \cdots < p(l) < n + 1. \tag{2.12}$$

Bedingung (2.12) besagt, daß nur die inneren Knoten $\tau_{k+1}, \ldots, \tau_n$ als freie Knoten zugelassen sind. Diese Beschränkung erscheint wegen der besonderen Wahl der Knoten (2.1) natürlich, da Informationen über das Meßintervall $[a, b]$ zumeist vorliegen. Es gilt $l \leq n - k$.

Vereinbarung
Da im weiteren nur die freien Knoten verändert werden, wird die Abhängigkeit der Größen von den Knoten $\boldsymbol{\tau}$ durch den Vektor $\mathbf{t}$ der freien Knoten kenntlich gemacht. Falls aus dem Zusammenhang deutlich wird, auf welche Knoten sich die Größen beziehen, verzichten wir gegebenenfalls auch auf die gesonderte Kennzeichnung der freien Knoten, z.B. $\mathbf{B}(\boldsymbol{\tau}) \to \mathbf{B}(\mathbf{t}) \to \mathbf{B}$.

2.2.4 Anordnungsbedingung für die freien Knoten

Ein absolutes Distanzmaß

Für eine Knotenfolge $\boldsymbol{\tau}$ muß laut Definition gelten $\tau_{j+1} - \tau_j \geq 0$, $j = 1, \ldots, n + k - 1$. Da bei der Minimierung des Glättungsfunktionals nur die freien Knoten $\mathbf{t}$ verändert werden, lautet diese Bedingung

$$\begin{aligned} \tau_{p(j)} - \tau_{p(j)-1} &\geq 0 \\ \tau_{p(j)+1} - \tau_{p(j)} &\geq 0 \qquad j = 1, \ldots, l. \end{aligned} \tag{2.13}$$

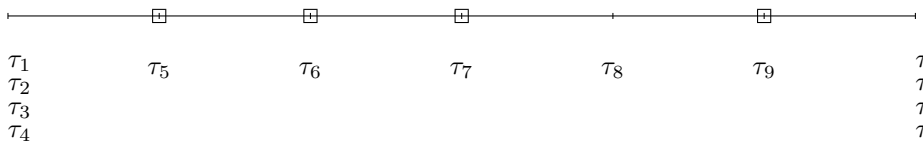
Für die numerische Praxis ist diese Bedingung ungünstig ((2.6) bzw. (2.7) wird nicht gesichert). Wir fordern daher

$$\begin{aligned} \tau_{p(j)} - \tau_{p(j)-1} &\geq \epsilon \\ \tau_{p(j)+1} - \tau_{p(j)} &\geq \epsilon \qquad j = 1, \ldots, l \end{aligned} \tag{2.14}$$

mit einer vorgegebenen Konstanten $\epsilon > 0$. Die Konstante ϵ sollte in Abhängigkeit von der mittleren absoluten Distanz zweier benachbarter Knoten gewählt werden und wird daher als *absolutes Distanzmaß* bezeichnet.

Bedingung (2.14) kann äquivalent in Matrixform $\mathbf{Ct} \geq \mathbf{h}$ mit $\mathbf{C} \in \mathbb{R}^{ncstr,l}$, $\mathbf{t} \in \mathbb{R}^l$, $\mathbf{h} \in \mathbb{R}^{ncstr}$ angegeben werden. Die Anzahl der linear unabhängigen Nebenbedingungen sei *ncstr*. Die Matrix $\mathbf{C}$ und der Vektor $\mathbf{h}$ hängen von der *Lagebeziehung* der Knoten ab, d. h. ob der linke oder rechte Nachbarknoten frei oder fest ist, und können leicht algorithmisch erzeugt werden. Zur Illustration möge ein kleines Beispiel dienen:

Beispiel 2.1. $k = 4$, $n = 9$, $l = 4$, $p(1) = 5$, $p(2) = 6$, $p(3) = 7$, $p(4) = 9$



$$\begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & -1 & 1 & \\ & & -1 & \\ & & & 1 \\ & & & -1 \end{bmatrix} \begin{bmatrix} \tau_5 \\ \tau_6 \\ \tau_7 \\ \tau_9 \end{bmatrix} \geq \begin{bmatrix} \tau_4 + \epsilon \\ \epsilon \\ \epsilon \\ -\tau_8 + \epsilon \\ \tau_8 + \epsilon \\ -\tau_{10} + \epsilon \end{bmatrix}$$

Es gilt $l+1 \leq ncstr \leq 2l$. Die Matrix $\mathbf{C}$ enthält maximal zwei Nichtnullelemente in jeder Zeile. Mit der Wahl eines beliebig kleinen, aber positiven ϵ ist – zumindest theoretisch in rundungsfehlerfreier Arithmetik – die maximale Glattheit des Splines an den Stellen $\tau_{p(j)}$, $j = 1, \ldots, l$, sichergestellt, da die freien Knoten dann einfache Knoten sind.

Ein relatives Distanzmaß

Bisher betrachteten wir die Anordnungsbedingung (2.13) bzw. (2.14), um das Zusammenfallen von Knoten zu verhindern. Der Nachteil der Bedingung (2.14) ist, daß ϵ eine *absolute* minimale Distanz zwischen den Knoten vorschreibt und deshalb i. allg. schwierig zu wählen ist, insbesondere bei sehr ungleichmäßig verteilten Knoten.

Besser ist die folgende *relative* Bedingung, welche auch von de Boor/Rice [dBR68] verwendet wird. Sie verlangen, daß

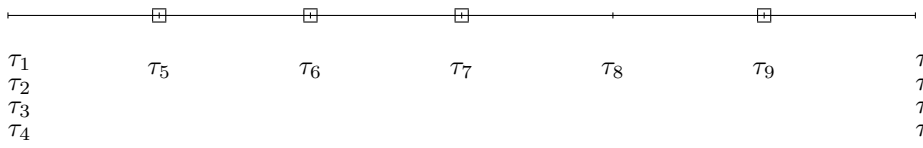
$$\tau_{p(j)-1} + \epsilon \left(\tau_{p(j)+1} - \tau_{p(j)-1}\right) \leq \tau_{p(j)} \leq \tau_{p(j)+1} - \epsilon \left(\tau_{p(j)+1} - \tau_{p(j)-1}\right),$$

d. h. die relative Distanz aufeinanderfolgender Knoten wird nach unten beschränkt gemäß

$$\begin{aligned} \tau_{p(j)} - \tau_{p(j)-1} &\geq \epsilon \left(\tau_{p(j)+1} - \tau_{p(j)-1}\right) \\ \tau_{p(j)+1} - \tau_{p(j)} &\geq \epsilon \left(\tau_{p(j)+1} - \tau_{p(j)-1}\right) \qquad j = 1, \ldots, l. \end{aligned} \tag{2.15}$$

Als *relatives Distanzmaß* wird von de Boor/Rice $\epsilon = 0.0625$ gewählt. Auch hier soll zur Illustration ein Beispiel dienen:

Beispiel 2.2. $k = 4$, $n = 9$, $l = 4$, $p(1) = 5$, $p(2) = 6$, $p(3) = 7$, $p(4) = 9$



$$\begin{bmatrix} 1 & -\epsilon & & \\ -1 & 1-\epsilon & & \\ \epsilon - 1 & 1 & -\epsilon & \\ \epsilon & -1 & 1-\epsilon & \\ & \epsilon - 1 & 1 & \\ & \epsilon & -1 & \\ & & & 1 \\ & & & -1 \end{bmatrix} \begin{bmatrix} \tau_5 \\ \tau_6 \\ \tau_7 \\ \tau_9 \end{bmatrix} \geq \begin{bmatrix} (1-\epsilon)\,\tau_4 \\ -\epsilon\,\tau_4 \\ 0 \\ 0 \\ \epsilon\,\tau_8 \\ (\epsilon - 1)\,\tau_8 \\ (1-\epsilon)\,\tau_8 + \epsilon\,\tau_{10} \\ -\epsilon\,\tau_8 + (\epsilon - 1)\,\tau_{10} \end{bmatrix}$$

Formal erhalten wir eine ähnliche Matrixformulierung der Nebenbedingungen wie im Fall des absoluten Distanzmaßes: $\mathbf{Ct} \geq \mathbf{h}$ mit $\mathbf{C} \in \mathbb{R}^{2l,l}$, $\mathbf{t} \in \mathbb{R}^l$, $\mathbf{h} \in \mathbb{R}^{2l}$. Man beachte, daß jetzt $ncstr = 2l$ und unabhängig von der Lagebeziehung der freien Knoten ist. Die Matrix $\mathbf{C} \in \mathbb{R}^{2l,l}$ ist wiederum schwachbesetzt, kann jetzt jedoch maximal drei Nichtnullelemente je Zeile enthalten und ist abhängig von ϵ.

Sowohl für das absolute als auch für das relative Distanzmaß ϵ erhalten wir also lineare Ungleichheitsnebenbedingungen für die freien Knoten in der Form $\mathbf{Ct} \geq \mathbf{h}$. Die Matrix $\mathbf{C} \in \mathbb{R}^{ncstr,l}$ ist konstant, schwach besetzt und hat nach Konstruktion Vollrang. Der Vektor $\mathbf{h} \in \mathbb{R}^{ncstr}$ ist von $\boldsymbol{\tau}\backslash\{\mathbf{t}\}$ abhängig. Für die Anzahl $ncstr$ linear unabhängiger Nebenbedingungen gilt $l + 1 \leq ncstr \leq 2l$.

2.2.5 Formulierung des vollständigen Optimierungsproblems

Nachdem wir das Schoenberg-Funktional in Abhängigkeit von den Knoten und den Koeffizienten dargestellt sowie die Anordnungsbedingung der freien Knoten formuliert haben, können wir nun das vollständige Optimierungsproblem definieren.

Definition 2.5 (Vollständiges Glättungsproblem)**.** Für die Splineknoten $\boldsymbol{\tau}$ gelte $\tau_j < \tau_{j+k-q}$, $j = q+1, \ldots, n$, und $r \in \{0, \ldots, q\}$. Das Optimierungsproblem

$$\mathfrak{f}(\boldsymbol{\alpha}, \mathbf{t}) := \frac{1}{2} \left\| \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} - \begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}) \end{bmatrix} \boldsymbol{\alpha} \right\|^2 \to \min_{\boldsymbol{\alpha},\mathbf{t}} \quad \text{bei} \quad \mathbf{Ct} - \mathbf{h} \geq \mathbf{0} \tag{2.16}$$

heißt *vollständiges Glättungsproblem* (*full smoothing problem*, FSP).

Die auftretenden Vektoren, Matrizen und Matrixfunktionen haben die folgenden Dimensionen: $\mathbf{y} \in \mathbb{R}^m$; $\boldsymbol{\alpha} \in \mathbb{R}^n$; $\mathbf{t} \in \mathbb{R}^l$; $\mathbf{h} \in \mathbb{R}^{ncstr}$; $\mathbf{C} \in \mathbb{R}^{ncstr,l}$; $\mathbf{B}(.) : \mathbf{t} \in \mathbb{R}^l \to \mathbf{B}(\mathbf{t}) \in \mathbb{R}^{m,n}$; $\mathbf{S}_r(.) : \mathbf{t} \in \mathbb{R}^l \to \mathbf{S}_r(\mathbf{t}) \in \mathbb{R}^{n-r,n}$. Wenn wir den Fall der Splineapproximation ($\mu = 0$) getrennt betrachten, so können wir analog ein vollständiges Approximationsproblem (*full approximation problem*, FAP) definieren.

2.3 Separable Quadratmittelprobleme

Das vollständige Glättungsproblem (2.16) ist ein nichtlineares Quadratmittelproblem in den Variablen $\boldsymbol{\alpha}$ und $\mathbf{t}$, wobei jedoch die Variable $\boldsymbol{\alpha}$ nur linear auftritt. Solche Probleme werden als linear separabel bezeichnet.

In diesem Abschnitt lösen wir uns von dem speziellen Problem der Splineapproximation mit freien Knoten und betrachten allgemeine Optimierungsprobleme dieser speziellen Struktur. Im folgenden seien deshalb $\mathbf{y} \in \mathbb{R}^m$ ein beliebiger Vektor und $\mathbf{B}(.) : \mathbf{t} \in \mathbb{R}^l \to \mathbf{B}(\mathbf{t}) \in \mathbb{R}^{m,n}$ eine beliebige glatte Matrixfunktion, welche zusammen das *Ausgangsproblem* definieren.

Ausgangsproblem

$$\min \left\{ \mathfrak{f}(\boldsymbol{\alpha}, \mathbf{t}) := \frac{1}{2} \|\mathbf{y} - \mathbf{B}(\mathbf{t})\boldsymbol{\alpha}\|^2 = \frac{1}{2} \|\mathfrak{F}(\boldsymbol{\alpha}, \mathbf{t})\|^2 \,:\, \boldsymbol{\alpha} \in \mathbb{R}^n, \mathbf{t} \in \mathbb{R}^l \right\}. \tag{2.17}$$

Es gelte $m \geq n + l$, d. h. das System (2.17) sei überbestimmt.

Separable Quadratmittelprobleme sind ein Spezialfall sog. *reduzibler* nichtlinearer Optimierungsprobleme, bei denen eine natürliche Unterscheidung der Variablen in zwei Gruppen auftritt. Die Variablen, welche den Vektor $\boldsymbol{\alpha}$ darstellen, sind so gewählt, daß das n-dimensionale Subproblem

$$\min\left\{\mathfrak{f}(\boldsymbol{\alpha},\mathbf{t}_c)\ :\ \boldsymbol{\alpha}\in\mathbb{R}^n,\mathbf{t}_c\in\mathbb{R}^l \text{ fest}\right\}$$

effizient und akkurat gelöst werden kann.

Der unrestringierte Quadratmittelfall wurde erstmals von Golub/Pereyra [GP73] im Detail untersucht. Später verallgemeinerten Ruhe/Wedin [RW80] diese Ideen auf allgemeine unrestringierte nichtlineare Optimierungsprobleme, während Parks [Par85] restringierte Probleme betrachtet. Wir werden zunächst die Vorgehensweise im Fall des unrestringierten separablen Quadratmittelproblems (2.17) erläutern, bevor in Abschnitt 3.3 der restringierte Fall aus [Par85] behandelt wird.

2.3.1 Gauß-Newton-ähnliche Verfahren für separable Quadratmittelprobleme

Wir benutzen das Gauß-Newton-Verfahren, um die Ideen zur effektiven Lösung von (2.17) darzulegen. Ausgehend von der aktuellen Iterierten $\mathbf{x}=(\boldsymbol{\alpha},\mathbf{t})$ wird der Gauß-Newton-Schritt $\mathbf{s}\in\mathbb{R}^{n+l}$ zur Minimierung von $\frac{1}{2}\left\|\mathfrak{F}(\boldsymbol{\alpha},\mathbf{t})\right\|^2$ als Lösung des quadratischen Ersatzproblems

$$\min\left\{\mu_{GN}(\mathbf{x}+\mathbf{s})=\frac{1}{2}\left\|\mathfrak{F}+\mathfrak{J}\mathbf{s}\right\|^2\ :\ \mathbf{s}\in\mathbb{R}^{n+l}\right\} \tag{2.18}$$

bestimmt. Unter der Voraussetzung $\operatorname{rank}\mathfrak{J}=n+l$, d. h. die Jacobi-Matrix $\mathfrak{J}=\mathfrak{F}'\in\mathbb{R}^{m,n+l}$ hat vollen Spaltenrang, ist die Lösung von (2.18) eindeutig, erfüllt die Normalgleichungen $\mathfrak{J}^T\mathfrak{J}\mathbf{s}=-\mathfrak{J}^T\mathfrak{F}$ und definiert damit das ungedämpfte Gauß-Newton-Verfahren $\mathbf{x}^+=\mathbf{x}+\mathbf{s}$. Die Jacobi-Matrix $\mathfrak{J}\in\mathbb{R}^{m,n+l}$ der Residuumsfunktion $\mathfrak{F}$ ist $\mathfrak{J}=[\mathfrak{J}_\alpha\ \mathfrak{J}_t]$ mit $\mathfrak{J}_\alpha=-\mathbf{B}$ und $\mathfrak{J}_t=-\boldsymbol{\partial}\mathbf{B}\boldsymbol{\alpha}$, dabei bezeichne $\boldsymbol{\partial}$ den Operator der Fréchet-Ableitung bez. $\mathbf{t}$.

Wir können den Gauß-Newton-Schritt $\mathbf{s}\in\mathbb{R}^{n+l}$ in eine Komponente $\mathbf{s}_\alpha\in\mathbb{R}^n$ bezüglich der Variablen $\boldsymbol{\alpha}$ und eine Komponente $\mathbf{s}_t\in\mathbb{R}^l$ bezüglich $\mathbf{t}$ aufspalten. Schreiben wir die Normalgleichungen aus, so erhalten wir

$$\begin{bmatrix}\mathfrak{J}_\alpha^T\mathfrak{J}_\alpha & \mathfrak{J}_\alpha^T\mathfrak{J}_t\\ \mathfrak{J}_t^T\mathfrak{J}_\alpha & \mathfrak{J}_t^T\mathfrak{J}_t\end{bmatrix}\begin{pmatrix}\mathbf{s}_\alpha\\ \mathbf{s}_t\end{pmatrix}=\begin{pmatrix}-\mathfrak{J}_\alpha^T\mathfrak{F}\\ -\mathfrak{J}_t^T\mathfrak{F}\end{pmatrix}$$

bzw.

$$\mathbf{B}^T\mathbf{B}\mathbf{s}_\alpha-\mathbf{B}^T\mathfrak{J}_t\mathbf{s}_t=\mathbf{B}^T\mathfrak{F} \tag{2.19}$$

$$-\mathfrak{J}_t^T\mathbf{B}\mathbf{s}_\alpha+\mathfrak{J}_t^T\mathfrak{J}_t\mathbf{s}_t=-\mathfrak{J}_t^T\mathfrak{F}. \tag{2.20}$$

Indem wir die Gleichung (2.19) von links mit $\mathfrak{J}_t^T\left(\mathbf{B}^+\right)^T$ multiplizieren und das Resultat zu (2.20) addieren, erhalten wir wegen der Beziehung $\mathbf{B}^{+T}\mathbf{B}^T\mathbf{B}=\left(\mathbf{B}\mathbf{B}^+\right)^T\mathbf{B}\overset{(P3)}{=}\mathbf{B}\mathbf{B}^+\mathbf{B}\overset{(P1)}{=}\mathbf{B}$ die Gleichung

$$\mathfrak{J}_t^T\left(\mathbf{I}-\mathbf{B}\mathbf{B}^+\right)\mathfrak{J}_t\mathbf{s}_t=-\mathfrak{J}_t^T\left(\mathbf{I}-\mathbf{B}\mathbf{B}^+\right)\mathfrak{F}.$$

Definieren wir den orthogonalen Projektor $\mathbf{P}_B^{\perp} := \mathbf{I} - \mathbf{B}\mathbf{B}^{+}$, so ergibt sich

$$\mathfrak{J}_t^T \mathbf{P}_B^{\perp} \mathfrak{J}_t \mathbf{s}_t = -\mathfrak{J}_t^T \mathbf{P}_B^{\perp} \mathfrak{F}.$$

Wegen der Symmetrie und Idempotenz eines Projektors ist dazu äquivalent

$$\left(\mathbf{P}_B^{\perp}\mathfrak{J}_t\right)^T \left(\mathbf{P}_B^{\perp}\mathfrak{J}_t\right) \mathbf{s}_t = -\left(\mathbf{P}_B^{\perp}\mathfrak{J}_t\right)^T \mathbf{P}_B^{\perp}\mathfrak{F}.$$

Dies sind aber genau die Normalgleichungen zum Gleichungssystem $\mathbf{P}_B^{\perp}\mathfrak{J}_t\mathbf{s}_t \cong -\mathbf{P}_B^{\perp}\mathfrak{F}$. Die eindeutige Normallösung ist $\mathbf{s}_t = -\left(\mathbf{P}_B^{\perp}\mathfrak{J}_t\right)^{+} \mathbf{P}_B^{\perp}\mathfrak{F}$. Aus der Gleichung (2.19) erhalten wir dann $\mathbf{s}_\alpha = \mathbf{B}^{+}\left(\mathfrak{F} + \mathfrak{J}_t\mathbf{s}_t\right)$. Jetzt sind wir in der Lage, das Gauß-Newton-Verfahren für das unseparierte Problem (2.17) zu formulieren:

Algorithmus G (Gauß-Newton-Verfahren für das unseparierte Problem).S1: Wähle Startpunkt $\mathbf{t}^{(0)} \in \mathbb{R}^l$, $\boldsymbol{\alpha}^{(0)} \in \mathbb{R}^n$

S2: For $\nu = 0, 1, \ldots$ do

S2.1: Konvergenztest

S2.2: Berechne $\mathbf{s}_t^{(\nu)} := -\left(\mathbf{P}_B^{\perp}\mathfrak{J}_t\right)^{+} \mathbf{P}_B^{\perp}\mathfrak{F}, \quad \mathbf{s}_\alpha^{(\nu)} := \mathbf{B}^{+}\left(\mathfrak{F} + \mathfrak{J}_t\mathbf{s}_t^{(\nu)}\right)$

S2.3: Setze $\mathbf{t}^{(\nu+1)} := \mathbf{t}^{(\nu)} + \mathbf{s}_t^{(\nu)}, \quad \boldsymbol{\alpha}^{(\nu+1)} := \boldsymbol{\alpha}^{(\nu)} + \mathbf{s}_\alpha^{(\nu)}$

Das separable Quadratmittelproblem (2.17) besitzt offensichtlich mehr Struktur als ein beliebiges nichtlineares Quadratmittelproblem. Das wesentliche Merkmal separabler Quadratmittelprobleme ist die Tatsache, daß für feste $\mathbf{t}_c \in \mathbb{R}^l$ die Lösung des *Subproblems*

$$\min\left\{\frac{1}{2}\left\|\mathbf{y} - \mathbf{B}(\mathbf{t}_c)\boldsymbol{\alpha}\right\|^2 \,:\, \boldsymbol{\alpha} \in \mathbb{R}^n, \mathbf{t}_c \in \mathbb{R}^l \quad \text{fest}\right\} \tag{2.21}$$

explizit angegeben und sehr effizient berechnet werden kann. Das Subproblem (2.21) ist ein lineares Quadratmittelproblem der Dimension n, welches mit Standardorthogonalisierungstechniken gelöst werden kann.

Eine naheliegende Veränderung des Gauß-Newton-Verfahrens besteht darin, abwechselnd über einer Variablengruppe zu minimieren, während die Variablen der anderen Gruppe konstant gehalten werden. Diese Methode ist in der Literatur als *Methode der alternierenden Variablen* bekannt. Sie entspricht dem Algorithmus NIPALS, welcher von Wold und Lyttkens [WL69] stammt und unter Statistikern verbreitet ist.

Algorithmus I (Methode der alternierenden Variablen).S1: Wähle Startpunkt $\mathbf{t}^{(0)} \in \mathbb{R}^l$

S2: For $\nu = 0, 1, \ldots$ do

S2.1: Konvergenztest

S2.2: Berechne $\boldsymbol{\alpha}^{(\nu)}$ als Lösung des linearen Quadratmittelproblems

$$\min\left\{\frac{1}{2}\left\|\mathbf{y} - \mathbf{B}\left(\mathbf{t}^{(\nu)}\right)\boldsymbol{\alpha}\right\|^2 : \boldsymbol{\alpha} \in \mathbb{R}^n\right\}$$

S2.3: Berechne $\mathbf{s}_t^{(\nu)} := -\mathfrak{J}_t^{+}\mathfrak{F}$

S2.4: Setze $\mathbf{t}^{(\nu+1)} := \mathbf{t}^{(\nu)} + \mathbf{s}_t^{(\nu)}$

Betrachten wir nun das lineare Quadratmittelproblem in Schritt S2.2 des obigen Algorithmus. Dieses lineare Quadratmittelproblem hat im allgemeinen, rankdefizienten Fall

unendlich viele Lösungen. Für jedes feste $\mathbf{t} \in \mathbb{R}^l$ ist die eindeutige Minimum-Norm-Lösung gegeben durch $\boldsymbol{\alpha} = \boldsymbol{\alpha}_{opt}(\mathbf{t}) := \mathbf{B}(\mathbf{t})^+\mathbf{y}$. Ersetzen wir die Variable $\boldsymbol{\alpha}$ im Ausgangsproblem (2.17) durch ihren Optimalwert $\boldsymbol{\alpha}_{opt}(\mathbf{t})$, so erhalten wir folgendes *reduzierte Problem*:

Reduziertes Problem

$$\min\left\{ f(\mathbf{t}) := \frac{1}{2}\left\| \left[\mathbf{I} - \mathbf{B}(\mathbf{t})\mathbf{B}(\mathbf{t})^+\right]\mathbf{y}\right\|^2 = \frac{1}{2}\|\mathbf{F}(\mathbf{t})\|^2 \;:\; \mathbf{t} \in \mathbb{R}^l \right\} \tag{2.22}$$

Da $\mathbf{P}_B^\perp := \mathbf{I} - \mathbf{B}(\mathbf{t})\mathbf{B}(\mathbf{t})^+$ der orthogonale Projektor auf $\operatorname{im}\mathbf{B}(\mathbf{t})^\perp$ ist, wird das Funktional f auch als *variable projection functional* und die beschriebene Reduktionstechnik als *Verfahren der variablen Projektion* bezeichnet, siehe [GP73].

Das reduzierte Problem (2.22) enthält nur noch die l nichtlinear auftretenden Parameter $\mathbf{t}$, ist dafür aber von einer größeren Komplexität als das Ausgangsproblem. Für die Anwendung des Verfahrens der variablen Projektion bleibt die Zulässigkeit der Reduktion zu zeigen. Außerdem muß für die Anwendung von Gauß-Newton-ähnlichen Verfahren die Jacobi-Matrix $\mathbf{F}'$ berechnet werden.

2.3.2 Der Übergang zum reduzierten Funktional

Das folgende Theorem von Golub/Pereyra [GP73] rechtfertigt den Übergang vom Ausgangsfunktional $\mathfrak{f}(\boldsymbol{\alpha}, \mathbf{t})$ zum reduzierten Funktional $f(\mathbf{t})$. Da wir vom reduzierten Funktional i. allg. nur kritische Punkte und keine globalen Minimalstellen berechnen können, ist besonders die Beziehung zwischen stationären Punkten der beiden Probleme interessant.

Theorem 2.1 (Golub/Pereyra [GP73, Theorem 2.1])**.**
Seien $\mathfrak{f}(\boldsymbol{\alpha}, \mathbf{t})$ und $f(\mathbf{t})$ wie oben definiert. Wir nehmen an, daß $\mathbf{B}(\mathbf{t})$ in der offenen Menge $\Omega \subset \mathbb{R}^l$ konstanten Rang $r \le \min(m, n)$ hat.

(a) Wenn $\mathbf{t}^*$ ein kritischer Punkt (oder eine globale Minimumstelle in Ω) von $f(\mathbf{t})$ ist und

$$\boldsymbol{\alpha}^* = \mathbf{B}(\mathbf{t}^*)^+\mathbf{y}, \tag{2.23}$$

so ist $(\boldsymbol{\alpha}^*, \mathbf{t}^*)$ ein kritischer Punkt (oder eine globale Minimumstelle für $\mathbf{t} \in \Omega$) von $\mathfrak{f}(\boldsymbol{\alpha}, \mathbf{t})$, und es gilt $f(\mathbf{t}^*) = \mathfrak{f}(\boldsymbol{\alpha}^*, \mathbf{t}^*)$.

(b) Wenn $(\boldsymbol{\alpha}^*, \mathbf{t}^*)$ eine globale Minimumstelle von $\mathfrak{f}(\boldsymbol{\alpha}, \mathbf{t})$ für $\mathbf{t} \in \Omega$ ist, so ist $\mathbf{t}^*$ eine globale Minimumstelle von $f(\mathbf{t})$ in Ω, und es gilt $f(\mathbf{t}^*) = \mathfrak{f}(\boldsymbol{\alpha}^*, \mathbf{t}^*)$. Wenn es ein eindeutiges $\boldsymbol{\alpha}^*$ unter den Minimumstellen von $\mathfrak{f}(\boldsymbol{\alpha}, \mathbf{t})$ gibt, so muß $\boldsymbol{\alpha}^*$ die Beziehung (2.23) erfüllen.

Da die Moore-Penrose-Inverse $\mathbf{B}^+$ bei rangerhöhenden Störungen nicht stetig ist (siehe z.B. [KS88]), ist die Forderung nach einem lokal konstanten Rang der Matrix $\mathbf{B}$ natürlich. Wir machen daher im weiteren die Voraussetzung:

Die Matrix $\mathbf{B}(\mathbf{t})$ habe lokal konstanten Rang $r \le \min(m, n)$ für alle $\mathbf{t} \in \Omega \subset \mathbb{R}^l$, wobei Ω eine offene Menge ist, welche die gesuchte Lösung enthält.

2.3.3 Die Bestimmung der Jacobi-Matrix

Zur Berechnung der Jacobi-Matrix $\mathbf{F}'$ benötigen wir zunächst die Fréchet-Ableitung eines orthogonalen Projektors $\mathbf{P}_B$ einer differenzierbaren $m \times n$ Matrix-Funktion $\mathbf{B}(\mathbf{t})$ von lokal konstantem Rang r. Wir werden dabei zunächst nicht die Moore-Penrose-Inverse $\mathbf{B}^+$ und den zugehörigen Projektor $\mathbf{P}_B = \mathbf{B}\mathbf{B}^+$ betrachten, sondern eine verallgemeinerte Inverse $\mathbf{B}^-$, welche nicht alle vier Moore-Penrose-Bedingungen erfüllt. Der zugehörige Projektor $\mathbf{P}_B = \mathbf{B}\mathbf{B}^-$ wird dabei nicht geändert.

Lemma 2.6 (Golub/Pereyra [GP73, Lemma 4.1])**.**
Sei $\mathbf{B}^-(\mathbf{t})$ *eine* $n \times m$ *Matrixfunktion, so daß (P1)* $\mathbf{B}\mathbf{B}^-\mathbf{B} = \mathbf{B}$ *und (P3)* $\left(\mathbf{B}\mathbf{B}^-\right)^T = \mathbf{B}\mathbf{B}^-$*. Dann gilt*

$$\boldsymbol{\partial}\mathbf{P}_B = \mathbf{P}_B^\perp \left(\boldsymbol{\partial}\mathbf{B}\right) \mathbf{B}^- + \left(\mathbf{P}_B^\perp \left(\boldsymbol{\partial}\mathbf{B}\right) \mathbf{B}^-\right)^T.$$

Das Lemma gilt natürlich erst recht, wenn die verallgemeinerte Inverse $\mathbf{B}^-$ durch die Moore-Penrose-Inverse $\mathbf{B}^+$ ersetzt wird. Wegen $\mathbf{P}_B^\perp = \mathbf{I} - \mathbf{P}_B$ gilt $\boldsymbol{\partial}\mathbf{P}_B^\perp = -\boldsymbol{\partial}\mathbf{P}_B$.

Wir sind nun in der Lage, die Fréchet-Ableitung des reduzierten Funktionals

$$f(\mathbf{t}) = \frac{1}{2}\left\|\mathbf{P}_B^\perp \mathbf{y}\right\|^2 = \frac{1}{2}\left(\mathbf{P}_B^\perp \mathbf{y}\right)^T \left(\mathbf{P}_B^\perp \mathbf{y}\right)$$

zu berechnen. Es gilt

$$\begin{aligned}
\boldsymbol{\partial} f &= \left(\mathbf{P}_B^\perp \mathbf{y}\right)^T \boldsymbol{\partial}\left(\mathbf{P}_B^\perp \mathbf{y}\right) \\
&= -\left(\mathbf{P}_B^\perp \mathbf{y}\right)^T \left[\mathbf{P}_B^\perp (\boldsymbol{\partial}\mathbf{B})\mathbf{B}^- + \left(\mathbf{B}^-\right)^T (\boldsymbol{\partial}\mathbf{B})^T \mathbf{P}_B^\perp\right] \mathbf{y} \\
&= -\mathbf{y}^T \mathbf{P}_B^\perp \left[\mathbf{P}_B^\perp (\boldsymbol{\partial}\mathbf{B})\mathbf{B}^- + \left(\mathbf{B}^-\right)^T (\boldsymbol{\partial}\mathbf{B})^T \mathbf{P}_B^\perp\right] \mathbf{y}.
\end{aligned}$$

Man beachte, daß diese Darstellung für die Fréchet-Ableitung des reduzierten Funktionals für beliebige orthogonale Projektoren $\mathbf{P}_B^\perp = \mathbf{I} - \mathbf{B}\mathbf{B}^-$ gilt, sofern $\mathbf{B}^-$ die Moore-Penrose-Bedingungen (P1) und (P3) erfüllt.

Wir nehmen jetzt an, daß $\mathbf{B}^-$ zusätzlich die Bedingung (P2) $\mathbf{B}^-\mathbf{B}\mathbf{B}^- = \mathbf{B}^-$ erfüllt. Dann gilt $\mathbf{P}_B^\perp (\mathbf{B}^-)^T = \mathbf{0}$. Falls $\mathbf{B}^-$ die Moore-Penrose-Bedingungen (P1), (P2) und (P3) erfüllt, erhalten wir also für die Fréchet-Ableitung des reduzierten Funktionals

$$\boldsymbol{\partial} f = -\mathbf{y}^T \mathbf{P}_B^\perp \left(\boldsymbol{\partial}\mathbf{B}\right) \mathbf{B}^- \mathbf{y}. \tag{2.24}$$

Kommen wir nun zur angekündigten Berechnung der Jacobi-Matrix $\mathbf{F}' = \boldsymbol{\partial}\left(\mathbf{P}_B^\perp \mathbf{y}\right) = -\mathbf{P}_B^\perp (\boldsymbol{\partial}\mathbf{B})\mathbf{B}^+\mathbf{y} - \left(\mathbf{P}_B^\perp (\boldsymbol{\partial}\mathbf{B})\mathbf{B}^+\right)^T \mathbf{y}$. Wir erhalten für beliebige $\mathbf{s} \in \mathbb{R}^l$

$$\mathbf{F}'(\mathbf{t})\mathbf{s} = \left\{\mathcal{A}(\mathbf{t})[\mathbf{s}] + \left(\mathcal{A}(\mathbf{t})[\mathbf{s}]\right)^T\right\} \mathbf{y} \tag{2.25}$$

mit

$$\begin{aligned}
\mathcal{A}(\mathbf{t})[\mathbf{s}] &:= -\mathbf{P}_B^\perp \left(\boldsymbol{\partial}\mathbf{B}(\mathbf{t})[\mathbf{s}]\right) \mathbf{B}(\mathbf{t})^+ \\
&= -\left[\mathbf{I} - \mathbf{B}(\mathbf{t})\mathbf{B}(\mathbf{t})^+\right] \left(\boldsymbol{\partial}\mathbf{B}(\mathbf{t})[\mathbf{s}]\right) \mathbf{B}(\mathbf{t})^+.
\end{aligned} \tag{2.26}$$

Nun können wir das Gauß-Newton-Modell für das reduzierte Problem (2.22) bilden:

$$\mu_{GP}(\mathbf{t}+\mathbf{s}) := \frac{1}{2}\left\|\mathbf{F}(\mathbf{t})+\mathbf{F}'(\mathbf{t})\mathbf{s}\right\|^2. \tag{2.27}$$

Im weiteren werden wir dieses Modell als Golub/Pereyra-Modell bezeichnen. In der Originalarbeit von Golub/Pereyra [GP73] wird anstatt mit der Pseudoinversen $\mathbf{B}^+$ mit einer symmetrischen verallgemeinerten Inversen $\mathbf{B}^-$, welche nur die Bedingungen (P1), (P2) und (P3) erfüllt, gearbeitet. Diese verallgemeinerte Inverse wird über eine QR-Transformation auf Trapezform berechnet.

Damit sind wir in der Lage, das Gauß-Newton-Verfahren für das reduzierte Problem (2.22) zu formulieren.

Algorithmus II (Golub/Pereyra-Modell).S1: Wähle Startpunkt $\mathbf{t}^{(0)} \in \mathbb{R}^l$

S2: For $\nu = 0, 1, \ldots$ do

S2.1: Konvergenztest, Falls Konvergenz goto S3;

S2.2: Berechne $\mathbf{s}_t^{(\nu)} := -\left\{-\left[\mathbf{P}_B^{\perp}(\boldsymbol{\partial}\mathbf{B})\mathbf{B}^+ + \left(\mathbf{P}_B^{\perp}(\boldsymbol{\partial}\mathbf{B})\mathbf{B}^+\right)^T\right]\mathbf{y}\right\}^+ \mathbf{P}_B^{\perp}\mathbf{y}$

S2.3: Setze $\mathbf{t}^{(\nu+1)} := \mathbf{t}^{(\nu)} + \mathbf{s}_t^{(\nu)}$

S3: Setze $\boldsymbol{\alpha}^* := \mathbf{B}(\mathbf{t}^*)^+\mathbf{y}$

Eine effiziente Implementierung dieses Verfahrens findet sich in Krogh [Kro74]. Dort wird angenommen, daß $\mathbf{B}$ Vollrang hat. Sowohl in den Implementierungen von Golub/Pereyra als auch in der von Krogh muß der Tensor $\boldsymbol{\partial}\mathbf{B}$ gebildet werden, da er einmal als $\boldsymbol{\partial}\mathbf{B}$ als auch als $(\boldsymbol{\partial}\mathbf{B})^T$ auftritt. Dies erschwert eine effiziente Implementierung erheblich. In den meisten Fällen ist noch dazu $\boldsymbol{\partial}\mathbf{B}$ schwach besetzt. Wie in (2.24) gezeigt wurde, trägt der Term $(\mathcal{A}(\mathbf{t}))^T\mathbf{y}$ jedoch gar nicht zum Gradienten ∇f bei. Eine genaue Analyse zeigt, daß für das Modell (2.27) gilt

$$\left\|\mathbf{F}(\mathbf{t})+\mathbf{F}'(\mathbf{t})\mathbf{s}\right\|^2 = \left\|\mathbf{F}(\mathbf{t})+\mathcal{A}(\mathbf{t})[\mathbf{s}]\mathbf{y}\right\|^2 + \mathbf{s}^T\mathbf{T}_F\mathbf{s},$$

wobei die Matrix $\mathbf{T}_F$ gemäß

$$\mathbf{u}^T\mathbf{T}_F\mathbf{v} := \left\{(\mathcal{A}[\mathbf{u}])^T\mathbf{y}\right\}^T\left\{(\mathcal{A}[\mathbf{v}])^T\mathbf{y}\right\} = \mathbf{F}^T\left(\boldsymbol{\partial}\mathbf{B}[\mathbf{u}]\right)\mathbf{B}^+\mathbf{B}^{+T}\left(\boldsymbol{\partial}\mathbf{B}[\mathbf{v}]\right)^T\mathbf{F} \qquad \forall \mathbf{u},\mathbf{v} \in \mathbb{R}^l$$

definiert ist. Der Beitrag des Terms $(\mathcal{A}[\mathbf{s}])^T\mathbf{y}$ zur Hesse-Matrix $\nabla^2 f$ des reduzierten Funktionals ist $\mathbf{T}_F$, dieser ist aber von der Größenordnung $\mathcal{O}\left(\|\mathbf{F}\|^2\right)$. Er ist somit für kleines $\|\mathbf{F}\|$ kleiner als der beim Gauß-Newton-Verfahren ohnehin weggelassene Term $\mathbf{S} = \mathbf{F}\circ\mathbf{F}''$. Entsprechend der Gauß-Newton-Philosophie wird der Term $\mathbf{T}_F$ daher ebenfalls vernachlässigt. Dies führt auf das vereinfachte Modell

$$\mu_K(\mathbf{t}+\mathbf{s}) := \frac{1}{2}\left\|\mathbf{F}(\mathbf{t})+\mathbf{J}_K(\mathbf{t})\mathbf{s}\right\|^2 \tag{2.28}$$

mit

$$\mathbf{J}_K(\mathbf{t})\mathbf{s} := \mathcal{A}(\mathbf{t})[\mathbf{s}]\mathbf{y} = -\left[\mathbf{I}-\mathbf{B}(\mathbf{t})\mathbf{B}(\mathbf{t})^+\right]\left(\boldsymbol{\partial}\mathbf{B}(\mathbf{t})[\mathbf{s}]\right)\mathbf{B}(\mathbf{t})^+\mathbf{y}. \tag{2.29}$$

Es gilt $\nabla f = \mathbf{F}'^T\mathbf{F} = \mathbf{J}_K^T\mathbf{F}$, $\nabla^2 f = \mathbf{F}'^T\mathbf{F}' + \mathbf{F} \circ \mathbf{F}'' = \mathbf{J}_K^T\mathbf{J}_K + \mathbf{T}_F + \mathbf{F} \circ \mathbf{F}''$. Die soeben beschriebene Vereinfachung wurde erstmals von Kaufman [Kau75] vorgenommen. Sie wird dort unter Benutzung einer differenzierbaren QR-Transformation auf Trapezform hergeleitet. Eine direkte Herleitung, an welche sich die obige Darstellung anlehnt, findet sich in [RW80]. Wir werden das obige Modell im weiteren als Kaufman-Modell und die Approximation $\mathbf{J}_K$ für die Jacobi-Matrix $\mathbf{F}'$ als Kaufman-Approximation bezeichnen. Der Kaufman-Zugang eignet sich nach [Bjö96] besser zur Verallgemeinerung auf restringierte Probleme, siehe auch [KP78].

Algorithmus III (Kaufman-Modell).S1: Wähle Startpunkt $\mathbf{t}^{(0)} \in \mathbb{R}^l$

S2: For $\nu = 0, 1, \ldots$ do

S2.1: Konvergenztest, Falls Konvergenz goto S3;

S2.2: Berechne $\mathbf{s}_t^{(\nu)} := -\left\{-\left[\mathbf{P}_B^{\perp}(\boldsymbol{\partial}\mathbf{B})\mathbf{B}^{+}\right]\mathbf{y}\right\}^{+}\mathbf{P}_B^{\perp}\mathbf{y}$

S2.3: Setze $\mathbf{t}^{(\nu+1)} := \mathbf{t}^{(\nu)} + \mathbf{s}_t^{(\nu)}$

S3: Setze $\boldsymbol{\alpha}^* := \mathbf{B}(\mathbf{t}^*)^{+}\mathbf{y}$

2.3.4 Konvergenzraten und Aufwand

Ruhe/Wedin [RW80] haben die asymptotischen Konvergenzraten der vorgestellten Algorithmen untersucht und geben hinreichende Bedingungen für die lokale Konvergenz der ungedämpften Verfahren an. Sei

$$\epsilon := \frac{\|\mathbf{y} - \mathbf{B}(\mathbf{t})\,\boldsymbol{\alpha}\|_2}{\|\mathbf{B}(\mathbf{t})\,\boldsymbol{\alpha}\|_2} = \frac{\left\|\left[\mathbf{I} - \mathbf{B}(\mathbf{t})\,\mathbf{B}(\mathbf{t})^{+}\right]\mathbf{y}\right\|_2}{\left\|\mathbf{B}(\mathbf{t})\,\mathbf{B}(\mathbf{t})^{+}\,\mathbf{y}\right\|_2}$$

ein Maß für die relative Größe des Residuums im Lösungspunkt. Das Residuum wird als klein betrachtet, wenn ϵ klein ist. Für jede Fixpunktiteration $z_{k+1} := h(z_k)$, $\lim_{k\to\infty} z_k = z^*$ ist die R-Konvergenzrate als $R = \rho\left(h'(z^*)\right)$ definiert (ρ – Spektralradius).

Satz 2.7 (Asymptotische Konvergenzraten, [RW80, Corollary 3.3])**.**
Die R-Konvergenzraten der Algorithmen hängen in folgender Weise von ϵ ab:

$$\begin{array}{rll} R_G &= R_{II} + \mathcal{O}(\epsilon^2) & \textit{Gauß-Newton unseparliertes Problem} \\ R_I &= \mathcal{O}(1) & \textit{Alternierende Variablen} \\ R_{II} &= \mathcal{O}(\epsilon) & \textit{Golub/Pereyra-Modell} \\ R_{III} &= R_{II} + \mathcal{O}(\epsilon^2) & \textit{Kaufman-Modell} \end{array}$$

Jeder der vier Algorithmen ist also i. allg. R-linear konvergent. Wenn der Gauß-Newton-Algorithmus für das unseparierte Problem superlinear konvergiert, so konvergieren auch die Algorithmen II und III superlinear. Die Methode der alternierenden Variablen dagegen konvergiert stets nur R-linear.

In Kaufman [Kau75] und Ruhe/Wedin [RW80] werden die Algorithmen für freie Optimierungsprobleme verglichen und festgestellt, daß der Kaufman-Algorithmus mit weniger arithmetischen Operationen nicht mehr Iterationen als der ursprüngliche Algorithmus von Golub/Pereyra benötigt. Bei praktischen Problemen konvergieren beide i. allg. schneller als der Gauß-Newton-Algorithmus für das unseparierte Problem. Dabei ist jedoch zu beachten, daß stets nur bestimmte Implementierungen der Algorithmen verglichen werden, und daß das Abschneiden der Algorithmen stark vom Aufwand zur Berechnung der Residuumsfunktion abhängig ist.

2.3.5 Separable Quadratmittelprobleme mit Nebenbedingungen

Bisher wurde die Zulässigkeit des Übergangs vom Ausgangs- zum reduzierten Problem lediglich für den unrestringierten Fall (2.17) nachgewiesen. Kaufman zeigte, daß diese Reduktion ebenfalls für separable Quadratmittelprobleme mit linearen Ungleichheitsnebenbedingungen an die nichtlinearen Variablen $\mathbf{t}$ zulässig ist.

Theorem 2.2 (Kaufman, zitiert nach [Par85, S. 41f], siehe auch [GP76])**.**
Ein Problem der Form

$$\min\left\{\mathfrak{f}(\boldsymbol{\alpha},\mathbf{t}) := \frac{1}{2}\|\mathbf{y}-\mathbf{B}(\mathbf{t})\boldsymbol{\alpha}\|^2 \,:\, \mathbf{C}\mathbf{t}\geq\mathbf{h}\,,\, \boldsymbol{\alpha}\in\mathbb{R}^n, \mathbf{t}\in\mathbb{R}^l\right\}$$

ist äquivalent zu

$$\min\left\{f(\mathbf{t}) := \frac{1}{2}\left\|\left[\mathbf{I}-\mathbf{B}(\mathbf{t})\mathbf{B}(\mathbf{t})^+\right]\mathbf{y}\right\|^2 = \frac{1}{2}\left\|\mathbf{P}_B^{\perp}(\mathbf{t})\mathbf{y}\right\|^2 \,:\, \mathbf{C}\mathbf{t}\geq\mathbf{h}\,,\, \mathbf{t}\in\mathbb{R}^l\right\}$$

gefolgt von der Rücksubstitution $\boldsymbol{\alpha} = \mathbf{B}(\mathbf{t})^+\mathbf{y}$.

Die Aussage des Satzes ist anschaulich klar, da die variable Projektion bezüglich der linearen Variablen $\boldsymbol{\alpha}$ erfolgt und nicht bezüglich der restringierten nichtlinearen Variablen $\mathbf{t}$. Eine exakte Formulierung der Äquivalenz im Sinne von Theorem 2.1 folgt außerdem aus der allgemeineren Aussage [Par85, Theorem 4.7], siehe Theorem 3.1.

2.4 Splineglättung mit freien Knoten

In diesem Abschnitt wollen wir die beschriebene Reduktionstechnik auf die Splineapproximation durch Splines mit freien Knoten anwenden. Offensichtlich bleiben die Aussagen des letzten Abschnitts richtig, wenn man die *beliebigen* Vektoren, Matrizen und Matrixfunktionen durch die *speziellen* Bezeichnungen des vollständigen Approximationsproblems FAP ersetzt. Wir erhalten das *reduzierte Approximationsproblem* (reduced approximation problem, RAP):

Reduziertes Approximationsproblem

$$\min\left\{f(\mathbf{t}) := \tfrac{1}{2}\left\|\left[\mathbf{I}-\mathbf{B}(\mathbf{t})\mathbf{B}(\mathbf{t})^+\right]\mathbf{y}\right\|^2 = \tfrac{1}{2}\|\mathbf{F}(\mathbf{t})\|^2 \,:\, \mathbf{C}\mathbf{t}\geq\mathbf{h}\,,\, \mathbf{t}\in\mathbb{R}^l\right\} \tag{2.30}$$

Zu den Aussagen für das vollständige Glättungsproblem FSP (2.16) gelangt man, indem man die Größen $\{\mathbf{B}(\mathbf{t}),\mathbf{y}\}$ durch das Paar

$$\left\{\begin{bmatrix}\mathbf{B}(\mathbf{t})\\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t})\end{bmatrix}, \begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix}\right\}$$

ersetzt. Man erhält damit das *reduzierte Glättungsproblem* (reduced smoothing problem, RSP):

Reduziertes Glättungsproblem

$$\min \left\{ f(\mathbf{t}) := \frac{1}{2} \left\| \left\{ \begin{bmatrix} \mathbf{I}_m \\ \mathbf{I}_{n-r} \end{bmatrix} - \begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}) \end{bmatrix} \begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}) \end{bmatrix}^+ \right\} \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} \right\|^2 \right. \\ \left. = \tfrac{1}{2} \left\| \mathbf{F}(\mathbf{t}) \right\|^2 \,:\, \mathbf{Ct} \geq \mathbf{h}\,,\, \mathbf{t} \in \mathbb{R}^l \right\} \tag{2.31}$$

Leider erfüllt die Matrix $\mathbf{B}(\mathbf{t})$ des reduzierten Approximationsproblems (2.30) für beliebige Daten nicht die Bedingung eines konstanten Ranges aus Theorem 2.1 für alle zulässigen Knoten $\mathbf{t} \in \mathbb{R}^l$, während diese Bedingung für die Systemmatrix aus dem reduzierten Glättungsproblem (2.31) erfüllt ist, sofern $m \geq r$ und $\mu > 0$. Bei den existierenden Algorithmen zur Lösung von (2.30) wurde stets angenommen, daß die Schoenberg-Whitney-Bedingung (2.3), welche den Vollrang von $\mathbf{B}(\mathbf{t})$ sichert, für alle auftretenden Knoten erfüllt ist. Falls dies nicht der Fall ist, so kann man zwar die eindeutige Minimum-Norm-Lösung $\mathbf{B}(\mathbf{t})^+\mathbf{y}$ verwenden, verliert jedoch im rangdefizienten Fall die Stetigkeit und Differenzierbarkeit des reduzierten Funktionals f und die Äquivalenz von vollständigem und reduziertem Problem. Der Übergang vom Approximationsproblem zum Glättungsproblem — in anderen Gebieten der Splineapproximation weit verbreitet — kann also als Regularisierung verstanden werden und scheint im Zusammenhang mit Splines mit freien Knoten neu zu sein. Im weiteren werden wir daher das Glättungsproblem (2.16), (2.31) betrachten und nur gelegentlich auf die Unterschiede zum Approximationsproblem eingehen.

2.4.1 Existenz von Lösungen des reduzierten Glättungsproblems

Es ist allgemein bekannt, daß die B-Splines stetig von der Lage der Knoten abhängen. Genauer gilt:

Theorem 2.3 (Schumaker [Sch81, Theorem 4.26])**.**
Sei $\tau_j \leq \cdots \leq \tau_{j+k}$ eine Knotenfolge, und sei $\tau_j^{(\nu)} \leq \cdots \leq \tau_{j+k}^{(\nu)}$ eine Folge von Punkten mit $\tau_i^{(\nu)} \to \tau_i$, $i = j, \ldots, j+k$ für $\nu \to \infty$. Seien Q_j^k und $Q_{j,\nu}^k$ die zugehörigen nichtnormalisierten B-Splines der Ordnung k. Dann gilt für alle $q = 0, \ldots, k-1$,

$$D_+^q Q_{j,\nu}^k(x) \to D_+^q Q_j^k(x) \quad \textit{für alle } x \in \mathbb{R} \backslash J_j^k,$$

wobei $J_j^k := \{\tau_i : \tau_i$ ist ein Knoten von Q_j^k der Vielfachheit $k-q$ oder höher$\}$. Die Konvergenz ist gleichmäßig auf jeder abgeschlossenen Menge, welche J_j^k nicht enthält.

Der Operator D_+^q ist dabei der rechtsseitige Ableitungsoperator der Ordnung q, die nichtnormalisierten B-Splines Q_j^k sind durch die Beziehung $B_{j,k,\boldsymbol{\tau}} = (\tau_{j+k} - \tau_j) Q_j^k$ definiert.

Unter der Voraussetzung $\tau_j < \tau_{j+k}$, $j = 1, \ldots, n$, an die Knoten haben wir daher die Stetigkeit der B-Splines und damit der Matrixfunktion $\mathbf{B}(.)$ als Funktion der Knoten für alle $x \in [a, b]$.

Theorem 2.4 (Existenz einer Lösung des reduzierten Glättungsproblems)**.**
Die Menge der zulässigen Knoten $\left\{\mathbf{t} \in \mathbb{R}^l \,:\, \mathbf{Ct} - \mathbf{h} \geq \mathbf{0}\right\}$ sei nichtleer. Für ein festes $r \in \{0, \ldots, q\}$, $0 \leq q < k$ gelte:

($V1$) *Die Knoten erfüllen die Bedingung* $\tau_j < \tau_{j+k-q}$ $(j = q+1, \ldots, n)$.

($V2$) *Die Regularitätsbedingung* $m \geq r$ *und* $\mu > 0$ *ist erfüllt.*

Dann besitzt das reduzierte Glättungsproblem RSP (2.31) *eine Lösung* $\mathbf{t}^*$.

Beweis. Voraussetzung (V1) sichert die Existenz und Stetigkeit der Matrixfunktionen $\mathbf{B}(.)$ und $\mathbf{S}_r(.)$ als Funktionen der (freien) Knoten (wegen der Anordnungsnebenbedingung $\mathbf{Ct} - \mathbf{h} \geq \mathbf{0}$ ändert sich die Knotenvielfachheit nicht). Nach Lemma 2.5 hat die Systemmatrix $\mathbf{B}_\mu$ unter der Voraussetzung (V2) Vollrang n. Somit ist auch die Moore-Penrose-Inverse $\mathbf{B}_\mu^+$ eine stetige Funktion der Knoten, siehe etwa [KS88]. Das stetige Funktional f nimmt daher über der abgeschlossenen (wegen $\mathbf{Ct} - \mathbf{h} \geq \mathbf{0}$) und beschränkten (wegen $a \leq \tau_{p(1)}$ und $\tau_{p(l)} \leq b$) Menge der zulässigen Knoten $\left\{\mathbf{t} \in \mathbb{R}^l : \mathbf{Ct} - \mathbf{h} \geq \mathbf{0}\right\}$ nach dem Satz von Weierstraß ihr Minimum an einer Stelle $\mathbf{t}^*$ an. □

2.4.2 Äquivalenz von vollständigem und reduziertem Problem

Für den Nachweis der Äquivalenz von vollständigem und reduziertem Problem benötigen wir die differenzierbare Abhängigkeit der Matrixfunktionen $\mathbf{B}(.)$ und $\mathbf{S}_r(.)$ bezüglich der freien Knoten. Aus der Definition von $\mathbf{S}_r(.)$ ist klar, daß diese Matrixfunktion stetig differenzierbar ist, falls $\tau_j < \tau_{j+k-q}$ $(j = q+1, \ldots, n)$, $r \in \{0, \ldots, q\}$. Es bleibt also die Ableitung eines B-Splines nach den Knoten zu untersuchen.

Eine zu Definition 2.2 äquivalente Darstellung der B-Splines kann mittels dividierter Differenzen angegeben werden (siehe [Sch81] zur Schreibweise und zur Definition dividierter Differenzen).

$$B_{j,k,\boldsymbol{\tau}}(x) := \begin{cases} (\tau_{j+k} - \tau_j)\,(-1)^k\,[\tau_j, \ldots, \tau_{j+k}]\,(x-y)_+^{k-1}, & \text{falls} \quad \tau_j < \tau_{j+k}, \\ 0, & \text{sonst.} \end{cases}$$

Sei $\boldsymbol{\Theta}$ die feinste streng monoton wachsende Teilfolge von $\{\tau_j, \ldots, \tau_{j+k}\}$, d. h.

$$\tau_j \leq \tau_{j+1} \leq \cdots \leq \tau_{j+k} \quad \widehat{=} \quad \underbrace{\theta_1}_{\nu_1} < \underbrace{\theta_2}_{\nu_2} < \cdots < \underbrace{\theta_d}_{\nu_d}.$$

Jedem τ_j ist ein θ_i mit der entsprechenden Vielfachheit ν_i zugeordnet. Eine alternative Schreibweise des B-Splines lautet daher

$$B_{j,k,\boldsymbol{\tau}}(x) := \begin{cases} (\tau_{j+k} - \tau_j)\,(-1)^k \begin{bmatrix} \nu_1, \nu_2, ..., \nu_d \\ \theta_1, \theta_2, ..., \theta_d \end{bmatrix} (x-y)_+^{k-1}, & \text{falls} \quad \tau_j < \tau_{j+k}, \\ 0, & \text{sonst.} \end{cases}$$

Damit können wir nach [Sch81, Theorem 4.27] die Ableitung eines B-Splines nach den Knoten angeben:

Theorem 2.5 (Ableitung eines B-Splines nach den Knoten)**.**
Sei $\tau_j \leq \tau_{j+1} \leq \cdots \leq \tau_{j+k} \quad \widehat{=} \quad \underbrace{\theta_1}_{\nu_1} < \underbrace{\theta_2}_{\nu_2} < \cdots < \underbrace{\theta_d}_{\nu_d}$. *Sei* $1 \leq i \leq d$ *fest und gelte*

$\nu_i \leq k-2$. *Dann gilt*

$$\frac{\partial B_{j,k,\boldsymbol{\tau}}(x)}{\partial \theta_i} \tag{2.32}$$

$$= \begin{cases} (-1)^{k-1}\left[\begin{array}{c} \nu_1+1, \nu_2, ..., \nu_d - 1 \\ \theta_1, \theta_2, ..., \theta_d \end{array}\right](x-y)_+^{k-1}, & \text{für } i=1 \text{ und } \nu_1 = 1 \\ (-1)^{k}\left[\begin{array}{c} \nu_1 - 1, ..., \nu_{d-1}, \nu_d+1 \\ \theta_1, ..., \theta_{d-1}, \theta_d \end{array}\right](x-y)_+^{k-1}, & \text{für } i=d \text{ und } \nu_d = 1 \\ \nu_i\,(\tau_{j+k}-\tau_j)\,(-1)^{k}\left[\begin{array}{c} \nu_1, ..., \nu_i+1, ..., \nu_d \\ \theta_1, ..., \theta_i, ..., \theta_d \end{array}\right](x-y)_+^{k-1}, & \text{sonst.} \end{cases}$$

Die Ableitung $\partial/\partial\theta_i$ ist als rechtsseitige Ableitung zu verstehen, wenn $\theta_i = \theta_{i-1}$ und als linksseitige, wenn $\theta_i = \theta_{i+1}$. Die gleichen Formel gelten, falls $\nu_i = k-1$ oder $\nu_i = k$ für alle x außer $x = \theta_i$.

Theorem 2.5 zeigt die Existenz der Ableitungen eines B-Splines nach den Randknoten $\tau_1 = \cdots = \tau_k$ bzw. $\tau_{n+1} = \cdots = \tau_{n+k}$ an allen Stellen außer τ_1 bzw. τ_{n+1}. In den weiteren Ausführungen wird jedoch lediglich die Ableitung eines B-Splines nach den inneren Knoten benötigt.

Korollar 2.8 (Ableitung eines B-Splines nach den inneren Knoten)**.**
Sei $\tau_1 = \cdots = \tau_k = a < \tau_{k+1} \leq \cdots \leq \tau_n < b = \tau_{n+1} = \cdots = \tau_{n+k}$ eine Knotenfolge mit $\tau_j < \tau_{j+k-2}$ $(j = 3, \ldots, n)$, d. h. $\#\tau_j \leq k-2$. Dann existieren die Ableitungen der B-Splines nach den inneren Knoten $\partial B_{j,k,\boldsymbol{\tau}}/\partial\tau_i$ $(j = 1, \ldots, n; i = k+1, \ldots, n)$ für alle $x \in [a, b]$.

Die Bedingung $\#\tau_j \leq k-2$ impliziert $k \geq 3$, d. h. mindestens quadratische B-Splines. Der letzte Teil der Formel (2.32) liefert den bekannten Zusammenhang zwischen der Ableitung eines B-Splines der Ordnung k nach den inneren Knoten und der Ableitung eines B-Splines der Ordnung $k+1$ nach dem Argument

$$\frac{\partial}{\partial\tau_i} B_{j,k}\left(x \,\middle|\, \begin{array}{c} \nu_1, ..., \nu_i, ..., \nu_d \\ \theta_1, ..., \theta_i, ..., \theta_d \end{array}\right) = -\frac{\nu_i}{k}\frac{\partial}{\partial x} B_{j,k+1}\left(x \,\middle|\, \begin{array}{c} \nu_1, ..., \nu_i+1, ..., \nu_d \\ \theta_1, ..., \theta_i, ..., \theta_d \end{array}\right) \quad \text{für } 1 < i < d.$$

Ableitung eines Splines nach den Knoten

In den weiteren Ausführungen benötigen wir nicht nur die Ableitung eines einzelnen B-Splines, sondern die Ableitung eines Splines. Gesucht ist

$$\frac{\partial s}{\partial \tau_{j_0}} = \sum_{j=1}^{n} \frac{\partial B_{j,k,\boldsymbol{\tau}}}{\partial \tau_{j_0}} \alpha_j \qquad \text{mit } k < j_0 < n+1 \text{ und } \#\tau_{j_0} \leq k-2.$$

Nach Theorem 2.5 gilt

$$\frac{\partial}{\partial \tau_{j_0}} B_{j,k,\boldsymbol{\tau}} = -\frac{\#\tau_{j_0}}{k}\frac{\partial}{\partial x} B_{j,k+1,\boldsymbol{\tau}'}$$

für $j < j_0 < j+k$ mit der Knotenfolge $\boldsymbol{\tau}' = (\tau_1, \ldots, \tau_{j_0}, \tau_{j_0}, \ldots \tau_{n+k})^T \in \mathbb{R}^{n+k+1}$, d. h. $\tau'_j = \tau_j$, $j = 1, \ldots, j_0$ und $\tau'_j = \tau_{j-1}$, $j = j_0+1, \ldots, n+k+1$. Bekanntlich gilt für die Ableitung eines B-Splines nach seinem Argument

$$\frac{\partial}{\partial x} B_{j,k,\boldsymbol{\tau}} = (k-1)\left(\frac{B_{j,k-1,\boldsymbol{\tau}}}{\tau_{j+k-1}-\tau_j} - \frac{B_{j+1,k-1,\boldsymbol{\tau}}}{\tau_{j+k}-\tau_{j+1}}\right),$$

also

$$\frac{\partial}{\partial x} B_{j,k+1,\boldsymbol{\tau}'} = k \left(\frac{B_{j,k,\boldsymbol{\tau}'}}{\tau'_{j+k} - \tau'_j} - \frac{B_{j+1,k,\boldsymbol{\tau}'}}{\tau'_{j+k+1} - \tau'_{j+1}} \right).$$

Im weiteren setzen wir voraus, daß die Vielfachheit von τ_{j_0} gleich 1 ist ($\#\tau_{j_0} = 1$). Da wir ausschließlich die Ableitung eines Splines nach den freien Knoten benötigen, ist diese Voraussetzung erfüllt, falls die freien Knoten einfache Knoten sind (bzw. einfache Knoten bleiben). Wir erhalten damit

$$\begin{aligned} \frac{\partial}{\partial \tau_{j_0}} B_{j,k,\boldsymbol{\tau}} &= -\frac{1}{k} k \left(\frac{B_{j,k,\boldsymbol{\tau}'}}{\tau'_{j+k} - \tau'_j} - \frac{B_{j+1,k,\boldsymbol{\tau}'}}{\tau'_{j+k+1} - \tau'_{j+1}} \right) \\ &= \frac{B_{j+1,k,\boldsymbol{\tau}'}}{\tau'_{j+k+1} - \tau'_{j+1}} - \frac{B_{j,k,\boldsymbol{\tau}'}}{\tau'_{j+k} - \tau'_j} \qquad \text{für} \quad j < j_0 < j + k. \end{aligned} \tag{2.33}$$

Zur Berechnung von $\partial s / \partial \tau_{j_0}$ benötigen wir noch $\partial B_{j,k,\boldsymbol{\tau}} / \partial \tau_{j_0}$ für $j_0 \le j$ und $j + k \le j_0$. Da in die Definition von $B_{j,k,\boldsymbol{\tau}}$ nur die Knoten $\tau_j, \ldots, \tau_{j+k}$ eingehen, gilt

$$\frac{\partial B_{j,k,\boldsymbol{\tau}}}{\partial \tau_{j_0}} \equiv 0 \qquad \text{für } j_0 < j \text{ und } j + k < j_0. \tag{2.34}$$

Es bleiben also $\partial B_{j,k,\boldsymbol{\tau}} / \partial \tau_j$ und $\partial B_{j,k,\boldsymbol{\tau}} / \partial \tau_{j+k}$ zu bestimmen. Nach der ersten Formel von (2.32) gilt

$$\begin{aligned} \frac{\partial B_{j,k,\boldsymbol{\tau}}}{\partial \tau_j} &= (-1)^{k-1} \left[\begin{array}{c} \nu_1 + 1, \nu_2, ..., \nu_d - 1 \\ \theta_1, \theta_2, ..., \theta_d \end{array} \right] (x - y)_+^{k-1} \\ &= \frac{-1}{\tau_{j+k-1} - \tau_j} (-1)^k (\tau_{j+k-1} - \tau_j) \left[\begin{array}{c} \nu_1 + 1, \nu_2, ..., \nu_d - 1 \\ \theta_1, \theta_2, ..., \theta_d \end{array} \right] (x - y)_+^{k-1} \\ &= -\frac{1}{\tau'_{j+k} - \tau'_j} B_{j,k,\boldsymbol{\tau}'}. \end{aligned} \tag{2.35}$$

Analog erhalten wir nach der zweiten Formel von (2.32)

$$\begin{aligned} \frac{\partial B_{j,k,\boldsymbol{\tau}}}{\partial \tau_{j+k}} &= (-1)^k \left[\begin{array}{c} \nu_1 - 1, ..., \nu_{d-1}, \nu_d + 1 \\ \theta_1, ..., \theta_{d-1}, \theta_d \end{array} \right] (x - y)_+^{k-1} \\ &= \frac{1}{\tau_{j+k} - \tau_{j+1}} (-1)^k (\tau_{j+k} - \tau_{j+1}) \left[\begin{array}{c} \nu_1 - 1, ..., \nu_{d-1}, \nu_d + 1 \\ \theta_1, ..., \theta_{d-1}, \theta_d \end{array} \right] (x - y)_+^{k-1} \\ &= \frac{1}{\tau'_{j+k+1} - \tau'_{j+1}} B_{j+1,k,\boldsymbol{\tau}'}. \end{aligned} \tag{2.36}$$

Berücksichtigen wir nun (2.34), so ergibt sich

$$\frac{\partial s}{\partial \tau_{j_0}} = \sum_{j=j_0-k}^{j_0} \frac{\partial B_{j,k,\boldsymbol{\tau}}}{\partial \tau_{j_0}} \times \alpha_j$$

und mit (2.33), (2.35) und (2.36) schließlich

$$\begin{aligned}\frac{\partial s}{\partial \tau_{j_0}} &= \frac{1}{\tau'_{j_0+1}-\tau'_{j_0-k+1}} B_{j_0-k+1,k,\boldsymbol{\tau}'} \times \alpha_{j_0-k} \\ &\quad + \sum_{j=j_0-k+1}^{j_0-1} \left(\frac{B_{j+1,k,\boldsymbol{\tau}'}}{\tau'_{j+k+1}-\tau'_{j+1}} - \frac{B_{j,k,\boldsymbol{\tau}'}}{\tau'_{j+k}-\tau'_{j}} \right) \times \alpha_j \\ &\quad - \frac{1}{\tau'_{j_0+k}-\tau'_{j_0}} B_{j_0,k,\boldsymbol{\tau}'} \times \alpha_{j_0} \\ &= \sum_{j=j_0-k+1}^{j_0} \frac{\alpha_{j-1}-\alpha_j}{\tau'_{j+k}-\tau'_j} \times B_{j,k,\boldsymbol{\tau}'}.\end{aligned}$$

Lemma 2.9 (Ableitung eines Splines nach den Knoten)**.**
Sei $s = \sum_{j=1}^{n} B_{j,k,\boldsymbol{\tau}} \alpha_j$ ein Spline der Ordnung k mit der Knotenfolge $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{n+k})^T$. Sei τ_{j_0} ein Splineknoten mit $\#\tau_{j_0} = 1 \leq k-2$ und $k < j_0 < n+1$. Dann existiert die Ableitung des Splines s bez. des Knoten τ_{j_0} für alle $x \in [a,b]$ und es gilt

$$\frac{\partial s}{\partial \tau_{j_0}} = \sum_{j=j_0-k+1}^{j_0} \frac{\alpha_{j-1}-\alpha_j}{\tau'_{j+k}-\tau'_j} \times B_{j,k,\boldsymbol{\tau}'} \tag{2.37}$$

mit der Knotenfolge $\boldsymbol{\tau}' = (\tau_1, \ldots, \tau_{j_0}, \tau_{j_0}, \ldots \tau_{n+k})^T \in \mathbb{R}^{n+k+1}$, d. h.

$$\begin{aligned}\tau'_j &= \tau_j \qquad && j = 1, \ldots, j_0 \\ \tau'_j &= \tau_{j-1} \qquad && j = j_0+1, \ldots, n+k+1.\end{aligned}$$

Als Folgerung haben wir sofort $\partial s(x_i)/\partial \tau_{j_0} = 0$, falls $x_i \leq \tau_{j_0-k+1}$ oder $x_i \geq \tau_{j_0+k-1}$.

Kommen wir nun zum angekündigten Nachweis der Äquivalenz von vollständigem und reduziertem Glättungsproblem.

Theorem 2.6 (Äquivalenz von vollständigem und reduziertem Problem)**.**
Sei $\mathbf{t}^*$ eine zulässige Knotenfolge, d. h. $\mathbf{t}^* \in \left\{\mathbf{t} \in \mathbb{R}^l : \mathbf{C}\mathbf{t} - \mathbf{h} \geq \mathbf{0}\right\}$. Für ein festes $r \in \{0, \ldots, q\}$, $0 \leq q < k$ gelte:

(V1) Die Knoten erfüllen die Bedingung $\tau_j < \tau_{j+k-q}$ $(j = q+1, \ldots, n)$.

(V2) Die Regularitätsbedingung $m \geq r$ und $\mu > 0$ ist erfüllt.

(V3) Die freien Knoten $\mathbf{t}^*$ sind einfache Knoten, d. h. $\#\tau^*_{p(j)} = 1$ $(j = 1, \ldots, l)$. Es gilt $k \geq 3$.

Dann gelten für das vollständige Glättungsproblem FSP (2.16) und das reduzierte Glättungsproblem RSP (2.31) die folgenden Beziehungen: Die reduzierte Funktion $\mathbf{F}$ von RSP ist glatt im zulässigen Bereich $\left\{\mathbf{t} \in \mathbb{R}^l : \mathbf{C}\mathbf{t} - \mathbf{h} \geq \mathbf{0}\right\}$.

(a) Wenn $\mathbf{t}^*$ ein kritischer Punkt (oder eine globale Minimumstelle) von RSP ist und

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \mathbf{B}(\mathbf{t}^*) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}^*) \end{bmatrix}^+ \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}, \tag{2.38}$$

so ist $(\boldsymbol{\alpha}^*, \mathbf{t}^*)$ ein kritischer Punkt (oder eine globale Minimumstelle) von FSP. Es gilt $f(\mathbf{t}^*) = \mathfrak{f}(\boldsymbol{\alpha}^*, \mathbf{t}^*)$.

(b) Wenn $(\boldsymbol{\alpha}^*, \mathbf{t}^*)$ *eine globale Minimumstelle von* FSP *ist, so ist* $\mathbf{t}^*$ *globale Minimumstelle von* RSP. *Es gilt* $f(\mathbf{t}^*) = \mathfrak{f}(\boldsymbol{\alpha}^*, \mathbf{t}^*)$ *und* (2.38).

Beweis. Voraussetzung (V1) sichert die Existenz der Matrixfunktionen $\mathbf{B}(.)$ und $\mathbf{S}_r(.)$. Wegen (V2) hat die Systemmatrix $\mathbf{B}_\mu(\mathbf{t}) \in \mathbb{R}^{m+n-r,n}$ nach Lemma 2.5 Vollrang n für alle zulässigen $\mathbf{t}$. Unter Verwendung von (V1) und (V3) erhalten wir aus Korollar 2.8 die Differenzierbarkeit der B-Splines und damit der Matrixfunktion $\mathbf{B}(.)$ bez. der freien Knoten. Die Differenzierbarkeit der Glättungsmatrix $\mathbf{S}_r(.)$ folgt unmittelbar aus der Definition. Die Systemmatrix $\mathbf{B}_\mu(.)$ ist also eine Fréchet-differenzierbare Matrixfunktion mit konstantem Rang, erfüllt also alle Voraussetzungen von Theorem 2.1 bzw. 2.2. Die Aussage folgt durch unmittelbare Anwendung dieser Sätze. Dabei beachte man, daß die Menge Ω den gesamten zulässigen Bereich $\left\{\mathbf{t} \in \mathbb{R}^l \,:\, \mathbf{Ct} - \mathbf{h} \geq \mathbf{0}\right\}$ einschließt und daß $\boldsymbol{\alpha}^*$ in Theorem 2.1 (b) wegen des Vollrangs von $\mathbf{B}_\mu(.)$ eindeutig ist. □

Bemerkung 2.1. (i) Im Fall der Splineapproximation muß (V2) durch die Bedingung

($V2'$) Die Schoenberg-Whitney-Bedingung (2.3) ist für alle $\mathbf{t}$ aus einer offenen Umgebung Ω^* von $\mathbf{t}^*$ erfüllt.

ersetzt werden. Man erhält dann nur lokale Aussagen bez. dieser Umgebung und die globale Glattheit des reduzierten Funktionals ist nicht gesichert.

(ii) Eigentlich ist nur die Bedingung $\#\tau_{p(j)} \leq k-2$ $(j = 1, \ldots, l)$ für die Differenzierbarkeit nach den Knoten an allen Stellen $\mathbf{t}$ erforderlich. Der Einfachheit halber, um einheitlich die Formel (2.37) verwenden zu können, beschränken wir uns auf $\#\tau_{p(j)} = 1$ und demzufolge $k \geq 3$. Für die Fälle $k = 1, 2$ existieren spezielle Verfahren zur L_2-Approximation von stetigen Funktionen, siehe z.B. [LW91] und [Bai94], welche sich sinngemäß auf die l_2-Approximation übertragen lassen. Loach und Wathen berichten jedoch, daß selbst im Fall $k = 2$ eine direkte, und somit teure Suche nach einem guten Startpunkt auf einem genügend feinen Gitter nötig ist, um einen robusten Algorithmus zu erhalten. Baines betrachtet unstetige stückweise konstante und lineare L_2-Approximationen.

(iii) Läßt man zusammenfallende Knoten zu, so erhält man Mannigfaltigkeiten mit Spitzen, d. h. die Ableitungen nach den Knoten sind bei mehrfachen Knoten nicht mehr regulär. Man muß dann mit Tangentialkegeln [Cro79] bzw. erweiterten Tangentialkegeln [Mul92] arbeiten. In [Cro79] findet man dazu ein numerisches Verfahren für den Fall der L_2- und L_∞-Approximation von Funktionen.

Theorem 2.6 zeigt, daß der Übergang vom vollständigen zum reduzierten Glättungsproblem keine stationären Punkte erzeugt und daß die Lösung des Originalproblems nicht ausgeschlossen wird. Wir werden uns daher im weiteren darauf konzentrieren, das reduzierte Problem numerisch zu lösen.

2.5 Numerische Lösung des reduzierten Problems

Das reduzierte Problem ist ein nichtlineares Quadratmittelproblem mit linearen Ungleichheitsnebenbedingungen. Es besitzt nur l unabhängige Variable gegenüber $n + l$ beim Ausgangsproblem, ist jedoch von einer größeren Komplexität. Bei der Auswahl eines numerischen Verfahrens zur Lösung dieses Problems ist darauf zu achten, daß

(i) die Quadratmittelstruktur ausgenutzt wird,

(ii) das Verfahren nur mit zulässigen Punkten arbeitet,

(iii) die Bandstruktur der definierenden Matrizen ausgenutzt wird.

Während in fast allen bekannten Verfahren das restringierte reduzierte Problem in ein unrestringiertes Problem transformiert wird, vgl. Einführung, behandeln wir das restringierte Problem direkt durch ein verallgemeinertes Gauß-Newton-Verfahren.

2.5.1 Ein verallgemeinertes Gauß-Newton-Verfahren

Sei $\mathbf{F} \in \mathbb{R}^{m+n-r}$ die Residuumsfunktion des reduzierten Funktionals $f = \frac{1}{2}\|\mathbf{F}\|^2$ und sei $\mathbf{t}^\nu$ die aktuelle Iterierte. Im ν-ten Schritt eines verallgemeinerten Gauß-Newton-Verfahrens ist das quadratische Ersatzproblem

$$\min\left\{\mu(\mathbf{t}^\nu+\mathbf{s}) = \frac{1}{2}\,\|\mathbf{F}(\mathbf{t}^\nu)+\mathbf{J}(\mathbf{t}^\nu)\mathbf{s}\|^2 \;:\; \mathbf{C}\mathbf{s} \geq \mathbf{h}-\mathbf{C}\mathbf{t}^\nu \;:\; \mathbf{s}\in\mathbb{R}^l\right\} \tag{2.39}$$

zu lösen, wobei $\mu(\mathbf{t}^\nu+\mathbf{s}) \approx f(\mathbf{t}^\nu+\mathbf{s})$ ein quadratisches Modell an das Funktional f ist. Wir nennen $\mu = \mu_{GP}$ mit $\mathbf{J}(\mathbf{t}^\nu) = \mathbf{F}'(\mathbf{t}^\nu)$ Golub/Pereyra-Modell und $\mu = \mu_K$ mit $\mathbf{J}(\mathbf{t}^\nu) = \mathbf{J}_K(\mathbf{t}^\nu)$ Kaufman-Modell, vgl. Abschnitt 2.3. Wenn die Jacobi-Matrix oder deren Approximation Vollrang l hat, so besitzt (2.39) eine eindeutige Lösung $\mathbf{s}^\nu$, welche die nächste Iterierte $\mathbf{t}^{\nu+1} := \mathbf{t}^\nu + \mathbf{s}^\nu$ $(\nu = 0, 1, \ldots)$ definiert.

In [Boc87] wurde eine Konvergenztheorie für verallgemeinerte Gauß-Newton-Verfahren entwickelt. Ähnlich wie im unrestringierten Fall kann man lokale Q-lineare Konvergenz für Probleme mit „kleinen Residuen“ nachweisen.

Lösung der linearen Quadratmittelprobleme mit Nebenbedingungen

In jedem Schritt des verallgemeinerten Gauß-Newton-Verfahrens sind lineare Quadratmittelprobleme mit linearen Ungleichheitsnebenbedingungen der Form (2.39) zu lösen. Die Matrix $\mathbf{J}$ (Jacobi-Matrix oder Kaufman-Approximation) ist i. allg. vollbesetzt, obwohl die sie definierenden Matrizen $\mathbf{B}$ und $\mathbf{S}_r$ Bandgestalt haben. Dies ist anschaulich klar, da der Wert der Splinefunktion – und somit F_i – von *jedem* Splineknoten τ_j abhängt.

Wir untersuchen daher kurz die Lösung von vollbesetzten Quadratmittelproblemen mit Nebenbedingungen und betrachten

Problem LSI:

$$\min\{\|\mathbf{A}\mathbf{x}-\mathbf{b}\| \;:\; \mathbf{C}\mathbf{x}\geq\mathbf{h},\, \mathbf{x}\in\mathbb{R}^n\}$$

Problem LDP (Least Distance Programming):

$$\min\{\|\mathbf{x}\| \;:\; \mathbf{C}\mathbf{x}\geq\mathbf{h},\, \mathbf{x}\in\mathbb{R}^n\}$$

Problem NNLS (Nonnegative Least Squares):

$$\min\{\|\mathbf{A}\mathbf{x}-\mathbf{b}\| \;:\; \mathbf{x}\geq\mathbf{0},\, \mathbf{x}\in\mathbb{R}^n\}$$

mit Matrizen $\mathbf{A} \in \mathbb{R}^{m,n}$, $\mathbf{C} \in \mathbb{R}^{ncstr,n}$ und Vektoren $\mathbf{h} \in \mathbb{R}^{ncstr}$, $\mathbf{b} \in \mathbb{R}^m$.

Das Problem LSI mit einer Vollrangmatrix $\mathbf{A}$ wird durch eine orthogonale Transformation in das Problem LDP überführt. Das Problem LDP schließlich wird in das Problem NNLS transformiert, welches mit einer aktiven Mengenstrategie gelöst wird (siehe [LH95, Chapter 23]). Die Prozeduren zur Lösung von LDP und NNLS entnehmen wir den gleichnamigen FORTRAN-Programmen in [LH95]. Wir geben daher nur die Transformation von Problem LSI in Problem LDP an.

Transformation von Problem LSI in Problem LDP

Sei $\mathbf{AP} = \mathbf{Q}\hat{\mathbf{R}}$ eine QR-Faktorisierung der Vollrangmatrix $\mathbf{A} \in \mathbb{R}^{m,n}$ mit

$$\begin{aligned}
&\mathbf{Q} = (\mathbf{Q}_1, \mathbf{Q}_2) \in \mathbb{R}^{m,m} \quad \text{orthogonal, } \mathbf{Q}_1 \in \mathbb{R}^{m,n},\ \mathbf{Q}_2 \in \mathbb{R}^{m,m-n} \\
&\hat{\mathbf{R}} = \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix} \in \mathbb{R}^{m,n}, \mathbf{R} \in \mathbb{R}^{n,n} \quad \text{reguläre obere Dreiecksmatrix} \\
&\mathbf{P} \in \mathbb{R}^{n,n} \quad \text{Permutationsmatrix.}
\end{aligned}$$

Dann gilt

$$\begin{aligned}
\frac{1}{2}\|\mathbf{Ax} - \mathbf{b}\|^2 &= \frac{1}{2}\left\|\left(\mathbf{Q}^T\mathbf{AP}\right)\left(\mathbf{P}^T\mathbf{x}\right) - \mathbf{Q}^T\mathbf{b}\right\|^2 \\
&= \frac{1}{2}\left\|\begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix}\left(\mathbf{P}^T\mathbf{x}\right) - \begin{pmatrix} \mathbf{Q}_1^T\mathbf{b} \\ \mathbf{Q}_2^T\mathbf{b} \end{pmatrix}\right\|^2,
\end{aligned}$$

nach Einführung neuer Variablen $\mathbf{y} := \mathbf{P}^T\mathbf{x} \in \mathbb{R}^n$, $\mathbf{c}_1 := \mathbf{Q}_1^T\mathbf{b} \in \mathbb{R}^n$, $\mathbf{c}_2 := \mathbf{Q}_2^T\mathbf{b} \in \mathbb{R}^{m-n}$

$$= \frac{1}{2}\|\mathbf{Ry} - \mathbf{c}_1\|^2 + \frac{1}{2}\|\mathbf{c}_2\|^2$$

und nach erneuter Variablentransformation $\mathbf{z} := \mathbf{Ry} - \mathbf{c}_1 \in \mathbb{R}^n$ schließlich

$$= \frac{1}{2}\|\mathbf{z}\|^2 + \frac{1}{2}\|\mathbf{c}_2\|^2.$$

Durch Transformation der Nebenbedingungen erhalten wir $\mathbf{CPR}^{-1}\mathbf{z} \geq \mathbf{h} - \mathbf{CPR}^{-1}\mathbf{c}_1$. Das Problem LSI ist damit äquivalent zu dem Least Distance Problem

$$\min\left\{\|\mathbf{z}\| \ :\ \mathbf{CPR}^{-1}\mathbf{z} \geq \mathbf{h} - \mathbf{CPR}^{-1}\mathbf{c}_1, \mathbf{z} \in \mathbb{R}^n\right\}.$$

Die notwendige QR-Faktorisierung erfolgt durch Householder-Spiegelungen mit Spaltenpivotisierung. Vor der Lösung der Gleichungssysteme mit der oberen Dreiecksmatrix $\mathbf{R}$ erfolgt eine Schätzung der Konditionszahl. Wenn $cond(\mathbf{R}) > 1 / \sqrt{macheps}$ gilt, brechen wir den Algorithmus ab.

Ist die Matrix $\mathbf{J}$ in obigem Sinne (fast) rangdefizient, so ersetzen wir das Originalproblem (2.39) durch das regularisierte Problem

$$\min\left\{\frac{1}{2}\left\|\begin{pmatrix} \mathbf{F} \\ \mathbf{0} \end{pmatrix} + \begin{bmatrix} \mathbf{J} \\ \sqrt{\lambda}\mathbf{I} \end{bmatrix}\mathbf{s}\right\|^2 \ :\ \mathbf{Cs} \geq \mathbf{h} - \mathbf{Ct}^\nu,\ \mathbf{s} \in \mathbb{R}^l\right\}$$

mit $\lambda = \sqrt{l \times \text{macheps}}\left\|\mathbf{J}^T\mathbf{J}\right\|_1$. Diese Strategie wird von Dennis/Schnabel [DS83, S. 151] vorgeschlagen und in dem zugehörigen Programmpaket realisiert. Dort findet sich auch eine theoretische Fundierung für die Wahl der Größe des Lagrange-Parameters λ.

Globalisierung

Das bisher betrachtete verallgemeinerte Gauß-Newton-Verfahren konvergiert nur lokal und bei Problemen mit kleinen Residuen. Wir verwenden eine Dämpfungsstrategie, um globale Konvergenz zu erreichen. Da die Nebenbedingungen linear sind, ist es möglich, nur mit zulässigen Punkten zu arbeiten, und die Zielfunktion $f(\mathbf{t}) = \frac{1}{2}\|\mathbf{F}(\mathbf{t})\|^2$ kann als Niveaufunktion bei der Strahlsuche verwendet werden. Globale Konvergenzaussagen zu gedämpften verallgemeinerten Gauß-Newton-Verfahren findet man in [Boc87, Satz 3.2.6].

Wir haben eine abgesicherte Strahlsuche nach Armijo/Goldstein mit gemischter quadratischer und kubischer Interpolation implementiert, siehe [DS83, S. 128ff] für Details. Die Informationen, welche sowieso aus der Lösung der linearen Quadratmittelprobleme anfallen (siehe nächster Abschnitt), erlauben es, den Funktionswert $f(\mathbf{t}^\nu + \gamma\mathbf{s}^\nu) = \frac{1}{2}\|\mathbf{F}(\mathbf{t}^\nu + \gamma\mathbf{s}^\nu)\|^2$ *ohne* explizite Bildung des Residuumsvektors $\mathbf{F}(\mathbf{t}^\nu + \gamma\mathbf{s}^\nu)$ zu berechnen. Es sei angemerkt, daß die bei Quadratmittelproblemen weit verbreitete Strahlsuche nach Al-Baali/Fletcher [ABF86], bei der jede individuelle Komponente der Residuumsfunktion durch ein Polynom approximiert wird, die explizite Kenntnis des Residuumsvektors erfordert. Bei den durchgeführten numerischen Tests war die Anzahl abgelehnter Schritte bei Verwendung unserer Strahlsuche relativ gering.

Abschließend wollen wir den kompletten Algorithmus formulieren. Dieser Algorithmus stellt das Basisverfahren für die Lösung der Probleme in diesem und den folgenden Kapiteln dar. Lediglich die Berechnung von Residuumsfunktion und Jacobi-Matrix ändert sich.

Algorithmus 2.1 (Gedämpftes verallgemeinertes Gauß-Newton-Verfahren).S1: Wähle zulässige Startknotenfolge $\mathbf{t}^0 \in \mathbb{R}^l$, wähle $\delta \in (0, \frac{1}{4})$, setze $\nu := 0$

S2: Repeat

S2.1: $\mathbf{F} := \mathbf{F}(\mathbf{t}^\nu)$ {Residuumsfunktion des reduzierten Funktionals}
$\mathbf{J} :\approx \boldsymbol{\partial}\mathbf{F}(\mathbf{t}^\nu)$ {Approximation der Jacobi-Matrix (Golub/Pereyra, Kaufman-Modell)}

S2.2: Berechne Abstiegsrichtung $\mathbf{s}^\nu$ aus

$$\min\left\{\tfrac{1}{2}\left\|\mathbf{F}+\mathbf{J}\mathbf{s}\right\|^2 \,:\, \mathbf{C}\mathbf{s} \geq \mathbf{h} - \mathbf{C}\mathbf{t}^\nu\,,\, \mathbf{s}\in\mathbb{R}^l\right\}$$

Falls das Problem schlecht konditioniert ist, berechne $\mathbf{s}^\nu$ aus dem regularisierten Problem

$$\min\left\{\tfrac{1}{2}\left\|\begin{pmatrix}\mathbf{F}\\ \mathbf{0}\end{pmatrix}+\begin{bmatrix}\mathbf{J}\\ \sqrt{\lambda}\mathbf{I}\end{bmatrix}\mathbf{s}\right\|^2 \,:\, \mathbf{C}\mathbf{s} \geq \mathbf{h} - \mathbf{C}\mathbf{t}^\nu\,,\, \mathbf{s}\in\mathbb{R}^l\right\}$$

mit $\lambda = \sqrt{l \times \text{macheps}}\left\|\mathbf{J}^T\mathbf{J}\right\|_1$

S2.3: Strahlsuche
$\gamma := 1.0$
while $f(\mathbf{t}^\nu) - f(\mathbf{t}^\nu + \gamma\mathbf{s}^\nu) < -\gamma\delta\nabla f(\mathbf{t}^\nu)^T\mathbf{s}^\nu$ do
$\gamma := \gamma * \alpha$
{α wird so gewählt, daß $\gamma * \alpha$ die Minimumstelle eines quadratischen oder kubischen Polynoms ist, welches $f(\mathbf{t} + \gamma\mathbf{s}^\nu)$ approximiert. Dabei wird sichergestellt, daß α nicht zu nah an den Rändern 0 oder 1 ist.}

S2.4: $\mathbf{t}^{\nu+1} := \mathbf{t}^\nu + \gamma\mathbf{s}^\nu$
$\nu := \nu + 1$

until $\nu > \nu max$ or *Konvergenz*

2.5.2 Die Berechnung der Residuumsfunktion

Zur Berechnung der Residuumsfunktion $\mathbf{F}$ sowie der Matrizen $\mathbf{F}'$ oder $\mathbf{J}_K$ des verwendeten quadratischen Modells benötigen wir ein schnelles und stabiles Verfahren zur Lösung der linearen Quadratmittelprobleme

$$\min\left\{\frac{1}{2}\left\|\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix}-\begin{bmatrix}\mathbf{B}\\ \sqrt{\mu}\mathbf{S}_r\end{bmatrix}\boldsymbol{\alpha}\right\|^2 : \boldsymbol{\alpha}\in\mathbb{R}^n\right\}. \tag{2.40}$$

Wir benutzen Algorithmen aus [Cox81] und [Eld84] (siehe auch [SK93]) und reduzieren die Systemmatrix auf obere Dreiecksform unter Benutzung zeilenweiser Givens-Drehungen.

In einem ersten Schritt wird die Matrix $\mathbf{B}$ durch Linksmultiplikation mit einer Folge von geeigneten Givens-Drehungen auf obere Dreiecksform transformiert:

Beispiel 2.3. $k = 4$, $n = 9$, $m = 12$

$$\mathbf{B} = \begin{bmatrix}
x & x & x & x & & & & & \\
x & x & x & x & & & & & \\
x & x & x & x & & & & & \\
 & x & x & x & x & & & & \\
 & & x & x & x & x & & & \\
 & & x & x & x & x & & & \\
 & & & x & x & x & x & & \\
 & & & x & x & x & x & & \\
 & & & & x & x & x & x & \\
 & & & & x & x & x & x & \\
 & & & & & x & x & x & x \\
 & & & & & x & x & x & x
\end{bmatrix} \xrightarrow{\text{Givens}} \begin{bmatrix}\mathbf{R}_0\\ \mathbf{0}\end{bmatrix} = \begin{bmatrix}
x & x & x & x & & & & & \\
 & x & x & x & x & & & & \\
 & & x & x & x & x & & & \\
 & & & x & x & x & x & & \\
 & & & & x & x & x & x & \\
 & & & & & x & x & x & x \\
 & & & & & & x & x & x \\
 & & & & & & & x & x \\
 & & & & & & & & x \\
 & & & & \mathbf{0} & & & & \\
 & & & & & & & &
\end{bmatrix}$$

Formal erhalten wir $\mathbf{Q}_0^T\mathbf{B} = \begin{bmatrix}\mathbf{R}_0\\ \mathbf{0}\end{bmatrix}$, $\mathbf{Q}_0 \in \mathbb{R}^{m,m}$ orthogonal, $\mathbf{R}_0 \in \mathbb{R}^{n,n}$ obere Dreiecksmatrix der Bandbreite k. Die rechte Seite $\mathbf{y} \in \mathbb{R}^m$ kann simultan gemäß $\mathbf{Q}_0^T\mathbf{y} = \begin{pmatrix}\mathbf{c}\\ \mathbf{d}\end{pmatrix}$, $\mathbf{c} \in \mathbb{R}^n$, $\mathbf{d} \in \mathbb{R}^{m-n}$ transformiert werden. Die Matrix $\mathbf{B}$ kann zeilenweise abgearbeitet werden, und es ist nur nötig, Speicherplatz für $\mathbf{R}_0$ bereitzuhalten. Die einzelnen Zeilen von $\mathbf{B}$, d. h. alle nichtverschwindenden B-Splines an einer Stelle x_i, werden simultan mittels der Rekursionsformel berechnet, siehe [dB78, S. 134f]. Für die wiederholte Lösung solcher Quadratmittelprobleme mit verschiedenen rechten Seiten wird die relevante Information zur Rekonstruktion der Givens-Parameter im Band von $\mathbf{B}$ gespeichert.

Zur Berechnung der zeilenweisen QR-Faktorisierung einer Matrix $\mathbf{B} \in \mathbb{R}^{m,n}$ (k – Zeilenbandbreite) benötigt man nach [Cox81] (siehe auch [Bjö96]) die folgende Anzahl von arithmetischen Operationen:

Matrix-Struktur	Standardform	Nicht-Standardform
volle Matrix	$2mn^2$	$2mn^2$
Bandmatrix	$2mk^2$	$2mnk$

Eine Matrix ist in Standardform, wenn der Index des am weitesten rechts stehenden Nichtnullelementes jeder Zeile eine nichtfallende Funktion der Zeilennummer ist. Offenbar ist die Matrix $\mathbf{B}$ in Standardform, die Systemmatrix $\mathbf{B}_\mu$ dagegen nicht.

Nach dieser orthogonalen Transformation kann das Quadratmittelproblem (2.40) äquivalent dargestellt werden als

$$\min\left\{\frac{1}{2}\left\|\begin{pmatrix}\mathbf{c}\\ \mathbf{0}\end{pmatrix}-\begin{bmatrix}\mathbf{R}_0\\ \sqrt{\mu}\mathbf{S}_r\end{bmatrix}\boldsymbol{\alpha}\right\|^2+\frac{1}{2}\|\mathbf{d}\|^2\ :\ \boldsymbol{\alpha}\in\mathbb{R}^n\right\}.$$

Im nächsten Schritt wird die erweiterte Matrix $\begin{bmatrix}\mathbf{R}_0\\ \sqrt{\mu}\mathbf{S}_r\end{bmatrix}$ auf obere Dreiecksform gebracht. Dazu werden die Zeilen zunächst so permutiert, daß im wesentlichen die Zeilen von $\mathbf{R}_0$ und $\sqrt{\mu}\mathbf{S}_r$ abwechselnd behandelt werden – mit Ausnahme der letzten Zeilen. Wir verdeutlichen dies am Beispiel der approximierten Glättungsmatrix (Bandbreite $r+1$!):

Beispiel 2.4. $k=4$, $n=9$, $r=2$

$$\begin{bmatrix}
x&x&x&x&&&&&\\
&x&x&x&x&&&&\\
&&x&x&x&x&&&\\
&&&x&x&x&x&&\\
&&&&x&x&x&x&\\
&&&&&x&x&x&x\\
&&&&&&x&x&x\\
&&&&&&&x&x\\
&&&&&&&&x\\
x&x&x&&&&&&\\
&x&x&x&&&&&\\
&&x&x&x&&&&\\
&&&x&x&x&&&\\
&&&&x&x&x&&\\
&&&&&x&x&x&\\
&&&&&&x&x&x
\end{bmatrix}
\rightarrow
\begin{bmatrix}
x&x&x&&&&&&\\
x&x&x&x&&&&&\\
&x&x&x&&&&&\\
&x&x&x&x&&&&\\
&&x&x&x&&&&\\
&&x&x&x&x&&&\\
&&&x&x&x&&&\\
&&&x&x&x&x&&\\
&&&&x&x&x&&\\
&&&&x&x&x&x&\\
&&&&&x&x&x&x\\
&&&&&&x&x&x\\
&&&&&&&x&x\\
&&&&&&&&x
\end{bmatrix}$$

In Matrixnotation erhalten wir für diese Transformation $\tilde{\mathbf{Q}}^T\begin{bmatrix}\mathbf{R}_0\\ \sqrt{\mu}\mathbf{S}_r\end{bmatrix}=\begin{bmatrix}\tilde{\mathbf{R}}\\ \mathbf{0}\end{bmatrix}$, $\tilde{\mathbf{Q}}\in\mathbb{R}^{2n-r,2n-r}$ orthogonal, $\tilde{\mathbf{R}}\in\mathbb{R}^{n,n}$ obere Dreiecksmatrix der Bandbreite k. Unter der Voraussetzung $m\geq r$ und $\mu>0$ ist $\tilde{\mathbf{R}}$ regulär. Transformieren wir die rechte Seite gemäß $\tilde{\mathbf{Q}}^T\begin{pmatrix}\mathbf{c}\\ \mathbf{0}\end{pmatrix}=\begin{pmatrix}\tilde{\mathbf{c}}\\ \tilde{\mathbf{d}}\end{pmatrix}$, so ist (2.40) äquivalent zu

$$\min\left\{\frac{1}{2}\|\tilde{\mathbf{c}}-\tilde{\mathbf{R}}\boldsymbol{\alpha}\|^2+\frac{1}{2}\|\tilde{\mathbf{d}}\|^2+\frac{1}{2}\|\mathbf{d}\|^2\ :\ \boldsymbol{\alpha}\in\mathbb{R}^n\right\},$$

dessen Lösung wir unmittelbar aus dem oberen Dreieckssystem $\tilde{\mathbf{R}}\boldsymbol{\alpha}=\tilde{\mathbf{c}}$ berechnen können. Man beachte, daß nur die zweite Transformation bei Änderung des Glättungsparameters μ erneut ausgeführt werden muß. Das Residuum $\|\tilde{\mathbf{d}}\|^2+\|\mathbf{d}\|^2$ des Quadratmittelproblems fällt bei der orthogonalen Transformation automatisch an.

2.5.3 Die Berechnung der Jacobi-Matrix

Eine effiziente Berechnung der Jacobi-Matrix unter Ausnutzung der Schwachbesetztheitsstrukturen ist wesentlich für einen schnellen Algorithmus zur Berechnung von Splines mit

freien Knoten. Formeln zur Berechnung der Jacobi-Matrix $\mathbf{F}'$ oder der Kaufman-Approximation $\mathbf{J}_K$ im Fall der Spline-Glättung erhält man, indem man die Größen $\{\mathbf{B}(\mathbf{t}), \mathbf{y}\}$ in (2.25), (2.26) bzw. (2.29) durch das Paar

$$\left\{ \begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}) \end{bmatrix}, \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} \right\}$$

ersetzt. Im Fall der Kaufman-Approximation erhalten wir

$$\mathbf{J}_K\mathbf{s} := -\left\{ \begin{bmatrix} \mathbf{I}_m \\ \mathbf{I}_{n-r} \end{bmatrix} - \begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix} \begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix}^+ \right\} \boldsymbol{\partial} \begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix} [\mathbf{s}] \begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix}^+ \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}$$

für alle $\mathbf{s} \in \mathbb{R}^l$. Die Matrixfunktionen sind dabei jeweils an der Stelle $\mathbf{t}$ zu berechnen. Wir berechnen die Kaufman-Approximation $\mathbf{J}_K \in \mathbb{R}^{m+n-r,l}$ spaltenweise, d. h. $\mathbf{J}_K\mathbf{e}^\kappa \in \mathbb{R}^{m+n-r}$ ($\kappa = 1, \ldots, l$), $\mathbf{e}^\kappa \in \mathbb{R}^l$ – Einheitsvektor.

Betrachten wir zunächst die Matrizen, welche Ableitungen bez. der freien Knoten enthalten

$$\boldsymbol{\partial} \begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}) \end{bmatrix} [\mathbf{e}^\kappa] \underbrace{\begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}) \end{bmatrix}^+ \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}}_{=:\boldsymbol{\alpha}(\mathbf{t})} = \boldsymbol{\partial} \begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}) \end{bmatrix} [\mathbf{e}^\kappa]\boldsymbol{\alpha}(\mathbf{t}).$$

Ableitung der Beobachtungsmatrix nach den freien Knoten

Offensichtlich gilt

$$\left(\boldsymbol{\partial} \begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}) \end{bmatrix} [\mathbf{e}^\kappa]\boldsymbol{\alpha}(\mathbf{t}) \right)_{i=1,\ldots,m} = \left(\sum_{j=1}^{n} \frac{\partial B_{j,k,\boldsymbol{\tau}}(x_i)}{\partial \tau_{p(\kappa)}} \alpha_j \right)_{i=1,\ldots,m} = \left(\frac{\partial s(x_i)}{\partial \tau_{p(\kappa)}} \right)_{i=1,\ldots,m},$$

d. h. wir können mittels Lemma 2.9 die ersten m Komponenten berechnen. Wegen $\frac{\partial s(x_i)}{\partial \tau_{p(\kappa)}} = 0$ für $x_i \le \tau_{p(\kappa)-k+1}$ oder $x_i \ge \tau_{p(\kappa)+k-1}$ und der Struktur der Ableitung (2.37) ist dies auf sehr effiziente Weise möglich.

Ableitung der Glättungsmatrix nach den freien Knoten

Für die letzten $n - r$ Komponenten benötigen wir die Ableitung der Glättungsmatrix $\mathbf{S}_r(\mathbf{t}) = \mathbf{F}_r(\mathbf{t})\mathbf{D}_r(\mathbf{t})$. Es gilt $\boldsymbol{\partial}\mathbf{S}_r = (\boldsymbol{\partial}\mathbf{F}_r)\mathbf{D}_r + \mathbf{F}_r(\boldsymbol{\partial}\mathbf{D}_r)$. Während $\boldsymbol{\partial}\mathbf{F}_r = \boldsymbol{\partial}\tilde{\mathbf{F}}_r$ im Fall der approximierten Glättungsmatrix einfach berechnet werden kann, erscheint es sehr schwierig, $\boldsymbol{\partial}\mathbf{F}_r = \boldsymbol{\partial}\bar{\mathbf{F}}_r$ im Fall der exakten Glättungsmatrix anzugeben ($\bar{\mathbf{F}}_r$ – Cholesky-Faktor der Gramschen Matrix von B-Splines der Ordnung $k - r$). In diesem Fall können Methoden des automatischen Differenzierens angewendet werden.

Wir wollen zunächst die Fréchet-Ableitung der Matrixfunktion $\mathbf{D}_r(.)$ bez. der freien Knoten $\mathbf{t}$ bestimmen, siehe Lemma 2.3 zur Definition von $\mathbf{D}_r$. Offensichtlich gilt $\mathbf{H}_\nu = \mathbf{H}_\nu(\mathbf{t})$, $\boldsymbol{\partial}\mathbf{L}_\nu = \mathbf{0}$ ($\nu = 1, \ldots, r$). Für $\boldsymbol{\partial}\mathbf{D}_r \in \mathfrak{L}\left(\mathbb{R}^l, \mathfrak{L}\left(\mathbb{R}^n, \mathbb{R}^{n-r}\right)\right)$ hat man

$$\boldsymbol{\partial}\mathbf{D}_0 = \mathbf{0}, \quad \boldsymbol{\partial}\mathbf{D}_1 = (\boldsymbol{\partial}\mathbf{H}_1)\,\mathbf{L}_1, \quad \boldsymbol{\partial}\mathbf{D}_2 = (\boldsymbol{\partial}\mathbf{H}_2)\,\mathbf{L}_2\mathbf{D}_1 + \mathbf{H}_2\mathbf{L}_2\,(\boldsymbol{\partial}\mathbf{D}_1)\,.$$

Allgemein zeigt man leicht durch vollständige Induktion

$$\boldsymbol{\partial}\mathbf{D}_r = \begin{cases} \mathbf{0} & \text{für } r = 0, \\ (\boldsymbol{\partial}\mathbf{H}_r)\,\mathbf{L}_r\mathbf{D}_{r-1} + \mathbf{H}_r\mathbf{L}_r\,(\boldsymbol{\partial}\mathbf{D}_{r-1}) & \text{für } r \ge 1. \end{cases}$$

Damit erhalten wir $\boldsymbol{\partial}\mathbf{D}_r[\mathbf{e}^\kappa]\mathbf{c}^n = (\boldsymbol{\partial}\mathbf{H}_r)\,[\mathbf{e}^\kappa]\mathbf{L}_r\mathbf{D}_{r-1}\mathbf{c}^n + \mathbf{H}_r\mathbf{L}_r\,(\boldsymbol{\partial}\mathbf{D}_{r-1})\,[\mathbf{e}^\kappa]\mathbf{c}^n \in \mathbb{R}^{n-r}$ mit dem Einheitsvektor $\mathbf{e}^\kappa \in \mathbb{R}^l$, $\kappa = 1,\ldots,l$ sowie einem beliebigen Vektor $\mathbf{c}^n \in \mathbb{R}^n$. Wir betrachten nun $\boldsymbol{\partial}\mathbf{H}_\nu[\mathbf{e}^\kappa]\mathbf{c}^{n-\nu}$ für $\nu = 1,\ldots,r$; $\kappa = 1,\ldots,l$ und $\mathbf{c}^{n-\nu} \in \mathbb{R}^{n-\nu}$. Es gilt

$$\boldsymbol{\partial}\mathbf{H}_\nu[\mathbf{e}^\kappa]\mathbf{c}^{n-\nu} = \left(\sum_{j=1}^{n-\nu} \frac{\partial H_{ij}(\mathbf{t})}{\partial \tau_{p(\kappa)}} c_j\right)_{i=1,\ldots,n-\nu}. \tag{2.41}$$

Laut Definition von $(\mathbf{H}_\nu)_{ij} = H_{ij}$ gilt

$$H_{ij}(\mathbf{t}) = \begin{cases} 0 & \text{falls } i \neq j, \\ \frac{k-\nu}{\tau_{k+j}-\tau_{\nu+j}} & \text{falls } i = j; i,j = 1,\ldots,n-\nu, \end{cases}$$

also

$$\frac{\partial H_{ij}(\mathbf{t})}{\partial \tau_{p(\kappa)}} = 0 \quad \text{falls } i \neq j, \qquad \frac{\partial H_{jj}(\mathbf{t})}{\partial \tau_{p(\kappa)}} = \begin{cases} -\frac{k-\nu}{\left(\tau_{k+j}-\tau_{\nu+j}\right)^2} & \text{falls } p(\kappa) = k+j, \\ \frac{k-\nu}{\left(\tau_{k+j}-\tau_{\nu+j}\right)^2} & \text{falls } p(\kappa) = \nu+j, \\ 0 & \text{sonst.} \end{cases} \tag{2.42}$$

Mittels (2.41) und (2.42) kann $\mathbf{c}^{n-\nu}$ direkt mit $\boldsymbol{\partial}\mathbf{H}_\nu[\mathbf{e}^\kappa]\mathbf{c}^{n-\nu}$ überschrieben werden, wobei nur zwei Elemente verschieden von Null sind.

Nun sind wir in der Lage, einen Algorithmus zur Berechnung der Ableitung der Matrixfunktion $\mathbf{D}_r(.)$ nach den Knoten anzugeben:

Algorithmus 2.2 (Berechnung von $\mathbf{v} := \boldsymbol{\partial}\mathbf{D}_r[\mathbf{e}^\kappa]\boldsymbol{\alpha} \in \mathbb{R}^{n-r}$).
$\mathbf{v} := \mathbf{0}$; $\qquad \{(\boldsymbol{\partial}\mathbf{D}_0)\,[\mathbf{e}^\kappa]\boldsymbol{\alpha} = \mathbf{0} \in \mathbb{R}^n\}$
for $\nu := 1$ to r do

begin

$\mathbf{v}^1 := \mathbf{L}_\nu\mathbf{D}_{\nu-1}\boldsymbol{\alpha}$; $\qquad \{\mathbf{v}^1 = \mathbf{L}_\nu\mathbf{D}_{\nu-1}\boldsymbol{\alpha} \in \mathbb{R}^{n-\nu}\}$

$\mathbf{v}^1 := (\boldsymbol{\partial}\mathbf{H}_\nu)\,[\mathbf{e}^\kappa]\mathbf{v}^1$; $\qquad \{\mathbf{v}^1 = (\boldsymbol{\partial}\mathbf{H}_\nu)\,[\mathbf{e}^\kappa]\mathbf{L}_\nu\mathbf{D}_{\nu-1}\boldsymbol{\alpha} \in \mathbb{R}^{n-\nu}\}$

$\mathbf{v}^2 := \mathbf{v}$; $\qquad \{\mathbf{v}^2 = (\boldsymbol{\partial}\mathbf{D}_{\nu-1})\,[\mathbf{e}^\kappa]\boldsymbol{\alpha} \in \mathbb{R}^{n-\nu+1}\}$

$\mathbf{v}^2 := \mathbf{H}_\nu\mathbf{L}_\nu\mathbf{v}^2$; $\qquad \{\mathbf{v}^2 = \mathbf{H}_\nu\mathbf{L}_\nu\,(\boldsymbol{\partial}\mathbf{D}_{\nu-1})\,[\mathbf{e}^\kappa]\boldsymbol{\alpha} \in \mathbb{R}^{n-\nu}\}$

$\mathbf{v} := \mathbf{v}^1 + \mathbf{v}^2$; $\qquad \{\mathbf{v} = (\boldsymbol{\partial}\mathbf{D}_\nu)\,[\mathbf{e}^\kappa]\boldsymbol{\alpha} \in \mathbb{R}^{n-\nu}\}$

end;

Wir betrachten nun die Fréchet-Ableitung der Matrix $\tilde{\mathbf{F}}_r$, d. h. $\boldsymbol{\partial}\tilde{\mathbf{F}}_r[\mathbf{e}^\kappa]\mathbf{c}^{n-r}$ für $\kappa = 1,\ldots,l$ und $\mathbf{c}^{n-r} \in \mathbb{R}^{n-r}$. Es gilt

$$\boldsymbol{\partial}\tilde{\mathbf{F}}_r[\mathbf{e}^\kappa]\mathbf{c}^{n-r} = \left(\sum_{j=1}^{n-r} \frac{\partial \tilde{F}_{ij}(\mathbf{t})}{\partial \tau_{p(\kappa)}} c_j\right)_{i=1,\ldots,n-r.} \tag{2.43}$$

Laut Definition von $\left(\tilde{\mathbf{F}}_r\right)_{ij} = \tilde{F}_{ij}$ gilt

$$\tilde{F}_{ij}(\mathbf{t}) = \begin{cases} 0 & \text{falls } i \neq j, \\ \sqrt{\frac{\tau_{k+j}-\tau_{r+j}}{k-r}} & \text{falls } i = j; i,j = 1,\ldots,n-r, \end{cases}$$

also

$$(2.44)\quad \frac{\partial \tilde{F}_{ij}(\mathbf{t})}{\partial \tau_{p(\kappa)}} = 0 \qquad \text{falls } i \neq j, \qquad \frac{\partial \tilde{F}_{jj}(\mathbf{t})}{\partial \tau_{p(\kappa)}} = \begin{cases} -\frac{1}{2}\sqrt{\frac{k-r}{\tau_{k+j}-\tau_{r+j}}} & \text{falls } p(\kappa) = k+j, \\ \frac{1}{2}\sqrt{\frac{k-r}{\tau_{k+j}-\tau_{r+j}}} & \text{falls } p(\kappa) = r+j, \\ 0 & \text{sonst.} \end{cases}$$

Erneut wird $\mathbf{c}^{n-r}$ direkt mit $\boldsymbol{\partial}\tilde{\mathbf{F}}_r[\mathbf{e}^\kappa]\mathbf{c}^{n-r}$ überschrieben, wobei wiederum nur zwei Elemente verschieden von Null sind.

Schließlich geben wir den Algorithmus zur Berechnung der Ableitung der approximierten Glättungsmatrix nach den Knoten an. Für $\kappa \in \{1, \ldots, l\}$ und $\boldsymbol{\alpha} \in \mathbb{R}^n$ gilt

$$\left(\boldsymbol{\partial}\tilde{\mathbf{S}}_r\right)[\mathbf{e}^\kappa]\boldsymbol{\alpha} = \left(\boldsymbol{\partial}\tilde{\mathbf{F}}_r\right)[\mathbf{e}^\kappa]\mathbf{D}_r\boldsymbol{\alpha} + \tilde{\mathbf{F}}_r\left(\boldsymbol{\partial}\mathbf{D}_r\right)[\mathbf{e}^\kappa]\boldsymbol{\alpha} \in \mathbb{R}^{n-r}.$$

Algorithmus 2.3 (Berechnung von $\mathbf{v} := \left(\boldsymbol{\partial}\tilde{\mathbf{S}}_r\right)[\mathbf{e}^\kappa]\boldsymbol{\alpha} \in \mathbb{R}^{n-r}$).S1: Berechne $\mathbf{v}^1 := \left(\boldsymbol{\partial}\mathbf{D}_r\right)[\mathbf{e}^\kappa]\boldsymbol{\alpha} \in \mathbb{R}^{n-r}$ mittels Algorithmus 2.2;

S2: Berechne $\mathbf{v}^1 := \tilde{\mathbf{F}}_r\mathbf{v}^1 \in \mathbb{R}^{n-r}$;

S3: Berechne $\mathbf{v}^2 := \mathbf{D}_r\boldsymbol{\alpha} \in \mathbb{R}^{n-r}$;

S4: Berechne $\mathbf{v}^2 := \left(\boldsymbol{\partial}\tilde{\mathbf{F}}_r\right)[\mathbf{e}^\kappa]\mathbf{v}^2 \in \mathbb{R}^{n-r}$;

S5: Setze $\mathbf{v} := \mathbf{v}^1 + \mathbf{v}^2$; $\qquad \{\mathbf{v} = \left(\boldsymbol{\partial}\tilde{\mathbf{S}}_r\right)[\mathbf{e}^\kappa]\boldsymbol{\alpha} \in \mathbb{R}^{n-r}\}$

Mittels Algorithmus 2.3 werden die letzten $n-r$ Komponenten der gesuchten Ableitung nach den freien Knoten berechnet. Unter Benutzung der bereits berechneten QR-Faktorisierungen kann damit die Kaufman-Approximation $\mathbf{J}_K$ spaltenweise aufgebaut werden. Wenn die benötigten Ableitungen in der obigen Weise bereitgestellt werden, so bezeichnen wir den Algorithmus mit RSP-Ka-ED (reduced smoothing problem, Kaufman model, exact derivatives).

Bei Benutzung des Golub/Pereyra-Modells approximieren wir die Jacobi-Matrix $\mathbf{F}'(\mathbf{t})$ spaltenweise durch finite Differenzen

$$\mathbf{F}'(\mathbf{t})\mathbf{e}^\kappa \approx \frac{\mathbf{F}(\mathbf{t} + h_\kappa\mathbf{e}^\kappa) - \mathbf{F}(\mathbf{t})}{h_\kappa} \qquad (\kappa = 1, \ldots, l),$$

wobei die Schrittweite gemäß $h_\kappa = \epsilon_1(|\tau_{p(\kappa)}| + \epsilon_2)$, $\epsilon_1 \approx \sqrt{macheps}$ gewählt wird. Da hier die Diskretisierung unter Vernachlässigung der inneren Struktur der Jacobi-Matrix erfolgt, nennen wir diese Approximation äußere Diskretisierung und bezeichnen den Algorithmus mit RSP-GP-OD (reduced smoothing problem, Golub/Pereyra model, outer discretization). Man beachte, daß diese Diskretisierung die zusätzliche Lösung von l Quadratmittelproblemen des Typs (2.40) erfordert. Eine Alternative – etwa bei Verwendung der exakten Glättungsmatrix – besteht in der Diskretisierung im Inneren des Ausdrucks für die Kaufman-Approximation. Die Ableitungen der B-Splines nach den freien Knoten können z.B. gemäß

$$\left(\boldsymbol{\partial}\mathbf{B}(\mathbf{t})\,[\mathbf{e}^\kappa]\right)_{i,j} \approx \frac{B_j\,(x_i, \mathbf{t} + h_\kappa\mathbf{e}^\kappa) - B_j\,(x_i, \mathbf{t})}{h_\kappa}$$

berechnet werden, analog für $\boldsymbol{\partial}\mathbf{S}_r(\mathbf{t})[\mathbf{e}^\kappa]$. Den entsprechenden Algorithmus bezeichnen wir mit RSP-Ka-ID (reduced smoothing problem, Kaufman model, inner discretization).

Für einen Vergleich zwischen dem diskretisierten Golub/Pereyra-Modell und dem Kaufman-Modell sind sowohl die Kosten der linearen Algebra als auch die Approximationsqualität zu beachten. Letztere ist vergleichbar, wie numerische Tests verschiedener Autoren zeigen. Die Kosten der linearen Algebra hängen sehr stark von der Matrix $\mathbf{B}$ oder ihrer Regularisierung $\mathbf{B}_\mu$ ab. Falls die Matrix vollbesetzt ist, so ist i. allg. das Kaufman-Modell billiger, da die l zusätzlichen Quadratmittelprobleme mit verschiedenen Systemmatrizen des Golub/Pereyra-Modells durch l Quadratmittelprobleme mit verschiedenen rechten Seiten, aber derselben Matrix ersetzt werden. Andererseits ist das diskretisierte Golub/Pereyra-Modell wesentlich einfacher zu implementieren, da es die Feinstruktur vernachlässigt und lediglich Code zur Berechnung der Residuumsfunktion $\mathbf{F}(\mathbf{t})$ benötigt wird.

Zum Vergleich mit existierenden Algorithmen zur Berechnung von Splines mit freien Knoten, welche stets den Approximationsfall, d. h. $\mu = 0$, betrachten, haben wir eigens Algorithmen für diesen Fall implementiert. Dabei ergeben sich wesentliche Vereinfachungen. Die entsprechenden Modelle heißen RAP-Ka-ED (reduced approximation problem, Kaufman model, exact derivatives) bzw. RAP-GP-OD (reduced approximation problem, Golub/Pereyra model, outer discretization).

Leider können wir an dieser Stelle nicht auf den sehr interessanten Aspekt der optimalen Wahl des Glättungsparameters μ eingehen. In den Beispielen wurde μ interaktiv bestimmt. Eine Standardmethode zur optimalen Wahl aus statistischer Sicht besteht in der Minimierung des GCV-Funktionals, siehe [Wah90] und [Eub88] sowie [Wah82] für den restringierten Fall. Betrachtet man die Aufgabe als Regularisierung eines diskreten schlechtgestellten Problems, so kommt die L-curve method, siehe [Han92], in Frage.

2.5.4 FREE – Ein Programm zur Berechnung von Splines mit freien Knoten

Die Algorithmen dieses und des nächsten Kapitels wurden in einem umfangreichen Programmpaket FREE zur Berechnung und Visualisierung von Splines mit freien Knoten implementiert. Die Quellen des Programms – etwa 14000 Zeilen Pascal-Quelltext – stehen zusammen mit einigen Testdaten zur freien Verfügung[1]. Eine ausführliche Beschreibung findet man in [Sch96].

Da die Bandstruktur der beteiligten Matrizen vollständig ausgenutzt wird, sind die Anforderungen an Speicherplatz und Rechenzeit relativ gering. Besonders bei Verwendung des regularisierenden Glättungsterms ist der Algorithmus sehr robust und wurde bereits an realen Daten erfolgreich eingesetzt.

2.6 Numerische Tests

Dieser Abschnitt enthält einige Beispiele aus der Literatur, welche die Leistungsfähigkeit des entwickelten Verfahrens zeigen sollen. Alle Testrechnungen wurden auf einem Pentium-PC in IEEE double Arithmetik mit einer relativen Maschinengenauigkeit von $macheps = 2.2\,\text{E-}16$ durchgeführt. Es wurden die Abbruchkriterien aus Tabelle 2.1 mit $\varepsilon_0^t = \varepsilon_1^t = \varepsilon_2^t = \varepsilon_5^t = 1.0\,\text{E-}10$, $\varepsilon_3^t = 1.0\,\text{E-}06$ und $\varepsilon_4^t = 1.0\,\text{E-}03$ verwendet. Man beachte, daß diese ziemlich harten Kriterien ausschließlich für Testzwecke verwendet wurden. In realen Anwendungen würde man etwa $\varepsilon_3^t = 10^{-2} \ldots 10^{-3}$ wählen. Zur Sicherung der Anordnungsbedingungen der freien

[1] `http://www.math.tu-dresden.de/~schuetze/free.zip`

Knoten haben wir das relative Distanzmaß $\epsilon = 0.0625$ verwendet, das ebenfalls von de Boor/Rice benutzt wurde.

1	$\left\|\mathbf{F}^{\nu+1}\right\| \leq \varepsilon_0^t$
2	$\left\|\mathbf{J}^{\nu T}\mathbf{F}^{\nu}\right\| \leq \varepsilon_1^t$
3	$\left\|\mathbf{F}^{\nu T}\mathbf{J}^{\nu}\mathbf{s}^{\nu}\right\| \leq \varepsilon_2^t$
4	$\left\|\mathbf{t}^{\nu+1} - \mathbf{t}^{\nu}\right\| \leq \varepsilon_3^t \left(\left\|\mathbf{t}^{\nu}\right\| + \varepsilon_4^t\right)$
5	$\left\|\left\|\mathbf{F}^{\nu+1}\right\| - \left\|\mathbf{F}^{\nu}\right\|\right\| \leq \varepsilon_5^t \left\|\mathbf{F}^{\nu}\right\|$
6	$\nu > \nu max$
7	failed otherwise

Tabelle 2.1: Rückgabewerte und Abbruchtests

Zum Vergleich des entwickelten Verfahrens mit Standardoptimierungsverfahren führen wir noch eine Minimierung des reduzierten Funktionals mit der MATLAB-Routine CONSTR durch. Die Abbruchkonstanten wurden dabei auf den vorgegebenen Standardwerten belassen.

2.6.1 Titanium Heat Data

Eines der am besten untersuchten Beispiele innerhalb der Approximation durch Splines mit freien Knoten sind die sog. *Titanium Heat Data*, siehe [dBR68], [dB78] und [Jup78]. Diese realen Daten beschreiben eine Eigenschaft von Titan in Abhängigkeit von der Temperatur. Sie sind mit einem nicht zu vernachlässigenden stochastischen Fehler behaftet. Die $m = 49$ Meßpunkte sind äquidistant im Intervall $[595, 1075]$ verteilt. Die Meßwerte verlaufen relativ flach bis auf den typischen Peak im Bereich $x = 900$.

Um einen direkten Vergleich mit den Resultaten von [dBR68] und [Jup78] zu ermöglichen, wollen wir diese Daten durch $n = 9$ B-Splines der Ordnung $k = 4$ *approximieren*, d. h. wir benutzen keine Glättung. Alle $l = 5$ innere Knoten werden als frei betrachtet. Der Optimierungsprozeß wird mit den folgenden drei Knotenfolgen, welche auch von Jupp verwendet wurden, gestartet:

$\mathbf{t}_1$ Startpunkt nahe der inneren Optimalstelle,
$\mathbf{t}_2$ Startpunkt von [dBR68],
$\mathbf{t}_3$ äquidistante innere Knoten.

Jupp [Jup78] erwähnt, daß es vier stationäre Punkte zu diesem Problem gibt, welche dort wahrscheinlich in einfacher Arithmetik gefunden wurden, siehe Tabelle 2.2.

Tabelle 2.3 zeigt einen Vergleich der Routinen. Man erkennt, daß sowohl unser Algorithmus als auch die MATLAB-Routine ausgehend von den Startpunkten $\mathbf{t}_1$ und $\mathbf{t}_2$ das globale Optimum im Inneren des zulässigen Bereiches finden, obwohl die MATLAB-Routine im zweiten Fall die gewünschte Genauigkeit nicht erreicht und vorzeitig abbricht. Der neuentwickelte Algorithmus benötigt jedoch stets nur einen Bruchteil der Zeit. Die Abbildungen 2.1 und

	$\mathbf{t}^*$ innere Optimalstelle	$\mathbf{t}^1$ lokales Minimum	$\mathbf{t}^2$ Sattelpunkt	$\mathbf{t}^3$ lokales Minimum
τ_5	835.9670	803.5453	796.9648	836.2518
τ_6	876.4016	866.2870	852.4528	878.7197
τ_7	898.1462	866.3510	885.2259	890.6851
τ_8	916.3146	901.2944	885.9946	905.0262
τ_9	973.9075	905.5165	910.5168	905.0902
$\|\mathbf{F}\|$	8.752539 E-02	2.440692 E-01	2.493385 E-01	2.510164 E-01

Tabelle 2.2: Titanium Heat Data: Stationäre Punkte

2.2 zeigen die Splines zu der Startknotenfolge $\mathbf{t}_2$ und zur optimalen Knotenfolge $\mathbf{t}^{*2}$.

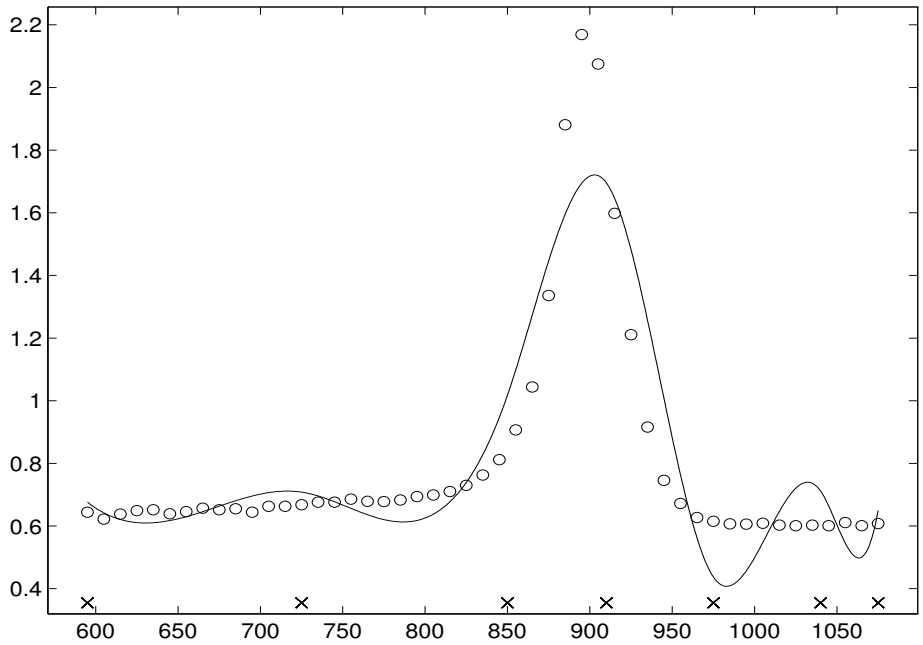


Abbildung 2.1: Titanium Heat Data: Startknotenfolge $\mathbf{t}_2$

Die Lage der Daten läßt bereits vermuten, daß eine äquidistante Knotenverteilung in diesem Beispiel eine schlechte Approximation liefert. Dies wird durch die numerischen Tests bestätigt. MATLAB liefert eine Lösung, welche jedoch weit entfernt von einem der stationären Punkte dieses Problems ist. Auch das Golub/Pereyra-Modell bricht in diesem Beispiel wegen zu kleiner Schritte ab. Einzig das Kaufman-Modell liefert eine Lösung, welche sich in der Nähe des lokalen Minimums $\mathbf{t}^1$ befindet.

2.6.2 Ein Algorithmus zur Datenreduktion

Ein Motiv für die Entwicklung des vorgestellten Algorithmus war eine optimale Plazierung der Knoten, um gegebene Daten mit möglichst wenig B-Splines zu approximieren und eine

[2]Die Splineknoten τ_j werden in den folgenden Abbildungen durch ein Kreuz $\times$, die Meßdaten (x_i, y_i) durch einen kleinen Kreis $\circ$ dargestellt.

	$\mathbf{t}_1$	MATLAB	RAP-Ka-ED	RAP-GP-OD
τ_5	838.2	835.467	835.457	835.457
τ_6	876.6	876.464	876.506	876.506
τ_7	895.8	898.282	898.166	898.166
τ_8	915.0	916.140	916.280	916.280
τ_9	979.0	974.384	974.017	974.017
$\|\mathbf{F}\|$	1.011427 E-01	8.748693 E-02	8.748003 E-02	8.748003 E-02
Schritte		469 funct. calls	10	10
Zeit [s]		7.58	0.201	0.259
$\|\mathbf{F}^T\mathbf{Js}\|$			9.389011 E-11	9.411498 E-11
$\|\mathbf{J}^T\mathbf{F}\|$			1.038693 E-07	1.041373 E-07
Ret. Code		successfully	3	3

	$\mathbf{t}_2$	MATLAB	RAP-Ka-ED	RAP-GP-OD
τ_5	725	835.462	835.457	835.457
τ_6	850	876.520	876.506	876.506
τ_7	910	898.150	898.167	898.166
τ_8	975	916.285	916.280	916.280
τ_9	1040	974.018	974.017	974.017
$\|\mathbf{F}\|$	1.008965 E+00	8.748019 E-02	8.748003 E-02	8.748003 E-02
Schritte		504 funct. calls	16	13
Zeit [s]		8.43	0.319	0.332
$\|\mathbf{F}^T\mathbf{Js}\|$			3.045682 E-11	5.660611 E-11
$\|\mathbf{J}^T\mathbf{F}\|$			6.548293 E-08	8.749910 E-08
Ret. Code		terminated	3	3

	$\mathbf{t}_3$	MATLAB	RAP-Ka-ED	RAP-GP-OD
τ_5	675	689.382	803.307	610.104
τ_6	755	839.776	864.915	836.667
τ_7	835	877.214	869.022	877.900
τ_8	915	897.039	897.214	897.006
τ_9	995	908.162	908.326	908.131
$\|\mathbf{F}\|$	1.235202 E+00	2.515806 E-01	2.450116 E-01	2.525756 E-01
Schritte		387 funct. calls	11	12
Zeit [s]		6.45	0.227	0.308
$\|\mathbf{F}^T\mathbf{Js}\|$			2.014991 E-11	1.571338 E-11
$\|\mathbf{J}^T\mathbf{F}\|$			1.060476 E-04	9.329192 E-05
Ret. Code		successfully	3	4

Tabelle 2.3: Titanium Heat Data: Vergleich des Algorithmus mit MATLAB-Routine und drei verschiedenen Startpunkten

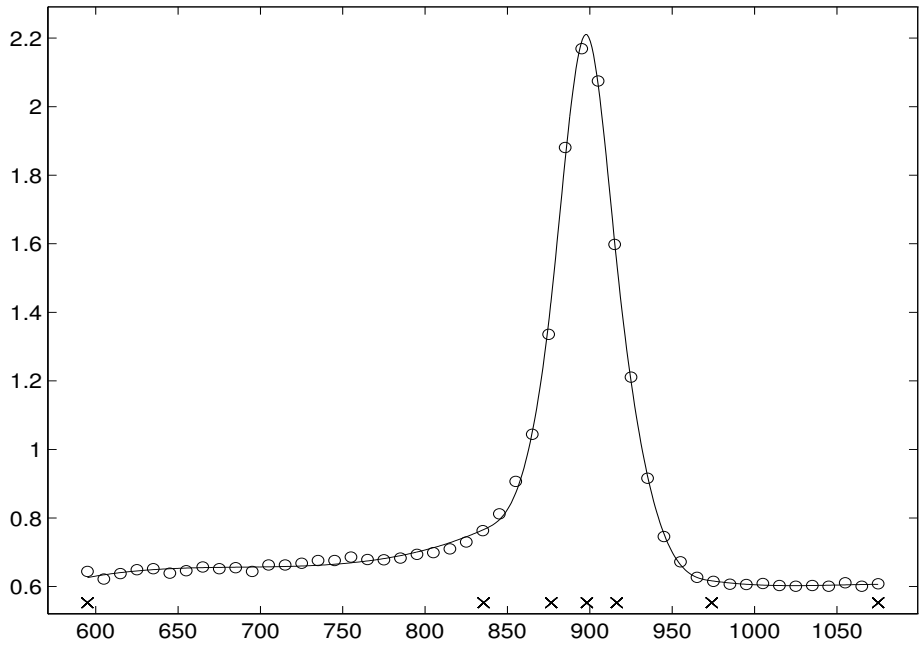


Abbildung 2.2: Titanium Heat Data: Optimalstelle $\mathbf{t}^*$

maximale Datenreduktion zu erreichen. Wir werden daher jetzt das neue Verfahren mit einem Algorithmus zur Knotenreduktion kombinieren.

Sei $\Delta > 0$ eine gegebene Schätzung des Fehlerniveaus der Daten, welches durch den Quadratmittelfehler $\|\mathbf{F}\|$ gemessen wird. Wir nennen eine Knotenfolge $\mathbf{t}$ *akzeptabel*, wenn der Quadratmittelfehler des zugehörigen Splines nicht größer als die vorgegebene Schranke Δ ist, d. h. falls $\|\mathbf{F}\| \leq \Delta$. Unsere Knotenreduktionsstrategie ist ähnlich der von Lyche/Mørken [LM88]. Wir starten den Prozeß mit einer akzeptablen Knotenfolge, welche aus einer relativ großen Anzahl von Knoten besteht. Im Gegensatz zu Lyche/Mørken müssen diese Knoten jedoch nicht an Datenstellen lokalisiert sein. Ein Schritt des iterativen Prozesses zur Knotenreduktion besteht zunächst in der Optimierung der Lage der Knoten mittels unseres Algorithmus. Danach bewerten wir die Knoten nach ihrer Bedeutung für die Approximationsgüte, wir verwenden die Größe des Sprungs in der ersten unstetigen Ableitung als Gütemaß. Falls das Entfernen des Knotens mit dem kleinsten Sprung zu einer inakzeptablen Knotenfolge führt, wird die Iteration beendet und die letzte, akzeptable Knotenfolge wiederhergestellt. Andernfalls wird die Iteration mit der neuen, um einen Knoten verkürzten Knotenfolge fortgesetzt.

Der Unterschied zu anderen bekannten Verfahren zur Knotenreduktion besteht darin, daß wir versuchen, in jedem Schritt eine optimale Knotenverteilung zu finden, während in anderen Verfahren überhaupt keine Lageoptimierung stattfindet. Da die Knotenoptimierung in *jedem* Schritt jedoch sehr teuer ist, haben wir eine zweistufige Strategie implementiert: In einer ersten Stufe beginnen wir mit einer festen, akzeptablen Knotenfolge. Wir entfernen iterativ Knoten wie oben beschrieben, jedoch ohne die Lage der Knoten zu optimieren. Dieser billige Prozeß wird solange ausgeführt, wie die Knoten akzeptabel bleiben. Geht die Akzeptabilität verloren, so beginnen wir die zweite Stufe und versuchen, weiter Knoten zu entfernen, indem wir die Lage der Knoten in jedem Schritt optimieren.

2.6.3 Ein Beispiel von Hu

Wir möchten diese zweistufige Strategie anhand eines Beispiels illustrieren, welches wir Hu [Hu93] entnommen haben. Wir berechnen die rationale Funktion g mit $g(x) = 10x/(1 + 100x^2)$ an $m = 90$ äquidistanten Punkten innerhalb des Intervalls $[-2, 2]$. Im Gegensatz zu Hu, welcher exakte Daten betrachtet und die $\|.\|_\infty$-Norm verwendet, betrachten wir verrauschte Daten, indem wir mittels Pseudozufallszahlen $\{\epsilon_i\}$ mit $-0.05 \leq \epsilon_i \leq 0.05$ die gestörten Meßwerte $y_i = g(x_i) + \epsilon_i$, $i = 1, \ldots, m$ erzeugen.

Wir wollen diese Daten durch einen glättenden Spline der Ordnung $k = 5$ mit einer Ordnung $r = 2$ im Glättungsterm approximieren. Als Glättungsparameter μ benutzen wir $\mu = 1.0\,\text{E-}10$, die vorgegebene Schranke für den Quadratmittelfehler $\|\mathbf{F}\|$ sei $\Delta = 0.3$.

Die Approximation wird mit $l = 15$ äquidistanten inneren Knoten gestartet, welche jedoch nicht akzeptabel sind. Daher führen wir zu Beginn eine vorgeschaltete Optimierung der Knoten durch und erhalten die Approximation in Abbildung 2.3. Diese Knotenfolge wird als Ausgangspunkt für die erste Stufe des Knotenreduktionsalgorithmus benutzt. Wir erhalten einen Spline mit $l = 12$ inneren Knoten. Der Wechsel zur zweiten Stufe ergibt eine weitere Reduktion auf $l = 5$ innere Knoten. Abbildung 2.4 zeigt die abschließende Knotenfolge und den zugehörigen Spline. Die Ergebnisse des Knotenreduktionsalgorithmus sind in Tabelle 2.4 zusammengefaßt. Bei der Optimierung der Knoten haben wir das Modell RSP-GP-OD benutzt.

äquidistante innere Knoten	$n = 20$	$\|\mathbf{F}\|$	$= 5.3641288\,\text{E-}01$
	$l = 15$	$\|\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}\|$	$= 5.3641269\,\text{E-}01$
optimierte Knotenfolge	$n = 20$	$\|\mathbf{F}\|$	$= 2.2395534\,\text{E-}01$
(Abb. 2.3)	$l = 15$	$\|\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}\|$	$= 2.2395367\,\text{E-}01$
Knotenreduktionsalgorithmus, Stufe I	$n = 17$	$\|\mathbf{F}\|$	$= 2.6486235\,\text{E-}01$
	$l = 12$	$\|\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}\|$	$= 2.6486158\,\text{E-}01$
Knotenreduktionsalgorithmus, Stufe II	$n = 10$	$\|\mathbf{F}\|$	$= 2.7195137\,\text{E-}01$
(Abb. 2.4)	$l = 5$	$\|\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}\|$	$= 2.7194998\,\text{E-}01$

Tabelle 2.4: Beispiel von Hu: Knotenreduktionsalgorithmus

Die gleiche Prozedur haben wir mit dem Kaufman-Modell RSP-Ka-ED durchgeführt. Hier führt der vorgeschaltete Optimierungsalgorithmus zu einer Knotenfolge, dessen Quadratmittelfehler etwas größer ist, nämlich $\|\mathbf{F}\| = 2.2486127\,\text{E-}01$. Trotzdem wurde die Anzahl der inneren Knoten in der ersten Stufe auf $l = 9$ reduziert, der Fehler beträgt $\|\mathbf{F}\| = 2.7767981\,\text{E-}01$. In der zweiten Stufe erreichen wir ebenfalls $l = 5$ innere Knoten und den gleichen Fehler wie mit dem Modell RSP-GP-OD. Abschließend sei bemerkt, daß die MATLAB-Routine CONSTR bei diesem Beispiel abbricht, da die Anordnungsnebenbedingungen zwischenzeitlich nicht erfüllt sind. Es zeigt sich insbesondere, daß die MATLAB-Routine relativ empfindlich bei der Verwendung des absoluten Distanzmaßes reagierte, während sich unser Algorithmus sehr robust verhielt. Dies zeigt, daß wir tatsächlich einen Algorithmus benötigen, welcher nur mit zulässigen Punkten arbeitet.

Das letzte Beispiel zeigt einen wesentlichen Vorteil der kombinierten Knotenreduktions- und Optimierungsstrategie: Der erhaltene Spline reproduziert nicht nur die Daten innerhalb des Fehlerniveaus mit $l = 5$ an Stelle von $l = 15$ inneren Knoten, sondern er vermeidet auch das sog. „Overfitting“ durch eine zu große Anzahl von Parametern. Die reduzierte Anzahl von Freiheitsgraden bewirkt eine Art Regularisierung durch Dimensionsreduktion, siehe

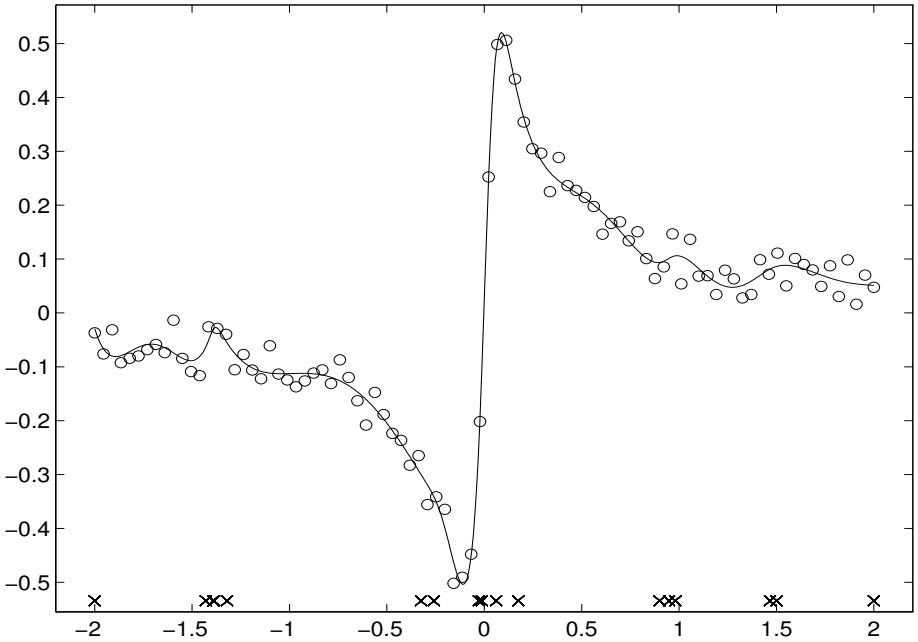


Abbildung 2.3: Beispiel von Hu: Vorgeschalteter Optimierungsschritt ausgehend von äquidistanten inneren Knoten, $n = 20$, $l = 15$

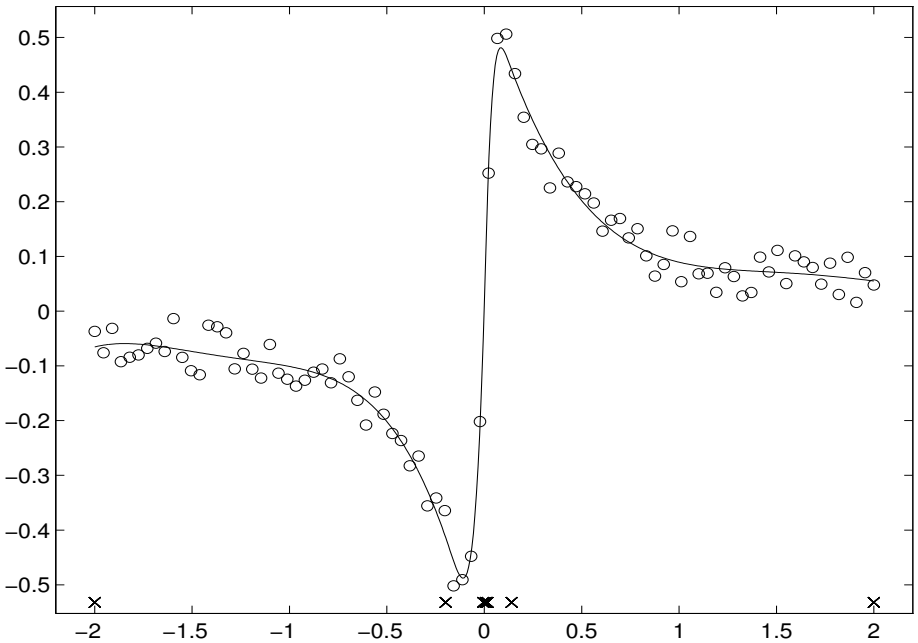


Abbildung 2.4: Beispiel von Hu: Knotenreduktionsalgorithmus, Ergebnis von Stufe II, $n = 10$, $l = 5$

[Var82] und [Wah90] für eine Diskussion und Ausnutzung dieses Effekts.

Die wenigen angeführten Beispiele machen bereits deutlich, daß der Algorithmus zur Knotenoptimierung verbunden mit einer Strategie zur Reduktion der Anzahl der Knoten ein effektives Werkzeug zur Approximation von fehlerbehafteten Daten durch Splines mit freien Knoten ist. Insbesondere bei Benutzung des regularisierenden Glättungsterms verhält sich der Algorithmus sehr robust, da die Durchführbarkeit unabhängig von der Lage der freien Knoten gesichert ist.

Kapitel 3

Univariate Splines mit Ungleichheitsnebenbedingungen an Ableitungen

3.1 Einleitung

In diesem Kapitel untersuchen wir erneut die Approximation fehlerbehafteter Meßwerte $y_i = g(x_i) + \epsilon_i$ $(i = 1, \ldots, m)$ einer unbekannten glatten Funktion $g \in W_2^q[a, b]$. Diese Meßwerte wollen wir in bewährter Weise durch einen Spline $s \in \mathcal{S}_{k,\boldsymbol{\tau}}$ approximieren, dessen Parameter durch Minimierung des Schoenberg-Funktionals

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

bestimmt werden.

In Erweiterung der Problemstellung aus Kapitel 2 seien jetzt zusätzliche Informationen über die Form der Funktion g bekannt, z.B. $g^{(p)}(x) \geq 0$ für alle $x \in [a, b]$ mit einer vorgegebenen Ableitungsordnung $p \in \{0, \ldots, q\}$. Später werden wir noch allgemeinere Nebenbedingungen an Ableitungen zulassen. Es sollen sich also bestimmte geometrische Eigenschaften der Funktion g auf die Approximation s übertragen.

Die formerhaltende Approximation ist von großer praktischer Bedeutung: In einigen Anwendungen führt erst die Einhaltung bestimmter Nebenbedingungen zu physikalisch oder technisch sinnvollen Lösungen (nichtnegative Drücke, monoton wachsende Konzentrationen bei einem chemischen Prozeß usw.), in anderen Anwendungen können Restriktionen an Ableitungen zur Vermeidung unerwünschter Oszillationen benutzt werden (obere Schranken für die Krümmung). In den letzten zwei Jahrzehnten erschien eine sehr große Anzahl von Arbeiten zur formerhaltenden Interpolation und -approximation, siehe etwa die Überblicksarbeiten [Gre91], [Utr91] und die Monographie [Spä95]. Basierend auf einem Variationszugang untersuchen Micchelli/Utreras [MU88], [MU91] Existenz und Eindeutigkeit von Splineinterpolation und -approximation in einer konvexen Teilmenge eines Hilbertraums. Elfving/Andersson [EA88] betrachten den Fall $r = 2$ und Konvexitätsnebenbedingungen $s''(x) \geq \delta(x)$, δ gegeben.

Während diese Autoren den Variationszugang bevorzugen, wird in den folgenden Arbeiten direkt von Splines als Ansatzfunktionen ausgegangen: Dierckx [Die80] betrachtet

konvexe kubische Splines und löst das zugrundeliegende Optimierungsproblem mit der simplexartigen Theil/van de Panne-Prozedur. Cox/Jones [CJ87] untersuchen den Fall der formerhaltenden l_1-Approximation, während Schmidt [Sch90] die Glättung durch monotone quadratische Splines betrachtet. In Schmidt/Scholz [SS90] wird ein effizientes Verfahren zur Glättung ($r = 2$) mit kubischen Splines unter den verallgemeinerten Konvexitätsnebenbedingungen $\delta(x) \leq s''(x) \leq \epsilon(x)$, ($\delta, \epsilon$ - lineare $\mathbb{C}^0$-Splines) entwickelt, bei der das unrestringierte duale Problem an Stelle des partiell separablen primalen Problems gelöst wird[1]. In Schwetlick/Kunert [SK93] wird das allgemeine Problem $\delta(x) \leq s^{(p)}(x) \leq \epsilon(x)$ (δ, ϵ - stückweise konstante $\mathbb{C}^0$-Splines; $p, r \in \{0, \ldots, q\}$) mittels Orthogonalisierungstechniken gelöst. Es sei bemerkt, daß bei den dualen Zugängen eine möglichst genaue numerische Lösung des dualen Programms erforderlich ist, da ein vorzeitiger Abbruch eine primal unzulässige Lösung liefert. Bei dem Problem der formerhaltenden Approximation kommt jedoch der Einhaltung der Nebenbedingungen eine vergleichsweise hohe Bedeutung zu, während die genaue Form des Glättungsfunktionals oft nicht a priori gegeben ist.

In diesem Kapitel wollen wir daher die formerhaltende Approximation mit der Quadratmittelapproximation durch Splines mit freien Knoten verbinden. Wir beziehen erneut eine Teilfolge der Knoten, die sog. *freien Knoten*, in den Optimierungsprozeß ein. Dies resultiert in einem nichtlinearen Quadratmittelproblem in den Koeffizienten und den freien Knoten mit linearen Ungleichheitsnebenbedingungen an die freien Knoten (Anordnungsnebenbedingungen) sowie *nichtlinearen* Ungleichheitsnebenbedingungen an die Koeffizienten und Knoten (formerhaltende Nebenbedingungen). Letztere Nebenbedingungen sind linear in den Koeffizienten, wenn man die Splineknoten festhält. Das Originalproblem ist demzufolge ein Spezialfall sog. restringierter semi-linearer Quadratmittelprobleme (constrained semi-linear least squares problems), einer Verallgemeinerung der wohlbekannten separablen Quadratmittelprobleme, vgl. Abschnitt 2.3. Unter Benutzung von Ergebnissen aus der PhD-Thesis von Parks [Par85] über solche speziellen Optimierungsprobleme leiten wir ein reduziertes Problem her, in welchem wiederum nur die freien Knoten auftreten. Wir untersuchen, unter welchen Bedingungen im Rahmen der Splineglättung Originalproblem und reduziertes Problem äquivalent sind.

Das reduzierte Problem wollen wir erneut durch ein verallgemeinertes Gauß-Newton-Verfahren lösen. Da die Struktur der Jacobi-Matrix im restringierten Fall sehr kompliziert ist, neben den Ableitungen nach den B-Splines gehen Ableitungen nach den Nebenbedingungen und verschiedene Projektoren ein, verallgemeinern wir die Ideen von Kaufman [Kau75] für separable Quadratmittelprobleme auf den restringierten Fall. Wir verwenden eine billiger zu berechnende Approximation an die Jacobi-Matrix, die Kaufman-Approximation, und untersuchen sowohl den qualitativen als auch quantitativen Einfluß dieser Approximation. Nach jüngsten Aussagen der Autorin T. A. Parks und eigenen Erkenntnissen wurden damit die theoretischen Ergebnisse aus [Par85] erstmals praktisch umgesetzt und die Verwendung von exakter Jacobi-Matrix und Kaufman-Approximation numerisch getestet.

Wir entwickeln also in diesem Kapitel ein Verfahren, welches zu gegebener Anzahl und Anfangsposition der Knoten eine optimale Plazierung der Knoten in Abhängigkeit von den Daten $\{x_i, y_i\}$ und den gegebenen formerhaltenden Nebenbedingungen sucht. Da das Problem nichtkonvex ist, kann Eindeutigkeit i. allg. nicht garantiert werden.

Im Gegensatz zu Kapitel 2 sind uns bei der *formerhaltenden* Approximation durch Spli-

[1] Für einen Überblick über partiell separable Optimierungsprobleme und ihre Anwendung bei der restringierten Splineapproximation siehe [Sch92a].

nes mit freien Knoten keine existierenden Verfahren bekannt, welche *direkt* das Schoenberg-Funktional oder den Quadratmittelfehler als Funktion der freien Knoten minimieren. Es existiert lediglich das „heuristische“ Verfahren CONCON aus der FITPACK-Bibliothek [Die87], bei welchem durch schrittweises Einfügen von Knoten eine „optimale“ Plazierung der Knoten erreicht wird. Bei den numerischen Tests werden wir unser Verfahren zur Knotenbestimmung mit dieser adaptiven Strategie vergleichen.

Betrachtet man Algorithmen zur Knotenreduktion bei restringierten Splines, so stehen etwa Verfahren von Arge et.al. [ADLM90] und Schumaker/Stanley [SS96] zur Auswahl, letzteres allerdings nur für quadratische Splines.

3.2 Problemformulierung

3.2.1 Glättungsfunktional und Anordnungsnebenbedingungen

Nach den Vorarbeiten im letzten Kapitel können wir das Glättungsfunktional und die Anordnungsnebenbedingungen an die freien Knoten unmittelbar formulieren.

Für die Splineknoten $\boldsymbol{\tau}$ gelte $\tau_j < \tau_{j+k-q}$, $j = q+1, \ldots, n$ und $r \in \{0, \ldots, q\}$. Wir betrachten das Optimierungsproblem

$$\mathfrak{f}(\boldsymbol{\alpha}, \mathbf{t}) := \frac{1}{2} \left\| \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} - \begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}) \end{bmatrix} \boldsymbol{\alpha} \right\|^2 \to \min_{\boldsymbol{\alpha}, \mathbf{t}} \tag{3.1}$$

unter den Anordnungsnebenbedingungen an die freien Knoten

$$\mathbf{C}\mathbf{t} - \mathbf{h} \geq \mathbf{0} \tag{3.2}$$

und weiteren Nebenbedingungen an Ableitungen des Splines.

3.2.2 Nebenbedingungen an Ableitungen

Eine in der Einleitung formulierte Forderung an die Funktion s war die Formerhaltung. Zu diesem Zweck drücken wir jetzt Nebenbedingungen an Ableitungen des Splines mittels der Splineknoten und der Splinekoeffizienten aus.

Die Knotenfolge $\boldsymbol{\tau}$ erfülle die Bedingung (2.7). Es gelte $p \in \{0, \ldots, q\}$. Dann existiert $s^{(p)}$ und besitzt die Matrixdarstellung

$$s^{(p)}(x) = \boldsymbol{\beta}_p^T(x, \boldsymbol{\tau})\mathbf{D}_p(\boldsymbol{\tau})\boldsymbol{\alpha} = \boldsymbol{\beta}_p^T(x, \mathbf{t})\mathbf{D}_p(\mathbf{t})\boldsymbol{\alpha}.$$

Damit sind wir in der Lage, die Nebenbedingung, welche den „shape“ des Splines charakterisiert, zu formulieren. Wir betrachten zunächst die Positivitätsforderung an $s^{(p)}$ im ganzen Intervall $[a, b]$.

Positivitätsforderung an die Ableitung des Splines

$$\begin{aligned} s^{(p)}(x) &\geq 0 \\ \boldsymbol{\beta}_p^T(x, \mathbf{t})\mathbf{D}_p(\mathbf{t})\boldsymbol{\alpha} &\geq 0 \qquad \forall x \in [t_k, t_{n+1}] \end{aligned} \tag{3.3}$$

Die Nebenbedingung (3.3) an die p-te Ableitung des Splines ergibt zusammen mit der Anordnungsbedingung (3.2) für die freien Knoten und dem zu minimierenden Funktional (3.1) ein semiinfinites Optimierungsproblem.

Zumindest für kleine $k-p$ lassen sich – in der Regel nichtlineare – notwendige und hinreichende Bedingungen für die Positivität der Polynome der Ordnung $k-p$ in den Teilintervallen angeben. Damit läßt sich in diesen Fällen das semiinfinite direkt in ein finites Optimierungsproblem überführen.

Um die Lösung des semiinfiniten Optimierungsproblems zu vermeiden, verwenden wir statt der notwendigen und hinreichenden Bedingung (3.3) die – wegen der Nichtnegativität der B-Splines $\boldsymbol{\beta}_p(x,\mathbf{t})$ – hinreichende Bedingung $\boldsymbol{\alpha}^{(p)} = \mathbf{D}_p(\mathbf{t})\boldsymbol{\alpha} \geq \mathbf{0}$. Wir erhalten daher die für die Positivität hinreichende Nebenbedingung

$$\mathbf{D}_p(\mathbf{t})\boldsymbol{\alpha} \geq 0. \tag{3.4}$$

Wir betrachten jetzt eine feste Knotenfolge $\boldsymbol{\tau}$. Da das Funktional (3.1) unter der Voraussetzung $m \geq r$ und $\mu > 0$ streng konvex ist, sind sowohl das semiinfinite Problem (3.3) als auch das finite Problem (3.4) eindeutig lösbar. Die Optimallösung $\boldsymbol{\alpha}^{\mathrm{f}}$ des finiten Problems ist wegen $B_{j,k-p,\boldsymbol{\tau}}(x) \geq 0$ zulässig für das semiinfinite Problem. Falls $k-p \leq 2$, so stimmen die Lösungen der beiden Probleme überein, $\boldsymbol{\alpha}^{\mathrm{f}} = \boldsymbol{\alpha}^{\mathrm{if}}$. Allgemein kann man feststellen, daß $\boldsymbol{\alpha}^{\mathrm{if}}$ durch $\boldsymbol{\alpha}^{\mathrm{f}}$ desto besser approximiert wird, je feiner die Knotenfolge $\boldsymbol{\tau}$ ist.

Hinreichende Nebenbedingungen der obigen Form werden von einer Reihe von Autoren verwendet, z.B. untersuchen Cox und Jones [CJ87] das Problem $\min\{\|\mathbf{y}-\mathbf{B}\boldsymbol{\alpha}\|_1 : \boldsymbol{\alpha}^{(p)} = \mathbf{D}_p\boldsymbol{\alpha} \geq 0, \boldsymbol{\alpha} \in \mathbb{R}^n\}$. Andere Autoren, z.B. [WD95], bevorzugen notwendige und hinreichende Bedingungen, haben dann allerdings numerische Schwierigkeiten, da selbst die Probleme zu festen Knoten nichtlinear werden. Unsere numerischen Erfahrungen zeigen, daß die hinreichenden Bedingungen bei sorgfältiger Wahl der Splineknoten durchaus gute Ergebnisse liefern, d. h. daß sie nicht zu einschränkend sind. Im Zweifelsfall können die mit hinreichenden Nebenbedingungen berechneten Splines als sehr guter Startpunkt für Algorithmen mit notwendigen und hinreichenden Bedingungen genutzt werden.

Man beachte, daß die Form der hinreichenden Nebenbedingungen für das weitere Vorgehen wesentlich ist, genauer gesagt, benötigen wir Nebenbedingungen, welche bei festen Knoten $\boldsymbol{\tau}$ linear in den Splinekoeffizienten $\boldsymbol{\alpha}$ sind.

Einfache Schranken für die Ableitung des Splines

Oftmals wird neben der unteren Schranke Null an $s^{(p)}$ gleichzeitig eine obere Schranke vorgegeben, z.B. um eine allzu große Krümmung des Splines zu verhindern, oder es werden auf verschiedenen Intervallen verschiedene „shape constraints“ vorgegeben, z.B. konvex-konkaver Datenabgleich (siehe [SS90]).

Allgemeiner fordern wir daher nun

$$l_i^{(p)} \leq s^{(p)}(x) \leq u_i^{(p)} \qquad \forall x \in [t_i, t_{i+1}), \quad i = k, \ldots, n \tag{3.5}$$

mit $2(n-k+1)$ Konstanten

$$\mathbf{l} := \left(l_k^{(p)}, \ldots, l_n^{(p)}\right)^T \in \mathbb{R}^{n-k+1}, \mathbf{u} := \left(u_k^{(p)}, \ldots, u_n^{(p)}\right)^T \in \mathbb{R}^{n-k+1}.$$

Da die B-Splines eine nichtnegative Partition der Eins bilden, gilt

$$\min\left\{\alpha_j^{(p)} \,:\, j \in K_i\right\} \leq s^{(p)}(x) = \sum_{j \in K_i} B_{j,k-p}(x)\alpha_j^{(p)} \leq \max\left\{\alpha_j^{(p)} \,:\, j \in K_i\right\}$$

für $x \in [t_i, t_{i+1})$ und $K_i := \{i-k+p+1, \ldots, i\}$. Eine hinreichende Bedingung für (3.5) ist deshalb

$$l_i^{(p)} \le \min\left\{\alpha_j^{(p)} \,:\, j \in K_i\right\} \text{ und } \max\left\{\alpha_j^{(p)} \,:\, j \in K_i\right\} \le u_i^{(p)} \quad i = k, \ldots, n.$$

Dies können wir äquivalent folgendermaßen formulieren

$$L_j^{(p)} \le \alpha_j^{(p)} \le U_j^{(p)} \qquad j = p+1, \ldots, n$$

mit den $2(n-p)$ Konstanten

$$L_j^{(p)} := \max\left\{l_i^{(p)} \,:\, i \in W_j\right\}, U_j^{(p)} := \min\left\{u_i^{(p)} \,:\, i \in W_j\right\}$$

und der Indexmenge $W_j := \left\{\max\{j,k\}, \ldots, \min\{j+k-p-1, n\}\right\}$ bzw. in Matrixform

$$\mathbf{L} \le \boldsymbol{\alpha}^{(p)} \le \mathbf{U}$$

$$\mathbf{L} := \left(L_{p+1}^{(p)}, \ldots, L_n^{(p)}\right)^T \in \mathbb{R}^{n-p}, \mathbf{U} := \left(U_{p+1}^{(p)}, \ldots, U_n^{(p)}\right)^T \in \mathbb{R}^{n-p}.$$

Eine hinreichende Bedingung für (3.5) ist also

$$\mathbf{L} \le \mathbf{D}_p(\mathbf{t})\boldsymbol{\alpha} \le \mathbf{U}. \tag{3.6}$$

In den weiteren Ausführungen seien $-\infty$ und $+\infty$ formal als untere bzw. obere Schranken zugelassen. Damit ergibt sich die praktisch wichtige Positivitätsforderung an $s^{(p)}$ als Spezialfall der einfachen Schranken. Die verwendeten Algorithmen sind in der Lage, diese Fälle ebenfalls zu behandeln.

Noch allgemeinere Nebenbedingungen an Ableitungen – etwa die simultane Forderung nach Konvexität und Monotonie – führen auf ähnlich strukturierte Nebenbedingungen (siehe [Kun95]). Um unsere entwickelten Techniken anwenden zu können, sind von diesen verallgemeinerten Nebenbedingungen gegebenenfalls in einem „preprocessing step“ redundante Nebenbedingungen zu entfernen.

3.2.3 Konsistenz der Nebenbedingungen

Für feste zulässige Knotenfolgen, d. h. $\mathbf{Ct} \ge \mathbf{h}$, ist der zulässige Bereich des Optimierungsproblems (3.1), (3.2), (3.6) genau dann nicht leer, wenn $\mathbf{L} \le \mathbf{U}$.

Definition 3.1 (Konsistenz, strikte Konsistenz)**.** Die Nebenbedingungen

$$l_i^{(p)} \le s^{(p)}(x) \le u_i^{(p)} \qquad \forall x \in [t_i, t_{i+1}), \quad i = k, \ldots, n$$

heißen *konsistent*, falls

$$L_j^{(p)} \le U_j^{(p)} \qquad j = p+1, \ldots, n \qquad (\mathbf{L} \le \mathbf{U}). \tag{3.7}$$

Sie heißen *strikt konsistent*, falls

$$L_j^{(p)} < U_j^{(p)} \qquad j = p+1, \ldots, n \qquad (\mathbf{L} < \mathbf{U}). \tag{3.8}$$

Beispiel 3.1. $k = 4$, $n = 9$, $p = 2$, $l_i^{(2)} \le s''(x) \le u_i^{(2)}\ \forall x \in [t_i, t_{i+1})\ i = 4, \ldots, 9$

$$\begin{array}{l} \mathbf{l} = (0,0,0,0,-\infty,-\infty) \\ \mathbf{u} = (+\infty,+\infty,+\infty,+\infty,-1,-1) \end{array} \Longrightarrow \begin{array}{l} \mathbf{L} = (0,0,0,0,0,-\infty,-\infty) \\ \mathbf{U} = (+\infty,+\infty,+\infty,+\infty,-1,-1,-1) \end{array}$$

Die Konsistenzbedingung $\mathbf{L} \le \mathbf{U}$ ist verletzt ($0 = L_7 > U_7 = -1$), obwohl $\mathbf{l} < \mathbf{u}$.

Beispiel 3.2. $k = 4$, $n = 9$, $p = 1$, $s'(x) \ge 0\ \forall x \in [t_4, t_8)$, $s'(x) \le 0\ \forall x \in [t_9, t_{10})$

$$\begin{array}{l} \mathbf{l} = (0,0,0,0,-\infty,-\infty) \\ \mathbf{u} = (+\infty,+\infty,+\infty,+\infty,+\infty,0) \end{array} \Longrightarrow \begin{array}{l} \mathbf{L} = (0,0,0,0,0,0,-\infty,-\infty) \\ \mathbf{U} = (+\infty,+\infty,+\infty,+\infty,+\infty,0,0,0) \end{array}$$

Die Nebenbedingungen sind konsistent, aber nicht strikt konsistent.

Beispiel 3.3. $k = 4$, $n = 9$, $p = 2$, $s''(x) \ge 0\ \forall x \in [t_4, t_8)$, $s''(x) \le 0\ \forall x \in [t_9, t_{10})$

$$\begin{array}{l} \mathbf{l} = (0,0,0,0,-\infty,-\infty) \\ \mathbf{u} = (+\infty,+\infty,+\infty,+\infty,+\infty,0) \end{array} \Longrightarrow \begin{array}{l} \mathbf{L} = (0,0,0,0,0,-\infty,-\infty) \\ \mathbf{U} = (+\infty,+\infty,+\infty,+\infty,+\infty,0,0) \end{array}$$

Die Nebenbedingungen sind strikt konsistent.

Strikte Konsistenz

In den späteren Anwendungen benötigen wir die strikte Konsistenz der Nebenbedingungen. Im Falle einseitiger Schranken an $s^{(p)}$ auf dem ganzen Intervall ist diese Bedingung trivialerweise erfüllt. Wir untersuchen jetzt den wichtigen Fall von einseitigen Schranken an $s^{(p)}$ auf verschiedenen Teilintervallen. Wir beschränken uns o.B.d.A. auf zwei Teilintervalle mit Schranken an Ableitungen und betrachten die Nebenbedingung

$$s^{(p)}(x) \ge 0 \qquad \text{für alle } x \in [a, t_\iota) \quad \text{und} \quad s^{(p)}(x) \le 0 \qquad \text{für alle } x \in [t_\kappa, b)$$

mit $\iota, \kappa \in \{k+1, \ldots, n\}$ und $\iota \le \kappa$. Die Knoten t_ι und t_κ sollen o.B.d.A. hinreichend weit im Inneren des Intervalls $[t_k, t_{n+1}]$ liegen, d. h. die Indexmengen W_j werden nicht von den Randknoten beeinflußt. Dann gilt:

$$\begin{array}{lll} l_i^{(p)} = 0, & u_i^{(p)} = +\infty, & i = k, \ldots, \iota - 1, \\ l_i^{(p)} = -\infty, & u_i^{(p)} = +\infty, & i = \iota, \ldots, \kappa - 1, \\ l_i^{(p)} = -\infty, & u_i^{(p)} = 0, & i = \kappa, \ldots, n. \end{array}$$

Für die Schranken $L_j^{(p)}$ und $U_j^{(p)}$ für die Splinekoeffizienten $\alpha_j^{(p)}$ $(j = p+1, \ldots, n)$ erhalten wir

$$\begin{array}{lll} L_j^{(p)} = 0, & U_j^{(p)} = +\infty, & j = p+1, \ldots, \iota - k + p, \\ L_j^{(p)} = -\infty, & U_j^{(p)} = 0, & j = \kappa, \ldots, n. \end{array}$$

Die obigen Schranken werden allein durch die Vorgaben $s^{(p)} \ge 0$ auf $[a, t_\iota)$ bzw. $s^{(p)} \le 0$ auf $[t_\kappa, b]$ bestimmt. Für die Schranken $L_j^{(p)}$, $U_j^{(p)}$ $(j = \iota - k + p + 1, \ldots, \kappa - 1)$ beeinflussen sich diese Nebenbedingungen gegenseitig.

a) Untere Schranken $j = \iota - 1$ ist der letzte Index, für welchen gilt $\iota - 1 \in W_j$. Es gilt $W_{\iota-1} = \{\iota - 1, \ldots, \iota + k - p - 2\}$, $L^{(p)}_{\iota-1} = 0$ sowie $W_\iota = \{\iota, \ldots, \iota + k - p - 1\}$, $L^{(p)}_{\iota-1} = -\infty$.

b) Obere Schranken $j = \kappa - k + p + 1$ ist der erste Index, für welchen gilt $\kappa \in W_j$. Es gilt $W_{\kappa-k+p} = \{\kappa - k + p, \ldots, \kappa - 1\}$, $U^{(p)}_{\kappa-k+p} = +\infty$, $W_{\kappa-k+p+1} = \{\kappa - k + p + 1, \ldots, \kappa\}$ und $U^{(p)}_{\kappa-k+p+1} = 0$. Die Nebenbedingungen sind also genau dann strikt konsistent, wenn $\iota - 1 < \kappa - k + p + 1$, d. h. $\kappa - \iota \geq k - p - 1$.

Lemma 3.1 (Strikte Konsistenz der Nebenbedingungen)**.**
Sei $\kappa - \iota \geq k - p - 1$ *und* $\iota \leq \kappa$. *Für die Nebenbedingung*

$$s^{(p)}(x) \geq 0 \qquad \textit{für alle } x \in [a, t_\iota) \quad \textit{und} \quad s^{(p)}(x) \leq 0 \qquad \textit{für alle } x \in [t_\kappa, b)$$

gilt dann

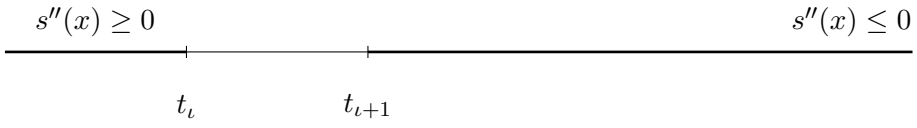
$$\begin{aligned} &L^{(p)}_j = 0, && j = p+1, \ldots, \iota - 1, && U^{(p)}_j = +\infty, && j = p+1, \ldots, \kappa - k + p, \\ &L^{(p)}_j = -\infty, && j = \iota, \ldots, n, && U^{(p)}_j = 0, && j = \kappa - k + p + 1, \ldots, n. \end{aligned}$$

Die strikte Konsistenzbedingung $\mathbf{L} < \mathbf{U}$ *ist erfüllt.*

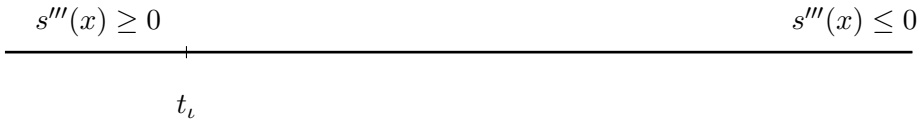
Beispiel 3.4 (Strikte Konsistenz für kubische Splines).
$p = 0$: $\kappa \geq \iota + 3$

$p = 1$: $\kappa \geq \iota + 2$

$p = 2$: $\kappa \geq \iota + 1$



$p = 3$: $\kappa \geq \iota$



3.2.4 Vollständiges restringiertes Glättungsproblem

Abschließend können wir das vollständige Problem der Splineglättung durch Splines mit freien Knoten unter Nebenbedingungen an Ableitungen formulieren.

Definition 3.2 (Vollständiges restringiertes Glättungsproblem)**.** Für die Splineknoten $\boldsymbol{\tau}$ gelte $\tau_j < \tau_{j+k-q}$, $j = q+1, \ldots, n$, und $p, r \in \{0, \ldots, q\}$. Das Optimierungsproblem

$$\mathfrak{f}(\boldsymbol{\alpha}, \mathbf{t}) := \frac{1}{2} \left\| \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} - \begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}) \end{bmatrix} \boldsymbol{\alpha} \right\|^2 \to \min_{\boldsymbol{\alpha}, \mathbf{t}} \tag{3.9}$$

bei

$$\mathbf{C}\mathbf{t} - \mathbf{h} \geq \mathbf{0} \quad \text{und} \quad \mathbf{L} \leq \mathbf{D}_p(\mathbf{t})\boldsymbol{\alpha} \leq \mathbf{U} \tag{3.10}$$

heißt *vollständiges restringiertes Glättungsproblem* (*full constrained smoothing problem*, FCSP).

Die auftretenden Vektoren, Matrizen und Matrixfunktionen haben die folgenden Dimensionen: $\mathbf{y} \in \mathbb{R}^m$; $\boldsymbol{\alpha} \in \mathbb{R}^n$; $\mathbf{t} \in \mathbb{R}^l$; $\mathbf{h} \in \mathbb{R}^{ncstr}$; $\mathbf{L}, \mathbf{U} \in \mathbb{R}^{n-p}$; $\mathbf{C} \in \mathbb{R}^{ncstr,l}$; $\mathbf{B}(.) : \mathbf{t} \in \mathbb{R}^l \to \mathbf{B}(\mathbf{t}) \in \mathbb{R}^{m,n}$; $\mathbf{S}_r(.) : \mathbf{t} \in \mathbb{R}^l \to \mathbf{S}_r(\mathbf{t}) \in \mathbb{R}^{n-r,n}$; $\mathbf{D}_p(.) : \mathbf{t} \in \mathbb{R}^l \to \mathbf{D}_p(\mathbf{t}) \in \mathbb{R}^{n-p,n}$.

3.3 Restringierte semi-lineare Quadratmittelprobleme

3.3.1 Vollständiges und reduziertes Problem

Das Problem FCSP ist ein nichtlineares Quadratmittelproblem, wobei die Variable $\boldsymbol{\alpha}$ stets linear auftritt. In diesem Abschnitt betrachten wir allgemeine Probleme solchen Typs und beginnen mit dem vollständigen Problem.
Vollständiges Problem

$$\mathfrak{f}(\boldsymbol{\alpha}, \mathbf{t}) := \frac{1}{2} \|\mathfrak{F}(\boldsymbol{\alpha}, \mathbf{t})\|^2 = \frac{1}{2} \|\mathbf{y} - \mathbf{B}(\mathbf{t})\boldsymbol{\alpha}\|^2 \longrightarrow \min_{\boldsymbol{\alpha} \in \mathbb{R}^n, \mathbf{t} \in \mathbb{R}^l} \tag{3.11}$$

bei

$$\mathbf{C}\mathbf{t} - \mathbf{h} \geq \mathbf{0} \quad \text{und} \quad \begin{bmatrix} \mathbf{D}_p(\mathbf{t}) \\ -\mathbf{D}_p(\mathbf{t}) \end{bmatrix} \boldsymbol{\alpha} - \begin{pmatrix} \mathbf{L} \\ -\mathbf{U} \end{pmatrix} \geq \mathbf{0}. \tag{3.12}$$

Hierbei sind $\mathbf{B}$ und $\mathbf{D}_p$ beliebige glatte Matrixfunktionen und die verbleibenden Größen $\mathbf{y}$, $\mathbf{h}$, $\mathbf{L}$, $\mathbf{U}$, $\mathbf{C}$ konstante Vektoren und Matrizen.

Wenn die Variable $\mathbf{t}$ im Problem (3.11), (3.12) festgehalten wird, so erhalten wir ein lineares Quadratmittelproblem in $\boldsymbol{\alpha}$, welches wir als *Subproblem (A)* bezeichnen und dessen Lösung $\boldsymbol{\alpha}(\mathbf{t})$ sei. Durch Ersetzen der Variable $\boldsymbol{\alpha}$ im vollständigen Problem durch ihren Optimalwert $\boldsymbol{\alpha}(\mathbf{t})$ erhält man ein *reduziertes Problem* in der Variable $\mathbf{t}$. Diese Reduktionstechnik ist eine Verallgemeinerung der Methode der variablen Projektion von Golub/Pereyra im unrestringierten Fall. Wir nennen Probleme solchen Typs nach Parks [Par85] *restringierte semi-lineare Quadratmittelprobleme* und definieren die folgenden verwandten Optimierungsprobleme:

Reduziertes Problem

$$f(\mathbf{t}) := \frac{1}{2} \|\mathbf{F}(\mathbf{t})\|^2 = \frac{1}{2} \|\mathbf{y} - \mathbf{B}(\mathbf{t})\boldsymbol{\alpha}(\mathbf{t})\|^2 \longrightarrow \min_{\mathbf{t} \in \mathbb{R}^l} \tag{3.13}$$

bei

$$\mathbf{Ct} - \mathbf{h} \geq \mathbf{0}, \tag{3.14}$$

wobei $\mathbf{F}(\mathbf{t}) := \mathbf{y} - \mathbf{B}(\mathbf{t})\boldsymbol{\alpha}(\mathbf{t})$ und $\boldsymbol{\alpha}(\mathbf{t})$ löst

Subproblem (A)

$$\frac{1}{2}\left\|\mathbf{y} - \mathbf{B}(\mathbf{t})\boldsymbol{\alpha}\right\|^2 \longrightarrow \min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \tag{3.15}$$

bei

$$\begin{bmatrix} \mathbf{D}_p(\mathbf{t}) \\ -\mathbf{D}_p(\mathbf{t}) \end{bmatrix} \boldsymbol{\alpha} - \begin{pmatrix} \mathbf{L} \\ -\mathbf{U} \end{pmatrix} \geq \mathbf{0}. \tag{3.16}$$

Es sei bemerkt, daß das vollständige Problem (3.11), (3.12) nicht die allgemeine Form eines restringierten semi-linearen Quadratmittelproblems darstellt, da z.B. die Gleichheitsrestriktionen fehlen. Aus Gründen der Vereinfachung der Notation beschränken wir uns jedoch auf diesen benötigten Spezialfall.

In den nächsten Abschnitten werden wir die Frage behandeln, wann der Übergang vom vollständigen zum reduzierten Problem zulässig ist, und wie Kenntnisse über die Struktur von Subproblem (A) bei der effektiven Lösung des reduzierten Problems angewendet werden können. Die Darstellung im Rest des Abschnittes folgt weitgehend dem Vorgehen in [Par85]. Da die Ergebnisse dieser PhD Thesis nicht allgemein bekannt zu sein scheinen und die dort eingeführten Bezeichnungen im weiteren noch intensiv verwendet werden, geben wir die Hauptergebnisse dieser Arbeit hier an.

3.3.2 Äquivalenz von vollständigem und reduziertem Problem

Das folgende Theorem [Par85, Theorem 4.7] zeigt, unter welchen Voraussetzungen der Übergang vom vollständigem zum reduzierten Problem berechtigt ist, und in welcher Beziehung die Lösungen der beiden Probleme stehen. Es sichert, daß der Wechsel keine kritischen Punkte hinzufügt und daß die Lösung des Originalproblems nicht ausgeschlossen wird.

Theorem 3.1 (Äquivalenz von vollständigem und reduziertem Problem)**.** *1. Die Funktion* $\mathfrak{f}$ *sei zweimal stetig differenzierbar bez.* $\boldsymbol{\alpha}$*, ihr Gradient* $\nabla_{\boldsymbol{\alpha}}\mathfrak{f}$ *sei stetig differenzierbar bez.* $\mathbf{t}$*.*

2. Jede Nebenbedingung sei stetig differenzierbar bez. ihrer Argumente.

3. Für jedes $\mathbf{t}$ *habe das Subproblem (A) eine Lösung* $\boldsymbol{\alpha}(\mathbf{t})$*, so daß*

(a) die hinreichenden Optimalitätsbedingungen zweiter Ordnung für ein lokales Minimum von Subproblem (A) an der Stelle $\boldsymbol{\alpha}(\mathbf{t})$ *gelten,*

(b) die Gradienten bez. $\boldsymbol{\alpha}$ *für die aktiven Nebenbedingungen von Subproblem (A) an der Stelle* $\boldsymbol{\alpha}(\mathbf{t})$ *linear unabhängig sind,*

(c) strikte Komplementarität für Subproblem (A) an der Stelle $\boldsymbol{\alpha}(\mathbf{t})$ *gilt.*

Dann gilt für das vollständige und das reduzierte Problem

(i) Sei $(\boldsymbol{\alpha}^*, \mathbf{t}^*)$ *eine globale Minimumstelle des vollständigen Problems. Dann erfüllt* $\boldsymbol{\alpha}^*$ *die notwendigen Optimalitätsbedingungen erster Ordnung für das Subproblem (A),* $\mathbf{t}^*$ *ist globale Minimumstelle des reduzierten Problems und es gilt* $f(\mathbf{t}^*) = \mathfrak{f}(\boldsymbol{\alpha}^*, \mathbf{t}^*)$*. Wenn es ein eindeutiges* $\boldsymbol{\alpha}^*$ *unter allen Paaren* $(\boldsymbol{\alpha}^*, \mathbf{t}^*)$*, welche* $\mathfrak{f}$ *minimieren und denselben Minimalwert ergeben, gibt, so gilt* $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}(\mathbf{t}^*)$*.*

(ii) Wenn $\mathbf{t}^*$ *die notwendigen Optimalitätsbedingungen erster Ordnung für das reduzierte Problem erfüllt, so erfüllt* $(\boldsymbol{\alpha}(\mathbf{t}^*), \mathbf{t}^*)$ *die notwendigen Optimalitätsbedingungen erster Ordnung für das vollständige Problem.*

Das obige Theorem wurde von Parks für allgemeine nichtlineare Optimierungsprobleme der folgenden Form (sog. *reduzible Optimierungsprobleme*) bewiesen:

Vollständiges Problem

$$\mathfrak{f}(\boldsymbol{\alpha}, \mathbf{t}) \longrightarrow \min_{\boldsymbol{\alpha}\in\mathbb{R}^n, \mathbf{t}\in\mathbb{R}^l} \quad \text{bei}$$

$$\begin{aligned} &g_i(\boldsymbol{\alpha}, \mathbf{t}) \geq 0,\ i = 1, \ldots, p_1, && c_i(\boldsymbol{\alpha}) \geq 0,\ i = 1, \ldots, p_3, && r_i(\mathbf{t}) \geq 0\ i = 1, \ldots, p_5, \\ &h_i(\boldsymbol{\alpha}, \mathbf{t}) = 0,\ i = 1, \ldots, p_2, && d_i(\boldsymbol{\alpha}) = 0,\ i = 1, \ldots, p_4, && s_i(\mathbf{t}) = 0\ i = 1, \ldots, p_6. \end{aligned}$$

Reduziertes Problem

$$f(\mathbf{t}) := \mathfrak{f}(\boldsymbol{\alpha}(\mathbf{t}), \mathbf{t}) \longrightarrow \min_{\mathbf{t}\in\mathbb{R}^l} \quad \text{bei}$$

$$r_i(\mathbf{t}) \geq 0,\ i = 1, \ldots, p_5, \quad s_i(\mathbf{t}) = 0,\ i = 1, \ldots, p_6,$$

wobei $\boldsymbol{\alpha}(\mathbf{t})$ das folgende Subproblem (A) löst:

Subproblem (A)

$$\mathfrak{f}(\boldsymbol{\alpha}; \mathbf{t}) \longrightarrow \min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \quad \text{bei}$$

$$\begin{aligned} &g_i(\boldsymbol{\alpha}, \mathbf{t}) \geq 0,\ i = 1, \ldots, p_1, && c_i(\boldsymbol{\alpha}) \geq 0,\ i = 1, \ldots, p_3, \\ &h_i(\boldsymbol{\alpha}, \mathbf{t}) = 0,\ i = 1, \ldots, p_2, && d_i(\boldsymbol{\alpha}) = 0,\ i = 1, \ldots, p_4. \end{aligned}$$

Die Schlüsselidee beim Beweis von Theorem 3.1 liegt in der Anwendung des Hauptsatzes der Sensitivitätstheorie [Fia76] auf Subproblem (A). Er liefert Sensitivitätsaussagen erster Ordnung für ein lokales Minimum, welches die Optimalitätsbedingungen zweiter Ordnung erfüllt, siehe auch [Fia83]. Tatsächlich stellen die Bedingungen 3(a)–(c) die Voraussetzungen dieses Satzes dar.

Theorem 3.1 ist von ähnlicher Bedeutung wie das entsprechende Theorem für separable Quadratmittelprobleme, siehe [GP73, Theorem 2.1], und ist eine direkte Verallgemeinerung. Das Auftreten der sog. gemischten Nebenbedingungen $g_i(\boldsymbol{\alpha}, \mathbf{t}) \geq 0$, $h_i(\boldsymbol{\alpha}, \mathbf{t}) = 0$ kompliziert die Ausdrücke für Gradient und Hesse-Matrix des reduzierten Funktionals erheblich. Der Fall von semi-linearen Gleichheitsnebenbedingungen $\mathbf{H}(\mathbf{t})\boldsymbol{\alpha} - \boldsymbol{\delta}(\mathbf{t}) = \mathbf{0}$ wurde in [KP78] und [Cor81] behandelt. Als Spezialfall des obigen Satzes erhalten wir Theorem 2.2 von Kaufman mit einer präzisen Formulierung der Äquivalenz.

3.3.3 Quantitative Analyse von Subproblem (A) und reduziertem Problem

Zur späteren Beschreibung der Lösungsmethode benötigen wir selbstverständlich neben der Äquivalenz der Probleme eine genaue quantitative Analyse von Subproblem (A) und reduziertem Problem, insbesondere Gradient, Jacobi-Matrix und Hesse-Matrix des reduzierten Funktionals f.

Für die Lagrange-Funktion l von Subproblem (A) haben wir

$$l(\boldsymbol{\alpha}, \mathbf{u}; \mathbf{t}) := \frac{1}{2} \left\| \mathbf{y} - \mathbf{B}(\mathbf{t})\boldsymbol{\alpha} \right\|^2 - \mathbf{u}^T \mathbf{g}(\boldsymbol{\alpha}; \mathbf{t})$$

mit den Lagrange-Parametern $\mathbf{u} \in \mathbb{R}_+^{2(n-p)}$ und dem Vektor der Nebenbedingungen

$$\mathbf{g} = \mathbf{g}(\boldsymbol{\alpha}; \mathbf{t}) := \begin{bmatrix} \mathbf{D}_p(\mathbf{t}) \\ -\mathbf{D}_p(\mathbf{t}) \end{bmatrix} \boldsymbol{\alpha} - \begin{pmatrix} \mathbf{L} \\ -\mathbf{U} \end{pmatrix} \in \mathbb{R}^{2(n-p)}.$$

Die vorzeichenbehafteten Gradienten der Nebenbedingungen sind

$$\begin{aligned} \mathbf{R} &:= -\left(\nabla_{\boldsymbol{\alpha}} \mathbf{g}\right)^T = -\begin{bmatrix} \mathbf{D}_p(\mathbf{t}) \\ -\mathbf{D}_p(\mathbf{t}) \end{bmatrix} \in \mathbb{R}^{2(n-p),n}, \\ \boldsymbol{\Gamma} &:= -\left(\nabla_{\mathbf{t}} \mathbf{g}\right)^T = -\left(\nabla_{\mathbf{t}} \begin{bmatrix} \mathbf{D}_p(\mathbf{t}) \\ -\mathbf{D}_p(\mathbf{t}) \end{bmatrix} \boldsymbol{\alpha}\right)^T \in \mathbb{R}^{2(n-p),l}. \end{aligned}$$

Größen, welche zu aktiven Nebenbedingungen von Subproblem (A) gehören, werden wir im weiteren durch einen Strich kennzeichnen, z.B. $\bar{\mathbf{R}} := -\left(\nabla_{\boldsymbol{\alpha}} \mathbf{g}_i\right)_{i \in \mathcal{I}}^T \in \mathbb{R}^{nact,n}$, $\bar{\boldsymbol{\Gamma}} := -\left(\nabla_{\mathbf{t}} \bar{\mathbf{g}}_i\right)_{i \in \mathcal{I}}^T \in \mathbb{R}^{nact,l}$, wobei $\mathcal{I} := \{i \in \{1, \ldots, 2(n-p)\} \,|\, g_i(\boldsymbol{\alpha}; \mathbf{t}) = 0\}$ und $nact := \#\mathcal{I}$ die Indexmenge und Anzahl der aktiven Restriktionen bezeichnen.

Sei $\boldsymbol{\partial} = \nabla_{\mathbf{t}}^T$ der Operator der Fréchet-Ableitung bez. $\mathbf{t}$. Auf Grund der Regularitätsbedingung 3(b) von Theorem 3.1 hat die Matrix $\bar{\mathbf{R}}$ vollen Zeilenrang $nact$. Sei $\mathbf{N} \in \mathbb{R}^{n,n-nact}$ eine Basis für den Nullraum von $\bar{\mathbf{R}}$, und sei $\bar{\mathbf{R}}^+$ die Moore-Penrose-Inverse von $\bar{\mathbf{R}}$. Parks zeigt, daß für allgemeine reduzible nichtlineare Optimierungsprobleme gilt

$$\begin{bmatrix} \nabla_{\boldsymbol{\alpha}}^2 l & \bar{\mathbf{R}}^T \\ \bar{\mathbf{R}} & \mathbf{0} \end{bmatrix} \begin{pmatrix} \boldsymbol{\partial}\boldsymbol{\alpha} \\ \boldsymbol{\partial}\bar{\mathbf{u}} \end{pmatrix} = -\begin{pmatrix} \nabla_{\boldsymbol{\alpha}\mathbf{t}}^2 l \\ \bar{\boldsymbol{\Gamma}} \end{pmatrix} \tag{3.17}$$

mit

$$\begin{bmatrix} \nabla_{\boldsymbol{\alpha}}^2 l & \bar{\mathbf{R}}^T \\ \bar{\mathbf{R}} & \mathbf{0} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{12}^T & \mathbf{W}_{22} \end{bmatrix} \tag{3.18}$$

und

$$\begin{aligned} \mathbf{W}_{11} &= \mathbf{N} \left[\mathbf{N}^T \left(\nabla_{\boldsymbol{\alpha}}^2 l\right) \mathbf{N}\right]^{-1} \mathbf{N}^T \in \mathbb{R}^{n,n} \\ \mathbf{W}_{12} &= \left[\mathbf{I} - \mathbf{W}_{11} \left(\nabla_{\boldsymbol{\alpha}}^2 l\right)\right] \bar{\mathbf{R}}^+ \in \mathbb{R}^{n,nact} \\ \mathbf{W}_{22} &= -\mathbf{W}_{12}^T \left(\nabla_{\boldsymbol{\alpha}}^2 l\right) \mathbf{W}_{12} \in \mathbb{R}^{nact,nact}. \end{aligned}$$

Sie erhält damit Gradient und Hesse-Matrix des reduzierten Funktionals [Par85, Lemma 4.4, 4.6] im allgemeinen Fall.

Lemma 3.2 (Gradient und Hesse-Matrix des reduzierten Funktionals)**.**

$$\begin{aligned}\nabla_{\mathbf{t}} f &= \nabla_{\mathbf{t}} \mathfrak{f}(\boldsymbol{\alpha}(\mathbf{t}), \mathbf{t}) + \boldsymbol{\Gamma}^T \mathbf{u} \\ \nabla^2_{\mathbf{t}} f &= \nabla^2_{\mathbf{t}} l - \left(\nabla^2_{\mathbf{t}\boldsymbol{\alpha}} l \,|\, \bar{\boldsymbol{\Gamma}}^T\right) \begin{bmatrix} \nabla^2_{\boldsymbol{\alpha}} l & \bar{\mathbf{R}}^T \\ \bar{\mathbf{R}} & \mathbf{0} \end{bmatrix}^{-1} \begin{pmatrix} \nabla^2_{\boldsymbol{\alpha}\mathbf{t}} l \\ \bar{\boldsymbol{\Gamma}} \end{pmatrix} \\ &= \left(\boldsymbol{\partial\alpha}^T \,|\, \mathbf{I}\right) \begin{bmatrix} \nabla^2_{\boldsymbol{\alpha}} l & \nabla^2_{\mathbf{t}\boldsymbol{\alpha}} l \\ \nabla^2_{\boldsymbol{\alpha}\mathbf{t}} l & \nabla^2_{\mathbf{t}} l \end{bmatrix} \begin{pmatrix} \boldsymbol{\partial\alpha} \\ \mathbf{I} \end{pmatrix}\end{aligned}$$

Die Größen auf der rechten Seite werden an der Stelle $(\boldsymbol{\alpha}(\mathbf{t}), \mathbf{t})$ *berechnet.*

Dieses Lemma zeigt auch, warum der Fall von gemischten Nebenbedingungen zusätzliche Schwierigkeiten bereitet. Bei Fehlen dieser Nebenbedingungen hat man $\nabla_{\mathbf{t}} f(\mathbf{t}) = \nabla_{\mathbf{t}} \mathfrak{f}(\boldsymbol{\alpha}(\mathbf{t}), \mathbf{t})$, d. h. die Berechnung des Gradienten ändert sich nicht.

Für die Hesse-Matrix von l erhalten wir in unserem Spezialfall

$$\begin{aligned}\nabla^2_{\boldsymbol{\alpha}} l &= \mathbf{B}^T\mathbf{B} \in \mathbb{R}^{n,n} & \nabla^2_{\boldsymbol{\alpha}\mathbf{t}} l &= -\mathbf{B}^T\mathfrak{J}_{\mathbf{t}} + \mathbf{K} \in \mathbb{R}^{n,l} \\ \nabla^2_{\mathbf{t}\boldsymbol{\alpha}} l &= -\mathfrak{J}_{\mathbf{t}}^T\mathbf{B} + \mathbf{K}^T \in \mathbb{R}^{l,n} & \nabla^2_{\mathbf{t}} l &= \mathfrak{J}_{\mathbf{t}}^T\mathfrak{J}_{\mathbf{t}} + \mathbf{S}_{\mathbf{t}} \in \mathbb{R}^{l,l}\end{aligned}$$

mit

$$\begin{aligned}\mathfrak{J}_{\mathbf{t}} &:= \boldsymbol{\partial}\mathfrak{F} = -\boldsymbol{\partial}\mathbf{B}\boldsymbol{\alpha} \in \mathbb{R}^{m,l} \\ \mathbf{K} &:= -\boldsymbol{\partial}\mathbf{B}^T(\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}) + \boldsymbol{\partial}\mathbf{R}^T\mathbf{u} \in \mathbb{R}^{n,l} \\ \mathbf{S}_{\mathbf{t}} &:= \mathbf{u}^T\boldsymbol{\partial}^2\mathbf{R}\boldsymbol{\alpha} - \left(\boldsymbol{\partial}^2\mathbf{B}\boldsymbol{\alpha}\right)^T(\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}) \in \mathbb{R}^{l,l}.\end{aligned}$$

Man beachte, daß der Term $\left(\boldsymbol{\partial}^2\mathbf{B}\boldsymbol{\alpha}\right)^T(\mathbf{y} - \mathbf{B}\boldsymbol{\alpha})$ in $\mathbf{S}_{\mathbf{t}}$ in der Arbeit [Par85] fälschlicherweise fehlt. Dies beeinträchtigt jedoch die weiteren Resultate nicht, da dieser Term als einziger Term mit zweiten Ableitungen sowieso später vernachlässigt wird.

Aus (3.17) und (3.18) erhalten wir

$$\boldsymbol{\partial\alpha} = -\mathbf{W}_{11}\left(\nabla^2_{\boldsymbol{\alpha}\mathbf{t}} l\right) - \mathbf{W}_{12}\bar{\boldsymbol{\Gamma}} = \mathbf{W}_{11}\mathbf{B}^T\mathfrak{J}_{\mathbf{t}} - \mathbf{W}_{11}\mathbf{K} - \left(\mathbf{I} - \mathbf{W}_{11}\mathbf{B}^T\mathbf{B}\right)\bar{\mathbf{R}}^+\bar{\boldsymbol{\Gamma}}.$$

Mit den orthogonalen Projektoren $\mathbf{P}_{BN} := (\mathbf{BN})(\mathbf{BN})^+ \in \mathbb{R}^{m,m}$ und $\mathbf{P}^{\perp}_{BN} := \mathbf{I} - \mathbf{P}_{BN} \in \mathbb{R}^{m,m}$ gilt daher

$$\begin{aligned}\mathbf{BW}_{11}\mathbf{B}^T &= \mathbf{BN}(\mathbf{N}^T\mathbf{B}^T\mathbf{BN})^{-1}\mathbf{N}^T\mathbf{B}^T = \mathbf{P}_{BN}, \\ \mathbf{B}(\mathbf{I} - \mathbf{W}_{11}\mathbf{B}^T\mathbf{B}) &= \mathbf{P}^{\perp}_{BN}\mathbf{B},\end{aligned}$$

sowie

$$\begin{aligned}\mathbf{BW}_{11}\mathbf{K} &= \mathbf{BN}(\mathbf{N}^T\mathbf{B}^T\mathbf{BN})^{-1}\mathbf{N}^T\mathbf{K} \\ &= \mathbf{BN}(\mathbf{N}^T\mathbf{B}^T\mathbf{BN})^{-T}\mathbf{N}^T\mathbf{K} \\ &= \left[(\mathbf{BN})^+\right]^T\mathbf{N}^T\mathbf{K}.\end{aligned}$$

Für die Jacobi-Matrix des reduzierten Funktionals ergibt sich $\mathbf{J}(\mathbf{t}) := \boldsymbol{\partial}\mathbf{F}(\mathbf{t}) = \boldsymbol{\partial}(\mathbf{y} - \mathbf{B}(\mathbf{t})\boldsymbol{\alpha}(\mathbf{t})) = \mathfrak{J}_{\mathbf{t}} - \mathbf{B}\boldsymbol{\partial\alpha}$. Setzen wir $\mathbf{B}\boldsymbol{\partial\alpha} = \mathbf{P}_{BN}\mathfrak{J}_{\mathbf{t}} - \left[(\mathbf{BN})^+\right]^T\mathbf{N}^T\mathbf{K} - \mathbf{P}^{\perp}_{BN}\mathbf{B}\bar{\mathbf{R}}^+\bar{\boldsymbol{\Gamma}}$ ein, so erhalten wir unmittelbar [Par85, Lemma 6.2]:

Lemma 3.3 (Jacobi-Matrix des reduzierten Funktionals)**.**
Die Jacobi-Matrix $\mathbf{J}(\mathbf{t}) = \boldsymbol{\partial}\mathbf{F}(\mathbf{t})$ *von* $\mathbf{F}(\mathbf{t}) = \mathbf{y} - \mathbf{B}(\mathbf{t})\boldsymbol{\alpha}(\mathbf{t})$ *ist gegeben durch*

$$\mathbf{J}(\mathbf{t}) = \mathbf{P}_{\bar{B}N}^{\perp}(\mathbf{t})\Big(\mathfrak{J}_{\mathbf{t}}(\mathbf{t}) + \mathbf{B}(\mathbf{t})\bar{\mathbf{R}}^{+}(\mathbf{t})\bar{\boldsymbol{\Gamma}}(\mathbf{t})\Big) + \mathbf{P}_{BN}(\mathbf{t})\left[(\mathbf{B}(\mathbf{t})\mathbf{N}(\mathbf{t}))^{+}\right]^{T}\mathbf{N}^{T}(\mathbf{t})\mathbf{K}(\mathbf{t}),$$

wobei $\mathbf{K}(\mathbf{t}) = \mathbf{K}(\boldsymbol{\alpha}(\mathbf{t}), \mathbf{t})$ *und* $\mathbf{K}(\boldsymbol{\alpha}, \mathbf{t}) := -\boldsymbol{\partial}\mathbf{B}^{T}(\mathbf{t})\left(\mathbf{y} - \mathbf{B}(\mathbf{t})\boldsymbol{\alpha}\right) + \boldsymbol{\partial}\mathbf{R}^{T}(\mathbf{t})\mathbf{u}$.

Für die Residuumsfunktion des reduzierten Funktionals erhält man nach [Par85, Lemma 6.1] (dort ist $\mathfrak{F}(\boldsymbol{\alpha}, \mathbf{t}) = \mathbf{B}(\mathbf{t})\boldsymbol{\alpha} - \mathbf{y}$!):

Lemma 3.4 (Residuumsfunktion des reduzierten Funktionals)**.**

$$\mathbf{F}(\mathbf{t}) = \mathbf{P}_{\bar{B}N}^{\perp}(\mathbf{t})\Big(\mathbf{y} - \mathbf{B}(\mathbf{t})\bar{\mathbf{R}}^{+}(\mathbf{t})\bar{\boldsymbol{\xi}}\Big) \quad \textit{mit} \quad \bar{\boldsymbol{\xi}} := -\begin{pmatrix} \mathbf{L} \\ -\mathbf{U} \end{pmatrix}_{i\in\mathcal{I}} \in \mathbb{R}^{nact}.$$

3.3.4 Struktur der Jacobi-Matrix, Kaufman-Approximation

In den letzten Aussagen kann man mit den bisherigen Bezeichnungen eine Struktur nur schwer erkennen. Wir definieren daher

$$\boldsymbol{\psi} := \mathfrak{J}_{\mathbf{t}} + \mathbf{B}\bar{\mathbf{R}}^{+}\bar{\boldsymbol{\Gamma}}, \qquad \boldsymbol{\phi} := \left((\mathbf{BN})^{+}\right)^{T}\mathbf{N}^{T}\mathbf{K}, \qquad \mathbf{v} := \mathbf{y} - \mathbf{B}\bar{\mathbf{R}}^{+}\bar{\boldsymbol{\xi}}.$$

Mit $\mathbf{P} = \mathbf{P}_{BN}$ lauten die letzten Aussagen in den neuen Bezeichnungen $\mathbf{F} = \mathbf{P}^{\perp}\mathbf{v}$ und $\mathbf{J} = \mathbf{P}^{\perp}\boldsymbol{\psi} + \mathbf{P}\boldsymbol{\phi}$. Der Term $\mathbf{P}\boldsymbol{\phi} = \mathbf{P}_{BN}\left((\mathbf{BN})^{+}\right)^{T}\mathbf{N}^{T}\mathbf{K}$ erschwert die Berechnung der Jacobi-Matrix enorm, insbesondere die Ausnutzung der Schwachbesetztheitsstruktur. Glücklicherweise gilt

$$\mathbf{J}^{T}\mathbf{F} = \boldsymbol{\psi}^{T}\left(\mathbf{P}^{\perp}\right)^{T}\mathbf{P}^{\perp}\mathbf{v} \qquad \text{und} \qquad \mathbf{J}^{T}\mathbf{J} = \boldsymbol{\psi}^{T}\left(\mathbf{P}^{\perp}\right)^{T}\mathbf{P}^{\perp}\boldsymbol{\psi} + \boldsymbol{\phi}^{T}\mathbf{P}^{T}\mathbf{P}\boldsymbol{\phi},$$

d. h. der Term $\mathbf{P}\boldsymbol{\phi}$ trägt nicht zum Gradienten $\mathbf{J}^{T}\mathbf{F}$ bei und sein Beitrag zu $\mathbf{J}^{T}\mathbf{J}$ ist $\boldsymbol{\phi}^{T}\mathbf{P}^{T}\mathbf{P}\boldsymbol{\phi}$, wie Parks bemerkt. Eine Untersuchung der Größenordnung dieses Terms fand nicht statt. Im unrestringierten Fall gilt darüberhinaus $\boldsymbol{\phi}^{T}\mathbf{P}^{T}\mathbf{P}\boldsymbol{\phi} = \mathcal{O}(\|\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}\|^{2})$, so daß dieser Term im Rahmen von Gauß-Newton-Verfahren vernachlässigt werden kann, vgl. die Arbeit von Kaufman [Kau75]. Wir verallgemeinern diese Philosophie auf den restringierten Fall und definieren

Definition 3.3 (Kaufman-Approximation)**.** Die Approximation

$$\mathbf{J}_{K} := \mathbf{P}^{\perp}\boldsymbol{\psi} = \mathbf{P}_{\bar{B}N}^{\perp}\left(\mathfrak{J}_{\mathbf{t}} + \mathbf{B}\bar{\mathbf{R}}^{+}\bar{\boldsymbol{\Gamma}}\right) \in \mathbb{R}^{m,l}$$

für die Jacobi-Matrix

$$\mathbf{J} = \mathbf{P}^{\perp}\boldsymbol{\psi} + \mathbf{P}\boldsymbol{\phi} = \mathbf{P}_{\bar{B}N}^{\perp}\left(\mathfrak{J}_{\mathbf{t}} + \mathbf{B}\bar{\mathbf{R}}^{+}\bar{\boldsymbol{\Gamma}}\right) + \mathbf{P}_{\bar{B}N}^{\perp}\left((\mathbf{BN})^{+}\right)^{T}\mathbf{N}^{T}\mathbf{K}$$

des reduzierten Funktionals heißt *Kaufman-Approximation.*

Im weiteren verwenden wir diese billiger zu berechnende Approximation an die Jacobi-Matrix und untersuchen sowohl den qualitativen als auch quantitativen Einfluß dieser Approximation. Die theoretischen Ergebnisse aus [Par85] werden nach jüngsten Aussagen der Autorin T. A. Parks erstmals praktisch umgesetzt und numerisch getestet.

Wir untersuchen die Größenordnung des Fehlers in der Hesse-Matrix $\mathbf{J}^T\mathbf{J}$ des Gauß-Newton-Modells, wenn wir an Stelle von $\mathbf{J}$ mit der Kaufman-Approximation $\mathbf{J}_K$ arbeiten. Es gilt $\mathbf{J}^T\mathbf{J} = \mathbf{J}_K^T\mathbf{J}_K + \boldsymbol{\phi}^T\mathbf{P}^T\mathbf{P}\boldsymbol{\phi}$ mit

$$\begin{aligned}\boldsymbol{\phi}^T\mathbf{P}^T\mathbf{P}\boldsymbol{\phi} &= \left(((\mathbf{BN})^+)^T\mathbf{N}^T\mathbf{K}\right)^T \mathbf{P}_{BN}^T\mathbf{P}_{BN}((\mathbf{BN})^+)^T\mathbf{N}^T\mathbf{K} \\ &= \left(((\mathbf{BN})^+)^T\mathbf{N}^T\mathbf{K}\right)^T ((\mathbf{BN})^+)^T\mathbf{N}^T\mathbf{K} \\ &= \mathbf{K}^T\mathbf{N}\left(\mathbf{N}^T\mathbf{B}^T\mathbf{BN}\right)^{-1}\mathbf{N}^T\mathbf{K}.\end{aligned}$$

Für die Matrix $\mathbf{K}$ gilt $\mathbf{K} = \left(-\boldsymbol{\partial}\mathbf{B}^T\right)(\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}) + \mathbf{R}^T\mathbf{u}$, d. h. der erste Term ist von der Größenordnung $\mathcal{O}(\|\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}\|)$. Aus den notwendigen Optimalitätsbedingungen für Subproblem (A) erhalten wir $\nabla_{\boldsymbol{\alpha}} l = (-\mathbf{B})^T(\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}) + \mathbf{R}^T\mathbf{u} = \mathbf{0}$ und wegen $(u_i)_{i\notin\mathcal{I}} = 0$ schließlich $\bar{\mathbf{R}}^T\bar{\mathbf{u}} = \mathbf{B}^T(\mathbf{y} - \mathbf{B}\boldsymbol{\alpha})$. Die Matrix $\bar{\mathbf{R}} \in \mathbb{R}^{nact,n}$ hat nach den Voraussetzungen von Theorem 3.1 Vollrang *nact*, so daß

$$\bar{\mathbf{u}} = (\bar{\mathbf{R}}\bar{\mathbf{R}}^T)^{-1}\bar{\mathbf{R}}\mathbf{B}^T(\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}) = \bar{\mathbf{R}}^+\mathbf{B}^T(\mathbf{y} - \mathbf{B}\boldsymbol{\alpha})$$

und der zweite Term der Matrix $\mathbf{K}$ ebenfalls von der Größenordnung $\mathcal{O}(\|\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}\|)$ ist. Es gilt also auch im restringierten Fall $\mathbf{J}^T\mathbf{J} = \mathbf{J}_K^T\mathbf{J}_K + \mathcal{O}\left(\|\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}\|^2\right)$, so daß dieser Term im Rahmen von verallgemeinerten Gauß-Newton-Verfahren vernachlässigt werden kann und sich qualitativ die gleichen Konvergenzeigenschaften wie im unrestringierten Fall ergeben. Die numerischen Tests bestätigen diese Aussage.

3.4 Splineglättung mit freien Knoten und Ungleichheitsnebenbedingungen an Ableitungen

Wir können die Ergebnisse aus Abschnitt 3.3 unmittelbar auf das vollständige restringierte Glättungsproblem FCSP anwenden, wenn wir das Paar $\{\mathbf{B}(\mathbf{t}), \mathbf{y}\}$ durch die Größen

$$\left\{\begin{bmatrix}\mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t})\end{bmatrix}, \begin{pmatrix}\mathbf{y} \\ \mathbf{0}\end{pmatrix}\right\}.$$

ersetzen. Das entsprechende reduzierte Problem heißt *reduziertes restringiertes Glättungsproblem* RCSP und ist definiert durch

$$\min\left\{f(\mathbf{t}) := \frac{1}{2}\|\mathbf{F}(\mathbf{t})\|^2 = \frac{1}{2}\left\|\begin{pmatrix}\mathbf{y} \\ \mathbf{0}\end{pmatrix} - \begin{bmatrix}\mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t})\end{bmatrix}\boldsymbol{\alpha}(\mathbf{t})\right\|^2 : \mathbf{Ct} \geq \mathbf{h}\,,\ \mathbf{t} \in \mathbb{R}^l\right\}, \tag{3.19}$$

wobei $\mathbf{F}(\mathbf{t}) := \begin{pmatrix}\mathbf{y} \\ \mathbf{0}\end{pmatrix} - \begin{bmatrix}\mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t})\end{bmatrix}\boldsymbol{\alpha}(\mathbf{t})$, und $\boldsymbol{\alpha}(\mathbf{t})$ löst *Subproblem (A)*

$$\min\left\{\frac{1}{2}\left\|\begin{pmatrix}\mathbf{y} \\ \mathbf{0}\end{pmatrix} - \begin{bmatrix}\mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t})\end{bmatrix}\boldsymbol{\alpha}\right\|^2 : \mathbf{L} \leq \mathbf{D}_p(\mathbf{t})\boldsymbol{\alpha} \leq \mathbf{U}\,,\ \boldsymbol{\alpha} \in \mathbb{R}^n\right\}. \tag{3.20}$$

In diesem Abschnitt zeigen wir, daß das Problem RCSP immer eine Lösung besitzt, und untersuchen die Beziehungen zwischen Lösungen von FCSP und RCSP. Um die Existenz von Lösungen für RCSP zu beweisen, benötigen wir einen schwächeren Störungssatz als den

Hauptsatz der Störungstheorie von Fiacco. Aussagen über die stetig differenzierbare Abhängigkeit der Lösung von den Störungen benötigen die strikte Komplementarität. Man kann jedoch zeigen, siehe [Dan73], daß für die Lösung der quadratischen Optimierungsprobleme

$$\mathbf{x}^* = \operatorname{argmin}\left\{\mathbf{b}^T\mathbf{x} + \tfrac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} \,:\, \mathbf{H}^T\mathbf{x} + \mathbf{h}^0 = \mathbf{0}, \mathbf{G}^T\mathbf{x} + \mathbf{g}^0 \geq \mathbf{0}, \mathbf{x} \in \mathbb{R}^n\right\}$$
$$\tilde{\mathbf{x}}^* = \operatorname{argmin}\left\{\tilde{\mathbf{b}}^T\mathbf{x} + \tfrac{1}{2}\mathbf{x}^T\tilde{\mathbf{A}}\mathbf{x} \,:\, \tilde{\mathbf{H}}^T\mathbf{x} + \tilde{\mathbf{h}}^0 = \mathbf{0}, \tilde{\mathbf{G}}^T\mathbf{x} + \tilde{\mathbf{g}}^0 \geq \mathbf{0}, \mathbf{x} \in \mathbb{R}^n\right\}$$

die Beziehung

$$\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\| \leq C \max\left\{\|\mathbf{A} - \tilde{\mathbf{A}}\|, \|\mathbf{H} - \tilde{\mathbf{H}}\|, \|\mathbf{G} - \tilde{\mathbf{G}}\|, \|\mathbf{b} - \tilde{\mathbf{b}}\|, \|\mathbf{h}^0 - \tilde{\mathbf{h}}^0\|, \|\mathbf{g}^0 - \tilde{\mathbf{g}}^0\|\right\}$$

gilt, wenn die Störungen $\|\mathbf{A} - \tilde{\mathbf{A}}\|$, ... nur hinreichend klein sind ($\mathbf{A}$ symmetrisch, positiv definit, $\mathbf{H}$ spaltenregulär, $\exists\, \mathbf{x}^0 : \mathbf{G}^T\mathbf{x}^0 + \mathbf{g}^0 > \mathbf{0}$ (Slater-Bedingung)).

Satz 3.5 (Lineare Unabhängigkeit der Gradienten der Nebenbedingungen)**.**
Sei $\tau_j < \tau_{j+k-p}$ $(j = p+1, \ldots, n)$ und seien

$$\mathbf{g} = \mathbf{g}(\boldsymbol{\alpha}; \mathbf{t}) := \begin{bmatrix} \mathbf{D}_p(\mathbf{t}) \\ -\mathbf{D}_p(\mathbf{t}) \end{bmatrix} \boldsymbol{\alpha} - \begin{pmatrix} \mathbf{L} \\ -\mathbf{U} \end{pmatrix} \geq \mathbf{0}$$

die Nebenbedingungen von Subproblem (A). Die Gradienten der aktiven Nebenbedingungen sind linear unabhängig, d. h.

$$\bar{\mathbf{R}} := -\left(\nabla_{\boldsymbol{\alpha}} \mathbf{g}_i^T\right)_{i \in \mathcal{I}} = -\begin{bmatrix} \mathbf{D}_p(\mathbf{t}) \\ -\mathbf{D}_p(\mathbf{t}) \end{bmatrix}_{i \in \mathcal{I}} \in \mathbb{R}^{nact,n}$$

hat Vollrang $\operatorname{rank} \bar{\mathbf{R}} = nact = \#\mathcal{I}$ *genau dann, wenn die* strikte Konsistenzbedingung $L_i < U_i$ $(i = 1, \ldots, n-p)$ $(\mathbf{L} < \mathbf{U})$ *gilt.*

Beweis. In einem ersten Schritt betrachten wir die Nebenbedingung $\mathbf{g} = \mathbf{D}_p(\mathbf{t})\boldsymbol{\alpha} - \mathbf{L} \geq \mathbf{0}$. Offensichtlich sind die Gradienten der aktiven Nebenbedingungen linear unabhängig, wenn *alle* Zeilen von $\mathbf{D}_p$ linear unabhängig sind, d. h. $\operatorname{rank} \mathbf{D}_p = n - p$. Da $\mathbf{D}_p$ eine obere Dreiecksmatrix ist, gilt

$$\operatorname{rank} \mathbf{D}_p = n - p \quad \Longleftrightarrow \quad (\mathbf{D}_p)_{ii} \neq 0 \quad (i = 1, \ldots, n-p).$$

Aus (2.4) folgt die lineare Unabhängigkeit der Gradienten der Nebenbedingungen zu $\mathbf{g} = \mathbf{D}_p(\mathbf{t})\boldsymbol{\alpha} - \mathbf{L} \geq \mathbf{0}$, falls $\tau_j < \tau_{j+k-p}$ $(j = p+1, \ldots, n)$.

Betrachten wir nun die ursprünglichen Nebenbedingungen. Offensichtlich kann im Fall $L_i < U_i$ eine Nebenbedingung entweder an L_i oder an U_i aktiv werden, aber niemals gleichzeitig. Die Matrix $\bar{\mathbf{R}}$ enthält deshalb keine zwei identischen Zeilen von $\mathbf{D}_p$ (identisch bis auf einen Faktor -1). Daher sind die Zeilen von $\bar{\mathbf{R}}$ linear unabhängig, falls $\tau_j < \tau_{j+k-p}$ $(j = p+1, \ldots, n)$. Umgekehrt enthält $\bar{\mathbf{R}}$ im Fall $L_i = U_i$ zwei bis auf Vorzeichen identische Zeilen von $\mathbf{D}_p$, d. h. $\bar{\mathbf{R}}$ hat keinen Vollrang. □

Bemerkung 3.1. Man beachte, daß die Bedingung $l_i^{(p)} < u_i^{(p)}$ für die Nebenbedingung $l_i^{(p)} \leq s^{(p)}(x) \leq u_i^{(p)}$ für alle $x \in [\tau_i, \tau_{i+1})$, $(i = k, \ldots, n)$ *nicht hinreichend* für die strikte Konsistenz der Nebenbedingungen ist, ja noch nicht einmal hinreichend für die Konsistenz, siehe die Beispiele 3.1–3.3. In unserem Fall ist die strikte Konsistenzbedingung äquivalent zur Slater-Bedingung, d. h. $\exists \boldsymbol{\alpha}^0 \in \mathbb{R}^n$ mit $\mathbf{g}(\boldsymbol{\alpha}^0; \mathbf{t}) > \mathbf{0}$.

Theorem 3.2 (Existenz einer Lösung von RCSP)**.**
Die Menge der zulässigen Knoten $\left\{\mathbf{t} \in \mathbb{R}^l : \mathbf{Ct} - \mathbf{h} \geq \mathbf{0}\right\}$ *sei nichtleer. Für feste* $p, r \in \{0, \ldots, q\}$, $0 \leq q < k$ *gelte:*

$(V1)$ *Die Knoten erfüllen die Bedingung* $\tau_j < \tau_{j+k-q}$ $(j = q+1, \ldots, n)$.

$(V2)$ *Die Regularitätsbedingung* $m \geq r$ *und* $\mu > 0$ *ist erfüllt.*

$(V4)$ *Die Nebenbedingungen* $\mathbf{L} \leq \mathbf{D}_p(\mathbf{t})\boldsymbol{\alpha} \leq \mathbf{U}$ *erfüllen die strikte Konsistenzbedingung* $\mathbf{L} < \mathbf{U}$.

Dann besitzt das reduzierte restringierte Glättungsproblem RCSP (3.19), (3.20) *eine Lösung* $\mathbf{t}^*$.

Beweis. Voraussetzung (V1) sichert die Existenz der Matrizen für alle zulässigen Knoten. Darüberhinaus sind die Matrixfunktionen $\mathbf{B}(.)$, $\mathbf{S}_r(.)$ und $\mathbf{D}_p(.)$ stetige Funktionen der (freien) Knoten. Wegen Lemma 2.5 sichert (V2) die Vollrangeigenschaft der Systemmatrix $\mathbf{B}_\mu$ unabhängig von der Position der Knoten, d. h. die Hesse-Matrix von Subproblem (A) ist positiv definit. Schließlich liefert die strikte Konsistenzbedingung (V4) die Exsitenz eines Parameters $\boldsymbol{\alpha}^0$, so daß $\mathbf{L} < \mathbf{D}_p(\mathbf{t})\boldsymbol{\alpha}^0 < \mathbf{U}$, d. h. die Slater-Bedingung für die formerhaltenden Nebenbedingungen ist erfüllt, vgl. Bemerkung 3.1. Mit

$$\mathbf{A}(\mathbf{t}) := \begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}) \end{bmatrix}^T \begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}) \end{bmatrix}, \qquad \mathbf{b}(\mathbf{t}) := -\begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}) \end{bmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix},$$
$$\mathbf{G}(\mathbf{t}) := \begin{bmatrix} \mathbf{D}_p(\mathbf{t}) \\ -\mathbf{D}_p(\mathbf{t}) \end{bmatrix}^T, \qquad \mathbf{g}^0 := -\begin{pmatrix} \mathbf{L} \\ -\mathbf{U} \end{pmatrix},$$

ist Subproblem (A) äquivalent zu

$$\mathbf{b}(\mathbf{t})^T\boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{A}(\mathbf{t})\boldsymbol{\alpha} \to \min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \quad \text{bei} \quad \mathbf{G}(\mathbf{t})^T\boldsymbol{\alpha} + \mathbf{g}^0 \geq \mathbf{0}.$$

Wenn wir den Parameter $\mathbf{t}$ durch $\mathbf{t} + \delta\mathbf{t}$ ersetzen, erhalten wir ein gestörtes quadratisches Optimierungsproblem. Auf Grund der Stetigkeit der Matrixfunktionen $\mathbf{B}(.)$, $\mathbf{S}_r(.)$ und $\mathbf{D}_p(.)$ sind die Änderungen von $\mathbf{b}$, $\mathbf{A}$, $\mathbf{G}$ und $\mathbf{g}^0$ klein bei kleinen Störungen des Parameters. Wenden wir nun den Störungssatz für definite quadratische Optimierungsprobleme [Dan73] an, so erhalten wir die Lipschitz-stetige Änderung der Lösung $\boldsymbol{\alpha}(.)$ von Subproblem (A), d. h. das reduzierte Funktional f ist stetig. Wir können daher wiederum den Satz von Weierstraß auf das stetige Funktional f über der abgeschlossenen (wegen $\mathbf{Ct}-\mathbf{h} \geq \mathbf{0}$) und beschränkten (wegen $a \leq \tau_{p(1)}$ und $\tau_{p(l)} \leq b$) Menge der zulässsigen Knoten $\{\mathbf{t} \in \mathbb{R}^l : \mathbf{Ct} - \mathbf{h} \geq \mathbf{0}\}$ anwenden. □

Man beachte, daß sich die Voraussetzungen des Satzes in natürlicher Weise global für alle Knoten erfüllen lassen.

Kommen wir nun zur angekündigten Äquivalenz der Lösungen von vollständigem und reduziertem Glättungsproblem:

Theorem 3.3 (Äquivalenz von vollständigem und reduziertem Problem)**.**
Sei $\mathbf{t}^*$ *eine zulässige Knotenfolge,* $\mathbf{t}^* \in \left\{\mathbf{t} \in \mathbb{R}^l : \mathbf{Ct} - \mathbf{h} \geq \mathbf{0}\right\}$. *Für feste* $p, r \in \{0, \ldots, q\}$, $0 \leq q < k$ *gelte:*

($V1$) *Die Knoten erfüllen die Bedingung* $\tau_j < \tau_{j+k-q}$ $(j = q+1, \ldots, n)$.

($V2$) *Die Regularitätsbedingung* $m \geq r$ *und* $\mu > 0$ *ist erfüllt.*

($V3$) *Die freien Knoten* $\mathbf{t}^*$ *sind einfache Knoten, d. h.* $\#\tau^*_{p(j)} = 1$ $(j = 1, \ldots, l)$. *Es gilt* $k \geq 3$.

($V4$) *Die Nebenbedingungen* $\mathbf{L} \leq \mathbf{D}_p(\mathbf{t})\boldsymbol{\alpha} \leq \mathbf{U}$ *erfüllen die strikte Konsistenzbedingung* $\mathbf{L} < \mathbf{U}$.

($V5$) *Die Lagrange-Parameter* $\mathbf{u}^*$ *von Subproblem (A) an der Stelle* $\boldsymbol{\alpha}(\mathbf{t}^*)$ *sind strikt komplementär.*

Dann gelten für das vollständige restringierte Glättungsproblem FCSP (3.9), (3.10) *und das reduzierte restringierte Glättungsproblem* RCSP (3.19), (3.20) *die folgenden Beziehungen:*

(i) Wenn $(\boldsymbol{\alpha}^*, \mathbf{t}^*)$ *eine globale Minimumstelle des Ausgangsproblems* FCSP *ist, dann erfüllt* $\boldsymbol{\alpha}^*$ *die notwendigen (und für quadratische definite Probleme auch hinreichende) Optimalitätsbedingungen erster Ordnung für das Subproblem (A),* $\mathbf{t}^*$ *ist eine globale Minimumstelle des reduzierten Problems* RCSP *und es gilt* $f(\mathbf{t}^*) = \mathfrak{f}(\boldsymbol{\alpha}^*, \mathbf{t}^*)$, $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}(\mathbf{t}^*)$.

(ii) Wenn $\mathbf{t}^*$ *die notwendigen Optimalitätsbedingungen erster Ordnung für das reduzierte Problem* RCSP *erfüllt, so erfüllt* $(\boldsymbol{\alpha}(\mathbf{t}^*), \mathbf{t}^*)$ *die notwendigen Optimalitätsbedingungen erster Ordnung für das Ausgangsproblem* FCSP.

Beweis. Sei $\mathbf{t}^*$ eine zulässige Knotenfolge, d. h. $\mathbf{Ct} - \mathbf{h} \geq \mathbf{0}$.
1.Differenzierbarkeit der Problemfunktionen

Aus den Voraussetzungen (V1) und (V3) folgt die zweimalige stetige Differenzierbarkeit von

$$\mathfrak{f}(\boldsymbol{\alpha}, \mathbf{t}) = \frac{1}{2} \left\| \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} - \begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}) \end{bmatrix} \boldsymbol{\alpha} \right\|^2$$

bezüglich $\boldsymbol{\alpha}$ sowie die stetige Differenzierbarkeit des Gradienten

$$\nabla_{\boldsymbol{\alpha}} \mathfrak{f}(\boldsymbol{\alpha}, \mathbf{t}) = - \begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}) \end{bmatrix}^T \left(\begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} - \begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}) \end{bmatrix} \boldsymbol{\alpha} \right)$$

bezüglich $\mathbf{t}$ in einer Umgebung $\mathcal{U}_1^*$ von $\mathbf{t}^*$.
2.Differenzierbarkeit der Nebenbedingungen

Die Nebenbedingung $\mathbf{Ct} - \mathbf{h} \geq \mathbf{0}$ ist stetig differenzierbar bez. $\mathbf{t}$. Die Nebenbedingung $\mathbf{L} \leq \mathbf{D}_p(\mathbf{t})\boldsymbol{\alpha} \leq \mathbf{U}$ ist wegen (V1) und (V3) stetig differenzierbar bez. $\boldsymbol{\alpha}$ und $\mathbf{t}$ in einer Umgebung $\mathcal{U}_2^*$ von $\mathbf{t}^*$.
3.Voraussetzungen an Subproblem (A)

(a) Unter der Voraussetzung (V2) hat die Systemmatrix $\mathbf{B}_\mu$ von Subproblem (A) Vollrang n für alle zulässigen Knotenfolgen $\mathbf{t}$. Das quadratische Optimierungsproblem hat daher eine positiv definite Hesse-Matrix und besitzt folglich über dem nichtleeren zulässigen Bereich $\mathbf{L} \leq \mathbf{D}_p(\mathbf{t})\boldsymbol{\alpha} \leq \mathbf{U}$ eine eindeutig bestimmte Lösung $\boldsymbol{\alpha}(\mathbf{t})$. Es gelten die hinreichenden Optimalitätsbedingungen zweiter Ordnung an der Stelle $\boldsymbol{\alpha}(\mathbf{t})$.

(b) Aus Voraussetzung (V4) folgt nach Satz 3.5 die lineare Unabhängigkeit der Gradienten der aktiven Restriktionen von Subproblem (A).

(c) Falls die eindeutig bestimmten Lagrange-Parameter $\mathbf{u}^*$ von Subproblem (A) an der Stelle $\mathbf{t}^*$ strikt komplementär sind, vgl. (V5), dann existiert eine Umgebung $\mathcal{U}_4^*$ von $\mathbf{t}^*$ so, daß die Lagrange-Parameter $\mathbf{u}$ von Subproblem (A) an der Stelle $\mathbf{t} \in \mathcal{U}_4^*$ (bzw. $\boldsymbol{\alpha}(\mathbf{t})$) ebenfalls strikt komplementär sind.

Im nichtleeren Durchschnitt der Mengen $\mathcal{U}_1^*, \ldots, \mathcal{U}_4^*$ sind somit alle Voraussetzungen von Theorem 3.1 erfüllt. Die Aussagen ergeben sich durch unmittelbare Anwendung des Theorems. Für den Teil (i) der Behauptung gilt zudem:

Da das Subproblem (A) ein quadratisches, definites Optimierungsproblem ist, existiert ein eindeutiges $\boldsymbol{\alpha}^*$ unter allen Paaren $(\boldsymbol{\alpha}^*, \mathbf{t}^*)$, welche $\mathfrak{f}$ minimieren und denselben Minimalwert ergeben, und es gilt $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}(\mathbf{t}^*)$. □

Korollar 3.6.
Die Voraussetzungen von Theorem 3.3 seien in einer Umgebung Ω der zulässigen Knotenfolge $\mathbf{t}^*$ erfüllt. Wenn $\mathbf{t}^*$ eine globale Minimumstelle des reduzierten Funktionals $f(\mathbf{t})$ in Ω ist, so ist $(\boldsymbol{\alpha}(\mathbf{t})^*, \mathbf{t}^*)$ globale Minimumstelle von $\mathfrak{f}(\boldsymbol{\alpha}, \mathbf{t})$ für $\mathbf{t} \in \Omega$.

Beweis. Sei $\mathbf{t}^*$ eine globale Minimumstelle von $f(\mathbf{t})$ in Ω und sei $\boldsymbol{\alpha}(\mathbf{t}^*)$ die zugehörige eindeutige Lösung von Subproblem (A). Klar ist, daß $\mathfrak{f}(\boldsymbol{\alpha}(\mathbf{t}^*), \mathbf{t}^*) = f(\mathbf{t}^*)$. Wir nehmen an, daß eine Knotenfolge $\mathbf{t}^\dagger \in \Omega$ und Koeffizienten $\boldsymbol{\alpha}^\dagger$ existieren mit $\mathfrak{f}\left(\boldsymbol{\alpha}^\dagger, \mathbf{t}^\dagger\right) < \mathfrak{f}(\boldsymbol{\alpha}(\mathbf{t}^*), \mathbf{t}^*)$.

Nach Definition des reduzierten Problems gilt aber $\mathfrak{f}(\boldsymbol{\alpha}, \mathbf{t}) \geq f(\mathbf{t})$ für alle $\mathbf{t} \in \Omega$, denn $\mathfrak{f}(\boldsymbol{\alpha}, \mathbf{t}) := \frac{1}{2} \|\mathbf{y} - \mathbf{B}(\mathbf{t})\boldsymbol{\alpha}\|^2 + \frac{1}{2} \|\mathbf{S}_r(\mathbf{t})\boldsymbol{\alpha}\|^2$ mit „beliebigem“ $\boldsymbol{\alpha}$ und $f(\mathbf{t}) := \frac{1}{2} \|\mathbf{y} - \mathbf{B}(\mathbf{t})\boldsymbol{\alpha}(\mathbf{t})\|^2 + \frac{1}{2} \|\mathbf{S}_r(\mathbf{t})\boldsymbol{\alpha}(\mathbf{t})\|^2$ mit „optimalem“ $\boldsymbol{\alpha}(\mathbf{t})$.

Es folgt $f(\mathbf{t}^\dagger) \leq \mathfrak{f}\left(\boldsymbol{\alpha}^\dagger, \mathbf{t}^\dagger\right) < \mathfrak{f}(\boldsymbol{\alpha}(\mathbf{t}^*), \mathbf{t}^*) = f(\mathbf{t}^*)$. Dies ist ein Widerspruch zur Voraussetzung, daß $\mathbf{t}^*$ globale Minimumstelle von $f(\mathbf{t})$ in Ω ist. Also ist $(\boldsymbol{\alpha}(\mathbf{t}^*), \mathbf{t}^*)$ globale Minimumstelle von $\mathfrak{f}(\boldsymbol{\alpha}, \mathbf{t})$ für $\mathbf{t} \in \Omega$. □

Abbildung 3.1 verdeutlicht die Beziehungen zwischen Ausgangsproblem und reduziertem Problem. Theorem 3.3 stellt zusammen mit Korollar 3.6 eine vollständige Entsprechung von Theorem 2.6 auf den restringierten Fall dar.

Die Voraussetzungen (V1)–(V3) wurden schon im Anschluß an Theorem 2.6 diskutiert. Wir möchten daher nur noch einmal auf die Voraussetzungen (V4) und (V5) eingehen: Ist die strikte Konsistenzbedingung $\mathbf{L} < \mathbf{U}$ nicht erfüllt, so erhält man für einige Koeffizienten $\alpha_j^{(p)}$ Gleichheitsrestriktionen. Dies ist zwar prinzipiell möglich (Theorem 3.1 gilt für allgemeine reduzible Optimierungsprobleme, also auch für Gleichheitsrestriktionen), wurde jedoch aus Gründen der Vereinfachung in unserem Rahmen nicht untersucht. Die Ausdrücke für die Jacobi-Matrix werden in diesem Fall natürlich noch komplizierter.

Bei nichtstrikter Komplementarität der Lagrange-Parameter hat man nach dem Satz von Daniel nur die Lipschitz-Stetigkeit von $\boldsymbol{\alpha}(\mathbf{t})$ und damit $\mathbf{F}(\mathbf{t})$. In diesem nichtgenerischen Fall kann man Methoden der nichtglatten Optimierung anwenden, verliert jedoch die Äquivalenz der Probleme im Sinne von Theorem 3.3. Für Lipschitz-stetige Funktionale im $\mathbb{R}^n$ hat man nach dem Satz von Rademacher zumindest die Differenzierbarkeit fast überall.

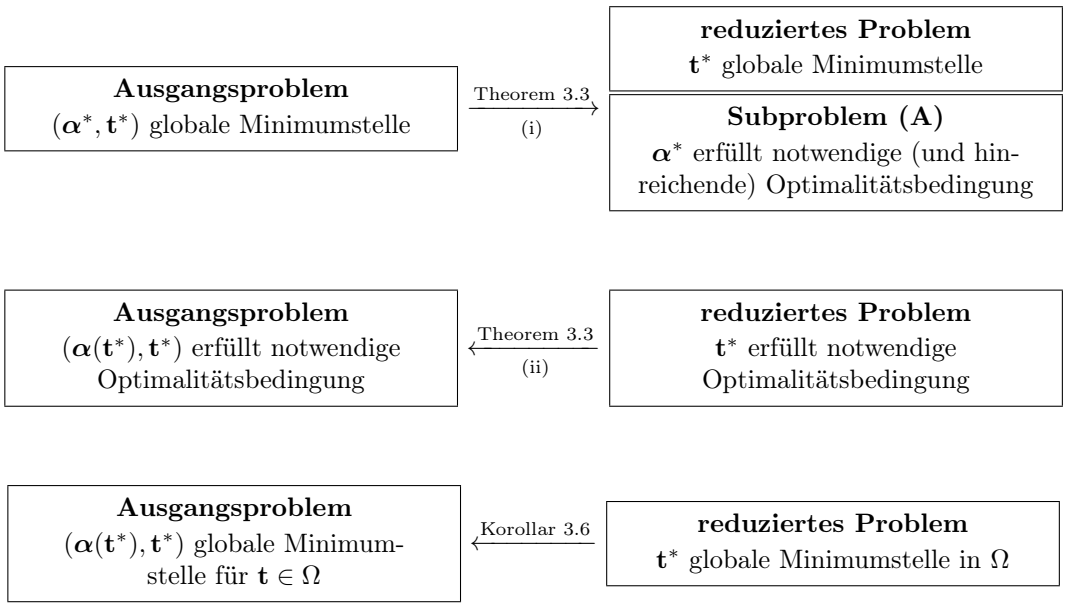


Abbildung 3.1: Äquivalenz von vollständigem restringierten Glättungsproblem FCSP und reduziertem restringierten Glättungsproblem RCSP

3.5 Numerische Lösung des reduzierten Problems

Nachdem wir die Äquivalenz von Ausgangsproblem und reduziertem Problem im Sinne von Theorem 3.3 gezeigt haben, widmen wir uns nun der numerischen Lösung des reduzierten Problems. Die Lösung des reduzierten Problems RCSP hat einige Vorteile gegenüber der Lösung des Ausgangsproblems FCSP:

- Die Anzahl unabhängiger Variabler ist l gegenüber $n + l$. Es werden keine Startwerte für $\boldsymbol{\alpha}$ benötigt.
- Die Nebenbedingungen von RCSP sind linear, während FCSP nichtlineare Nebenbedingungen besitzt.
- Existierende Software zur Lösung von Subproblem (A), vgl. [SK93], kann unmittelbar eingesetzt werden.
- Die Bandstruktur der Matrizen kann voll ausgenutzt werden.

Dem stehen die folgenden Nachteile gegenüber:

- Die Struktur von Gradient, Hesse- und Jacobi-Matrix des reduzierten Funktionals ist relativ kompliziert.
- Die Koeffizienten $\boldsymbol{\alpha}$ sind bei nichtstrikter Komplementarität lediglich Lipschitz-stetig, aber nicht stetig differenzierbar bez. $\mathbf{t}$.

Das reduzierte Problem ist erneut ein nichtlineares Quadratmittelproblem mit linearen Ungleichheitsnebenbedingungen, welches wir mit dem Basisalgorithmus 2.1 lösen können. Gegenüber dem Kapitel 2 ändert sich lediglich die Berechnung der Residuumsfunktion, also im wesentlichen die Lösung von Subproblem (A), sowie die Berechnung der Jacobi-Matrix.

3.5.1 Die Lösung von Subproblem (A)

Die Lösung von Subproblem (A)

$$\min\left\{\frac{1}{2}\left\|\mathbf{y}-\mathbf{B}\boldsymbol{\alpha}\right\|^2+\mu\frac{1}{2}\left\|\mathbf{S}_r\boldsymbol{\alpha}\right\|^2 \,:\, \mathbf{L}\le\mathbf{D}_p\boldsymbol{\alpha}\le\mathbf{U},\boldsymbol{\alpha}\in\mathbb{R}^n\right\} \tag{3.21}$$

ist wesentlich zur Berechnung der Residuumsfunktion des reduzierten Problems. Die Methode zur Lösung dieses Problems ist ausführlich in [SK93] dargelegt. Wir beschreiben kurz die wesentlichen Schritte: Zunächst wird unter Benutzung der QR-Faktorisierungen

$$\mathbf{Q}_0^T\mathbf{B}=\begin{bmatrix}\mathbf{R}_0\\ \mathbf{0}\end{bmatrix},\quad \mathbf{Q}_0\in\mathbb{R}^{m,m},\ \mathbf{R}_0\in\mathbb{R}^{n,n}\text{ obere Dreiecksmatrix} \tag{3.22}$$

$$\tilde{\mathbf{Q}}^T\begin{bmatrix}\mathbf{R}_0\\ \sqrt{\mu}\mathbf{S}_r\end{bmatrix}=\begin{bmatrix}\tilde{\mathbf{R}}\\ \mathbf{0}\end{bmatrix},\quad \tilde{\mathbf{Q}}\in\mathbb{R}^{2n-r,2n-r},\ \tilde{\mathbf{R}}\in\mathbb{R}^{n,n}\text{ obere Dreiecksmatrix} \tag{3.23}$$

das Problem (3.21) äquivalent in das Problem

$$\min\left\{\frac{1}{2}\|\tilde{\mathbf{c}}-\tilde{\mathbf{R}}\boldsymbol{\alpha}\|^2+\frac{1}{2}\|\tilde{\mathbf{d}}\|^2+\frac{1}{2}\|\mathbf{d}\|^2 \,:\, \mathbf{L}\le\mathbf{D}_p\boldsymbol{\alpha}\le\mathbf{U},\boldsymbol{\alpha}\in\mathbb{R}^n\right\}$$

transformiert, siehe Abschnitt 2.5.2. Durch Einführung der neuen Variablen

$$\boldsymbol{\beta}:=\mathbf{T}_p\boldsymbol{\alpha}=\left[\begin{array}{c|c}\multicolumn{2}{c}{\mathbf{D}_p}\\ \hline \mathbf{0} & \mathbf{I}_p\end{array}\right]\left(\begin{array}{c}\alpha_1\\ \vdots\\ \alpha_{n-p}\\ \hline \alpha_{n-p+1}\\ \vdots\\ \alpha_n\end{array}\right),$$

$\mathbf{T}_p\in\mathbb{R}^{n,n}$ reguläre obere Dreiecksmatrix, wird das Problem in ein Problem mit einfachen Schranken überführt

$$\min\left\{\frac{1}{2}\|\tilde{\mathbf{c}}-\tilde{\mathbf{R}}\mathbf{T}_p^{-1}\boldsymbol{\beta}\|^2+\frac{1}{2}\|\tilde{\mathbf{d}}\|^2+\frac{1}{2}\|\mathbf{d}\|^2 \,:\, \tilde{\mathbf{L}}\le\boldsymbol{\beta}\le\tilde{\mathbf{U}},\boldsymbol{\beta}\in\mathbb{R}^n\right\}, \tag{3.24}$$

wobei die Schranken $\mathbf{L}$, $\mathbf{U}$ geeignet ergänzt werden. Man beachte, daß die transformierte Matrix $\tilde{\mathbf{R}}\mathbf{T}_p^{-1}\in\mathbb{R}^{n,n}$ zwar obere Dreiecksform, aber keine Bandgestalt mehr hat.

Das Problem (3.24) wird schließlich mit dem Algorithmus BLS aus [Bjö96] gelöst, einer aktiven Mengenstrategie, welche speziell für den Fall von einfachen Schranken geeignet ist. Wir bemerken, daß die Bandgestalt lediglich im letzten Schritt zerstört wird, welcher jedoch nur noch von der Dimension n ist – im Gegensatz zu einer naiven Anwendung einer aktiven Mengenstrategie auf die Ausgangsformulierung (3.21), welche von der Dimension $m+n-r$ ist.

3.5.2 Die Berechnung der Jacobi-Matrix

Wir benutzen die Kaufman-Approximation an die Jacobi-Matrix, welche sich wesentlich einfacher berechnen läßt als die exakte Jacobi-Matrix (Golub/Pereyra-Modell). Im Fall der

Splineglättung erhalten wir für die Kaufman-Approximation

$$\mathbf{J}_K := \mathbf{P}^{\perp}_{\left[\begin{smallmatrix} B \\ \sqrt{\mu}S \end{smallmatrix}\right]N} \left(\mathfrak{J}_{\mathbf{t}} + \begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix} \bar{\mathbf{R}}^{+} \bar{\mathbf{\Gamma}} \right) \in \mathbb{R}^{m+n-r,l} \tag{3.25}$$

mit

$$\mathfrak{J}_{\mathbf{t}} := -\boldsymbol{\partial} \begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}) \end{bmatrix} \boldsymbol{\alpha}(\mathbf{t}) \in \mathbb{R}^{m+n-r,l}, \tag{3.26}$$

$$\mathbf{P}^{\perp}_{\left[\begin{smallmatrix} B \\ \sqrt{\mu}S \end{smallmatrix}\right]N} := \mathbf{I}_{m+n-r} - \left(\begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix} \mathbf{N} \right) \left(\begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix} \mathbf{N} \right)^{+} \in \mathbb{R}^{m+n-r,m+n-r} \tag{3.27}$$

sowie

$$\bar{\mathbf{R}} := -\begin{bmatrix} \mathbf{D}_p(\mathbf{t}) \\ -\mathbf{D}_p(\mathbf{t}) \end{bmatrix}_{i\in\mathcal{I}} \in \mathbb{R}^{nact,n}, \qquad \bar{\mathbf{\Gamma}} := -\boldsymbol{\partial} \begin{bmatrix} \mathbf{D}_p(\mathbf{t}) \\ -\mathbf{D}_p(\mathbf{t}) \end{bmatrix}_{i\in\mathcal{I}} \boldsymbol{\alpha}(\mathbf{t}) \in \mathbb{R}^{nact,l}. \tag{3.28}$$

Die Matrix $\mathbf{J}_K \in \mathbb{R}^{m+n-r,l}$ wird erneut spaltenweise berechnet, d. h. $\mathbf{J}_K\mathbf{e}^{\kappa} \in \mathbb{R}^{m+n-r}$ $(\kappa = 1, \ldots, l)$.

Wir bemerken zunächst, daß sich die Berechnung der Matrizen, welche Ableitungen bez. der freien Knoten enthalten, gegenüber dem unrestringierten Fall nicht geändert hat. Die Berechnung von $\mathfrak{J}_{\mathbf{t}}$ kann unmittelbar aus Abschnitt 2.5.3 übernommen werden, wobei jetzt jedoch $\boldsymbol{\alpha}$ das Subproblem (A) löst. Die Matrix $\bar{\mathbf{\Gamma}}$ erhalten wir aus Algorithmus 2.2, indem wir die Komponenten, welche zu aktiven Nebenbedingungen gehören, mit entsprechenden Vorzeichen versehen.

Berechnung einer Nullraumbasis N bzw. von Vektoren $\bar{\mathbf{R}}^{+}\mathbf{v}$

Zur Berechnung des orthogonalen Projektors (3.27) wird eine Basis $\mathbf{N}$ des Nullraumes von $\bar{\mathbf{R}}$ benötigt. Eng damit verbunden ist die Berechnung von Vektoren $\mathbf{x} = \bar{\mathbf{R}}^{+}\mathbf{v}$ für verschiedene $\mathbf{v} \in \mathbb{R}^{nact}$, genauer von $\mathbf{x}_{\kappa} = \bar{\mathbf{R}}^{+}\bar{\mathbf{\Gamma}}\mathbf{e}^{\kappa}$ $(\kappa = 1, \ldots, l)$. Für die Matrix $\bar{\mathbf{R}} \in \mathbb{R}^{nact,n}$ gilt $0 \le nact \le n$, $\operatorname{rank}\bar{\mathbf{R}} = nact$ und die Zeilenbandbreite beträgt $p + 1$.

Beispiel 3.5. $n = 10$, $p = 2$, $nact = 4$, $\mathcal{I} = \{4, 6, 10, 15\}$

$$\bar{\mathbf{R}} = \begin{bmatrix} & & & x & x & x & & & & \\ & & & & & x & x & x & & \\ & x & x & x & & & & & & \\ & & & & & & x & x & x & \end{bmatrix} \in \mathbb{R}^{4,10}$$

Der Vektor $\mathbf{x}$ kann als Lösung des *Minimum-Norm-Problems* $\|\mathbf{x}\| \to \min$ bei $\bar{\mathbf{R}}\mathbf{x} = \mathbf{v}$ interpretiert werden. Ein naheliegende Möglichkeit zu dessen Lösung ist die Methode der Normalgleichungen 2.Art, d. h. die Berechnung der Cholesky-Faktorisierung $\bar{\mathbf{R}}\bar{\mathbf{R}}^T = \mathbf{L}\mathbf{L}^T$, Lösung des Gleichungssystems $\mathbf{L}\mathbf{L}^T\mathbf{w} = \mathbf{v}$ und Bildung von $\bar{\mathbf{R}}^T\mathbf{w}$. Die Fehlerschranke der berechneten Lösung hängt von $\operatorname{cond}(\bar{\mathbf{R}})^2$ ab.

Ein Verfahren mit besseren Stabilitätseigenschaften erhält man mittels einer QR-Faktorisierung von $\bar{\mathbf{R}}^T$. Wegen $\mathcal{N}\left(\bar{\mathbf{R}}\right) = \mathcal{R}\left(\bar{\mathbf{R}}^T\right)^{\perp}$ kann damit gleichzeitig eine orthonormale Basis $\mathbf{N}$ des Nullraumes von $\bar{\mathbf{R}}$ berechnet werden. Sei $\mathbf{Q}_1 = (\mathbf{Q}_{11} \,|\, \mathbf{Q}_{12}) \in \mathbb{R}^{n,n}$ mit $\mathbf{Q}_{11} \in \mathbb{R}^{n,nact}$ und $\mathbf{Q}_{12} \in \mathbb{R}^{n,n-nact}$ eine orthogonale Matrix mit

$$\mathbf{Q}_1^T\bar{\mathbf{R}}^T = \begin{bmatrix} \mathbf{Q}_{11}^T \\ \mathbf{Q}_{12}^T \end{bmatrix} \bar{\mathbf{R}}^T = \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{R}_1 \in \mathbb{R}^{nact,nact} \text{ reguläre obere Dreiecksmatrix.}$$

Es gilt $\mathcal{R}\left(\bar{\mathbf{R}}^T\right)^{\perp} = \mathcal{R}\left(\mathbf{Q}_{12}\right)$, d. h. $\mathbf{N} := \mathbf{Q}_{12} \in \mathbb{R}^{n,n-nact}$ ist die gesuchte Nullraummatrix. Die Matrix $\bar{\mathbf{R}}^T$ enthält in jeder Spalte $p+1$ Nichtnullelemente, allerdings (abhängig von der Indexmenge $\mathcal{I}$) in ungeordneter Weise. Da die Matrix $\mathbf{L} = \mathbf{R}_1^T$ (bis auf Vorzeichen) der Dreiecksfaktor der Cholesky-Faktorisierung von $\bar{\mathbf{R}}\bar{\mathbf{R}}^T$ ist, können wir $\mathbf{x} = \bar{\mathbf{R}}^{+}\mathbf{v} = \bar{\mathbf{R}}^T\left(\mathbf{R}_1^T\mathbf{R}_1\right)^{-1}\mathbf{v}$ über die Methode der *Semi-Normalgleichungen* berechnen.

Algorithmus 3.1 (Berechnung von $\mathbf{x} = \bar{\mathbf{R}}^{+}\mathbf{v}$ und $\mathbf{N}$, Semi-Normalgleichungen).S1: Berechne QR-Faktorisierung mittels Householder-Transformationen

$$\begin{bmatrix} \mathbf{Q}_{11}^T \\ \mathbf{Q}_{12}^T \end{bmatrix} \bar{\mathbf{R}}^T = \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{bmatrix}, \qquad \bar{\mathbf{R}}^T \in \mathbb{R}^{n,nact}$$

Speichere den regulären oberen Dreiecksfaktor $\mathbf{R}_1 \in \mathbb{R}^{nact,nact}$

S2: Setze $\mathbf{N} := \mathbf{Q}_{12} \in \mathbb{R}^{n,n-nact}$ (Basis des Nullraumes von $\bar{\mathbf{R}}$)

S3: Für jede rechte Seite $\mathbf{v} \in \mathbb{R}^{nact}$

S3.1: Löse $\mathbf{R}_1^T\mathbf{R}_1\mathbf{w} = \mathbf{v}$; $\mathbf{w} \in \mathbb{R}^{nact}$

S3.2: Bilde $\mathbf{x} := \bar{\mathbf{R}}^T\mathbf{w} \in \mathbb{R}^n$

Bei den Semi-Normalgleichungen muß zusätzlich zu den ohnehin benötigten Größen $\mathbf{N}$ und $\mathbf{R}_1$ nur die schwach besetzte Originalmatrix $\bar{\mathbf{R}}^T$ an Stelle von $\mathbf{Q}_{11}$ gespeichert werden.

Für überbestimmte Systeme wird in [Bjö87] gezeigt, daß der Fehler der berechneten Lösung bei der Methode der Semi-Normalgleichungen ebenfalls vom Quadrat der Konditionszahl abhängt, obwohl der berechnete Dreiecksfaktor von „besserer“ Qualität als der Dreiecksfaktor der Normalgleichungen ist. Bei der von Saunders [Sau72] vorgeschlagenen Methode der Semi-Normalgleichungen für das Minimum-Norm-Problem tritt dieser Effekt nicht auf. In [Pai73] wurde gezeigt, daß „the bound on the error in $\mathbf{x}$ is proportional to κu rather than $\kappa^2 u$ as has often been thought“ (κ-Konditionszahl, u-Maschinengenauigkeit), siehe [Hig96] für eine ausführliche Diskussion dieser Problematik.

Berechnung des Projektors $\mathbf{P}_{\left[\frac{B}{\sqrt{\mu}S}\right]N}$

Abschließend betrachten wir die Berechnung des orthogonalen Projektors (3.27). Es sei bemerkt, daß der übliche Weg – nämlich die Bildung von $\left[\frac{B}{\sqrt{\mu}S}\right]N$ und anschließende orthogonale Faktorisierung – in diesem speziellen Fall sehr ineffizient ist, da die Bandstruktur von $\mathbf{B}$ und $\mathbf{S}_r$ zerstört wird. An Stelle dessen benutzen wir die zwei orthogonalen Zerlegungen (3.22) und (3.23), welche bereits bei der Lösung von Subproblem (A) berechnet wurden und die Bandstruktur erhalten.

Für spaltenreguläre Matrizen $\mathbf{A}$ gilt $\mathbf{A}^{+} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T$. Wir erhalten daher

$$\left(\begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix}\mathbf{N}\right)^{+} = \left(\mathbf{N}^T\begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix}^T\begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix}\mathbf{N}\right)^{-1}\mathbf{N}^T\begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix}^T.$$

Unter Benutzung von (3.22) und (3.23) gilt

$$\begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix}^T\begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix} = \tilde{\mathbf{R}}^T\tilde{\mathbf{R}},$$

also

$$\mathbf{P}_{\left[\begin{smallmatrix}B\\ \sqrt{\mu}S\end{smallmatrix}\right]} = \begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix} \mathbf{N} \left(\mathbf{N}^T \tilde{\mathbf{R}}^T \tilde{\mathbf{R}} \mathbf{N} \right)^{-1} \mathbf{N}^T \begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix}^T.$$

Wir berechnen nun die QR-Faktorisierung von $\tilde{\mathbf{R}}\mathbf{N}$ mittels Householder-Transformationen. Man beachte, daß $\tilde{\mathbf{R}}\mathbf{N} \in \mathbb{R}^{n,n-nact}$ als Produkt einer oberen Dreiecksmatrix und einer i. allg. vollbesetzten Nullraumbasis vollbesetzt ist.

$$\mathbf{Q}_2^T \left(\tilde{\mathbf{R}}\mathbf{N} \right) = \begin{bmatrix} \mathbf{R}_2 \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{Q}_2 \in \mathbb{R}^{n,n}, \mathbf{R}_2 \in \mathbb{R}^{n-nact,n-nact} \quad \text{reguläre obere Dreiecksmatrix}$$

Damit erhalten wir $\mathbf{N}^T \tilde{\mathbf{R}}^T \tilde{\mathbf{R}} \mathbf{N} = \mathbf{R}_2^T \mathbf{R}_2$ und schließlich

$$\begin{aligned} \mathbf{N} \left(\mathbf{N}^T \tilde{\mathbf{R}}^T \tilde{\mathbf{R}} \mathbf{N} \right)^{-1} \mathbf{N}^T &= \tilde{\mathbf{R}}^{-1} \mathbf{Q}_2 \underbrace{\begin{bmatrix} \mathbf{R}_2 \\ \mathbf{0} \end{bmatrix} \mathbf{R}_2^{-1}}_{\begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix}} \underbrace{\mathbf{R}_2^{-T} \left[\mathbf{R}_2^T | \mathbf{0} \right]}_{[\mathbf{I}|\mathbf{0}]} \mathbf{Q}_2^T \tilde{\mathbf{R}}^{-T} \\ &= \tilde{\mathbf{R}}^{-1} \mathbf{Q}_2 \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{Q}_2^T \tilde{\mathbf{R}}^{-T}. \end{aligned}$$

Lemma 3.7 (Berechnung des Projektors $\mathbf{P}_{\left[\begin{smallmatrix}B\\ \sqrt{\mu}S\end{smallmatrix}\right]N}$)**.**

Seien $\mathbf{B} \in \mathbb{R}^{m,n}$ und $\mathbf{S}_r \in \mathbb{R}^{n-r,n}$ gegebene Matrizen, so daß $\begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix} \in \mathbb{R}^{m+n-r,n}$ für $\mu > 0$ und $m \geq r$ Vollrang n besitzt. Ferner seien die folgenden QR-Faktorisierungen bekannt

$$\mathbf{Q}_0^T \mathbf{B} = \begin{bmatrix} \mathbf{R}_0 \\ \mathbf{0} \end{bmatrix}, \quad \tilde{\mathbf{Q}}^T \begin{bmatrix} \mathbf{R}_0 \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{R}} \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{Q}_2^T \left(\tilde{\mathbf{R}}\mathbf{N} \right) = \begin{bmatrix} \mathbf{R}_2 \\ \mathbf{0} \end{bmatrix}$$

wobei $\mathbf{N} \in \mathbb{R}^{n,n-nact}$ spaltenregulär sei. Dann gilt für beliebiges $\mathbf{v} \in \mathbb{R}^{m+n-r}$

$$\mathbf{P}^{\perp}_{\left[\begin{smallmatrix}B\\ \sqrt{\mu}S\end{smallmatrix}\right]N} \mathbf{v} = \left[\mathbf{I}_{m+n-r} - \begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix} \tilde{\mathbf{R}}^{-1} \mathbf{Q}_2 \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{Q}_2^T \tilde{\mathbf{R}}^{-T} \begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix}^T \right] \mathbf{v}.$$

Berechnung der Kaufman-Approximation

Zusammenfassend können wir den Algorithmus zur spaltenweisen Berechnung der Kaufman-Approximation $\mathbf{J}_K$ für die Jacobi-Matrix $\mathbf{F}'(\mathbf{t})$ angeben:

Algorithmus 3.2 (Berechnung der Jacobi-Matrix, Kaufman-Approximation).S1: Berechne QR-Faktorisierung mittels zeilenweiser Givens-Drehungen

$$\mathbf{Q}_0^T\mathbf{B} = \begin{bmatrix} \mathbf{R}_0 \\ \mathbf{0} \end{bmatrix}, \qquad \mathbf{B} \in \mathbb{R}^{m,n}$$

S2: Berechne QR-Faktorisierung mittels zeilenweiser Givens-Drehungen

$$\tilde{\mathbf{Q}}^T \begin{bmatrix} \mathbf{R}_0 \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{R}} \\ \mathbf{0} \end{bmatrix}, \qquad \begin{bmatrix} \mathbf{R}_0 \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix} \in \mathbb{R}^{2n-r,n}$$

S3: Berechne $\boldsymbol{\alpha}(\mathbf{t})$ als Lösung von

$$\min\left\{\frac{1}{2}\left\|\begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} - \begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}) \end{bmatrix}\boldsymbol{\alpha}(\mathbf{t})\right\|^2 : \mathbf{L} \le \mathbf{D}_p(\mathbf{t})\boldsymbol{\alpha} \le \mathbf{U} : \boldsymbol{\alpha} \in \mathbb{R}^n\right\}$$

unter Benutzung der QR-Faktorisierungen aus S1 und S2, siehe [SK93]

S4: Setze

$$\bar{\mathbf{R}} := -\begin{bmatrix} \mathbf{D}_p(\mathbf{t}) \\ -\mathbf{D}_p(\mathbf{t}) \end{bmatrix}_{i\in\mathcal{I}} \in \mathbb{R}^{nact,n}$$

S5: Berechne QR-Faktorisierung mittels Householder-Transformationen

$$\mathbf{Q}_1^T\bar{\mathbf{R}}^T = \begin{bmatrix} \mathbf{Q}_{11}^T \\ \mathbf{Q}_{12}^T \end{bmatrix}\bar{\mathbf{R}}^T = \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{bmatrix}, \qquad \bar{\mathbf{R}}^T \in \mathbb{R}^{n,nact}$$

S6: $\mathbf{N} := \mathbf{Q}_{12} \in \mathbb{R}^{n,n-nact}$ Basis des Nullraums von $\bar{\mathbf{R}}$

S7: Berechne QR-Faktorisierung mittels Householder-Transformationen

$$\mathbf{Q}_2^T\left(\tilde{\mathbf{R}}\mathbf{N}\right) = \begin{bmatrix} \mathbf{R}_2 \\ \mathbf{0} \end{bmatrix}, \qquad \tilde{\mathbf{R}}\mathbf{N} \in \mathbb{R}^{n,n-nact}$$

S8: for $\kappa := 1$ to l do

S8.1: Berechne $\mathbf{v}^1 := \mathfrak{J}_{\mathbf{t}}\mathbf{e}^\kappa = -\boldsymbol{\partial}\begin{bmatrix} \mathbf{B}(\mathbf{t}) \\ \sqrt{\mu}\mathbf{S}_r(\mathbf{t}) \end{bmatrix}[\mathbf{e}^\kappa]\,\boldsymbol{\alpha}(\mathbf{t}) \in \mathbb{R}^{m+n-r}$;

S8.2: Berechne $\mathbf{v}^2 := \bar{\boldsymbol{\Gamma}}\mathbf{e}^\kappa = -\boldsymbol{\partial}\bar{\mathbf{R}}(\mathbf{t})\,[\mathbf{e}^\kappa]\,\boldsymbol{\alpha}(\mathbf{t}) \in \mathbb{R}^{nact}$;

S8.3: Löse das System $\mathbf{R}_1^T\mathbf{R}_1\mathbf{v}^3 = \mathbf{v}^2$; ($\mathbf{R}_1$ reguläre obere Dreiecksmatrix, $\mathbf{v}^3 \in \mathbb{R}^{nact}$);

S8.4: $\mathbf{v}^4 := \bar{\mathbf{R}}^T\mathbf{v}^3 \in \mathbb{R}^n$;

S8.5: $\mathbf{v}^5 := \begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix}\mathbf{v}^4 \in \mathbb{R}^{m+n-r}$;

S8.6: $\mathbf{v}^6 := \mathbf{v}^1 + \mathbf{v}^5 \in \mathbb{R}^{m+n-r}$;

S8.7: Berechne $\mathbf{v}^7 \in \mathbb{R}^{m+n-r}$

$$\mathbf{v}^7 = \left[\mathbf{I}_{m+n-r} - \begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix}\tilde{\mathbf{R}}^{-1}\mathbf{Q}_2\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\mathbf{Q}_2^T\tilde{\mathbf{R}}^{-T}\begin{bmatrix} \mathbf{B} \\ \sqrt{\mu}\mathbf{S}_r \end{bmatrix}^T\right]\mathbf{v}^6;$$

S8.8: $\mathbf{J}_K\mathbf{e}^\kappa := \mathbf{v}^7$;

Wenn die Kaufman-Approximation in der obigen Weise berechnet wird, so nennen wir das Verfahren RCSP-Ka-ED (reduced constrained smoothing problem, Kaufman model, exact derivatives). Analog verfahren wir bei der Berechnung über finite Differenzen RCSP-GP-OD (reduced constrained smoothing problem, Golub/Pereyra model, outer discretization).

3.6 Numerische Tests

In diesem Abschnitt wollen wir die Fähigkeiten des entwickelten Verfahrens an einigen Beispielen aus der Literatur demonstrieren. Wie schon im unrestringierten Fall haben wir sowohl das Kaufman-Modell als auch das Golub/Pereyra-Modell innerhalb des Programmpaketes FREE implementiert.

3.6.1 Titanium Heat Data

Unser erstes Beispiel sind die wohlbekannten *Titanium Heat Data*. Diesmal wollen wir die $m = 49$ Datenpunkte durch $n = 11$ B-Splines der Ordnung $k = 4$ approximieren. Wir verwenden die Ordnung $r = 2$ im Glättungsterm und einen Glättungsparameter $\mu = 1.0$, welchen wir interaktiv bestimmt haben. Wir fixieren die Knoten $\tau_7 = 835$ und $\tau_{10} = 955$, d. h. wir haben schließlich $l = 5$. Die verbleibenden freien Knoten τ_5 und τ_6 bzw. τ_8, τ_9 und τ_{11} verteilen wir äquidistant in den Intervallen $[595, 835]$, $[835, 955]$ und $[955, 1075]$. Wir fordern die Konvexität des Splines in den Intervallen $[595, 835)$ und $[955, 1075)$ und erhalten

$$\begin{aligned}\mathbf{L} &= (0, 0, 0, 0, -\infty, -\infty, 0, 0, 0)^T \\ \mathbf{U} &= (+\infty, +\infty, +\infty, +\infty, +\infty, +\infty, +\infty, +\infty, +\infty)^T.\end{aligned}$$

Tabelle 3.1 zeigt das Ergebnis der Splineglättung mit freien Knoten unter diesen Konvexitätsnebenbedingungen. Dabei bezeichnet RSP-Ka-ED eine Methode aus dem zweiten Kapitel. In diesem Fall haben wir zuerst die Lage der Knoten ohne Berücksichtigung der formerhaltenden Nebenbedingungen optimiert. Abschließend wurde zu diesen festen Knoten ein formerhaltender Spline berechnet.

	$\mathbf{t}^0$	RCSP-Ka-ED	RCSP-GP-OD	RSP-Ka-ED
τ_5	675.0	7.975133 E+02	7.822991 E+02	5.959958 E+02
τ_6	755.0	8.110142 E+02	7.947857 E+02	6.109336 E+02
τ_8	875.0	8.751572 E+02	8.755310 E+02	8.767428 E+02
τ_9	915.0	8.810366 E+02	8.804978 E+02	8.816339 E+02
τ_{11}	1015.0	9.625000 E+02	9.625000 E+02	9.625000 E+02
$\|\mathbf{F}\|$	1.027722 E+00	3.469246 E-01	3.460394 E-01	3.544604 E-01
Schritte		7	13	9
Zeit [s]		0.269	0.504	0.239
$\|\mathbf{F}^T\mathbf{J}\mathbf{s}\|$		2.491557 E-03	3.592591 E-11	5.363720 E-10
$\|\mathbf{J}^T\mathbf{F}\|$		2.797501 E-03	2.988107 E-03	1.749176 E-03
Ret. Code		4	3	4

Tabelle 3.1: Titanium Heat Data: Glättung mit Nebenbedingungen ($\mu = 1.0$)

Die Abbildungen 3.2 und 3.3 zeigen den Graph des Splines vor und nach der Optimierung. Die Verbesserung sowohl in der Approximationsgüte als auch dem Aussehen ist deutlich erkennbar. In der Abbildung 3.4 wird schließlich die zweite Ableitung des Splines gezeigt.

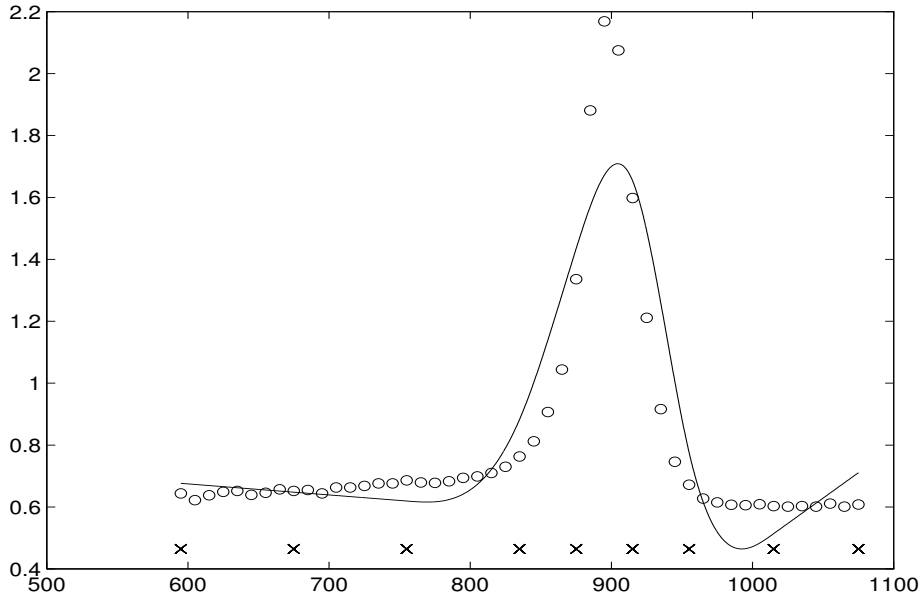


Abbildung 3.2: Titanium Heat Data: Spline s, Startknotenfolge

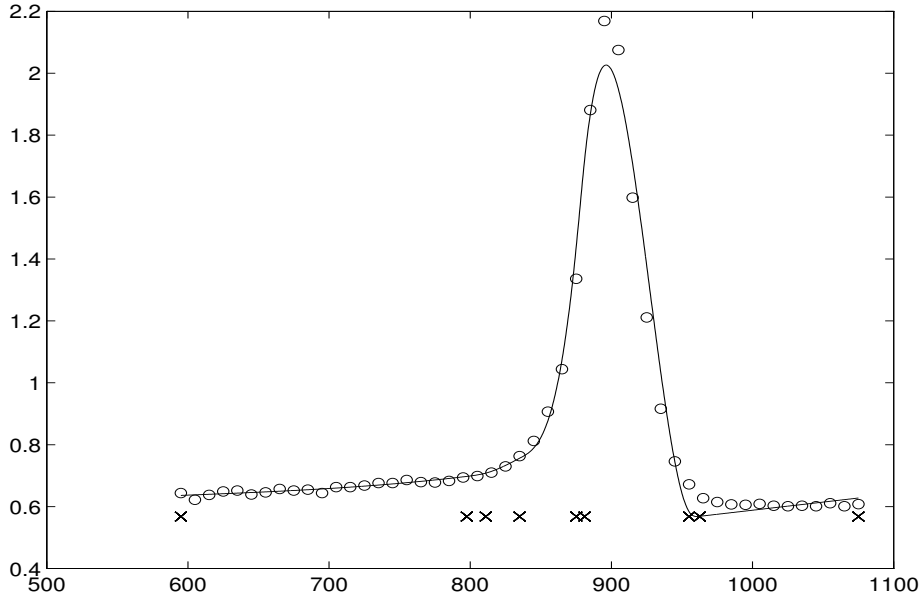


Abbildung 3.3: Titanium Heat Data: Spline s, Optimierte Knotenfolge, RCSP-Ka-ED

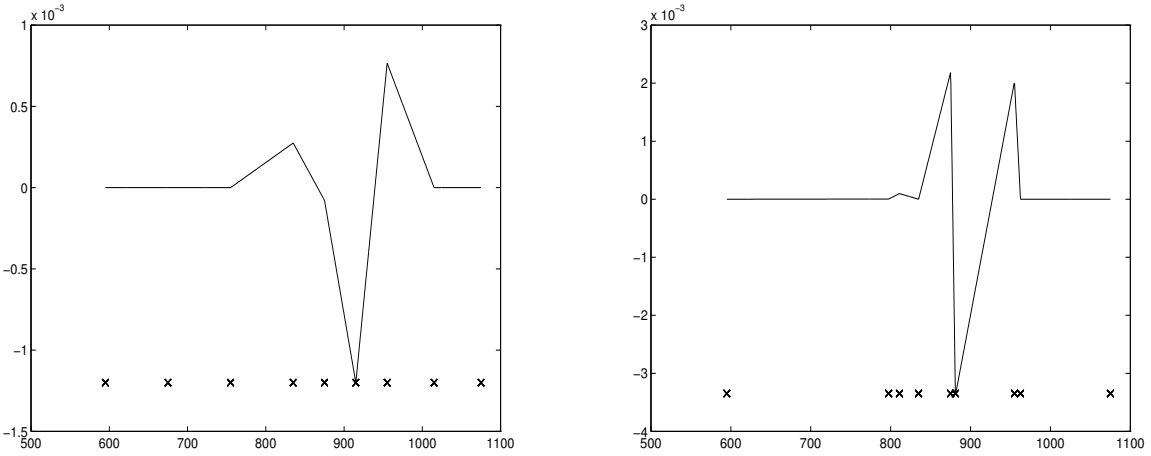


Abbildung 3.4: Titanium Heat Data: s'' zur Startknotenfolge und zur optimierten Knotenfolge (RCSP-Ka-ED)

Wir wollen nun unseren Algorithmus mit einer adaptiven Strategie zur Knotenplazierung bei kubischen Splines unter Konvexitätsnebenbedingungen vergleichen, welche in der Routine CONCON des weitverbreiteten FITPACK-Paketes [Die87] implementiert wurde, siehe auch [Die80] und [Die93]. In dieser Routine werden ausgehend von einer weniger dichten Knotenfolge schrittweise Knoten eingefügt, so daß die formerhaltenden Nebenbedingungen erfüllt sind und das Residuum unter einer vorgegebenen Schranke S ist. Man beachte, daß das Ziel von CONCON nicht die „optimale“ Plazierung der Knoten ist, sondern eine schnelle und „gute“ automatische Knotenwahl.

Gibt man die Schranke $S = \|\mathbf{F}\|^2 = 0.25$ für den Quadratmittelfehler vor und benutzt obige Nebenbedingungen, so berechnet CONCON einen Spline mit $n = 11$ B-Splines, siehe Tabelle 3.2 für die Lage der inneren Knoten. Wir wollen erneut einen formerhaltenden Spline mittels unseres Algorithmus berechnen. Man beachte, daß es nicht erforderlich ist, die Intervalle, auf denen die formerhaltenden Nebenbedingungen vorgegeben sind, fix vorzugeben. Deshalb verlangen wir nun $s''(x) \geq 0$ für alle $x \in [595, t_7), [t_{11}, 1075)$, wobei *alle* $l = 7$ innere Knoten freie Knoten sind. Unser Algorithmus wurde mit äquidistanten inneren Knoten gestartet. Das Ergebnis ist in Tabelle 3.2 und Abbildung 3.5 zusammengefaßt.

Methode	τ_5	τ_6	τ_7	τ_8	τ_9	τ_{10}	τ_{11}	$\|\mathbf{F}\|$
CONCON	715.0	835.0	865.0	875.0	895.0	925.0	955.0	1.11664 E-01
RCAP-Ka-ED	777.8	819.7	864.8	867.8	897.9	916.7	974.3	5.72718 E-02

Tabelle 3.2: Titanium Heat Data: CONCON — RCAP-Ka-ED

Das obige Beispiel zeigt, daß der Algorithmus auch sehr gut zur automatischen Bestimmung von Wendepunkten eingesetzt werden kann. Über diese aktuelle Problematik gibt es – aus statistischer Sicht – einige interessante theoretische Arbeiten, siehe [Rie95], jedoch noch keinerlei numerische Erfahrungen.

Man beachte, daß die derzeit frei per Netlib[2] verfügbare Version von CONCON auf Grund

[2] `http://www.netlib.org/dierckx`

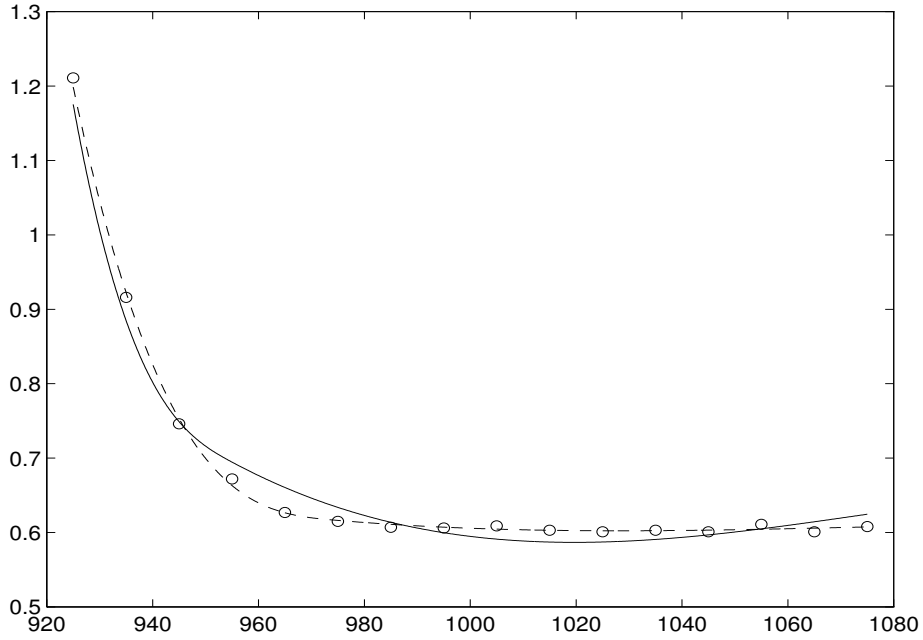


Abbildung 3.5: Titanium Heat Data: CONCON (—), RCAP-Ka-ED (- - -), $x \in [925, 1075]$

eines Fehlers in der Knotenplazierungsstrategie bei diesem Beispiel nach zwei Knoten mit einer Fehlermeldung abbricht, da „adding one or more knots will not further reduce the value of SQ (sum of squared residuals)“. Vom Autor P. Dierckx ist eine korrigierte Version erhältlich.

Verkleinert man in dem Titan-Beispiel schrittweise die obere Schranke S, um für verschiedene n die optimale Lage der Knoten zu bestimmen, so liefert CONCON selbst nach sehr langer Zeit (mehrere Tage) keine Lösung. Grund dafür scheint der Algorithmus zur Lösung des zugrundeliegenden quadratischen Optimierungsproblems, die Theil/van de Panne-Prozedur, zu sein.

3.6.2 Arctan-Daten

In einem zweiten Beispiel möchten wir die Splineglättung durch monotone Splines mit freien Knoten illustrieren. Wir betrachten Daten, welche durch Auswertung der Funktion $g(x) = \arctan(10x)$ an $m = 41$ äquidistanten Punkten x_i im Intervall $[-10, 10]$ entstanden sind. Unter Benutzung von Pseudozufallszahlen ϵ_i, $-0.075 \leq \epsilon_i \leq 0.075$ erzeugen wir die gestörten Werte $y_i = g(x_i)(1 + \epsilon_i)$, $i = 1, \ldots, m$.

Wir approximieren diese Daten durch $n = 8$ B-Splines der Ordnung $k = 4$ und fordern, daß der Spline s monoton in $[-10, 10]$ ist, d. h.

$$\mathbf{L} = (0, 0, 0, 0, 0, 0, 0)^T$$
$$\mathbf{U} = (+\infty, +\infty, +\infty, +\infty, +\infty, +\infty, +\infty)^T$$

Wir wählen $l = 4$ äquidistante innere Knoten als Startpunkt unseres Verfahrens. Die Tabellen 3.3 und 3.4 zeigen die Ergebnisse der Splineglättung und Approximation. Im Fall der Glättung haben wir $\mu = 1.0\,\text{E-}03$ und $r = 2$ verwendet.

	$\mathbf{t}^0$	RCSP-Ka-ED	RCSP-GP-OD	RSP-Ka-ED
τ_5	−6.0	−8.760175 E-01	−9.652265 E-01	−8.838273 E-01
τ_6	−2.0	−2.301281 E-01	−3.258263 E-01	−2.760825 E-01
τ_7	2.0	2.743868 E-01	3.454588 E-01	2.779473 E-01
τ_8	6.0	8.822404 E-01	9.809856 E-01	1.090103 E-01
$\|\mathbf{F}\|$	2.359790 E+00	5.230920 E-01	5.098921 E-01	5.283729 E-01
Schritte		4	100	13
Zeit [s]		0.116	3.259	0.257
$\|\mathbf{F}^T\mathbf{J}\mathbf{s}\|$		8.407664 E-04	4.345827 E-06	1.0626114 E-10
$\|\mathbf{J}^T\mathbf{F}\|$		4.336670 E-02	1.270197 E-03	7.8463565 E-03
Ret. Code		4	6	5

Tabelle 3.3: Arctan-Daten: Splineglättung ($\mu = 1.0$ E-03)

	$\mathbf{t}^0$	RCAP-Ka-ED	RCAP-GP-OD	RAP-Ka-ED
τ_5	−6.0	−8.100201 E-01	−8.074899 E-01	−6.794899 E-01
τ_6	−2.0	−1.689546 E-01	−1.946541 E-01	−5.812258 E-02
τ_7	2.0	1.598707 E-01	1.855516 E-01	5.200732 E-03
τ_8	6.0	8.773101 E-01	8.745770 E-01	9.550505 E-01
$\|\mathbf{F}\|$	2.359717 E+00	4.268960 E-01	4.268960 E-01	4.423301 E-01
Schritte		6	8	8
Zeit [s]		0.135	0.202	0.125
$\|\mathbf{F}^T\mathbf{J}\mathbf{s}\|$		1.664365 E-14	7.873466 E-09	1.571400 E-12
$\|\mathbf{J}^T\mathbf{F}\|$		1.727962 E-07	7.822950 E-07	1.278808 E-02
Ret. Code		3	4	3

Tabelle 3.4: Arctan-Daten: Splineapproximation ($\mu = 0$)

Wie in Abbildung 3.6 beobachtet werden kann, wird die schnelle Änderung der Krümmung um den Nullpunkt herum durch äquidistante Knoten nicht richtig wiedergegeben. Dies erkennt man deutlich in Abbildung 3.7, in der die erste Ableitung des Splines s und der zugrundeliegenden Funktion g gezeigt wird. Durch die Optimierung der Lage der Knoten wird das Residuum um 80% reduziert. Obwohl die Residuen von Approximation und Glättung für den Startpunkt fast identisch sind, kann man nach der Optimierung einen wesentlichen Unterschied feststellen. Dies wird verursacht durch den größeren Einfluß des Glättungsterms für nichtäquidistante Knoten.

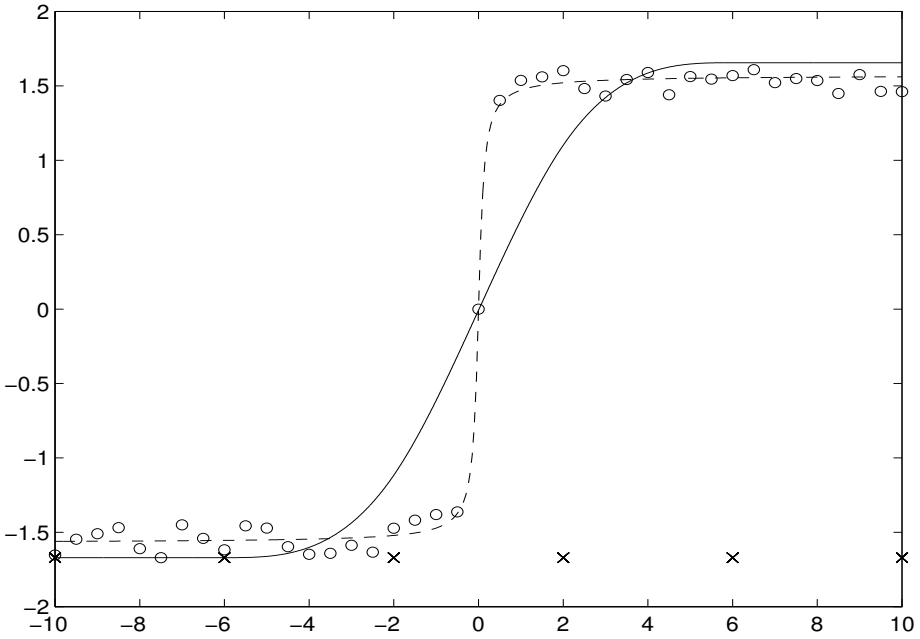


Abbildung 3.6: Arctan-Daten: Spline s (—) und Funktion g (- - -), Startknotenfolge

3.6.3 Volumetric moisture content data

Als letztes Beispiel haben wir einen Datensatz benutzt, welcher im Testprogramm des FITPACK-Paketes Verwendung findet, siehe [Die87, S. 79ff] und [Die93, S. 129f]. Geben wir die Schranke $S = \|\mathbf{F}\|^2 = 0.0002$ für den Quadratmittelfehler vor, so berechnet CONCON eine konkave Approximation mit $n = 7$ kubischen Splines mit dem Residuum $\|\mathbf{F}\| = 0.012097$ zu den $m = 16$ Datenpunkten. Unser Algorithmus RCAP-Ka-ED wurde mit äquidistanten Knoten gestartet und liefert $\|\mathbf{F}\| = 0.010675$, siehe Tabelle 3.5. Die Qualität beider Approximationen ist vergleichbar, wie Abbildung 3.9 zeigt.

Die numerischen Tests zeigen, daß unsere Methode ein leistungsfähiger Algorithmus zur Berechnung von formerhaltenden Splines mit freien Knoten ist. Der Algorithmus liefert auch im Vergleich mit der vielbenutzten CONCON-Routine aus dem FITPACK-Paket sehr gute Ergebnisse, erweitert jedoch dessen Funktionalität, d. h. allgemeinere Nebenbedingungen als Konvexität-Konkavität, beliebige Splineordnung und Einbeziehung eines Glättungsterms. Im Gegensatz zu CONCON minimiert unser Algorithmus direkt das Schoenberg-Funktional. Da die Schwachbesetztheitsstruktur der Matrizen soweit als möglich ausgenutzt wird, ist der Algorithmus auch für größere Datenmengen effektiv.

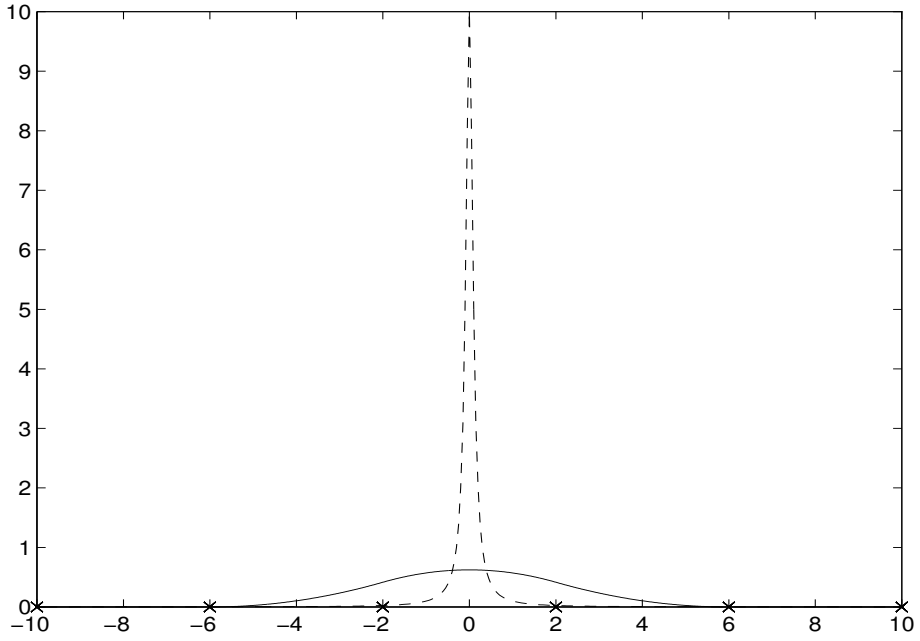


Abbildung 3.7: Arctan-Daten: Erste Ableitungen s' (—) und g' (- - -), Startknotenfolge

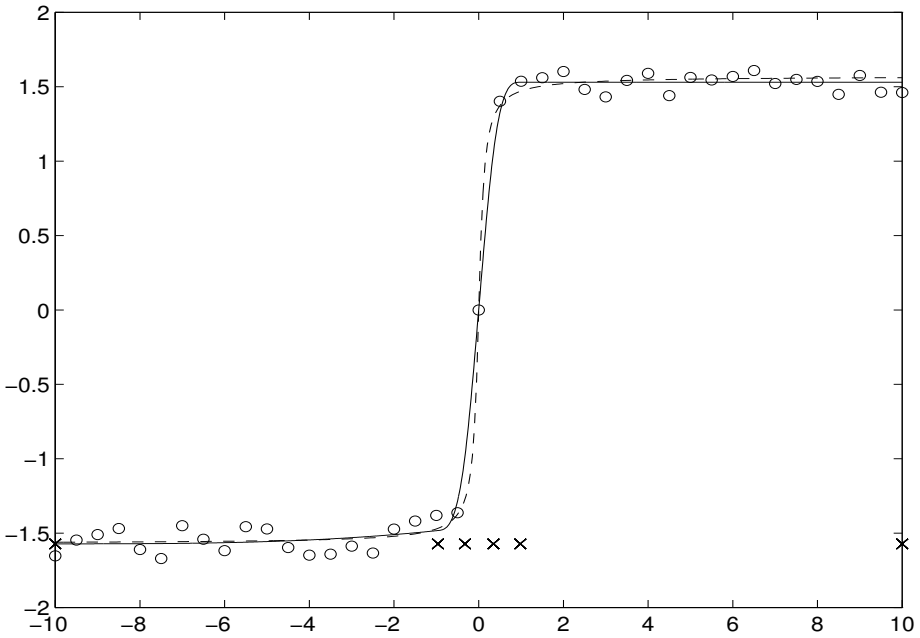


Abbildung 3.8: Arctan-Daten: Spline s (—) und Funktion g (- - -), optimierte Knoten, RCAP-GP-OD

	$\mathbf{t}^0$	CONCON	RCAP-Ka-ED
τ_5	2.45	0.30	0.14
τ_6	4.80	0.70	0.83
τ_7	7.15	2.25	4.01
$\|\mathbf{F}\|$	0.064072	0.012709	0.010675

Tabelle 3.5: Volumetric Moisture Content Data

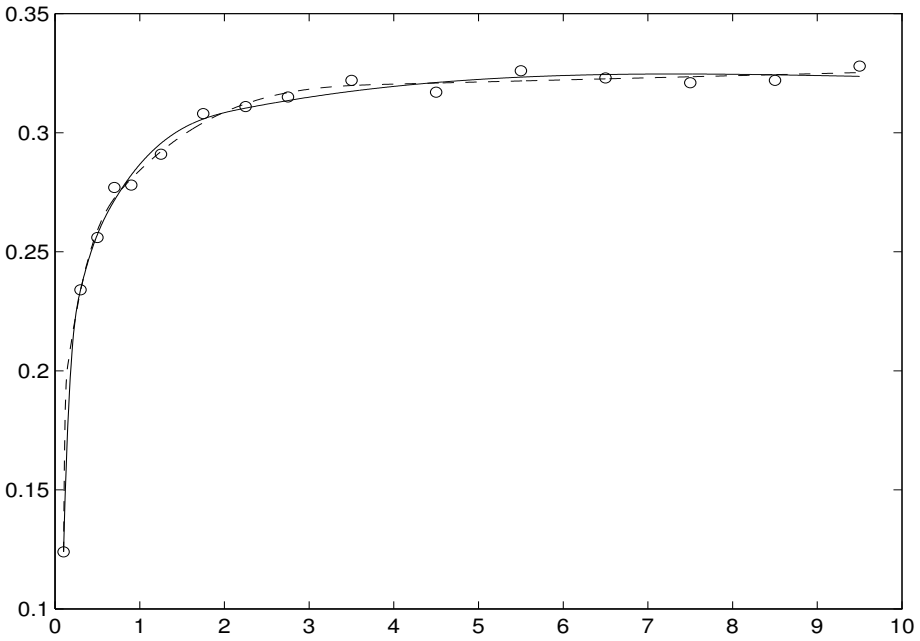


Abbildung 3.9: Volumetric Moisture Content Data: CONCON (—), RCAP-Ka-ED (- - -)

Kapitel 4

Bivariate Tensorprodukt-Splines

4.1 Einleitung und Problemstellung

In diesem Kapitel verallgemeinern wir die Ergebnisse der Quadratmittelapproximation durch univariate Splines mit freien Knoten auf den Fall der Approximation mit bivariaten Tensorprodukt-Splines bei Daten auf Rechteckgittern.

Seien $\mathbf{Z} = \{z_{i_1,i_2} : i_1 = 1,\dots,m_1; i_2 = 1,\dots,m_2\}$ fehlerbehaftete Meßwerte einer unbekannten Funktion $g \in W_2^{q_1,q_2}[a_1,b_1]\times[a_2,b_2]$, welche auf einem Rechteckgitter $[x_1,\dots,x_{m_1}]\times[y_1,\dots,y_{m_2}]$ gegeben sind, d. h. es gilt

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

mit den Abszissen

$$\begin{aligned} a_1 = x_1 < \cdots < x_{m_1} = b_1 \\ a_2 = y_1 < \cdots < y_{m_2} = b_2 \end{aligned}$$

und den Fehlern ε_{i_1,i_2}. Die stochastischen Fehler ε_{i_1,i_2} seien unabhängig und identisch verteilt.

Diese Daten wollen wir durch Tensorprodukt-Splines approximieren. Diese Splines haben einerseits eine einfache Gestalt und gestatten andererseits, das zweidimensionale Problem durch zwei Folgen eindimensionaler Probleme zu lösen, sofern die Daten auf einem Rechteckgitter liegen. Außerdem übertragen sich durch Tensorprodukt-Darstellungen viele Eigenschaften des Eindimensionalen auf das Zweidimensionale.

Seien $\mathcal{S}_1$ und $\mathcal{S}_2$ lineare Räume reellwertiger Funktionen, die auf den Mengen $\mathbf{X}$ und $\mathbf{Y}$ definiert sind. Das *Tensorprodukt zweier Funktionen* $s^1 \in \mathcal{S}_1$ und $s^2 \in \mathcal{S}_2$ ist eine Funktion auf $\mathbf{X}\times\mathbf{Y}$, die durch

$$s^1 \otimes s^2(x,y) := s^1(x)\cdot s^2(y) \qquad x \in \mathbf{X}, y \in \mathbf{Y} \tag{4.1}$$

definiert ist. Das *Tensorpodukt* $\mathcal{S}_1 \otimes \mathcal{S}_2$ *der linearen Funktionenräume* $\mathcal{S}_1$ und $\mathcal{S}_2$ wird dann als die Menge aller endlichen Linearkombinationen der Form (4.1) definiert, d. h.

$$\mathcal{S}_1 \otimes \mathcal{S}_2 := \left\{ \sum_{j=1}^{n} \alpha_j \left(s_j^1 \otimes s_j^2\right) \;:\; \alpha_j \in \mathbb{R}, s_j^1 \in \mathcal{S}_1, s_j^2 \in \mathcal{S}_2, j = 1,\dots,n \right\}.$$

Die Extremaleigenschaften eindimensionaler Splines kann man sowohl auf interpolierende als auch glättende Tensorprodukt-Splines übertragen. So ist etwa der natürliche bikubische glättende Spline die Lösung des Variationsproblems

$$\min \Bigg\{ \int_{a_1}^{b_1} \int_{a_2}^{b_2} [D^{2,2}s(x,y)]^2 \, dx\, dy + \mu_1 \sum_{i_1=1}^{m_1} \int_{a_2}^{b_2} [D^{0,2}s(x_{i_1},y)]^2 \, dy + \mu_2 \sum_{i_2=1}^{m_2} \int_{a_1}^{b_1} [D^{2,0}s(x,y_{i_2})]^2 \, dx + \mu_1\mu_2 \sum_{i_1=1}^{m_1} \sum_{i_2=1}^{m_2} [z_{i_1,i_2} - s(x_{i_1},y_{i_2})]^2 \Bigg\} \tag{4.2}$$

für $s \in W_2^{2,2}[a_1,b_1] \times [a_2,b_2]$, siehe [HS85]. Der Operator D^{r_1,r_2} bezeichnet die partielle Ableitung der Ordnung r_1 bez. der Variablen x und der Ordnung r_2 bez. der Variablen y. Die Parameter $\mu_1 > 0$ und $\mu_2 > 0$ heißen Glättungsparameter. Die obige Extremalcharakterisierung hat den Nachteil, daß die Knoten mit den Datenstellen zusammenfallen müssen. Es ist also keine Datenreduktion möglich.

Aus diesem Grund verwenden wir die Quadratmittelapproximation durch Tensorprodukt-B-Splines, d. h. die unbekannte Funktion g soll durch einen bivariaten Spline $s \in \mathcal{S}_{k_1,\boldsymbol{\tau}^1} \otimes \mathcal{S}_{k_2,\boldsymbol{\tau}^2}$ approximiert werden. Der Raum $\mathcal{S}_{k_1,\boldsymbol{\tau}^1} \otimes \mathcal{S}_{k_2,\boldsymbol{\tau}^2}$ sei dabei das Tensorprodukt der univariaten Splineräume $\mathcal{S}_{k_1,\boldsymbol{\tau}^1}$ und $\mathcal{S}_{k_2,\boldsymbol{\tau}^2}$ der polynomialen Splines der Ordnung k_1 bzw. k_2 zur Knotenfolge $\boldsymbol{\tau}^1$ bzw. $\boldsymbol{\tau}^2$. Die Knotenfolgen seien wie folgt

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

Die Parameter des Splines s sollen so gewählt werden, daß der Quadratmittelfehler

$$\varphi(s) := \frac{1}{2} \sum_{i_1=1}^{m_1} \sum_{i_2=1}^{m_2} [z_{i_1,i_2} - s(x_{i_1},y_{i_2})]^2 \tag{4.3a}$$

minimiert wird. Es ist bekannt, daß die stets existierende Lösung von (4.3a) *nur* dann eindeutig ist, wenn die Daten die Schoenberg-Whitney-Bedingung erfüllen. Im Falle beliebiger Daten ist die Eindeutigkeit i. allg. nicht gesichert. Diesem Problem wird wiederum durch Verwendung von glättenden Splines begegnet.

Durch die Verwendung des „thin plate functionals“

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

als Glättungsterm, welches physikalisch eine mittlere Biegeenergie verkörpert, ist die Eindeutigkeit unabhängig von den Daten gesichert. Ein Nachteil der Verwendung des „thin plate functionals“ ist, daß das Funktional keine Tensorprodukt-Struktur besitzt und somit im Fall von Daten auf Rechteckgittern keine Separation der Lösung mehr möglich ist.

Durch Verwendung eines gegenüber (4.2) geringfügig abgeänderten Glättungsterms erreichen wir wieder eine Separation in eine Folge univariater Probleme. Wir betrachten das

Funktional ϕ mit

$$\begin{aligned}(4.3b)\quad \phi(s) := \frac{1}{2}\sum_{i_1=1}^{m_1}\sum_{i_2=1}^{m_2}\left[z_{i_1,i_2} - s(x_{i_1},y_{i_2})\right]^2 + \mu_1\frac{1}{2}\sum_{i_2=1}^{m_2}\int_{a_1}^{b_1}\left[D^{r_1,0}s(x,y_{i_2})\right]^2\,dx \\ + \mu_2\frac{1}{2}\sum_{i_1=1}^{m_1}\int_{a_2}^{b_2}\left[D^{0,r_2}s(x_{i_1},y)\right]^2\,dy + \mu_1\mu_2\frac{1}{2}\int_{a_1}^{b_1}\int_{a_2}^{b_2}\left[D^{r_1,r_2}s(x,y)\right]^2\,dy\,dx.\end{aligned}$$

Das Glättungsfunktional (4.3b) – eine Verallgemeinerung des univariaten Schoenberg-Funktionals – wurde dabei so konstruiert, daß die entstehenden Quadratmittelprobleme in eine Folge von univariaten Problemen zerfallen. Es ist zwar nicht mehr so schön physikalisch interpretierbar, wir benutzen den Glättungsterm jedoch vorwiegend zur Regularisierung. Die Verwendung eines separablen Glättungsterms kann man historisch weit zurückverfolgen. So wird in [Die81] zunächst ein nichtzerfallender Glättungsterm verwendet, bevor in [Die82] der passende zerfallende Term benutzt wird. Nachdem in [HS85] natürliche bikubische glättende Splines untersucht wurden, untersuchen die Autoren in [HS86] komplette glättende Splines. Der abstrakte Fall von interpolierenden und glättenden Tensorprodukt-Splines – aus dem sich viele der bisher betrachteten Spezialfälle ableiten lassen – wird in [EMM89] behandelt. In den Arbeiten [Mul90], [Bre90] und [Pig91] wird schließlich der Fall der Splineapproximation (4.3b) untersucht. Man beachte, daß in letzteren Arbeiten – im Gegensatz zu den Variationszugängen – eine Datenreduktion möglich ist, da Splineknoten und Datenstellen unabhängig voneinander gewählt werden können.

4.1.1 Darstellung des Zielfunktionals

Als Basis für die univariaten Splineräume $\mathcal{S}_{k_1,\boldsymbol{\tau}^1}$ und $\mathcal{S}_{k_2,\boldsymbol{\tau}^2}$ benutzen wir die polynomialen B-Splines der Ordnung k_1 bzw. k_2 zur Knotenfolge $\boldsymbol{\tau}^1$ bzw. $\boldsymbol{\tau}^2$. Sie seien analog zu Definition 2.2 gebildet und bezeichnet mit $B_{j_1,k_1,\boldsymbol{\tau}^1}$ $(j_1 = 1,\ldots,n_1)$ bzw. $B_{j_2,k_2,\boldsymbol{\tau}^2}$ $(j_2 = 1,\ldots,n_2)$. Damit erhalten wir die Darstellung

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

mit den Splinekoeffizienten α_{j_1,j_2} für einen Tensorprodukt-Spline $s \in \mathcal{S}_{k_1,\boldsymbol{\tau}^1} \otimes \mathcal{S}_{k_2,\boldsymbol{\tau}^2}$. Setzt man dies in das Zielfunktional ein, so erhält man die folgenden Darstellungen

$$(4.4a)\qquad \frac{1}{2}\sum_{i_1=1}^{m_1}\sum_{i_2=1}^{m_2}\left[z_{i_1,i_2} - \sum_{j_1=1}^{n_1}\sum_{j_2=1}^{n_2} B_{j_1,k_1,\boldsymbol{\tau}^1}(x_{i_1})B_{j_2,k_2,\boldsymbol{\tau}^2}(y_{i_2})\alpha_{j_1,j_2}\right]^2 \to \min_{\alpha_{j_1,j_2}}$$

$$\begin{aligned}(4.4b)\quad &\frac{1}{2}\sum_{i_1=1}^{m_1}\sum_{i_2=1}^{m_2}\left[z_{i_1,i_2} - \sum_{j_1=1}^{n_1}\sum_{j_2=1}^{n_2} B_{j_1,k_1,\boldsymbol{\tau}^1}(x_{i_1})B_{j_2,k_2,\boldsymbol{\tau}^2}(y_{i_2})\alpha_{j_1,j_2}\right]^2 \\ &\quad + \mu_1\frac{1}{2}\sum_{i_2=1}^{m_2}\int_{a_1}^{b_1}\left[\sum_{j_1=1}^{n_1}\sum_{j_2=1}^{n_2} B^{(r_1)}_{j_1,k_1,\boldsymbol{\tau}^1}(x)B_{j_2,k_2,\boldsymbol{\tau}^2}(y_{i_2})\alpha_{j_1,j_2}\right]^2 dx \\ &\quad + \mu_2\frac{1}{2}\sum_{i_1=1}^{m_1}\int_{a_2}^{b_2}\left[\sum_{j_1=1}^{n_1}\sum_{j_2=1}^{n_2} B_{j_1,k_1,\boldsymbol{\tau}^1}(x_{i_1})B^{(r_2)}_{j_2,k_2,\boldsymbol{\tau}^2}(y)\alpha_{j_1,j_2}\right]^2 dy\end{aligned}$$

$$+ \mu_1\mu_2 \frac{1}{2} \int_{a_1}^{b_1} \int_{a_2}^{b_2} \left[B^{(r_1)}_{j_1,k_1,\boldsymbol{\tau}^1}(x) B^{(r_2)}_{j_2,k_2,\boldsymbol{\tau}^2}(y) \alpha_{j_1,j_2} \right]^2 dy\, dx \to \min_{\alpha_{j_1,j_2}} .$$

Matrixformulierung

Zur besseren Übersichtlichkeit gehen wir jetzt zur Matrixnotation über. Wir vereinbaren zunächst die folgenden Bezeichnungen:

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

Für den Spline s erhalten wir die Matrixdarstellung $s(x,y) = \boldsymbol{\beta}^1(x,\boldsymbol{\tau}^1)^T \mathbf{A} \boldsymbol{\beta}^2(y,\boldsymbol{\tau}^2)$ und für das Quadratmittelproblem (4.4a) unter Benutzung der Frobeniusnorm

$$\frac{1}{2} \left\| \mathbf{Z} - \mathbf{B}_1(\boldsymbol{\tau}^1) \mathbf{A} \mathbf{B}_2(\boldsymbol{\tau}^2)^T \right\|_F^2 \to \min_{\mathbf{A} \in \mathbb{R}^{n_1,n_2}} . \tag{4.5a}$$

Verwenden wir die Glättungsmatrizen

$$\bar{\mathbf{S}}^1_{r_1}(\boldsymbol{\tau}^1) := \bar{\mathbf{F}}^1_{r_1}(\boldsymbol{\tau}^1) \mathbf{D}^1_{r_1}(\boldsymbol{\tau}^1) \in \mathbb{R}^{n_1-r_1,n_1}, \qquad \bar{\mathbf{S}}^2_{r_2}(\boldsymbol{\tau}^2) := \bar{\mathbf{F}}^2_{r_2}(\boldsymbol{\tau}^2) \mathbf{D}^2_{r_2}(\boldsymbol{\tau}^2) \in \mathbb{R}^{n_2-r_2,n_2}$$

mit den analog zu (2.8) definierten Matrizen bzw. die approximierten Glättungsmatrizen

$$\tilde{\mathbf{S}}^1_{r_1}(\boldsymbol{\tau}^1) := \tilde{\mathbf{F}}^1_{r_1}(\boldsymbol{\tau}^1) \mathbf{D}^1_{r_1}(\boldsymbol{\tau}^1) \in \mathbb{R}^{n_1-r_1,n_1}, \qquad \tilde{\mathbf{S}}^2_{r_2}(\boldsymbol{\tau}^2) := \tilde{\mathbf{F}}^2_{r_2}(\boldsymbol{\tau}^2) \mathbf{D}^2_{r_2}(\boldsymbol{\tau}^2) \in \mathbb{R}^{n_2-r_2,n_2}$$

mit $\tilde{\mathbf{F}}^1_{r_1}(\boldsymbol{\tau}^1)$ und $\tilde{\mathbf{F}}^2_{r_2}(\boldsymbol{\tau}^1)$ gemäß (2.9), so zeigt man leicht, daß (4.4b) äquivalent ist zu

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

bzw.

$$\frac{1}{2} \left\| \begin{bmatrix} \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{B}_1(\boldsymbol{\tau}^1) \\ \sqrt{\mu_1} \mathbf{S}^1_{r_1}(\boldsymbol{\tau}^1) \end{bmatrix} \mathbf{A} \begin{bmatrix} \mathbf{B}_2(\boldsymbol{\tau}^2) \\ \sqrt{\mu_2} \mathbf{S}^2_{r_2}(\boldsymbol{\tau}^2) \end{bmatrix}^T \right\|_F^2 \to \min_{\mathbf{A} \in \mathbb{R}^{n_1,n_2}}, \tag{4.5b}$$

siehe [Bre90] und [Pig91] für eine ausführliche Herleitung.

Bevor wir uns der Charakterisierung der Lösung von (4.5a) und (4.5b) widmen, stellen wir einige technische Hilfsmittel bereit. Das *Kronecker-Produkt* von $\mathbf{A} \in \mathbb{R}^{m,n}$ und $\mathbf{B} \in \mathbb{R}^{p,q}$ ist definiert durch

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

Der *vec-Operator* $vec : \mathbb{R}^{m,n} \to \mathbb{R}^{m\cdot n}$ ist definiert durch

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

für Matrizen $\mathbf{A} \in \mathbb{R}^{m,n}$.

Die folgenden Rechenregeln, siehe etwa [Die93, S. 169ff] und [Bjö96, S. 336ff], fassen die wichtigsten Eigenschaften von vec-Operator und Kronecker-Produkt zusammen.

$$\begin{aligned} (\mathbf{A}\otimes\mathbf{B})^T &= \mathbf{A}^T\otimes\mathbf{B}^T & \operatorname{vec}(\mathbf{A}_{p,s}\mathbf{X}_{s,t}) &= (\mathbf{I}_t\otimes\mathbf{A})\operatorname{vec}(\mathbf{X}) \\ (\mathbf{A}\otimes\mathbf{B})^{-1} &= \mathbf{A}^{-1}\otimes\mathbf{B}^{-1} & \operatorname{vec}(\mathbf{X}_{s,t}\mathbf{B}_{t,q}) &= (\mathbf{B}^T\otimes\mathbf{I}_s)\operatorname{vec}(\mathbf{X}) \\ (\mathbf{A}\otimes\mathbf{B})^{+} &= \mathbf{A}^{+}\otimes\mathbf{B}^{+} & \operatorname{vec}(\mathbf{A}_{p,s}\mathbf{X}_{s,t}\mathbf{B}_{t*q}) &= (\mathbf{B}^T\otimes\mathbf{A})\operatorname{vec}(\mathbf{X}) \end{aligned}$$

$$\begin{aligned} (\mathbf{AB})\otimes(\mathbf{CD}) &= (\mathbf{A}\otimes\mathbf{C})\cdot(\mathbf{B}\otimes\mathbf{D}) \\ (\mathbf{A}+\mathbf{B})\otimes(\mathbf{C}+\mathbf{D}) &= \mathbf{A}\otimes\mathbf{C}+\mathbf{B}\otimes\mathbf{C}+\mathbf{A}\otimes\mathbf{D}+\mathbf{B}\otimes\mathbf{D} \end{aligned}$$

Unter Benutzung dieser Rechenregeln zeigt man leicht den Zusammenhang $\|\mathbf{AXB}-\mathbf{C}\|_F^2 = \left\|\left(\mathbf{B}^T\otimes\mathbf{A}\right)\operatorname{vec}(\mathbf{X}) - \operatorname{vec}(\mathbf{C})\right\|_2^2$. Damit ist Problem (4.5a) äquivalent zu

$$\frac{1}{2}\left\|\operatorname{vec}(\mathbf{Z}) - (\mathbf{B}_2(\boldsymbol{\tau}^2)\otimes\mathbf{B}_1(\boldsymbol{\tau}^1))\operatorname{vec}(\mathbf{A})\right\|_2^2 \to \min_{\operatorname{vec}(\mathbf{A})\in\mathbb{R}^{n_1 n_2}}, \tag{4.6}$$

dessen Normallösung $\mathbf{A}_{opt}(\boldsymbol{\tau}^1,\boldsymbol{\tau}^2)$ bei festen Knoten $\boldsymbol{\tau}^1$ und $\boldsymbol{\tau}^2$ durch $\operatorname{vec}(\mathbf{A}_{opt}) = (\mathbf{B}_2\otimes\mathbf{B}_1)^+\operatorname{vec}(\mathbf{Z}) = (\mathbf{B}_2^+\otimes\mathbf{B}_1^+)\operatorname{vec}(\mathbf{Z}) = \operatorname{vec}\left(\mathbf{B}_1^+\mathbf{Z}(\mathbf{B}_2^+)^T\right)$, also

$$\mathbf{A}_{opt}(\boldsymbol{\tau}^1,\boldsymbol{\tau}^2) := \mathbf{B}_1(\boldsymbol{\tau}^1)^+\mathbf{Z}\left(\mathbf{B}_2(\boldsymbol{\tau}^2)^+\right)^T \tag{4.7a}$$

gegeben ist. Sie kann durch aufeinanderfolgende Lösung der beiden folgenden univariaten Quadratmittelprobleme berechnet werden

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

$$\text{(A)}\qquad \frac{1}{2}\left\|\mathbf{F}^T-\mathbf{B}_2(\boldsymbol{\tau}^2)\mathbf{A}^T\right\|_F^2 \to \min_{\mathbf{A}\in\mathbb{R}^{n_1,n_2}}, \qquad \mathbf{A}^T = \mathbf{B}_2(\boldsymbol{\tau}^2)^+\mathbf{F}^T.$$

Analog erhält man die Normallösung zum Problem (4.5b)

$$\mathbf{A}_{opt}(\boldsymbol{\tau}^1,\boldsymbol{\tau}^2) := \begin{bmatrix} \mathbf{B}_1(\boldsymbol{\tau}^1) \\ \sqrt{\mu_1}\mathbf{S}_{r_1}^1(\boldsymbol{\tau}^1) \end{bmatrix}^+ \begin{bmatrix} \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \left(\begin{bmatrix} \mathbf{B}_2(\boldsymbol{\tau}^2) \\ \sqrt{\mu_2}\mathbf{S}_{r_2}^2(\boldsymbol{\tau}^2) \end{bmatrix}^+\right)^T, \tag{4.7b}$$

welche durch aufeineinderfolgende Lösung der folgenden Quadratmittelprobleme berechnet werden kann

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

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

Allgemeine Untersuchungen zur Strukturausnutzung bei großen linearen Quadratmittelproblemen, welche auf Kronecker-Produkten beruhen, findet man in [FF94].

Die Formulierung (4.6) ist der Ausgangspunkt für die Tensorprodukt-Approximation bei unregelmäßig verteilten Daten. Dort hat man $\frac{1}{2}\|\tilde{\mathbf{z}} - \tilde{\mathbf{B}}(\boldsymbol{\tau}^1, \boldsymbol{\tau}^2)\,\mathrm{vec}\,(\mathbf{A})\,\|_2^2 \to \min$, wobei $\tilde{z}_i$ $(i = 1, \ldots, M)$ die Datenwerte und $\tilde{\mathbf{B}}(\boldsymbol{\tau}_1, \boldsymbol{\tau}^2) \in \mathbb{R}^{M, n_1 n_2}$ die B-Spline-Matrix bezeichnet. Dieser Fall wurde erstmals von [HH74] untersucht. Im Gegensatz zum Problem mit Daten auf Rechteckgittern zerfällt dieses Problem nicht. Außerdem tritt hier die Rangdefizienz von $\tilde{\mathbf{B}}$ bei praktischen Beispielen sehr oft auf und es gibt keinen simplen Test wie die Schoenberg-Whitney-Bedingung. Um die Untersuchungen auf Splines mit freien Knoten auszudehnen, *müssen* wir also regularisieren. Obwohl sich die theoretischen Ergebnisse dann sicherlich übertragen lassen, haben wir den Fall der Tensorprodukt-Glättung durch Splines mit freien Knoten bei unregelmäßig verteilten Daten nicht näher untersucht, da schon die Lösung der linearen Probleme zu festen Knoten nicht zerfällt und demzufolge teuer ist.

4.1.2 Vollständiges und reduziertes Approximationsproblem

Unserer allgemeinen Zielstellung gemäß beziehen wir jetzt die Knoten des Splines in den Optimierungsprozeß ein. Betrachtet man im Problem (4.5a) eine Teilmenge $(\mathbf{t}^1, \mathbf{t}^2)$ der inneren Knoten als variabel, so erhält man das *vollständige Approximationsproblem*

$$\mathfrak{f}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}) := \frac{1}{2}\left\|\mathbf{Z} - \mathbf{B}_1(\mathbf{t}^1)\mathbf{A}\mathbf{B}_2(\mathbf{t}^2)^T\right\|_F^2 \to \min_{\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}} \tag{4.8}$$

mit linearen Nebenbedingungen der Form

$$\mathbf{C}_1\mathbf{t}^1 - \mathbf{h}^1 \geq \mathbf{0}, \quad \mathbf{C}_2\mathbf{t}^2 - \mathbf{h}^2 \geq \mathbf{0}. \tag{4.9}$$

Die Nebenbedingungen verhindern dabei wie im univariaten Fall das Zusammenfallen der Knoten.

Durch Einsetzen der Normallösung (4.7a) in das Funktional $\mathfrak{f}$ erhält man das *reduzierte Approximationsproblem*

$$f(\mathbf{t}^1, \mathbf{t}^2) := \frac{1}{2}\left\|\mathbf{Z} - \mathbf{B}_1(\mathbf{t}^1)\mathbf{B}_1(\mathbf{t}^1)^+\mathbf{Z}\left(\mathbf{B}_2(\mathbf{t}^2)^+\right)^T\mathbf{B}_2(\mathbf{t}^2)^T\right\|_F^2 \to \min_{\mathbf{t}^1, \mathbf{t}^2} \tag{4.10}$$

mit den linearen Ungleichheitsnebenbedingungen (4.9). Durch Umformung erhalten wir

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

mit den orthogonalen Projektoren $\mathbf{P}_{B_1} := \mathbf{B}_1(\mathbf{t}^1)\mathbf{B}_1(\mathbf{t}^1)^+$ und $\mathbf{P}_{B_2} := \mathbf{B}_2(\mathbf{t}^2)\mathbf{B}_2(\mathbf{t}^2)^+$.

4.1.3 Vollständiges und reduziertes Glättungsproblem

Betrachtet man im Problem (4.5b) eine Teilmenge $(\mathbf{t}^1, \mathbf{t}^2)$ der inneren Knoten als variabel, so erhält man das *vollständige Glättungsproblem*

$$\mathfrak{f}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}) := \frac{1}{2}\left\|\begin{bmatrix}\mathbf{Z} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix} - \begin{bmatrix}\mathbf{B}_1(\boldsymbol{\tau}^1)\\ \sqrt{\mu_1}\mathbf{S}_{r_1}^1(\boldsymbol{\tau}^1)\end{bmatrix}\mathbf{A}\begin{bmatrix}\mathbf{B}_2(\boldsymbol{\tau}^2)\\ \sqrt{\mu_2}\mathbf{S}_{r_2}^2(\boldsymbol{\tau}^2)\end{bmatrix}^T\right\|_F^2 \to \min_{\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}} \tag{4.11}$$

mit den linearen Ungleichheitsnebenbedingungen (4.9).

Setzt man hier die Normallösung (4.7b) in das Funktional ein, so erhält man das *reduzierte Glättungsproblem*

$$f(\mathbf{t}^1,\mathbf{t}^2) := \frac{1}{2}\left\| \begin{bmatrix} \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} - \mathbf{P}_{\left[\frac{B}{\sqrt{\mu}S}\right]1} \begin{bmatrix} \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{P}_{\left[\frac{B}{\sqrt{\mu}S}\right]2} \right\|_F^2 \to \min_{\mathbf{t}^1,\mathbf{t}^2} \tag{4.12}$$

mit den Nebenbedingungen (4.9). Die orthogonalen Projektoren sind definiert als

$$\mathbf{P}_{\left[\frac{B}{\sqrt{\mu}S}\right]1} := \begin{bmatrix} \mathbf{B}_1(\boldsymbol{\tau}^1) \\ \sqrt{\mu_1}\mathbf{S}^1_{r_1}(\boldsymbol{\tau}^1) \end{bmatrix} \begin{bmatrix} \mathbf{B}_1(\boldsymbol{\tau}^1) \\ \sqrt{\mu_1}\mathbf{S}^1_{r_1}(\boldsymbol{\tau}^1) \end{bmatrix}^+$$
$$\mathbf{P}_{\left[\frac{B}{\sqrt{\mu}S}\right]2} := \begin{bmatrix} \mathbf{B}_2(\boldsymbol{\tau}^2) \\ \sqrt{\mu_2}\mathbf{S}^2_{r_2}(\boldsymbol{\tau}^2) \end{bmatrix} \begin{bmatrix} \mathbf{B}_2(\boldsymbol{\tau}^2) \\ \sqrt{\mu_2}\mathbf{S}^2_{r_2}(\boldsymbol{\tau}^2) \end{bmatrix}^+ .$$

Unser Ziel für den Rest des Kapitels ist der Nachweis der Äquivalenz zwischen vollständigem und reduziertem Problem sowie die Entwicklung eines Algorithmus zur Lösung des reduzierten Problems. Besonders interessiert uns dabei, ob die entstehenden Probleme zerfallen und wie sich die Techniken aus dem univariaten Fall nutzen lassen. Man beachte, daß eine vollständige Lokalisierung mittels Tensorprodukt-Splines prinzipbedingt nicht möglich ist, z. B. kann man Peaks oder diagonal verlaufende Wellenfronten nur schlecht approximieren. Einen Ausweg aus dieser Situation bieten Splines auf Triangulationen, gekrümmte Knotenlinien oder die Verwendung von hierarchischen B-Splines [Kra94]. Trotz dieser Nachteile haben Tensorprodukt-Splines auch heute noch eine weite Verbreitung in der Praxis (Fahrzeugbau, CAGD-Systeme), da sie einfach zu handhaben und billig zu berechnen sind, sofern die Knotenlinien einmal bestimmt sind.

Während es im univariaten Fall eine große Auswahl an Algorithmen zur direkten Minimierung des Quadratmittelfehlers als Funktion der freien Knoten gab, sind uns im bivariaten Tensorprodukt-Fall keine solchen Algorithmen bekannt. In [Die93] berichtet der Autor lediglich über einen Algorithmus aus der unveröffentlichten PhD-Thesis [Die79], welcher wie im Univariaten auf der Lösung eines Barriereproblems mittels CG-Verfahren beruht. Beschränkt man sich auf das heuristische, adaptive Einfügen von Knotenlinien, so steht der Algorithmus REGRID aus [Die89] zur Verfügung. Im Fall der Chebyshev-Approximation von Funktionen wurden in [MNW96] kürzlich erste Ergebnisse erzielt.

4.2 Separable Quadratmittelprobleme mit Tensorprodukt-Struktur

In diesem Abschnitt lösen wir uns vom Problem der Approximation durch Splines und betrachten allgemeine Probleme der Form

Vollständiges Problem

$$\mathfrak{f}(\mathbf{t}^1,\mathbf{t}^2,\mathbf{A}) := \frac{1}{2}\,\|\mathfrak{F}(\mathbf{t}^1,\mathbf{t}^2,\mathbf{A})\|_F^2 \to \min_{\mathbf{t}^1,\mathbf{t}^2,\mathbf{A}} \tag{4.13}$$

mit $\mathfrak{F}(\mathbf{t}^1,\mathbf{t}^2,\mathbf{A}) := \mathbf{Z} - \mathbf{B}_1(\mathbf{t}^1)\mathbf{A}\mathbf{B}_2(\mathbf{t}^2)^T$ unter den Nebenbedingungen

$$\mathbf{C}_1\mathbf{t}^1 - \mathbf{h}^1 \geq \mathbf{0}, \quad \mathbf{C}_2\mathbf{t}^2 - \mathbf{h}^2 \geq \mathbf{0}. \tag{4.14}$$

Hierbei seien $\mathbf{B}_1$ und $\mathbf{B}_2$ beliebige glatte Matrixfunktionen und die verbleibenden Größen $\mathbf{Z}$, $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{h}^1$ sowie $\mathbf{h}^2$ konstante Vektoren und Matrizen. Die *variable Projektion*, d. h. die Optimallösung $\mathbf{A}_{opt}$ von (4.13) bei festen $\mathbf{t}^1$ und $\mathbf{t}^2$, ist gegeben durch

$$\mathbf{A}_{opt}(\mathbf{t}^1,\mathbf{t}^2) := \mathbf{B}_1(\mathbf{t}^1)^+\mathbf{Z}\left(\mathbf{B}_2(\mathbf{t}^2)^+\right)^T. \tag{4.15}$$

Reduziertes Problem

$$f(\mathbf{t}^1,\mathbf{t}^2) := \frac{1}{2}\left\|\mathbf{F}(\mathbf{t}^1,\mathbf{t}^2)\right\|_F^2 \to \min_{\mathbf{t}^1,\mathbf{t}^2} \tag{4.16}$$

mit $\mathbf{F}(\mathbf{t}^1,\mathbf{t}^2) := \mathfrak{F}(\mathbf{t}^1,\mathbf{t}^2,\mathbf{A}_{opt}(\mathbf{t}^1,\mathbf{t}^2)) = \mathbf{P}_{B_1}^{\perp}\mathbf{Z}\mathbf{P}_{B_2}$ unter den Nebenbedingungen (4.14). Eine äquivalente Darstellung des reduzierten Funktionals $f(\mathbf{t}^1,\mathbf{t}^2)$ ist offensichtlich durch

$$f(\mathbf{t}^1,\mathbf{t}^2) = \frac{1}{2}\left\|\mathbf{B}_1(\mathbf{t}^1)\mathbf{B}_1(\mathbf{t}^1)^+\mathbf{Z}\left(\mathbf{B}_2(\mathbf{t}^2)^+\right)^T\mathbf{B}_2(\mathbf{t}^2)^T - \mathbf{Z}\right\|_F^2 = \frac{1}{2}\left\|\mathbf{P}_{B_1}\mathbf{Z}\mathbf{P}_{B_2}^{\perp}\right\|_F^2$$

gegeben.

Wir wollen nun in Verallgemeinerung der Ergebnisse von [GP73] untersuchen, inwieweit vollständiges und reduziertes Problem äquivalent sind. Im folgenden benötigen wir oft die Fréchet-Ableitung von gewissen Funktionalen, welche durch die Frobeniusnorm definiert sind. Es bezeichne $\boldsymbol{\partial}$ den Operator der Fréchet-Ableitung.

Lemma 4.1 (Fréchet-Ableitungen von Frobeniusnormen)**.**
Sei $\mathbf{A}:\mathbb{R}^l \to \mathfrak{L}(\mathbb{R}^n,\mathbb{R}^m)$, $\mathbf{x}\in\mathbb{R}^l \to \mathbf{A}(\mathbf{x})\in\mathbb{R}^{m,n}$ *eine Fréchet-differenzierbare Matrixfunktion und sei*

$$f:\mathbb{R}^l \to \mathfrak{L}(\mathbb{R}), \mathbf{x}\in\mathbb{R}^l \to f(\mathbf{x}) := \frac{1}{2}\left\|\mathbf{A}(\mathbf{x})\right\|_F^2 \in \mathbb{R}.$$

Dann gilt

$$\begin{aligned}\boldsymbol{\partial} f(\mathbf{x})[\Delta\mathbf{x}] &= \operatorname{tr}\Big\{\left(\boldsymbol{\partial}\mathbf{A}(\mathbf{x})[\Delta\mathbf{x}]\right)^T\mathbf{A}(\mathbf{x})\Big\} = \operatorname{tr}\Big\{\mathbf{A}(\mathbf{x})^T\left(\boldsymbol{\partial}\mathbf{A}(\mathbf{x})[\Delta\mathbf{x}]\right)\Big\}\\ &= \operatorname{tr}\Big\{\left(\boldsymbol{\partial}\mathbf{A}(\mathbf{x})[\Delta\mathbf{x}]\right)\mathbf{A}(\mathbf{x})^T\Big\} = \operatorname{tr}\Big\{\mathbf{A}(\mathbf{x})\left(\boldsymbol{\partial}\mathbf{A}(\mathbf{x})[\Delta\mathbf{x}]\right)^T\Big\} \qquad \textit{für alle } \Delta\mathbf{x}\in\mathbb{R}^l.\end{aligned}$$

Beweis. Bekanntlich gilt $f(\mathbf{x}) = \frac{1}{2}\left\|\mathbf{A}(\mathbf{x})\right\|_F^2 = \frac{1}{2}\operatorname{tr}\left\{\mathbf{A}(\mathbf{x})^T\mathbf{A}(\mathbf{x})\right\}$. Man hat

$$\begin{aligned}f(\mathbf{x}+\Delta\mathbf{x}) &= \frac{1}{2}\operatorname{tr}\Big\{\mathbf{A}(\mathbf{x}+\Delta\mathbf{x})^T\mathbf{A}(\mathbf{x}+\Delta\mathbf{x})\Big\}\\ &= \frac{1}{2}\operatorname{tr}\Big\{\left(\mathbf{A}(\mathbf{x})+\boldsymbol{\partial}\mathbf{A}(\mathbf{x})[\Delta\mathbf{x}]+\mathcal{O}(\|\Delta\mathbf{x}\|)\right)^T\left(\mathbf{A}(\mathbf{x})+\boldsymbol{\partial}\mathbf{A}(\mathbf{x})[\Delta\mathbf{x}]+\mathcal{O}(\|\Delta\mathbf{x}\|)\right)\Big\}\end{aligned}$$

und wegen der Linearität der Spur

$$f(\mathbf{x}+\Delta\mathbf{x}) - f(\mathbf{x}) = \frac{1}{2}\operatorname{tr}\Big\{\left(\boldsymbol{\partial}\mathbf{A}(\mathbf{x})[\Delta\mathbf{x}]\right)^T\mathbf{A}(\mathbf{x}) + \mathbf{A}(\mathbf{x})^T\left(\boldsymbol{\partial}\mathbf{A}(\mathbf{x})[\Delta\mathbf{x}]\right) + \mathcal{O}(\|\Delta\mathbf{x}\|)\Big\}.$$

Für quadratische Matrizen $\mathbf{B}$ gilt $\operatorname{tr}(\mathbf{B}) = \operatorname{tr}(\mathbf{B}^T)$, also wegen $\left(\boldsymbol{\partial}\mathbf{A}(\mathbf{x})[\Delta\mathbf{x}]\right)^T\mathbf{A}(\mathbf{x}) \in \mathbb{R}^{n,n}$ schließlich

$$\begin{aligned}f(\mathbf{x}+\Delta\mathbf{x}) - f(\mathbf{x}) &= \operatorname{tr}\Big\{\left(\boldsymbol{\partial}\mathbf{A}(\mathbf{x})[\Delta\mathbf{x}]\right)^T\mathbf{A}(\mathbf{x}) + \mathcal{O}(\|\Delta\mathbf{x}\|)\Big\}, \text{ d. h.}\\ \boldsymbol{\partial} f(\mathbf{x})[\Delta\mathbf{x}] &= \operatorname{tr}\Big\{\left(\boldsymbol{\partial}\mathbf{A}(\mathbf{x})[\Delta\mathbf{x}]\right)^T\mathbf{A}(\mathbf{x})\Big\} = \operatorname{tr}\Big\{\mathbf{A}(\mathbf{x})^T\left(\boldsymbol{\partial}\mathbf{A}(\mathbf{x})[\Delta\mathbf{x}]\right)\Big\}.\end{aligned}$$

Aus der äquivalenten Darstellung $f(\mathbf{x}) = \frac{1}{2}\left\|\mathbf{A}(\mathbf{x})\right\|_F^2 = \frac{1}{2}\operatorname{tr}\left\{\mathbf{A}(\mathbf{x})\mathbf{A}(\mathbf{x})^T\right\}$ erhält man in analoger Weise den restlichen Teil der Behauptung. □

4.2.1 Die Fréchet-Ableitung des vollständigen Funktionals

Sei $\boldsymbol{\partial}_1$ der Operator der Fréchet-Ableitung bez. $\mathbf{t}^1$. Mittels Lemma 4.1 erhalten wir

$$\boldsymbol{\partial}_1 \mathfrak{f}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})[\Delta\mathbf{t}^1] = \operatorname{tr}\left\{ \left(\boldsymbol{\partial}_1 \mathfrak{F}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})[\Delta\mathbf{t}^1]\right)^T \mathfrak{F}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}) \right\}.$$

Im folgenden werden wir das Argument der Matrixfunktionen $\mathbf{B}_1$ und $\mathbf{B}_2$ zwecks Vereinfachung der Schreibweise weglassen.

Offensichtlich gilt $\boldsymbol{\partial}_1 \mathfrak{F}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})[\Delta\mathbf{t}^1] = -\boldsymbol{\partial}_1 \mathbf{B}_1[\Delta\mathbf{t}^1]\mathbf{A}\mathbf{B}_2^T$. Damit haben wir

$$\boldsymbol{\partial}_1 \mathfrak{f}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})[\Delta\mathbf{t}^1] = \operatorname{tr}\left\{ \left(-\boldsymbol{\partial}_1 \mathbf{B}_1[\Delta\mathbf{t}^1]\mathbf{A}\mathbf{B}_2^T\right)^T \left(\mathbf{Z} - \mathbf{B}_1\mathbf{A}\mathbf{B}_2^T\right) \right\},$$

also schließlich den folgenden Ausdruck für die Fréchet-Ableitung des vollständigen Funktionals bez. $\mathbf{t}^1$

$$\boldsymbol{\partial}_1 \mathfrak{f}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})[\Delta\mathbf{t}^1] = -\operatorname{tr}\left\{ \mathbf{B}_2\mathbf{A}^T \left(\boldsymbol{\partial}_1 \mathbf{B}_1[\Delta\mathbf{t}^1]\right)^T \left(\mathbf{Z} - \mathbf{B}_1\mathbf{A}\mathbf{B}_2^T\right) \right\}. \tag{4.17}$$

Sei $\boldsymbol{\partial}_2$ der Operator der Fréchet-Ableitung bez. $\mathbf{t}^2$. Unter Benutzung der Darstellung

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

mit $\mathfrak{F}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}) := \mathbf{B}_1\mathbf{A}\mathbf{B}_2^T - \mathbf{Z}$ erhalten wir aus Lemma 4.1

$$\boldsymbol{\partial}_2 \mathfrak{f}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})[\Delta\mathbf{t}^2] = \operatorname{tr}\left\{ \mathfrak{F}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}) \left(\boldsymbol{\partial}_2 \mathfrak{F}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})[\Delta\mathbf{t}^2]\right)^T \right\}.$$

Es gilt $\boldsymbol{\partial}_2 \mathfrak{F}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})[\Delta\mathbf{t}^2] = \mathbf{B}_1\mathbf{A} \left(\boldsymbol{\partial}_2 \mathbf{B}_2[\Delta\mathbf{t}^2]\right)^T$ und daher

$$\boldsymbol{\partial}_2 \mathfrak{f}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})[\Delta\mathbf{t}^2] = \operatorname{tr}\left\{ \left(\mathbf{B}_1\mathbf{A}\mathbf{B}_2^T - \mathbf{Z}\right) \left(\boldsymbol{\partial}_2 \mathbf{B}_2[\Delta\mathbf{t}^1]\right) \mathbf{A}^T\mathbf{B}_1^T \right\} \tag{4.18}$$

für die Fréchet-Ableitung des vollständigen Funktionals bez. $\mathbf{t}^2$.

4.2.2 Die Fréchet-Ableitung des reduzierten Funktionals

Wir berechnen nun die Fréchet-Ableitung des reduzierten Funktionals f bez. $\mathbf{t}^1$. Wir erhalten zunächst wegen Lemma 4.1

$$\boldsymbol{\partial}_1 f(\mathbf{t}^1, \mathbf{t}^2)[\Delta\mathbf{t}^1] = \operatorname{tr}\left\{ \left(\boldsymbol{\partial}_1 \mathbf{F}(\mathbf{t}^1, \mathbf{t}^2)[\Delta\mathbf{t}^1]\right)^T \mathbf{F}(\mathbf{t}^1, \mathbf{t}^2) \right\}$$

und mit Lemma 2.6 für die Fréchet-Ableitung der Residuumsfunktion

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

Es folgt

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

und wegen $\mathbf{B}_1^+ \mathbf{P}_{B_1}^\perp = \mathbf{0}$ schließlich

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

Es gilt $\left(\mathbf{P}_{B_1}^\perp\right)^T \mathbf{P}_{B_1}^\perp = \mathbf{P}_{B_1}^\perp$ und daher

$$\boldsymbol{\partial}_1 f(\mathbf{t}^1, \mathbf{t}^2)[\Delta\mathbf{t}^1] = -\operatorname{tr}\left\{ \mathbf{P}_{B_2} \mathbf{Z}^T \left((\boldsymbol{\partial}_1 \mathbf{B}_1)[\Delta\mathbf{t}^1]) \mathbf{B}_1^+ \right)^T \mathbf{P}_{B_1}^\perp \mathbf{Z} \mathbf{P}_{B_2} \right\}. \tag{4.19}$$

Betrachten wir abschließend die Fréchet-Ableitung des reduzierten Funktionals f bez. $\mathbf{t}^2$. Wir verwenden die Darstellung

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

mit $\mathbf{F}(\mathbf{t}^1, \mathbf{t}^2) := \mathbf{P}_{B_1} \mathbf{Z} \mathbf{P}_{B_2}^\perp$. Aus Lemma 4.1 folgt

$$\boldsymbol{\partial}_2 f(\mathbf{t}^1, \mathbf{t}^2)[\Delta\mathbf{t}^2] = \operatorname{tr}\left\{ \mathbf{F}(\mathbf{t}^1, \mathbf{t}^2) \left(\boldsymbol{\partial}_2 \mathbf{F}(\mathbf{t}^1, \mathbf{t}^2)[\Delta\mathbf{t}^2] \right)^T \right\}.$$

Unter Benutzung von Lemma 2.6 erhalten wir

$$\begin{aligned} \boldsymbol{\partial}_2 \mathbf{F}(\mathbf{t}^1, \mathbf{t}^2)[\Delta\mathbf{t}^2] &= \mathbf{P}_{B_1} \mathbf{Z} \left(\boldsymbol{\partial}_2 \mathbf{P}_{B_2}^\perp [\Delta\mathbf{t}^2] \right) \\ &= -\mathbf{P}_{B_1} \mathbf{Z} \left\{ \mathbf{P}_{B_2}^\perp (\boldsymbol{\partial}_2 \mathbf{B}_2[\Delta\mathbf{t}^2]) \mathbf{B}_2^+ + \left(\mathbf{P}_{B_2}^\perp (\boldsymbol{\partial}_2 \mathbf{B}_2[\Delta\mathbf{t}^2]) \mathbf{B}_2^+ \right)^T \right\}, \end{aligned}$$

also

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

Es gilt $\mathbf{P}_{B_2}^\perp (\mathbf{B}_2^+)^T = \mathbf{0}$, d. h.

$$\boldsymbol{\partial}_2 f(\mathbf{t}^1, \mathbf{t}^2)[\Delta\mathbf{t}^2] = -\operatorname{tr}\left\{ \mathbf{P}_{B_1} \mathbf{Z} \mathbf{P}_{B_2}^\perp (\boldsymbol{\partial}_2 \mathbf{B}_2[\Delta\mathbf{t}^2]) \mathbf{B}_2^+ \mathbf{Z}^T \mathbf{P}_{B_1} \right\}. \tag{4.20}$$

4.2.3 Beziehungen zwischen den Fréchet-Ableitungen

Lemma 4.2.
Es gelte die Bedingung: Die Matrixfunktionen $\mathbf{B}_1$ *und* $\mathbf{B}_2$ *besitzen an der Stelle* $\mathbf{t}^1$ *und* $\mathbf{t}^2$ *lokal konstanten Rang. Sei*

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

die zugehörige variable Projektion. Dann gilt

$$\boldsymbol{\partial}_1 f(\mathbf{t}^1, \mathbf{t}^2) = \boldsymbol{\partial}_1 \mathfrak{f}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}_{opt}(\mathbf{t}^1, \mathbf{t}^2)) \quad \textit{und} \quad \boldsymbol{\partial}_2 f(\mathbf{t}^1, \mathbf{t}^2) = \boldsymbol{\partial}_2 \mathfrak{f}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}_{opt}(\mathbf{t}^1, \mathbf{t}^2)).$$

Beweis. (i) Wir betrachten zunächst die Fréchet-Ableitung bez. $\mathbf{t}^1$. Setzt man die variable Projektion in die Formel (4.17) ein, so erhält man

$$\begin{aligned}\boldsymbol{\partial}_1 \mathfrak{f}\left(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}_{opt}\left(\mathbf{t}^1, \mathbf{t}^2\right)\right)\left[\Delta \mathbf{t}^1\right] &= \operatorname{tr}\left\{-\mathbf{B}_2 \mathbf{B}_2^+ \mathbf{Z}^T (\mathbf{B}_1^+)^T (\boldsymbol{\partial}_1 \mathbf{B}_1[\Delta \mathbf{t}^1])^T \left(\mathbf{Z} - \mathbf{B}_1 \mathbf{B}_1^+ \mathbf{Z} (\mathbf{B}_2^+)^T \mathbf{B}_2^T\right)\right\} \\ &= \operatorname{tr}\left\{-\mathbf{P}_{B_2} \mathbf{Z}^T ((\boldsymbol{\partial}_1 \mathbf{B}_1[\Delta \mathbf{t}^1]) \mathbf{B}_1^+)^T \mathbf{P}_{B_1}^{\perp} \mathbf{Z} \mathbf{P}_{B_2}\right\} \\ &= \boldsymbol{\partial}_1 f\left(\mathbf{t}^1, \mathbf{t}^2\right)\left[\Delta \mathbf{t}^1\right] \qquad \text{(siehe (4.19))}.\end{aligned}$$

(ii) Widmen wir uns nun der Fréchet-Ableitung bez. $\mathbf{t}^2$. Einsetzen der variablen Projektion in die Formel (4.18) liefert

$$\begin{aligned}\boldsymbol{\partial}_2 \mathfrak{f}\left(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A}_{opt}\left(\mathbf{t}^1, \mathbf{t}^2\right)\right)\left[\Delta \mathbf{t}^2\right] &= \operatorname{tr}\left\{\left(\mathbf{B}_1 \mathbf{B}_1^+ \mathbf{Z} (\mathbf{B}_2^+)^T \mathbf{B}_2^T - \mathbf{Z}\right) (\boldsymbol{\partial}_2 \mathbf{B}_2[\Delta \mathbf{t}^2]) \mathbf{B}_2^+ \mathbf{Z}^T (\mathbf{B}_1^+)^T \mathbf{B}_1^T\right\} \\ &= \operatorname{tr}\left\{-\mathbf{P}_{B_1} \mathbf{Z} \mathbf{P}_{B_2}^{\perp} (\boldsymbol{\partial}_2 \mathbf{B}_2[\Delta \mathbf{t}^2]) \mathbf{B}_2^+ \mathbf{Z}^T \mathbf{P}_{B_1}\right\} \\ &= \boldsymbol{\partial}_2 f\left(\mathbf{t}^1, \mathbf{t}^2\right)\left[\Delta \mathbf{t}^2\right] \qquad \text{(siehe (4.20))}.\end{aligned}$$

□

Die Bedeutung des obigen Lemmas ebenso wie die des nächsten Satzes liegt nicht so sehr in deren Aussage an sich – die man erwarten konnte, sondern in den Darstellungen von Gradient und Jacobi-Matrizen des reduzierten Funktionals. Das folgende Theorem ist damit eine natürliche Verallgemeinerung von [GP73, Theorem 2.1] auf den Tensorprodukt-Fall. Man beachte jedoch, daß wir die linearen Ungleichheitsnebenbedingungen an $\mathbf{t}^1$ und $\mathbf{t}^2$ einbezogen haben.

4.2.4 Äquivalenz von vollständigem und reduziertem Problem

Theorem 4.1 (Äquivalenz von vollständigem und reduziertem Problem)**.**
Seien vollständiges und reduziertes Problem wie oben definiert. Sei weiterhin vorausgesetzt, daß die Matrixfunktionen $\mathbf{B}_1(\mathbf{t}^1)$ *(bzw.* $\mathbf{B}_2(\mathbf{t}^2)$*) konstanten Rang auf der offenen Menge* $\Omega_1 \subset \mathbb{R}^{l_1}$ *(bzw.* $\Omega_2 \subset \mathbb{R}^{l_2}$*) besitzen.*

(i) Ist $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$ *ein kritischer Punkt (oder eine globale Minimumstelle auf* $\Omega_1 \times \Omega_2$*) des reduzierten Problems und gilt*

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

so ist $(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*}))$ *ein kritischer Punkt (oder eine globale Minimumstelle für* $(\mathbf{t}^1, \mathbf{t}^2) \in \Omega_1 \times \Omega_2$*) des vollständigen Problems und es gilt*

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

(ii) Ist $(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$ *eine globale Minimumstelle des vollständigen Problems für* $(\mathbf{t}^1, \mathbf{t}^2) \in \Omega_1 \times \Omega_2$*, so ist* $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$ *eine globale Minimumstelle des reduzierten Problems auf* $\Omega_1 \times \Omega_2$ *und es gilt*

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

Gibt es ein eindeutiges $\mathbf{A}^*$ *unter allen minimierenden Paaren von* $\mathfrak{f}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})$*, so muß gelten*

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

Beweis. Wir definieren die Lagrange-Funktionen des vollständigen und reduzierten Problems

$$L_I\left(\mathbf{t}^1,\mathbf{t}^2,\mathbf{A},\mathfrak{w}^1,\mathfrak{w}^2\right) := \mathfrak{f}\left(\mathbf{t}^1,\mathbf{t}^2,\mathbf{A}\right) - \sum_{i=1}^{ncstr_1} \mathfrak{w}_i^1 r_i^1(\mathbf{t}^1) - \sum_{i=1}^{ncstr_2} \mathfrak{w}_i^2 r_i^2(\mathbf{t}^2)$$

$$L_{II}\left(\mathbf{t}^1,\mathbf{t}^2,\mathbf{w}^1,\mathbf{w}^2\right) := f\left(\mathbf{t}^1,\mathbf{t}^2\right) - \sum_{i=1}^{ncstr_1} w_i^1 r_i^1(\mathbf{t}^1) - \sum_{i=1}^{ncstr_2} w_i^2 r_i^2(\mathbf{t}^2)$$

mit nichtnegativen Multiplikatoren $\mathfrak{w}_i^1$, w_i^1 $(i=1,\ldots,ncstr_1)$ und $\mathfrak{w}_i^2$, w_i^2 $(i=1,\ldots,ncstr_2)$ und den Nebenbedingungen $\mathbf{r}^1(\mathbf{t}^1) := \mathbf{C}_1\mathbf{t}^1 - \mathbf{h}^1 \geq \mathbf{0}$ und $\mathbf{r}^2(\mathbf{t}^2) := \mathbf{C}_2\mathbf{t}^2 - \mathbf{h}^2 \geq \mathbf{0}$.

(i) Sei $(\mathbf{t}^{1*},\mathbf{t}^{2*})$ ein kritischer Punkt des reduzierten Problems und sei $\mathbf{A}_{opt}\left(\mathbf{t}^{1*},\mathbf{t}^{2*}\right) = \mathbf{B}_1\left(\mathbf{t}^{1*}\right)^+ \mathbf{Z}\left(\mathbf{B}_2\left(\mathbf{t}^{2*}\right)^+\right)^T$. Die notwendigen Optimalitätsbedingungen erster Ordnung liefern dann die Existenz von Multiplikatoren $\mathbf{w}^{1*}$ und $\mathbf{w}^{2*}$, so daß

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

Ferner sind constraint qualifications an der Stelle $(\mathbf{t}^{1*},\mathbf{t}^{2*})$ erfüllt. Es gilt

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

Mittels Lemma 4.2 und durch die Identifikation entsprechender Multiplikatoren erhalten wir

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

Analog erhalten wir $\mathbf{0} = \nabla_{\mathbf{t}^1} L_I\left(\mathbf{t}^{1*},\mathbf{t}^{2*},\mathbf{A}_{opt}\left(\mathbf{t}^{1*},\mathbf{t}^{2*}\right),\mathfrak{w}^{1*},\mathfrak{w}^{2*}\right)$. Zusammen mit der Zulässigkeit der Nebenbedingungen und der Komplementarität (nach Identifikation entsprechender Mulitplikatoren) ergibt dies die notwendigen Optimalitätsbedingungen erster Ordnung für das vollständige Problem. Die constraint qualification überträgt sich auf das vollständige Problem, da die Nebenbedingungen unverändert bleiben. Man beachte, daß $\nabla_{\mathbf{A}} L_I\left(\mathbf{t}^{1*},\mathbf{t}^{2*},\mathbf{A}_{opt}\left(\mathbf{t}^{1*},\mathbf{t}^{2*}\right),\mathfrak{w}^{1*},\mathfrak{w}^{2*}\right) = \mathbf{0}$ auf Grund der Definition von $\mathbf{A}_{opt}\left(\mathbf{t}^{1*},\mathbf{t}^{2*}\right)$.

Also ist $\left(\mathbf{t}^{1*},\mathbf{t}^{2*},\mathbf{A}_{opt}\left(\mathbf{t}^{1*},\mathbf{t}^{2*}\right),\mathfrak{w}^{1*},\mathfrak{w}^{2*}\right)$ ein kritischer Punkt des vollständigen Problems. Nach Definition des reduzierten Problems hat man

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

Der Rest des Beweises folgt den Grundideen des Beweises von Golub/Pereyra. Der Vollständigkeit halber sei er hier angeführt.

Sei $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$ eine globale Minimumstelle des reduzierten Problems in $\Omega_1 \times \Omega_2$ und sei $\mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*}) = \mathbf{B}_1(\mathbf{t}^{1*})^+ \mathbf{Z}\left(\mathbf{B}_2(\mathbf{t}^{2*})^+\right)^T$. Dann gilt $\mathfrak{f}(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*})) = f(\mathbf{t}^{1*}, \mathbf{t}^{2*})$. Angenommen es existieren $\left(\mathbf{t}^{1\dagger}, \mathbf{t}^{2\dagger}, \mathbf{A}^\dagger\right)$ mit $\mathbf{t}^{1\dagger} \in \Omega_1$, $\mathbf{t}^{2\dagger} \in \Omega_2$, so daß $\mathfrak{f}\left(\mathbf{t}^{1\dagger}, \mathbf{t}^{2\dagger}, \mathbf{A}^\dagger\right) < \mathfrak{f}(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*}))$. Für alle $(\mathbf{t}^1, \mathbf{t}^2)$ gilt $f(\mathbf{t}^1, \mathbf{t}^2) \le \mathfrak{f}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})$, also

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

im Widerspruch zur Annahme, daß $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$ eine globale Minimumstelle des reduzierten Problems in $\Omega_1 \times \Omega_2$ ist. Folglich ist $(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*}))$ eine globale Minimumstelle des vollständigen Problems in $\Omega_1 \times \Omega_2$.

(ii) Sei $(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$ eine globale Minimumstelle des vollständigen Problems für $(\mathbf{t}^1, \mathbf{t}^2) \in \Omega_1 \times \Omega_2$ und sei $\mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*}) = \mathbf{B}_1(\mathbf{t}^{1*})^+ \mathbf{Z}\left(\mathbf{B}_2(\mathbf{t}^{2*})^+\right)^T$. Es gilt $f(\mathbf{t}^{1*}, \mathbf{t}^{2*}) \le \mathfrak{f}(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$. Nach Definition des reduzierten Funktionals folgt $f(\mathbf{t}^{1*}, \mathbf{t}^{2*}) = \mathfrak{f}(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*})) \le \mathfrak{f}(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$. Da $(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$ eine globale Minimumstelle ist, gilt das Gleichheitszeichen, d. h.

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

Gibt es ein eindeutiges $\mathbf{A}^*$ unter allen minimierenden Paaren von $\mathfrak{f}(\mathbf{t}^1, \mathbf{t}^2, \mathbf{A})$, so muß gelten $\mathbf{A}^* = \mathbf{A}_{opt}(\mathbf{t}^{1*}, \mathbf{t}^{2*})$.

Wir nehmen nun an, daß $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$ *keine* globale Minimumstelle des reduzierten Problems auf $\Omega_1 \times \Omega_2$ ist, d. h. es existiert $\left(\mathbf{t}^{1\dagger}, \mathbf{t}^{2\dagger}\right) \in \Omega_1 \times \Omega_2$, so daß $f\left(\mathbf{t}^{1\dagger}, \mathbf{t}^{2\dagger}\right) < f(\mathbf{t}^{1*}, \mathbf{t}^{2*})$. Mit $\mathbf{A}^\dagger = \mathbf{B}_1\left(\mathbf{t}^{1\dagger}\right)^+ \mathbf{Z}\left(\mathbf{B}_2\left(\mathbf{t}^{2\dagger}\right)^+\right)^T$ gilt dann

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

im Widerspruch zur Voraussetzung, daß $(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$ eine globale Minimumstelle des vollständigen Problems ist. □

Aus der Beweisführung erkennt man unmittelbar, daß sich die Aussage des Theorems auch auf Probleme mit nichtlinearen Gleichheitsnebenbedingungen $\mathbf{s}^1(\mathbf{t}^1) = \mathbf{0}$ und $\mathbf{s}^2(\mathbf{t}^2) = \mathbf{0}$ übertragen läßt.

Den Herleitungen der Fréchet-Ableitungen entnimmt man, daß sich die von Kaufman ausgenutzte Struktur auf den bivariaten Fall überträgt.

4.3 Bivariate Tensorprodukt-Splines mit freien Knoten

Nach diesen Vorarbeiten können wir die Reduktionstechnik unmittelbar auf das vollständige Glättungsproblem (4.11), (4.9) anwenden. Durch Stetigkeitsargumente analog dem univariaten Fall ohne Nebenbedingungen erhalten wir zunächst

Theorem 4.2 (Existenz einer Lösung des reduzierten Glättungsproblems)**.**
Die Menge der zulässigen Knoten $\left\{(\mathbf{t}^1, \mathbf{t}^2) \in \mathbb{R}^{l_1} \times \mathbb{R}^{l_2} : \mathbf{C}_1\mathbf{t}^1 - \mathbf{h}^1 \ge \mathbf{0}, \mathbf{C}_1\mathbf{t}^1 - \mathbf{h}^1 \ge \mathbf{0}\right\}$ *sei nichtleer. Für feste* $r_1 \in \{0, \ldots, q_1\}$, $0 \le q_1 < k_1$ *und* $r_2 \in \{0, \ldots, q_2\}$, $0 \le q_2 < k_2$ *gelte:*

(*V*1) *Die Knoten erfüllen die Bedingung* $\tau^1_{j_1} < \tau^1_{j_1+k_1-q_1}$ $(j_1 = q_1 + 1, \ldots, n_1)$ *und* $\tau^2_{j_2} < \tau^2_{j_2+k_2-q_2}$ $(j_2 = q_2 + 1, \ldots, n_2)$.

(*V*2) *Die Regularitätsbedingung* $m_1 \ge r_1$, $\mu_1 > 0$ *und* $m_2 \ge r_2$, $\mu_2 > 0$ *ist erfüllt.*

Dann besitzt das reduzierte Glättungsproblem (4.12), (4.9) *eine Lösung* $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$.

Wir erhalten ferner die Glattheit des reduzierten Funktionals und durch Anwendung von Theorem 4.1 die Äquivalenz von vollständigem und reduziertem Glättungsproblem.

Theorem 4.3 (Äquivalenz von vollständigem und reduziertem Problem)**.**
Sei $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$ *eine zulässige Knotenfolge, d. h.*

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

Für feste $r_1 \in \{0, \ldots, q_1\}$, $0 \leq q_1 < k_1$ *und* $r_2 \in \{0, \ldots, q_2\}$, $0 \leq q_2 < k_2$ *gelte:*

$(V1)$ *Die Knoten erfüllen die Bedingung* $\tau^1_{j_1} < \tau^1_{j_1+k_1-q_1}$ $(j_1 = q_1 + 1, \ldots, n_1)$ *und* $\tau^2_{j_2} < \tau^2_{j_2+k_2-q_2}$ $(j_2 = q_2 + 1, \ldots, n_2)$.

$(V2)$ *Die Regularitätsbedingung* $m_1 \geq r_1$, $\mu_1 > 0$ *und* $m_2 \geq r_2$, $\mu_2 > 0$ *ist erfüllt.*

$(V3)$ *Die freien Knoten* $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$ *sind einfache Knoten. Es gilt* $k_1 \geq 3$ *und* $k_2 \geq 3$.

Dann gelten für das vollständige Glättungsproblem (4.11), (4.9) *und das reduzierte Glättungsproblem* (4.12), (4.9) *die Beziehungen: Die reduzierte Funktion* $\mathbf{F}$ *ist glatt im zulässigen Bereich* $\left\{ (\mathbf{t}^1, \mathbf{t}^2) \in \mathbb{R}^{l_1} \times \mathbb{R}^{l_2} \,:\, \mathbf{C}_1\mathbf{t}^1 - \mathbf{h}^1 \geq \mathbf{0}, \mathbf{C}_1\mathbf{t}^1 - \mathbf{h}^1 \geq \mathbf{0} \right\}$.

(a) Wenn $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$ *ein kritischer Punkt (oder eine globale Minimumstelle) von* (4.12), (4.9) *ist und* $\mathbf{A}^*$ *erfüllt* (4.7b) *an der Stelle* $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$, *so ist* $(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$ *ein kritischer Punkt (oder eine globale Minimumstelle) von* (4.11), (4.9) *und es gilt* $f(\mathbf{t}^{1*}, \mathbf{t}^{2*}) = \mathfrak{f}(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$.

(b) Wenn $(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$ *eine globale Minimumstelle von* (4.11), (4.9) *ist, so ist* $(\mathbf{t}^{1*}, \mathbf{t}^{2*})$ *globale Minimumstelle von* (4.12), (4.9). *Es gilt* $f(\mathbf{t}^{1*}, \mathbf{t}^{2*}) = \mathfrak{f}(\mathbf{t}^{1*}, \mathbf{t}^{2*}, \mathbf{A}^*)$ *sowie* (4.7b).

Die Aussagen der letzten beiden Sätzen sind nach den univariaten Vorbetrachtungen nicht sonderlich überraschend. Man beachte jedoch, daß der konkreten Gestalt von Gradient und Jacobi-Matrix aus Abschnitt 4.2 eine mindestens ebenso große praktische Bedeutung zukommt.

Die Aussagen gelten sinngemäß auch im Fall der Approximation von unregelmäßig verteilten Daten durch Tensorprodukt-B-Splines. In diesem Fall *muß* man jedoch glätten, um die Differenzierbarkeit des reduzierten Funktionals zu sichern. In [HH74] und anderen Arbeiten findet man Beispiele aus realen Anwendungen, die zeigen, daß die Vollrangeigenschaft von $\tilde{\mathbf{B}}(\mathbf{t}^1, \mathbf{t}^2)$ *nicht* erfüllt ist. Durch die in der Literatur beschriebenen Techniken kann zwar eine eindeutige Lösung zu festen Knoten bestimmt werden, dies sichert jedoch nicht den konstanten Rang für alle zulässigen Knoten.

4.4 Numerische Lösung des reduzierten Problems

Nachdem wir die Äquivalenz von vollständigem und reduziertem Problem im Sinne von Theorem 4.3 gezeigt haben, widmen wir uns nun der numerischen Lösung des reduzierten Problems. Das reduzierte Problem ist wiederum ein nichtlineares Quadratmittelproblem,

diesmal in den Variablen $\mathbf{t}^1$ und $\mathbf{t}^2$, mit linearen Ungleichheitsnebenbedingungen, welches wir mit unserem Basisalgorithmus 2.1 lösen können.

Das vollständige Problem hat formal große Ähnlichkeit mit einem separablen Quadratmittelproblem mit mehreren rechten Seiten, welches in der Form

$$\frac{1}{2}\,\|\mathbf{Z}-\mathbf{B}_1(\mathbf{t}^1)\mathbf{A}\|_F^2 \to \min_{\mathbf{t}^1,\mathbf{A}}$$

dargestellt werden kann, siehe [GL79] und [KS92]. Eine naive Realisierung überführt in unserem Fall das Problem $\frac{1}{2}\|\mathbf{Z}-\mathbf{B}_1(\mathbf{t}^1)\mathbf{A}\mathbf{B}_2^T(\mathbf{t}^2)\|_F^2 \to \min$ in „Standardform“ $\frac{1}{2}\|\operatorname{vec}(\mathbf{Z}) - (\mathbf{B}_2(\mathbf{t}^2)\otimes\mathbf{B}_1(\mathbf{t}^1))\operatorname{vec}(\mathbf{A})\|_2^2 \to \min$ und wendet an dieser Stelle die Reduktionstechnik an. Zur Berechnung der Jacobi-Matrix des reduzierten Problems ist dann die Matrix $\mathbf{B}_2\otimes\mathbf{B}_1 \in \mathbb{R}^{m_1m_2,n_1n_2}$ zu faktorisieren. Betrachtet man jedoch die obige Struktur genauer, so erkennt man, daß in den einzelnen Blöcken jeweils die gleiche Fréchet-Ableitung vorkommt. Diese Struktur kann (und muß) man bei realen Problemen ausnutzen. Für separable Quadratmittelprobleme mit mehreren rechten Seiten wurde dies in [GL79] erstmals durchgeführt.

In jedem Schritt des verallgemeinerten Gauß-Newton-Verfahrens ist das quadratische Modellproblem

$$\mu_{GP}(\mathbf{t}^1+\Delta\mathbf{t}^1,\mathbf{t}^2+\Delta\mathbf{t}^2) := \frac{1}{2}\,\|\mathbf{F}(\mathbf{t}^1,\mathbf{t}^2)+\boldsymbol{\partial}_1\mathbf{F}(\mathbf{t}^1,\mathbf{t}^2)[\Delta\mathbf{t}^1]+\boldsymbol{\partial}_2\mathbf{F}(\mathbf{t}^1,\mathbf{t}^2)[\Delta\mathbf{t}^2]\|_F^2 \to \min_{\Delta\mathbf{t}^1\in\mathbb{R}^{l_1},\Delta\mathbf{t}^2\in\mathbb{R}^{l_2}}$$

mit den Nebenbedingungen

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

zu lösen. Für das Gauß-Newton-Modell μ_{GP} erhalten wir

$$\begin{aligned}
\mu_{GP} &= \frac{1}{2}\,\|\mathbf{F}+\boldsymbol{\partial}_1\mathbf{F}\Delta\mathbf{t}^1+\boldsymbol{\partial}_2\mathbf{F}\Delta\mathbf{t}^2\|_F^2\\
&= \frac{1}{2}\left\|\mathbf{F}+\sum_{\kappa=1}^{l_1}\boldsymbol{\partial}_1\mathbf{F}[\mathbf{e}^\kappa]\Delta\mathbf{t}^1_\kappa+\sum_{\kappa=1}^{l_2}\boldsymbol{\partial}_2\mathbf{F}[\mathbf{e}^\kappa]\Delta\mathbf{t}^2_\kappa\right\|_F^2\\
&= \frac{1}{2}\left\|\operatorname{vec}(\mathbf{F})+\sum_{\kappa=1}^{l_1}\operatorname{vec}\left(\boldsymbol{\partial}_1\mathbf{F}[\mathbf{e}^\kappa]\right)\Delta\mathbf{t}^1_\kappa+\sum_{\kappa=1}^{l_2}\operatorname{vec}\left(\boldsymbol{\partial}_2\mathbf{F}[\mathbf{e}^\kappa]\right)\Delta\mathbf{t}^2_\kappa\right\|_2^2\\
&= \frac{1}{2}\left\|\operatorname{vec}(\mathbf{F})+\mathbf{J}\begin{pmatrix}\Delta\mathbf{t}^1\\ \Delta\mathbf{t}^2\end{pmatrix}\right\|_2^2
\end{aligned}$$

mit

$$\mathbf{J} := \begin{bmatrix} | & & | & | & & | \\ \operatorname{vec}\left(\boldsymbol{\partial}_1\mathbf{F}[\mathbf{e}^1]\right) & \cdots & \operatorname{vec}\left(\boldsymbol{\partial}_1\mathbf{F}[\mathbf{e}^{l_1}]\right) & \operatorname{vec}\left(\boldsymbol{\partial}_2\mathbf{F}[\mathbf{e}^1]\right) & \cdots & \operatorname{vec}\left(\boldsymbol{\partial}_2\mathbf{F}[\mathbf{e}^{l_2}]\right) \\ | & & | & | & & | \end{bmatrix}$$

Die Jacobi-Matrix $\mathbf{J}\in\mathbb{R}^{(m_1+n_1-r_1)(m_2+n_2-r_2),l_1+l_2}$ im Glättungsfall (bzw. $\mathbf{J}\in\mathbb{R}^{m_1m_2,l_1+l_2}$ im Approximationsfall) kann spaltenweise berechnet werden. Sie ist i. allg. vollbesetzt.

Berechnet man die Jacobi-Matrix auf die obige Art, so sind die notwendigen QR-Faktorisierungen von $\mathbf{B}_1$ und $\mathbf{B}_2$ für $\boldsymbol{\partial}_1\mathbf{F}$ und $\boldsymbol{\partial}_2\mathbf{F}$ nur jeweils einmal durchzuführen und dann auf verschiedene rechte Seiten anzuwenden. Sämtliche Algorithmen – einschließlich der Verwendung der Kaufman-Approximation und der Ausnutzung der Schwachbesetztheitsstrukturen – können direkt vom univariaten Fall übernommen werden. Wir erhalten etwa für $\boldsymbol{\partial}_1\mathbf{F}$ im Fall der Spline-Approximation

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

und für die Kaufman-Approximation

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

Diese Verfahrensweise hat große Ähnlichkeit mit der Berechnung der Jacobi-Matrix für separable Quadratmittelprobleme mit mehreren rechten Seiten. Setzen wir $\mathbf{B}_2(\mathbf{t}^2) = \mathbf{I}$, so erhalten wir unmittelbar die Ergebnisse von Golub/LeVeque als Spezialfall. Die Strukturausnutzung bei solchen Problemen wurde in den letzten Jahren intensiv untersucht [KS92], [KSW94], [GK92], [SB92]. Kaufman/Sylvester berichten über eine drastische Aufwandsreduzierung bei realen Problemen mit mehreren Tausend Parametern und Millionen von Beobachtungen, wobei die linearen Quadratmittelprobleme vollbesetzt waren. Die zusätzliche Schwachbesetztheit der Beobachtungsmatrix untersuchen Soo/Bates, u. a. auch am Beispiel der „self-modeling free-knot splines“, allerdings nicht im Tensorprodukt-Fall.

Die Algorithmen zur Berechnung der Jacobi-Matrix und somit das verallgemeinerte Gauß-Newton-Verfahren lassen sich also ohne große Probleme auf den bivariaten Fall übertragen.

4.5 Numerische Tests

Für die numerischen Tests zur Berechnung von bivariaten Tensorprodukt-Splines mit freien Knoten wurde ein Verfahren zur Lösung des reduzierten Problems – vorerst ohne Ausnutzung der Feinstruktur, d. h. der Schwachbesetztheit der Systemmatrizen – in MATLAB implementiert. Auf Grund der schlechten Verfügbarkeit von Software zur Lösung von nichtlinearen Quadratmittelproblemen mit linearen Nebenbedingungen haben wir das Problem als allgemeines nichtlineares Optimierungsproblem behandelt. Wir haben den Optimierungsalgorithmus CONSTR aus der MATLAB Optimization Toolbox [Gra90] sowie das kommerzielle Paket NPSOL [GMSW86] für unsere Tests verwendet. Beide Verfahren sind SQP-Verfahren, welche einen BFGS-Update der Hesse-Matrix durchführen. Die Gradienten wurden über finite Differenzen berechnet. Der wesentliche Unterschied zwischen beiden Verfahren besteht darin, daß NPSOL die linearen Nebenbedingungen ausnutzen kann und deshalb nur mit zulässigen Punkten arbeitet.

Obwohl die Algorithmen – im Gegensatz zu den univariaten Verfahren – ohne Ausnutzung der Feinstruktur implementiert wurden, ergeben sich durchaus akzeptable Rechenzeiten. Eine weitere Verbesserung wird durch die Verwendung eines Algorithmus erwartet, welcher die Quadratmittelstruktur ausnutzt. Leider stand das in Frage kommende Verfahren NLSSOL zum Zeitpunkt des Tests noch nicht zur Verfügung.

4.5.1 Bivariate Titanium Heat Data

In einem ersten Beispiel betrachten wir die bekannten Titanium Heat Data. Durch *Tensorisieren* der univariaten Daten, d.h. $z_{i_1,i_2} = y_{i_1} \times y_{i_2}$, erhalten wir $m_1 = 49 \times m_2 = 49$ Punkte im Bereich $[595, 1075] \times [595, 1075]$, siehe Abbildung 4.1. Diese 2401 Datenpunkte wollen wir durch $n_1 = 11$ kubische B-Splines in x-Richtung und $n_2 = 9$ kubische B-Splines in y-Richtung approximieren. Wir erhalten $l_1 = 7$ freie Knoten in x-Richtung bzw. $l_2 = 5$ in y-Richtung.

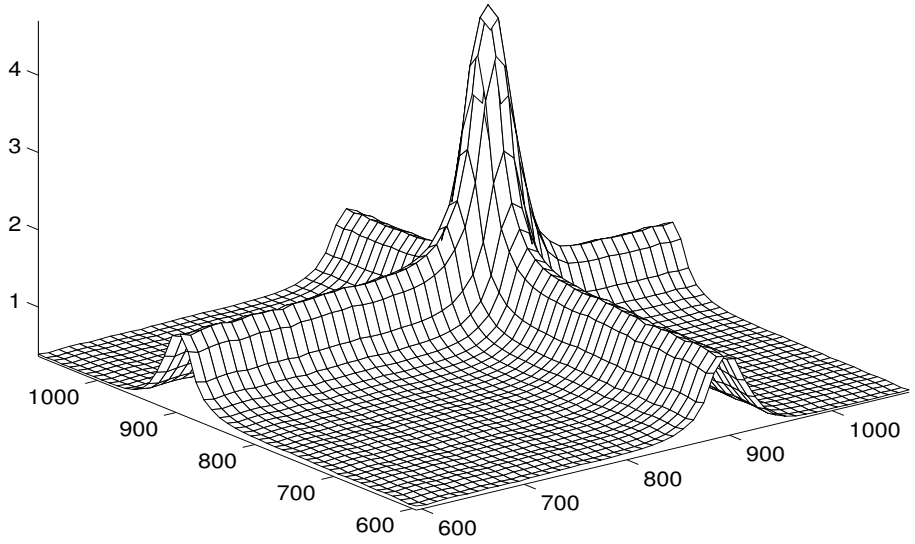


Abbildung 4.1: Bivariate Titanium Heat Data: Datenpunkte

Verwenden wir äquidistante innere Knoten als Startpunkt, so erhalten wir die Approximation in Abbildung 4.2. Neben den großen Oszillationen in den flachen Bereichen erkennt man, daß insbesondere der Peak sehr schlecht wiedergegeben wird. Durch Optimierung der Knoten verschwinden diese Oszillationen und das Residuum sinkt auf ca. 17%. In Tabelle 4.1 sind die Ergebnisse zusammengefaßt. Die MATLAB-Routine kann die geforderte Genauigkeit nicht erreichen. Der resultierende Spline nach der Optimierung mit NPSOL ist in Abbildung 4.3 dargestellt. Die Lage der Knoten sowie die zugehörigen Contour-Plots vor und nach der Optimierung zeigt Abbildung 4.4.

	Startknotenfolge	CONSTR	NPSOL
$\|\mathbf{F}\|$	9.049841 E+00	1.580966 E+00	1.560459 E+00
Schritte		93	101
func. calls		1200	1466
Zeit [s]		51.26	59.27
Ret. Code		max. no. iterations	successful

Tabelle 4.1: Bivariate Titanium Heat Data: Vergleich von CONSTR und NPSOL

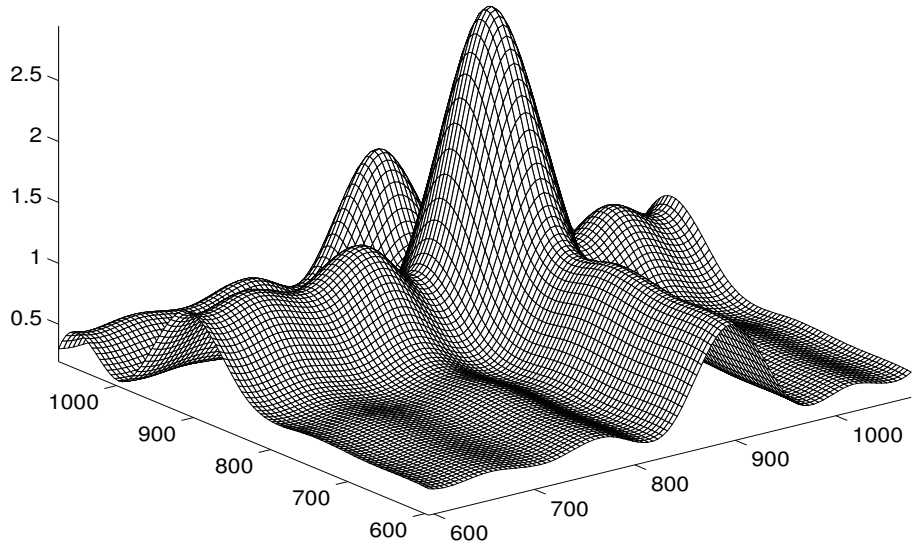


Abbildung 4.2: Bivariate Titanium Heat Data: Spline s, Startknotenfolge

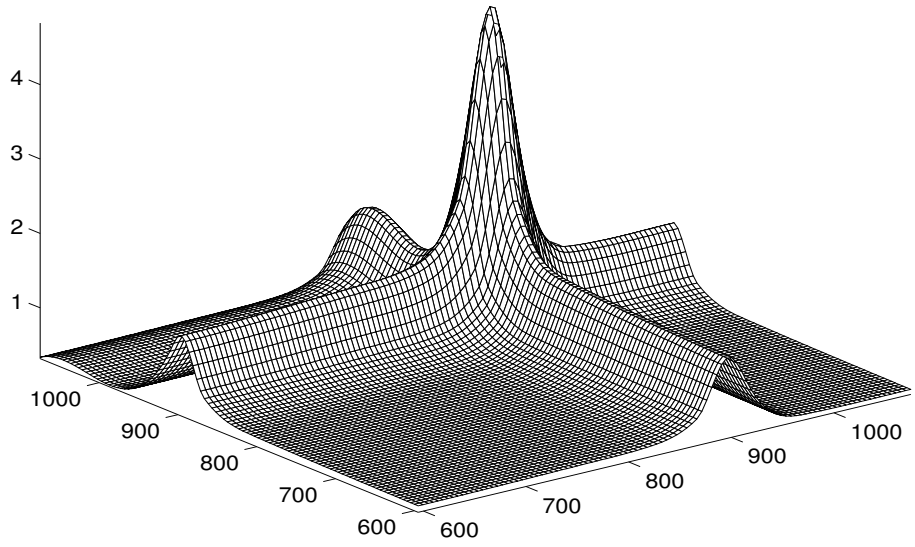


Abbildung 4.3: Bivariate Titanium Heat Data: Optimierte Knoten, NPSOL

Das obige Beispiel macht noch einmal die Bedeutung einer guten Knotenwahl deutlich. Obwohl sich die Residuen nicht um Größenordnungen unterscheiden, ist der Approximant zu äquidistanten inneren Knoten auf Grund der hohen Oszillationen praktisch unbrauchbar. Durch Minimierung des Quadratmittelfehlers bez. der freien Knoten verringert sich nicht nur dieser Fehler, sondern auch die unnötigen Oszillationen verschwinden weitgehend. Eine

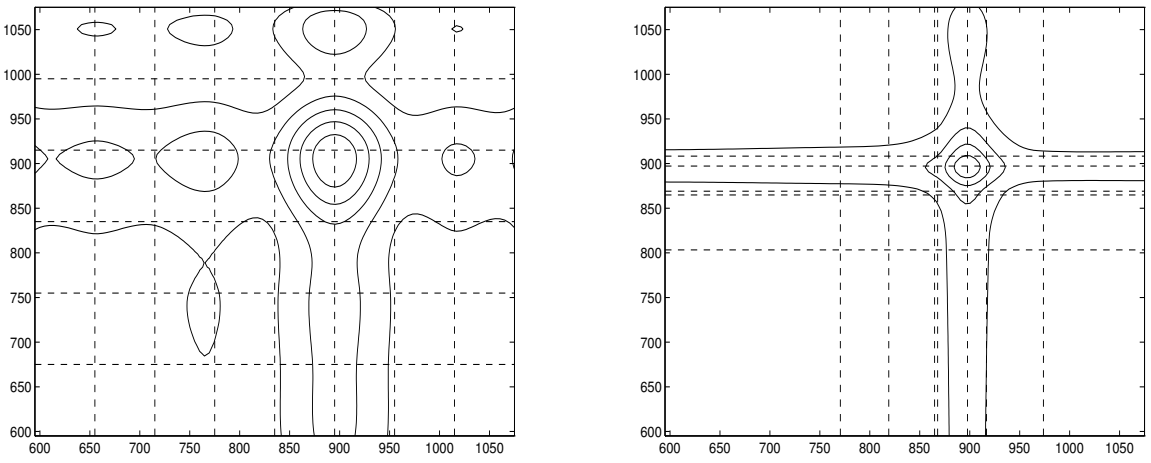


Abbildung 4.4: Bivariate Titanium Heat Data: Contour-Linien und Knoten vor und nach der Optimierung

	Startknotenfolge	CONSTR	NPSOL
$\|\mathbf{F}\|$	1.587922 E+01	1.228815 E+00	1.228815 E+00
Schritte		80	52
func. calls		705	573
Zeit [s]		18.82	14.17
Ret. Code		max. no. iterations	solution found, accuracy problems

Tabelle 4.2: EOS Aluminium Daten: Vergleich von CONSTR und NPSOL

Tendenz bei der Bewertung der beiden Optimierungsverfahren zeichnet sich schon in diesem Beispiel ab: Die Routine NPSOL ist wesentlich robuster als CONSTR. Wenn beide Routinen funktionieren, so unterscheiden sich die gefundenen lokalen Minima i. allg. nicht wesentlich. Der Vorteil der größeren Robustheit von NPSOL liegt in der Ausnutzung der *linearen* Nebenbedingungen begründet. Dies wird besonders an dem nächsten Beispiel deutlich.

4.5.2 EOS Aluminium Daten

Jetzt betrachten wir ein Standardbeispiel zur bivariaten *restringierten* Approximation, siehe z. B. [CF85]. Die $m_1 = 10 \times m_2 = 6$ Punkte im Bereich $[-0.07, 1.13] \times [-2.3, 0]$ beschreiben eine Zustandsfläche (equation of state, EOS) von Aluminium. Dargestellt ist die Spannung als Funktion von Dichte und Temperatur auf einer log-log-Skale. Die Daten sind in monotoner Lage, vgl. Abbildung 4.5.

Obwohl dieser Datensatz relativ klein ist, erweist er sich als schwierig zu approximieren. Wir benutzen $n_1 = 8$ quadratische B-Splines in x-Richtung, $n_2 = 5$ quadratische B-Splines in y-Richtung und die Glättungsfaktoren $\mu_1 = \mu_2 = 1.0\,\text{E-}08$ sowie $r_1 = r_2 = 2$. Wählt man die $l_1 = 5$, $l_2 = 2$ freien inneren Knoten äquidistant, so erhält man die unbefriedigende Approximation in Abbildung 4.6. Wir bemerken, daß beide Verfahren im Fall der Splineapproximation ($\mu_1 = \mu_2 = 0$) abbrechen, da zwischenzeitlich rangdefiziente Beobachtungsmatrizen – und damit Verlust der Differenzierbarkeit! – auftreten.

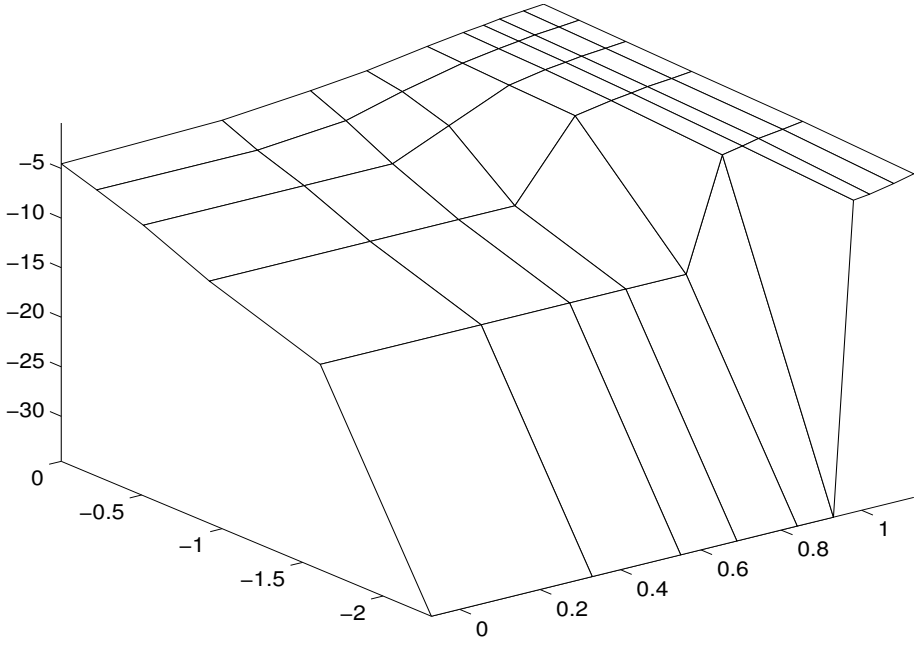


Abbildung 4.5: EOS Aluminium Daten: Datenpunkte

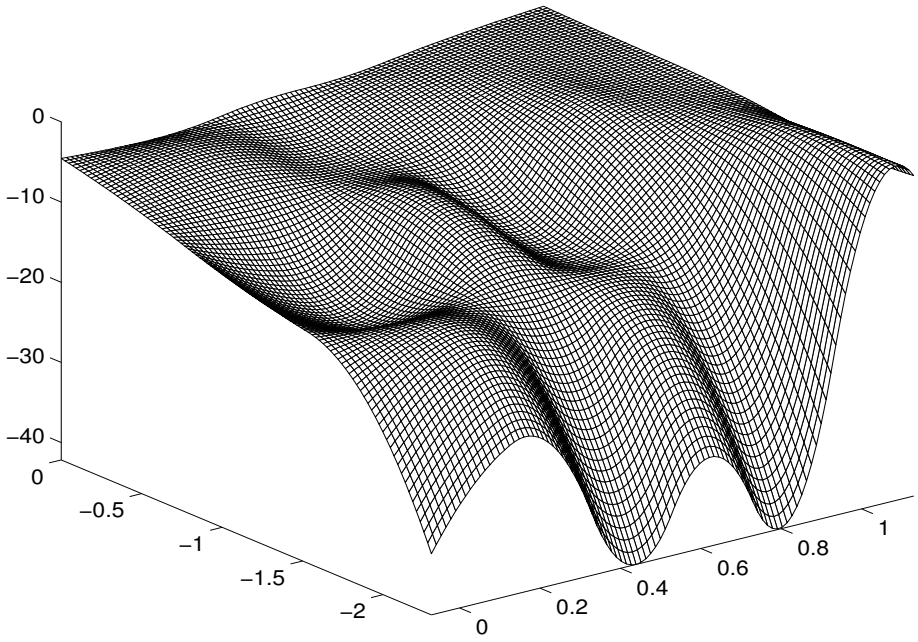


Abbildung 4.6: EOS Aluminium Daten: Spline s, Startknotenfolge

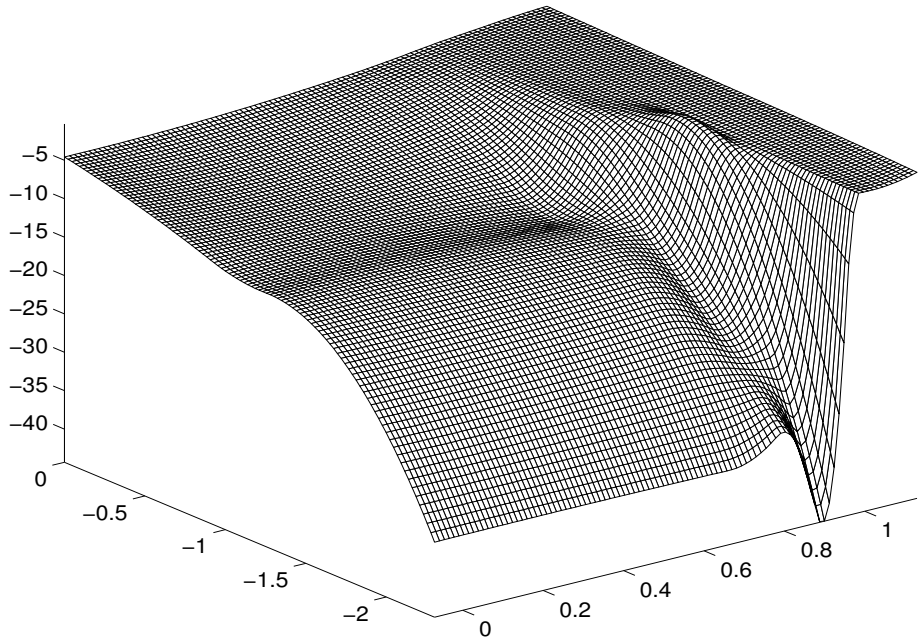


Abbildung 4.7: EOS Aluminium Daten: Optimierte Knoten, CONSTR

Optimiert man die Lage der Knoten, so erhält man etwa mit CONSTR den Spline in Abbildung 4.7. Die wesentlichen Oszillationen sind bei diesem Spline verschwunden. Tabelle 4.2 zeigt die Residuen der entsprechenden Splines, die Lage der Knoten wird in Abbildung 4.8 gezeigt. Man beachte, daß der resultierende Spline-Approximant *fast* monoton ist, es gilt z. B. $\min s_y \approx -8.03$, $\max s_y \approx 86.62$!

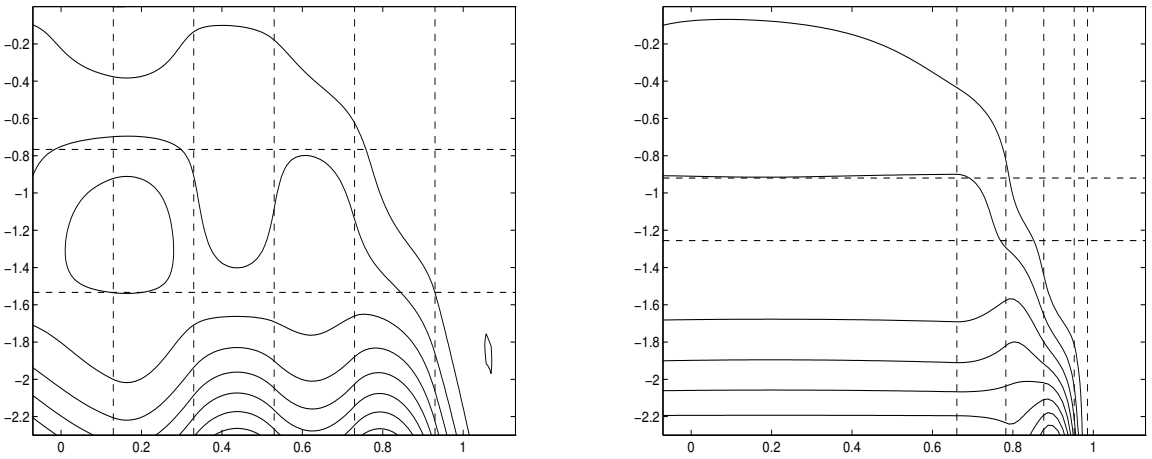


Abbildung 4.8: EOS Aluminium Daten: Contour-Linien und Knoten vor und nach der Optimierung

Die letzte Bemerkung legt eine weitere Anwendungsmöglichkeit der bivariaten Splines mit freien Knoten nahe: Bei der sog. *fit-and-modify*-Methode zur restringierten Interpolation werden gute Schätzwerte für Ableitungen benötigt. Die Parameter des restringierten Splines werden dann so bestimmt, daß der Spline möglichst wenig von dem vorgegebenen Spline abweicht, jedoch die Nebenbedingungen erfüllt. Da die Ableitungswerte der bivariaten Spli-

Approximationsverfahren	Approximation	Interpolation
Typ der Optimierungsprobleme	nichtlinear	quadratisch
Nebenbedingungen an Ableitungen	unrestringiert	Monotonie
Rechenzeit [s]	14.07	218.61
$\min s_y$	-8	0
$\max s_y$	86	905

Tabelle 4.3: Vergleich der Verfahren von Schütze und Walther

nes mit freien Knoten i. allg. eine sehr gute Qualität haben, eignen sich diese Splines gut als Startverfahren für *fit-and-modify*-Methoden der restringierten Interpolation. Abbildung 4.9 zeigt einen monotonen interpolierenden Spline, welcher mit Methoden aus [SW97] berechnet wurde. Tabelle 4.3 vergleicht die beiden Verfahren am Beispiel der EOS Aluminium Daten. Von einer Verbindung der beiden Verfahren werden wesentliche Vorteile erwartet, da die unrestringierten Ausgangssplines *fast optimal* sind (bessere Qualität der Splines) und ein guter Startpunkt für das Iterationsverfahren bekannt ist (kürzere Rechenzeit).

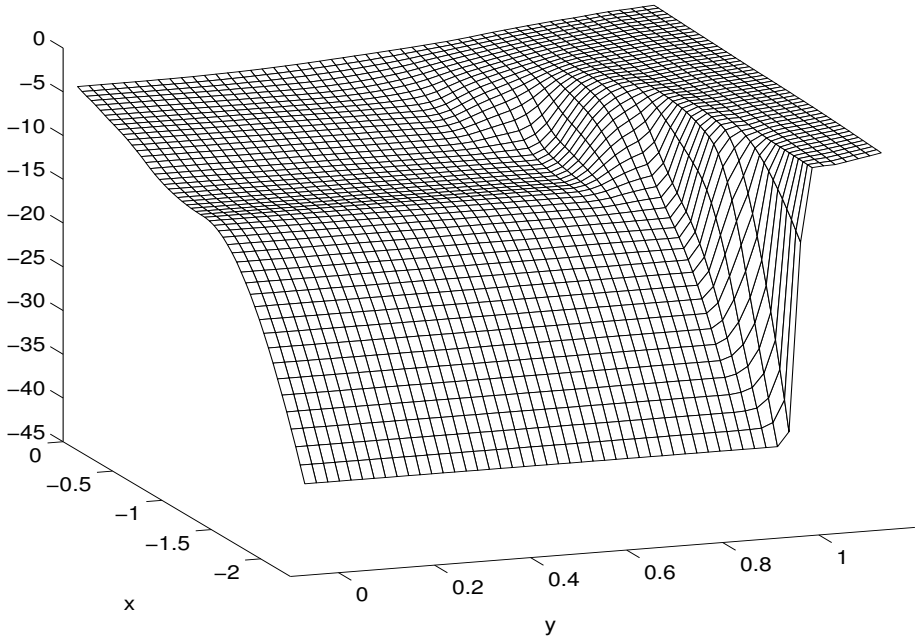


Abbildung 4.9: EOS Aluminium Daten: Verfahren von Walther, monotoner Spline (fit-and-modify)

Kapitel 5

Zusammenfassung und Ausblick

In dieser Dissertation werden Verfahren zur diskreten Quadratmittelapproximation durch Splines mit freien Knoten vorgestellt. Wir wählen dabei den direkten Zugang zur Splinetheorie und verwenden a priori den Raum $\mathcal{S}_{k,\boldsymbol{\tau}}$ der polynomialen Splines der Ordnung k zu den Knoten $\boldsymbol{\tau}$. Als Basis für diesen Raum benutzen wir die bekannten polynomialen B-Splines. Diese Wahl gestattet eine stabile Berechnung des Splines und liefert Matrizen mit Bandstruktur für die entsprechenden Optimierungsprobleme. Die Parameter des Splines, d. h. die Koeffizienten und (freien) Knoten, werden nun so bestimmt, daß das Schoenberg-Funktional minimal wird. Durch die Betrachtung von Splines mit freien Knoten erreicht man insbesondere bei Daten von nichtglatten Funktionen eine wesentliche Verbesserung der Approximation.

In Kapitel 2 untersuchen wir zunächst die Glättung durch univariate Splines mit freien Knoten ohne Nebenbedingungen an Ableitungen. Nachdem das Schoenberg-Funktional in Abhängigkeit von Koeffizienten und freien Knoten ausgedrückt wurde, formulieren wir Anordnungsnebenbedingungen an die Knoten, welche das Zusammenfallen der Knoten verhindern. Unter Benutzung von Ergebnissen der Theorie separabler Quadratmittelprobleme wird ein reduziertes Problem in den freien Knoten entwickelt. Durch Verwendung des Schoenberg-Funktionals an Stelle des Quadratmittelfehlers kann die Äquivalenz von vollständigem und reduziertem Problem unabhängig von der Lage der Knoten gezeigt werden. Das reduzierte Problem wird schließlich mit einem verallgemeinerten Gauß-Newton-Verfahren gelöst, wobei der effizienten Berechnung der Residuumsfunktion und der Jacobi-Matrix besondere Bedeutung zukommt, insbesondere im Hinblick auf die Ausnutzung der Bandstruktur. Der Algorithmus zur Optimierung der Lage der Knoten wird mit einem Datenreduktionsalgorithmus kombiniert.

Kapitel 3 liefert den Hauptbeitrag dieser Arbeit. Es beschäftigt sich mit der Minimierung des Schoenberg-Funktionals unter Beachtung von Schrankennebenbedingungen an Ableitungen. Wir verwenden hinreichende Bedingungen, welche linear in den Koeffizienten sind. Nachdem zunächst die numerische Berechnung des restringierten Splines zu festen Knoten zusammengefaßt wird, untersuchen wir Bedingungen, welche die strikte Konsistenz der Nebenbedingungen sichern. Unter Verwendung von Ergebnissen aus [Par85] für restringierte semi-lineare Quadratmittelprobleme kann ebenfalls die Äquivalenz der Probleme nachgewiesen werden. Da die Struktur der Jacobi-Matrix im restringierten Fall sehr kompliziert ist, verwenden wir eine Approximation, welche wesentlich billiger zu berechnen ist und außerdem die Ausnutzung der Schwachbesetztheit erlaubt. Es wird gezeigt, daß diese Ap-

proximation die gleichen qualitativen Eigenschaften wie im unrestringierten Fall hat.

Im Kapitel 4 wird schließlich die Problemstellung auf die bivariate Glättung durch Tensorprodukt-Splines mit Daten auf Rechteckgittern verallgemeinert. Es wird ein separabler Glättungsterm verwendet, bei welchem die linearen Probleme zu festen Knoten zerfallen. Eine Verallgemeinerung auf Probleme mit unregelmäßig verteilten Daten ist einfach.

Alle Algorithmen dieser Arbeit wurden implementiert und an zahlreichen Beispielen getestet. Die numerischen Tests zeigen, daß die Algorithmen zur Knotenoptimierung ein effizientes und robustes Werkzeug zur Quadratmittelapproximation durch Splines mit freien Knoten sind. Die Verfahren aus den Kapiteln 2 und 3 sind in einem umfangreichen Programmpaket [Sch96] implementiert und werden bereits von Dritten erfolgreich eingesetzt.

Wir möchten an dieser Stelle auf mögliche Verbesserungen und Vorschläge für die weitere Arbeit eingehen: Zunächst einmal sollte eine portable Implementierung der Verfahren angestrebt werden, da die derzeitige Fassung auf Grund von systemspezifischen Eigenheiten auf 100–200 Datenpunkte beschränkt ist. Zudem sollte bei der bivariaten Approximation auch die Feinstruktur der Matrizen ausgenutzt werden. Die Ergebnisse für diesen Fall liegen vor und lassen sich unmittelbar übertragen.

Ein nächster Schritt wäre die Ausdehnung der Tensorprodukt-Approximation auf unregelmäßig verteilte Daten. Obwohl hier keine theoretischen Schwierigkeiten zu erwarten sind, muß man bei der numerischen Realisierung beachten, daß die Probleme zu festen Knoten nicht mehr zerfallen.

Ein naheliegender nächster Schritt ist dann natürlich die Verbindung von Nebenbedingungen an Ableitungen und Splines mit freien Knoten. In diesem Fall liegen selbst für die Approximation durch Tensorprodukt-Splines mit *festen* Knoten und formerhaltenden Nebenbedingungen noch keine Ergebnisse vor. Ein möglicher Ansatzpunkt ist die Übertragung der Tensorprodukt-Techniken bei der restringierten Interpolation, wie sie in einer Reihe von Arbeiten [MS94], [Sch92b], [SW97] betrachtet wurden, auf den Quadratmittelfall. Unter Benutzung eines Ergebnisses aus [MS94] (siehe auch [Mul97]), dem sog. *Nichtnegativitätslemma*, ist es möglich, aus hinreichenden Nebenbedingungen im univariaten Fall hinreichende Bedingungen für den bivariaten Fall zu erhalten. Wir möchten den Ansatz kurz am Beispiel der bi-monotonen Approximation erläutern:

Zunächst betrachten wir die univariaten Splines

$$s^1(x) = \sum_{j_1=1}^{n_1} B_{j_1,k_1,\boldsymbol{\tau}^1}(x)\alpha^1_{j_1} \quad \text{und} \quad s^2(y) = \sum_{j_2=1}^{n_2} B_{j_2,k_2,\boldsymbol{\tau}^2}(y)\alpha^2_{j_2}.$$

Hinreichende Nebenbedingungen für die Nichtnegativität können durch die Funktionale ϕ_j, $\tilde{\phi}_j$ mit den Werten

$$\phi_{j_1}(s^1) = \alpha^1_{j_1} \quad (j_1 = 1,\ldots,n_1) \quad \text{und} \quad \tilde{\phi}_{j_2}(s^2) = \alpha^2_{j_2} \quad (j_2 = 1,\ldots,n_2)$$

beschrieben werden, d. h. aus $\phi_{j_1}(s^1) \geq 0\ \forall j_1$ folgt $s^1(x) \geq 0$, analog für $\tilde{\phi}$. Hinreichende Nebenbedingungen für die Monotonie werden durch

$$\varphi_{j_1}(s^1) = \frac{\alpha^1_{j_1} - \alpha^1_{j_1-1}}{\frac{\tau^1_{j_1+k_1-1} - \tau^1_{j_1}}{k_1-1}} \quad (j_1 = 2,\ldots,n_1) \quad \text{und} \quad \tilde{\varphi}_{j_2}(s^2) = \frac{\alpha^2_{j_2} - \alpha^2_{j_2-1}}{\frac{\tau^2_{j_2+k_2-1} - \tau^2_{j_2}}{k_2-1}} \quad (j_2 = 2,\ldots,n_2)$$

dargestellt. Nach dem Nichtnegativitätslemma sind dann

$$(5.1) \quad (\varphi_{j_1} \otimes \tilde{\phi}_{j_2})s \geq 0 \quad (j_1 = 2,\ldots,n_1; j_2 = 1,\ldots,n_2) \quad (s^{1\prime}(x) \geq 0, s^2(y) \geq 0, \frac{\partial s}{\partial x} \geq 0),$$

d. h. die Monotonie in x-Richtung wird mit der Nichtnegativität in y-Richtung kombiniert und liefert die Monotonie des bivariaten Splines in x-Richtung, sowie

$$(5.2) \quad (\phi_{j_1} \otimes \tilde{\varphi}_{j_2})s \geq 0 \quad (j_1 = 1, \ldots, n_1; j_2 = 2, \ldots, n_2) \quad (s^1(x) \geq 0, s^{2\prime}(y) \geq 0, \frac{\partial s}{\partial y} \geq 0)$$

hinreichend für einen bi-monotonen Spline. Aus (5.1) erhält man z. B. die Bedingung

$$\frac{\alpha_{j_1,j_2} - \alpha_{j_1-1,j_2}}{\frac{\tau^1_{j_1+k_1-1} - \tau^1_{j_1}}{k_1-1}} \geq 0 \qquad (j_1 = 2, \ldots, n_1; j_2 = 1, \ldots, n_2)$$

an die Koeffizienten. Es ist möglich, dieses Vorgehen auf den Durchschnitt von verschobenen Kegeln und damit auf Schrankennebenbedingungen zu erweitern.

Die effiziente numerische Berechnung von Tensorprodukt-Splines (zu festen Knoten) im unrestringierten Fall beruht auf der Tatsache, daß der Interpolations- bzw. Approximationsoperator gerade der Tensorprodukt-Operator der *linearen* univariaten Operatoren ist, siehe [dB78]. Bei der formerhaltenden Approximation sind die entsprechenden Operatoren jedoch i. allg. nicht mehr linear. Mittels des Nichtnegativitätslemmas kann man zwar hinreichende Bedingungen aufstellen, es ist jedoch nicht klar, ob und wie die Tensorprodukt-Struktur bei der numerischen Berechnung des Splines ausgenutzt werden kann.

Ein weiteres Gebiet für zukünftige Forschung sind Splines mit freien Knoten auf Triangulierungen. Während es gerade in der FEM-Literatur eine Fülle von „heuristischen“ Verfahren (wie Gleichverteilung des Fehlers) gibt, wurde die direkte Minimierung eines Fehlerfunktionals als Funktion der Knoten bisher kaum untersucht. Ein erster Ansatz findet sich in [TB97], welche die unstetige stückweise L_2-Approximation von Funktionen betrachten. Jedoch wird in dieser Arbeit beim „Mesh Tangling“ mehr oder weniger willkürlich das entsprechende Dreieck entfernt. Die geeignete Formulierung der Bedingungen an Knoten, welche das Zusammenfallen verhindern und eine *allgemeine* Lage sichern, stellt bereits ein Problem dar.

Literaturverzeichnis

[ABF86] M. Al-Baali and R. Fletcher. An efficient line search for nonlinear least squares. *J. Optim. Theory Appl.*, 48(3):359–377, 1986.

[ADLM90] E. Arge, M. Dæhlen, T. Lyche, and K. Mørken. Constrained spline approximation of functions and data based on constrained knot removal. In J. C. Mason and M. G. Cox, editors, *Algorithms for Approximation II*, pages 4–20. Chapman and Hall, 1990.

[AE87] L. E. Andersson and T. Elfving. An algorithm for constrained interpolation. *SIAM J. Sci. Statist. Comp.*, 8(6):1012–1025, 1987.

[AE91] L. E. Andersson and T. Elfving. Interpolation and approximation by monotone cubic splines. *J. Approx. Theory*, 66:302–333, 1991.

[AE95] L. E. Andersson and T. Elfving. Best constrained approximation in Hilbert space and interpolation by cubic splines subject to obstacles. *SIAM J. Sci. Comput.*, 16(5):1209–1232, 1995.

[Bai94] M. J. Baines. Algorithms for optimal discontinuous piecewise linear and constant L_2 fits to continuous functions with adjustable nodes in one and two dimensions. *Math. Comput.*, 62(206):645–669, 1994.

[Bjö87] Å. Björck. Stability analysis of the method of seminormal equations for linear least squares problems. *Linear Algebra Appl.*, 88/89:31–48, 1987.

[Bjö96] Å. Björck. *Numerical methods for least squares problems.* SIAM, Philadelphia, 1996.

[Boc87] H. G. Bock. *Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen (Habilitationsschrift).* Bonner Mathematische Schriften Nr.183, Universität Bonn, 1987.

[Bre90] K. Brenner. Bivariate Quadratmittelapproximation mittels glättender Tensorprodukt-B-Splines. Diplomarbeit, Martin-Luther-Universität Halle/Wittenberg, 1990.

[BS78] D. L. Barrow and P. W. Smith. Asymptotic properties of best $L_2[0,1]$ approximation by splines with variable knots. *Quart. Appl. Math.*, 36:293–304, 1978.

[CF85] R. E. Carlson and F. N. Fritsch. Monotone piecewise bicubic interpolation. *SIAM J. Numer. Anal.*, 22(2):386–400, 1985.

[CJ87] M. G. Cox and H. Jones. Shape-preserving spline approximation in the l_1 norm. In J. C. Mason and M. G. Cox, editors, *Algorithms for approximation*, pages 115–129. Clarendon Press, Oxford, 1987.

[Cor81] C. Corradi. A note on the solution of separable nonlinear least squares problems with separable nonlinear equality constraints. *SIAM J. Numer. Anal.*, 18:1134–1138, 1981.

[Cox81] M. G. Cox. The least squares solution of overdetermined linear equations having band or augmented band structure. *IMA J. Numer. Anal.*, 1:3–22, 1981.

[Cox82] M. G. Cox. Practical spline approximation. In P. R. Turner, editor, *Topics in Numerical Analysis*, volume 965 of *Lect. Notes Math.*, pages 79–112. Springer, 1982.

[Cro79] L. J. Cromme. *Approximation auf Mannigfaltigkeiten mit Spitzen – Theorie und numerische Methoden.* Habilitationsschrift, Göttingen, 1979.

[Dan73] J. W. Daniel. Stability of definite quadratic programs. *Math. Prog.*, 5:41–53, 1973.

[dB76] C. de Boor. Splines as linear combinations of B-splines. In G. G. Lorentz, C. K. Chui, and L. L. Schumaker, editors, *Approximation Theory II*, pages 1–47. Academic Press, New York, 1976.

[dB78] C. de Boor. *A practical guide to splines.* Springer-Verlag, New York, Heidelberg, Berlin, 1978.

[dBH87] C. de Boor and K. Höllig. *B*-splines without divided differences. In G. E. Farin, editor, *Geometric Modeling: Algorithms and New Trends*, pages 21–27. SIAM Publications, Philadelphia, 1987.

[dBLS76] C. de Boor, T. Lyche, and L. L. Schumaker. On calculating with B-splines ii: Integration. In L. Collatz, H. Werner, and G. Meinardus, editors, *Numerische Methoden der Approximationstheorie*, pages 123–146. Birkhäuser, Basel, 1976.

[dBR68] C. de Boor and J. R. Rice. Least squares cubic spline approximation II – variable knots. Technical Report CSD TR 21, Computer Science Department, Purdue University, 1968.

[Die79] P. Dierckx. *Het aanpassen van krommen en oppervlakken aan meetpunten met behulp van spline funkties.* PhD thesis, Katholieke Universiteit Leuven, 1979.

[Die80] P. Dierckx. An algorithm for cubic spline fitting with convexity constraints. *Computing*, 24:349–371, 1980.

[Die81] P. Dierckx. An algorithm for surface-fitting with spline functions. *IMA J. Numer. Anal.*, 1:267–283, 1981.

[Die82] P. Dierckx. A fast algorithm for smoothing data on a rectangular grid while using spline functions. *SIAM J. Numer. Anal.*, 19(6):1286–1305, 1982.

[Die87] P. Dierckx. FITPACK user guide, part 1: Curve fitting routines. Technical Report TW Report 89, Department of Computer Science, Katholieke Universiteit Leuven, Belgium, 1987.

[Die89] P. Dierckx. FITPACK user guide, part 2: Surface fitting routines. Technical Report TW Report 122, Department of Computer Science, Katholieke Universiteit Leuven, Belgium, 1989.

[Die93] P. Dierckx. *Curve and Surface Fitting with Splines.* Oxford University Press, 1993.

[DS83] J. E. Dennis and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice Hall, Inc., 1983.

[DYS91] N. Dyn and I. Yad-Shalom. Optimal distribution of knots for tensor-product spline approximation. *Quart. Appl. Math.*, XLIX(1):19–27, 1991.

[EA88] T. Elfving and L. E. Andersson. An algorithm for computing constrained smoothing spline functions. *Numer. Math.*, 52:583–595, 1988.

[Eld84] L. Eldén. An algorithm for the regularization of ill-conditioned banded least squares problems. *SIAM J. Sci. Stat. Comput.*, 5:237–254, 1984.

[EMM89] S. Ewald, H. Mühlig, and B. Mulansky. Bivariate interpolating and smoothing tensor product splines. In J. W. Schmidt and H. Späth, editors, *Splines in Numerical Analysis, Proceed. ISAM-89*, volume 52 of *Mathematical Research*, pages 55–68. Akademie-Verlag, Berlin, 1989.

[Eub88] R. L. Eubank. *Spline Smoothing and Nonparametric Regression*, volume 90 of *Statistics, textbooks and monographs*. Marcel Dekker, New York, 1988.

[FF94] D. W. Fausett and C. T. Fulton. Large least squares problems involving Kronecker products. *SIAM J. Matrix Anal. Appl.*, 15(1):219–227, 1994.

[Fia76] A. V. Fiacco. Sensitivity analysis for nonlinear programming using penalty methods. *Math. Prog.*, 10:287–311, 1976.

[Fia83] A. V. Fiacco. *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*. Academic Press, 1983.

[FOP91] B. Fischer, G. Opfer, and M. L. Puri. A local algorithm for constructing non-negative cubic splines. *J. Approx. Theory*, 64:1–16, 1991.

[GK92] 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, pages 179–183, 1992.

[GL79] 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.

[GMSW86] 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, Standford University, 1986.

[GP73] G. H. Golub and V. Pereyra. The differentiation of pseudoinverses and nonlinear least squares problems whose variables separate. *SIAM J. Numer. Anal.*, 10:413–432, 1973.

[GP76] G. H. Golub and V. Pereyra. Differentiation of pseudoinverses, separable nonlinear least squares problems and other tales. In M. Nashed, editor, *Generalized Inverses and Applications*, pages 303–324. Academic Press, New York, 1976.

[Gra90] A. Grace. *Optimization Toolbox User's Guide*. The MathWorks, Inc., 1990.

[Gre91] H. Greiner. A survey on univariate data interploation and approximation by splines of given shape. *Math. Comput. Modelling*, 15:97–106, 1991.

[Han92] P. C. Hansen. Analysis of discrete ill-posed problems by means of the *L*-curve. *SIAM Review*, 34(4):561–580, 1992.

[HF79] J. N. Holt and R. Fletcher. An algorithm for constrained nonlinear least squares. *J. Inst. Math. Appl.*, 23:449–463, 1979.

[HH74] J. G. Hayes and J. Halliday. The least-squares fitting of cubic spline surfaces to general data sets. *J. Inst. Math. Appl.*, 14:89–103, 1974.

[Hig96] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, Philadelphia, 1996.

[Hor78] U. Hornung. Monotone Spline-Interpolation. In *Numerische Methoden der Approximationstheorie*, volume 42 of *ISNM*, pages 172–191. Birkhäuser, Basel, 1978.

[HS85] C. L. Hu and L. L. Schumaker. Bivariate natural spline smoothing. In G. Meinardus and G. Nürnberger, editors, *Delay Equations, Approximation and Applications, Mannheim 1994*, volume 74 of *ISNM*, pages 165–179. Birkhäuser, 1985.

[HS86] C. L. Hu and L. L. Schumaker. Complete spline smoothing. *Numer. Math.*, 49(1):1–10, 1986.

[Hu93] Y. Hu. An algorithm for data reduction using splines with free knots. *IMA J. Numer. Anal.*, 13(3):365–381, 1993.

[Jup75] D. L. B. Jupp. The lethargy theorem – a property of approximation by γ-polynomials. *J. Approx. Theory*, 14:204–217, 1975.

[Jup78] D. L. B. Jupp. Approximation to data by splines with free knots. *SIAM J. Numer. Anal.*, 15(2):328–343, 1978.

[Kau75] L. Kaufman. A variable projection method for solving separable nonlinear least squares problems. *BIT*, 15(4):49–57, 1975.

[KP78] L. Kaufman and V. Pereyra. A method for separable nonlinear least squares problems with separably nonlinear equality constraints. *SIAM J. Numer. Anal.*, 15:12–20, 1978.

[Kra94] R. Kraft. Hierarchical B-splines. Preprint 94-14, Universität Stuttgart, 1994.

[Kro74] F. T. Krogh. Efficient implementation of a variable projection algorithm for nonlinear least squares problems. *Comm. ACM*, 17(3):167–169, 1974.

[KS88] A. Kiełbasinski and H. Schwetlick. *Numerische Lineare Algebra*. Deutscher Verlag der Wissenschaften, Berlin, 1988.

[KS92] L. Kaufman and G. S. Sylvester. Separable nonlinear least squares with multiple right-hand sides. *SIAM J. Matrix Anal. Appl.*, 13(1):68–89, 1992.

[KSW94] 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(4):847–871, 1994.

[Kun95] V. Kunert. Ein Algorithmus zur Splineapproximation unter Nebenbedingungen. Manuskript, 1995.

[LH95] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. SIAM Publications, Philadelphia, 1995. Reprint of edition by Prentice Hall, 1974.

[LM87] T. Lyche and K. Mørken. Knot removal for parametric B-spline curves and surfaces. *Computer Aided Geometric Design*, 4:217–230, 1987.

[LM88] T. Lyche and K. Mørken. A data reduction strategy for splines with applications to the approximation of functions and data. *IMA J. Numer. Anal.*, 8:185–208, 1988.

[LW91] P. D. Loach and A. J. Wathen. On the best least squares approximation of continuous functions using linear splines with free knots. *IMA J. Numer. Anal.*, 11(3):393–409, 1991.

[MNSS89] G. Meinardus, G. Nürnberger, M. Sommer, and H. Strauss. Algorithms for piecewise polynomials and splines with free knots. *Math. Comp.*, 53(187):235–247, 1989.

[MNW96] G. Meinardus, G. Nürnberger, and G. Walz. Bivariate segment approximation and splines. *Adv. in Comput. Math.*, 6:25–45, 1996.

[MS94] B. Mulansky and J. W. Schmidt. Nonnegative interpolation by biquadratic splines on refined rectangular grids. In P. J. Laurent, A. Le Méhauté, and L. L. Schumaker, editors, *Wavelets, Images and Surface Fitting, Chamonix 1993*, pages 379–386. A K Peters Wellesley, 1994.

[MSSW85] C. A. Micchelli, P. W. Smith, J. Swetits, and J. D. Ward. Constrained L_p approximation. *Constr. Approx.*, 1:93–102, 1985.

[MU88] C. A. Micchelli and F. I. Utreras. Smoothing and interpolation in a convex subset of a Hilbert space. *SIAM J. Sci. Stat. Comput.*, 9(4):728–746, 1988.

[MU91] C. A. Micchelli and F. I. Utreras. Smoothing and interpolation in a convex subset of a Hilbert space: II. the semi-norm case. *Mathematical Modelling and Numerical Analyis*, 25(4):425–440, 1991.

[Mul90] B. Mulansky. Glättung mittels zweidimensionaler Tensorprodukt-Splinefunktionen. *Wiss. Z. Tech. Univ. Dresden*, 39:187–190, 1990.

[Mul92] B. Mulansky. Necessary conditions for local best Chebyshev approximations by splines with free knots. In D. Braess and L. L. Schumaker, editors, *Numerical Methods of Approximation Theory*, volume 9 of *International Series of Numerical Mathematics*, pages 195–206. Birkhäuser, 1992.

[Mul97] B. Mulansky. Tensor products of convex cones. In G. Nürnberger, J. W. Schmidt, and G. Walz, editors, *Multivariate Approximation and Splines*, ISNM, pages 167–176. Birkhäuser, Basel, 1997.

[Nür96] G. Nürnberger. Bivariate segment approximation and free knot splines: Research problems 96-4. *Constr. Approx.*, 12:555–558, 1996.

[OO88] G. Opfer and H. J. Oberle. The derivation of cubic splines with obstacles by methods of optimization and optimal control. *Numer. Math.*, 52:17–31, 1988.

[Pai73] C. C. Paige. An error analysis of a method for solving matrix equations. *Math. Comp.*, 27:355–359, 1973.

[Par85] T. A. Parks. *Reducible Nonlinear Programming Problems.* PhD thesis, Houston Univ., Dept. of Mathematics, Houston, 1985.

[Pig91] T. Pigorsch. Bivariate Quadratmittelapproximation unter Verwendung von Tensorprodukt-B-Splines im Falle von Rechteckgitterdaten. Diplomarbeit, Martin-Luther-Universität Halle-Wittenberg, 1991.

[Rei67] C. H. Reinsch. Smoothing by spline functions. *Numer. Math.*, 10:177–183, 1967.

[Rei71] C. H. Reinsch. Smoothing by spline functions II. *Numer. Math.*, 16:451–454, 1971.

[Ric69] J. R. Rice. *The Approximation of Functions II.* Addison Wesley, Reading, Massachusetts, 1969.

[Rie95] K. S. Riedel. Piecewise convex function estimation and model selection. In C. K. Chui and L. L. Schumaker, editors, *Proc. of Approximation Theory VIII*, pages 467–475. World Scientific Pub., 1995.

[RW80] A. Ruhe and P. Å. Wedin. Algorithms for separable nonlinear least squares problems. *SIAM Rev.*, 22(3):318–337, 1980.

[Sau72] M. A. Saunders. Large-scale linear programming using the Cholesky factorization. Technical Report Report No. CS252, Computer Science Dept., Stanford Univ., 1972.

[SB92] Y. W. Soo and D. M. Bates. Loosely coupled nonlinear least squares. *Computational Statistics and Data Analysis*, 14:249–259, 1992.

[Sch64] I. J. Schoenberg. On interpolation by spline functions and its minimal properties. In P. L. Butzer and J. Korevaar, editors, *On Approximation Theory*, volume 5 of *Internat. Ser. Numer. Math.*, pages 109–129. Birkhäuser, Basel-Stuttgart, 1964.

[Sch81] L. L. Schumaker. *Spline Functions: Basic Theory.* John Wiley and Sons, New York, 1981. Reprint Edition by Krieger Publ., 1993.

[Sch90] J. W. Schmidt. Monotone data smoothing by quadratic splines via dualization. *Z. Angew. Math. Mech.*, 70:299–307, 1990.

[Sch91] H. Schwetlick. Nichtlineare Parameterschätzung: Modelle, Schätzkriterien und numerische Algorithmen. *GAMM-Mitteilungen*, 2/91:13–51, 1991.

[Sch92a] J. W. Schmidt. Dual algorithms for solving convex partially separable optimization problems. *Jber. d. Dt. Math.-Verein.*, 94:40–62, 1992.

[Sch92b] J. W. Schmidt. Positive, monotone, and S-convex C^1-interpolation on rectangular grids. *Computing*, 48:363–371, 1992.

[Sch96] T. Schütze. FREE – A program for constrained approximation by splines with free knots. Preprint MATH-NM-04-1996, Technical University of Dresden, 1996.

[SK93] H. Schwetlick and V. Kunert. Spline smoothing under constraints on derivatives. *BIT*, 33:512–528, 1993.

[Spä95] H. Späth. *One Dimensional Spline Interpolation Algorithms.* AK Peters, Wellesley MA, 1995.

[SS90] J. W. Schmidt and I. Scholz. A dual algorithm for convex-concave data smoothing by cubic C^2-splines. *Numer. Math.*, 57:333–350, 1990.

[SS95] H. Schwetlick and T. Schütze. Least squares approximation by splines with free knots. *BIT*, 35(3):361–384, 1995.

[SS96] L. L. Schumaker and S. S. Stanley. Shape-preserving knot removal. *Computer Aided Geometric Design*, 13:851–872, 1996.

[SS97] T. Schütze and H. Schwetlick. Constrained approximation by splines with free knots. *BIT*, 37(1):105–137, 1997.

[Suc91] P. Suchomski. Method of optimal variable-knot spline interpolation in the L_2 discrete norm. *Internat. J. Systems Sci.*, 22(11):2263–2274, 1991.

[SW53] I. J. Schoenberg and A. Whitney. On Pólya frequency functions III: The positivity of translation determinants with an application to the interpolation problem by spline curves. *Trans. Amer. Math. Soc.*, 74:246–259, 1953.

[SW97] J. W. Schmidt and M. Walther. Gridded data interpolation with restrictions on the first order derivatives. In G. Nürnberger, J. W. Schmidt, and G. Walz, editors, *Multivariate Approximation and Splines*, ISNM, pages 291–307. Birkhäuser, Basel, 1997.

[TB97] Y. Tourigny and M. J. Baines. Analysis of an algorithm for generating locally optimal meshes for L_2 approximation by discontinous piecewise polynomials. *Math. Comp.*, pages 623–650, 1997.

[Utr91] F. I. Utreras. The variational approach to shape preservation. In P. J. Laurent, A. le Méhauté, and L. L. Schumaker, editors, *Curves and Surfaces, Proceedings Chamonix 1990*, pages 461–476. Academic Press, Boston, 1991.

[Var82] J. M. Varah. A spline least squares method for numerical parameter estimation in differential equations. *SIAM J. Sci. Statist. Comput.*, 3:28–46, 1982.

[VBH92] A. H. Vermeulen, R. H. Bartels, and G. R. Heppler. Integrating products of B-splines. *SIAM J. Sci. Statist. Comput.*, 13(4):1025–1038, 1992.

[Wah82] G. Wahba. Constrained regularization for ill-posed linear operator equations, with applications to meteorology and medicine. In S. S. Gupta and J. O. Bergers, editors, *Statistical Decision Theory and Related Topics III*, pages 383–418. Academic Press, New York, 1982.

[Wah90] G. Wahba. *Spline Models for Observational Data.* SIAM Publications, Philadelphia, 1990.

[WD95] K. Willemans and P. Dierckx. Nonnegative surface fitting with Powell-Sabin splines. *Numer. Algorith.*, 9:263–276, 1995.

[Wev89] U. Wever. *Darstellung von Kurven und Flächen mittels datenreduzierender Algorithmen.* Dissertation, TU München, 1989.

[WL69] H. Wold and E. Lyttkens. Nonlinear iterative partial least squares (NIPALS) estimation procedures. *Bull. ISI*, 43:29–51, 1969.

Versicherung

Hiermit versichere ich, daß ich die vorliegende Arbeit ohne unzulässsige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Die vorgelegte Dissertation wurde am Institut für Numerische Mathematik der Technischen Universität Dresden unter der wissenschaftlichen Betreuung von Herrn Prof. Dr. rer. nat. habil. H. Schwetlick angefertigt.

Dresden, den 2. September 1997 .....