Interpolationsbasierte Reduzierte-Basis-Modellierung von Lösungskurven mit Umkehrpunkten

Vom Promotionsausschuss der
Technischen Universität Hamburg-Harburg

zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigte Dissertation

von
Hagen Eichel

aus
Magdeburg

2016

Gutachter:
Prof. Dr. Wolfgang Mackens
Prof. Dr. Alexander Düster

Tag der mündlichen Prüfung:
25. Januar 2016

It is hard work and great art to make life not so serious.
- John Irving

Danksagung

Zunächst möchte ich meinem Doktorvater, Professor Wolfgang Mackens, für seine Betreuung danken. Ohne seine Unterstützung, Motivation und Hilfe bei der Bekämpfung gelegentlicher Panikattacken wäre diese Arbeit nicht zustande gekommen. Dem zweiten Gutachter dieser Arbeit, Professor Alexander Düster, gilt ebenfalls mein Dank.

Desweiteren möchte ich meinen Kollegen an der Technischen Universität Hamburg-Harburg danken, insbesonders Peter Baasch, Torge Schmidt und Nicolai Rehbein, die den Forschungs- und Lehralltag mit Diskussionen von Politik bis Prominenz ein gutes Stück unterhaltsamer gemacht haben.

Mein Dank gilt außerdem allen Freunden, die ich seit meiner Ankunft in Hamburg gewinnen konnte und deren Verdienst es ist, dass ich diese Stadt mein Zuhause nennen konnte. Dabei möchte ich mich insbesondere bei Eike Ketzler bedanken, einem wahren Freund auf dessen Rückhalt ich immer zählen konnte und der maßgeblich dazu beigetragen hat, die Zeit in Hamburg unvergesslich zu machen. An dieser Stelle soll außerdem Christian Baumann erwähnt werden, der nicht nur geholfen hat, einige Lücken in meinen Programmierfähigkeiten zu schließen, sondern auch in anderen Lebensbereichen stets wertvolle Ratschläge bereitgehalten hat.

Besonderer Dank gilt meinen Eltern, die mich immer unterstützt haben und denen möglicherweise nicht einmal völlig klar ist, wie groß ihr Beitrag zum Gelingen der Promotion war. Dem Rest meiner Familie sei an dieser Stelle ebenfalls gedankt.

Zum Schluss möchte ich meinem Partner Jon danken. Die Aussicht, nach langer Zeit der Hin- und Herreise endlich ein gemeinsames Leben beginnen zu können, hat sehr viel zur Motivation, die Arbeit abzuschließen, beigetragen.

Inhaltsverzeichnis

Abbildungsverzeichnis **iii**

1 Einleitung **1**
1.1 Motivation, Ziel und Ergebnisse 1
1.2 Aufbau der Arbeit 6

2 Lösungsmengen parameterabhängiger nichtlinearer Gleichungen **9**
2.1 Aussagen zur Lösungsexistenz 10
2.2 Einparametrige nichtlineare Gleichungen 14
2.2.1 Lösungsexistenz und Umkehrpunkte 14
2.2.2 Das Tangentialfeld 19

3 RB-Methoden **23**
3.1 Lokale RB-Methoden 23
3.1.1 Lokale Galerkin-Diskrektisierung 23
3.1.2 Ansatz- und Testraum 26
3.2 Globale RB-Methoden 28
3.2.1 Reduktion mittels Snapshots 29
3.2.2 Offline-Online-Berechnungen 32

4 Globale Ansatzräume für einparametrige nichtlineare Gleichungen **35**
4.1 Lagrange-Ansatzraum 36
4.2 POD-Ansatzraum 39
4.2.1 Proper Orthogonal Decomposition 39
4.2.2 Aufstellen eines Ansatzraums mittels POD 40
4.2.3 Abschätzung des Projektionsfehlers 42
4.2.4 Bestmögliche affine Verschiebung 44

5 Testräume für einparametrige nichtlineare Gleichungen **49**
5.1 Aufbau eines Testraumes mittels POD 51
5.2 Aufbau eines Testraumes mittels Tangentialfeld 53

5.3 Aufbau Lipschitz-stetiger Basen 62

6 Interpolationsbasierte Reduktion 71
6.1 Zerfallende Lösungskurven 71
6.2 Grundidee 73
6.3 Gewichtsfunktionen 77
6.4 Existenz einer Lösung 80
6.4.1 Interpolation mittels Lagrange-Ansatzraum 80
6.4.2 Interpolation mittels inexakter Knoten 91

7 Zweiparametrige Systeme und numerische Untersuchungen 103
7.1 Das verallgemeinerte Bratu-Problem 107
7.2 Exotherme Reaktion 112
7.3 Bifurkationspunkte 117

8 Abschließende Betrachtungen 123
8.1 Empirische Interpolation 123
8.2 Weiterführende Betrachtungen 126

Literaturverzeichnis 128

Lebenslauf 137

Abbildungsverzeichnis

2.1 Schrittweises Berechnen der Lösung von $F(u, \lambda) = 0$ aus Beispiel 2.2.1 16
2.2 Lösung von $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ und $\mathbf{F}(\mathbf{x}) = \mathbf{b}$ und das Tangentialfeld von $\mathbf{F}$ 22

3.1 Lokale RB-Methode 26

6.1 Erste Komponente der Lösung von $\mathbf{G}(\mathbf{u}, \lambda) = \mathbf{0}$ in Abhängigkeit von λ 73
6.2 Zerfallen der Lösungskurve der Reduktion in mehrere Lösungsäste 74
6.3 Lösungskurve, Interpolationsknoten und dazugehörige Träger der Gewichtsfunktionen 76
6.4 Graph der Funktion H 78
6.5 Skizze zum Beweis des Satzes 6.4.7 92

7.1 Skizze zum Beweis des Satzes 7.0.17 104
7.2 Erste Komponente der Lösung von $\mathbf{G}(\mathbf{u}, \lambda, 0) = \mathbf{0}$ in Abhängigkeit von λ 108
7.3 Auswahl von $d = 3$ und $d = 10$ Interpolationsknoten mittels Greedy-Algorithmus 109
7.4 Fehler der Approximation bei $d = 3$ Interpolationsknoten und steigender Dimension des Ansatzraumes 110
7.5 Verlauf der Lösung von $\mathbf{G}(\mathbf{u}, \lambda, \mu) = \mathbf{0}$ für $\mu \in [-1, 1]$. 111
7.6 Lösung und Approximation von $\mathbf{G}(\mathbf{u}, \lambda, \mu) = \mathbf{0}$ für $\mu \in [-1, 1]$ mittels für $\mu = 0$ bestimmter interpolationsbasierter Reduktion . 111
7.7 Fehler der Approximation bei $d = 3$ Interpolationsknoten für $\mu \in [-1, 1]$. 112
7.8 Korrektur der Interpolationsknoten 113
7.9 Lösungskurven für verschiedene $\mu \in [0, 2.1]$. 114
7.10 Vergleich zwischen Reduktion mit festem Testraum und interpolationsbasierter Reduktion 115
7.11 Lösungskurve für $\mu = 1.0$ mit 4 gleichverteilten Interpolationsknoten 116
7.12 Gestörter Bifurkationspunkt 116

7.13 Interpolation mit 7 Knoten ohne Bifurkation 117
7.14 Lösungen der interpolationsbasierten Reduktion für $\mu \in [0.5, 1.5]$ 118
7.15 Fehler der Approximation bei $d = 7$ Interpolationsknoten für $\mu \in [0, 1.5]$ 119
7.16 Lösungskurve von (7.2) 119
7.17 Oberer Teil der Lösungskurve und Interpolationsknoten 120
7.18 Approximation der Lösungskurve 121
7.19 Lösungskurven für verschiedene μ 122
7.20 Lösungskurven der interpolationsbasierten Reduktion für verschiedene μ 122

Kapitel 1

Einleitung

1.1 Motivation, Ziel und Ergebnisse

Diese Arbeit enthält Beiträge zur effizienten Berechnung mehrfach parametrisierter großer nichtlinearer Gleichungssysteme

$$\mathbf{F}(\mathbf{u}, \alpha) = \mathbf{0}. \tag{1.1}$$

mit $\mathbf{F} \in C^r(\Omega, \mathbb{R}^n), (r \geq 2)$, wobei $\Omega \in \mathbb{R}^n \times \mathbb{R}^k, n, k \in \mathbb{N}$ offen ist und auf ganz Ω $\operatorname{Rang}(\mathbf{DF}) = n$ gilt. Unter diesen Voraussetzungen ist die Lösungsgesamtheit in Ω eine glatte k-dimensionale Mannigfaltigkeit.

Für die direkte numerische Berechnung solcher Mengen für $k \geq 2$ sind relativ viele Methoden entwickelt worden, [57], wobei sich zusätzlich zu der Berechnung auch noch das Problem einer geeigneten Darstellung ergibt. In dieser Arbeit ist speziell der Fall $k = 2$ von Interesse, wobei zusätzlich einer der Parameter, hier mit λ bezeichnet, von primärem Interesse ist. Dies kann z.B. die Größe einer Last sein, die bei einem Problem aus der Mechanik zur Deformation eines betrachteten Körpers führt.

Der zweiten Parameter, bezeichnet mit μ, wird als zusätzlichen Parameter angesehen, bei dessen Variation die Veränderung der Gestalt der λ-abhängigen Lösungsgesamtheit untersucht werden soll. Die Zweiparameterstudien (1.1) wird so als Menge gestörter Einparameterprobleme

$$\mathbf{F}(\mathbf{u}, \lambda, \mu_i) = 0 \tag{1.2}$$

für eine Sequenz fester Parameter $\mu_1, \mu_2, \ldots$ behandelt. Dies hat mehrere Vorteile:

1. Es gibt heute zuverlässige Verfahren, eindimensionale Lösungsmannigfaltigkeiten zu berechnen (Siehe [2] für eine einführende Übersicht bis zum Jahr 2003, bzw. [31]).

2. Durch die Berechnung von eindimensionalen Lösungskurven wird auf der Lösungsfläche eine Art Koordinatensystem eingeführt, mit der die Fläche oft einfacher zu analysieren ist.

3. Die Aufteilung in wichtigere Parameter (λ) und nachgeordnete Parameter (μ) entspricht sehr oft auch den behandelten Anwendungssituationen, bei denen Strukturveränderungen der Lösungsgesamtheit im $(\mathbf{u}, \lambda)$-Raum bei Variation von μ von Interesse sind, [60].

Weil Systeme der Form (1.1) heute vorwiegend durch Diskretisierung partieller Differentialgleichungen entstehen, kommt zur Schwierigkeit der Parameterabhängigkeit zusätzlich noch das Problem einer großen Dimension n hinzu. Strukturell besonders einfache Herangehensweisen an die Aufwandreduktion insbesondere bei parameterabhängigen Probleme sind die Reduzierte-Basis-Methoden (RB-Methoden). Diese gehören zu den Galerkin-Methoden, die man vor allem aus der Finite-Elemente-Methode kennt, bei der Operatorgleichungen in Hilberträumen betrachtet werden. Diese haben die Form $Au - f = 0$, mit einem Operator A und einer gesuchten Lösung u. Bei der Approximation dieser Lösung u wird nach Näherungslösungen $u_{\mathcal{Z}}$ in (endlichdimensionalen) Unterräumen $\mathcal{Z}$ gesucht, sodass das auftretende Residuum orthogonal zu einem geeigneten Testraum $\mathcal{V}$ steht. Es ergeben sich so Gleichungen der Form

$$\langle v, Au_{\mathcal{Z}} - f \rangle = 0, \quad \forall v \in \mathcal{V},$$

welche zu einem linearen Gleichungssystem für den Koeffizientenvektor $\mathbf{u}_{\mathcal{Z}}$ von $u_{\mathcal{Z}}$ der Form

$$\mathbf{V}^T(\mathbf{A}\mathbf{u}_{\mathcal{Z}} - \mathbf{f}) = \mathbf{0}$$

führen. Bei der RB-Methode wird ein solcher Ansatz auf die hier betrachteten nichtlinearen Gleichungssysteme angewendet und die Gleichung (1.2) durch einen Galerkin-Ansatz

$$\mathbf{V}^T\mathbf{F}(\mathbf{u}_0 + \mathbf{Z}\hat{\mathbf{u}}, \lambda, \mu_i) = \mathbf{0} \tag{1.3}$$

approximiert, wobei die Spalten von $\mathbf{Z}$ und $\mathbf{V}$ jeweils eine Basis von $\mathcal{Z}$, bzw. $\mathcal{V}$ bilden, die in einer Analysephase dem Problem angepasst entwickelt werden. Da die Lösungsgesamtheit sich im allgemeinen nicht nach λ parametrisieren lässt, hat es sich für die Astverfolgungsverfahren als günstiger herausgestellt den Parameter λ zusammen mit dem Zustandsvektor $\mathbf{u}$ in einen übergeordneten Zustandsvektor

$$\mathbf{x} = (\mathbf{u}^T, \lambda)^T \in \mathbb{R}^{n+1}$$

zu integrieren, [60, 49], und (1.2) und (1.3) als

$$\begin{aligned} \mathbf{F}(\mathbf{x}, \mu_i) &= \mathbf{0}, \text{ bzw.} \\ \mathbf{V}^T\mathbf{F}(\mathbf{x}_0 + \mathbf{Z}\hat{\mathbf{x}}, \mu_i) &= \mathbf{0}, \ i = 1, 2, \ldots \end{aligned} \tag{1.4}$$

mit einem angepassten Matrix $\mathbf{Z}$ umzuformulieren. Der Punkt $\mathbf{x}_0 = (\mathbf{u}_0^T, \lambda_0)$ stellt dabei eine bekannte Lösung von $\mathbf{F}(\mathbf{x}, \mu_i) = \mathbf{0}$ dar (diese muss natürlich nicht für alle μ_i die selbe sein). In den achtziger Jahren wurde - besonders in den Ingenieurwissenschaften - viele solcher Rechenmethoden entwickelt, [42]. Eine erste mathematische Analyse findet sich in [60, 20] und [49].

Dabei wurde die Aufmerksamkeit darauf gelegt, die Lösungsgesamtheit in einer Umgebung eines Startpunktes in der Lösungsmenge zu finden. Diese Methoden kann man also unter dem Begriff lokale RB-Methoden zusammen fassen. Die Basisfunktionen werden hier basierend auf den differentialgeometrischen Eigenschaften der Lösungskurve, wie Tangente, Krümmungsvektor und Torsion, aufgebaut. Zudem wurden notwendige und hinreichende Bedingungen an $\mathbf{V}$ entwickelt unter denen das reduzierte System ebenfalls eine eindimensionale Lösungsgesamtheit hat, [20, 49, 7, 38] und es wurde in einer lokalen Fehleranalyse gezeigt, dass die Lösungskurve durch die Approximationen im wesentlichen wie durch entsprechende Taylorentwicklungen approximiert wurden.

Seit etwa 2000 interessiert man sich verstärkt für parameterabhängige Differential- und Evolutionsgleichungen, wobei hier globale RB-Methoden mit einem Ansatz- und Testraum, der nicht an eine Lösung $\mathbf{x}_0$ gekoppelt ist, verwendet werden, [47, 63]. Diese basieren auf der sogenannten Snapshot-Methode, bei der zunächst eine Menge von Snapshots gesammelt wird. Diese stellen Lösungen des volldimensionalen Ausgangsproblems für eine Auswahl von Parametern dar, mit deren Hilfe dann ein globaler Ansatzraum $\mathcal{Z}$ erzeugt wird. Besonderes Augenmerk wird dabei auf die Unterscheidung in Off- und Online-Anteil der Berechnung der Reduktion gelegt, wobei die aufwändige Arbeit (Sammeln der Snapshots, Aufstellen des Ansatz- und Testraumes) im Offline-Teil erfolgt und die vom Aufwand her kleine Berechnung der reduzierten Lösung für gegebene Parameter im Online-Teil. Diese Unterteilung bietet einen Vorteil gegenüber den lokalen Methoden, da dort die Berechnung von Ansatz- und Testraum nicht von der Berechnung der reduzierten Lösung getrennt werden kann.

Um mit einem globalen Ansatz- und Testraum rechnen zu können, wird die nichtlineare Gleichung in der Form (1.1) betrachtet, das heißt die Lösung $\mathbf{u}$ und die Parameter α bleiben strikt voneinander getrennt und die Gleichung wird durch

$$\mathbf{V}^T\mathbf{F}(\mathbf{Z}\hat{\mathbf{u}}, \alpha) = \mathbf{0} \tag{1.5}$$

approximiert. Die Nachteile dieser Methode bestehen darin, dass keine Umkehrpunkte zugelassen sind, sowie in den starken Restriktionen an die zu re-

duzierenden Gleichungssysteme, die notwendig sind, um die Lösbarkeit der Reduktion zu garantieren.

In dieser Arbeit werden Möglichkeiten gezeigt, die globalen Ansätze auf allgemeine nichtlineare Probleme, die vorher nur lokal reduziert wurden, anzuwenden. Ziel ist es dabei mit globalen Ansatz- und Testräumen eine Reduktion aufzubauen, die Wendepunkte bezüglich λ zulässt und mit deren Hilfe effektiv Parameterstudien für den zweiten auftretenden Parameter μ durchgeführt werden können. Die dafür an die Funktion zu stellenden Bedingungen sind dabei so allgemein wie möglich gehalten, um eine breite Anwendbarkeit zu ermöglichen.

Es zeigt sich, dass eine direkte Anwendung der Snapshot-Methode, das heißt der Aufbau eines festen Ansatz- und Testraumes für ein allgemeines nichtlineares Problem der Form (1.2) nicht ohne weiteres möglich ist. Dabei sind weniger die Auswahl der Snapshots und der Aufbau des Ansatzraumes $\mathcal{Z}$ das Problem, sondern das Erzeugen eines geeigneten Testraums $\mathcal{V}$. Dieser muss garantieren, dass das reduzierte Problem der Form (1.3) überhaupt eine Lösung besitzt, bzw. eine sinnvolle Reduktion ermöglichen.

Dieses Problem wird bei globalen RB-Methoden durch aus der Finiten-Elemente-Methode bekannte Bedingungen, die an das Problem gestellt werden, wie Koerzivität und Stetigkeit der Bilinearform, gelöst. Diese Bedingungen stellen jedoch eine starke Einschränkung dar, bzw. lassen sich für den Fall eines nicht ausgezeichneten Parameters nicht garantieren, da selbst das volldimensionale Problem (1.1) für ein festes λ keine eindeutige Lösung besitzen muss.

Im Fall einer lokalen RB-Reduktion lässt sich die Lösbarkeit von (1.3) direkt dadurch gewährleisten, dass $\mathcal{V}$ in Abhängigkeit des bezüglich $\mathbf{x}_0$ erzeugten Ansatzraumes $\mathcal{Z}$ aufgebaut wird, [38]. In den meisten Fällen ist dabei die Ableitung $\mathbf{c}'$ in $\mathbf{x}_0$ Teil von $\mathcal{Z}$, was jedoch für einen über Snapshots aufgebauten globalen Ansatzraum nicht gewährleistet werden kann. Numerische Beispiele zeigen, dass es ohne weitere Einschränkungen im Allgemeinen nicht möglich ist, ein $\mathcal{V}$ zu finden, das eine sinnvolle Reduktion für einen globalen über die Snapshot-Methode erzeugten Ansatzraum $\mathcal{Z}$ liefert. Es kann sogar passieren, dass ein $\mathcal{V}$, das in einem Snapshot eine geeignete Reduktion ermöglicht, die reduzierte Lösungskurve an einer anderen Stelle aufbrechen und in mehrere voneinander getrennte Äste zerfallen lässt.

Um dieses Problem in den Griff zu bekommen, wird in dieser Arbeit eine neue RB-Methode präsentiert, die auf einem festen Ansatzraum $\mathcal{Z}$, aber lokalen Testräumen $\mathcal{V}_i$ basiert, die für eine Menge von Punkten $\mathbf{x}_i \in \mathcal{Z}$, Interpolationsknoten genannt, einzeln erzeugt werden. Um daraus schließlich eine globale Reduktion zu erzeugen, werden die die $\mathcal{V}_i$ repräsentierenden Matrizen $\mathbf{V}_i$ mittels einer über eine Zerlegung der Eins aufgebauten Interpolation miteinander

verbunden. Diese orientiert sich an einer in [56] vorgestellten Idee und hat die Gestalt

$$\sum_{i=0}^{d} w_i(\mathbf{Z}\hat{\mathbf{x}})\mathbf{V}_i^T\mathbf{F}(\mathbf{Z}\hat{\mathbf{x}}) = \mathbf{0} \tag{1.6}$$

wobei w_i Gewichtsfunktionen mit beschränktem Träger beschreiben. Diese Methode verbindet lokale Reduktionen zu einer globalen, die eine reduzierte Lösungskurve in einem festen Ansatzraum $\mathcal{Z}$ liefert und wird als interpolationsbasierte Reduktion bezeichnet.

Um zu gewährleisten, dass eine Reduktion der Form (1.6) eine Lösung besitzt werden Bedingungen an die Matrizen $\mathbf{V}_i$ hergeleitet, sowie Verfahren vorgestellt, Matrizen aufzubauen, die diese Bedingungen erfüllen. Für den Fall, dass der von den Interpolationsknoten aufgespannte Raum mit $\mathcal{Z}$ überein stimmt, führt dies zum Beweis eines Satzes, der garantiert, dass die Reduktion (1.6) unter gewissen Regularitätsbedingungen an $\mathbf{F}$ eine zusammenhängende Lösungskurve im Lagrange-Raum $\mathcal{Z}$ besitzt, so lange die ausgewählten Interpolationsknoten genügend nah beieinander liegen.

Eine populäre Methode, einen globalen Ansatzraum aufzubauen, besteht darin, diesen über eine Singulärwertzerlegung der Snapshots zu erzeugen, [34, 53]. Diese sogenannte POD-Methode nutzt aus, dass ein über die Singulärwertzerlegung erzeugter Raum den Projektionsfehler der Snapshots minimiert und in diesem Sinne den bestmöglichen Raum liefert, um die Bewegung der Lösungskurve widerzuspiegeln. Der Unterschied zum bisher verwendeten Lagrange-Raum besteht hauptsächlich darin, dass der Ansatzraum keinen Lösungspunkt des volldimensionalen Ausgangsproblems (1.2) mehr enthalten muss, und bei Verwendung der interpolationsbasierten Reduktion die Knoten nicht mehr auf der Lösungskurve liegen müssen. Dadurch entsteht eine gewisse Unabhängigkeit von $\mathcal{Z}$ und den Interpolationsknoten. Ähnlich zum Lagrange-Ansatzraum wird in dieser Arbeit ein Satz bewiesen, der die Existenz einer reduzierten Lösungskurve für den Fall eines über die POD-Methode aufgebauten Ansatzraumes $\mathcal{Z}$ garantiert, so lange $\mathbf{F}$ gewissen Regularitätsbedingungen genügt, die Dimension des Ansatzraumes groß genug ist und die Knoten nah genug beeinander liegen.

Mit einer auf diese Weise aufgebauten Reduktion können Parameterstudien bezüglich eines zweiten Parameters μ durchgeführt werden. Zu diesem Zweck werden Kurven betrachtet, die bezüglich μ eine gewisse numerische Stabilität aufweisen und bewiesen, dass es stets ein Intervall gibt, sodass eine Lösungskurve von 1.4 für alle μ_i in diesem Intervall existiert. Anhand numerischer Beispiele zeigt sich, dass die einmal aufgestellte Reduktion in einigen Fällen ohne Anpassung weiter verwendet werden kann.

Im Ausblick beschäftigt sich diese Arbeit noch mit der Empirischen Interpo-

lation. Diese für globale RB-Methoden genutzte Methode hat zwei Aufgaben. Einerseits wird sie verwendet, nichtaffine Parameterabhängigkeiten aufzulösen, in dem sie den nichtlinearen Anteil einer Funktion approximiert, [14, 22]. Dies ist für die Offline-Online-Zerlegung von Bedeutung, da für diese eine affine Parameterabhängigkeit notwendig ist. Andererseits kann sie allgemein dafür verwendet werden, den Rechenaufwand eines bereits reduzierten Problems zu verringern, was vor allem für den Fall, dass die Auswertung von $\mathbf{F}$ selbst sehr rechenaufwändig ist, von Bedeutung ist, [14]. Es werden Möglichkeiten präsentiert, die empirische Interpolation auf das hier vorgestellte reduzierte Problem (1.6) anzuwenden.

1.2 Aufbau der Arbeit

In Kapitel 2 wird zunächst die grundlegende Theorie über die Lösungsexistenz für Gleichungen der Form (1.1) erläutert. Dabei wird eine Version des Satzes über implizite Funktionen basierend auf [9] bewiesen. Danach werden die Regularitätsbedingungen, die an die Funktion $\mathbf{F}$ gestellt werden müssen, genauer erklärt.

In Kapitel 3 werden die lokalen und globalen RB-Methoden zusammen mit ihren Vor- und Nachteilen beleuchtet. Dabei wird vor allem auf die Existenz von Umkehrpunkten eingegangen, die für lokale Methoden kein Problem darstellen, bei den globalen Methoden jedoch ausgeschlossen sind. Die populäre Snapshot-Methode, mit Lagrange- und POD-Ansatzraum, wird dabei genauer betrachtet.

In den Kapiteln 4 und 5 wird sich der Frage zugewendet, inwiefern sich die Snapshot-Methode auf ein allgemeines nichtlineares Problem übertragen lässt. Dabei werden Möglichkeiten zum Aufbau des Ansatzraumes $\mathcal{Z}$ als Lagrange-Raum, sowie mittels POD präsentiert. Es wird erläutert, wo die Schwierigkeiten beim Anwenden von globalen Methoden auf allgemeine nichtlineare Gleichungen liegen. Des Weiteren wird gezeigt, wie sich Testräume und deren Basen bezüglich der Punkte im Raum aufbauen lassen, um deren Stetigkeit zu gewährleisten.

In Kapitel 6 wird die neue RB-Methode (1.6) basierend auf einem Interpolationsansatz vorgestellt. Es wird bewiesen, dass unter gewissen Stetigkeitsforderungen an die die Testräume $\mathcal{V}_i$ repräsentierenden Matrizen $\mathbf{V}_i$, die Interpolationsknoten stets so gewählt werden können, dass das reduzierte System eine Lösung besitzt. Dies wird sowohl für den Fall, dass ein Lagrange-Ansatzraum als auch für den Fall, dass ein POD-Ansatzraum verwendet wird, bewiesen.

Schließlich wird in Kapitel 7 ein zusätzlicher Parameter μ betrachtet. Es werden Bedingungen an $\mathbf{F}$ hergeleitet, die garantieren, dass ein Intervall D existiert, sodass für alle $\mu \in D$ eine Lösungskurve existiert. Des Weiteren werden numerische Beispiele präsentiert, die bestimmte auftretende Phänomene

bei einer interpolationsbasierten Reduktion zeigen, sowie ihre Anwendung zur Durchführung von Parameterstudien bezüglich μ.

in Kapitel 8 wird schließlich als Ausblick die empirische Interpolation betrachtet und gezeigt, wie diese sich auf das interpolationsbasierte RB-Verfahren (1.6) anwenden lässt.

Kapitel 2

Lösungsmengen parameterabhängiger nichtlinearer Gleichungen

Die Problemstellung, die Lösungsmenge einer nichtlinearen Gleichung

$$\mathbf{F}(\mathbf{u}, \lambda) = \mathbf{0} \tag{2.1}$$

mit $\mathbf{F} \in C^r(\Omega, \mathbb{R}^n), r \geq 1$, mit einer offenen Teilmenge $\Omega \subset \mathbb{R}^n \times \mathbb{R}^k$, zu beschreiben wurde vielfach untersucht, [57, 9], und lässt sich vereinfacht ausgedrückt wie folgt zusammen fassen:

Ist in einem Punkt $(\mathbf{u}_0, \lambda_0)$ mit $\mathbf{F}(\mathbf{u}_0, \lambda_0) = \mathbf{0}$ die Jacobimatrix $\mathbf{D_u F}(\mathbf{u}_0, \lambda_0)$ invertierbar, so lässt sich die Lösung nach $\mathbf{u}$ auflösen und bildet in einer Umgebung von $(\mathbf{u}_0, \lambda_0)$ eine k-dimensionale Mannigfaltigkeit.

Dies ist das Ergebnis der Anwendung des Satzes über implizite Funktionen. Da dieser Satz im späteren Verlauf der Arbeit einige Male Verwendung findet, wird er im Folgenden in einer an die hier untersuchten Probleme angepassten Version nach [9] wiedergegeben. Zusätzlich werden einige hilfreiche daraus resultierende Aussagen getroffen. Diese stellen die notwendigen Werkzeuge dar, um die Lösungsmenge der in der Arbeit auftretenden verschiedenen reduzierten Systeme zu untersuchen.

Die Größe der Umgebung um $(\mathbf{u}_0, \lambda_0)$, in der die Existenz einer Lösungsmannigfaltigkeit gesichert werden kann, hängt von den Eigenschaften von $\mathbf{F}$, bzw. $\mathbf{DF}$ ab. In der hier verwendeten Version werden diese Eigenschaften benannt und die Zusammenhänge mit der Größe der Lösungsumgebung genau dargelegt. Dies ist notwendig um den Satz über implizite Funktionen auf die in Kapitel 6 vorgestellte interpolationsbasierte Reduktion anwenden zu können.

Im Gegensatz zu den gängigsten Varianten des Satzes ist es in der Variante nach [9] nicht notwendig, dass der Ausgangspunkt $(\mathbf{u}_0, \lambda_0)$ selbst Lösung des nichtlinearen Gleichungssystems ist. Diese Bedingung wird durch eine notwendige Beschränktheit der Norm von $\mathbf{F}(\mathbf{u}_0, \lambda_0)$ ersetzt. Dies ist vor allem für die

in Kapitel 6.4.2 betrachtete Reduktion von Bedeutung, da dort der Ansatzraum nicht notwendigerweise Lösungen des volldimensionalen Systems (2.1) enthält.

2.1 Aussagen zur Lösungsexistenz

Seien mit

$$B(\mathbf{u}_0;\delta) := \{\mathbf{u} \in \mathbb{R}^n : \|\mathbf{u}-\mathbf{u}_0\| \leq \delta\} \text{ und}$$
$$B(\mathbf{u}_0,\lambda_0;\delta) := \{(\mathbf{u},\lambda) \in \mathbb{R}^n \times \mathbb{R}^k : \sqrt{\|\mathbf{u}-\mathbf{u}_0\|^2 + \|\lambda-\lambda_0\|^2} \leq \delta\}$$

die abgeschlossenen Kugeln um $\mathbf{u}_0$, bzw. $(\mathbf{u}_0,\lambda_0) \in \mathbb{R}^n \times \mathbb{R}^k$ mit Radius δ bezeichnet, wobei mit $\|\cdot\|$ die jeweilige euklidische Norm gemeint ist. Bevor Aussagen über die Lösungsmenge parameterabhängiger Gleichungen getroffen werden können, benötigt man den Satz über inverse Funktionen.

Lemma 2.1.1 (Satz über inverse Funktionen)**.** *Sei $\mathbf{u}_0 \in \mathbb{R}^n$ und $\mathbf{H}$ eine in einer Umgebung von $\mathbf{u}_0$ definierte, stetig differenzierbare Funktion mit Bild in $\mathbb{R}^n$. Sei weiterhin die Jacobimatrix $\mathbf{DH}(\mathbf{u}_0)$ regulär mit*

$$\|\mathbf{DH}(\mathbf{u}_0)^{-1}\| \leq M. \tag{2.2}$$

Wählt man $\delta > 0$ so, dass für alle $\mathbf{u} \in B(\mathbf{u}_0;\delta)$

$$\|\mathbf{DH}(\mathbf{u}) - \mathbf{DH}(\mathbf{u}_0)\| \leq \frac{1}{2M} \tag{2.3}$$

erfüllt ist, dann existiert mit $\mathbf{y}_0 := \mathbf{H}(\mathbf{u}_0)$ eine eindeutige C^1-Funktion $\mathbf{q} : B(\mathbf{y}_0;\delta/(2M)) \to B(\mathbf{u}_0;\delta)$, sodass

$$\mathbf{H}(\mathbf{q}(\mathbf{y})) = \mathbf{y}$$

mit $\mathbf{q}(\mathbf{y}_0) = \mathbf{u}_0$ gilt. Des Weiteren ergibt sich für alle $\mathbf{y} \in B(\mathbf{y}_0;\delta/(2M))$

$$\|\mathbf{q}(\mathbf{y}) - \mathbf{q}(\mathbf{y}_0)\| \leq 2M\|\mathbf{y}-\mathbf{y}_0\|. \tag{2.4}$$

Beweis. Beweise für dieses Lemma finden sich zum Beispiel in [9] oder [45]. □

Die folgende Version des Satzes über implizite Funktionen nach [9] gibt Bedingungen an, unter denen eine Lösung des Problems (2.1) für $\mathbf{F}$ existiert.

Satz 2.1.2 (Satz über implizite Funktionen)**.** *Sei $\mathbf{F}$ eine um einen Punkt $(\mathbf{u}_0,\lambda_0) \in \mathbb{R}^n \times \mathbb{R}^k$ definierte, stetig differenzierbare Funktion mit Bild in $\mathbb{R}^n$. Seien weiterhin die folgenden Bedigungen erfüllt:*

(i) Die Matrix $\mathbf{D_u F}(\mathbf{u}_0, \lambda_0)$ *ist regulär mit*

$$\|\mathbf{D_u F}(\mathbf{u}_0, \lambda_0)^{-1}\| \leq c_0; \tag{2.5}$$

(ii) es gilt

$$\|\mathbf{D}_\lambda \mathbf{F}(\mathbf{u}_0, \lambda_0)\| \leq c_1. \tag{2.6}$$

Sei außerdem $\beta > 0$ *so gewählt, dass für alle* $(\mathbf{u}, \lambda) \in B(\mathbf{u}_0, \lambda_0; \beta)$ *und* $M := \sqrt{2}\max(c_0, 1 + c_0 c_1)$

$$\|\mathbf{DF}(\mathbf{u}, \lambda) - \mathbf{DF}(\mathbf{u}_0, \lambda_0)\| \leq \frac{1}{2M}$$

gilt. Gilt dann für den Funktionswert in $(\mathbf{u}_0, \lambda_0)$

$$\|\mathbf{F}(\mathbf{u}_0, \lambda_0)\| \leq \delta$$

mit $\delta := \beta/(2\sqrt{2}M)$*, dann existiert für* $\alpha := \beta/(2\sqrt{2}M)$ *eine eindeutige* C^1*-Funktion* $\mathbf{g} : B(\lambda_0; \alpha) \to B(\mathbf{u}_0; \beta)$ *mit*

$$\mathbf{F}(\mathbf{g}(\lambda), \lambda) = \mathbf{0}, \ \lambda \in B(\lambda_0; \alpha).$$

Des Weiteren gilt für alle $\lambda \in B(\lambda_0; \alpha)$ *die Ungleichung*

$$\|\mathbf{g}(\lambda) - \mathbf{g}(\lambda_0)\| \leq 2M(\|\lambda - \lambda_0\| + \|\mathbf{F}(\mathbf{u}_0, \lambda_0)\|). \tag{2.7}$$

Beweis. Man betrachte die Funktion $\mathbf{H}$, definiert auf einer Umgebung von $(\mathbf{u}_0, \lambda_0)$ mit

$$\mathbf{H}(\mathbf{u}, \lambda) := \begin{pmatrix} \mathbf{F}(\mathbf{u}, \lambda) \\ \lambda \end{pmatrix}.$$

Ziel ist es, den Satz über inverse Funktionen (Lemma 2.1.1) auf $\mathbf{H}$ anzuwenden. Dazu müssen die Bedingungen (2.2) und (2.3) erfüllt sein. Zunächst hält man fest, dass $\mathbf{DH}(\mathbf{u}_0, \lambda_0)$ invertierbar ist, denn aus der Regularität von $\mathbf{DF}(\mathbf{u}_0, \lambda_0)$ folgt

$$\begin{aligned} \mathbf{DH}(\mathbf{u}_0, \lambda_0)^{-1} &= \begin{pmatrix} \mathbf{D_u F}(\mathbf{u}_0, \lambda_0) & \mathbf{D}_\lambda \mathbf{F}(\mathbf{u}_0, \lambda_0) \\ \mathbf{0} & \mathbf{I}_k \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \mathbf{D_u F}(\mathbf{u}_0, \lambda_0)^{-1} & -\mathbf{D_u F}(\mathbf{u}_0, \lambda_0)^{-1}\mathbf{D}_\lambda \mathbf{F}(\mathbf{u}_0, \lambda_0) \\ \mathbf{0} & \mathbf{I}_k \end{pmatrix}. \end{aligned}$$

Somit folgt aus (2.5) und (2.6) und $\mathbf{z} = (\mathbf{u}^T, \lambda)^T \in \mathbb{R}^{n+k}$

$$\begin{aligned} \|\mathbf{DH}(\mathbf{u}_0, \lambda_0)^{-1}\mathbf{z}\| &= \left\| \begin{pmatrix} \mathbf{D_u F}(\mathbf{u}_0, \lambda_0)^{-1}(\mathbf{u} - \mathbf{D}_\lambda \mathbf{F}(\mathbf{u}_0, \lambda_0)\lambda) \\ \lambda \end{pmatrix} \right\| \\ &\leq \|\lambda\| + \|\mathbf{D_u F}(\mathbf{u}_0, \lambda_0)^{-1}(\mathbf{u} - \mathbf{D}_\lambda \mathbf{F}(\mathbf{u}_0, \lambda_0)\lambda)\| \\ &\leq \|\lambda\| + c_0(\|\mathbf{u}\| + c_1\|\lambda\|) \leq (1 + c_0 c_1)\|\lambda\| + c_0\|\mathbf{u}\| \end{aligned}$$

Mit $M := \sqrt{2}\max(c_0, 1 + c_0 c_1)$ und der Ungleichung $(a+b) \leq \sqrt{2(a^2+b^2)}$ für $a, b > 0$ erhält man dann

$$\|\mathbf{DH}(\mathbf{u}_0, \lambda_0)^{-1}\mathbf{z}\| \leq \frac{M}{\sqrt{2}}(\|\lambda\| + \|\mathbf{u}\|) \leq M\sqrt{\|\lambda\|^2 + \|\mathbf{u}\|^2} = M\|\mathbf{z}\|$$

und somit

$$\|\mathbf{DH}(\mathbf{u}_0, \lambda_0)^{-1}\| \leq M.$$

Weiterhin gilt

$$\mathbf{DH}(\mathbf{u}, \lambda) - \mathbf{DH}(\mathbf{u}_0, \lambda_0) = \begin{pmatrix} \mathbf{DF}(\mathbf{u}, \lambda) - \mathbf{DF}(\mathbf{u}_0, \lambda_0) \\ \mathbf{0} \end{pmatrix}.$$

und es lässt sich wegen der Stetigkeit von $\mathbf{DH}$ ein $\beta > 0$ finden, sodass für alle $(\mathbf{u}, \lambda) \in B(\mathbf{u}_0, \lambda_0; \beta)$ die Abschätzung

$$\|\mathbf{DH}(\mathbf{u}, \lambda) - \mathbf{DH}(\mathbf{u}_0, \lambda_0)\| \leq \frac{1}{2M} \tag{2.8}$$

gilt. Aus Lemma 2.1.1 folgt nun, dass eine C^1-Funktion
$\mathbf{q} : B(\mathbf{F}(\mathbf{u}_0, \lambda_0), \lambda_0; \beta/(2M)) \to B(\mathbf{u}_0, \lambda_0; \beta)$ mit $\mathbf{H}(\mathbf{q}(\mathbf{u}, \lambda)) = (\mathbf{u}, \lambda)$ existiert. Schreibt man $\mathbf{q}$ als $\mathbf{q}(\mathbf{u}, \lambda) := (\mathbf{q}_\lambda(\mathbf{u}, \lambda), \mathbf{q}_\mathbf{u}(\mathbf{u}, \lambda)^T)^T$, erhält man dann für alle $(\mathbf{u}, \lambda) \in B(\mathbf{F}(\mathbf{u}_0, \lambda_0), \lambda_0; \beta/(2M))$ die Identität

$$\begin{pmatrix} \mathbf{u} \\ \lambda \end{pmatrix} = \mathbf{H}(\mathbf{q}(\mathbf{u}, \lambda)) = \begin{pmatrix} \mathbf{F}(\mathbf{q}_\mathbf{u}(\mathbf{u}, \lambda), \mathbf{q}_\lambda(\mathbf{u}, \lambda)) \\ \mathbf{q}_\lambda(\mathbf{u}, \lambda) \end{pmatrix}.$$

Somit gilt $\mathbf{q}_\lambda(\mathbf{u}, \lambda) = \lambda$ und damit

$$\mathbf{u} = \mathbf{F}(\mathbf{q}_\mathbf{u}(\mathbf{u}, \lambda), \lambda), \ (\mathbf{u}, \lambda) \in B(\mathbf{F}(\mathbf{u}_0, \lambda_0), \lambda_0; \beta/(2M)) \tag{2.9}$$

Ziel ist es jetzt, $\mathbf{u} = \mathbf{0}$ zu setzen, um dann Lemma 2.1.1 anzuwenden. Dies ist allerdings nicht ohne Einschränkungen möglich. Wegen $\mathbf{F}(\mathbf{u}_0, \lambda_0) \neq \mathbf{0}$ muss der Punkt $(\mathbf{0}^T, \lambda)^T$ nicht in der Kugel $B(\mathbf{u}_0, \lambda_0; \beta/(2M))$ enthalten sein. Damit $(\mathbf{0}^T, \lambda_0)^T \in B(\mathbf{F}(\mathbf{u}_0, \lambda_0), \lambda_0; \beta/(2M))$ erfüllbar ist, muss die folgende Ungleichung gelten:

$$\|\lambda - \lambda_0\| \leq \sqrt{\frac{\beta^2}{4M^2} - \|\mathbf{F}(\mathbf{u}_0, \lambda_0)\|^2}.$$

$\|\mathbf{F}(\mathbf{u}_0, \lambda_0)\|$ muss nun also in jedem Fall echt kleiner als $\beta/(2M)$ sein. Je größer der Wert wird, umso kleiner wird die Kugel um λ_0 für die eine Lösung garantiert werden kann. Eine mögliche Wahl bietet die Forderung

$$\|\mathbf{F}(\mathbf{u}_0, \lambda_0)\| \leq \frac{\beta}{2\sqrt{2}M} =: \delta.$$

Es gilt dann $(\mathbf{0}^T, \lambda_0)^T \in B(\mathbf{F}(\mathbf{u}_0, \lambda_0), \lambda_0; \beta/(2M))$ für alle $\lambda \in B(\lambda_0; \beta/2\sqrt{2}M)$. Setzt man nun $\mathbf{u} = \mathbf{0}$ in (2.9) ein, erhält man mit

$$\alpha := \frac{\beta}{2\sqrt{2}M} \tag{2.10}$$

die gewünschte C^1-Funktion über

$$\mathbf{g} : \begin{cases} B(\lambda_0; \alpha) & \to B(\mathbf{u}_0; \beta) \\ \lambda & \mapsto \mathbf{q_u}(\mathbf{0}, \lambda) \end{cases},$$

für die gilt $\mathbf{F}(\mathbf{g}(\lambda), \lambda) = \mathbf{0}$.

Um nun noch die Ungleichung (2.7) zu zeigen, wird (2.4) auf die Funktion $\mathbf{q}$ angewendet und es ergibt sich so

$$\begin{aligned} \|\mathbf{g}(\lambda) - \mathbf{g}(\lambda_0)\| &\leq \|\mathbf{q}(\lambda, \mathbf{0}) - \mathbf{q}(\lambda_0, \mathbf{F}(\mathbf{u}_0, \lambda_0))\| \\ &\leq 2M(\|\lambda - \lambda_0\| + \|\mathbf{F}(\mathbf{u}_0, \lambda_0)\|) \end{aligned}$$

□

Bemerkung 2.1.3. *In Gleichung (2.10) kann wegen des linearen Zusammenhanges zwischen α und β (vergleiche (2.10) die Konstante β auch durch $\theta\beta$ mit $\theta \in [0, 1]$ ersetzt werden, woraus $g : B(\lambda_0, \theta\alpha) \to B(\mathbf{y}_0, \theta\beta)$ folgt. Es ist also möglich, innerhalb der Kugel $B(\lambda_0, \alpha)$ kleinere Gebiete $B(\lambda_0, \theta\alpha)$ zu finden, die garantieren, dass der Funktionswert $\mathbf{g}(t)$ für alle $t \in B(\lambda_0, \theta\alpha)$ in einer beliebig kleinen Kugel um $\mathbf{y}_0$ liegt. In diesem Fall verändert sich natürlich auch die maximal erlaubte Größe von $\|\mathbf{F}_0\|$, da im Grenzfall $\theta = 0$ der Punkt $\mathbf{x}_0$ Lösung von $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ sein muss.*

Das folgende Korollar stellt eine Version des Satzes über implizite Funtkionen für den Fall dar, dass $(\mathbf{u}_0, \lambda_0)$ eine Lösung von (2.1) ist und somit $\mathbf{F}(\mathbf{u}_0, \lambda_0) = \mathbf{0}$ gilt.

Korollar 2.1.4. Sei $\mathbf{F}$ wie im Satz 2.1.2 definiert und es seien im Punkt $(\mathbf{u}_0, \lambda_0)$ mit $\mathbf{F}(\mathbf{u}_0, \lambda_0) = \mathbf{0}$ die folgenden Bedingungen erfüllt:

(i) Die Matrix $\mathbf{D_u F}(\mathbf{u}_0, \lambda_0)$ ist regulär mit

$$\|\mathbf{D_u F}(\mathbf{u}_0, \lambda_0)^{-1}\| \leq c_0;$$

(ii) es gilt

$$\|\mathbf{D}_\lambda \mathbf{F}(\mathbf{u}_0, \lambda_0)\| \leq c_1.$$

Sei außerdem $\beta > 0$ so gewählt, dass für alle $(\mathbf{u}, \lambda) \in B(\mathbf{u}_0, \lambda_0; \beta)$ und $M := \sqrt{2}\max(c_0, 1 + c_0 c_1)$

$$\|\mathbf{DF}(\mathbf{u}, \lambda) - \mathbf{DF}(\mathbf{u}_0, \lambda_0)\| \leq \frac{1}{2M}$$

gilt. Dann existiert für $\alpha := \beta/(2M)$ eine eine eindeutige C^1-Funktion $\mathbf{g} : B(\lambda_0; \alpha) \to B(\mathbf{u}_0; \beta)$ mit

$$\mathbf{F}(\mathbf{g}(\lambda), \lambda) = \mathbf{0},\ \lambda \in B(\lambda_0; \alpha).$$

Für diese Funktion gilt dann zudem

$$\|\mathbf{g}(\lambda) - \mathbf{g}(\lambda_0)\| \leq 2M\|\lambda - \lambda_0\|$$

Beweis. Der Beweis verläuft analog zu dem des Satzes 2.1.2 mit $\delta = 0$. □

2.2 Einparametrige nichtlineare Gleichungen

In diesem Kapitel wird der Fall $\mathbf{F} \in C^1(\Omega, \mathbb{R}^n)$ mit einer offenen Menge $\Omega \subset \mathbb{R}^n \times \mathbb{R}$ genauer betrachtet. In diesem Fall existiert also nur ein einziger Parameter $\lambda \in \mathbb{R}$. Die auftretenden Lösungsmannigfaltigkeiten sind eindimensional und werden als Lösungskurven bezeichnet. Für die Berechnung von Approximationen dieser Lösungskurven existieren eine Vielzahl von numerischen Verfahren, vergleiche dazu [2, 61].

2.2.1 Lösungsexistenz und Umkehrpunkte

Zunächst werden Probleme, die durch eine feste Parametrisierung nach λ entstehen, betrachtet. Dies sei an einem kurzen Beispiel erläutert.

Beispiel 2.2.1. Sei $\mathbf{F} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ gegeben mit

$$F(u, \lambda) = u^2 + \lambda - 1.$$

Es ist leicht zu sehen, dass die Lösungsmenge von $F(u, \lambda) = 0$ als einparametrige Lösungskurve

$$\mathbf{c} : \begin{cases} \mathbb{R} & \to \mathbb{R}^2 \\ s & \mapsto (s, -s^2 + 1)^T \end{cases}$$

angegeben werden kann. Im Punkt $(0, 1)$ lässt sich der Satz über implizite Funktionen nicht anwenden, da die notwendige Bedingung der Regularität von $\mathbf{D}_u\mathbf{F}(u, \lambda) = 2u$ nicht erfüllt ist. Somit lässt sich die Lösung also nicht bezüglich λ fortsetzen. Obwohl also eine Lösungskurve existiert, lässt sich ihre Existenz auf die bisher betrachtete Weise nicht nachweisen.

Punkte, in denen die Kurve $\mathbf{c}$ bzgl. einer Variable ein Verhalten wie in Beispiel 2.2.1 gegenüber λ aufweist, werden als Umkehrpunkte (genauer λ-Umkehrpunkte) bezeichnet, [67, 13, 17, 60] und sind wie folgt definiert:

Definition 2.2.2. Eine Lösung $(\mathbf{u}_0, \lambda_0)$ von $\mathbf{F}(\mathbf{u}, \lambda) = 0$ mit

(i) $\text{Rang}([\mathbf{D_u F}(\mathbf{u}_0, \lambda_0), \mathbf{D}_\lambda \mathbf{F}(\mathbf{u}_0, \lambda_0)]) = n$

(ii) $\mathbf{D_u F}(\mathbf{u}_0, \lambda_0)$ ist singulär

heißt Umkehrpunkt.

Man beachte, dass sich die grundlegenden Eigenschaften der Lösungskurve $\mathbf{c}$ aus Beispiel 2.2.1 im Umkehrpunkt $(0, 1)$ nicht ändern: sie bleibt eine differenzierbare einparametrige Mannigfaltigkeit. Ein Umkehrpunkt und die damit auftretenden Probleme entstehen allein durch den Versuch einer Parametrisierung bezüglich des Parameters λ.

Umkehrpunkte, wie sie in Beispiel 2.2.1 auftreten, lassen darauf schließen, dass die ursprüngliche Gleichung für einen festen Parameter λ mehrere Lösungen $\mathbf{u}$ besitzt, von denen zwei in einem Umkehrpunkt zusammen laufen. Es lässt sich also auch abseits des Umkehrpunktes für gewisse λ keine globale Zuordnung $\lambda \mapsto \mathbf{u}(\lambda)$ treffen.

Da man üblicherweise an der Beschreibung des Zusammenhangs zwischen λ und $\mathbf{u}$ interessiert ist, entstehen durch Umkehrpunkte Probleme bei der numerischen Berechnung der Lösungskurve von $\mathbf{F}(\mathbf{u}, \lambda) = \mathbf{0}$. Eine einfache und naheliegende Möglichkeit eine solche Berechnung durchzuführen ist ausgehend von einer Lösung $(\mathbf{u}_0, \lambda_0)$ den Parameter λ stückweise zu erhöhren und dann das nichtlineare Gleichungssystem bezüglich $\mathbf{u}$ zu lösen. In Abbildung 2.1 ist diese Methode für Beispiel 2.2.1 skizziert. Man erkennt leicht, dass ein Umkehrpunkt die vollständige Analyse der Lösungskurve unmöglich macht, da das Verhalten der Kurve im Umkehrpunkt nicht erfasst wird.

Um solche Probleme in den Griff zu bekommen, kann ausgenutzt werden, dass die Matrix $[\mathbf{D_u F}(\mathbf{u}_0, \lambda_0), \mathbf{D}_\lambda \mathbf{F}(\mathbf{u}_0, \lambda_0)]$ bei vollem Zeilenrang stets eine invertierbare $n \times n$-Untermatrix enthält. Daher existiert immer eine Raumdimension bezüglich der die Lösungskurve parametrisiert werden kann, [49]. Eine Möglichkeit das Problem der Umkehrpunkte zu umgehen die ohne Umparametrisierung auskommt besteht darin, das Problem mittels $\mathbf{x} = (\mathbf{u}^T, \lambda)^T$ zu

$$\mathbf{F}(\mathbf{x}) := \mathbf{F}(\mathbf{u}, \lambda) = \mathbf{0} \tag{2.11}$$

umzuformulieren. Für $\mathbf{F}$ definiert man die Regularitätsmenge

$$\mathcal{R}(\mathbf{F}) := \{\mathbf{x} \in \Omega : \text{Rang}(\mathbf{DF}(\mathbf{x})) = n\}.$$

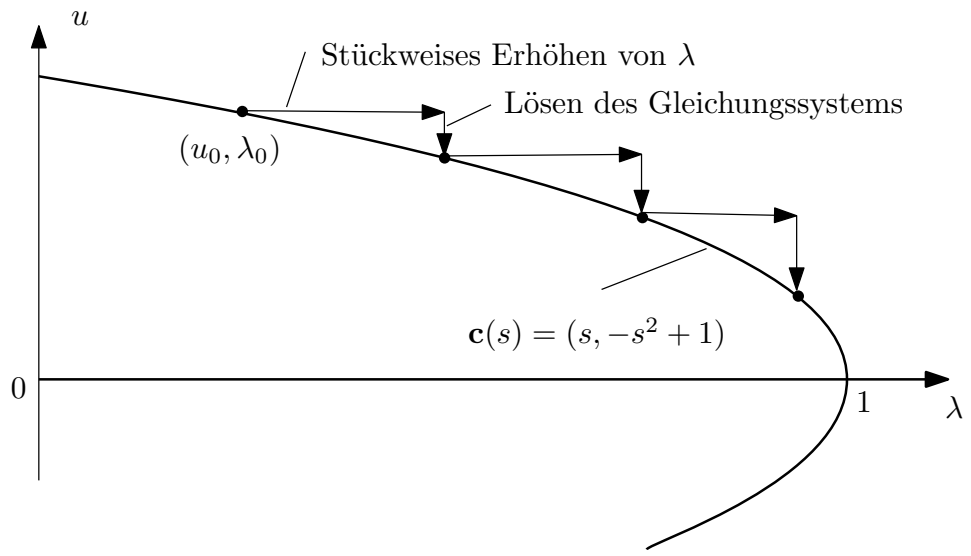

Abbildung 2.1: Schrittweises Berechnen der Lösung von $F(u, \lambda) = 0$ aus Beispiel 2.2.1

Die Eigenschaft $\text{Rang}(\mathbf{DF}(\mathbf{x})) = n$ stellt eine Verallgemeinerung von (2.5) dar, da nicht mehr die Regularität der Ableitung bezüglich $\mathbf{u}$ gefordert wird. Stattdessen ist ausreichend, dass die Jacobimatrix $\mathbf{DF}(\mathbf{x})$ eine reguläre $n \times n$ Untermatrix enthält.

Die Menge $\mathcal{R}(\mathbf{F})$ ist offen. Ein Beweis dafür findet sich zum Beispiel in [2] und basiert auf der Stetigkeit von $\mathbf{DF}(\mathbf{x})$ und dem Umstand, dass $\mathbf{x}$ genau dann in $\mathcal{R}(\mathbf{F})$ liegt, wenn $\det(\mathbf{DF}(\mathbf{x})\mathbf{DF}(\mathbf{x})^T) \neq 0$ gilt.

Aussagen über die Existenz einer Lösungskurve in einem Punkt $\mathbf{x}_0 \in \mathcal{R}(\mathbf{F})$ mit $\mathbf{F}(\mathbf{x}_0) = \mathbf{0}$ finden sich zum Beispiel in [49, 57] oder [60]. Für die in Kapitel 6 entwickelte Reduktion werden jedoch genauere Aussagen als die generelle Existenz einer Lösungskurve benötigt. Durch die Verallgemeinerung der Parameter muss die Kurve nicht mehr notwendigerweise bezüglich einer festen Raumrichtung parametrisiert werden. Dies führt außerdem dazu, dass die Norm der Inversen wie in Satz 2.1.2 nicht mehr als ausschlaggebende Größe verwendet werden kann und stattdessen die Singulärwerte von $\mathbf{DF}$ herangezogen werden.

Der folgende Satz gibt nun an, unter welchen Bedingungen und wie weit sich eine bekannte Lösung bezüglich einer gegebenen Raumrichtung fortsetzen lässt und dient als wichtiges Werkzeug für die in Kapitel 6 geführten Existenzbeweise.

Satz 2.2.3. *Sei für $\mathbf{F} \in C^1(\Omega, \mathbb{R}^n)$ mit offenem $\Omega \subset \mathbb{R}^{n+1}$ und $\mathbf{x}_0 \in \mathcal{R}(\mathbf{F})$ zwei Konstanten c_0 und c_1 gegeben, sodass für den kleinsten und größten Sin-*

gulärwert σ_n bzw. σ_1 in $\mathbf{DF}(\mathbf{x}_0)$ die Abschätzungen

$$\sigma_n^{-1} \leq c_0, \text{ sowie } \sigma_1 \leq c_1$$

gelten. Seien zudem zwei normierte Vektoren $\mathbf{T}(\mathbf{x}_0) \in Kern(\mathbf{DF}(\mathbf{x}_0))$ und $\mathbf{r}$, sowie eine Konstante $c_2 > 0$ mit

$$\langle \mathbf{T}(\mathbf{x}_0), \mathbf{r} \rangle \geq c_2$$

gegeben. Weiterhin sei $\mathbf{Y}$ eine Matrix, deren Spalten eine Orthornormalbasis von $R(\mathbf{r})^\perp$ enthalten.

Sei $\beta > 0$ so gewählt, sodass mit $M := \sqrt{2}\max(c_0 c_2^{-1}, 1 + c_0 c_2^{-1} c_1)$ für all $\mathbf{x} \in B(\mathbf{x}_0; \beta)$

$$\|\mathbf{DF}(\mathbf{x}) - \mathbf{DF}(\mathbf{x}_0)\| \leq \frac{1}{2M}$$

gilt und die Konstanten α und β über

$$\alpha := \frac{\beta}{2\sqrt{2}M}, \text{ und } \delta := \frac{\beta}{2\sqrt{2}M}$$

definiert. Gilt dann

$$\|\mathbf{F}(\mathbf{x}_0)\| \leq \delta,$$

dann existiert eine eindeutige C^1-Funktion $\mathbf{g} : B(0;\alpha) \to B(\mathbf{0};\beta)$, sodass für die C^1-Funktion

$$\mathbf{c} : \begin{cases} B(0;\alpha) & \to \mathbb{R}^{n+1} \\ s & \mapsto \mathbf{x}_0 + s\mathbf{r} + \mathbf{Y}\mathbf{g}(s) \end{cases}$$

und für alle $s \in B(0;\alpha)$

$$\mathbf{F}(\mathbf{c}(s)) = \mathbf{0}$$

gilt.

Beweis. Ziel ist es, den Satz über implizite Funktionen (Satz 2.1.2) auf die Funktion

$$\mathbf{G} : \begin{cases} Y \times S & \to \mathbb{R}^n, \\ (\mathbf{y}, s) & \mapsto \mathbf{F}(\mathbf{x}_0 + s\mathbf{r} + \mathbf{Y}\mathbf{y}) \end{cases}$$

anzuwenden. Dabei sollen Y und S offen sein und stets $\mathbf{x}_0 + s\mathbf{r} + \mathbf{Y}\mathbf{y} \in \Omega$ gelten. Es muss nun zunächst der Wert $\|\mathbf{D_y G}(\mathbf{0}, 0)^{-1}\|$ abgeschätzt werden. Dazu sei zunächst festgehalten, dass für $\mathbf{DF}$

$$\sigma_n = \min_{\|\mathbf{u}\|=1, \mathbf{u} \perp \mathrm{Kern}(\mathbf{DF}(\mathbf{x}_0))} \|\mathbf{DF}(\mathbf{x}_0)\mathbf{u}\| = \min_{\|\mathbf{u}\|=1, \mathbf{u} \perp \mathbf{T}(\mathbf{x}_0)} \|\mathbf{DF}(\mathbf{x}_0)\mathbf{u}\|$$

gilt.

Sei nun $\sigma_n^{\mathbf{G}}$ der kleinste Singulärwert von $\mathbf{DyG}(\mathbf{0},0)$, sowie $\mathbf{P}(\mathbf{x}_0)$ und $\mathbf{P}^{\perp}(\mathbf{x}_0)$ die orthognalen Projektoren auf $\mathbf{R}(\mathbf{T}(\mathbf{x}_0))$ bzw. dessen Orthogonalraum, dann gilt

$$\begin{aligned}\sigma_n^{\mathbf{G}} &= \min_{\|\mathbf{y}\|=1} \|\mathbf{D}_{\mathbf{y}}\mathbf{G}(\mathbf{0},0)\mathbf{y}\| = \min_{\|\mathbf{y}\|=1} \|\mathbf{DF}(\mathbf{x}_0)\mathbf{Y}\mathbf{y}\| \\ &= \min_{\|\mathbf{w}\|=1,\mathbf{w}\perp\mathbf{r}} \|\mathbf{DF}(\mathbf{x}_0)\mathbf{w}\| = \min_{\|\mathbf{w}\|=1,\mathbf{w}\perp\mathbf{r}} \|\mathbf{DF}(\mathbf{x}_0)(\mathbf{P}(\mathbf{x}_0)\mathbf{w} + \mathbf{P}^{\perp}(\mathbf{x}_0)\mathbf{w}\|\end{aligned}$$

Da $\mathrm{Kern}(\mathbf{DF}(\mathbf{x}_0)) = R(\mathbf{T}(\mathbf{x}_0))$ gilt, vereinfacht sich dies zu

$$\begin{aligned}\sigma_n^{\mathbf{G}} &= \min_{\|\mathbf{w}\|=1,\mathbf{w}\perp\mathbf{r}} \|\mathbf{DF}(\mathbf{x}_0)\mathbf{P}^{\perp}(\mathbf{x}_0)\mathbf{w}\| \\ &= \min_{\|\mathbf{w}\|=1,\mathbf{w}\perp\mathbf{r}} \|\mathbf{P}^{\perp}(\mathbf{x}_0)\mathbf{w}\| \left\|\mathbf{DF}(\mathbf{x}_0)\frac{\mathbf{P}^{\perp}(\mathbf{x}_0)\mathbf{w}}{\|\mathbf{P}^{\perp}(\mathbf{x}_0)\mathbf{w}\|}\right\| \\ &\geq \min_{\|\mathbf{w}\|=1,\mathbf{w}\perp\mathbf{r}} \|\mathbf{P}^{\perp}(\mathbf{x}_0)\mathbf{w}\| \min_{\|\mathbf{u}\|=1,\mathbf{u}\perp\mathbf{T}(\mathbf{x}_0)} \|\mathbf{DF}(\mathbf{x}_0)\mathbf{u}\| \\ &= |\langle\mathbf{T}(\mathbf{x}_0),\mathbf{r}\rangle|\sigma_n \geq \frac{c_2}{c_0}.\end{aligned}$$

Die letzte Gleichung ergibt sich daraus, dass für zwei $k-1$-dimensionale Unterräume U und V des $\mathbb{R}^k$ mit dazu jeweiligen senkrechten normierten Vektoren $\mathbf{n}_U$ und $\mathbf{n}_V$ für den orthogonalen Projektor $\mathbf{P}_V$ auf V die Gleichung

$$\min_{\|\mathbf{u}\|=1,\mathbf{u}\in U} \|\mathbf{P}_V\mathbf{u}\| = \langle\mathbf{n}_U,\mathbf{n}_V\rangle$$

gilt.

Für die Norm der Inversen ergibt sich so

$$\|\mathbf{D}_{\mathbf{y}}\mathbf{G}(\mathbf{0},0)^{-1}\| = (\sigma_n^{\mathbf{G}})^{-1} \leq c_0 c_2^{-1} =: c_0^{\mathbf{G}}$$

Für die Ableitung $\mathbf{D}_s\mathbf{G}(\mathbf{0},0)$ erhält man direkt

$$\|\mathbf{D}_s\mathbf{G}(\mathbf{0},0)\| = \|\mathbf{DF}(\mathbf{x}_0)\mathbf{r}\| \leq \sigma_1\|\mathbf{r}\| \leq c_1.$$

Sei nun $M := \sqrt{2}\max(c_0^{\mathbf{G}}, 1 + c_0^{\mathbf{G}}c_1)$, dann gilt wegen (2.12) für alle $(\mathbf{y},s) \in B((\mathbf{0},0);\beta)$

$$\begin{aligned}\|\mathbf{DG}(\mathbf{y},s) - \mathbf{DG}(\mathbf{0},0)\| &= \|(\mathbf{DF}(\mathbf{x}) - \mathbf{DF}(\mathbf{x}_0))(\mathbf{Y},\mathbf{r})\| \\ &= \|\mathbf{DF}(\mathbf{x}) - \mathbf{DF}(\mathbf{x}_0)\| \leq \frac{1}{2M}.\end{aligned}$$

Weiterhin ist mit $\delta := \beta/(2\sqrt{2}M)$ auch die Bedingung

$$\|\mathbf{G}(\mathbf{0},0)\| = \|\mathbf{F}(\mathbf{x}_0)\| \leq \delta$$

erfüllt. Nach Satz 2.1.2 existiert jetzt für $\alpha = \beta/(2\sqrt{2}M)$ eine eindeutige C^1-Funktion $\mathbf{g} : B(0;\alpha) \to B(\mathbf{0};\beta)$, sodass für alle $s \in B(0;\alpha)$

$$\mathbf{G}(\mathbf{g}(s), s) = \mathbf{0}$$

gilt. Setzt man nun $\mathbf{c}(s) := \mathbf{x}_0 + s\mathbf{r} + \mathbf{Y}\mathbf{g}(s)$ ergibt sich die gesuchte Funktion. □

Bemerkung 2.2.4. *Analog zu Korollar 2.1.4 wird die Konstante α des vorherigen Satzes für den Fall, dass $\mathbf{F}(\mathbf{x}_0) = \mathbf{0}$ gilt zu*

$$\alpha := \frac{\beta}{2M}.$$

Bemerkung 2.2.5. *Die Lösungskurve aus dem vorherigen Satz existiert für den Fall, dass $\mathbf{c}(\alpha) \in \mathcal{R}(\mathbf{F})$ gilt, natürlich über das durch das Lemma gesicherte Intervall $B(0;\alpha)$ hinaus, nur muss die Parametrisierung für den weiteren Verlauf der Lösungskurve angepasst werden, um die Lösung fortzusetzen.*

Aus diesem Grund benötigt man zur globalen Beschreibung der Kurve eine gemeinsame Parametrisierung. Hierfür hat es ich als günstig erwiesen, die Bogenlänge zu verwenden. Auf diese Weise lässt sich die Lösungskurve für ein offenes, die Null enthaltendes Intervall $S \subset \mathbb{R}$ als stetig differenzierbare Abbildung

$$\mathbf{c} : S \to \mathbb{R}^{n+1} \text{ mit } \|\mathbf{c}'(s)\| = 1, s \in S$$

darstellen.

2.2.2 Das Tangentialfeld

Im Folgenden wird das im späteren Verlauf benötigte Tangentialvektorfeld basierend auf [2] definiert.

Definition 2.2.6. Sei $\mathbf{DF}(\mathbf{x}) \in \mathbb{R}^{n,n+1}$ mit $\text{Rang}(\mathbf{DF}(\mathbf{x})) = n$, dann wird der Vektor $\mathbf{T}(\mathbf{x}) \in \mathbb{R}^{n+1}$ mit den Eigenschaften

(i) $\mathbf{DF}(\mathbf{x})\mathbf{T}(\mathbf{x}) = \mathbf{0}$,

(ii) $\|\mathbf{T}(\mathbf{x})\| = 1$,

(iii) $\det \begin{pmatrix} \mathbf{DF}(\mathbf{x}) \\ \mathbf{T}(\mathbf{x})^T \end{pmatrix} > 0$

Tangentialvektor im Punkt $\mathbf{x}$ genannt.

Der Tangentialvektor $\mathbf{T}$ hängt stetig differenzierbar von der Matrix $\mathbf{DF}$ ab. Dies lässt sich zeigen, in dem man man das Korollar 2.1.4 auf die Abbildung

$$\mathbf{M} : \begin{cases} \mathbb{R}^{n+1} \times \mathbb{R}^{n,n+1} & \to \mathbb{R}^{n+1} \\ (\mathbf{v}, \mathbf{A}) & \mapsto \begin{pmatrix} \mathbf{Av} \\ \frac{1}{2}(\mathbf{v}^T\mathbf{v} - 1) \end{pmatrix} \end{cases}$$

anwendet. Für einen beliebigen Punkt $\mathbf{x}_0 \in \mathcal{R}(\mathbf{F})$ gilt $\mathbf{M}(\mathbf{T}(\mathbf{x}_0), \mathbf{DF}(\mathbf{x}_0)) = \mathbf{0}$ und für die Jacobimatrix dieser Abbildung erhält man

$$\mathbf{DM}(\mathbf{v}, \mathbf{A}) = \begin{pmatrix} \mathbf{A} \\ \mathbf{v}^T \end{pmatrix}.$$

Diese Matrix ist in $(\mathbf{DF}(\mathbf{x}_0), \mathbf{T}(\mathbf{x}_0))$ wegen $\text{Kern}(\mathbf{DF}(\mathbf{x}_0)) = \mathbf{T}(\mathbf{x}_0)$ regulär und nach Korollar 2.1.4 gibt es eine Konstante $\alpha > 0$ und eine stetig differenzierbare Funktion

$$\mathbf{g} : B(\mathbf{DF}(\mathbf{x}_0); \alpha) \to \mathbb{R}^{n+1}$$

mit $\mathbf{M}(\mathbf{g}(\mathbf{A}), \mathbf{A}) = \mathbf{0},\ \mathbf{A} \in B(\mathbf{DF}(\mathbf{x}_0); \alpha)$, sowie $\mathbf{g}(\mathbf{DF}(\mathbf{x}_0)) = \mathbf{T}(\mathbf{x}_0)$.

Man betrachtet nun die Menge $D = \{\mathbf{x} \in \mathcal{R}(\mathbf{F}) : \mathbf{DF}(\mathbf{x}) \in B(\mathbf{DF}(\mathbf{x}_0); \alpha)\}$ genauer. Für alle $\mathbf{x} \in D$ gilt $\mathbf{DF}(\mathbf{x})\mathbf{g}(\mathbf{DF}(\mathbf{x})) = \mathbf{0}$ und $\|\mathbf{g}(\mathbf{DF}(\mathbf{x}))\| = 1$. Außerdem ist die Abbildung

$$\mathbf{DF}(\mathbf{x}) \mapsto \det \begin{pmatrix} \mathbf{DF}(\mathbf{x}) \\ \mathbf{g}(\mathbf{DF}(\mathbf{x}))^T \end{pmatrix}$$

stetig und besitzt keine Nullstellen; ihr Vorzeichen ist also konstant und wegen $\mathbf{g}(\mathbf{DF}(\mathbf{x}_0)) = \mathbf{T}(\mathbf{x}_0)$ in $\mathbf{DF}(\mathbf{x}_0)$ positiv. Somit erfüllen die Funktionswerte $\mathbf{g}(\mathbf{DF}(\mathbf{x}))$ für alle $\mathbf{x} \in D$ die Bedingungen der Definition 2.2.6 und stellen somit die Tangentialvektoren in $\mathbf{x}$ dar.

Nach Voraussetzung ist $\mathbf{DF} : \Omega \to \mathbb{R}^{n,n+1}$ stetig und über

$$\mathbf{T} : \begin{cases} \mathcal{R}(\mathbf{F}) & \to \mathbb{R}^{n+1} \\ \mathbf{x} & \mapsto \mathbf{g}(\mathbf{DF}(\mathbf{x})) \end{cases}$$

lässt sich daher eine stetige Funktion definieren, deren Funktionswerte auf $\mathcal{R}(\mathbf{F})$ ein stetiges Tangentialfeld bilden.

Bemerkung 2.2.7. *Die Regularität von* $\mathbf{F}$ *überträgt sich zum Teil auf das Tangentialfeld. So ist die Abbildung* $\mathbf{T}(\mathbf{x})$ *unter der Bedingung, dass* $\mathbf{F} \in C^2(\Omega, \mathbb{R}^n)$ *gilt (die Jacobimatrix* $\mathbf{DF}$ *also selbst wieder stetig differenzierbar ist), ebenfalls stetig differenzierbar. Gleiches gilt für die Lipschitzstetigkeit von* $\mathbf{DF}(\mathbf{x})$.

Bemerkung 2.2.8. *Die Bezeichnung Tangentialvektor leitet sich daraus ab, dass für die Ableitung von* $\mathbf{F}$ *entlang der Lösungskurve* c *für alle* $s \in S$

$$\frac{d}{ds}\mathbf{F}(\mathbf{c}(s)) = \mathbf{DF}(\mathbf{c}(s))\mathbf{c}'(s) = \mathbf{0} \tag{2.12}$$

gilt. Der eindimensionale Kern von $\mathbf{DF}(\mathbf{x})$ *wird also entlang der Lösungskurve von ihrer ersten Ableitung* $\mathbf{c}'$ *aufgespannt. Somit ist der Tangentialvektor* $\mathbf{T}(\mathbf{c}(s))$ *ein normiertes Vielfaches der Ableitung* $\mathbf{c}'(s)$.

Basierend auf Bemerkung 2.2.8 betrachtet man nun ein alternatives Problem

$$\mathbf{F}(\mathbf{x}) = \mathbf{b} \tag{2.13}$$

mit $\mathbf{b} \neq \mathbf{0}$. Liegt ein Punkt $\mathbf{x}_0$ mit $\mathbf{F}(\mathbf{x}_0) = \mathbf{b}$ in der Regularitätsmenge $\mathcal{R}(\mathbf{F})$, so lässt sich Satz 2.2.3 auf $\mathbf{F}^*(\mathbf{x}) = \mathbf{F}(\mathbf{x}_0) - \mathbf{b}$ anwenden, da $\mathbf{DF}^*(\mathbf{x}_0) = \mathbf{DF}(\mathbf{x}_0)$ gilt. Somit existiert eine Lösungskurve $\mathbf{c}^* : B(0;\alpha^*) \to \mathbb{R}^{n+1}$ mit $\mathbf{F}^*(\mathbf{c}^*(s)) = \mathbf{0}, s \in B(0;\alpha^*)$, und damit

$$\mathbf{F}(\mathbf{c}^*(s)) = \mathbf{b}.$$

So entsteht eine Lösungskurve, die einem ähnlichen Verlauf wie die des Problems $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ folgt, diese aber in der Regularitätsmenge nicht schneidet (da in einem solchen Bifurkationspunkt genannten Schnittpunkt die Bedingung $\text{Rang}(\mathbf{DF}(\mathbf{x})) = n$ nicht mehr erfüllt ist). Für das Problem (2.13) existiert also stets eine Lösungskurve, falls $\mathbf{b}$ in der Menge

$$\{\mathbf{b} \in \mathbb{R}^n : \exists \mathbf{x}_0 \in \mathcal{R}(\mathbf{F}), \text{ mit } \mathbf{F}(\mathbf{x}_0) = \mathbf{b}\}$$

liegt. Diese Lösungskurve erfüllt dann das Anfangswertproblem

$$\begin{cases} \mathbf{c}'(s) & = \mathbf{T}(\mathbf{c}(s)), \\ \mathbf{c}(0) & = \mathbf{x}_0, \text{ mit } \mathbf{F}(\mathbf{x}_0) = \mathbf{b}. \end{cases} \tag{2.14}$$

Auf diesem Sachverhalt basiert auch eine der Predictor-Corrector-Methoden in [2], die zur numerischen Berechnung der Lösungskurve $\mathbf{c}$ verwendet werden. In Abbildung 2.2 ist der Zusammenhang zwischen dem Tangentialfeld und der Lösungskurve von Problemen der Form (2.13) skizziert.

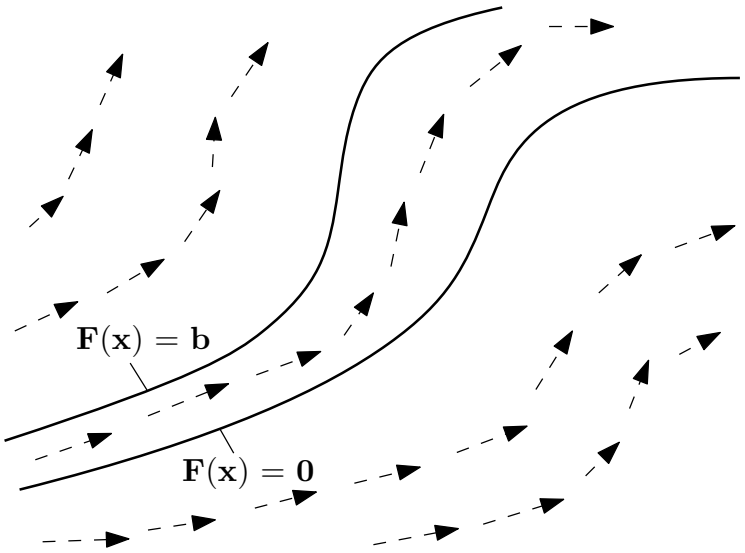

Abbildung 2.2: Lösung von $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ und $\mathbf{F}(\mathbf{x}) = \mathbf{b}$ und das Tangentialfeld von $\mathbf{F}$

Kapitel 3

RB-Methoden

3.1 Lokale RB-Methoden

Die lokalen RB-Methoden stellen historisch die ersten Methoden zur effizienten Basisreduktion großer nichtlinearer Systeme dar, eine Übersicht dazu findet sich zum Beispiel in [42]. Das Interesse liegt wie bei allen RB-Methoden auf dem effizienten Approximieren der Lösung eines parameterabhängigen nichtlinearen Gleichungssystems

$$\mathbf{F}(\mathbf{u}, \lambda) = \mathbf{0}, \tag{3.1}$$

wobei $\mathbf{F} \in C^r(\Omega, \mathbb{R}^n)$ mit $r \geq 1$ und $\Omega \subset \mathbb{R}^n \times \mathbb{R}^k$ gelten soll. Die lokalen RB-Methoden werden meist auf den Fall $k = 1$ angewendet, da sie sich dort mit einer Vielzahl von Astverfolgungsmethoden (Continuation Methods) kombinieren lassen, [60]. Eine Übersicht über solche Astverfolgungsmethoden findet sich zum Beispiel in [2] und [61]. Es existieren aber auch Untersuchungen für eine direkte Reduktion des mehrparametrigen Falles $k \geq 2$, siehe dazu [59] und [38].

Diese Arbeit beschäftigt sich mit der Reduktion einer zweiparametrigen nichtlinearen Funktion, wobei diese durch eine Erweiterung einer stückweise lokalen Reduktion bezüglich eines Parameters erfolgt. Basierend auf [38, 20, 49] wird im Folgenden daher zunächst die lokale RB-Methode für Einparameter-Systeme erläutert.

3.1.1 Lokale Galerkin-Diskrektisierung

Die zu Grunde liegende Problemstellung stellt die Approximation der Lösung von (3.1) für den Fall $\lambda \in \mathbb{R}$ dar. Wie in Kapitel 2.2.1 näher erläutert wurde, ist es sinnvoll, die Variablen $\mathbf{u}$ und λ zu $\mathbf{x} = (\mathbf{u}^T, \lambda)^T \in \mathbb{R}^{n+1}$ zusammenzufassen und das Problem als

$$\mathbf{F}(\mathbf{x}) = \mathbf{0} \tag{3.2}$$

umzuformulieren. Bei der lokalen RB-Methode wird davon ausgegangen, dass ein offenes Intervall $S \subset \mathbb{R}$ und eine stetig differenzierbare Lösungskurve $\mathbf{c} : S \to \mathbb{R}^{n+1}$ mit $\mathbf{F}(\mathbf{c}(s)) = 0, s \in S,\ \mathbf{c}'(s) \neq 0$ und $\mathbf{c}(0) = \mathbf{x}_0$ existiert.

Zur Approximation dieser Lösungskurve wird ein Galerkin-artiges Verfahren verwendet. Dazu wird ein Ansatzraum $\mathcal{Z}$ mit $\dim(\mathcal{Z}) = m+1 \ll n$ aufgebaut und eine Approximation $\mathbf{c}_R$ von $\mathbf{c}$ im affinen Unterraum $\mathbf{x}_0 + \mathcal{Z}$ gesucht. Dabei wird ausgenutzt, dass sich die Bewegung der Kurve lokal durch wenige Raumrichtungen näherungsweise beschreiben lässt (Taylor-Approximation). Diese Richtungen sollen durch den Raum $\mathcal{Z}$ möglichst gut wiedergegeben werden.

Da wegen der geringen Dimension von $\mathcal{Z}$ ein stark überbestimmtes System entsteht, benötigt man weiterhin einen m-dimensionalen Testraum $\mathcal{V}$ für den gefordert wird, dass die bei der Approximation auftretenden Residuen senkrecht zu ihm steht.

Seien mit $\mathbf{z}_1, \ldots, \mathbf{z}_{m+1} \in \mathbb{R}^{n+1}$ und $\mathbf{v}_1, \ldots, \mathbf{v}_m \in \mathbb{R}^n$ eine Basis von $\mathcal{Z}$ und $\mathcal{V}$ bezeichnet, dann werden nun Punkte

$$\mathbf{x}_R := \mathbf{x}_0 + \sum_{i=1}^{m+1} \mathbf{z}_i \hat{x}_i,\ \hat{x}_i \in \mathbb{R}$$

in $\mathbf{x}_0 + \mathcal{Z}$ gesucht, für die

$$\mathbf{v}_j^T \mathbf{F}(\mathbf{x}_R) = 0,\ j = 1, \ldots, m$$

gilt.

Fasst man die beiden Basen zu Matrizen $\mathbf{Z} := (\mathbf{z}_1, \ldots, \mathbf{z}_{m+1}) \in \mathbb{R}^{n+1,m+1}$ und $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_m) \in \mathbb{R}^{n,m}$ zusammen, so gelangt man zu dem nichtlinearen Gleichungssystem

$$\hat{\mathbf{F}}(\hat{\mathbf{x}}) := \mathbf{V}^T \mathbf{F}(\mathbf{x}_0 + \mathbf{Z}\hat{\mathbf{x}}) = \mathbf{0}. \tag{3.3}$$

Dieses besteht aus m nichtlinearen Gleichungen für $m+1$ Unbekannte $\hat{\mathbf{x}}$.

Die Funktion

$$\hat{\mathbf{F}}(\hat{\mathbf{x}}) : \begin{cases} \mathbb{R}^{m+1} & \to \mathbb{R}^m \\ \hat{\mathbf{x}} & \mapsto \mathbf{V}^T \mathbf{F}(\mathbf{x}_0 + \mathbf{Z}\hat{\mathbf{x}}) \end{cases} \tag{3.4}$$

wird lokale Reduktion von $\mathbf{F}$ genannt. Analog wird das Gleichungssystem (3.3) als das lokal reduzierte Problem bezeichnet.

Besitzt das lokale reduzierte Problem eine Lösung $\hat{\mathbf{c}} : \hat{S} \to \mathbb{R}^{m+1}$ mit $\hat{\mathbf{F}}(\hat{c}(s)) = \mathbf{0}, s \in \hat{S}$, so erhält man über

$$\mathbf{c}_R(s) := \mathbf{x}_0 + \mathbf{Z}\hat{\mathbf{c}}(s),\ s \in \hat{S}$$

die gesuchte Approximation der ursprünglichen Lösungskurve $\mathbf{c}$. Die Funktion $\hat{\mathbf{c}}$ wird als reduzierte Lösung bezeichnet.

Eine Bedingungen für die Lösbarkeit des lokal reduzierten Problems ergibt sich aus Lemma 2.1.1. Wegen $\hat{\mathbf{F}}(\mathbf{0}) = \mathbf{V}^T\mathbf{F}(\mathbf{x}_0) = \mathbf{0}$ ist mit $\hat{\mathbf{x}}_0 = \mathbf{0}$ eine Lösung bekannt. Für diesen Punkt gilt

$$\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0) = \mathbf{V}^T\mathbf{D}\mathbf{F}(\mathbf{x}_0)\mathbf{Z}.$$

Besitzt diese Matrix vollen Zeilenrang, liegt $\hat{\mathbf{x}}_0$ in der Regularitätsmenge von $\hat{\mathbf{F}}$. Für $\hat{\mathbf{F}}$ sind damit alle Bedingungen des Satzes 2.2.3 erfüllt und es existiert ein Intervall $\hat{S}$ und eine stetig differenzierbare Funktion $\hat{\mathbf{c}} : \hat{S} \to \mathbb{R}^{m+1}$ mit $\hat{\mathbf{F}}(\hat{\mathbf{c}}(s)) = \mathbf{0}, s \in \hat{S}$.

Eine ähnliche Aussagen lässt sich auch für den Fall $\mathbf{F}(\mathbf{x}_0) \neq \mathbf{0}$ treffen, so lange $\|\mathbf{F}(\mathbf{x}_0)\|$ klein genug ist. Bei lokalen RB-Methoden wird aber grundsätzlich davon ausgegangen, dass der Wert $\mathbf{x}_0$ in dem eine Reduktion aufgebaut wird, eine Lösung von (3.2) ist.

Vereinfacht ausgedrückt besitzt das lokal reduzierte Problem also eine Lösung, falls

$$\text{Rang}(\mathbf{V}^T\mathbf{D}\mathbf{F}(\mathbf{x}_0)\mathbf{Z}) = m \tag{3.5}$$

gilt.

Zur Berechnung der reduktionsbasierten Approximation wird die hier beschriebene RB-Methode gewöhnlich mit den bereits erwähnten Astverfolgungsmethoden kombiniert, die sich sowohl auf das ursprüngliche Problem (3.2), als auch dessen Reduktion (3.3) anwenden lassen.

Ausgehend von einer bekannten Lösung $\mathbf{x}_0$ wird eine lokale Reduktion aufgebaut und per Astverfolgungsmethoden die Approximation $\mathbf{c}_R$ berechnet, bis die Norm des Residuums $\mathbf{F}(\mathbf{c}_R(s))$ eine vorgegebene Toleranz übersteigt. Ist dies der Fall muss das volldimensionale System (3.2) gelöst werden, um mittels eines Korrektors (meist wird hier das Newton-Verfahren verwendet) wieder "zurück" auf die echte Lösungskurve $\mathbf{c}$ zu gelangen. Dort wird dann eine neue Reduktion aufgebaut und der Ablauf beginnt von vorn. In Abbildung 3.1 ist dieses Verfahren skizziert.

Im Zusammenhang mit den in [2] vorgestellten Astverfolgungsmethoden lässt sich die lokale RB-Methode auch als Prädiktor-Korrektor-Verfahren interpretieren. Dabei stellt die Berechnung der Approximation $\mathbf{c}_R$ den Prädiktor-Schritt dar, auf den (sobald die Approximationsgüte nicht mehr ausreichend ist) ein Korrektor-Schritt folgt.

Im Folgenden wird erläutert, wie $\mathcal{Z}$ und $\mathcal{V}$ sinnvoll gewählt werden können. Dabei soll $\mathcal{Z}$ die Bewegung der Lösungskurve $\mathbf{c}$ um $\mathbf{x}_0$ bestmöglich approximieren und $\mathcal{V}$ die Gültigkeit von (3.5) sichern.

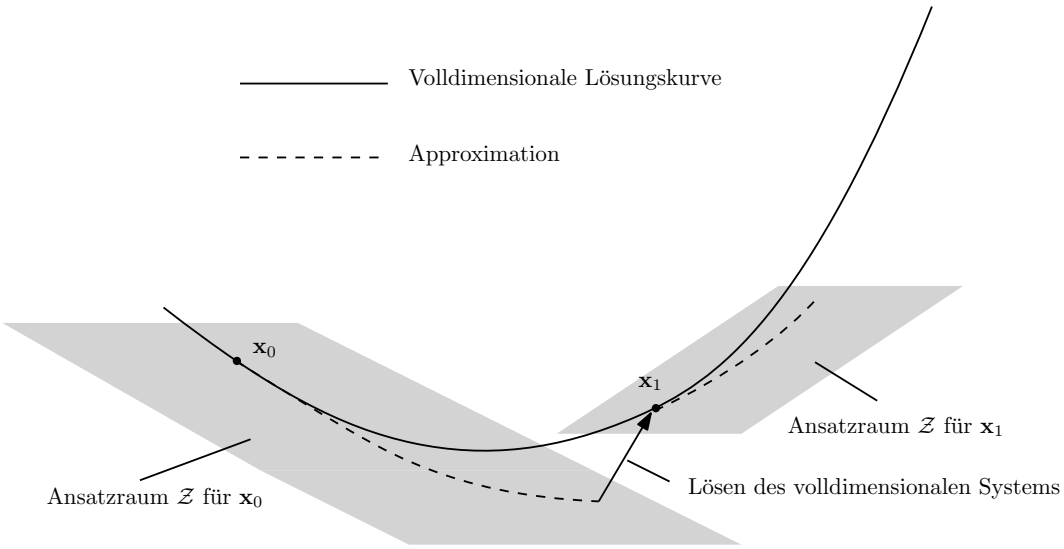

Abbildung 3.1: Lokale RB-Methode

3.1.2 Ansatz- und Testraum

Es existieren viele verschiedene Möglichkeiten, einen geeigneten Raum $\mathcal{Z}$ für die lokale RB-Methode aufzubauen, [42]. Bei allen wird versucht, Informationen über den lokalen Verlauf des Astes zu gewinnen und diese dann in den Aufbau von $\mathcal{Z}$ einfließen zu lassen. An dieser Stelle soll dabei auf zwei davon näher eingegangen werden.

Taylor-Ansatzraum

Basierend auf der Taylor-Approximation glatter Funktionen wird, unter Annahme, dass die ersten $m+1$ Ableitungen von $\mathbf{c}(s)$ im Punkt $s = 0$ existieren, der Ansatzraum über

$$\mathcal{Z} = \text{span}\{\mathbf{d}_i := \mathbf{c}^{(i)}(0), i = 1, \ldots, m+1\}$$

aufgebaut. In [44] und [43] wurde ein solcher Raum zur Basireduktion verwendet. Der Vektor $\mathbf{d}_1$ ist dabei ein Vielfaches des Tangentialvektors $\mathbf{T}(\mathbf{x}_0)$. Dieser lässt sich über das Lösen der Gleichung

$$\mathbf{DF}(\mathbf{x}_0)\mathbf{d}_1 = \mathbf{0}$$

berechnen. Die nächsten Ableitungen erfüllen dann

$$\begin{aligned}
&\mathbf{DF}(\mathbf{x}_0)\mathbf{d}_2 = -\mathbf{D}^2\mathbf{F}(\mathbf{x}_0)(\mathbf{d}_1, \mathbf{d}_1),\\
&\mathbf{DF}(\mathbf{x}_0)\mathbf{d}_3 = -3\mathbf{D}^2\mathbf{F}(\mathbf{x}_0)(\mathbf{d}_1, \mathbf{d}_2) - \mathbf{D}^3\mathbf{F}(\mathbf{x}_0)(\mathbf{d}_1, \mathbf{d}_1, \mathbf{d}_1),\\
&\mathbf{DF}(\mathbf{x}_0)\mathbf{d}_k = \cdots
\end{aligned}$$

wobei $\mathbf{D}^k\mathbf{F}$ die k-te Fréchet-Ableitung von $\mathbf{F}$ bezeichnen. Auf diese Weise lassen sich die Vektoren $\mathbf{d}^1, \ldots, \mathbf{d}^{m+1}$ rekursiv berechnen; jedoch ist die Auswertung der höheren Ableitungen von $\mathbf{F}$ numerisch sehr aufwändig. Es gibt allerdings Verfahren mit denen die höheren Ableitungen der Kurve $\mathbf{c}$ approximiert werden können, siehe dazu [38] und [68].

Lagrange-Ansatzraum

Eine Möglichkeit einen geeigneten lokalen Ansatzraum ohne die Ableitungen der Kurve aufzubauen stellt die Verwendung eines Lagrange-Ansatzraums dar. Hierbei werden eine Anzahl von Datenpunkten $\mathbf{x}_i := \mathbf{c}(s_i)$ in der Nähe von $\mathbf{x}_0$ gesammelt und der Ansatzraum mittels

$$\mathcal{Z} = \text{span}\{\mathbf{z}_i := \mathbf{x}_i - \mathbf{x}_0, i = 1, \ldots, m+1\}$$

aufgebaut. Wie bei der Lagrange-Interpolation wird das lokale Verhalten der Kurve über Knotenpunkte beschrieben. Die Güte dieser Approximation steigt, je Näher die Punkte $\mathbf{x}_i$ dem Startwert $\mathbf{x}_0$ sind. Der Lagrange-Ansatzraum nähert sich mit sinkender Entfernung der $\mathbf{x}_i$ zu $\mathbf{x}_0$ dem Taylor-Raum an, ist jedoch wesentlich einfacher zu berechnen.

Ein solcher Ansatz findet sich bereits in einer der ersten Arbeiten zur RB-Methode, [3], vergleiche dazu auch [49]. Da die vollständige Lösung $\mathbf{c}$ im Allgemeinen natürlich nicht bekannt ist, bedarf es zusätzlichen Überlegungen, wie eine Basis des Ansatzraums berechnet werden kann. Es können dafür zum Beispiel die bereits erwähnten Astverfolgungsmethoden herangezogen werden.

Bei den in Kapitel 3.2 besprochenen Methoden kommt ebenfalls ein Lagrange-Raum zum Einsatz. Dieser ist aber globaler Natur, das heißt die Punkte $\mathbf{x}_i$ werden innerhalb des gesamten Intervalls S (bzw. da dort der Parameter λ als separate Variable verbleibt, über das gesamte Parameterspektrum) gesammelt.

Testraum

Der Testraum $\mathcal{V}$, für den die Orthogonalität des Residuums $\mathbf{F}(\mathbf{x}_0 + \mathbf{Z}\hat{\mathbf{x}})$ gefordert wird muss die Eigenschaft

$$\text{Rang}(\mathbf{V}^T\mathbf{DF}(\mathbf{x}_0)\mathbf{Z}) = m$$

erfüllen, um zu garantieren, dass die Reduktion (3.3) eine Lösungskurve besitzt. Hierbei stellt $\mathbf{V} \in \mathbb{R}^{n,m}$ eine Matrix dar, deren Spalten eine Basis von $\mathcal{V}$ bilden. Für den Fall, dass $\mathbf{T}(\mathbf{x}_0) \in \mathcal{Z}$ gilt, lassen sich diese Matrizen $\mathcal{V}$ in eine Klasse zusammen fassen, wie in [38] durch folgendes Lemma bewiesen wurde.

Lemma 3.1.1. *Sei $\Omega \subset \mathbb{R}^{n+1}$ offen und $\mathbf{F} \in C^1(\Omega, \mathbb{R}^n)$. Sei weiterhin ein Punkt $\mathbf{x}_0 \in \mathcal{R}(\mathbf{F})$ gegeben, sowie $\mathbf{Z}^* = (\mathbf{z}_2, \ldots, \mathbf{z}_{m+1})$, sodass $\mathbf{T}(\mathbf{x}_0) \perp R(\mathbf{Z}^*)$ gilt, dann lässt sich jede Matrix $\mathbf{V} \in \mathbb{R}^{n,m}$ mit*

$$\textit{Rang}(\mathbf{V}^T\mathbf{DF}(\mathbf{x}_0)\mathbf{Z}^*) = m$$

in der Form

$$\mathbf{V} = \mathbf{A}\mathbf{D}\mathbf{F}(\mathbf{x}_0)\mathbf{Z}^*\mathbf{B} \tag{3.6}$$

darstellen, wobei $\mathbf{A} \in \mathbb{R}^{n,n}$ *symmetrisch positiv definit und* $\mathbf{B} \in \mathbb{R}^{m,m}$ *regulär ist.*

Die Matrix $\mathbf{Z}^*$ besitzt eine Spalte weniger als die bei der Reduktion (3.3) verwendete Matrix $\mathbf{Z}$, die den Ansatzraum $\mathcal{Z}$ repräsentiert. Eine solche Matrix $\mathbf{Z}^*$ erhält man für den Fall, dass $\mathbf{T}(\mathbf{x}_0) \in \mathcal{Z}$ gilt, da so eine Basis von $\mathcal{Z}$ über

$$\mathbf{Z} = (\mathbf{T}(\mathbf{x}_0), \mathbf{z}_2, \ldots, \mathbf{z}_{m+1})$$

existiert. Geht man nun davon aus, dass diese Vektoren eine Orthonormalbasis von $\mathcal{Z}$ sind, so stehen die Vektoren $\mathbf{z}_2, \ldots, \mathbf{z}_{m+1}$ alle senkrecht auf $\mathbf{T}(\mathbf{x}_0)$. Die reduzierte Jacobimatrix $\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)$ ergibt sich dann zu

$$\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0) = \mathbf{V}^T\mathbf{D}\mathbf{F}(\mathbf{x}_0)\mathbf{Z} = (\mathbf{0}, \mathbf{V}^T\mathbf{D}\mathbf{F}(\mathbf{x}_0)\mathbf{Z}^*)$$

und besitzt unter der Bedingung, dass $\mathbf{V}$ der Eigenschaft (3.6) genügt, vollen Zeilenrang. Die Grundidee basiert in diesem Fall darauf, dass $\mathbf{V}$ Informationen über die "Reaktion" des Funktionswertes $\mathbf{F}(\mathbf{x})$ orthogonal zum Tangentialvektor $\mathbf{T}(\mathbf{x}_0)$ in $\mathcal{Z}$ enthält.

Dabei ist die Forderung $\mathbf{T}(\mathbf{x}_0) \in \mathcal{Z}$ naheliegend, da der Tangentialvektor die beste eindimensionale Approximation der Kurvenbewegung in $\mathbf{x}_0$ darstellt, er also einen sinnvollen Anteil des Ansatzraumes darstellt. Des weiteren lässt sich der Tangentialvektor in $\mathbf{x}_0$ im Gegensatz zu den höheren Ableitungen mit dem einmaligen Lösen eines linearen $n \times n$-Gleichungssystems berechnen und ist für den Fall, dass beim Aufbau des Ansatzraumes das Newton-Verfahren zum Einsatz kommt, oft bereits bekannt.

3.2 Globale RB-Methoden

Der derzeitige Fokus hinsichtlich RB-Methoden liegt auf den globalen Methoden, bei denen die Lösbarkeit des reduzierten Systems global gesichert werden kann und der reine Rechenaufwand und weniger die Lösungsexistenz im Mittelpunkt der Forschung steht. Eine der ersten Anwendungen globaler RB-Methoden findet sich in [50], eine weiterführende Übersicht in [47] und [63].

Diese RB-Methoden werden dabei hauptsächlich auf diskretisierte parameterabhängige Differentialgleichungen angewendet, wobei die Anzahl der Parameter größer sein kann, als bei den lokalen Methoden. Die Diskretisierung der Differentialgleichung erfolgen zum größten Teil über die Finite-Elemente-Methode. In [25] wurde aber zum Beispiel auch die Anwendung der RB-Methode auf Finite-Volumen-Approximationen untersucht.

Die Gemeinsamkeit all dieser Methoden besteht darin, dass sich die untersuchten Gleichung

$$\mathbf{F}(\mathbf{u}, \lambda) = \mathbf{0}$$

stets nach $\mathbf{u}$ auflösen lassen, die Lösung also als Funktion von λ geschrieben werden kann. Es existiert somit eine direkte Zuordnung $\lambda \mapsto \mathbf{u}(\lambda)$, deren Zustandekommen im Folgenden erklärt wird. Dieser Sachverhalt ist entscheidend bei der Anwendung globaler RB-Methoden.

3.2.1 Reduktion mittels Snapshots

Das generelle Vorgehen bei einer Reduktion über globale RB-Methoden sowie die Frage nach der Existenz und Eindeutigkeit der Lösung des reduzierten Systems soll anhand eines einfachen (aber gebräuchlichen,[62, 39]) Beispiels erläutert werden. Sei dazu eine bezüglich u lineare Differentialgleichung gegeben, deren schwache Formulierung lautet: Finde $u \in X$, sodass für alle $v \in X$

$$a(u, v, \lambda) = f(v, \lambda) \tag{3.7}$$

gilt, wobei X einen Hilbertraum, $a : X \times X \times D \to \mathbb{R}$ eine parameterabhängige Bilinearform (bezüglich u und v) und $f : X \times D \to \mathbb{R}$ eine stetige Linearform (bezüglich v) darstellen. Die Menge D kann dabei mehrdimensional sein, das heißt, man ist nicht auf einen Parameter eingeschränkt. Ist a bezüglich $\lambda \in D$ gleichmäßig stetig ist und bezüglich u gleichmäßig koerziv, das heißt es gilt

$$\inf_{u \in X \setminus \{0\}} \frac{a(u, u, \lambda)}{\|u\|^2} > 0, \ \forall \lambda \in D,$$

dann besitzt das Problem (3.7) für alle $\lambda \in D$ eine eindeutige Lösung. Daher lässt sich (3.7) auch schreiben als

$$a(u(\lambda), v, \lambda) = f(v, \lambda).$$

Zur numerischen Approximation der Lösung einer solchen Gleichung wird zum Beispiel die Finite-Elemente-Methode benutzt. Dazu wird zunächst ein endlichdimensionaler Raum $X_h \subset X$ aufgestellt (der meistens aus stückweise stetigen Polynomen besteht). In diesem Raum wird dann eine Funktion u_h gesucht, sodass für alle $v \in X_h$

$$a(u_h, v, \lambda) = f(v, \lambda) \tag{3.8}$$

gilt. Sei nun $\mathbf{u}_h$ der Koeffizientenvektor der gesuchten Funktion u_h bezüglich einer Basis $\{\varphi_1, \ldots, \varphi_n\}$ von X_h und $\mathbf{A}(\lambda) \in \mathbb{R}^{n,n}$, $\mathbf{f} \in \mathbb{R}^{n,1}$ mit

$$\begin{aligned} A(\lambda)_{i,j} &:= a(\varphi_i, \varphi_j, \lambda), \ i, j = 1, \ldots, m, \\ f_i(\lambda) &:= f(\varphi_i, \lambda), \ i = 1, \ldots, m, \end{aligned}$$

so lässt sich (3.8) als lineares Gleichungssystem

$$\mathbf{A}(\lambda)\mathbf{u}_h = \mathbf{f}(\lambda) \tag{3.9}$$

schreiben. In dem hier betrachteten Fall ist die resultierende Finite-Elemente-Gleichung linear bezüglich $\mathbf{u}_h$. Ist aber die zu Grunde liegende Differentialgleichung nichtlinear, so wird es auch das diskretisierte Problem. Die Lösbarkeit von (3.9) leitet sich direkt aus der der ursprünglichen Gleichung (3.7) ab. Da $X_h \subset X$ gilt, überträgt sich die gleichmäßige Stetigkeit und Koerzivität auf das Finite-Elemente-Problem (3.8), wodurch sich eine eindeutige Lösung u_h (und damit auch $\mathbf{u}_h$) für jedes $\lambda \in D$ garantieren lässt. Schreibt man nun (3.9) mittels

$$\mathbf{F}(\mathbf{u}_h, \lambda) = \mathbf{A}(\lambda)\mathbf{u}_h - \mathbf{f}(\lambda) = \mathbf{0} \tag{3.10}$$

um, erhält man ein Problem der Form (1.1).

Solche Finiten-Elemente-Systeme sind in der Regel sehr groß und der Rechenaufwand, der für Parameterstudien bezüglich λ benötigt wird, kann die erlaubten Dimensionen (besonders im nichtlinearen Fall) schnell übersteigen.

Mit einer globalen RB-Methode wird (3.10) nun noch einmal reduziert. Dazu wird zunächst eine Menge von Parametern $\lambda_1, \ldots, \lambda_q$ ausgewählt und die dazugehörigen Lösungen $u_h(\lambda_i), i = 1, \ldots, q$ von (3.8) bestimmt. Diese werden "Snapshots" genannt. Basierend auf dieser Menge wird nun ein Ansatzraum $\mathcal{Z}$ bestimmt, zum Beispiel durch eine Auswahl von m Punkten aus der Snapshotmenge. Unter der Annahme, dass die ausgewählten Lösungen linear unabhängig sind, wird nun der Ansatzraum $\mathcal{Z}$ über

$$\mathcal{Z} := \operatorname{span}\{u_h(\lambda_i), i = 1, \ldots, m\} \subset X_h$$

erzeugt. Dies ist ein Lagrange-Ansatzraum, der nur eine Möglichkeit darstellt, den Ansatzraum aufzubauen. Alternativen bestehen zum Beispiel aus der Konstruktion mittels Proper Orthogonal Decomposition (POD), die in Kapitel 4.2.1 noch genauer betrachtet wird.

Dieser Ansatzraum führt zu dem reduzierten Problem: Finde $\hat{u} \in \mathcal{Z}$, sodass für alle $v \in \mathcal{Z}$

$$a(\hat{u}, v, \lambda) = f(v, \lambda) \tag{3.11}$$

gilt. Es ergeben sich so $m \ll n$ gleichungen für m Variablen und der Lösungsaufwand ist daher wesentlich geringer. Die Lösungsexistenz und -eindeutigkeit ergeben sich wieder direkt aus den Eigenschaften von a, die sich wegen $\mathcal{Z} \subset X_h$ auf den Ansatzraum $\mathcal{Z}$ übertragen. Sei nun $\{\hat{\varphi}_1, \ldots, \hat{\varphi}_m\}$ eine Basis von $\mathcal{Z}$ und $\hat{\mathbf{u}}$ der Koeffizientenvektor von $\hat{u}$ bezüglich dieser Basis, dann führt (3.11) analog zu (3.9) zu dem Gleichungssystem

$$\hat{\mathbf{A}}(\lambda)\hat{\mathbf{u}} = \hat{\mathbf{f}}(\lambda),$$

wobei sich $\hat{\mathbf{A}}(\lambda)$ und $\hat{\mathbf{f}}$ wiederum aus $A(\lambda)_{i,j} = a(\hat{\varphi}_i, \hat{\varphi}_j, \lambda)$ und $f_i(\lambda) = f(\hat{\varphi}_1, \lambda)$ ergeben. Um nun einen Zusammenhang mit (3.10) herzustellen wird die Matrix $\mathbf{Z} \in \mathbb{R}^{n,m}$ benötigt, deren Spalten $\mathbf{z}^j$ die Koeffizientenvektoren der Basisvektoren $\hat{\varphi}_j$ bezüglich der Basis $\{\varphi_i, i = 1, \ldots, n\}$ enthalten. Mit

$$\begin{aligned}\hat{A}(\lambda)_{i,j} = a(\hat{\varphi}_i, \hat{\varphi}_j, \lambda) &= a\left(\sum_{k=1}^{n} z_k^i \varphi_k, \sum_{l=1}^{n} z_l^j \varphi_l\right) \\ &= \sum_{k=1}^{n}\sum_{l=1}^{n} z_k^i z_l^j a(\varphi_k, \varphi_l, \lambda), \text{ und} \qquad (3.12)\end{aligned}$$

$$\hat{f}(\lambda)_i = f\left(\sum_{k=1}^{n} z_k^i \varphi_k, \lambda\right) = \sum_{k=1}^{n} z_k^i f(\varphi_k, \lambda) \qquad (3.13)$$

ergibt sich so $\hat{\mathbf{A}}(\lambda) = \mathbf{Z}^T\mathbf{A}(\lambda)\mathbf{Z}$ und $\hat{\mathbf{f}}(\lambda) = \mathbf{Z}^T\mathbf{f}(\lambda)$ und somit

$$\mathbf{Z}^T\mathbf{A}(\lambda)\mathbf{Z}\hat{\mathbf{u}} = \mathbf{Z}^T\mathbf{f}(\lambda).$$

In diesem einfachen linearen Beispiel bedeutet die eindeutige Lösbarkeit für jedes λ dass $\mathbf{Z}^T\mathbf{A}(\lambda)\mathbf{Z}$ für jedes $\lambda \in D$ regulär ist. Verwendet man wieder die Schreibweise (3.11) erhält man

$$\mathbf{Z}^T\mathbf{F}(\mathbf{Z}\hat{\mathbf{u}}, \lambda) = \mathbf{0},$$

also ein Problem der Form (1.5). Da hier ein einfacher Galerkin-Ansatz verwendet wurde, stimmen Ansatz- und Testraum überein. Wählt man die v in (3.11) aus einem zu $\mathcal{Z}$ verschiedenen Raum $\mathcal{V}$ ergibt sich ein Gleichungssystem der Form

$$\mathbf{V}^T\mathbf{F}(\mathbf{Z}\hat{\mathbf{u}}, \lambda) = \mathbf{0}.$$

Das Vorgehen bei der Anwendung der Finiten-Elemente-Methode auf eine partielle Differentialgleichung ist ähnlich zur globalen RB-Methode. In beiden Fällen wird ein niedrigdimensionaler Unterraum erzeugt, in dem dann eine Approximation gesucht wird. Bei den Finiten Elementen ist dies ein endlichdimensionaler Raum (meist bestehend aus stetigen stückweise polynomialen Funktionen), bei der RB-Methode ein mittels gesammelter Snapshots erzeugter Raum. Der Hauptunterschied zu den lokalen Methoden besteht darin, dass $\mathcal{Z}$ und $\mathcal{V}$ fest gewählt sind und die daraus resultierende Reduktion für alle $\lambda \in D$ verwendet werden kann.

Der Große Vorteil besteht nun darin, dass die Eigenschaften, die die Existenz und Eindeutigkeit der Lösung des Finiten-Elemente-Systems garantieren, auch für das RB-System gelten und daher für einen Parametervektor λ immer genau eine reduzierte Lösung $\hat{u}(\lambda)$ existiert. Auch für nichtlineare Probleme,

wie Differentialgleichungen, die sich aus der Fluiddynamik ergeben, [27], wird die Lösbarkeit des reduzierten Systems über vergleichbare funktionalanalytische Voraussetzungen gesichert, vergleiche dazu [34].

In den Ingenieurwissenschaften sind jedoch oftmals jene Probleme von besonderem Interesse, in denen eine eindeutige Zuordnung der Lösung einer Differentialgleichung zu den auftretenden Parametern nicht möglich ist (z.B. bei Hysterese-Effekten). In solchen Fällen ist eine Anwendung der oben beschriebenen Verfahren daher nicht möglich.

3.2.2 Offline-Online-Berechnungen

Einer der großen Vorteile, einen globalen Ansatz- und Testraum zu verwenden, ist die Möglichkeit, die anstehenden Berechnungen in einen (aufwändigen) Offline-Anteil und in einen (schnell zu berechnenden) Online-Anteil zu zerlegen. Auf diese Weise ist es möglich, Parameterstudien in einem kleinen System durchzuführen ohne eine neue Reduktion erstellen zu müssen. Dazu müssen die zu reduzierenden Gleichungen separierbar parametrisch sein. Für das im vorherigen Kapitel betrachtete Beispiel heißt das, dass es Funktionen $\Theta_a^q : D \to \mathbb{R}, q = 1, \ldots, Q_a$ und $\Theta_f^q : D \to \mathbb{R}, q = 1, \ldots Q_f$ geben muss, sodass sich a und f als

$$a(u, v, \lambda) = \sum_{q=1}^{Q_a} \Theta_a^q(\lambda) a^q(u, v), \text{ bzw.}$$

$$f(v, \lambda) = \sum_{q=1}^{Q_f} \Theta_f^q(\lambda) f^q(v),$$

schreiben lassen, wobei a^q und f^q parameterunabhängige Bilinear- bzw. Linearformen darstellen. Das daraus resultierende lineare Gleichungssystem hat dann die Form

$$\sum_{q=1}^{Q_a} \Theta_q^a(\lambda) \mathbf{Z}^T \mathbf{A}_q \mathbf{Z} \hat{\mathbf{u}} - \sum_{q=1}^{Q_f} \Theta_q^f(\lambda) \mathbf{Z}^T \mathbf{f}_q = \mathbf{0}, \tag{3.14}$$

wobei die parameterunabhängigen Matrizen $\mathbf{A}_q$ und $\mathbf{f}_q$ die Bilinear- und Linearformen b_q und f_q analog zu (3.12) und (3.13) repräsentieren. In der Offline-Phase werden nun folgende Schritte durchgeführt:

- Sammeln der für die Snapshots benötigten Parameter $\lambda_i,\ i = 1, \ldots, d$ und Berechnung der dazugehörigen Lösungen $\mathbf{u}(\lambda_i),\ i = 1, \ldots, q$
- Aufstellen eines Ansatzraumes $\mathcal{Z}$ und der dazugehörigen Matrix $\mathbf{Z}$
- Berechnung der Matrizen $\mathbf{Z}^T \mathbf{A}_q \mathbf{Z}, q = 1, \ldots, Q_a$ und $\mathbf{Z}^T \mathbf{f}_q, q = 1, \ldots, Q_f$

Diese Berechnungen stellen den aufwändigen Teil der RB-Methode dar.

In der Online-Phase findet nur das Lösen des Gleichungssystems (3.14) statt. In unserem Beispiel läuft dies für ein beliebiges λ auf das Lösen eines linearen Gleichungssystems der Größe $m \times m$ hinaus, für den nichtlinearen Fall wird meist die Newton-Methode zur Approximation von Nullstellen angewendet. In beiden Fällen ist der Rechenaufwand wegen $m \ll n$ jedoch nicht nur gering, sondern auch unabhängig von der Dimension n des ursprünglichen Problems (3.10). So lassen sich effizient Approximationen der Lösung für beliebige Parameter λ berechnen.

Für den Fall, dass die auftretenden Funktionale nicht parametrisch separierbar sind, wurde die Methode der Empirischen Interpolation entwickelt, die den nichtlinearen nichtseparierbaren Teil der Gleichung durch eine separierbare Approximation ersetzt. Dieses Verfahren findet sich das erste mal in [6] und wird in Kapitel 8.1 näher betrachtet.

Kapitel 4

Globale Ansatzräume für einparametrige nichtlineare Gleichungen

In dieser Arbeit werden neue Verfahren entwickelt, die in Kapitel 3.2 beschriebenen globalen RB-Methoden mit Aussagen der lokalen RB-Methode zu verknüpfen. Ziel ist es dabei, zweiparametrige nichtlineare Gleichungssysteme zu reduzieren, um effektive Parameterstudien zu ermöglichen. Zunächst wird dabei eine Reduktion bezüglich eines Parameters aufgebaut und diese dann für Parameterstudien bezüglich des zweiten Parameters genutzt. Dabei werden insbesondere Probleme, die bezüglich eines Parameters Umkehrpunkte nach Definition 2.2.2 besitzen und sich daher nicht über die bekannten globalen RB-Methoden reduzieren lassen, betrachtet.

Zunächst werden für das ursprüngliche einparametriges nichtlineare Problem der Form

$$\mathbf{F}(\mathbf{u}, \lambda) = \mathbf{0}$$

die Variablen $\mathbf{u}$ und λ wie in Kapitel 2.2.1 mittels $\mathbf{x} = (\mathbf{u}^T, \lambda)^T$ zusammegefasst.

Sei $\Omega \subset \mathbb{R}^{n+1}$ offen und $\mathbf{F} \in C^1(\Omega, \mathbb{R}^n)$. Es existiere nun eine Funktion $\mathbf{c} \in C^1(S, \mathcal{R}(\mathbf{F}))$ mit einem offenen Intervall $S \subset \mathbb{R}$ für die

$$\mathbf{F}(\mathbf{c}(s)) = \mathbf{0}, \ s \in S \tag{4.1}$$

gilt. Die Funktion $\mathbf{c}$ stellt eine einzelne Lösungskurve ohne Verzweigungen dar. Die Wahl ihrer Parametrisierung spielt in diesem Kapitel keine Rolle, man kann zum Beispiel wieder davon ausgehen, dass die Lösungskurve $\mathbf{c}$ nach der Bogenlänge parametrisiert ist.

In diesem Kapitel geht es um die Möglichkeiten einen geeigneten globalen

Ansatzraum für eine solche Lösungskurve aufzubauen. Man bestimmt zunächst eine Menge

$$X = \{\mathbf{x}_i = \mathbf{c}(s_i),\ s_i \in S,\ i = 1, \ldots, q\}. \tag{4.2}$$

Diese stellt das Äquivalent zu den bei den globalen RB-Methoden verwendeten Snapshots dar und kann zum Beispiel über eine auf das volldimensionale System angwendete Astverfolgung aufgebaut werden. Im Unterschied zu den lokalen Methoden wird das gesamte Intervall S betrachtet und nicht nur eine Umgebung um einen Startwert $\mathbf{x}_0 = \mathbf{c}(s_0)$. Ziel ist es jetzt einen Raum $\mathcal{Z}$ mit $\dim(\mathcal{Z}) = m + 1 < q$ zu finden, in dem eine sinnvolle reduzierte Lösung gefunden werden kann.

4.1 Lagrange-Ansatzraum

Eine Möglichkeit, einen Ansatzraum $\mathcal{Z}$ zu bestimmen, stellt der Aufbau eines Lagrange-Ansatzraumes dar. Dabei werden aus der Menge X direkt $m + 1$ Punkte ausgewählt und der Raum $\mathcal{Z}$ über deren Span erzeugt. Die Spalten der Matrix $\mathbf{Z} \in \mathbb{R}^{n+1,m+1}$ bilden also eine (Orthonormal-)Basis von

$$\mathcal{Z} = \text{span}\{\mathbf{x}_j \in X,\ j = 1, \ldots, m + 1\}.$$

Da unabhängig von der Wahl des Testraumes die Punkte $\mathbf{x}_j$ stets auch Lösung des reduzierten Problems sind, verläuft die gesuchte Approximation der Lösungskurve durch diese Punkte. Aufgrund dieser Interpolationseigenschaft wird ein so aufgebauter Ansatzraum in Anlehnung an die gleichnamige Interpolation Lagrange-Ansatzraum genannt.

Im Folgenden wird stets davon ausgegangen, dass die für den Ansatzraum ausewählten Punkte eine Menge linear unabhängiger Vektoren bildet. Da der durch diese Punkte aufgespannte Raum betrachtet wird, wird im Falle linearer Abhängigkeit lediglich eine Anpassung der Dimension m notwendig.

Der Aufbau des Lagrange-Ansatzraums hängt also direkt mit der Frage zusammen, wie die Punkte gewählt werden können, um die Menge X und damit den Verlauf der Lösungskurve sinnvoll wiederzugeben.

Beim Erstellen dieser Auswahl kommen meist Greedy-Algorithmen zum Einsatz. Der Raum wird dabei in jedem Schritt um einen Punkt $\mathbf{x}_i \in X$ erweitert, der bezüglich bestimmter Kriterien die größte Verbesserung für den Ansatzraum darstellt. Im Zusammenhang mit RB-Methoden fand ein Greedy-Algorithmus das erste Mal in [72] Verwendung. Eine breite Übersicht über solche Algorithmen findet sich zum Beispiel in [73].

Um zu erklären, welche Probleme bei der Anwendung des in [47] verwendete Greedy-Algorithmus für eine Lösungskurve (4.1) auftreten, sei dieser hier

kurz erklärt. Üblicherweise bleibt bei den bekannten globalen RB-Methoden der Parameter λ separiert, und das Problem $\mathbf{F}(\mathbf{u}, \lambda) = \mathbf{0}$ besitzt für jeden Parameter λ eine eindeutige Lösung $\mathbf{u}(\lambda)$. Die Menge der Snapshots hat dann die Gestalt

$$X = \{\mathbf{u}_i = \mathbf{u}(\lambda_j),\ j = 1, \ldots, q\}$$

und wird von einer Menge von Parametern $D = \{\lambda_j,\ j = 1, \ldots, q\}$ bestimmt.

Für jede Auswahl $X_k \subset X$ und damit verbundenen Unterraum

$$\mathcal{Z}_k := \text{span}\{\mathbf{x}_j,\ j = 1, \ldots, k|\ \mathbf{x}_j \in X_k\}$$

existiert nun eine Lösung $\mathbf{u}_{\mathcal{Z}_k}$ des reduzierten Problems. Insbesondere lässt sich die auftretende Fehlerfunktion über

$$e_{\mathcal{Z}_k}(\lambda) := \|\mathbf{u}(\lambda) - \mathbf{u}_{\mathcal{Z}_k}(\lambda)\| \tag{4.3}$$

angeben. Für diesen Fehler existieren effiziente Fehlerschätzer $\Delta_{\mathcal{Z}_k}(\lambda)$, die numerisch unabhängig von der Raumdimension n berechnet werden können.

Ausgehend von einer gegebene Parametermenge $D_j = \{\lambda_i^*,\ i = 1, \ldots, j\} \subset D$, sowie des dazugehörigen Unterraumes $\mathcal{Z}_j := \text{span}\{\mathbf{u}(\lambda_i^*),\ i = 1, \ldots, j\}$ sucht der Algorithmus nun nach dem $\lambda^* \in D$, das für den größten Fehler bezüglich der bestehenden Auswahl an Parametern sorgt, für das also der Fehlerschätzer $\Delta_{\mathcal{Z}_j}(\lambda)$ maximal wird. Die dazugehörige Lösung $\mathbf{u}(\lambda^*) \in X$ ist dann jene, die durch die Reduktion mit dem bisherigen Ansatzraum $\mathcal{Z}_j$ am schlechtetesten approximiert wird. Die Menge D_j wird dann um den Parameter λ^* und der bisherige Ansatzraum um $\text{span}\{\mathbf{u}(\lambda)\}$ ergänzt. Algorithmisch lässt sich dies wie folgt beschreiben:

for $j = 2 : m + 1$ **do**
  $\lambda_j^* = \arg\max_D \Delta_{\mathcal{Z}_{j-1}}(\lambda)$;
  $D_j = D_{j-1} \cup \lambda_j^*$;
  $\mathcal{Z}_j = \mathcal{Z}_{j-1} + \text{span}\{\mathbf{u}(\lambda_j^*)\}$;
end for

Die Schleife kann natürlich vorher beendet werden, falls $\Delta_{\mathcal{Z}_{j-1}}(\lambda_j^*)$ eine vorgegeben Fehlertoleranz unterschreitet. Für den Startwert λ_1^* werden dabei keine genauen Angaben gemacht, oft wird der Wert $\lambda_1 \in D$ gewählt, oder auch

$$\lambda_1^* = \arg\max_{\lambda \in D} \|\mathbf{u}(\lambda)\|. \tag{4.4}$$

Die erzeugten Räume $\mathcal{Z}_j$ sind hierarchisch, das heißt es gilt $\mathcal{Z}_{j-1} \subset \mathcal{Z}_j$, was für die Resultate bezüglich Konvergenz und Offline-Online-Zerlegung in [47] von Bedeutung ist.

Sei nun eine Funktion und Lösungskurve wie in (4.1) gegeben und die Menge

X wie in (4.2). Ein Fehlerschätzer, wie er im Greedy-Algorithmus Verwendung findet, ließ sich für diesen Fall aus zwei Gründen nicht herleiten. Zum einen lässt sich keine Fehlerfunktion der Form (4.3) finden, da keine gemeinsame Parametrisierung der volldimensionalen Lösungskurve $\mathbf{c}$ und ihrer Approximation $\mathbf{c}_R$ existieren muss. Des Weiteren ist die Existenz einer Approximation $\mathbf{c}_R$ für die mittels Greedy-Algorithmus gewählten Mengen X_j^* nicht gesichert, bzw. hängt maßgeblich von der Wahl des Testraumes $\mathcal{V}$ ab, der in Kapitel 5 näher betrachtet wird.

Statt eines Fehlerschätzers wird daher der Projektionsfehler als Entscheidungskriterium für die Auswahl der Snapshots benutzt, siehe dazu auch [8]. Anstatt also direkt nach einem Unterraum zu suchen, für den der Reduktionsfehler klein wird, begnügt man sich in dieser Arbeit mit einem, der die Menge X im Sinne des Projektionsfehlers gut genug approximiert.

Ausgehend von einem Startwert $D_1 = \{p_1\}$ mit $p_1 \in \{1, \ldots, q\}$ und $\mathcal{Z}_1 := \operatorname{span}\{\mathbf{x}_{p_1}\}$ arbeitet der Algorithmus dann wie folgt:

for $j = 2 : m+1$ **do**
$p_j = \arg\max_{k \in \{1,\ldots,q\}} \|\mathbf{x}_k - \mathbf{P}_{j-1}\mathbf{x}_k\|$;
$D_j = D_{j-1} \cup p_j$;
$\mathcal{Z}_j = \mathcal{Z}_{j-1} + \operatorname{span}\{\mathbf{x}_{p_j}\}$;
end for.

Hier bezeichnet $\mathbf{P}_{j-1}$ den orthogonalen Projektor auf den Raum $\mathcal{Z}_{j-1}$. Die Matrix $\mathbf{Z}$ ergibt sich dann als Orthonormalbasis des Raumes $\mathcal{Z}_{m+1}$.

Den Startpunkt $\mathbf{x}_1$ kann man zum Beispiel als

$$\mathbf{x}_1 := \arg\min_{\mathbf{x} \in X} \|\mathbf{x}\|$$

wählen. Analog zu (4.4) kann $\mathbf{x}_1$ auch als $\arg\max_{\mathbf{x} \in X} \|\mathbf{x}\|$ gewählt werden. Beide Möglichkeiten liefern aber ab einer gewissen Größe von q ähnliche Resultate.

Es ist oft sinnvoll die Werte aus X um deren Mittelwert in Richtung des Ursprungs zu verschieben, bevor der Unterraum $\mathcal{Z}$ aufgebaut wird. Der Mittelwert $\bar{\mathbf{x}}$ ergibt sich aus

$$\bar{\mathbf{x}} := \frac{1}{q} \sum_{i=1}^{q} \mathbf{x}_i$$

und erhält die neue Menge

$$X^* := \{\mathbf{x}_i^* := \mathbf{x}_i - \bar{\mathbf{x}},\ \mathbf{x}_i \in X,\ i = 1, \ldots, q\}.$$

Auf diese Menge wendet man nun wieder den oben beschriebenen Algorithmus an und erhält so einen Unterraum $\mathcal{Z}$ der Dimension $m+1$. Macht man nun einen affinen Ansatz und wählt als Ansatzraum

$$\bar{\mathbf{x}} + \mathcal{Z}$$

enthält dieser eine Auswahl an Punkten aus X und stellt einen affinen Lagrange-Ansatzraum dar. Alternativ kann man das gesamte Problem auch um den Wert $\bar{\mathbf{x}}$ verschieben und für

$$\mathbf{F}^*(\mathbf{x}) := \mathbf{F}(\mathbf{x} + \bar{\mathbf{x}})$$

wieder mit einem linearen Ansatzraum arbeiten.

4.2 POD-Ansatzraum

Benutzt man den Projektionsfehler als ausschlaggebendes Kriterium für die Güte eines Ansatzraumes, führt dies zu der Frage, welcher Unterraum $\mathcal{Z}$ für die Menge $X = \{\mathbf{x}_j,\ j = 1, \ldots, q\}$ das Minimierungsproblem

$$\min_{\dim(\mathcal{Z})=m+1} \sum_{j=1}^{q} \|\mathbf{x}_j - \mathbf{P}\mathbf{x}_j\|^2 \tag{4.5}$$

löst, wobei der $\mathbf{P}$ den orthogonalen Projektor auf $\mathcal{Z}$ darstellt. Hierbei ist es nun nicht mehr notwendig, dass Punkte der Menge X in $\mathcal{Z}$ liegen. Dieses Minimierungsproblem lässt sich mittels Proper Orthogonal Decomposition (POD) lösen, die im endlichdimensionalen Fall mit der Singulärwertzerlegung übereinstimmt. Da im Zusammenhang mit RB-Methoden gewöhnlich der Begriff POD verwendet wird, wird diesem in dieser Arbeit der Vorzug gegeben.

4.2.1 Proper Orthogonal Decomposition

Die POD, deren erste Erwähnung sich in [48] findet, ist eine Verallgemeinerung der Diagonalisierung und existiert für alle Matrizen $\mathbf{A} \in \mathbb{C}^{n,m}$, wobei in dieser Arbeit nur reelle Matrizen behandelt werden. In [4] findet sich eine detaillierte Herleitung der POD, sowie die Beweise für die folgenden Aussagen.

Sei $\mathbf{A} \in \mathbb{R}^{n,m}$ und ohne Beschränkung der Allgemeinheit $n > m$, dann existiert eine Zerlegung der Form

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{W}^T \tag{4.6}$$

wobei $\mathbf{U} \in \mathbb{R}^{n,n}$ und $\mathbf{W} \in \mathbb{R}^{m,m}$ zwei orthonormale Matrizen sind und die Matrix Σ die gestalt

$$\Sigma = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \\ \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{n,m}.$$

Die Werte $\sigma_j \geq 0$ werden dabei Singulärwerte genannt und erfüllen die Bedingung

$$\mathbf{A}\mathbf{u}_i = \sigma_i \mathbf{w}_i, \quad \mathbf{A}^T\mathbf{w}_i = \mathbf{u}_i, \ i = 1, \ldots, m,$$

wobei $\mathbf{u}_i$ und $\mathbf{w}_i$ jeweils die Spalten von $\mathbf{U}$ bzw. $\mathbf{W}$ bezeichnen. Außerdem sind sie die Wurzeln der Eigenwerte der Matrizen $\mathbf{A}\mathbf{A}^T$ bzw. $\mathbf{A}^T\mathbf{A}$. Daher gilt auch der Zusammenhang

$$\|\mathbf{A}\|_2 = \sigma_1.$$

Die Darstellung (4.6) ist dabei nicht eindeutig, die Singulärwerte allerdings schon. Traditionellerweise geht man davon aus, dass sie in der Reihenfolge $\sigma_1 \geq \cdots \geq \sigma_m$ geordnet sind und es gilt $\sigma_j = 0, \ j > \text{Rang}(\mathbf{A})$. Seien nun

$$\mathbf{U}_k := (\mathbf{u}_1, \ldots, \mathbf{u}_k), \quad \mathbf{W}_k := (\mathbf{w}_1, \ldots, \mathbf{w}_k), \ 1 \leq k \leq m, \tag{4.7}$$

dann lässt sich $\mathbf{A}$ mit $k = \text{Rang}(\mathbf{A})$ in der reduzierten POD über

$$\mathbf{A} = \mathbf{U}_k \Sigma_k \mathbf{W}_k^T, \text{ mit } \Sigma_k := \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{pmatrix} \tag{4.8}$$

darstellen. Die POD ist eng mit der Frage verknüpft, wie sich eine gegebene Matrix möglichst gut durch eine andere Matrix mit niedrigerem Rang approximieren lässt. Dies wurde zum Beispiel bereits in [19] untersucht. Bezüglich der euklidischen Norm erhält man eine Lösung des Minimierungsproblems

$$\min_{\mathbf{X} \in \mathbb{R}^{n,m}, \text{Rang}(\mathbf{X}) = j} \|\mathbf{A} - \mathbf{X}\|_F$$

über

$$\mathbf{X}^* = \mathbf{U}_j \Sigma_j \mathbf{W}_j^T.$$

Außerdem gilt

$$\|\mathbf{A} - \mathbf{X}^*\| = \sigma_{j+1}(\mathbf{A}).$$

4.2.2 Aufstellen eines Ansatzraums mittels POD

Basierend auf den Eigenschaften der POD lässt sich ein Zusammenhang mit dem Minimierungsproblem (4.5) herstellen. Dazu sei wieder die Diskretisierung der Lösungskurve durch $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_q\}$ und die Matrix $\mathbf{X} \in \mathbb{R}^{n+1,q}$ über $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_q)$ gegeben. Mit der POD

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{W}^T$$

und $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_{n+1})$ gilt dann

$$\arg \min_{\dim(\mathcal{Z})=m+1} \sum_{j=1}^{q} \|\mathbf{x}_j - \mathbf{P}\mathbf{x}_j\|^2 = \operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_{m+1}\}. \tag{4.9}$$

Mittels der POD ist es also möglich einen bezüglich der Summe der Projektionsfehler idealen Ansatzraum der Dimension $m+1$ zu finden.

Die diesen Ansatzraum repräsentierende Matrix $\mathbf{Z}$ kann also als

$$\mathbf{Z} := \mathbf{U}_{m+1}$$

gewählt werden, wobei $\mathbf{U}_{m+1}$ wie in (4.7) aufgebaut ist. Weiterhin gilt

$$\sum_{j=1}^{q} \|\mathbf{x}_j - \mathbf{P}\mathbf{x}_j\|^2 = \sum_{j=m+2}^{n+1} \sigma_j^2, \tag{4.10}$$

wobei $\mathbf{P}$ den orthogonalen Projektor auf $\mathbf{U}_{m+1}$ bezeichnet. Die Summe der Quadrate der Projektionsfehler wird also 0, sobald $m+1$ den Rang der Matrix $\mathbf{X}$ erreicht hat, weil dann $\sigma_{m+2} = \cdots = \sigma_{n+1} = 0$ gilt. In Gleichung (4.10) wird davon ausgegangen, dass $q < n+1$ gilt, $\mathbf{X}$ also mehr Zeilen als Spalten besitzt. Ist dies nicht der Fall, muss die obere Grenze der Summe der rechten Seite durch q ersetzt werden. Beweise für diese Aussagen finden sich zum Beispiel in [74].

Der mittels POD erzeugte Ansatzraum ist kein Lagrange-Ansatzraum, das heißt es ist nicht gesichert, dass sich Snapshots $\mathbf{x}_i \in X$ in $\mathcal{Z}$ befinden. Basisreduktionen mittels POD sind weit verbreitet und wurden erfolgreich auf Probleme der Fluiddynamik, [34], bei nichltinearen elliptischen und parabolischen Systemen, oder gewöhnlichen Differentialgleichungen, [35, 53], sowie dynamischen Systemen, [15] angewendet.

Genau wie bei den Lagrange-Ansatzräumen kann es sinnvoll sein, die Menge X wieder um den Mittelwert

$$\bar{\mathbf{x}} = \frac{1}{q} \sum_{i=1}^{q} \mathbf{x}_i$$

zu verschieben und die POD auf die Menge $X^* := \{\mathbf{x}_i - \bar{\mathbf{x}},\ \mathbf{x}_i \in X,\ i = 1, \ldots, q\}$ anzuwenden. Als Ansatzraum fungiert dann $\bar{\mathbf{x}} + \mathcal{Z}$, wobei $\mathcal{Z}$ sich nun aus der POD von X^* ergibt.

Ist die Kurve $\mathbf{c}$ genügend glatt, findet ihre Bewegung lokal in nur wenigen Raumdimensionen statt, was direkt aus der Taylor-Reihenentwicklung folgt. Dieser Sachverhalt überträgt sich auf die gesamte Kurve, sodass ein rapider Abfall der Projektionsfehler mit zunehmender Dimension des Ansatzraumes zu erwarten ist. Zwar existiert mit (4.10) eine exakte Angabe der summierten

Projektionsfehler, sie enthält jedoch keine Informationen, über den qualitativen Zusammenhang dieses Fehlers mit der Dimension des Ansatzraumes (sicher ist nur, dass er irgendwann 0 ist, wenn m groß genug gewählt wird). Zudem erfordert die Fehlerdarstellung 4.10 die aufwändige Berechnung der restlichen Singulärwerte.

Für den Fall, dass die Lösungskurve $\mathbf{c}$ zusätzlichen Glattheitskriterien genügt, wird im Folgenden eine obere Schranke für den Fehler hergeleitet, die genauere Aussagen bezüglich der Konvergenzgeschwindigkeit der Projektionsfehler zulässt.

4.2.3 Abschätzung des Projektionsfehlers

Bei der folgenden Abschätzung wird ausgenutzt, dass der POD-Ansatzraum das Minimierungsproblem (4.5) löst. Dadurch lässt sich der summierte Projektionsfehler bezüglich eines beliebigen anderen $(m+1)$-Dimensionalen Unterraums als obere Schranke verwenden. Hier wird dazu ein Lagrange-Raum verwendet, der eine kubische Spline-Interpolation bezüglich $(m+1)$ Knotenpunkten, die entlang der Kurve $\mathbf{c}$ gewählt werden, enthält. Ist $\mathbf{c}$ genügend glatt, kann der Projektionsfehler auf den auftretenden Interpolationsfehler abgeschätzt werden. Steigt die Dimension m des Unterraums, erhöht dies die Knotenanzahl des Splines, wodurch sich der Interpolationsfehler verringert. Auf diese Weise lässt sich eine qualitative Verbindung der summierten Projektionsfehler und der Dimension des Ansatzraumes herstellen.

Zunächst seien einige Eigenschaften eines interpolierenden kubischen Splines h festgehalten. Für eine Funktion $f \in C^4([a,b],\mathbb{R})$ genügt diese Interpolation in den Knotenpunkten $x_i \in [a,b],\ i=1,\dots,p$ den Bedingungen

$$h(x_i) = f(x_i),\ h^{(j)}(x_i - 0) = h^{(j)}(x_i + 0),\ i = 1,\dots,p-1,\ j = 1,2.$$

In den Randpunkten $x_0 = a$ und $x_p = b$ gelte weiterhin

$$h'(x_0) = h'(x_n),\ h''(x_0) = h''(x_n).$$

Es lässt sich zeigen (zum Beispiel in [64]), dass für den Interpolationsfehler

$$\|f - h\|_\infty \leq \frac{9}{8}|\Delta|^4 \|f^{(4)}\|_\infty \tag{4.11}$$

gilt, mit $|\Delta| = \max_{i=1,\dots,p}(x_{i+1} - x_i)$. Basierend auf dieser Abschätzung ist es nun möglich eine obere Schranke für die Summe der quadrierten Projektionsfehler, die bei der Verwendung eines POD-Ansatzraumes auftritt, anzugeben.

Satz 4.2.1. *Sei $S \subset \mathbb{R}$ offen und $\mathbf{c} \in C^4(S,\mathbb{R}^{n+1})$, sowie eine Menge $X \subset \mathbb{R}^{n+1}$ gegeben als*

$$X = \{\mathbf{x}_j = \mathbf{c}(s_j),\ j = 1,\dots,q\}.$$

Sei weiterhin $\mathcal{Z} = span\{\mathbf{u}_1, \ldots, \mathbf{u}_{m+1}\}$, *wobei* $\mathbf{u}_i$ *die Spalten der Matrix* $\mathbf{U}$ *der POD*

$$X = \mathbf{U}\Sigma\mathbf{W}^T$$

darstellen. Sei nun $\mathbf{P}$ *der orthogonale Projektor auf* $\mathcal{Z}$, *dann existiert eine Konstante* $C > 0$, *sodass für den Projektionsfehler*

$$\sum_{j=1}^{q} \|\mathbf{x}_j - \mathbf{P}\mathbf{x}_j\|^2 \leq C\frac{q}{m^4}$$

gilt.

Beweis. Sei $H := (s_q - s_1)/m$, sowie $\tilde{s}_j := s_1 + jH$ und $\tilde{\mathbf{x}}_j := \mathbf{c}(\tilde{s}_j)$, $j = 0, \ldots, m$. Sei $Y := \text{span}\{\tilde{\mathbf{x}}_j,\ j = 0, \ldots, m\}$ der Unterraum, der alle Punkte $\tilde{\mathbf{x}}_j$ enthält.

Es sei ohne Einschränkung angenommen, dass die Vektoren $\tilde{\mathbf{x}}_j$, $j = 0, \ldots, m$ linear unabhängig sind. Ist dies nicht der Fall, kann mit einer dynamischen Schrittweite H_j stets eine linear unabhängige Menge $\{\tilde{\mathbf{x}}_j = \mathbf{c}(s_{j-1} + H_j),\ j = 1, \ldots, m+1\}$ gefunden werden (bis $m + 1 = \text{Rang}(X)$ gilt, wodurch der abzuschätzende Fehler 0 wird). Die folgenden Abschätzungen können auch für eine dynamische Schrittweite getroffen werden, an dieser Stelle sei nur der Übersicht halber ein festes H verwendet.

Da $\mathcal{Z}$ das Minimierungsproblem (4.5) löst, gilt

$$\sum_{j=1}^{q} \|\mathbf{x}_j - \mathbf{P}\mathbf{x}_j\|^2 \leq \sum_{j=1}^{q} \|\mathbf{x}_j - \mathbf{P}_Y\mathbf{x}_j\|^2, \tag{4.12}$$

wobei $\mathbf{P}_Y$ den orthogonalen Projektor auf Y bezeichnet. Sei nun mit $h_i : [s_1, s_q] \to \mathbb{R}$ der kubische Spline bezeichnet, der die Datenpunkte $(\tilde{s}_j, (\tilde{\mathbf{x}}_j)_i)$, $j = 0, \ldots, m$ interpoliert. Diese seien in einem Vektor $\mathbf{h}(s) := (h_1(s), \ldots, h_n(s))^T$ zusammengefasst. Da sich alle Punkte $\mathbf{h}(s)$ als Linearkombinationen der $\tilde{\mathbf{x}}_j$ ergeben, gilt für alle $s \in [s_1, s_q]$

$$\mathbf{h}(s) \in Y.$$

Daher gilt für alle $s \in [s_1, s_q]$ die Abschätzung $\|\mathbf{c}(s) - \mathbf{P}_Y\mathbf{c}(s)\| \leq \|\mathbf{c}(s) - \mathbf{h}(s)\|$.

Zusammen mit (4.11), wobei hier nun $|\Delta| = H$ gilt, ergibt sich

$$\begin{aligned}\sum_{j=1}^{q} \|\mathbf{x}_j - \mathbf{P}_Y \mathbf{x}_j\|^2 = \sum_{j=1}^{q} \|\mathbf{c}(s_j) - \mathbf{P}_Y \mathbf{c}(s_j)\|^2 &\leq \sum_{j=1}^{q} \|\mathbf{c}(s_j) - \mathbf{h}(s_j)\|^2 \\ &\leq (n+1) \sum_{j=1}^{q} \max_{i=1,\ldots,n+1} |c_i(s_j) - h_i(s_j)|^2 \\ &\leq (n+1) \sum_{j=1}^{q} \max_{i=1,\ldots,n+1} \|c_i - h_i\|_\infty^2 \\ &\leq \frac{9(n+1)}{8} H^4 \sum_{j=1}^{q} \max_{i=1,\ldots,n+1} \|c_i^{(iv)}\|_\infty^2.\end{aligned}$$

Da $\mathbf{c} \in C^4(S, \mathbb{R}^{n+1})$ gilt, existiert eine Konstante k mit $\max_{j=1,\ldots,n+1} \|c_j^{(iv)}\|_\infty^2 \leq k$. Setzt man nun

$$C := \frac{9k}{8}(n+1)(s_q - s_1)^4$$

ergibt sich mit (4.12) und $H = (s_q - s_1)/m$

$$\sum_{j=1}^{q} \|\mathbf{x}_j - \mathbf{P}\mathbf{x}_j\|^2 \leq C \frac{q}{m^4}.$$

□

4.2.4 Bestmögliche affine Verschiebung

Die bisher betrachteten Verfahren zum Aufbau der Ansatzräume lieferten stets lineare Unterräume. In diesem Abschnitt soll daher näher auf mögliche Vorteile bei der Verwendung von affinen Ansatzräumen eingegangen werden. Bei lokalen RB-Methoden wird stets ein affiner Ansatzraum verwendet, da die lokale Reduktion bezüglich eines festen Punktes $\mathbf{x}_0$ aufgebaut wird (siehe Kapitel 3.1.2). Da im globalen Fall ein Ansatzraum für eine gesamte Lösungskurve gesucht wird, ist die Verwendung eines affinen Unterraums weniger zwingend.

In diesem Kapitel wurden bereits zwei Möglichkeiten beschrieben, geeignete Ansatzräume für eine gegebene (diskretisierte) Lösungskurve aufzubauen. Beide lassen die Möglichkeit zu, anstelle eines linearen Unterraums einen affinen zu verwenden. Die Menge X, die die Punkte der volldimensionalen Lösungskurve enthält, wird zunächst um den Mittelwert $\bar{\mathbf{x}}$ dieser Punkte in Richtung Ursprung verschoben um dann einen linearen Unterraum $\mathcal{Z}$ aufzubauen. Eine solche Verschiebung ist sinnvoll, da mit Hilfe des Ansatzraumes die Bewegung der Lösungskurve approximiert werden soll und für diese die Lage der Kurve bezüglich des Ursprunges keine Rolle spielt.

Um nun diesen Ansatzraum für eine Reduktion zu verwenden, kann dieser entweder als um den Mittelwert verschobener affiner Raum $\bar{\mathbf{x}} + \mathcal{Z}$ verwendet werden, oder man betrachtet ein äquivalentes Problem der Form

$$\mathbf{F}^*(\mathbf{x}) := \mathbf{F}(\mathbf{x} + \bar{\mathbf{x}}) = \mathbf{0}. \tag{4.13}$$

Beide Betrachtungsweisen sind äquivalent, das heißt die Resultate, die sich aus der Reduktion ergeben unterscheiden sich nicht.

Sei daher nun der Fall (4.13) betrachtet und $\mathcal{Z}$ ein Ansatzraum der über eines der in Kapitel 4.1 und 4.2 beschriebenen Verfahren aufgebaut wurde. Die Summe der Projektionsfehler diente bei diesem Aufbau als Hauptkriterium für die Güte des Unterraums. Der summierte Projektionsfehler kann nun durch die Verwendung einer zusätzlichen affinen Verschiebung noch einmal verringert werden. Die Frage, wie eine affine Verschiebung gewählt werden kann, um für einen gegeben Unterraum die Summe der Projektionsfehler bezüglich des resultierenden affinen Raums zu minimieren, wird durch folgendes Lemma beantwortet.

Lemma 4.2.2. *Sei $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_q\} \subset \mathbb{R}^{n+1}$ eine Menge von Punkten und $\mathcal{Z}$ ein $m+1$-dimensionaler Unterraum des $\mathbb{R}^{n+1}$. Sei weiterhin $\mathbf{P}$ der orthogonale Projektor auf diesen Raum und $\mathbf{P}_\mathbf{b}$ der orthogonale Projektor auf den affinen Raum $\mathbf{b} + \mathcal{Z}$ mit*

$$\mathbf{P}_\mathbf{b}\mathbf{x} = \mathbf{P}(\mathbf{x} - \mathbf{b}) + \mathbf{b},\ \mathbf{x} \in \mathbb{R}^n.$$

Dann wird das Minimierungsproblem

$$\min_{\mathbf{b} \in \mathbb{R}^{n+1}} \sum_{j=1}^{q} \|\mathbf{x}_j - \mathbf{P}_\mathbf{b}\mathbf{x}_j\|^2$$

durch den Vektor

$$\mathbf{b} = \frac{1}{q} \sum_{j=1}^{q} (\mathbf{x}_j - \mathbf{P}\mathbf{x}_j)$$

gelöst.

Beweis. Man kann ohne Einschränkung davon ausgehen, dass $\mathbf{b} \in \mathcal{Z}^\perp$, die affine Verschiebung also nur senkrecht bezüglich $\mathcal{Z}$ erfolgen soll. Es gilt $\mathbf{P}_\mathbf{b}\mathbf{x}_j = \mathbf{P}(\mathbf{x}_j - \mathbf{b}) + \mathbf{b}$ und somit

$$\sum_{j=1}^{q} \|\mathbf{x}_j - \mathbf{P}_\mathbf{b}\mathbf{x}_j\|^2 = \sum_{j=1}^{q} \|\mathbf{x}_j - \mathbf{P}\mathbf{x}_j - \mathbf{b}\|^2. \tag{4.14}$$

Der Gradient dieses Terms bezüglich $\mathbf{b}$ muss in einer Extremstelle in $\mathbf{b}$ den Wert $\mathbf{0}$ annehmen, somit ergibt sich

$$\nabla \left(\sum_{j=1}^{q} \|\mathbf{x}_j - \mathbf{P}\mathbf{x}_j - \mathbf{b}\|^2 \right) = -2 \sum_{j=1}^{q} (\mathbf{x} - \mathbf{P}\mathbf{x} - \mathbf{b}) = \mathbf{0}.$$

Umstellen dieser Gleichung ergibt

$$\mathbf{b} = \frac{1}{q} \sum_{i=1}^{q} (\mathbf{x}_j - \mathbf{P}\mathbf{x}_j). \tag{4.15}$$

Die zweite Ableitung des Fehlerterms (4.14) ist $2 \cdot \mathbf{I}$ und somit positiv definit, womit (4.15) ein Minimum ist. □

Der Vektor $\mathbf{b}$ aus Lemma 4.2.2 wird in dieser Arbeit als die bestmögliche affine Verschiebung von $\mathcal{Z}$ bezeichnet. Man beachte, dass wenn es sich bei $\mathcal{Z}$ um einen Lagrange-Raum handelt, der Raum also so aufgebaut wurde, dass er alle Punkte $\mathbf{x}_j$ (bzw. $(\mathbf{x}_j - \hat{\mathbf{x}})$ bei vorheriger Verschiebung um den Mittelwert) enthält, der affine Raum $\mathbf{b} + \mathcal{Z}$ diese Eigenschaft verliert. Keiner der in X enthaltenen Punkte muss also im affinen Ansatzraum liegen.

Der für den Aufbau eines Lagrange-Raums in Kapitel 4.1 verwendete Greedy-Algorithmus wird nun mit dem Resultat aus Lemma 4.2.2 zu einem neuen Algorithmus kombiniert um sukzessive einen affinen Ansatzraum aufzubauen. Sei dazu wieder ein Startindex $D_1 = \{p_1\} \subset \mathbb{R}$ und ein Startraum $\mathcal{Z}_1 := \text{span}\{\mathbf{x}_{p_1}\}$ gegeben. Dann ergibt sich der Vektor $\mathbf{b}_1$ für die bestmögliche affine Verschiebung von X_1 zu

$$\mathbf{b}_1 := \frac{1}{q} \sum_{j=1}^{q} (\mathbf{x}_j - \mathbf{P}_1\mathbf{x}_j),$$

wobei $\mathbf{P}_i$ den orthogonalen Projektor auf $\mathcal{Z}_i$ bezeichnet. Der Algorithmus arbeitet dann wie folgt:

for $j = 2 : m + 1$ **do**
  $p_j = \arg\max_{k \in \{1,\dots,q\}} \|\mathbf{x}_k - (\mathbf{P}_{j-1}(\mathbf{x}_k - \mathbf{b}_{j-1}) + \mathbf{b}_{j-1})\|$;
  $D_j = D_{j-1} \cup p_j$;
  $\mathcal{Z}_j = \mathcal{Z}_{j-1} + \text{span}\{\mathbf{x}_{p_j}\}$;
  $\mathbf{b}_j = \frac{1}{q} \sum_{i=1}^{q} (\mathbf{x}_i - \mathbf{P}_j\mathbf{x}_i)$;
end for.

Der Unterraum wird also sukzessive um den Span der Punkte aus X erweitert, die den größten Projektionsfehler bezüglich des affinen Raums aufweisen. Dieser wird stets mit der bestmöglichen affinen Verschiebung erzeugt, die ebenfalls in jedem Schritt für den entsprechenden Unterraum neu berechnet wird.

Der so erzeugte affine Unterraum ist eine Kombination des zum Aufbau des Lagrange-Raums aus Kapitel 4.1 verwendeten Algorithmus und dem POD-Raum aus Kapitel 4.2, da einerseits der schrittweise Aufbau mit Rücksicht auf den maximalen Projektionsfehler (wie beim Lagrange-Raum) erfolgt und andererseits durch die affine Verschiebung die Summe der Projektionsfehler (wie beim POD-Raum) minimiert wird. Aufgrund der Minimalitätseigenschaft (4.9) ist der summierte Projektionsfehler für diesen Raum sicherlich nicht kleiner als der eines affinen POD-Raumes, dafür entfällt die aufwändige Berechnung der Proper Orthogonal Decomposition von X.

Kapitel 5

Testräume für einparametrige nichtlineare Gleichungen

Nachdem im letzten Kapitel verschiedene globale Ansatzräume für die Reduktion einer Lösungskurve betrachtet wurden, werden nun mögliche Testräume untersucht. Dazu sei wieder eine offene Menge $\Omega \subset \mathbb{R}^{n+1}$ gegeben sowie die Funktion $\mathbf{F} \in C^1(\Omega, \mathbb{R}^n)$, für die das Problem

$$\mathbf{F}(\mathbf{x}) = \mathbf{0}$$

betrachtet wird. Es existiere ein offenes Intervall S und eine Lösungskurve $\mathbf{c} \in C^1(S, \mathcal{R}(\mathbf{F}))$, sodass für alle $s \in S$

$$\mathbf{F}(\mathbf{c}(s)) = \mathbf{0}$$

gilt. Die Art der Parametrisierung ist beliebig. Weiterhin sei eine Diskretisierung X von $\mathbf{c}$ gegeben, sowie ein daraus erzeugter globaler Ansatzraum $\mathcal{Z}$ der Dimension $m+1$, in dem eine Approximation der Lösungskurve $\mathbf{c}$ gesucht wird. Wir gehen im Folgenden der Übersichtlichkeit halber davon aus, dass ein linearer Ansatzraum verwendet wird. Alle Aussagen lassen sich aber problemlos auf affine Ansatzräume übertragen.

Ein global reduziertes Problem hat die Gestalt

$$\hat{\mathbf{F}}(\hat{\mathbf{x}}) = \mathbf{V}^T \mathbf{F}(\mathbf{Z}\hat{\mathbf{x}}) = \mathbf{0}, \tag{5.1}$$

wobei die Spalten der Matrizen $\mathbf{Z}$ und $\mathbf{V}$ jeweils eine Basis des Ansatzraumes $\mathcal{Z}$, bzw. des Testraumes $\mathcal{V}$ bilden.

Genau wie $\mathcal{Z}$ ist man an einem globalen $\mathcal{V}$ interessiert, das heißt für die Berechnung der Approximation der gesamten Kurve $\mathbf{c}$ über die RB-Methode soll ein fester Testraum $\mathcal{V}$ verwendet werden. Basierend auf numerischen Versuchen kann man davon ausgehen, dass nicht immer ein globaler Testraum $\mathcal{V}$ existiert, der eine sinnvolle Reduktion ermöglicht. Daher werden zunächst

Verfahren entwickelt, mit denen für einen globalen Ansatzraum $\mathcal{Z}$ Testräumen aufgebaut werden können, die eine sinnvolle lokale Reduktion in einem festen Punkt $\mathbf{x}_0 \in \mathcal{Z}$ ermöglichen.

In Kapitel 2.2.1 wurden Bedingungen angegeben, unter denen eine Lösungskurve in der Nähe eines Punktes $\mathbf{x}_0$ existiert (siehe Satz 2.2.3). Diese Aussagen werden nun auf das reduzierte Problem (5.1) und einen Punkt $\hat{\mathbf{x}}_0 \in \mathbb{R}^{m+1}$ angewendet. Damit eine Lösungskurve des reduzierten Problems in der Nähe von $\hat{\mathbf{x}}_0$ existiert, muss die Matrix $\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)$ vollen Zeilenrang haben und $\|\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)\|$ darf eine bestimmte (vom kleinsten Singulärwert von $\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)$ abhängige) Schranke nicht übersteigen. Zunächst wird $\|\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)\|$ genauer betrachtet.

Ergibt sich der Punkt $\hat{\mathbf{x}}_0$ aus der Projektion eines Punktes $\mathbf{x}_* \in X$ auf $\mathcal{Z}$, also über

$$\mathbf{Z}\hat{\mathbf{x}}_0 = \mathbf{P}\mathbf{x}_*,$$

wobei $\mathbf{P}$ den orthogonalen Projektor auf $\mathcal{Z}$ beschreibt, so erhält man für den Funktionswert $\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)$

$$\|\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)\| = \|\mathbf{V}^T\mathbf{F}(\mathbf{Z}\hat{\mathbf{x}}_0)\| = \|\mathbf{V}^T(\mathbf{F}(\mathbf{P}\mathbf{x}_*) - \mathbf{F}(\mathbf{x}_*))\| \leq C\|\mathbf{P}\mathbf{x}_* - \mathbf{x}_*\|.$$

Dabei ergibt sich die Konstante C zu $C := \|\mathbf{V}\| \sup_{x\in\Omega} \|\mathbf{DF}(\mathbf{x})\|$. Die Norm des Funktionswertes der Reduktion in $\hat{\mathbf{x}}_0$ steht also mit dem Projektionsfehler bezüglich $\mathcal{Z}$ in Verbindung und ist somit unabhängig von der Wahl von $\mathcal{V}$ (Bilden die Spalten von $\mathbf{V}$ eine Orthonormalbasis von $\mathcal{V}$ gilt sogar $\|\mathbf{V}\| = 1$). Dieser Projektionsfehler wird für die in Kapitel 4 angegebenen Ansatzräume mit steigender Dimension m beliebig klein.

Der Testraum $\mathcal{V}$ muss so aufgebaut werden, dass in einem Punkt $\hat{\mathbf{x}}_0$ mit $\mathbf{x}_0 = \mathbf{Z}\hat{\mathbf{x}}_0$

$$\text{Rang}(\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)) = \text{Rang}(\mathbf{V}^T\mathbf{DF}(\mathbf{x}_0)) = m \tag{5.2}$$

gilt. In Kapitel 3.1.1 wurde bereits gezeigt, wie $\mathcal{V}$ für den Fall, dass $\mathbf{T}(\mathbf{Z}\hat{\mathbf{x}}_0) \in \mathcal{Z}$ gilt, gewählt werden kann. Da der Tangentialvektor bei der Verwendung eines globalen Ansatzraumes nicht in $\mathcal{Z}$ liegen muss, werden in den nächsten beiden Kapiteln Verallgemeinerungen des Aufbaus (3.6) hergeleitet.

Des Weiteren wird der Zusammenhang zwischen den kleinsten Singulärwerten von $\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)$ und $\mathbf{DF}(\mathbf{x}_0)$ genauer beschrieben, da nach Satz 2.2.3 die Größe des Intervalls, für das die Existenz einer Lösungskurve lokal gesichert werden kann, direkt vom kleinsten Singulärwert der Jacobimatrix der betrachteten Funktion abhängt.

Zusätzlich wird die Stetigkeit der Testräume $\mathcal{V}$ bezüglich des Punktes $\mathbf{x}_0$ betrachtet, da diese für die in Kapitel 6 entwickelte Reduktion eine entscheidene Rolle spielt.

5.1 Aufbau eines Testraumes mittels POD

Um mittels POD einen geeigneten Testraum zu erzeugen, werden zunächst einige Hilflemmata benötigt.

Lemma 5.1.1 (Satz von Courant-Fischer)**.** *Seien $\lambda_1 \leq \cdots \leq \lambda_n$ die Eigenwerte der symmetrischen Matrix $\mathbf{A} \in \mathbb{R}^{n,n}$, dann gilt*

$$\lambda_i = \min_{\dim \mathcal{V}=i} \max_{\mathbf{x}\in\mathcal{V},\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}.$$

Beweis. Der Beweis nutzt das Rayleighsche Prinzip und findet sich zum Beispiel in [46]. □

Lemma 5.1.2. *Sei $\mathbf{A} \in \mathbb{R}^{n,n+1}$ und $\mathbf{Z} \in \mathbb{R}^{n+1,m+1}$ mit $\mathbf{Z}^T\mathbf{Z} = \mathbf{I}_{m+1}$ (mit $n > m$) und sei σ_n der kleinste Singulärwert von $\mathbf{A}$, dann gilt für den zweitkleinsten Singulärwert η_m von $\mathbf{AZ} \in \mathbb{R}^{n,m+1}$*

$$\eta_m \geq \sigma_n.$$

Beweis. Sei λ_2 der zweitkleinste Eigenwert von $\mathbf{A}^T\mathbf{A}$ (der kleinste ist $\lambda_1 = 0$), dann gilt $\sigma_n = \sqrt{\lambda_2}$. Seien nun $\mathbf{M} := \mathbf{Z}^T\mathbf{A}^T\mathbf{A}\mathbf{Z}$ und $\mathbf{M}\mathbf{w}_i = \mu_i\mathbf{w}_i$ die aufsteigend geordneten Eigenwerte und Eigenvektoren von $\mathbf{M}$. Des weiteren gelte $\mathcal{W} := \operatorname{span}\{\mathbf{w}_1, \mathbf{w}_2\}$ und $\mathcal{U} := \operatorname{span}\{\mathbf{Z}\mathbf{w}_1, \mathbf{Z}\mathbf{w}_2\}$. Man sucht nun eine Abschätzung für $\eta_m = \sqrt{\mu_2}$. Mit Hilfe von Lemma 5.1.1 ergibt sich

$$\begin{aligned}\lambda_2 &= \min_{\dim \mathcal{V}=2} \max_{\mathbf{x}\in\mathcal{V},\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \leq \max_{\mathbf{x}\in\mathcal{U}} \frac{\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \\ &= \max_{\mathbf{w}\in\mathcal{W},\mathbf{w}\neq\mathbf{0}} \frac{\mathbf{w}^T\mathbf{Z}^T\mathbf{A}^T\mathbf{A}\mathbf{Z}^T\mathbf{w}}{\mathbf{w}^T\mathbf{w}} = \max_{\mathbf{w}\in\mathcal{W},\mathbf{w}\neq\mathbf{0}} \frac{\mathbf{w}^T\mathbf{M}\mathbf{w}}{\mathbf{w}^T\mathbf{w}} = \mu_2.\end{aligned}$$

Somit gilt $\sigma_n \leq \eta_m$. □

Sei nun ein Punkt $\mathbf{x}_0 := \mathbf{Z}\hat{\mathbf{x}}_0 \in \Omega$ gegeben. Im Folgenden wird ein Verfahren hergeleitet, einen Testraum $\mathcal{V}$ mit Hilfe der POD der Matrix $\mathbf{DF}(\mathbf{x}_0)\mathbf{Z}$ aufzubauen, sodass die Matrix

$$\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}_0}) = \mathbf{V}^T\mathbf{DF}(\mathbf{x}_0)\mathbf{Z}$$

vollen Zeilenrang besitzt.

Satz 5.1.3. *Seien für $\mathbf{F} \in C^1(\Omega, \mathbb{R}^n)$ ein Ansatzraum $\mathcal{Z}$, sowie ein Punkt $\mathbf{x}_0 = \mathbf{Z}\hat{\mathbf{x}}_0 \in \mathcal{R}(\mathbf{F})$ gegeben, wobei die Spalten von $\mathbf{Z} \in \mathbb{R}^{n+1,m+1}$ eine Orthonormalbasis von $\mathcal{Z}$ bilden. Sei weiterhin die POD*

$$\mathbf{DF}(\mathbf{x}_0)\mathbf{Z} = \mathbf{U}_{m+1}\Sigma_{m+1}\mathbf{W}_{m+1}^T$$

nach (4.10) gegeben mit $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_{m+1})$. *Sei*

$$\mathcal{V} := span\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}, \text{ sowie } \mathbf{V} := (\mathbf{u}_1, \ldots, \mathbf{u}_m),$$

dann gilt für die Reduktion $\hat{\mathbf{F}}(\hat{\mathbf{x}}) = \mathbf{V}^T\mathbf{F}(\mathbf{Z}\hat{\mathbf{x}}_0)\mathbf{Z}$

$$Rang(\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)) = m.$$

Sei σ_n *der kleinste Singulärwert von* $\mathbf{DF}(\mathbf{x}_0)$, *dann gilt für den kleinsten Singulärwert* $\hat{\sigma}_m$ *von* $\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)$ *außerdem*

$$\hat{\sigma}_m^{-1} \leq \sigma_n^{-1}.$$

Beweis. Aus Lemma 5.1.2 folgt für die Matrix $\mathbf{DF}(\mathbf{x}_0)\mathbf{Z} \in \mathbb{R}^{n,m+1}$

$$\eta_m \geq \sigma_n, \tag{5.3}$$

wobei η_m der zweitkleinste Singulärwert von $\mathbf{DF}(\mathbf{x}_0)\mathbf{Z}$ und σ_n der kleinste Singulärwert von $\mathbf{DF}(\mathbf{x}_0)$ sind. Man betrachte nun die gegebene POD

$$\begin{aligned}\mathbf{DF}(\mathbf{x}_0)\mathbf{Z} &= \mathbf{U}_{m+1}\Sigma_{m+1}\mathbf{W}_{m+1}^T \\ &= (\mathbf{u}_1, \ldots, \mathbf{u}_{m+1}) \begin{pmatrix} \eta_1 & & \\ & \ddots & \\ & & \eta_{m+1} \end{pmatrix} (\mathbf{w}_1, \ldots, \mathbf{w}_{m+1})^T\end{aligned}$$

Mit $\mathbf{V} = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$ gilt für $\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)$

$$\begin{aligned}\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0) &= \mathbf{V}^T\mathbf{DF}(\mathbf{x}_0)\mathbf{Z} = (\mathbf{u}_1, \ldots, \mathbf{u}_m)^T\mathbf{U}_{m+1}\Sigma_{m+1}\mathbf{W}_{m+1}^T \\ &= (\mathbf{I}_m, \mathbf{0})\Sigma\mathbf{W}^T = \begin{pmatrix} \eta_1 & & 0 \\ & \ddots & \vdots \\ & \eta_m & 0 \end{pmatrix} \mathbf{W}^T.\end{aligned}$$

Dies ist die POD der Matrix $\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)$ und somit gilt für ihre Singulärwerte $\hat{\sigma}_i$:

$$\hat{\sigma}_i = \eta_i, \ i = 1, \ldots, m.$$

Insbesondere folgt aus (5.3)

$$\hat{\sigma}_m = \eta_m \geq \sigma_n.$$

Wegen $\mathbf{x}_0 \in \mathcal{R}(\mathbf{F})$ gilt $\text{Rang}(\mathbf{DF}(\mathbf{x}_0)) = n$ und damit $\sigma_n > 0$ und man erhält schließlich

$$\text{Rang}(\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)) = m, \text{ sowie } \hat{\sigma}_m^{-1} \leq \sigma_n^{-1}.$$

□

Bemerkung 5.1.4. *Für den Fall, dass* $\mathbf{T}(\mathbf{x}_0) \in \mathcal{Z}$ *gilt, ist der kleinste Singulärwert von* $\mathbf{DF}(\mathbf{x}_0)\mathbf{Z}$ *Null. Es existiert dann eine Matrix* $\mathbf{Z}^* = (\mathbf{z}_2, \ldots, \mathbf{z}_{m+1}) \in \mathbb{R}^{n+1,m}$ *und* $\mathbf{T}(\mathbf{x}_0) \perp R(\mathbf{Z}^*)$*, sodass*

$$\mathcal{Z} = span\{\mathbf{T}(\mathbf{x}_0), \mathbf{z}_2, \ldots, \mathbf{z}_{m+1}\}$$

gilt. Für einen wie in Satz 5.1.3 aufgebauten Testraum $\mathcal{V}$ *ergibt sich dann*

$$\mathcal{V} = span\{\mathbf{u}_1, \ldots, \mathbf{u}_m\} = R(\mathbf{DF}(\mathbf{x}_0)\mathbf{Z}^*).$$

Somit lässt sich $\mathbf{V}$ *als* $\mathbf{DF}(\mathbf{x}_0)\mathbf{Z}^*$ *schreiben, besitzt also eine Struktur wie in (3.6). Somit stellt das hier entwickelte Verfahren eine Verallgemeinerung der in Lemma 3.1.1 vorgestellten Methode dar.*

Bemerkung 5.1.5. *In Satz 2.2.3 wird gezeigt, dass der kleinste Singulärwert* σ_n *der Matrix* $\mathbf{DF}(\mathbf{x}_0)$ *die Größe des Intervalls bestimmt, für das eine Lösungskurve in der Nähe des Punktes* $\mathbf{x}_0$ *gesichert werden kann. Wird der Testraum wie in Satz 5.1.3 erzeugt und gilt* $\sigma_n^{-1} \leq c_0$*, so gilt für den kleinsten Singulärwert* $\hat{\sigma}_m$ *von* $\mathbf{D\hat{F}}(\hat{\mathbf{x}}_0)$ *aufgrund von Satz 5.1.3:*

$$\hat{\sigma}_m^{-1} \leq c_0.$$

Existiert also eine untere Schranke für den kleinsten Singulärwert von $\mathbf{DF}(\mathbf{x}_0)$*, gilt diese auch für den kleinsten Singulärwert der Jacobimatrix der Reduktion.*

Der hier aufgebaute Testraum hängt von dem Punkt $\mathbf{x}_0$, in dem die Jacobimatrix $\mathbf{DF}(\mathbf{x}_0)$ berechnet wird, ab und es zeigt sich, dass die Abbildung $\mathbf{x} \mapsto \mathcal{V}(\mathbf{x})$ für das hier beschriebene Verfahren unstetig ist (in dem Sinne, dass die Abbildung, die die orthogonale Projektion auf den Raum beschreibt, unstetig ist). Eine genauere Erklärung dieser Eigenschaft findet sich in Kapitel 5.3.

Im folgenden Abschnitt wird ein alternatives Verfahren zum Aufbau der Testräume $\mathcal{V}$ entwickelt, das ohne die POD auskommt und zusätzliche Stetigkeitskriterien erfüllt.

5.2 Aufbau eines Testraumes mittels Tangentialfeld

Im Folgenden wird ein zweites Verfahren entwickelt, einen geeigneten Testraum ohne die POD aufzubauen. Diese nutzt das Tangentialfeld, das in Kapitel 2.2.2 definiert wurde, um einen Testraum $\mathcal{V}$ (und eine damit verbundene Matrix $\mathbf{V}$) zu erzeugen, der für einen Punkt $\mathbf{x}_0 = \mathbf{Z}\hat{\mathbf{x}}_0 \in \mathcal{R}(\mathbf{F})$ die Bedingung (5.2) erfüllt.

Die Grundidee des im vorherigen Kapitel beschriebenen Aufbaus bestand darin, als Testraum das Bild von $\mathbf{DF}(\mathbf{x}_0)\mathbf{Z}$ zu verwenden, abzüglich der zum

kleinsten Singulärwert gehörenden Raumrichtung. Ist der Tangentialvektor $\mathbf{T}(\mathbf{x}_0)$ Teil des Ansatzraumes $\mathcal{Z}$, so ist dieser Singulärwert 0. Das Verfahren wird hier so interpretiert, dass der Anteil des Bildes von $\mathbf{DF}(\mathbf{x}_0)\mathbf{Z}$ vernachlässigt wird, der die Wirkung von $\mathbf{DF}(\mathbf{x}_0)$ auf die Richtung in $\mathcal{Z}$ darstellt, die dem Tangentialvektor in $\mathbf{x}_0$ am nächsten kommt.

Die in diesem Kapitel behandelte Alternative basiert auf einer ähnlichen Grundidee, wobei zunächst der orthogonal auf dem Tangentialraum stehender Unterraum $\mathcal{Q} \subset \mathcal{Z}$ betrachtet und $\mathcal{V}$ als das Bild von $\mathbf{DF}(\mathbf{x}_0)$ eingeschränkt auf $\mathcal{Q}$ erzeugt wird. Sei dazu

$$\mathcal{Q} := R(\mathbf{T}(\mathbf{x}_0))^\perp \cap \mathcal{Z}$$

und es sei zunächst festgehalten, dass $\mathcal{Q}$ genau dann die Dimension m hat, wenn

$$\mathbf{T}(\mathbf{x}_0) \notin \mathcal{Z}^\perp$$

gilt, das heißt der Tangentialvektor in $\mathbf{x}_0$ nicht orthogonal zum Ansatzraum steht. Diese Forderung ist sinnvoll, da der Ansatzraum $\mathcal{Z}$ als Approximation der diskretisierten Kurve $\mathbf{c}$ erzeugt wird und deren Bewegung im Raum möglichst gut widergespiegeln werden soll. Gilt $\mathbf{T}(\mathbf{x}_0) \in \mathcal{Z}^\perp$, würde $\mathcal{Z}$ diesen Zweck nicht erfüllen, da sich die zu reduzierende Lösungskurve im Punkt $\mathbf{x}_0$ senkrecht zum Ansatzraum bewegen würde.

Der Testraum $\mathcal{V}$ wird nun mittels

$$\mathcal{V} := \mathbf{DF}(\mathbf{x}_0)\mathcal{Q} = \{\mathbf{DF}(\mathbf{x}_0)\mathbf{q} \,|\mathbf{q} \in \mathcal{Q}\} \tag{5.4}$$

aufgebaut. Der folgende Satz zeigt, dass ein auf diese Weise aufgebauter Testraum die Bedingung (5.2) erfüllt. Dies bedeutet, dass die Jacobimatrix der Reduktion vollen Zeilenrang hat.

Satz 5.2.1. *Sei für $\mathbf{F} \in C^1(\Omega, \mathbb{R}^n)$ ein Ansatzraum $\mathcal{Z}$ der Dimension $m+1$ gegeben, sowie eine Matrix $\mathbf{Z} \in \mathbb{R}^{n+1,m+1}$, deren Spalten eine Basis von $\mathcal{Z}$ bilden, und es gelte für den Tangentialvektor in einem Punkt $\mathbf{x}_0 = \mathbf{Z}\hat{\mathbf{x}}_0 \in \mathcal{R}(\mathbf{F})$:*

$$\mathbf{T}(\mathbf{x}_0) \notin \mathcal{Z}^\perp.$$

Wählt man dann

$$\mathcal{Q} := R(\mathbf{T}(\mathbf{x}_0))^\perp \cap \mathcal{Z}, \text{ sowie}$$
$$\mathcal{V} := \mathbf{DF}(\mathbf{x}_0)\mathcal{Q} = \{\mathbf{DF}(\mathbf{x}_0)\mathbf{q} \,|\mathbf{q} \in \mathcal{Q}\},$$

und $\mathbf{V} \in \mathbb{R}^{n,m}$ als eine Matrix, deren Spalten eine Basis von $\mathcal{V}$ bilden, dann gilt für die Reduktion $\hat{\mathbf{F}}(\hat{\mathbf{x}}) = \mathbf{V}^T\mathbf{F}(\mathbf{Z}\mathbf{x})$

$$\mathit{Rang}(\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)) = m.$$

Beweis. Sei zunächst der Fall $\mathbf{T}(\mathbf{x}_0) \notin \mathcal{Z}$ betrachtet. Es gilt dann wegen $\text{Kern}(\mathbf{DF}(\mathbf{x}_0)) = R(\mathbf{T}(\mathbf{x}_0))$ und $\mathbf{T}(\mathbf{x}_0) \notin \mathcal{Q}$

$$\dim(\mathbf{DF}(\mathbf{x}_0)\mathcal{Z}) = m+1, \text{ sowie}$$
$$\dim(\mathcal{V}) = \dim(\mathbf{DF}(\mathbf{x}_0)\mathcal{Q}) = m,$$

wobei $\mathbf{DF}(\mathbf{x}_0)\mathcal{Z}$ den Unterraum $\{\mathbf{DF}(\mathbf{x}_0)\mathbf{z} : \ \mathbf{z} \in \mathcal{Z}\}$ bezeichnet. Da $\mathcal{V}$ ein Teilraum von $\mathbf{DF}(\mathbf{x}_0)\mathcal{Z}$ ist, ergibt sich

$$\dim(\mathcal{V}^\perp \cap \mathbf{DF}(\mathbf{x}_0)\mathcal{Z}) = 1 \tag{5.5}$$

Sei nun $\mathbf{w} \in \mathbb{R}^{m+1}$ ein Vektor aus $\text{Kern}(\mathbf{D\hat{F}}(\hat{\mathbf{x}}_0)) = \text{Kern}(\mathbf{V}^T\mathbf{DF}(\mathbf{x}_0)\mathbf{Z})$ und $\mathbf{u} := \mathbf{Zw}$, dann gilt wegen $\mathbf{u} \in \mathcal{Z}$ und

$$\mathbf{V}^T\mathbf{DF}(\mathbf{x}_0)\mathbf{u} = \mathbf{0}$$

für den Vektor $\mathbf{DF}(\mathbf{x}_0)\mathbf{u}$:

$$\mathbf{DF}(\mathbf{x}_0)\mathbf{u} \in \mathbf{DF}(\mathbf{x}_0)\mathcal{Z},$$
$$\mathbf{DF}(\mathbf{x}_0)\mathbf{u} \in \mathcal{V}^\perp.$$

Wegen (5.5) gilt also $\dim\{\mathbf{DF}(\mathbf{x}_0)\mathbf{Zw} : \ \mathbf{w} \in \text{Kern}(\mathbf{V}^T\mathbf{DF}(\mathbf{x}_0)\mathbf{Z})\} = 1$. Da $\mathbf{Z}$ vollen Spaltenrang besitzt und der eindimensionale Kern von $\mathbf{DF}(\mathbf{x}_0)$ nicht in $\mathcal{Z}$ enthalten ist, folgt daraus

$$\dim(\text{Kern}(\mathbf{V}^T\mathbf{DF}(\mathbf{x}_0)\mathbf{Z})) = \dim(\text{Kern}\mathbf{D\hat{F}}(\hat{\mathbf{x}}_0)) = 1$$

und der Satz ist bewiesen.

Für den Fall, dass $\mathbf{T}(\mathbf{x}_0) \in \mathcal{Z}$ gilt, der Tangentialvektor in $\mathbf{x}_0$ also Teil des Ansatzraumes ist, ergibt sich

$$\dim(\mathbf{DF}(\mathbf{x}_0)\mathcal{Z}) = m, \text{ sowie}$$
$$\dim(\mathcal{V}) = m.$$

Mit einer zum ersten Fall analogen Argumentation ergibt sich dann mit $\mathbf{V}^T\mathbf{DF}(\mathbf{x}_0)\mathbf{u} = \mathbf{0}$

$$\mathbf{DF}(\mathbf{x}_0)\mathbf{u} \in \mathcal{V}^\perp \cap \mathbf{DF}(\mathbf{x}_0)\mathcal{Z} = \{\mathbf{0}\}.$$

Für jeden Vektor $\mathbf{w} \in \text{Kern}(\mathbf{V}^T\mathbf{DF}(\mathbf{x}_0)\mathbf{Z})$ gilt also $\mathbf{DF}(\mathbf{x}_0)\mathbf{Zw} = \mathbf{0}$ und damit sofort $\mathbf{Zw} \in R(\mathbf{T}(\mathbf{x}_0))$. Der Kern von $\mathbf{D\hat{F}}(\hat{\mathbf{x}}_0)$ ist also wieder eindimensional und die Aussage somit bewiesen. □

Für die Reduktion $\hat{\mathbf{F}}(\hat{\mathbf{x}}) = \mathbf{V}^T\mathbf{F}(\mathbf{Z}\hat{\mathbf{x}})$ kann die Regularitätsmenge $\mathcal{R}(\hat{\mathbf{F}})$ und ein Tangentialfeld $\hat{\mathbf{T}}$ analog zu dem von $\mathbf{F}$ definiert werden.

Um wie im vorherigen Kapitel Aussagen über den kleinsten Singulärwert $\hat{\sigma}_m$ von $\mathbf{D\hat{F}}(\hat{\mathbf{x}}_0)$ treffen zu können, wird ein Lemma benötigt, das den Zusammenhang zwischen den Tangentialvektoren von $\mathbf{F}$ und $\hat{\mathbf{F}}$ im Punkt $\mathbf{x}_0 = \mathbf{Z}\hat{\mathbf{x}}_0$ herstellt.

Für dessen Beweis wird zunächst folgendes Hilfslemma gezeigt:

Lemma 5.2.2. *Sei* $\mathbf{A} \in \mathbb{R}^{n,n+1}$ *mit Rang*$(\mathbf{A}) = n$*, und sei* σ_n *der kleinste Singulärwert von* $\mathbf{A}$*, dann gilt*

$$\min_{\|\mathbf{x}\|=1, \mathbf{x} \perp Kern(\mathbf{A})} \|\mathbf{A}\mathbf{x}\| = \sigma_n.$$

Beweis. Sei $\mathbf{A} = \mathbf{U}\Sigma\mathbf{W}^T$ die POD von $\mathbf{A}$, dann gilt

$$R(\mathbf{A}^T) = R((\mathbf{w}_1, \ldots, \mathbf{w}_n)) = \mathrm{Kern}(\mathbf{A})^\perp,$$

wobei die $\mathbf{w}_j$ die paarweise orthonormalen ersten n Spalten von $\mathbf{W}$ bezeichnen. Somit lässt sich jeder Vektor $\mathbf{x} \in \mathrm{Kern}(\mathbf{A})^\perp$ als eine eindeutige Linearkombination

$$\mathbf{x} = \sum_{j=1}^{n} \alpha_j \mathbf{w}_j$$

schreiben. Bezeichnet $\alpha = (\alpha_1, \ldots, \alpha_n)^T \in \mathbb{R}^n$ nun den zugehörigen Koeffizientenvektor, ergibt sich mit den orthonormalen Vektoren $(\mathbf{u}_1, \ldots, \mathbf{u}_n) = \mathbf{U}$

$$\begin{aligned}\min_{\|\mathbf{x}\|=1, \mathbf{x} \perp \mathrm{Kern}(\mathbf{A})} \|\mathbf{A}\mathbf{x}\|^2 &= \min_{\|\alpha\|=1} \left\| \mathbf{A} \sum_{j=1}^{n} \alpha_j \mathbf{w}_j \right\|^2 = \min_{\|\alpha\|=1} \left\| \sum_{j=1}^{n} \alpha_j \sigma_j \mathbf{u}_j \right\|^2 \\ &= \min_{\|\alpha\|=1} \sum_{j=1}^{n} \alpha_j^2 \sigma_j^2 \|\mathbf{u}_j\|^2 \geq |\sigma_n|^2 \min_{\|\alpha\|=1} \sum_{j=1}^{n} \alpha_j^2 = \sigma_n^2.\end{aligned}$$

Des Weiteren gilt

$$\min_{\|\mathbf{x}\|=1, \mathbf{x} \perp \mathrm{Kern}(\mathbf{A})} \|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{A}\mathbf{w}_n\| = \|\sigma_n \mathbf{u}_n\| = \sigma_n,$$

womit das Lemma bewiesen ist. □

Lemma 5.2.3. *Seien für* $\mathbf{F} \in C^1(\Omega, \mathbb{R}^n)$ *der Ansatzraum* $\mathcal{Z}$*, sowie die Matrix* $\mathbf{Z}$ *wie in Satz (5.2.1) gegeben, wobei die Spalten von* $\mathbf{Z}$ *zusätzlich orthonormal sind. Für einen Punkt* $\mathbf{x}_0 = \mathbf{Z}\hat{\mathbf{x}}_0 \in \mathcal{R}(\mathbf{F})$ *sei der Testraum* $\mathcal{V}$ *wie in (5.4) über*

$$\begin{aligned}\mathcal{Q} &= R(\mathbf{T}(\mathbf{x}_0))^\perp \cap \mathcal{Z}, \\ \mathcal{V} &= \mathbf{DF}(\mathbf{x}_0)\mathcal{Q}\end{aligned} \tag{5.6}$$

aufgebaut. Sei weiterhin

$$q := \frac{\sigma_n}{\sigma_1}$$

der Quotient aus kleinstem und größtem Singulärwert von $\mathbf{DF}(\mathbf{x}_0)$ *und* $\hat{\mathbf{T}}(\hat{\mathbf{x}}_0)$ *der Tangentialvektor von* $\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0) = \mathbf{V}^T\mathbf{DF}(\mathbf{x}_0)\mathbf{Z}$ *im Punkt* $\hat{\mathbf{x}}_0$*, dann gilt*

$$|\langle \mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0), \mathbf{T}(\mathbf{x}_0) \rangle| \geq \sqrt{\frac{q^2}{1+q^2}} \|\mathbf{P}\mathbf{T}(\mathbf{x}_0)\|,$$

wobei $\mathbf{P}$ *den orthogonalen Projektor auf* $\mathcal{Z}$ *bezeichnet.*

Beweis. Zunächst sei festgehalten, dass wegen (5.6) für alle $\mathbf{q} \in \mathcal{Q} \subset \mathcal{Z}$

$$\langle \mathbf{PT}(\mathbf{x}_0), \mathbf{q} \rangle = \langle \mathbf{T}(\mathbf{x}_0), \mathbf{Pq} \rangle = \langle \mathbf{T}(\mathbf{x}_0), \mathbf{q} \rangle = 0$$

und somit $\mathbf{PT}(\mathbf{x}_0) \perp \mathcal{Q}$ gilt. Daher ist $\mathcal{Z}$ das Bild der Matrix $(\mathbf{PT}(\mathbf{x}_0), \mathbf{Q})$, wobei die Spalten von $\mathbf{Q}$ eine Basis von $\mathcal{Q}$ bilden. Die Reduktion $\hat{\mathbf{F}}$ lässt sich daher auch als Funktion $\mathbf{G}(\alpha, \hat{\mathbf{q}})$ mit

$$\hat{\mathbf{G}}(\alpha, \hat{\mathbf{q}}) = \mathbf{V}^T \mathbf{F}(\alpha \mathbf{PT}(\mathbf{x}_0) + \mathbf{Q}\hat{\mathbf{q}})$$

schreiben. Der Punkt $(\alpha_0, \hat{\mathbf{q}}_0)$ sei so gewählt, dass $\mathbf{x}_0 = \alpha_0 \mathbf{PT}(\mathbf{x}_0) + \mathbf{Q}\hat{\mathbf{q}}_0$ gilt. Sei nun der Vektor $\hat{\mathbf{T}}_* := (\alpha_*, \hat{\mathbf{q}}_*^T)^T$ gegeben mit

$$(\mathbf{PT}(\mathbf{x}_0), \mathbf{Q})\hat{\mathbf{T}}_* = \mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0).$$

Man beachte, dass der Vektor $\hat{\mathbf{T}}_*$ im Gegensatz zu $\hat{\mathbf{T}}(\hat{\mathbf{x}}_0)$ nicht normiert sein muss. Es gilt

$$\mathbf{D}\hat{\mathbf{G}}(\alpha_0, \hat{\mathbf{q}}_0)\hat{\mathbf{T}}_* = \mathbf{V}^T \mathbf{DF}(\mathbf{x}_0)(\mathbf{PT}(\mathbf{x}_0), \mathbf{Q})\hat{\mathbf{T}}_* = \mathbf{V}^T \mathbf{DF}(\mathbf{x}_0)\mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0) = \mathbf{0}$$

und somit

$$\mathbf{0} = \mathbf{D}\hat{\mathbf{G}}(\alpha_0, \hat{\mathbf{q}}_0)\hat{\mathbf{T}}_* = \alpha_* \mathbf{V}^T \mathbf{DF}(\mathbf{x}_0)\mathbf{PT}(\mathbf{x}_0) + \mathbf{V}^T \mathbf{DF}(\mathbf{x}_0)\mathbf{Q}\hat{\mathbf{q}}_*. \tag{5.7}$$

Zunächst werde der Fall $\mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0) \notin R(\mathbf{PT}(\mathbf{x}_0))$ betrachtet, beide Summanden sind dann ungleich $\mathbf{0}$. Multipliziert man die Gleichung (5.7) nun mit $\hat{\mathbf{q}}_*^T$ ergibt sich mit Hilfe der Cauchy-Schwarzschen Ungleichung

$$\begin{aligned} |\alpha_*| &= \left| \frac{\hat{\mathbf{q}}_*^T \mathbf{V}^T \mathbf{DF}(\mathbf{x}_0)\mathbf{Q}\hat{\mathbf{q}}_*}{\hat{\mathbf{q}}_*^T \mathbf{V}^T \mathbf{DF}(\mathbf{x}_0)\mathbf{PT}(\mathbf{x}_0)} \right| = \left| \frac{\hat{\mathbf{q}}_*^T \mathbf{Q}^T \mathbf{DF}(\mathbf{x}_0)^T \mathbf{DF}(\mathbf{x}_0)\mathbf{Q}\hat{\mathbf{q}}_*}{\hat{\mathbf{q}}_*^T \mathbf{Q}^T \mathbf{DF}(\mathbf{x}_0)\mathbf{DF}(\mathbf{x}_0)\mathbf{PT}(\mathbf{x}_0)} \right| \\ &\geq \frac{\|\mathbf{DF}(\mathbf{x}_0)\mathbf{Q}\hat{\mathbf{q}}_*\|^2}{\|\mathbf{DF}(\mathbf{x}_0)\mathbf{Q}\hat{\mathbf{q}}_*\| \|\mathbf{DF}(\mathbf{x}_0)\mathbf{PT}(\mathbf{x}_0)\|} = \frac{\|\mathbf{DF}(\mathbf{x}_0)\mathbf{Q}\hat{\mathbf{q}}_*\|}{\|\mathbf{DF}(\mathbf{x}_0)\mathbf{PT}(\mathbf{x}_0)\|}. \end{aligned} \tag{5.8}$$

Zähler und Nenner dieses Terms werden nun einzeln abgeschätzt.

Aus der Konstruktion von $\mathcal{Q}$ und $\mathrm{Kern}(\mathbf{DF}(\mathbf{x}_0)) = R(\mathbf{T}(\mathbf{x}_0))$ folgt, dass $\mathbf{Q}\hat{\mathbf{q}}_* \perp \mathrm{Kern}(\mathbf{DF}(\mathbf{x}_0))$ gilt. Somit ergibt sich zusammen mit Lemma 5.2.2 für den Zähler

$$\begin{aligned} \|\mathbf{DF}(\mathbf{x}_0)\mathbf{Q}\hat{\mathbf{q}}_*\| &= \|\mathbf{Q}\hat{\mathbf{q}}_*\| \left\| \mathbf{DF}(\mathbf{x}_0) \frac{\mathbf{Q}\hat{\mathbf{q}}_*}{\|\mathbf{Q}\hat{\mathbf{q}}_*\|} \right\| \\ &\geq \|\hat{\mathbf{Q}}\mathbf{q}_*\| \min_{\mathbf{q} \in \mathcal{Q}, \|\mathbf{q}\|=1} \|\mathbf{DF}(\mathbf{x}_0)\mathbf{q}\| \\ &\geq \|\mathbf{Q}\hat{\mathbf{q}}_*\| \min_{\mathbf{x} \perp \mathrm{Kern}(\mathbf{DF}(\mathbf{x}_0)), \|\mathbf{x}\|=1} \|\mathbf{DF}(\mathbf{x}_0)\mathbf{x}\| = \|\mathbf{Q}\hat{\mathbf{q}}_*\| \sigma_n. \end{aligned}$$

Für den Nenner erhält man

$$\|\mathbf{DF}(\mathbf{x}_0)\mathbf{PT}(\mathbf{x}_0)\| \leq \|\mathbf{DF}(\mathbf{x}_0)\|\|\mathbf{PT}(\mathbf{x}_0)\| = \sigma_1\|\mathbf{PT}(\mathbf{x}_0)\|.$$

Mit $q = \sigma_n/\sigma_1$ ergibt sich dann für (5.8)

$$|\alpha_*| \geq q\frac{\|\mathbf{Q}\hat{\mathbf{q}}_*\|}{\|\mathbf{PT}(\mathbf{x}_0)\|}. \tag{5.9}$$

Da $\hat{\mathbf{T}}(\hat{\mathbf{x}}_0)$ normiert ist, ergibt sich wegen der Orthonormalität der Spalten von $\mathbf{Z}$

$$1 = \|\mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0)\|^2 = \|(\mathbf{PT}(\mathbf{x}_0), \mathbf{Q})\hat{\mathbf{T}}_*\|^2 = \alpha_*^2\|\mathbf{PT}(\mathbf{x}_0)\|^2 + \|\mathbf{Q}\hat{\mathbf{q}}_*\|^2$$

und somit $\|\mathbf{Q}\hat{\mathbf{q}}_*\|^2 = 1 - \|\mathbf{PT}(\mathbf{x}_0)\|^2\alpha_*^2$. Eingesetzt in (5.9) führt dies zu

$$|\alpha_*|^2 \geq q^2\frac{1 - \|\mathbf{PT}(\mathbf{x}_0)\|^2\alpha_*^2}{\|\mathbf{PT}(\mathbf{x}_0)\|^2}$$

und schließlich zu

$$|\alpha_*| \geq \sqrt{\frac{q^2}{1+q^2}}\|\mathbf{PT}(\mathbf{x}_0)\|^{-1}.$$

Für den Wert $|\langle\mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0), \mathbf{T}(\mathbf{x}_0)\rangle|$ erhält man jetzt

$$\begin{aligned}
|\langle\mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0), \mathbf{T}(\mathbf{x}_0)\rangle| &= |\langle(\mathbf{PT}(\mathbf{x}_0), \mathbf{Q})\hat{\mathbf{T}}_*, \mathbf{T}(\mathbf{x}_0)\rangle| \\
&= |\langle\alpha_*\mathbf{PT}(\mathbf{x}_0), \mathbf{T}(\mathbf{x}_0)\rangle + \underbrace{\langle\mathbf{Q}\hat{\mathbf{q}}_*, \mathbf{T}(\mathbf{x}_0)\rangle}_{=0}| \\
&= |\alpha_*||\langle\mathbf{PT}(\mathbf{x}_0), \mathbf{T}(\mathbf{x}_0)\rangle| \\
&\geq \sqrt{\frac{q^2}{1+q^2}}\|\mathbf{PT}(\mathbf{x}_0)\|^{-1}|\langle\mathbf{PT}(\mathbf{x}_0), \mathbf{T}(\mathbf{x}_0)\rangle|
\end{aligned} \tag{5.10}$$

Das letzte Skalarprodukt lässt sich wie folgt umschreiben:

$$\begin{aligned}
\langle\mathbf{PT}(\mathbf{x}_0), \mathbf{T}(\mathbf{x}_0)\rangle &= \mathbf{T}(\mathbf{x}_0)^T\mathbf{Z}\mathbf{Z}^T\mathbf{T}(\mathbf{x}_0) = \|\mathbf{Z}^T\mathbf{T}(\mathbf{x}_0)\|^2 = \|\mathbf{Z}\mathbf{Z}^T\mathbf{T}(\mathbf{x}_0)\|^2 \\
&= \|\mathbf{PT}(\mathbf{x}_0)\|^2.
\end{aligned}$$

Setzt man dies in (5.10) ein erhält man

$$|\langle\mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0), \mathbf{T}(\mathbf{x}_0)\rangle| \geq \sqrt{\frac{q^2}{1+q^2}}\|\mathbf{PT}(\mathbf{x}_0)\| \tag{5.11}$$

und der Satz ist für den Fall, dass $\mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0) \notin R(\mathbf{PT}(\mathbf{x}_0))$ gilt, bewiesen. Ist dies nicht der Fall, gilt nach Gleichung (5.7) $\alpha_* = \|\mathbf{PT}(\mathbf{x}_0)\|^{-1}$ und die Abschätzung (5.11) wird zu

$$|\langle \mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0), \mathbf{T}(\mathbf{x}_0)\rangle| = \|\mathbf{PT}(\mathbf{x}_0)\|.$$

Da $\sqrt{q^2/(1+q^2)} \leq 1$ gilt, ist die Ungleichung (5.10) für diesen Fall also ebenfalls erfüllt. □

Mit Hilfe dieses Lemmas lässt sich nun im folgenden Satz eine direkte Beziehung zwischen dem kleinsten Singulärwert σ_n von $\mathbf{DF}(\mathbf{x}_0)$ und $\hat{\sigma}_m$ von $\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)$ herleiten. Nach Satz 2.2.3 ist der kleinste Singulärwert der Jacobimatrix einer nichtlinearen Funktion ausschlaggebend für die Größe des Gebietes für das eine Lösung von $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ gesichert werden kann. Existiert nun eine untere Schranke für diesen Singulärwert, lässt sich basierend auf dieser ebenfalls eine untere Schranke für den kleinsten Singulärwert der Jacobimatrix der Reduktion angeben.

Satz 5.2.4. *Sei für $\mathbf{F} \in C^1(\Omega, \mathbb{R}^n)$ ein Ansatzraum $\mathcal{Z}$ der Dimension $m+1$ gegeben, sowie eine Matrix $\mathbf{Z} \in \mathbb{R}^{n+1,m+1}$, deren Spalten eine Orthonormalbasis von $\mathcal{Z}$ bilden. Für den Tangentialvektor gelte in einem Punkt $\mathbf{x}_0 = \mathbf{Z}\hat{\mathbf{x}}_0 \in \mathcal{R}(\mathbf{F})$*

$$\mathbf{T}(\mathbf{x}_0) \notin \mathcal{Z}^{\perp}.$$

Mit diesen Vorgaben sei nun

$$\mathcal{Q} := R(\mathbf{T}(\mathbf{x}_0))^{\perp} \cap \mathcal{Z},$$
$$\mathcal{V} := \mathbf{DF}(\mathbf{x}_0)\mathcal{Q} = \{\mathbf{DF}(\mathbf{x}_0)\mathbf{q} \mid \mathbf{q} \in \mathcal{Q}\},$$

sowie $\mathbf{V} := \mathbf{DF}(\mathbf{x}_0)\mathbf{Q}$, wobei die Spalten von $\mathbf{Q}$ eine Orthonormalbasis von $\mathcal{Q}$ bilden. Werden mit $\sigma_1 \geq \cdots \geq \sigma_n$ und $\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_m$ die Singulärwerte von $\mathbf{DF}(\mathbf{x}_0)$ bzw. $\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)$ bezeichnet, so gilt mit $q := \sigma_n/\sigma_1$ und $\tau = \sqrt{q^2/(1+q^2)}$ die Abschätzung

$$\hat{\sigma}_m^{-1} \leq \sigma_n^{-2}\left(1 + \frac{1}{\tau^2\|\mathbf{PT}(\mathbf{x}_0)\|^2}\right)^{\frac{1}{2}}.$$

Hierbei werde mit $\mathbf{P}$ der orthogonale Projektor auf $\mathcal{Z}$ bezeichnet.

Beweis. Nach Satz 5.2.1 gilt $\mathrm{Rang}(\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)) = m$ und somit $\mathrm{Kern}(\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)) = R(\hat{\mathbf{T}}(\hat{\mathbf{x}}_0))$. Sei $\mathbf{w} = \mathbf{Z}\hat{\mathbf{w}}$ ein Vektor aus $\mathcal{Z}$ dann gilt für den Tangentialvektor $\hat{\mathbf{T}}(\hat{\mathbf{x}}_0)$

$$\langle \hat{\mathbf{w}}, \hat{\mathbf{T}}(\hat{\mathbf{x}}_0)\rangle = \hat{\mathbf{w}}^T\hat{\mathbf{T}}(\hat{\mathbf{x}}_0) = \hat{\mathbf{w}}^T\mathbf{Z}^T\mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0) = \langle \mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0), \mathbf{w}\rangle.$$

Mit Lemma 5.2.2 ergibt sich so für den kleinsten Singulärwert $\hat{\sigma}_m$ von $\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)$

$$\begin{aligned}\hat{\sigma}_m &= \min_{\hat{\mathbf{w}}\perp\hat{\mathbf{T}}(\hat{\mathbf{x}}_0),\|\hat{\mathbf{w}}\|=1} \|\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)\hat{\mathbf{w}}\| = \min_{\hat{\mathbf{w}}\perp\hat{\mathbf{T}}(\hat{\mathbf{x}}_0),\|\hat{\mathbf{w}}\|=1} \|\mathbf{V}^T\mathbf{D}\mathbf{F}(\mathbf{x}_0)\mathbf{Z}\hat{\mathbf{w}}\| \\ &= \min_{\mathbf{w}\perp\mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0),\|\mathbf{w}\|=1} \|\mathbf{V}^T\mathbf{D}\mathbf{F}(\mathbf{x}_0)\mathbf{w}\|.\end{aligned}$$

Sei nun ein Vektor $\mathbf{u} \in \mathcal{Z}$ mit $\|\mathbf{u}\| = 1$ und $\langle \mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0), \mathbf{u}\rangle = \mathbf{0}$ gegeben. Es existieren α und $\mathbf{q}$ mit $\mathbf{u} = \alpha\mathbf{T}(\mathbf{x}_0) + \mathbf{Q}\mathbf{q}$, wobei die Spalten von $\mathbf{Q}$ eine Basis von $\mathcal{Q}$ bilden. Aus $\mathbf{u} \perp \mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0)$ folgt

$$\begin{aligned}0 = \langle \mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0), \mathbf{u}\rangle &= \alpha\langle \mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0), \mathbf{T}(\mathbf{x}_0)\rangle + \langle \mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0), \mathbf{Q}\mathbf{q}\rangle \\ \Rightarrow \langle \mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0), \mathbf{Q}\mathbf{q}\rangle &= -\alpha\langle \mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0), \mathbf{T}(\mathbf{x}_0)\rangle \\ \Rightarrow \alpha &= -\frac{\langle \mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0), \mathbf{Q}\mathbf{q}\rangle}{\langle \mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0), \mathbf{T}(\mathbf{x}_0)\rangle}\end{aligned}$$

Aus Lemma 5.2.3 und der Cauchy-Schwarzschen Ungleichung folgt nun

$$|\alpha| = \frac{|\langle \mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0), \mathbf{Q}\mathbf{q}\rangle|}{|\langle \mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0), \mathbf{T}(\mathbf{x}_0)\rangle|} \leq \frac{\|\mathbf{Q}\mathbf{q}\|}{\tau\|\mathbf{P}\mathbf{T}(\mathbf{x}_0)\|}.$$

Des Weiteren gilt $1 = \|\mathbf{u}\|^2 = \|\alpha\mathbf{T}(\mathbf{x}_0) + \mathbf{Q}\mathbf{q}\|^2 = \alpha^2 + \|\mathbf{Q}\mathbf{q}\|^2$. Daraus ergibt sich

$$\begin{aligned}\|\mathbf{Q}\mathbf{q}\|^2 = 1 - \alpha^2 &\geq 1 - \frac{\|\mathbf{Q}\mathbf{q}\|^2}{\tau^2\|\mathbf{P}\mathbf{T}(\mathbf{x}_0)\|^2}, \\ \Rightarrow \|\mathbf{Q}\mathbf{q}\|^2 &\geq \frac{\tau^2\|\mathbf{P}\mathbf{T}(\mathbf{x}_0)\|^2}{\tau^2\|\mathbf{P}\mathbf{T}(\mathbf{x}_0)\|^2 + 1}. \qquad (5.12)\end{aligned}$$

Für den Ausdruck $\mathbf{V}^T\mathbf{D}\mathbf{F}(\mathbf{x}_0)\mathbf{u}$ erhält man

$$\begin{aligned}\|\mathbf{V}^T\mathbf{D}\mathbf{F}(\mathbf{x}_0)\mathbf{u}\| &= \|\mathbf{V}^T\mathbf{D}\mathbf{F}(\mathbf{x}_0)\mathbf{Q}\mathbf{q}\| \\ &= \|\mathbf{Q}^T\mathbf{D}\mathbf{F}(\mathbf{x}_0)^T\mathbf{D}\mathbf{F}(\mathbf{x}_0)\mathbf{Q}\mathbf{q}\| \qquad (5.13) \\ &\geq \|\mathbf{q}^T\mathbf{Q}^T\mathbf{D}\mathbf{F}(\mathbf{x}_0)^T\mathbf{D}\mathbf{F}(\mathbf{x}_0)\mathbf{Q}\mathbf{q}\|\|\mathbf{q}\|^{-1} \\ &= \|\mathbf{D}\mathbf{F}(\mathbf{x}_0)\mathbf{Q}\mathbf{q}\|^2\|\mathbf{q}\|^{-1} \\ &\geq \min_{\mathbf{w}\perp\mathbf{T}(\mathbf{x}_0),\|\mathbf{w}\|=1} \|\mathbf{D}\mathbf{F}(\mathbf{x}_0)\mathbf{w}\|^2\|\mathbf{Q}\mathbf{q}\|^2\|\mathbf{q}\|^{-1} \\ &= \sigma_n^2\|\mathbf{Q}\mathbf{q}\|^2\|\mathbf{q}\|^{-1}\end{aligned}$$

Da die Spalten von $\mathbf{Q}$ orthonormal sind, ergibt sich $\|\mathbf{q}\| = \|\mathbf{Q}\mathbf{q}\|$ und somit zusammen mit Lemma 5.2.2

$$\|\mathbf{V}^T\mathbf{D}\mathbf{F}(\mathbf{x}_0)\mathbf{u}\| \geq \sigma_n^2\|\mathbf{q}\| \geq \sigma_n^2\left(\frac{\tau^2\|\mathbf{P}\mathbf{T}(\mathbf{x}_0)\|^2}{\tau^2\|\mathbf{P}\mathbf{T}(\mathbf{x}_0)\|^2 + 1}\right)^{\frac{1}{2}}.$$

Dies führt für $\hat{\sigma}_m$ schließlich zu

$$\begin{aligned}\hat{\sigma}_m^{-1} &= \left(\min_{\hat{\mathbf{u}}\perp\hat{\mathbf{T}}(\hat{\mathbf{x}}_0),\|\hat{\mathbf{u}}\|=1} \|\mathbf{D}\hat{\mathbf{F}}(\hat{\mathbf{x}}_0)\hat{\mathbf{u}}\|\right)^{-1} \\ &= \left(\min_{\mathbf{u}\perp\mathbf{Z}\hat{\mathbf{T}}(\hat{\mathbf{x}}_0),\|\mathbf{u}\|=1} \|\mathbf{V}^T\mathbf{DF}(\mathbf{x}_0)\mathbf{u}\|\right)^{-1} \\ &\leq \sigma_n^{-2}\left(\frac{\tau^2\|\mathbf{PT}(\mathbf{x}_0)\|^2+1}{\tau^2\|\mathbf{PT}(\mathbf{x}_0)\|^2}\right)^{\frac{1}{2}} = \sigma_n^{-2}\left(1+\frac{1}{\tau^2\|\mathbf{PT}(\mathbf{x}_0)\|^2}\right)^{\frac{1}{2}}\end{aligned}$$

□

Bemerkung 5.2.5. *Der vorherige Satz setzt voraus, dass die Spalten der Matrix* $\mathbf{Q}$*, die zum Aufbau der* $\mathbf{V}$ *verwendet wird, eine Orthonormalbasis von* $\mathcal{Q}$ *bilden. Die Matrix* $\mathbf{V} = \mathbf{DF}(\mathbf{x}_0)\mathbf{Q}$ *besitzt daher selbst normalerweise keine orthonormalen Spalten. Ist man (zum Beispiel aus Gründen numerischer Stabilität) an einem* $\mathbf{V}$ *interessiert, das diese Eigenschaft besitzt, benötigt man eine neue Basis* $\tilde{\mathbf{q}}_1,\ldots,\tilde{\mathbf{q}}_m$ *des Raumes* $\mathcal{Q}$*, sodass mit* $\tilde{\mathbf{Q}} = (\tilde{\mathbf{q}}_1,\ldots,\tilde{\mathbf{q}}_m)$

$$\tilde{\mathbf{Q}}^T\mathbf{DF}(\mathbf{x}_0)^T\mathbf{DF}(\mathbf{x}_0)\tilde{\mathbf{Q}} = \mathbf{I}_m \tag{5.14}$$

gilt. Eine solche Basis existiert immer, da für die Matrix $\mathbf{DF}(\mathbf{x}_0)\mathbf{Q}$ *unendlich viele reguläre Matrizen* $\mathbf{R}$ *existieren, sodass die Matrix* $\mathbf{DF}(\mathbf{x}_0)\mathbf{QR}$ *orthonormale Spalten besitzt. Es gilt dann*

$$\mathbf{DF}(\mathbf{x}_0)\mathbf{QR} = \mathbf{V}$$

mit $\mathbf{V}^T\mathbf{V} = \mathbf{I}_m$*. Setzt man dann* $\tilde{\mathbf{Q}} = \mathbf{QR}$*, so gilt* $R(\tilde{\mathbf{Q}}) = R(\mathbf{QR}) = R(\mathbf{Q})$*, sodass die Spalten von* $\tilde{\mathbf{Q}}$ *also wieder eine Basis von* $\mathcal{Q}$ *bilden. Mit* $\mathbf{u} = \alpha\mathbf{T}(\mathbf{x}_0) + \tilde{\mathbf{Q}}\tilde{\mathbf{q}}$ *wird die Gleichung (5.13) dann zu*

$$\begin{aligned}\|\mathbf{V}^T\mathbf{DF}(\mathbf{x}_0)\mathbf{u}\| &= \sigma_n^2\|\tilde{\mathbf{Q}}\tilde{\mathbf{q}}\|^2\|\tilde{\mathbf{q}}\|^{-1} = \sigma_n^2\|\tilde{\mathbf{Q}}\tilde{\mathbf{q}}\|\|\mathbf{QR}\tilde{\mathbf{q}}\|\|\tilde{\mathbf{q}}\|^{-1} \\ &= \sigma_n^2\|\tilde{\mathbf{Q}}\tilde{\mathbf{q}}\|\|\mathbf{R}\tilde{\mathbf{q}}\|\|\mathbf{R}^{-1}\mathbf{R}\tilde{\mathbf{q}}\|^{-1} \\ &\geq \sigma_n^2\|\tilde{\mathbf{Q}}\tilde{\mathbf{q}}\|\|\mathbf{R}\tilde{\mathbf{q}}\|\|\mathbf{R}^{-1}\|^{-1}\|\mathbf{R}\tilde{\mathbf{q}}\|^{-1} \\ &= \sigma_n^2\|\tilde{\mathbf{Q}}\tilde{\mathbf{q}}\|\|\mathbf{R}\|\end{aligned}$$

und es ergibt sich

$$\|\mathbf{V}^T\mathbf{DF}(\mathbf{x}_0)\mathbf{u}\| \geq \sigma_n^2\|\mathbf{R}\|\left(\frac{\tau^2\|\mathbf{PT}(\mathbf{x}_0)\|^2}{\tau^2\|\mathbf{PT}(\mathbf{x}_0)\|^2+1}\right)^{\frac{1}{2}}.$$

Der Wert $\|\mathbf{R}\|$ *lässt sich wegen*

$$1 = \|\mathbf{V}\| = \|\mathbf{DF}(\mathbf{x}_0)\mathbf{QR}\| \leq \|\mathbf{DF}(\mathbf{x}_0)\|\|\mathbf{Q}\|\|\mathbf{R}\| = \sigma_1\|\mathbf{R}\|$$

über $\|\mathbf{R}\| \geq \sigma_1^{-1}$ *abschätzen, und man erhält schließlich für den Fall, dass* $\mathbf{V} = \mathbf{DF}(\mathbf{x}_0)\tilde{\mathbf{Q}}$ *orthonormale Spalten besitzt, die Abschätzung*

$$\hat{\sigma}_m^{-1} \leq \sigma_1\sigma_n^{-2}\left(1+\frac{1}{\tau^2\|\mathbf{PT}(\mathbf{x}_0)\|^2}\right)^{\frac{1}{2}}.$$

5.3 Aufbau Lipschitz-stetiger Basen

Bisher wurden in diesem Kapitel zwei Verfahren zum Aufbau eines Testraumes $\mathcal{V}$ bezüglich eines Punktes $\mathbf{x}_0 = \mathbf{Z}\hat{\mathbf{x}}_0 \in \mathcal{R}(\mathbf{F})$ hergeleitet, die garantieren, dass die Reduktion $\hat{\mathbf{F}}(\hat{\mathbf{x}}) = \mathbf{V}^T\mathbf{F}(\mathbf{Z}\hat{\mathbf{x}})$ eine Lösungskurve in der Nähe von $\hat{\mathbf{x}}_0$ besitzt. Die Spalten der Matrix $\mathbf{V}$ bilden dabei eine Basis des Testraumes $\mathcal{V}$. Da sowohl der Testraum, als auch die ihn repräsentierende Matrix vom Punkt $\mathbf{x}_0$ abhängig sind, ist über

$$\mathbf{V} : \begin{cases} \mathcal{Z} \cap \mathcal{R}(\mathbf{F}) & \to \mathbb{R}^{n,m} \\ \mathbf{x} & \mapsto \mathbf{V}(\mathbf{x}) \end{cases}$$

eine Abbildung definiert, wobei $\mathbf{V}(\mathbf{x})$ eine nach einem der Verfahren aus Kapitel 5.1 und 5.2 aufgebaute Matrix ist. Für die in Kapitel 6 entwickelte Reduktion werden zusätzliche Stetigkeitsanforderungen an die Abbildung $\mathbf{V}$ gestellt. Daher liegt das Interesse in diesem Kapitel auf der Frage, wie die Matrizen $\mathbf{V}$ aufgebaut werden können, um in einer Umgebung $U(\mathbf{x}_0)$ eines Punktes $\mathbf{x}_0$ die Existenz einer Konstanten $L_{\mathbf{V}} > 0$ zu garantieren, sodass

$$\|\mathbf{V}(\mathbf{x}) - \mathbf{V}(\mathbf{x}_0)\| \leq L_{\mathbf{V}}\|\mathbf{x} - \mathbf{x}_0\| \tag{5.15}$$

für alle $\mathbf{x} \in U(\mathbf{x}_0)$ gilt.

Zunächst wird der mittels der POD aufgebaute Testraum aus Kapitel 5.1 betrachtet. Es wird also die POD der Matrix $\mathbf{DF}(\mathbf{x}_0)\mathbf{Z}$ berechnet, wobei die Spalten von $\mathbf{Z}$ eine Basis des Testraumes $\mathcal{Z}$ bilden. Es ergibt sich

$$\begin{aligned} \mathbf{DF}(\mathbf{x}_0)\mathbf{Z} &= \mathbf{U}_{m+1}\Sigma_{m+1}\mathbf{W}_{m+1}^T \\ &= (\mathbf{u}_1, \ldots, \mathbf{u}_{m+1}) \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_{m+1} \end{pmatrix} (\mathbf{w}_1, \ldots, \mathbf{w}_m)^T. \end{aligned}$$

Der Testraum $\mathcal{V}$ wird dann mittels $\mathcal{V} := \operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ erzeugt, es werden also die ersten m Spalten der Matrix $\mathbf{U}_{m+1}$ verwendet, wobei die Reihenfolge durch die Konvention $\sigma_1 \geq \cdots \geq \sigma_{m+1}$ festgelegt ist. Um eine Matrix $\mathbf{V}$, deren Spalten eine Basis dieses Testraumes bilden sollen, zu erhalten, ist es naheliegend $\mathbf{V} := (\mathbf{u}_1, \ldots, \mathbf{u}_m)$ zu verwenden. Will man die Abhängigkeit dieser Matrix von dem Punkt $\mathbf{x}_0$ untersuchen, führt dies zu der Frage, auf welche Weise die Singulärwerte- und vektoren einer ortsabhängigen Matrix von $\mathbf{x}$ abhängen. Die Vektoren $\mathbf{u}_1, \ldots, \mathbf{u}_{m+1}$ sind die Eigenvektoren der Matrix $\mathbf{DF}(\mathbf{x}_0)\mathbf{Z}\mathbf{Z}^T\mathbf{DF}(\mathbf{x}_0)$, sodass

$$\mathbf{DF}(\mathbf{x}_0)\mathbf{Z}\mathbf{Z}^T\mathbf{DF}(\mathbf{x}_0)^T\mathbf{u}_j = \sigma_j^2\mathbf{u}_j, \quad j = 1, \ldots, m+1$$

gilt. Die übrigen Eigenwerte dieser Matrix sind 0. Die Frage nach der Abhängigkeit der $\mathbf{u}_j$ von $\mathbf{x}_0$ ist also verknüpft mit folgendem Problem:

Seien $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^{n,n}$ eine von $\mathbf{x} \in \mathbb{R}^n$ abhängige Matrix und $\lambda : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{v} : \mathbb{R}^n \to \mathbb{R}^n$ ihre Eigenwerte und -vektoren. Welcher Zusammenhang besteht zwischen den Eigenschaften der Abbildung $\mathbf{A}$ und denen der Abbildungen λ und $\mathbf{v}$? Diese Problemstellung wurde zum Beispiel in [11, 29] behandelt und es zeigte sich (vereinfacht gesagt), dass sich die Stetigkeit (bzw. Differenzierbarkeit) von $\mathbf{A}$ bezüglich ihrer Parameter auf die Eigenwerte und -vektoren übertragen lässt, falls keine doppelten Eigenwerte auftreten.

Solche doppelten Eigenwerte sorgen auch beim Aufbau der Testräume mittels POD für Unstetigkeiten, wie an dem folgenden kurzen Beispiel erläutert sei: Sei dazu die parameterabhängige Matrix

$$\mathbf{A}(t) = \begin{pmatrix} 3-t & 0 \\ 0 & t \end{pmatrix}$$

für das Intervall $t \in [0, 2]$ betrachtet. Diese ist in diesem Intervall positiv definit und besitzt die Eigenwerte $(3-t)$ und t, sowie die Eigenvektoren $(1, 0)^T$ und $(0, 1)^T$. Wählt man nun wie beim Aufbau des Testraumes alle Eigenvektoren, bis auf den zum kleinsten Eigenwert gehörenden, wären das für $t \in [0, 1)$ der Vektor $(1, 0)^T$ und für $t \in (1, 2]$ der Vektor $(0, 1)^T$. An der Stelle $t = 1$ befindet sich also eine Unstetigkeit.

Solche Unstetigkeiten können nur dann ausgeschlossen werden, wenn sich garantieren lässt, dass die Matrix $\mathbf{DF}(\mathbf{x})\mathbf{Z}$ keine doppelten Singulärwerte besitzt. Dies stellt jedoch eine zu starke Einschränkung an die betrachteten Funktionen $\mathbf{F}$, bzw. den Ansatzraum $\mathcal{Z}$ dar.

Aus diesen Gründen konnte für den mittels POD erzeugten Testraum kein Algorithmus zum Aufbau von Matrizen $\mathbf{V}$ gefunden werden, sodass diese die Bedingung (5.15) erfüllen.

Bei der zweiten in dieser Arbeit entwickelten Variante aus Kapitel 5.2 wird der Testraum $\mathcal{V}$ über

$$\begin{aligned} \mathcal{Q} &:= R(\mathbf{T}(\mathbf{x}_0))^\perp \cap \mathcal{Z}, \\ \mathcal{V} &:= \mathbf{DF}(\mathbf{x}_0)\mathcal{Q} \end{aligned}$$

erzeugt.

Bei dem Aufbau der Matrix $\mathbf{V}$, die eine Basis von $\mathcal{V}$ enthält, wurden Resultate für den Fall, dass $\mathbf{DF}(\mathbf{x}_0)$ auf eine Orthonormalbasis von $\mathcal{Q}$ angewendet wird, hergeleitet. Daher werden im Folgenden zunächst Verfahren entwickelt, mit denen für alle Punkte $\mathbf{x} \in U(\mathbf{x}_0)$ und den dazugehörigen Raum

$$\mathcal{Q}(\mathbf{x}) := R(\mathbf{T}(\mathbf{x}))^\perp \cap \mathcal{Z} \tag{5.16}$$

eine Orthonormalbasis $\{\mathbf{q}_1(\mathbf{x}), \ldots, \mathbf{q}_m(\mathbf{x})\}$ aufgebaut werden kann, sodass für die Matrix $\mathbf{Q} = (\mathbf{q}_1(\mathbf{x}), \ldots, \mathbf{q}_m(\mathbf{x}))$ eine Konstante $L_\mathbf{Q} > 0$ existiert, sodass für alle $\mathbf{x} \in U(\mathbf{x}_0)$

$$\|\mathbf{Q}(\mathbf{x}) - \mathbf{Q}(\mathbf{x}_0)\| \leq L_\mathbf{Q} \|\mathbf{x} - \mathbf{x}_0\| \tag{5.17}$$

gilt. Man beachte, dass in dieser Umgebung stets $\mathbf{T}(\mathbf{x}) \notin \mathcal{Z}^\perp$ gelten soll, da ohne diese Voraussetzung der Raum $\mathcal{Q}(\mathbf{x})$ nicht für den Aufbau eines sinnvollen Testraumes $\mathcal{V}$ verwendet werden kann.

Die naheliegendste Methode, eine Orthonormalbasis eines Raumes zu berechnen, stellt die QR-Zerlegung dar. Hierbei wird für eine gegebene Basis $\{\mathbf{b}_1, \ldots, \mathbf{b}_m\}$ von $\mathcal{Q}$ eine obere rechte Dreiecksmatrix $\mathbf{R}$, sowie eine Matrix $\mathbf{U}$ mit orthonormalen Spalten erzeugt, sodass

$$\mathbf{B} = \mathbf{UR}$$

gilt, wobei $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_m)$ bezeichnet. Die Spalten der Matrix $\mathbf{U}$ enthalten dann eine Orthonormalbasis von $\mathcal{Q}$. In [16] wurde gezeigt, dass die so erzeugten Matrizen die Bedingung (5.17) im Allgemeinen nicht erfüllen. Des Weiteren sind die dort vorgeschlagenen Modifizierungen der QR-Zerlegung mit Nachteilen verbunden, die sie für die Anwendung auf das hier betrachtete Problem ungeeignet machen. Daher werden im Folgenden zwei Alternativen hergeleitet, Matrizen $\mathbf{Q}$ zu konstruieren, die in $U(\mathbf{x}_0)$ die Eigenschaft (5.17) besitzen.

Zunächst sei festgehalten, dass nach Bemerkung 2.2.7 das Tangentialfeld Lipschitz-stetig ist, falls $\mathbf{DF}(\mathbf{x})$ Lipschitz-stetig ist. Daher sei vorausgesetzt, dass eine Konstante $L_\mathbf{F} > 0$ existiert, sodass für alle $\mathbf{x}, \mathbf{y} \in U(\mathbf{x}_0)$

$$\|\mathbf{DF}(\mathbf{x}) - \mathbf{DF}(\mathbf{y})\| \leq L_\mathbf{F} \|\mathbf{x} - \mathbf{y}\|$$

gilt. Der orthogonale Projektor $\mathbf{P}_\mathbf{Q}$ auf den Raum $\mathcal{Q}$ lässt sich über

$$\mathbf{P}_\mathbf{Q} : \begin{cases} \mathbb{R}^{n+1} & \to \mathbb{R}^{n+1} \\ \mathbf{x} & \mapsto (\mathbf{I}_{n+1} - \mathbf{T}(\mathbf{x})\mathbf{T}(\mathbf{x})^T)\mathbf{Z}\mathbf{Z}^T \end{cases}$$

beschreiben. Dabei wird angenommen, dass die Spalten von $\mathbf{Z}$ eine Orthonormalbasis des Ansatzraumes $\mathcal{Z}$ bilden. Die Abbildung $\mathbf{P}_\mathbf{Q}$ ist als Komposition von Lipschitz-stetigen Funktionen selbst wieder Lipschitz-stetig. Man benötigt zunächst folgendes Hilfslemma

Lemma 5.3.1. *Für alle* $\mathbf{x} \in U(\mathbf{x}_0)$ *gilt*

$$\mathbf{T}(\mathbf{x}_0)^T \mathbf{T}(\mathbf{x}) \neq 0,$$

genau dann wenn

$$\mathcal{Q}(\mathbf{x})^\perp \cap \mathcal{Q}(\mathbf{x}_0) = \{\mathbf{0}\}$$

erfüllt ist, wobei $\mathcal{Q}$ *wie in (5.16) aufgebaut ist.*

Beweis. Mit der Mengenbeziehung $(A \cap B)^\perp = A^\perp + B^\perp$ ergibt sich

$$\mathcal{Q}(\mathbf{x})^\perp = (R(\mathbf{T}(\mathbf{x}))^\perp \cap \mathcal{Z})^\perp = R(\mathbf{T}(\mathbf{x})) + \mathcal{Z}^\perp$$

und damit

$$\begin{aligned}\mathcal{Q}(\mathbf{x})^\perp \cap \mathcal{Q}(\mathbf{x}_0) &= (R(\mathbf{T}(\mathbf{x})) + \mathcal{Z}^\perp) \cap (R(\mathbf{T}(\mathbf{x}_0))^\perp \cap \mathcal{Z}) \\ &= R(\mathbf{T}(\mathbf{x})) \cap R(\mathbf{T}(\mathbf{x}_0))^\perp \cap \mathcal{Z}.\end{aligned}$$

Es gilt nun zwei Fälle zu unterscheiden. Liegt $\mathbf{T}(\mathbf{x})$ nicht in $\mathcal{Z}$ gilt sofort $\mathcal{Q}(\mathbf{x})^\perp \cap \mathcal{Q}(\mathbf{x}_0) = \{\mathbf{0}\}$. Gilt dies nicht, dann folgt

$$\mathcal{Q}(\mathbf{x})^\perp \cap \mathcal{Q}(\mathbf{x}_0) = R(\mathbf{T}(\mathbf{x})) \cap R(\mathbf{T}(\mathbf{x}_0))^\perp.$$

Dieser Schnitt enthält nun genau dann den Nullvektor, wenn $\mathbf{T}(\mathbf{x})^\perp \mathbf{T}(\mathbf{x}_0) \neq 0$ gilt. □

Bemerkung 5.3.2. *Wegen der Stetigkeit des Tangentialfeldes existiert stets eine Umgebung $U(\mathbf{x}_0)$, sodass $\mathbf{T}(\mathbf{x}_0)^T \mathbf{T}(\mathbf{x}) \neq \mathbf{0}$, für alle $\mathbf{x} \in U(\mathbf{x}_0)$ gilt.*

Die folgenden beiden Lemmata enthalten jeweils ein Verfahren zum Aufbau der Matrizen $\mathbf{Q}$.

Lemma 5.3.3. *Sei für $\Omega \subset \mathbb{R}^{n+1}$ offen eine Funktion $\mathbf{F} \in C^1(\Omega, \mathbb{R}^n)$, sowie ein Ansatzraum $\mathcal{Z} \subset \mathbb{R}^{n+1}$ gegeben. Es existieren weiterhin ein Punkt $\mathbf{x}_0 \in \mathcal{R}(\mathbf{F})$, sowie eine Umgebung $U(\mathbf{x}_0) \subset \mathcal{R}(\mathbf{F})$, sodass $\mathbf{DF}(\mathbf{x})$ in dieser Umgebung Lipschitz-stetig ist und für alle $\mathbf{x} \in U(\mathbf{x}_0)$*

$$\mathbf{T}(\mathbf{x})^T \mathbf{T}(\mathbf{x}_0) \neq 0 \tag{5.18}$$

gilt. Sei mit $\mathbf{P}_\mathbf{Q}(\mathbf{x})$ der orthogonale Projektor auf den Raum

$$\mathcal{Q} = R(\mathbf{T}(\mathbf{x}))^\perp \cap \mathcal{Z}$$

und mit $G : \mathbb{R}^{n+1,m} \to \mathbb{R}^{n+1,m}$ das Gram-Schmidt-Orthonormalisierungs-Verfahren angewendet auf die Spalten einer Matrix bezeichnet. Des Weiteren sei eine Matrix $\mathbf{Q}_0 \in \mathbb{R}^{n+1,m}$ gegeben, deren Spalten eine Basis von $\mathcal{Q}(\mathbf{x}_0)$ enthalten. Für die Abbildung

$$\mathbf{Q} : \begin{cases} U(\mathbf{x}_0) & \to \mathbb{R}^{n+1,m} \\ \mathbf{x} & \mapsto \mathbf{Q}(\mathbf{x}) := G(\mathbf{P}(\mathbf{x})\mathbf{Q}_0) \end{cases}$$

existiert dann eine Konstante $L_\mathbf{Q} > 0$, sodass für alle $\mathbf{x} \in U(\mathbf{x}_0)$ die Abschätzung

$$\|\mathbf{Q}(\mathbf{x}) - \mathbf{Q}(\mathbf{x}_0)\| \leq L_\mathbf{Q} \|\mathbf{x} - \mathbf{x}_0\| \tag{5.19}$$

erfüllt ist und

$$R(\mathbf{Q}(\mathbf{x})) = \mathcal{Q}(\mathbf{x}) \tag{5.20}$$

gilt.

Beweis. Zunächst wird (5.20) bewiesen, dass also durch die Projektion (und Orthonormalisierung) der Matrix $\mathbf{Q}_0$ auf den Raum $\mathcal{Q}(\mathbf{x})$ eine Matrix entsteht, deren Spalten eine Basis von $\mathcal{Q}(\mathbf{x})$ bilden. Die Orthonormalisierung spielt keine Rolle, da dabei der Raum, der durch die Spalten einer Matrix aufgespannt wird, nicht verändert wird. Es muss also nur gezeigt werden, dass $\text{Rang}(\mathbf{Q}_0) = \text{Rang}(\mathbf{P}_\mathbf{Q}\mathbf{Q}_0)$ gilt.

Angenommen dies wäre nicht der Fall, dann existiert ein von Null verschiedener Vektor $\mathbf{q} \in \mathcal{Q}(\mathbf{x}_0)$ für den $\mathbf{P}_\mathbf{Q}(\mathbf{x})\mathbf{q} = \mathbf{0}$ gilt. Für diesen ergibt sich $\mathbf{q} \in \mathcal{Q}(\mathbf{x})^\perp$ und damit $\mathbf{q} \in \mathcal{Q}(\mathbf{x}_0) \cap \mathcal{Q}(\mathbf{x})^\perp$. Wegen Voraussetzung (5.18) enthält dieser Schnitt nach Lemma 5.3.1 aber nur den Nullvektor, wodurch es zum Widerspruch kommt.

Um nun (5.19) zu zeigen, werden die verknüpften Abbildungen von $\mathbf{Q}(\mathbf{x})$ einzeln betrachtet.

Nach Bemerkung 2.2.7 überträgt sich die Lipschitz-stetigkeit von $\mathbf{DF}(\mathbf{x})$ auf das Tangentialfeld $\mathbf{T}(\mathbf{x})$ und somit auf den Projektor $\mathbf{P}_\mathbf{Q}(\mathbf{x})$. Es existiert also eine Konstante $L_\mathbf{P} > 0$, sodass für alle $\mathbf{x} \in U(\mathbf{x}_0)$ die Ungleichung $\|\mathbf{P}(\mathbf{x}) - \mathbf{P}(\mathbf{x}_0)\| \leq L_\mathbf{P}\|\mathbf{x} - \mathbf{x}_0\|$ gilt.

Zur Betrachtung der Abbildung G sei eine Matrix $\mathbf{A}(\mathbf{x}) = (\mathbf{a}_1(\mathbf{x}), \ldots, \mathbf{a}_m(\mathbf{x})) \in \mathbb{R}^{n+1,m}$ gegeben, deren Spalten Lipschitz-stetig von $\mathbf{x}$ abhängen. Die orthonormalen Spalten $\mathbf{b}_i(\mathbf{x})$ der Matrix $\mathbf{B}(\mathbf{x}) = G(\mathbf{A}(\mathbf{x}))$ ergeben sich dann aus

$$\begin{aligned}
\mathbf{b}_1(\mathbf{x}) &:= \frac{\mathbf{a}_1(\mathbf{x})}{\|\mathbf{a}_1(\mathbf{x})\|}, \\
\tilde{\mathbf{b}}_j(\mathbf{x}) &:= \mathbf{a}_j(\mathbf{x}) - \sum_{i=1}^{j-1} \mathbf{b}_i(\mathbf{x})^T \mathbf{a}_j(\mathbf{x}) \mathbf{b}_i(\mathbf{x}), \\
\mathbf{b}_j(\mathbf{x}) &= \frac{\tilde{\mathbf{b}}_j(\mathbf{x})}{\|\tilde{\mathbf{b}}_j(\mathbf{x})\|}.
\end{aligned}$$

Diese Operationen sind bezüglich der Vektoren $\mathbf{a}_i$ sogar stetig differenzierbar, so lange diese linear unabhängig sind. Daher existiert eine Konstante $L_G > 0$, sodass

$$\begin{aligned}
\|\mathbf{Q}(\mathbf{x}) - \mathbf{Q}(\mathbf{x}_0)\| &= \|G(\mathbf{P}_\mathbf{Q}(\mathbf{x})\mathbf{Q}_0) - G(\mathbf{P}_\mathbf{Q}(\mathbf{x}_0)\mathbf{Q}_0)\| \\
&\leq L_G\|(\mathbf{P}_\mathbf{Q}(\mathbf{x}) - \mathbf{P}_\mathbf{Q}(\mathbf{x}_0))\mathbf{Q}_0\| \leq L_G L_\mathbf{P}\|\mathbf{Q}_0\|\|\mathbf{x} - \mathbf{x}_0\|
\end{aligned}$$

gilt. Setzt man nun $L_\mathbf{Q} := L_G L_\mathbf{P}\|\mathbf{Q}_0\|$, so ist das Lemma bewiesen. □

Die zweite Variante zum Aufbau der Matrizen $\mathbf{Q}(\mathbf{x})$ stellt eine Variation eines in [58] vorgestellten Algorithmus dar. Wiederum wird von einer Matrix $\mathbf{Q}_0$ ausgegangen, deren Spalten dieses mal aber eine Orthonormalbasis von $\mathcal{Q}(\mathbf{x}_0)$ bilden. Für den Raum $\mathcal{Q}(\mathbf{x})$ sei nun bereits eine Orthonormalbasis

$\{\tilde{\mathbf{q}}_1(\mathbf{x}), \ldots, \tilde{\mathbf{q}}_m(\mathbf{x})\}$ bekannt. Für $\tilde{\mathbf{Q}}(\mathbf{x}) := (\tilde{\mathbf{q}}_1(\mathbf{x}), \ldots, \mathbf{q}_m(\mathbf{x}))$ wird nach einer regulären Matrix $\mathbf{D}(\mathbf{x})$ gesucht, die das Minimierungsproblem

$$\|(\tilde{\mathbf{Q}}(\mathbf{x})\mathbf{D}(\mathbf{x}))^T\mathbf{Q}_0 - \mathbf{I}_m\|_F = \min, \quad \mathbf{D}^T(\mathbf{x})\mathbf{D}(\mathbf{x}) = \mathbf{I}_m$$

löst. $\mathbf{D}(\mathbf{x})$ soll also so gewählt werden, dass die Spalten der Matrix $\tilde{\mathbf{Q}}(\mathbf{x})\mathbf{D}(\mathbf{x})$ möglichst gut in Richtung der Spalten von $\mathbf{Q}_0$ zeigen. Die Matrix $\mathbf{Q}(\mathbf{x})$ wird dann über $\mathbf{Q}(\mathbf{x}) := \tilde{\mathbf{Q}}(\mathbf{x})\mathbf{D}(\mathbf{x})$ konstruiert. Man möchte also für jeden Punkt $\mathbf{x}$ eine Orthonormalbasis finden, die bezüglich einer gegebenen im Punkt $\mathbf{x}_0$ im Sinne des Minimerungsproblems optimal ausgerichtet ist. Das folgende Lemma zeigt, wie $\mathbf{D}(\mathbf{x})$ konstruiert werden kann und dass die so definierten Matrizen $\mathbf{Q}(\mathbf{x})$ die Bedingung (5.17) erfüllen.

Lemma 5.3.4. *Sei für $\Omega \subset \mathbb{R}^{n+1}$ offen eine Funktion $\mathbf{F} \in C^1(\Omega, \mathbb{R}^n)$, sowie ein Ansatzraum $\mathcal{Z} \subset \mathbb{R}^{n+1}$ gegeben. Es seien weiterhin ein Punkt $\mathbf{x}_0 \in \mathcal{R}(\mathbf{F})$, sowie eine Umgebung $U(\mathbf{x}_0) \subset \mathcal{R}(\mathbf{F})$ gegeben, sodass $\mathbf{DF}(\mathbf{x})$ in dieser Umgebung Lipschitz-stetig ist und für alle $\mathbf{x} \in U(\mathbf{x}_0)$*

$$\mathbf{T}(\mathbf{x})^T\mathbf{T}(\mathbf{x}_0) \neq 0 \tag{5.21}$$

gilt. Sei mit $\mathbf{P}_\mathbf{Q}(\mathbf{x})$ der orthogonale Projektor auf den Raum

$$\mathcal{Q}(\mathbf{x}) = R(\mathbf{T}(\mathbf{x}))^\perp \cap \mathcal{Z}$$

bezeichnet und mit $\mathbf{Q}_0$ eine Matrix, deren Spalten eine Orthonormalbasis von $\mathcal{Q}(\mathbf{x}_0)$ ergeben. Sei für jeden Punkt $\mathbf{x} \in U(\mathbf{x}_0)$ eine Matrix $\tilde{\mathbf{Q}}(\mathbf{x})$ gegeben, deren Spalten eine beliebige Orthonormalbasis von $\mathcal{Q}(\mathbf{x})$ bilden. Ist dann $\mathbf{D}(\mathbf{x})$ konstruiert über

$$\mathbf{U}(\mathbf{x}) := \tilde{\mathbf{Q}}(\mathbf{x})^T\mathbf{Q}_0, \quad \mathbf{U}(\mathbf{x}) = \mathbf{A}(\mathbf{x})\Sigma(\mathbf{x})\mathbf{B}(\mathbf{x})^T, \quad \mathbf{D}(\mathbf{x}) := \mathbf{A}(\mathbf{x})\mathbf{B}(\mathbf{x})^T,$$

dann bilden die Spalten von $\mathbf{Q}(\mathbf{x}) := \tilde{\mathbf{Q}}(\mathbf{x})\mathbf{D}(\mathbf{x})$ eine Orthonormalbasis von $\mathcal{Q}(\mathbf{x})$ und es existiert eine Konstante $L_\mathbf{Q} > 0$, sodass für alle $\mathbf{x} \in U(\mathbf{x}_0)$

$$\|\mathbf{Q}(\mathbf{x}) - \mathbf{Q}(\mathbf{x}_0)\| \leq L_\mathbf{Q}\|\mathbf{x} - \mathbf{x}_0\|$$

gilt.

Beweis. Zunächst zeigt man, dass die Matrix $\mathbf{U}(\mathbf{x}) \in \mathbb{R}^{m,m}$ regulär ist. Sei dazu $\mathbf{z} \in \mathbb{R}^m$ so gewählt, dass $\mathbf{U}(\mathbf{x})\mathbf{z} = \mathbf{0}$ gilt, dann folgt daraus $\tilde{\mathbf{Q}}(\mathbf{x})^T\mathbf{Q}_0\mathbf{z} = \mathbf{0}$. Somit gilt $\mathbf{Q}_0\mathbf{z} \in \mathcal{Q}(\mathbf{x})^\perp$ aber auch $\mathbf{Q}_0\mathbf{z} \in \mathcal{Q}(\mathbf{x}_0)$. Aus (5.21) und Lemma 5.3.1 folgt nun sofort, dass $\mathbf{z} = \mathbf{0}$ gilt. Für die POD der Matrix $\mathbf{U}(\mathbf{x})$ ergibt sich nun

$$\mathbf{U}(\mathbf{x}) = \mathbf{A}(\mathbf{x})\Sigma(\mathbf{x})\mathbf{B}(\mathbf{x})^T = \mathbf{A}(\mathbf{x})\mathbf{B}(\mathbf{x})^T(\mathbf{B}(\mathbf{x})\Sigma(\mathbf{x})\mathbf{B}(\mathbf{x})^T).$$

Aus

$$\mathbf{B}(\mathbf{x})\Sigma(\mathbf{x})\mathbf{B}(\mathbf{x})^T = (\mathbf{U}(\mathbf{x})^T\mathbf{U}(\mathbf{x}))^{1/2} = (\mathbf{Q}_0^T\tilde{\mathbf{Q}}(\mathbf{x})\tilde{\mathbf{Q}}(\mathbf{x})^T\mathbf{Q}_0)^{1/2}$$

folgt nun

$$\mathbf{D}(\mathbf{x}) = \tilde{\mathbf{Q}}(\mathbf{x})^T\mathbf{Q}_0(\mathbf{Q}_0^T\tilde{\mathbf{Q}}(\mathbf{x})\tilde{\mathbf{Q}}(\mathbf{x})^T\mathbf{Q}_0)^{1/2}.$$

Für den orthogonalen Projektor $\mathbf{P}(\mathbf{x})$ auf $\mathcal{Q}(\mathbf{x})$ gilt $\mathbf{P}(\mathbf{x}) = \tilde{\mathbf{Q}}(\mathbf{x})\tilde{\mathbf{Q}}(\mathbf{x})^T$ und $\mathbf{Q}(\mathbf{x})$ ergibt sich zu

$$\mathbf{Q}(\mathbf{x}) = \tilde{\mathbf{Q}}(\mathbf{x})\mathbf{D} = \mathbf{P}(\mathbf{x})\mathbf{Q}_0(\mathbf{Q}_0^T\mathbf{P}(\mathbf{x})\mathbf{Q}_0)^{-1/2}. \tag{5.22}$$

Für diese Matrix gilt $\mathbf{Q}(\mathbf{x})^T\mathbf{Q}(\mathbf{x}) = \mathbf{I}_m$, da die Spalten von $\mathbf{Q}_0$ orthonormal sind. Des Weiteren folgt aus der Regularität der Matrix $\tilde{\mathbf{Q}}(\mathbf{x})^T\mathbf{Q}_0$, dass $\mathbf{P}(\mathbf{x})\mathbf{Q}_0$ vollen Spaltenrang hat, und die Spalten dieser Matrix eine Basis von $\mathcal{Q}(\mathbf{x})$ bilden. Da die Matrix $(\mathbf{Q}_0^T\mathbf{P}(\mathbf{x})\mathbf{Q}_0)^{-1/2}$ regulär ist, gilt also

$$R(\mathbf{Q}(\mathbf{x})) = \mathcal{Q}(\mathbf{x}).$$

Aus der Lipschitz-Stetigkeit der Jacobimatrix $\mathbf{DF}(\mathbf{x})$ folgt die des Tangentialfeldes $\mathbf{T}(\mathbf{x})$ und des orthogonalen Projektors $\mathbf{P}(\mathbf{x})$. Des Weiteren ist die Abbildung $\mathbf{A} \mapsto \mathbf{A}^{-1/2}$ im Raum der symmetrisch positiv definiten Matrizen stetig differenzierbar und somit ebenfalls Lipschitz-stetig. Daher ist (5.22) als Verknüpfung Lipschitz-stetiger Funktionen selbst wieder Lipschitz-stetig und es existiert ein $L_{\mathbf{Q}} > 0$ mit

$$\|\mathbf{Q}(\mathbf{x}) - \mathbf{Q}(\mathbf{x}_0)\| \leq L_{\mathbf{Q}}\|\mathbf{x} - \mathbf{x}_0\|.$$

□

Bemerkung 5.3.5. *Wie man Gleichung (5.22) entnehmen kann, ist die zweite Variante der ersten ähnlicher als es zunächst den Anschein hat. Beide Verfahren projizieren eine bekannte (Orthononormal-)Basis von $\mathcal{Q}(\mathbf{x}_0)$ auf den entsprechenden Raum $\mathcal{Q}(\mathbf{x})$ um dann entweder mittels Gram-Schmidt-Orthonormalisierungsverfahren (Lemma 5.3.3) oder der Multiplikation mit der Matrix $(\mathbf{Q}_0^T\mathbf{P}(\mathbf{x})\mathbf{Q}_0)^{-1/2}$ (Lemma 5.3.4) die Vektoren orthonormal auszurichten.*

Entscheidend ist für beide Verfahren, dass der Tangentialvektor $\mathbf{T}(\mathbf{x})$ nicht senkrecht auf dem Tangentialvektor $\mathbf{T}(\mathbf{x}_0)$ im Punkt $\mathbf{x}_0$ steht. Die Existenz einer Umgebung $U(\mathbf{x}_0)$, für die diese Bedingung erfüllt ist, ist wegen der (Lipschitz-)Stetigkeit des Tangentialfeldes gesichert. Liegen die Punkte $\mathbf{x}$, für die die $\mathbf{Q}(\mathbf{x})$ erzeugt werden sollen, alle auf einer Lösungskurve von $\mathbf{F}$, so reicht diese Umgebung bis zu dem Punkt auf der Kurve, in dem bezüglich des Tangentialvektors in $\mathbf{x}_0$ ein Umkehrpunkt auftritt, die Kurve sich also senkrecht zu ihrer Richtung in $\mathbf{x}_0$ bewegt.

Der folgende Satz zeigt nun, wie mit Hilfe der beiden vorangegangenen Lemmata Matrizen $\mathbf{V}(\mathbf{x})$, deren Spalten eine Basis von $\mathcal{V}(\mathbf{x})$ enthalten erzeugt werden können.

Satz 5.3.6. *Sei für $\Omega \subset \mathbb{R}^{n+1}$ offen eine Funktion $\mathbf{F} \in C^1(\Omega, \mathbb{R}^n)$, sowie ein Ansatzraum $\mathcal{Z} \subset \mathbb{R}^{n+1}$ gegeben. Es existiere weiterhin ein Punkt $\mathbf{x}_0 \subset \mathcal{R}(\mathbf{F})$, sowie eine Umgebung $U(\mathbf{x}_0) \subset \mathcal{R}(\mathbf{F})$, sodass $\mathbf{DF}(\mathbf{x})$ in dieser Umgebung Lipschitz-stetig ist und für alle $\mathbf{x} \in U(\mathbf{x}_0)$*

$$\mathbf{T}(\mathbf{x})^T \mathbf{T}(\mathbf{x}_0) \neq 0$$

gilt. Für alle $\mathbf{x} \in U(\mathbf{x}_0)$ seien Matrizen $\mathbf{Q}(\mathbf{x})$ gegeben, deren Spalten eine Orthonormalbasis des Raumes $\mathcal{Q}(\mathbf{x}) = \mathbf{R}(\mathbf{T}(\mathbf{x}))^\perp \cap \mathcal{Z}$ enthalten und die nach den Verfahren in Lemma 5.3.3 oder 5.3.4 konstruiert sind. Dann existiert für die Matrizen $\mathbf{V}(\mathbf{x}) := \mathbf{DF}(\mathbf{x})\mathbf{Q}(\mathbf{x})$ eine Konstante $L_\mathbf{V} > 0$, sodass für alle $\mathbf{x} \in U(\mathbf{x}_0)$

$$\|\mathbf{V}(\mathbf{x}) - \mathbf{V}(\mathbf{x}_0)\| \leq L_\mathbf{V} \|\mathbf{x} - \mathbf{x}_0\|$$

gilt.

Beweis. Nach Voraussetzungen existieren zwei Konstante $L_\mathbf{F}, L_\mathbf{Q} > 0$, sodass für alle $\mathbf{x} \in U(\mathbf{x}_0)$

$$\begin{aligned} \|\mathbf{DF}(\mathbf{x}) - \mathbf{DF}(\mathbf{x}_0)\| &\leq L_\mathbf{F} \|\mathbf{x} - \mathbf{x}_0\| \text{ und} \\ \|\mathbf{Q}(\mathbf{x}) - \mathbf{Q}(\mathbf{x}_0)\| &\leq L_\mathbf{Q} \|\mathbf{x} - \mathbf{x}_0\| \end{aligned}$$

gilt. Es ergibt sich dann

$$\begin{aligned} \|\mathbf{V}(\mathbf{x}) - \mathbf{V}(\mathbf{x}_0)\| &= \|\mathbf{DF}(\mathbf{x})\mathbf{Q}(\mathbf{x}) - \mathbf{DF}(\mathbf{x}_0)\mathbf{Q}(\mathbf{x}_0)\| \\ &\leq \|\mathbf{Q}(\mathbf{x})(\mathbf{DF}(\mathbf{x}) - \mathbf{DF}(\mathbf{x}_0))\| \\ &\quad + \|\mathbf{DF}(\mathbf{x}_0)(\mathbf{Q}(\mathbf{x}) - \mathbf{Q}(\mathbf{x}_0))\| \\ &\leq L_\mathbf{F} \|\mathbf{x} - \mathbf{x}_0\| + \sigma_1 L_\mathbf{Q} \|\mathbf{x} - \mathbf{x}_0\| \\ &\leq (L_\mathbf{F} + \sigma_1 L_\mathbf{Q}) \|\mathbf{x} - \mathbf{x}_0\|. \end{aligned}$$

Hierbei bezeichnet σ_1 wieder den größten Singulärwert von $\mathbf{DF}(\mathbf{x}_0)$. Setzt man $L_\mathbf{V} := L_\mathbf{F} + \sigma_1 L_\mathbf{Q}$ so ist der Satz bewiesen. □

Bemerkung 5.3.7. *In Bemerkung 5.5 wurde zusätzlich der Fall betrachtet, dass die Matrizen $\mathbf{V}(\mathbf{x})$ selbst orthonormale Spalten besitzen. Ist man an solchen Matrizen interessiert, kann man analog zu den für den Aufbau der $\mathbf{Q}(\mathbf{x})$ verwendeten Verfahren die Spalten von $\mathbf{DF}(\mathbf{x})\mathbf{Q}(\mathbf{x})$ noch einmal orthonormalisieren. Die Eigenschaft (5.17) geht dabei nicht verloren.*

Kapitel 6

Interpolationsbasierte Reduktion

In diesem Kapitel wird eine neue Reduktion entwickelt, die globale und lokale Eigenschaften vereint. Diese stellt den Kern der Arbeit dar und vereint die Resultate der beiden vorherigen Kapitel 4 und 5. Dort wurden effektive Methoden beschrieben, Ansatz- und Testräume für die Reduktion einer nichtlinearen Gleichung der Form $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ mit $\mathbf{F} \in C^1(\Omega, \mathbb{R}^n)$ aufzubauen. Dabei hat der Ansatzraum eine globale Natur, während die entwickelten Verfahren zum Aufbau der Testräume stets von einem Punkt $\mathbf{x}_0 \in \Omega \subset \mathbb{R}^{n+1}$ abhängen. Diese Räume werden in diesem Kapitel für eine Reduktion verwendet, die eine Art implizite Interpolation der Lösungskurve $\mathbf{c}$ darstellt.

Zunächst wird anhand eines Beispiels erläutert, wieso die lokal aufgebauten Ansatz- und Testräume für eine globale Reduktion ungeeignet sein können.

6.1 Zerfallende Lösungskurven

Durch die Abhängigkeit der Testräume vom Punkt $\mathbf{x}_0$ kann es passieren, dass ein für $\mathbf{x}_0$ aufgebauter Testraum zu einer Reduktion führt, die an anderen Stellen der Lösungskurve unpassend ist und die im ungünstigsten Fall ein Aufbrechen der reduzierten Lösungskurve in verschiedene Lösungsäste zur Folge hat. Dies sei an einem einfachen Beispiel erläutert.

Beispiel 6.1.1 (Das Bratu-Problem)**.** Diskretisiert man das Randwertproblem

$$\begin{cases} -u'' = \lambda \exp(u), & u \in \Omega = (0,1) \\ u(0) = u(1) = 0 \end{cases},$$

mittels der Finiten-Differenzen-Methode und einer äquidistanten Zerlegung des Intervalls $[0,1]$ mit Schrittweite h, so ergibt sich die nichtlineare Gleichung

$\mathbf{G}(\mathbf{u}, \lambda) = \mathbf{0}$ mit

$$\mathbf{G} : \begin{cases} \mathbb{R} \times \mathbb{R}^n \supseteq \Lambda \times Y & \to \mathbb{R}^n, \\ (\mathbf{u}^T, \lambda)^T & \mapsto \mathbf{D}\mathbf{u} - \lambda \exp(\mathbf{u}) \end{cases},$$

mit den offenen Mengen Λ und Y. Der Vektor $\mathbf{u} = (u_1, \ldots, u_n)^T$ enthält dabei die Funktionswerte von u in den einzelnen Knotenpunkten der Zerlegung. Es gilt dann $\exp(\mathbf{u}) := (\exp(u_1), \ldots, \exp(u_n))^T$. Die Matrix $\mathbf{D}$ repräsentiert die diskretisierte negativen zweite Ableitung und hat die Gestalt

$$\mathbf{D} = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}.$$

Fasst man nun die Variablen $\mathbf{u}$ und λ als $\mathbf{x} := (\mathbf{u}^T, \lambda)^T$ zusammen und setzt $\Omega := \Lambda \times Y$, so ergibt sich mit

$$\mathbf{F}(\mathbf{x}) := \mathbf{G}(\mathbf{u}, \lambda) = \mathbf{0}$$

ein überbestimmtes nichtlineares Gleichungssystem der Form (2.11).

Ausgehend von der bekannten Lösung $\mathbf{x}_0 = \mathbf{0}$ lässt sich die Lösungskurve mittels Astverfolgsungsmethoden numerisch berechnen und es ergibt sich ein Bild wie in Abbildung 6.1. Die Lösungskurve besitzt bezüglich λ bei $\lambda \approx 3.5$ einen Umkehrpunkt, weshalb eine Reduktion wie sie in Kapitel 3.2 beschrieben ist, nicht möglich ist.

Es werden nun 4 Punkte $\{\mathbf{x}_0, \ldots, \mathbf{x}_3\}$ (Snapshots) ausgewählt und ein Lagrange-Raum mittels

$$\mathcal{Z} = \operatorname{span}\{\mathbf{x}_i - \mathbf{x}_0,\ i = 1, 2, 3\}$$

aufgebaut, sowie eine Matrix $\mathbf{Z} \in \mathbb{R}^{n+1,3}$, deren Spalten eine Orthonormalbasis von $\mathcal{Z}$ bilden.

Abhängig von dem Snapshot für den der Testram $\mathcal{V}$ aufgebaut wird, kommt es nun zu sehr unterschiedlichen Resultaten für die Reduktion. Dabei spielt es keine Rolle, welches der beiden in diesem Kapitel vorgestellten Verfahren zum Aufbau von $\mathcal{V}$ verwendet wird. Bezeichnet man mit $\mathbf{V}_j$ die Matrix, deren Spalten eine Basis des im Punkt $\mathbf{x}_j$ aufgebauten Testraumes bilden, so zeigt sich, dass die Lösungsmenge der Reduktion

$$\hat{\mathbf{F}}_j(\hat{\mathbf{x}}) = \mathbf{V}_j^T \mathbf{F}(\mathbf{x}_0 + \mathbf{Z}\hat{\mathbf{x}}) = \mathbf{0}$$

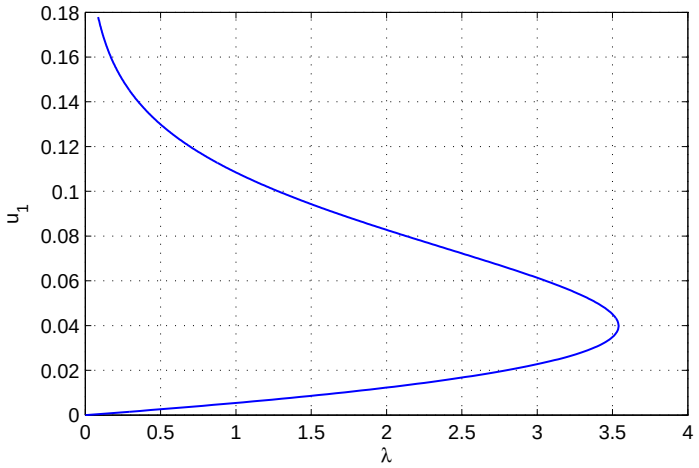

Abbildung 6.1: Erste Komponente der Lösung von $\mathbf{G}(\mathbf{u}, \lambda) = \mathbf{0}$ in Abhängigkeit von λ

für $j = 2, 3$ eine sinnvolle Approximation der Lösungskurve liefert, die durch alle Snapshots $\mathbf{x}_1, \ldots, \mathbf{x}_4$ verläuft. Für $j = 1, 4$ zerfällt die Lösungskurve jedoch in mehrere Lösungsäste. In Abbildung 6.2 ist dieses Zerfallen für den Fall $j = 4$ dargestellt. In diesem Fall ist es also nicht möglich, die in $\mathbf{x}_1$ und $\mathbf{x}_4$ aufgebauten Testräume für eine globale Reduktion zu verwenden.

6.2 Grundidee

Das vorherige Kapitel hat gezeigt, dass die lokal aufgebauten Testräume sich nicht für eine globale Reduktion eignen müssen. In einer Umgebung des Punktes, in dem sie aufgebaut wurden, liefern die Testräume jedoch Reduktionen, die zu sinnvollen Ergebnissen führen. Daher ist es naheliegend, die den verwendeten Testraum repräsentierende Matrix $\mathbf{V}$ über die Abbildung

$$\mathbf{V} : \begin{cases} \Omega & \to \mathbb{R}^{n,m}, \\ \mathbf{x} & \mapsto \mathbf{V}(\mathbf{x}) \end{cases},$$

ortsabhängig zu wählen. Die Matrix $\mathbf{V}(\mathbf{x})$ sei dabei nach einem der in Kapitel 5 entwickelten Verfahren aufgebaut. Die globale Reduktion hat dann die Gestalt

$$\hat{\mathbf{F}}(\hat{\mathbf{x}}) := \mathbf{V}(\mathbf{Z}\hat{\mathbf{x}})^T \mathbf{F}(\mathbf{Z}\hat{\mathbf{x}}).$$

Es gibt nun drei Gründe, wieso dieses Vorgehen, ungünstig ist.

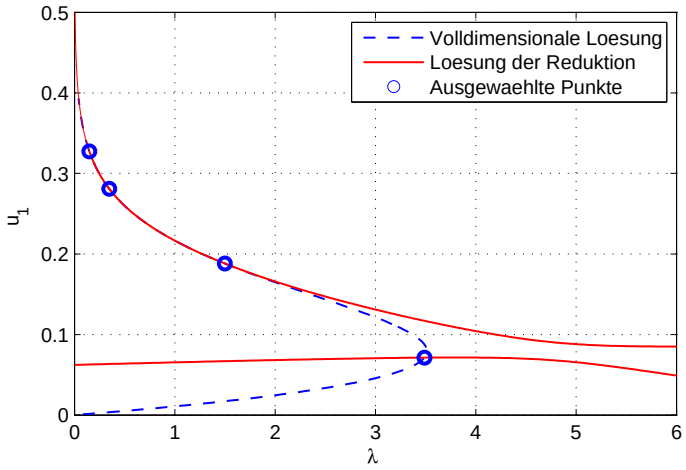

Abbildung 6.2: Zerfallen der Lösungskurve der Reduktion in mehrere Lösungsäste

Zum einen kann nicht mehr gesichert werden, dass die Abbildung $\hat{\mathbf{F}}$ stetig differenzierbar ist, da die beiden in Kapitel 5 beschriebenen Verfahren keine stetig differenzierbaren Matrizen $\mathbf{V}(\mathbf{Z}\hat{\mathbf{x}})$ liefern. Nutzt man die POD zum Aufbau der Testräume wie in Kapitel 5.1 ergeben sich noch nicht einmal stetige Matrizen, verwendet man die Alternative aus Kapitel 5.2, wäre die Abbildung $\mathbf{V}$ nur dann stetig differenzierbar, wenn $\mathbf{F} \in C^2(\Omega, \mathbb{R}^{n,n+1})$ gilt. Und selbst wenn dies der Fall wäre, müssten zur Sicherung der Lösungsexistenz von $\hat{\mathbf{F}}(\hat{\mathbf{x}}) = \mathbf{0}$ Abschätzungen für die zweite Ableitung von $\mathbf{F}$ getroffen werden, die wegen ihrer Tensorstruktur nur ausgesprochen umständlich möglich wären.

Zweitens kann die Regularität der Jacobimatrix der reduzierten Funktion nicht einmal in dem Snapshot $\mathbf{x}_0 = \mathbf{Z}\hat{\mathbf{x}}_0$, für den $\mathcal{V}$ erzeugt wurde, gesichert werden. Dies liegt daran, dass bei einer $\mathbf{x}$-Abhängigkeit der $\mathbf{V}$ diese Jacobimatrix zu

$$\mathbf{V}(\mathbf{x}_0)^T \mathbf{DF}(\mathbf{x}_0)\mathbf{Z} + R(\mathbf{x}_0)$$

Der Restterm $R(\mathbf{x}_0) \in \mathbb{R}^{m,m+1}$ hängt von der Ableitung der Abbildung $\mathbf{V}$ in $\mathbf{x}_0$ ab. Die Matrizen $\mathbf{V}$ wurden gerade so aufgebaut, dass der erste der beiden Summanden vollen Zeilenrang hat. Es kann nun aber nicht garantiert werden, dass dieser durch die Addition von $R(\mathbf{x}_0)$ erhalten bleibt.

Der dritte (und wichtigste) Grund, keine gänzlich ortsabhängige Matrix $\mathbf{V}$ zu verwenden, liegt in der kostspieligen Auswertung des Funktionswertes $\hat{\mathbf{F}}$. Für jeden neuen Wert $\hat{\mathbf{x}}$ muss die Matrix $\mathbf{V}(\mathbf{Z}\hat{\mathbf{x}})$ neu bestimmt werden, was schnell sehr aufwändig werden kann. Bei dem in Kapitel 5.2 beschriebenen Verfahren wird zum Beispiel für den Aufbau von $\mathbf{V}(\mathbf{Z}\hat{\mathbf{x}})$ der Tangentialvektor be-

nötigt, dessen Berechnung das Lösen eines linearen $n \times n+1$-Gleichungssystems einschließt. Somit ergäbe sich kein wirklicher Vorteil gegenüber dem direkten Lösen des volldimensionalen Systems.

Daher bedarf es eines Kompromisses zwischen einer festen globalen und einer vollkommen ortsabhängigen Reduktion. Ein solcher wird im Folgenden mit der interpolationsbasierten Reduktion entwickelt, bei der die Funktion $\mathbf{V}(\mathbf{x})$ durch eine Interpolierende ersetzt wird.

Man geht davon aus, dass ein globaler Ansatzraum $\mathcal{Z}$ gegeben ist, sowie eine Menge von Punkten $X^I = \{\mathbf{x}_1, \ldots, \mathbf{x}_d\} \subset \mathcal{Z}$. Diese Punkte werden als Interpolationsknoten bezeichnet. Im Beispiel 6.1.1 stimmt der Interpolationsknoten mit den für den Lagrange-Raum $\mathcal{Z}$ verwendeten Snapshots überein, dies muss aber nicht der Fall sein. Für $j = 1, \ldots, d$ liefern die Reduktionen

$$\hat{\mathbf{F}}_j(\hat{\mathbf{x}}) = \mathbf{V}_j^T \mathbf{F}(\mathbf{Z}\hat{\mathbf{x}})$$

mit $\mathbf{V}_j := \mathbf{V}(\mathbf{x}_j)$ sinnvolle Ergebnisse in einer Umgebung um den jeweiligen Punkt $\mathbf{x}_j$. Dabei liegen die sich ergebenen reduzierten Lösungskurven wegen des festen Ansatzraumes $\mathcal{Z}$ alle im selben Unterraum. Ziel ist es nun, diese lokalen Lösungskurven zu einer gemeinsamen Kurve zu verbinden.

Es seien dazu stetig differenzierbare Gewichtsfunktionen $w_i : \mathbb{R}^{n+1} \to \mathbb{R}$ gegeben, deren Träger beschränkt ist und die der Eigenschaft $w_i(\mathbf{x}_j) = \delta_{ij}$ genügen. Die Funktion

$$\hat{\mathbf{F}}_I(\hat{\mathbf{x}}) := \sum_{j=1}^{d} w_j(\mathbf{Z}\hat{\mathbf{x}}) \hat{\mathbf{F}}_j(\hat{\mathbf{x}}) = \sum_{j=1}^{d} w_j(\mathbf{Z}\hat{\mathbf{x}}) \mathbf{V}_j^T \mathbf{F}(\mathbf{Z}\hat{\mathbf{x}}) \tag{6.1}$$

wird als die interpolierte Reduktion von $\mathbf{F}$ bezeichnet und das nichtlineare Gleichungssystem

$$\hat{\mathbf{F}}_I(\hat{\mathbf{x}}) = \mathbf{0}$$

als das interpolierte Problem. In [56] wurde bereits ein ähnlicher Ansatz zur Verbindung von lokal linearisierten dynamischen Systemen verwendet.

Zur Veranschaulichung ist in Abbildung 6.3 eine Lösungskurve mit 5 Interpolationsknoten und den zu den jeweiligen Gewichtsfunktionen gehörenden Trägern dargestellt. Diese Träger sind dabei, wie im weiteren Verlauf der Arbeit auch, Kugeln.

Für die weiteren Betrachtungen dieser Reduktion seien die nötigen Voraussetzungen, die im Folgenden stets an $\mathbf{F}$ gestellt werden, zusammengefasst:

Voraussetzung 6.2.1.

(i) $\mathbf{F} \in C^1(\Omega, \mathbb{R}^n)$ mit $\Omega \subset \mathbb{R}^{n+1}$ offen.

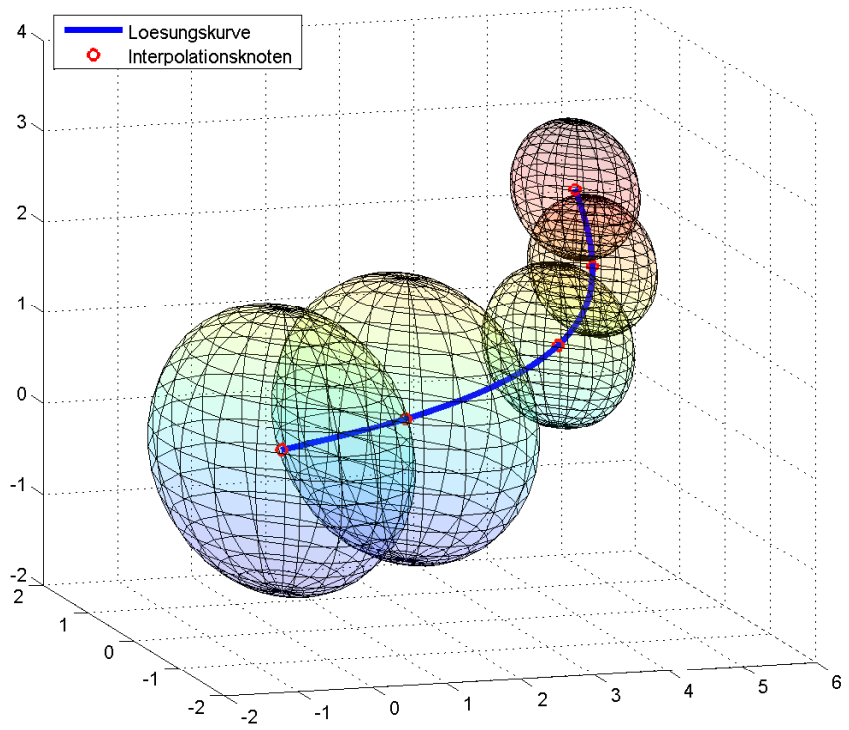

Abbildung 6.3: Lösungskurve, Interpolationsknoten und dazugehörige Träger der Gewichtsfunktionen

(ii) Es existiert eine Konstante $L_{\mathbf{F}} > 0$, sodass für alle $\mathbf{x}, \mathbf{y} \in \Omega$

$$\|\mathbf{DF}(\mathbf{x}) - \mathbf{DF}(\mathbf{y})\| \leq L_{\mathbf{F}} \|\mathbf{x} - \mathbf{y}\|$$

gilt.

(iii) Es existiert eine Funktion $\mathbf{c} \in C^1(S, \Omega)$ mit $S \subset \mathbb{R}$ offen, sodass für alle $s \in S$

$$\mathbf{F}(\mathbf{c}(s)) = \mathbf{0},\ \|\mathbf{c}'(s)\| = 1 \text{ und } \|\mathbf{c}''(s)\| \leq c_{\mathbf{c}}$$

gilt.

(iv) Für jeden Punkt $\mathbf{x}_* \in \Omega$ mit $\mathbf{F}(\mathbf{x}_*) = \mathbf{0}$ existiert ein $s_* \in S$, sodass $\mathbf{c}(s_*) = \mathbf{x}_*$ gilt. Es sollen also abseits der Lösungskurve $\mathbf{c}$ keine weiteren Lösungen von $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ in Ω existieren.

(v) Es existieren zwei Konstanten $c_0, c_1 > 0$, sodass für den kleinsten und größten Singulärwert $\sigma_1(\mathbf{x})$ und $\sigma_n(\mathbf{x})$ von $\mathbf{DF}(\mathbf{x})$ und für alle $\mathbf{x} \in \Omega$

$$\sigma_1(\mathbf{x}) \leq c_1 \text{ und } \sigma_n^{-1}(\mathbf{x}) \leq c_0$$

gilt.

Im Folgenden werden zunächst mögliche Gewichtsfunktionen w_i eingeführt.

6.3 Gewichtsfunktionen

Es gibt verschiedene Möglichkeiten, Gewichtsfunktionen zu finden, die für eine gegeben Menge von Punkten $\{\mathbf{x}_1, \ldots, \mathbf{x}_d\}$ die Bedingung $w_i(\mathbf{x}_j) = \delta_{ij}$ erfüllen. In [41] werden zum Beispiel stückweise lineare Funktionen verwendet, die allerdings nur schwach differenzierbar sind und daher für den in dieser Arbeit betrachteten Fall nicht in Frage kommen. Alternativ finden sich in [56] exponentielle Gewichtsfunktionen, diese besitzen zwar die nötige Glattheit, ihre Träger sind jedoch unbeschränkt. In dieser Arbeit werden daher sogenannte Bumper-Funktionen (vergleiche [71]) verwendet, die im Folgenden näher beschrieben werden. Ziel ist es, mit Hilfe dieser Funktionen eine Zerlegung der Eins zu erzeugen, zusätzlich zur stetigen Differenzierbarkeit, sollen die gesuchten Gewichtsfunktionen w_j in einem im Folgenden noch genauer beschriebenen Gebiet in der Summe stets 1 ergeben.

Seien die Funktionen f, g und H wie folgt definiert:

$$f : \mathbb{R} \to \mathbb{R} \text{ mit } f(t) = \begin{cases} \exp(-1/t), & \text{für } t > 0, \\ 0, & \text{für } t \leq 0 \end{cases},$$

$$g : \mathbb{R} \to \mathbb{R} \text{ mit } g(t) = \frac{f(t)}{f(t) + f(1-t)},$$

$$H : \mathbb{R}^{n+1} \to \mathbb{R} \text{ mit } H(\mathbf{x}) = g(2 - \|\mathbf{x}\|).$$

Das folgende Lemma enthält nun Aussagen über die Differenzierbarkeit und den Wertebereich von H. Ein Beweis findet sich in [71].

Lemma 6.3.1. *Die Funktionen f, g und H sind unendlich oft differenzierbar und es gilt für alle $\mathbf{x} \in \mathbb{R}^{n+1}$*

$$H(\mathbf{x}) = \begin{cases} 1 & \text{für } \|\mathbf{x}\| \leq 1, \\ g(2 - \|\mathbf{x}\|) & \text{für } \|\mathbf{x}\| \in (1,2), \\ 0 & \text{für } \|\mathbf{x}\| \geq 2, \end{cases}.$$

In Abbildung 6.4 ist die Funktion H für den eindimensionalen Fall abgebildet. Mittels

$$r_j := \tau \min_{i=1,\ldots,d, i \neq j} \{\|\mathbf{x}_i - \mathbf{x}_j\|\} \tag{6.2}$$

definiert man

$$\tilde{w}_j(\mathbf{x}) := H(2r_j^{-1}(\mathbf{x} - \mathbf{x}_j)). \tag{6.3}$$

Der Wert $\tau \in (0, 1]$ sollte für die im Folgenden mittels der w_j erzeugten Zerlegung der Eins idealerweise so gewählt werden, dass die Vereinigung der Kugeln $B(\mathbf{x}_j; r_j)$ zusammenhängend ist. Die Funktionen $\tilde{w}_j$ erfüllen nun die Voraussetzung $\tilde{w}_i(\mathbf{x}_j) = \delta_{ij}$ und sind außerdem als Komposition unendlich oft differenzierbarer Funktionen, selbst wieder unendlich oft differenzierbar. Ihre Träger sind dabei die Kugeln $B(\mathbf{x}_j; r_j)$, also beschränkt in $\mathbb{R}^{n+1}$.

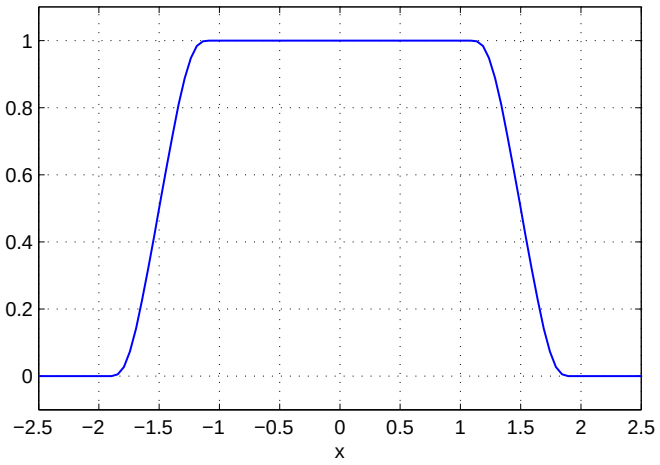

Abbildung 6.4: Graph der Funktion H

Seien nun mit $B^o(\mathbf{x}_j; r_j) := \{\mathbf{x} \in \mathbb{R}^{n+1} : \|\mathbf{x}-\mathbf{x}_j\| < r_j\}$ die offenen Kugeln um $\mathbf{x}_j$ mit Radius r_j bezeichnet, und die Menge Ω_w als

$$\Omega_w := \bigcup_{j=1,\ldots,d} B^o(\mathbf{x}_j; \mathbf{r}_j)$$

definiert. Die Gewichtsfunktionen w sollen zusätzlich noch für alle $\mathbf{x} \in \Omega_w$ die Bedingung

$$\sum_{j=1}^{d} w(\mathbf{x}) = 1,$$

erfüllen. Diese Eigenschaft wird für spätere Abschätzungen beim Beweis der Lösungsexistenz des interpolierten Problems von Bedeutung sein. Sei $s(\mathbf{x}) := \sum_{j=1}^{d} \tilde{w}_j(\mathbf{x})$ die Summe der in (6.3) beschriebenen Funktionen $\tilde{w}_j$ im Punkt $\mathbf{x}$. Man beachte, dass die Menge Ω_w offen ist und für alle $\mathbf{x} \in \Omega_w$ immer mindestens ein $k \in \{1, \ldots, j\}$ existiert, sodass $w_k(\mathbf{x}) \neq 0$ gilt. Daraus folgt $s(\mathbf{x}) \neq 0$ in Ω_w und es lassen sich die Funktionen

$$w_j(\mathbf{x}) : \begin{cases} \Omega_w & \to \mathbb{R}, \\ \mathbf{x} & \mapsto \tilde{w}(\mathbf{x})/s(\mathbf{x}) \end{cases}. \tag{6.4}$$

definieren. Ihre Eigenschaften sind im folgenden Lemma zusammen gefasst.

Lemma 6.3.2. *Seien die Funktionen $w_j, j = 1, \ldots, d$ wie in (6.4) gegeben, dann sind sie unendlich oft differenzierbar in Ω_w und es gilt*

$$\sum_{j=1}^{d} w_j(\mathbf{x}) = 1.$$

Sei weiterhin $\Omega_I \subset \Omega_w$ eine abgeschlossene Teilmenge, dann existiert eine Konstante $c_w > 0$, sodass für alle $\mathbf{x} \in \overline{\Omega^I}$

$$\|\nabla w_j(\mathbf{x})\| \leq c_w r_j^{-1}$$

gilt.

Beweis. Da der Nenner $s(\mathbf{x})$ in Ω_w nicht Null wird, sind die w_j als Quotient unendlich oft differenzierbarer Funktionen selbst wieder unendlich oft differenzierbar. Weiterhin gilt für alle $\mathbf{x} \in \Omega_w$

$$\sum_{j=1}^{d} w_j(\mathbf{x}) = \sum_{j=1}^{d} \frac{\tilde{w}_j(\mathbf{x})}{s(\mathbf{x})} = \frac{s(\mathbf{x})}{s(\mathbf{x})} = 1.$$

Für die Untersuchung des Gradienten sei zunächst festgehalten, dass wegen der stetigen Differenzierbarket von H ein $c_h > 0$ existiert, sodass für alle $\mathbf{x} \in \mathbb{R}^{n+1}$ die Abschätzung $\|\nabla H(\mathbf{x})\| \leq c_H$ gilt. Für den Gradienten von $\tilde{w}_j$ ergibt sich so

$$\begin{aligned}\|\nabla \tilde{w}_j(\mathbf{x})\| &= \|\nabla H(2r_j^{-1}(\mathbf{x}-\mathbf{x}_j))\| = \|\nabla H(2r_j^{-1}(\mathbf{x}-\mathbf{x}_j))2r_j^{-1}\| \\ &\leq 2c_H r_j^{-1}.\end{aligned}$$

Da Ω_I abgeschlossen und beschränkt ist, existiert ein $s_{min} > 0$ mit $s(\mathbf{x}) \geq s_{min}$ in Ω_I. Somit ergibt sich

$$\begin{aligned}\|\nabla w_j(\mathbf{x})\| &= \|s(\mathbf{x})^{-2}(\nabla \tilde{w}_j(\mathbf{x})s(\mathbf{x}) - \nabla s(\mathbf{x})\tilde{w}_j(\mathbf{x}))\| \\ &\leq \frac{1}{s_{min}}\|\nabla \tilde{w}_j(\mathbf{x})\| + \frac{1}{s_{min}^2}\|\nabla s(\mathbf{x})\|\|\tilde{w}_j(\mathbf{x})\| \\ &\leq \frac{2c_H}{s_{min}}r_j^{-1} + \frac{2pc_H}{s_{min}^2}r_j^{-1}.\end{aligned}$$

Die natürliche Zahl p ist dabei die Anzahl der sich im Punkt $\mathbf{x}$ überschneidenden Träger. Mit $c_w := \max(2c_H s_{min}^{-1}, 2pc_h s_{min}^{-2})$ folgt daraus

$$\|\nabla w_j(\mathbf{x})\| \leq c_w r_j^{-1}.$$

□

6.4 Existenz einer Lösung

In diesem Kapitel werden Bedingungen für die Existenz einer Lösungskurve des interpolierten Problems

$$\hat{\mathbf{F}}_I(\hat{\mathbf{x}}) = \sum_{j=1}^{d} w_j(\mathbf{Z}\hat{\mathbf{x}})\mathbf{V}_j^T\mathbf{F}(\mathbf{Z}\hat{\mathbf{x}}) = \mathbf{0} \tag{6.5}$$

hergeleitet. Dabei wird die Menge der Interpolationsknoten $X^I = \{\mathbf{x}_1, \ldots, \mathbf{x}_d\}$ im Folgenden noch näher spezifiziert. Ziel ist es, zu beweisen, dass für eine Funktion nach Voraussetzung 6.2.1 stets ein interpoliertes Problem gefunden werden kann, dass eine zur Approximation von $\mathbf{c}$ geeignete Lösungskurve besitzt.

6.4.1 Interpolation mittels Lagrange-Ansatzraum

Als erstes wird der Fall betrachtet, dass die Interpolationsknoten X^I auf der Lösungskurve liegen und gleichzeitig als die Punkte dienen, mit denen der

Ansatzraum als Lagrange-Raum aufgebaut wird. Dieser Ansatzraum sei mit $\mathcal{Z}$ bezeichnet und es gilt

$$\mathcal{Z} = \text{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_{m+1}\}. \tag{6.6}$$

Ohne Einschränkung wird angenommen, dass stets $\dim(\mathcal{Z}) = m + 1$ gilt. Die diesen Raum repräsentierende Matrix sei mit $\mathbf{Z}$ bezeichnet und besitze orthonormale Spalten. Weiterhin seien mit $\hat{\mathbf{x}}_j$ die Koeffizientenvektoren der Knoten $\mathbf{x}_j$ bezüglich $\mathbf{Z}$ bezeichnet; es gilt also $\mathbf{x}_j = \mathbf{Z}\hat{\mathbf{x}}_j$. Da die $\mathbf{x}_j$ alle Teil der Lösungskurve sind, ist $\mathbf{F}(\mathbf{x}_j) = \mathbf{0},\ j = 1, \ldots, m + 1$.

Zunächst wird näher beschrieben welche Eigenschaften für die Interpolation gefordert werden.

Definition 6.4.1. Sei $\mathbf{F}$ nach Voraussetzung 6.2.1, sowie $X_k^I := \{\mathbf{x}_1, \ldots, \mathbf{x}_{m_k+1}\}$ mit $1 \leq m_k < m_{k+1}$ und $\mathbf{x}_j = \mathbf{c}(s_j),\ j = 1, \ldots, m_k + 1$ eine Folge von Interpolationsystemen. Die Mengen X_k^I seien hierarchisch, d.h. es gelte $X_k^I \subset X_{k+1}^I$. Für die Menge X_k^I sei der Wert $h_k > 0$ mit $h_k := \max\{|s_{i+1} - s_i|,\ i = 1, \ldots, m_k\}$ definiert und weiterhin eine Folge von Gewichtsfunktionen $W_k^I := \{w_j : \mathbb{R}^{n+1} \to \mathbb{R},\ j = 1, \ldots, m_k+1\}$ mit beschränktem offenen konvexen Träger Ω_k^j gegeben. Die Menge $\{\Omega_k^I, X_k^I, W_k^I\}$ wird als eine Folge von zulässigen Interpolationen von $\mathbf{c}$ bezeichnet, wenn weiterhin gilt:

(i) Ω_k^I ist eine Folge abgeschlossener Mengen mit $\Omega_k^I \subset \bigcup \Omega_k^j$ und $\mathbf{c} \subset \Omega_k^I$ und $\Omega_k^I \subset \mathcal{R}(\mathbf{F})$,

(ii) es gilt $\Omega_k^j \cap \Omega_k^{j+l} = \emptyset$ für $l \notin \{-1, 0, 1\}$ und es existiert eine Konstante $c_\Omega > 0$ unabhängig von k, sodass $\text{diam}(\Omega_k^j) \leq c_\Omega h,\ j = 1, \ldots, m_k + 1$ gilt,

(iii) für die Gewichtsfunktionen gilt $w \in C^1(\Omega_k^I)$, $\sum_{j=1}^{m_k+1} w_j(\mathbf{x}) = 1$, für alle $\mathbf{x} \in \Omega_k^I$, sowie $w_j(\mathbf{x}_i) = \delta_{ij},\ i, j = 1, \ldots, m_k + 1$. Des Weiteren existiert eine Konstante $c_w > 0$ unabhängig von k mit $\|\nabla w_i(\mathbf{x})\| \leq c_w h_k^{-1}$ für $i = 1, \ldots, m_k + 1$,

(iv) es existieren Konstanten $c_h, c_s > 0$ unabhängig von k, sodass $\|\mathbf{x}_{j+1} - \mathbf{x}_j\| \geq c_h h_k,\ j = 1, \ldots, m_k$ und $h_k \leq c_s m_k^{-1}$ gilt,

(v) es existieren $\gamma_j > 0$, sodass $\mathbf{x}_{j-1}, \mathbf{x}_{j+1} \in B(\mathbf{x}_j; \gamma_j) \subset \Omega,\ j = 2, \ldots, m_k$ gilt und für alle $\mathbf{x} \in B(\mathbf{x}_j; \gamma_j)$ die Bedingung $\mathbf{T}(\mathbf{x})^T \mathbf{T}(\mathbf{x}_j) \neq 0$ erfüllt ist,

(vi) es existiert ein $\tau \in (0, 1)$ unabhängig von k, sodass für alle $j = 1, \ldots, m_k$ gilt: $B(t\mathbf{x}_j + (1 - t)\mathbf{x}_{j+1}; \tau h) \subset \Omega_k^I,\ t \in [0, 1]$.

Bemerkung 6.4.2. *Man beachte zunächst, dass die Gewichtsfunktionen aus Kapitel 6.3 die sich auf die Gewichtsfunktionen beziehenden Teile der Definition 6.4.1 erfüllen. Weiterhin sind die Bedingungen (i),(ii),(v) und (vi) für*

eine bezüglich s äquidistante Wahl der Knoten entlang der Lösungskurve bei genügend kleiner Schrittweite erfüllt.

Ziel dieses Kapitels ist es, zu beweisen, dass für hinreichend große k, ein mittels einer zulässigen Interpolation nach Definition 6.4.1 erzeugtes interpoliertes Problem eine Lösungskurve besitzt, die durch alle Interpolationsknoten läuft.

Um diese Aussage zu beweisen, wird Satz 2.2.3 verwendet. Vorher wird mit einigen Hilfslemmata gezeigt, dass in diesem Satz auftretende wichtige Größen, wie die kleinsten Singulärwerte oder die Lipschitzkonstante der interpolierten Funktion unabhängig von k abgeschätzt werden können.

Lemma 6.4.3. *Sei* $\mathbf{F}$ *nach Voraussetzung 6.2.1 und* $\{\Omega_k^I, X_k^I, W_k^I\}$ *eine Folge zulässiger Interpolationen von* $\mathbf{c}$ *nach Definition 6.4.1, sowie der Ansatzraum* $\mathcal{Z}_k$ *ein Lagrange-Ansatzraum mit* $\mathcal{Z}_k := span\{\mathbf{x}_j,\ j = 1, \ldots, m_k + 1,\ \mathbf{x}_j \in X_k^I\}$. *Seien weiterhin Matrizen* $\mathbf{V}_j,\ j = 1, \ldots, m_k + 1$ *mittels eines der beiden Verfahren aus Kapitel 5 aufgebaut. Dann existieren zwei Konstanten* $\hat{c}_0, \hat{c}_1 > 0$, *unabhängig von* k, *sodass für alle* $j = 1, \ldots, m_k + 1$ *und die kleinsten und größten Singulärwerte* $\hat{\sigma}_{m_k}^j$ *bzw.* $\hat{\sigma}_1^j$ *von* $\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}_j)$ *die Abschätzung*

$$(\hat{\sigma}_{m_k}^j)^{-1} \leq \hat{c}_0, \text{ und } \hat{\sigma}_1^j \leq \hat{c}_1$$

gilt.

Beweis. Zunächst sei festgehalten, dass für die interpolierte Reduktion $\hat{\mathbf{F}}_I$ von $\mathbf{F}$

$$\begin{aligned}\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}_j) &= \sum_{i=1}^{m_k+1} \left(\underbrace{w_i(\mathbf{x}_j)}_{=\delta_{ij}} \mathbf{V}_i^T \mathbf{D}\mathbf{F}(\mathbf{x}_j)\mathbf{Z} + \mathbf{V}_i^T \underbrace{\mathbf{F}(\mathbf{x}_j)}_{=0} \nabla w_i(\mathbf{x}_j) \right) \\ &= \mathbf{V}_j^T \mathbf{D}\mathbf{F}(\mathbf{x}_j)\mathbf{Z}\end{aligned}$$

gilt.

Nach Voraussetzung 6.2.1 existiert ein $c_0 > 0$, sodass für den kleinsten Singulärwert $\sigma_n(\mathbf{x})$ von $\mathbf{D}\mathbf{F}(\mathbf{x})$ und alle $\mathbf{x} \in \Omega_k^I$ gilt: $\sigma_n(\mathbf{x})^{-1} \leq c_0$. Sind die Matrizen $\mathbf{V}_j$ mittels des Verfahrens aus Satz 5.1.3 aufgebaut gilt wegen Bemerkung 5.1.5 direkt

$$(\hat{\sigma}_{m_k}^j)^{-1} \leq c_0 =: \hat{c}_0.$$

Nutzt man stattdessen das Verfahren aus Satz 5.2.4 ergeben sich die Abschätzungen

$$(\hat{\sigma}_{m_k}^j)^{-1} \leq \sigma_n^{-2}(\mathbf{x}_j) \left(1 + \frac{1}{\tau^2 \|\mathbf{P}_k \mathbf{T}(\mathbf{x}_j)\|^2}\right)^{1/2}, \text{ bzw.}$$

$$(\hat{\sigma}_{m_k}^j)^{-1} \leq \sigma_1(\mathbf{x}_j)\sigma_n^{-2}(\mathbf{x}_j) \left(1 + \frac{1}{\tau^2 \|\mathbf{P}_k \mathbf{T}(\mathbf{x}_j)\|^2}\right)^{1/2},$$

abhängig davon, ob die $\mathbf{V}_j$ orthonormale Spalten besitzen. $\mathbf{P}_k$ bezeichnet hierbei den orthogonalen Projektor auf den Ansatzraum $\mathcal{Z}_k$. Nach Voraussetzung 6.2.1 gilt für alle $\mathbf{x} \in \Omega_k^I$ $\sigma_1(\mathbf{x}_j) \leq c_1$. Die oberen Schranken hängen also nur über den Wert $\|\mathbf{P}_k\mathbf{T}(\mathbf{x}_j)\|$ von k ab, da der Wert τ ebenfalls nur vom größten und kleinsten Singulärwert von $\mathbf{DF}(\mathbf{x}_j)$ abhängt. Die hierarschischen Struktur der X_k^I überträgt sich auf die $\mathcal{Z}_k$, womit

$$0 < \|\mathbf{P}_k\mathbf{T}(\mathbf{x}_j)\| \leq \|\mathbf{P}_{k+1}\mathbf{T}(\mathbf{x}_j)\| \leq 1, k = 1, \ldots$$

gilt, das heißt es existiert eine von k unabhängige obere Schranke für den Ausdruck $1/\|\mathbf{P}_k\mathbf{T}(\mathbf{x}_j)\|^2$. Somit existiert also auch in diesem Fall ein von k unabhängiges $\hat{c}_0$ mit

$$(\hat{\sigma}_{m_k}^j)^{-1} \leq \hat{c}_0.$$

Für den größten Singulärwert $\hat{\sigma}_1^j$ ergibt sich direkt

$$\hat{\sigma}_1^j = \|\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}_j)\| = \|\mathbf{V}_j^T\mathbf{DF}(\mathbf{x}_j)\mathbf{Z}\| \leq \|\mathbf{V}_j\|\sigma_1(\mathbf{x}_j) \leq c_1\|\mathbf{V}_j\|.$$

Abhängig davon, wie die $\mathbf{V}_j$ erzeugt werden, gilt entweder $\|\mathbf{V}_j\| = 1$ oder $\|\mathbf{V}_j\| \leq c_1$. Somit existiert mit $\hat{c}_1 := c_1$ bzw. $\hat{c}_1 := c_1^2$ eine k-unabhängige obere Schranke für $\hat{\sigma}_1^j$ und das Lemma ist bewiesen. □

Bemerkung 6.4.4. *Die Abschätzung für den größten Singulärwert $\hat{\sigma}_1^j$ kann wegen Bedingung (ii) aus Definition 6.4.1 auch unabhängig vom Punkt $\hat{\mathbf{x}}_j$ getroffen werden. Allgemein existiert also eine von k unabhängige Konstante $\hat{c}_1 > 0$, sodass für den größten Singulärwert $\hat{\sigma}_1$ von $\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}})$ mit $\hat{\mathbf{x}} \in \Omega_k^I$ gilt:*

$$\hat{\sigma}_1 \leq \hat{c}_1.$$

Für den kleinsten und den größten Singulärwert von $\mathbf{D}\hat{\mathbf{F}}_I$ in den Interpolationsknoten existieren also (untere bzw. obere) Schranken, die unabhängig von k gewählt werden können. Das folgende Lemma zeigt nun, dass die interpolierte Funktion einer k-unabhängigen Lipschitzbedingung genügt.

Lemma 6.4.5. *Sei $\mathbf{F}$ nach Voraussetzung 6.2.1, sowie eine Folge von zulässigen Interpolationen $\{\Omega_k^I, X_k^I, W_k^I\}$ von $\mathbf{c}$ nach Definition 6.4.1 gegeben. Es sei weiterhin eine Folge von Ansatzräumen $\mathcal{Z}_k$ gegeben und die für das interpolierte Problem verwendeten Matrizen $\mathbf{V}_j$ seien wie in Satz 5.3.6 aufgebaut. Sei weiterhin $\hat{\Omega}_k^j = \{\hat{\mathbf{x}} \in \mathbb{R}^{m_k+1} : \ \mathbf{Z}_k\hat{\mathbf{x}} \in \Omega_k^j \cap \Omega_k^I\}$.*

Dann existiert ein $\hat{L} > 0$ unabhängig von k, sodass für alle $\hat{\mathbf{x}} \in \hat{\Omega}_k^j$

$$\|\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}) - \mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}_j)\| \leq \hat{L}\|\hat{\mathbf{x}} - \hat{\mathbf{x}}_j\|$$

gilt.

Beweis. Durch die Anwendung der Produktregel erhält man für $\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}})$

$$\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}) = \sum_{j=1}^{m_k+1} \left(w_j(\mathbf{Z}_k\hat{\mathbf{x}})\mathbf{V}_j^T\mathbf{DF}(\mathbf{Z}_k\hat{\mathbf{x}})\mathbf{Z}_k + \mathbf{V}_j^T\mathbf{F}(\mathbf{Z}_k\hat{\mathbf{x}})\mathbf{Z}_k^T\nabla w_j(\mathbf{Z}\hat{\mathbf{x}})^T \right).$$

Für die Differenz $\|\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}) - \mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}_j)\|$ werden unter Ausnutzung der Dreiecksungleichung die beiden Terme auf

$$\left\| \sum_{i=1}^{m_k+1} \left(w_i(\mathbf{Z}_k\hat{\mathbf{x}})\mathbf{V}_i^T\mathbf{DF}(\mathbf{Z}_k\hat{\mathbf{x}})\mathbf{Z}_k - w_i(\mathbf{Z}_k\hat{\mathbf{x}}_j)\mathbf{V}_i^T\mathbf{DF}(\mathbf{Z}_k\hat{\mathbf{x}}_j)\mathbf{Z}_k \right) \right\| \text{ und}$$
$$\left\| \sum_{i=1}^{m_k+1} \left(\mathbf{V}_i^T\mathbf{F}(\mathbf{Z}_k\hat{\mathbf{x}})\mathbf{Z}_k^T\nabla w_i(\mathbf{Z}_k\hat{\mathbf{x}})^T - \mathbf{V}_i^T\mathbf{F}(\mathbf{Z}_k\hat{\mathbf{x}}_j)\mathbf{Z}_k^T\nabla w_i(\mathbf{Z}_k\hat{\mathbf{x}}_j)^T \right) \right\|$$

abgeschätzt. Zunächst betrachtet man die Terme zur Vereinfachung im volldimensionalen System in $\tilde{\Omega}_k^j := \Omega_k^j \cap \Omega_k^I$.Diese Gebiet enthält alle $\mathbf{Z}_k\hat{\mathbf{x}}$ mit $\hat{\mathbf{x}} \in \hat{\Omega}_j^k$, daher gelten hier getroffene Abschätzungen auch für das Ausgangsgebiet. Die zu untersuchenden Terme sind dann

$$A := \left\| \sum_{i=1}^{m_k+1} \left(w_i(\mathbf{x})\mathbf{V}_i^T\mathbf{DF}(\mathbf{x})\mathbf{Z}_k - w_i(\mathbf{x}_j)\mathbf{V}_i^T\mathbf{DF}(\mathbf{x}_j)\mathbf{Z}_k \right) \right\| \text{ und}$$
$$B := \left\| \sum_{i=1}^{m_k+1} \left(\mathbf{V}_i^T\mathbf{F}(\mathbf{x})\mathbf{Z}_k^T\nabla w_i(\mathbf{x})^T - \mathbf{V}_i^T\mathbf{F}(\mathbf{x}_j)\mathbf{Z}_k^T\nabla w_i(\mathbf{x}_j)^T \right) \right\|.$$

Jeder Summand von A wird nun um den Wert
$0 = (w_i(\mathbf{x}_j)\mathbf{V}_i^T\mathbf{DF}(\mathbf{x})\mathbf{Z}_k - w_i(\mathbf{x}_j)\mathbf{V}_i^T\mathbf{DF}(\mathbf{x})\mathbf{Z}_k)$ erweitert und man erhält

$$A = \left\| \sum_{i=1}^{m_k+1} \Bigg((w_i(\mathbf{x}) - w_i(\mathbf{x}_j))\mathbf{V}_i^T\mathbf{DF}(\mathbf{x})\mathbf{Z}_k \right.$$
$$\left. + w_i(\mathbf{x}_j)\mathbf{V}_i^T(\mathbf{DF}(\mathbf{x}) - \mathbf{DF}(\mathbf{x}_j))\mathbf{Z}_k \Bigg) \right\|.$$

Diese Summe wird mittels Dreiecksungleichung in $A \leq A_1 + A_2$ mit

$$A_1 := \left\| \sum_{i=1}^{m_k+1} \left((w_i(\mathbf{x}) - w_i(\mathbf{x}_j))\mathbf{V}_i^T\mathbf{DF}(\mathbf{x})\mathbf{Z}_k \right) \right\|,$$
$$A_2 := \left\| \sum_{i=1}^{m_k+1} \left(w_i(\mathbf{x}_j)\mathbf{V}_i^T(\mathbf{DF}(\mathbf{x}) - \mathbf{DF}(\mathbf{x}_j))\mathbf{Z}_k \right) \right\|$$

aufgeteilt. Wendet man den Mittelwertsatz

$$(w_i(\mathbf{x}) - w_i(\mathbf{x}_j)) = \int_0^1 \nabla w_i(t\mathbf{x} + (1-t)\mathbf{x}_j)^T(\mathbf{x} - \mathbf{x}_j)dt$$

auf A_1 an, ergibt sich

$$\begin{aligned} A_1 &= \left\| \sum_{i=1}^{m_k+1} \left(\int_0^1 \nabla w_i(t\mathbf{x} + (1-t)\mathbf{x}_j)^T (\mathbf{x} - \mathbf{x}_j) dt \mathbf{V}_i^T \mathbf{DF}(\mathbf{x}) \mathbf{Z}_k \right) \right\| \\ &\leq \|\mathbf{DF}(\mathbf{x})\| \left\| \sum_{i=1}^{m_k+1} \left(\int_0^1 \nabla w_i(t\mathbf{x} + (1-t)\mathbf{x}_j)^T (\mathbf{x} - \mathbf{x}_j) dt \mathbf{V}_i^T \right) \right\|. \end{aligned}$$

Zu diesem Summanden addiert man

$$\mathbf{0} = - \sum_{i=1}^{m_k+1} \left(\int_0^1 \nabla w_i(t\mathbf{x} + (1-t)\mathbf{x}_j)^T (\mathbf{x} - \mathbf{x}_j) dt \mathbf{V}_j^T \right),$$

der deshalb $\mathbf{0}$ ist, da für alle $\mathbf{x} \in \Omega_k^I$ gilt: $\sum_{i=1}^{m_k+1} \nabla w_i(\mathbf{x}) = \nabla \left(\sum_{i=1}^{m_k+1} w_i(\mathbf{x}) \right) = \nabla(1) = \mathbf{0}$. Zusammen mit $\|\mathbf{DF}(\mathbf{x})\| \leq c_1$ führt dies zu

$$\begin{aligned} A_1 &\leq c_1 \left\| \sum_{i=1}^{m_k+1} \left(\int_0^1 \nabla w_i(t\mathbf{x} + (1-t)\mathbf{x}_j)^T (\mathbf{x} - \mathbf{x}_j) dt (\mathbf{V}_i - \mathbf{V}_j)^T \right) \right\| \\ &\leq c_1 \sum_{i=1}^{m_k+1} \|\nabla w_i(\mathbf{x}^*)\| \|\mathbf{V}_i - \mathbf{V}_j\| \|\mathbf{x} - \mathbf{x}_j\| \end{aligned}$$

mit $x^* = t^*\mathbf{x} + (1-t^*)\mathbf{x}_j \in \tilde{\Omega}_k^j$, wobei $t^* \in [0,1]$ der Wert ist, an dem die Funktion

$$t \mapsto \sum_{i=1}^{m_k+1} \|\nabla w_i(t\mathbf{x} + (1-t)\mathbf{x}_j)\| \|\mathbf{V}_i - \mathbf{V}_j\|$$

einen maximalen Wert annimmt. Nach Bedingung (ii) aus Definition 6.4.1 folgt nun, dass wegen $\mathbf{x}^* \in \tilde{\Omega}_k^j$

$$\nabla w_i(\mathbf{x}^*) = \mathbf{0}, \ i \notin \{j-1, j, j+1\}$$

gilt. Es ergibt sich dann für A_1

$$A_1 \leq c_1 \sum_{k=-1}^{1} \|\nabla w_{j+k}(\mathbf{x}^*)\| \|\mathbf{V}_{j+k} - \mathbf{V}_j\| \|\mathbf{x} - \mathbf{x}_j\|.$$

Nach Bedingung (v) von Definition 6.4.1 und 5.3.6 existiert dann eine Konstante L_V unabhängig von k, sodass

$$\|\mathbf{V}_{j+k} - \mathbf{V}_j\| \leq L_V \|\mathbf{x}_{j+l} - \mathbf{x}_j\|, \ l \in \{-1, 0, 1\}$$

gilt. Aus Bedingung (iii) von Definition 6.4.1 folgt nun $\|\nabla w_{j+l}(\mathbf{x}^*)\| \leq c_w h^{-1}$. Zusammen mit der Ungleichung $\|\mathbf{x}_{i+1}-\mathbf{x}_i\| = \|\mathbf{c}(s_{i+1})-\mathbf{c}(s_i)\| \leq |s_{i+1}-s_i| \leq h$ ergibt sich so

$$A_1 \leq 3c_1c_wL_V\|\mathbf{x}-\mathbf{x}_j\|. \tag{6.7}$$

Wegen $w_i(\mathbf{x}_j) = \delta_{ij}$ ergibt sich für den Term A_2 zunächst

$$A_2 = \|\mathbf{V}_j^T\left(\mathbf{DF}(\mathbf{x}) - \mathbf{DF}(\mathbf{x}_j)\right)\mathbf{Z}_k\|.$$

Da $\mathbf{DF}$ nach Voraussetzung 6.2.1 Lipschitz-stetig ist, existiert eine Konstante $L_{\mathbf{F}} > 0$ mit $\|\mathbf{DF}(\mathbf{x}) - \mathbf{DF}(\mathbf{x}_j)\| \leq L_{\mathbf{F}}\|\mathbf{x}-\mathbf{x}_j\|$. Abhängig von dem Verfahren, dass zum Aufbau der $\mathbf{V}_j$ verwendet wurde, gilt außerdem entweder $\|\mathbf{V}_j\| \leq c_1$ (Aufbau nach Satz 5.2.4) oder $\|\mathbf{V}_j\| = 1$ (Aufbau nach Bemerkung 5.2.5). Mit $\hat{c}_1 := \max\{1, c_1\}$ erhält man dann

$$A_2 \leq \|\mathbf{V}_j\|\|\mathbf{DF}(\mathbf{x}) - \mathbf{DF}(\mathbf{x}_j)\|\|\mathbf{Z}_k\| \leq c_V L_{\mathbf{F}}\|\mathbf{x}-\mathbf{x}_j\|.$$

Analog zum Term A lässt sich B ebenfalls aufsplitten in

$$B_1 := \left\| \sum_{i=1}^{m_k+1} \left(\mathbf{V}_i^T(\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_j))\mathbf{Z}_k^T\nabla w_i(\mathbf{x})^T\right)\right\| \text{ und}$$
$$B_2 := \left\| \sum_{i=1}^{m_k+1} \left(\mathbf{V}_i^T\mathbf{F}(\mathbf{x}_j)\mathbf{Z}_k^T(\nabla w_i(\mathbf{x}) - \nabla w_i(\mathbf{x}_j))^T\right)\right\|.$$

B_2 verschwindet direkt wegen $\mathbf{F}(\mathbf{x}_j) = \mathbf{0}$. Für B_1 ergibt sich eine zu A_1 analoge Betrachtung und unter Anwendung des Mittelwertsatzes auf $\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_j)$

$$B_1 \leq c_1 \sum_{i=1}^{m_k+1} \|\mathbf{V}_i - \mathbf{V}_j\|\|\nabla w_i(\mathbf{x})\|\|\mathbf{x}-\mathbf{x}_j\| \leq 3c_1c_wL_{\mathbf{V}}\|\mathbf{x}-\mathbf{x}_j\|$$

Wählt man nun die Konstante $\hat{L} := (6c_1c_wL_{\mathbf{V}} + \hat{c}_1L_{\mathbf{F}})$ so gilt unabhängig von k

$$\|\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}) - \mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}_j)\| \leq \hat{L}\|\mathbf{x}-\mathbf{x}_j\|, j = 0, \ldots, m_k$$

und damit auch

$$\|\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}) - \mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}_j)\| \leq \hat{L}\|\hat{\mathbf{x}}-\hat{\mathbf{x}}_j\|, j = 0, \ldots, m_k.$$

□

Lemma 6.4.6. *Sei* $\mathbf{F}$ *nach Voraussetzung 6.2.1 und* $\{\Omega_k^I, X_k^I, W_k^I\}$ *eine Folge zulässiger Interpolationen von* $\mathbf{c}$ *nach Definition 6.4.1, sowie der Ansatzraum* $\mathcal{Z}_k$ *ein Lagrange-Ansatzraum mit* $\mathcal{Z}_k := span\{\mathbf{x}_j,\ j = 1, \ldots, m_k + 1,\ \mathbf{x}_j \in X_k^I\}$. *Seien weiterhin Matrizen* $\mathbf{V}_j,\ j = 1, \ldots, m_k + 1$ *mittels eines der beiden Verfahren aus Kapitel 5 aufgebaut.*

Sei $\hat{\mathbf{r}}_j := (\hat{\mathbf{x}}_{j+1} - \hat{\mathbf{x}}_j)/\|\hat{\mathbf{x}}_{j+1} - \hat{\mathbf{x}}_j\|,\ j = 1, \ldots, m_k$ *gegeben, sowie mit* $\hat{\mathbf{T}}(\hat{\mathbf{x}}_j)$ *der Tangentialvektor der interpolierten Reduktion in* $\hat{\mathbf{x}}_j$ *bezeichnet.*

Dann existiert ein $\tilde{k} \in \mathbb{N}$, *sodass für alle* $k \geq \tilde{k}$ *eine Konstante* $\hat{c}_2 > 0$ *unabhängig von* k *existiert mit*

$$|\langle \hat{\mathbf{T}}(\hat{\mathbf{x}}_j), \hat{\mathbf{r}}_j \rangle| \geq \hat{c}_2,\ j = 1, \ldots, m_k.$$

Beweis. Seien die Vektoren $\mathbf{r}_j := (\mathbf{x}_{j+1} - \mathbf{x}_j)/\|\mathbf{x}_{j+1} - \mathbf{x}_j\|,\ j = 1, \ldots, m_k$ definiert. Die folgenden Aussagen werden ohne Einschränkung für $\mathbf{x}_1$ und $\mathbf{r}_1$, bzw. $\hat{\mathbf{x}}_1$ und $\hat{\mathbf{r}}_1$ hergeleitet. Aus $X_k \subset \mathcal{Z}_k$ folgt zunächst $\mathbf{r}_1 \in \mathcal{Z}_k$. Sei nun mit $\mathbf{P}$ der Projektor auf $\text{span}\{\mathbf{Z}_k \hat{\mathbf{T}}(\hat{\mathbf{x}}_1), \mathbf{r}_1\}$ bezeichnet, wobei $\mathbf{Z}_k$ eine Matrix darstellt, deren Spalten eine Orthonormalbasis von $\mathcal{Z}_k$ bilden.

Für den Winkel α_1 zwischen $\mathbf{Z}_k \hat{\mathbf{T}}(\mathbf{x}_1)$ und $\mathbf{PT}(\mathbf{x}_1)$ ergibt sich

$$\begin{aligned}\cos(\alpha_1) &= \frac{|\langle \mathbf{Z}_k \hat{\mathbf{T}}(\hat{\mathbf{x}}_1), \mathbf{PT}(\mathbf{x}_1) \rangle|}{\|\mathbf{PT}(\mathbf{x}_1)\|} = \frac{|\langle \mathbf{Z}_k \hat{\mathbf{T}}(\hat{\mathbf{x}}_1), \mathbf{T}(\mathbf{x}_1) \rangle|}{\|\mathbf{PT}(\mathbf{x}_1)\|} \\ &\geq \frac{|\langle \mathbf{Z}_k \hat{\mathbf{T}}(\hat{\mathbf{x}}_1), \mathbf{T}(\mathbf{x}_1) \rangle|}{\|\mathbf{P}_k \mathbf{T}(\mathbf{x}_1)\|},\end{aligned}$$

wobei $\mathbf{P}_k$ den orthogonalen Projektor auf $\mathcal{Z}_k$ bezeichnet. Nach Voraussetzung 6.2.1 und Lemma 5.2.3 existiert ein von den Singulärwerten von $\mathbf{DF}$ abhängiges (und damit k-unabhängiges) $q > 0$ mit

$$|\langle \mathbf{Z}_k \hat{\mathbf{T}}(\hat{\mathbf{x}}_j), \mathbf{T}(\mathbf{x}_j) \rangle| \geq \sqrt{\frac{q^2}{1+q^2}} \|\mathbf{P}_k \mathbf{T}(\mathbf{x}_j)\|.$$

Somit ergibt sich mit $\tau := \sqrt{q^2/(1+q^2)}$

$$\cos(\alpha_1) \geq \tau.$$

Da $0 < \tau < 1$ gilt, folgt daraus $\alpha_1 < \pi/2$.

Sei mit α_2 der Winkel zwischen $\mathbf{PT}(\mathbf{x}_1)$ und $\mathbf{r}_1$ bezeichnet. Für diesen ergibt sich

$$\cos(\alpha_2) = \frac{|\langle \mathbf{PT}(\mathbf{x}_1), \mathbf{r}_1 \rangle|}{\|\mathbf{PT}(\mathbf{x}_1)\|} = \frac{|\langle \mathbf{T}(\mathbf{x}_1), \mathbf{r}_1 \rangle|}{\|\mathbf{PT}(\mathbf{x}_1)\|} \geq |\langle \mathbf{T}(\mathbf{x}_1), \mathbf{r}_1 \rangle|.$$

Für $k \to \infty$ gilt $\mathbf{r}_1 \to \mathbf{T}(\mathbf{x}_1)$, somit existiert ein $\tilde{k} \in \mathbb{N}$, sodass für $k \geq \tilde{k}$ und den Winkel α_2

$$\arccos(\tau) + \alpha_2 \leq \arccos\left(\frac{\tau}{2}\right)$$

gilt. Der maximale Wert, den der Winkel zwischen $\hat{\mathbf{T}}(\hat{\mathbf{x}}_1)$ und $\hat{\mathbf{r}}_1$ annehmen kann, ist $\alpha_1 + \alpha_2$. Mit $\hat{c}_2 := \tau/2$ ergibt sich so

$$|\langle \hat{\mathbf{T}}(\hat{\mathbf{x}}_1), \hat{\mathbf{r}}_1 \rangle| \geq \cos(\alpha_1 + \alpha_2) \geq \cos\left(\arccos\left(\frac{\tau}{2}\right)\right) = \frac{\tau}{2} = \hat{c}_2.$$

□

Mit Hilfe der vorangegangenen Lemmata wird nun die Hauptaussage, das für eine zulässige Interpolation bei genügend großem k die interpolierte Reduktion eine Lösungskurve durch die Interpolationsknoten besitzt, bewiesen.

Da die interpolierte Reduktion in den Knoten mit einer lokalen Reduktion übereinstimmt, ist die Existenz einer Lösungskurve in den einzelnen Interpolationsknoten sicher. Die Grundidee des Beweises besteht in der Sicherung der Existenz einer Lösung im Schnitt der Träger der Interpolationsknoten, durch den Nachweis, dass die benötigten Voraussetzungen für Satz 2.2.3 für genügend große k (bzw. genügend kleine h_k) stets erfüllt sind.

Satz 6.4.7. *Sei* $\mathbf{F}$ *nach Voraussetzung 6.2.1, sowie eine Folge von zulässigen Interpolationen* $\{\Omega_k^I, X_k^I, W_k^I\}$ *von* $\mathbf{c}$ *nach Definition 6.4.1 gegeben. Der Ansatzraum* $\mathcal{Z}_k$ *sei über* $\mathcal{Z}_k := span\{\mathbf{x}_j,\ j = 1, \ldots, m_k + 1,\ \mathbf{x}_j \in X_k^I\}$ *und die Matrizen* $\mathbf{V}_j$ *nach einem der Verfahren aus Kapitel 5.2 aufgebaut.*

Dann existiert ein $\tilde{k} \in \mathbb{N}$*, sodass das interpolierte Problem*

$$\hat{\mathbf{F}}_I(\hat{\mathbf{x}}) = \mathbf{0}$$

für $k \geq \tilde{k}$ *eine Lösungskurve besitzt, die durch alle Interpolationsknoten* $\hat{\mathbf{x}}_j,\ j = 1, \ldots, m_k + 1$ *verläuft.*

Beweis. In Abbildung 6.5 sind die im Folgenden auftretenden Größen zwecks besseren Verständnisses skizziert.

Ohne Einschränkung werden die zwei Punkte $\hat{\mathbf{x}}_1$ und $\hat{\mathbf{x}}_2$ betrachtet. Nach Lemma 6.4.3 existieren $\hat{c}_0, \hat{c}_1 > 0$ unabhängig von k, sodass

$$(\hat{\sigma}_{m_k}^j)^{-1} \leq \hat{c}_0, \text{ und } \hat{\sigma}_1^j \leq \hat{c}_1$$

gilt. Nach Lemma 6.4.6 existiert, falls $\tilde{k}$ groß genug ist, außerdem ein k-unabhängiges $\hat{c}_2 > 0$ mit

$$\langle \hat{\mathbf{T}}(\hat{\mathbf{x}}_1), \hat{\mathbf{r}}_1 \rangle \geq \hat{c}_2,$$

wobei $\hat{\mathbf{T}}(\hat{\mathbf{x}}_1)$ der Tangentialvektor von $\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}_1)$ und $\hat{\mathbf{r}}_1 := (\hat{\mathbf{x}}_2 - \hat{\mathbf{x}}_1)/\|\hat{\mathbf{x}}_2 - \hat{\mathbf{x}}_1\|$ ist.

Sei jetzt $\hat{\mathbf{x}}^*$ ein beliebiger Punkt auf der Strecke $\hat{\mathbf{x}}_1 + t\hat{\mathbf{r}}_1,\ t \in [0, \|\hat{\mathbf{x}}_2 - \hat{\mathbf{x}}_1\|]$ und $\mathbf{x}^* := \mathbf{Z}_k\hat{\mathbf{x}}^*$, wobei die Spalten der Matrix $\mathbf{Z}_k$ eine Orthonormalbasis von $\mathcal{Z}_k$ bilden.

Für $\mathbf{F}_I(\hat{\mathbf{x}}^*)$ werden die Bedingungen von Satz 2.2.3 betrachtet und gezeigt, dass in der Nähe von $\hat{\mathbf{x}}^*$ eine bezüglich $\hat{\mathbf{r}}_1$ parametrisierte Lösung von $\hat{\mathbf{F}}_I(\hat{\mathbf{x}}) = \mathbf{0}$ existiert.

Sei nun der Fall $\hat{\mathbf{x}}^* \to \hat{\mathbf{x}}_1$, bzw. $\mathbf{x}^* \to \mathbf{x}_1$ betrachtet. Ist $\mathbf{V}^* := w_1(\mathbf{x}^*)\mathbf{V}_1 + w_2(\mathbf{x}^*)\mathbf{V}_2$, dann gilt nach Bedingung (v) von Definition 6.4.1 $\mathbf{V}^* \to \mathbf{V}_1$ und damit $\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}^*) \to \mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}_1)$. Aufgrund dieser Stetigkeit und der Unabhängigkeit der Größen $\hat{c}_0, \hat{c}_1$ und $\hat{c}_2$ von k, existiert ein $\tilde{k} \in \mathbb{N}$, sodass für $k \geq \tilde{k}$ der Punkt $\mathbf{x}^*$ so nah an $\mathbf{x}_1$ liegt, dass die Abschätzungen

$$\begin{aligned}(\hat{\sigma}^*_{m_k})^{-1} &\leq \sqrt{2}\hat{c}_0,\\ \langle \hat{\mathbf{T}}(\hat{\mathbf{x}}^*), \hat{\mathbf{r}}_1\rangle &\geq \frac{1}{\sqrt{2}}c_2\end{aligned}$$

gelten, wobei mit $\hat{\sigma}^*_{m_k}$ der kleinste Singulärwert von $\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}^*)$ bezeichnet ist. Des Weiteren ergibt sich nach Bemerkung 6.4.4 für den größten Singulärwert $\hat{\sigma}^*_1$ von $\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}^*)$

$$\hat{\sigma}^*_1 = \|\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}^*)\| \leq \hat{c}_1.$$

Sei nun $M^* := \sqrt{2}\max(2\hat{c}_0\hat{c}_2^{-1}, 1 + 2\hat{c}_0\hat{c}_1\hat{c}_2^{-1})$. Dieser Wert ist, da nur abhängig von $\hat{c}_0, \hat{c}_1$ und $\hat{c}_2$ unabhängig von k. Die Konstante β^* wird nun so bestimmt, dass für alle $\hat{\mathbf{x}} \in B(\hat{\mathbf{x}}^*; \beta^*)$

$$\|\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}) - \mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}^*)\| \leq \frac{1}{2M^*} \tag{6.8}$$

gilt. Nach Lemma 6.4.5 existiert eine k-unabhängige Konstante $\hat{L} > 0$, sodass für alle $\hat{\mathbf{x}} \in \hat{\Omega}^1_k$

$$\|\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}) - \mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}_1)\| \leq \hat{L}\|\hat{\mathbf{x}} - \hat{\mathbf{x}}_1\|$$

gilt. Ohne Einschränkung geht man davon aus, dass $\mathbf{x}^* \in \Omega^1_k$ liegt. Ist dies nicht der Fall, gilt $\mathbf{x}^* \in \Omega^2_k$ und die folgenden Abschätzungen können mit $\hat{\mathbf{x}}_2$ statt $\hat{\mathbf{x}}_1$ durchgeführt werden. Es ergibt sich

$$\begin{aligned}\|\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}) - \mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}^*)\| &\leq \|\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}) - \mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}_1)\| + \|\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}^*) - \mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}_1)\|\\ &\leq \hat{L}(\|\hat{\mathbf{x}} - \hat{\mathbf{x}}_1\| + \|\hat{\mathbf{x}}^* - \hat{\mathbf{x}}_1\|)\\ &= \hat{L}(\|\mathbf{x} - \mathbf{x}_1\| + \|\mathbf{x}^* - \mathbf{x}_1\|) \leq \hat{2}\hat{L}h_k.\end{aligned}$$

Ist k also groß genug (und h_k damit klein genug), ist die Bedingung (6.8) für $\hat{\mathbf{x}} \in B(\hat{\mathbf{x}}^*; \beta)$ immer erfüllt. Da die Funktion $\mathbf{F}_I$ nur innerhalb von Ω^I_k definiert ist, muss β^* zudem so klein sein, dass $B(\mathbf{x}^*; \beta^*) \subset \Omega^1_k \cup \Omega^2_k$ gilt. Nach Bedingung (vi) von Definition 6.4.1 existiert ein $\tau \in (0, 1)$, sodass $B(\mathbf{x}^*; \tau h_k) \subset \Omega^1_k \cup \Omega^2_k$ gilt. Daher sei β^* nun als

$$\beta^* := \tau h_k$$

und α^* und δ^* als

$$\alpha^* = \delta^* := \frac{\beta^*}{2\sqrt{2}M^*} = \frac{\tau h_k}{2\sqrt{2}M^*}$$

gewählt. Nach Satz 2.2.3 existiert dann eine Lösungskurve in der Nähe von $\hat{\mathbf{x}}^*$, wenn die Bedingung

$$\|\hat{\mathbf{F}}_I(\hat{\mathbf{x}}^*)\| \leq \delta$$

erfüllt ist. Es gilt zunächst

$$\begin{aligned}\|\hat{\mathbf{F}}_I(\hat{\mathbf{x}}^*)\| &= \|(w_1(\mathbf{x}^*)\mathbf{V}_1 + w_2(\mathbf{x}^*)\mathbf{V}_2)^T\mathbf{F}(\mathbf{x}^*)\| \\ &\leq (\|\mathbf{V}_1\| + \|\mathbf{V}_2\|)\|\mathbf{F}(\mathbf{x}^*)\|.\end{aligned}$$

Unabhängig von dem Verfahren, dass zum Aufbau der $\mathbf{V}_j$ verwendet wurde, gilt $\|\mathbf{V}_j\| \leq \hat{c}_1$.

Da $\mathcal{Z}_k$ ein Lagrange-Raum ist, gilt $\mathbf{x}_1 = \mathbf{c}(s_1)$ und $\mathbf{x}_2 = \mathbf{c}(s_2)$. Die Funktion $\mathbf{I}(s) = \mathbf{x}_1 + s\mathbf{r}_1$ mit $s \in [0, \|\mathbf{x}_2 - \mathbf{x}_1\|]$ stellt die lineare Interpolation von $\mathbf{c}$ durch die Knoten $\mathbf{x}_1$ und $\mathbf{x}_2$ dar. Nach Bedingung (iii) von Voraussetzung 6.2.1 existiert ein $c_\mathbf{c} > 0$ mit $\|\mathbf{c}''(s)\| \leq c_\mathbf{c}$. Für alle $s \in [0, \|\mathbf{x}_2 - \mathbf{x}_1\|]$ gilt dann

$$\|\mathbf{I}(s) - \mathbf{c}(s)\| \leq c_\mathbf{c} s(s - \|\mathbf{x}_2 - \mathbf{x}_1\|) \leq c_\mathbf{c}\|\mathbf{x}_2 - \mathbf{x}_1\|^2 \leq c_\mathbf{c} h_k^2.$$

Sei nun $s^* \in [0, \|\mathbf{x}_2 - \mathbf{x}_1\|]$ so, dass $\mathbf{x}^* = \mathbf{I}(s^*)$ gilt, dann ergibt sich

$$\begin{aligned}\|\hat{\mathbf{F}}_I(\hat{\mathbf{x}}^*)\| &\leq 2\hat{c}_1\|\mathbf{F}(\mathbf{x}^*)\| = 2\hat{c}_1\|\mathbf{F}(\mathbf{x}^*) - \mathbf{F}(\mathbf{c}(s^*))\| \leq 2\hat{c}_1 c_1\|\mathbf{x}^* - \mathbf{c}(s^*)\| \\ &= 2\hat{c}_1 c_1\|\mathbf{I}(s^*) - \mathbf{c}(s^*)\| \leq 2\hat{c}_1 c_1 c_\mathbf{c} h_k^2.\end{aligned}$$

Sei dann $\tilde{k}$ so groß, dass für $k \geq \tilde{k}$

$$h_k \leq \frac{\tau}{4\sqrt{2}M^*\hat{c}_1 c_1 c_\mathbf{c}}$$

gilt, dann ist die Bedingung $\|\hat{\mathbf{F}}_I(\hat{\mathbf{x}}^*)\| \leq \delta$ erfüllt. Somit existiert eine Lösungskurve

$$\hat{\mathbf{c}}^*(s) = \hat{\mathbf{x}}^* + s\hat{\mathbf{r}}_1 + \hat{\mathbf{Y}}\hat{\mathbf{g}}^*(s)$$

mit einer eindeutigen stetig differenzierbaren Funktion
$\hat{\mathbf{g}}^* : B(0;\alpha^*) \to B(\mathbf{0};\tau h_k)$ und $\hat{\mathbf{F}}_I(\hat{\mathbf{c}}^*(s)) = \mathbf{0}$. Die Spalten der Matrix $\hat{\mathbf{Y}}$ enthält dabei eine Orthonormalbasis von $R(\hat{\mathbf{r}}_1)^\perp$.

Eine solche Lösungskurve existiert für jeden einzelnen Punkt auf der Strecke zwischen $\hat{\mathbf{x}}_1$ und $\hat{\mathbf{x}}_2$. Es bleibt nun noch zu zeigen, dass diese verbunden sind. Seien dazu $\mathbf{x}_0^*$ und $\mathbf{x}_1^*$ zwei Punkte auf dieser Strecke. Man beachte, dass α^*

und β^* nicht von der Wahl des Punktes auf der Strecke zwischen $\hat{\mathbf{x}}_1$ und $\hat{\mathbf{x}}_2$ abhängen. Daher existieren zwei Lösungkurven

$$\hat{\mathbf{c}}_i^* : \begin{cases} B(0;\alpha^*) & \to \mathbb{R}^{m+1} \\ s & \mapsto \hat{\mathbf{x}}_i^* + s\mathbf{r}_1 + \hat{\mathbf{Y}}\hat{\mathbf{g}}_i^*(s) \end{cases}, \quad i = 0,1$$

und $\hat{\mathbf{F}}_I(\hat{\mathbf{c}}_i^*(s)) = \mathbf{0}$. Die Kurve $\hat{\mathbf{c}}_1^*(s)$ lässt sich nun auch bezüglich $\hat{\mathbf{x}}_0^*$ schreiben als

$$\hat{\mathbf{c}}_1^*(s) = \hat{\mathbf{x}}_0^* + s\mathbf{r}_1 + \hat{\mathbf{Y}}\hat{\mathbf{g}}_1^*(s),\ s \in B(\|\hat{\mathbf{x}}_2 - \hat{\mathbf{x}}_1\|;\alpha^*)$$

Ist nun $\|\hat{\mathbf{x}}_2 - \hat{\mathbf{x}}_1\|$ klein genug, existiert ein $\bar{s} \in B(0;\alpha^*) \cap B(\|\hat{\mathbf{x}}_2 - \hat{\mathbf{x}}_1\|;\alpha^*)$. Aufgrund des Bildbereiches von $\hat{\mathbf{g}}_1^*$ gilt für den Wert $\bar{\mathbf{y}}_1 := \hat{\mathbf{g}}_1^*(\bar{s})$ dann

$$\|\bar{\mathbf{y}}_1\| \leq \tau h_k$$

und damit $\bar{\mathbf{y}}_1 \in B(\mathbf{0};\tau h_k)$. Wegen der Eindeutigkeit der Funktionen $\hat{\mathbf{g}}_0^*$ und $\hat{\mathbf{g}}_1^*$ ergibt sich $\hat{\mathbf{g}}_0^*(\bar{s}) = \bar{\mathbf{y}}_1$ und damit

$$\begin{aligned} \hat{\mathbf{c}}_1^*(\bar{s}) &= \hat{\mathbf{x}}_1 + \bar{s}\hat{\mathbf{r}}_1 + \hat{\mathbf{Y}}\hat{\mathbf{g}}_1^*(\bar{s}) = \hat{\mathbf{x}}_1 + \bar{s}\hat{\mathbf{r}}_1 + \hat{\mathbf{Y}}\bar{\mathbf{y}}_1 \\ &= \hat{\mathbf{x}}_1 + \bar{s}\hat{\mathbf{r}}_1 + \hat{\mathbf{Y}}\hat{\mathbf{g}}_0^*(\bar{s}) = \hat{\mathbf{c}}_0^*(\bar{s}). \end{aligned}$$

Somit sind die beiden Lösungkurven verbunden. Dies lässt sich auf jedes beliebige Punktepaar auf der Strecke zwischen $\hat{\mathbf{x}}_1$ und $\hat{\mathbf{x}}_2$ anwenden, womit die Existenz einer Lösungskurve, die zudem durch die Punkte $\hat{\mathbf{x}}_1$ und $\hat{\mathbf{x}}_2$ verläuft, bewiesen ist.

□

Bemerkung 6.4.8. *Für die in diesem Kapitel getroffenen Aussagen ist es nicht zwingend notwendig, dass die Anzahl der Interpolationsknoten stets mit der Dimension des Lagrange-Ansatzraumes $\mathcal{Z}$ überein stimmt. Liegen weitere Punkte der volldimensionalen Lösungskurve $\mathbf{c}$ bereits im Ansatzraum, können diese als zusätzliche Interpolationsknoten herangezogen werden. Die Anzahl der Knoten kann also erhöht werden, ohne die Güte des Ansatzraumes zu verändern.*

6.4.2 Interpolation mittels inexakter Knoten

Die im vorherigen Kapitel betrachtete Reduktion verwendet als Ansatzraum einen Lagrange-Raum, der über die Interpolationsknoten aufgespannt wird. Nachteil hierbei ist, dass ein Hinzufügen zusätzlicher Knoten (sollten diese nicht bereits in $\mathcal{Z}$ liegen) die Dimension des Ansatzraumes erhöht. Des Weiteren lässt sich das Verfahren nicht auf einen mittels POD erzeugten Ansatzraum anwenden (vergleiche Kapitel 4.2).

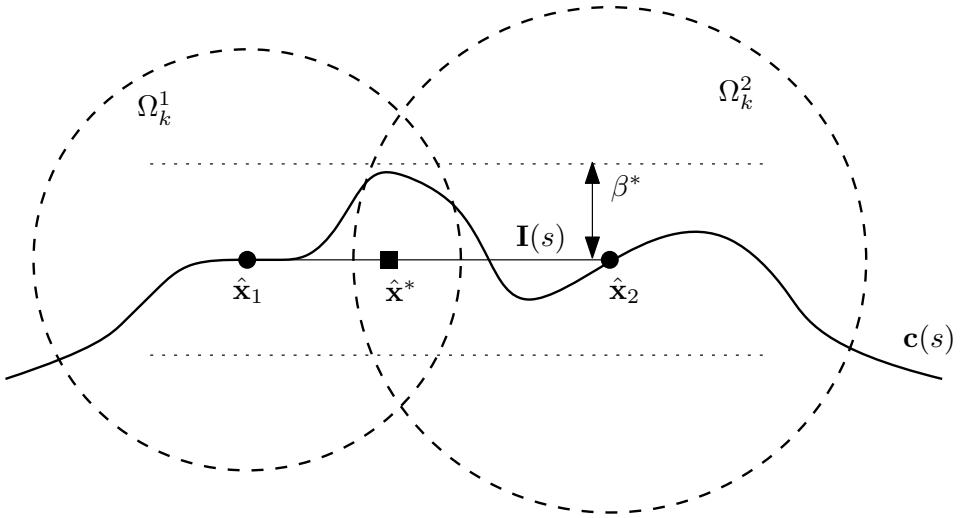

Abbildung 6.5: Skizze zum Beweis des Satzes 6.4.7

Ziel dieses Kapitels ist es, eine Interpolation für den Fall, dass die Interpolationsknoten keine Lösungen von $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ darstellen, zu entwickeln. Die Bedingung $\mathbf{F}(\mathbf{x}_i) = \mathbf{0}$ wird dabei ersetzt durch eine Beschränkung der Norm des Funktionswertes von $\mathbf{F}$ in den Interpolationsknoten. Diese sollen also nur noch "in der Nähe" der Lösungskurve liegen, daher wird hier von inexakten Knoten gesprochen. Des Weiteren stimmt der durch die Interpolationsknoten aufgespannte Raum nun nicht mehr mit dem Ansatzraum überein, was zu einer größeren Flexibilität beim Aufstellen der Interpolation führt.

Im Folgenden wird zunächst definiert, was unter einer zulässigen Folge von Interpolationen mit inexakten Knoten zu verstehen ist.

Definition 6.4.9. Sei $\mathbf{F}$ nach Voraussetzung 6.2.1, sowie eine Folge von Ansatzräumen $\mathcal{Z}_k$ der Dimension m_k+1 mit $\mathcal{Z}_k \subset \mathcal{Z}_{k+1}$ sowie der orthogonale Projektor $\mathbf{P}_k$ auf diese Räume gegeben. Sei $X_k^I := \{\mathbf{P}_k\mathbf{c}(s_j),\ s_j \in S, j = 1, \ldots, d_k\}$ eine Folge von Interpolationssystemen. Für die Menge X_k^I sei der Wert $h_k > 0$ mit $h_k := \max\{|s_{j+1} - s_j|,\ j = 1, \ldots, d_k - 1\}$ definiert und es existiere eine Konstante $c_h > 0$ mit $|s_{j+1} - s_j| \geq c_h h_k^2$ und weiterhin eine Folge von Gewichtsfunktionen $W_k^I := \{w_j : \mathbb{R}^{n+1} \to \mathbb{R},\ j = 1, \ldots, d_k\}$ mit beschränkten offenen Trägern Ω_k^j gegeben. Die Menge $\{\Omega_k^I, X_k^I, W_k^I, \mathcal{Z}_k\}$ wird dann eine Folge zulässiger inexakter Interpolationen genannt, wenn weiterhin gilt:

(i) Die Menge Ω_k^I ist abgeschlossen mit $\Omega_k^I \subset \bigcup \Omega_k^j$ mit $\mathbf{P}_k\mathbf{c} \subset \Omega_k^I$ und $\Omega_k^I \subset \mathcal{R}(\mathbf{F})$,

(ii) es gilt $\Omega_k^j \cap \Omega_k^{j+l} = \emptyset$ für $l \notin \{-1, 0, 1\}$ und es existiert eine Konstante $c_\Omega > 0$ unabhängig von k, sodass $\operatorname{diam}(\Omega_k^j) \leq c_\Omega h,\ j = 1, \ldots, d_k$ gilt,

(iii) für die Gewichtsfunktionen gilt $w \in C^1(\Omega_k^I)$, $\sum_{j=1}^{d_k} w_j(\mathbf{x}) = 1$, für alle $\mathbf{x} \in \Omega_k^I$, sowie $w_j(\mathbf{x}_i) = \delta_{ij},\ i,j = 1,\dots,d_k$. Des Weiteren seien zwei Konstante $c_w^1, c_w^2 > 0$ unabhängig von d_k gegeben mit $\|\nabla w_i(\mathbf{x})\| \leq c_w^1 h_k^{-1}$ und $\|\nabla^2 w_i(\mathbf{x})\| \leq c_w^2 h_k^{-2}$ für $i = 1,\dots,d_k$,

(iv) es existiert eine Konstante $c_s > 0$ unabhängig von d_k, sodass $h_k \leq c_s d_k^{-1}$ gilt,

(v) es existieren $\gamma_j > 0$, sodass $\mathbf{x}_{j-1}, \mathbf{x}_{j+1} \in B(\mathbf{x}_j;\gamma_j) \subset \Omega,\ j = 2,\dots,d_k - 1$ gilt und für alle $\mathbf{x} \in B(\mathbf{x}_j;\gamma_j)$ die Bedingung $\mathbf{T}(\mathbf{x})^T\mathbf{T}(\mathbf{x}_j) \neq 0$ erfüllt ist,

(vi) es existiert ein $\tau \in (0,1)$ unabhängig von k, sodass für alle $j = 1,\dots,d_k - 1$ gilt: $B(t\mathbf{x}_j + (1-t)\mathbf{x}_{j+1};\tau h) \subset \Omega_k^I,\ t \in [0,1]$,

(vii) es existiert eine Konstante $c_X > 0$ unabhängig von d_k, sodass $\|\mathbf{x}_j - \mathbf{c}(s_j)\| = \|\mathbf{P}_k\mathbf{c}(s_j) - \mathbf{c}(s_j)\| \leq c_X h_k^2,\ j = 1,\dots,d_k$ gilt.

Die Definition einer zulässigen inexakten Interpolation ist der der Lagrange-artigen Interpolation des vorherigen Kapitels also recht ähnlich. Der Hauptunterschied besteht darin, dass die volldimensionale Lösungskurve $\mathbf{c}$ nicht mehr in der Vereinigung der Träger Ω_k^I liegen muss und die Interpolationsknoten keine Lösung des ursprünglichen Gleichungssystems mehr sein müssen. Zusätzlich werden Bedingungen an die zweite Ableitung der Gewichtsfunktionen w gestellt, die von den in Kapitel 6.3 beschriebenen Funktionen erfüllt werden.

Bemerkung 6.4.10. *Bedingung (vii), die die Forderung $\mathbf{F}(\mathbf{x}_i) = \mathbf{0}$ ersetzt wird von dem in Kapitel 4.2 entwickelten POD-Raum erfüllt, so lange $m_k + 1 > \sqrt{d_k}$ gilt, da nach Satz 4.2.1 der Projektionsfehler $\|\mathbf{P}_k\mathbf{c}(s_j) - \mathbf{c}(s_j)\|$ mit einer Ordnung $(m_k+1)^{-4}$ und damit d_k^{-2}, also h_k^2 fällt.*

Ziel ist es einen zu Satz 6.4.7 ähnlichen Satz zu beweisen. Hierfür wird wieder der Satz 2.2.3 herangezogen und Abschätzungen für die dort auftretenden Größen für eine Interpolation mit inexakten Knoten gezeigt.

Analog zu 6.4.3 wird nun zunächst gezeigt, dass für eine Folge von zulässigen inexakten Interpolationen eine von k unabhängige Abschätzung für die kleinsten und größten Singulärwerte der Jacobimatrix der interpolierten Reduktion $\hat{\mathbf{F}}_I$ getroffen werden kann.

Lemma 6.4.11. *Sei $\mathbf{F}$ nach Voraussetzung 6.2.1 und $\{\Omega_k^I, X_k^I, W_k^I, \mathcal{Z}_k\}$ eine Folge zulässiger inexakter Interpolationen von $\mathbf{c}$ nach Definition 6.4.9. Zudem seien Matrizen $\mathbf{V}_j, j = 1,\dots,d_k$ mittels eines der beiden Verfahren aus Kapitel 5 aufgebaut. Dann existiert ein $\tilde{k} \in \mathbb{N}$ und zwei Konstanten $\hat{c}_0, \hat{c}_1 > 0$ unabhängig von k, sodass für $k \geq \tilde{k}$ und alle $j = 1,\dots,d_k$, sowie den kleinsten und größten Singulärwerten $\hat{\sigma}_{m_k}^j$ und $\hat{\sigma}_1^j$ von $\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}_j)$*

$$(\hat{\sigma}_{m_k}^j)^{-1} \leq \hat{c}_0 \text{ und } \hat{\sigma}_1^j \leq c_1$$

gelten.

Beweis. Zunächst sei festgehalten, dass für die interpolierte Reduktion $\hat{\mathbf{F}}_I$ von $\mathbf{F}$

$$\begin{aligned}\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}_j) &= \sum_{i=1}^{d_k} \left(\underbrace{w_i(\mathbf{x}_j)}_{=\delta_{ij}} \mathbf{V}_i^T \mathbf{DF}(\mathbf{x}_j)\mathbf{Z}_k + \mathbf{V}_i^T \mathbf{F}(\mathbf{x}_j) \underbrace{\nabla w_i(\mathbf{x}_j)}_{=0} \right) \\ &= \mathbf{V}_j^T \mathbf{DF}(\mathbf{x}_j)\mathbf{Z}_k\end{aligned}$$

gilt.

Nach Voraussetzung 6.2.1 existiert ein $c_0 > 0$, sodass für den kleinsten Singulärwert $\sigma_n(\mathbf{x})$ von $\mathbf{DF}(\mathbf{x})$ und alle $\mathbf{x} \in \Omega_k^I$ gilt: $\sigma_n(\mathbf{x})^{-1} \leq c_0$. Sind die Matrizen $\mathbf{V}_j$ mittels des Verfahrens aus Satz 5.1.3 aufgebaut gilt wegen Bemerkung 5.1.5 direkt

$$(\hat{\sigma}_{m_k}^j)^{-1} \leq c_0 =: \hat{c}_0.$$

Nutzt man stattdessen das Verfahren aus Satz 5.2.4 ergeben sich die Abschätzungen

$$\begin{aligned}(\hat{\sigma}_{m_k}^j)^{-1} &\leq \sigma_n^{-2}(\mathbf{x}_j) \left(1 + \frac{1}{\tau^2 \|\mathbf{P}_k \mathbf{T}(\mathbf{x}_j)\|^2}\right)^{1/2}, \text{ bzw.} \\ (\hat{\sigma}_{m_k}^j)^{-1} &\leq \sigma_1(\mathbf{x}_j)\sigma_n^{-2}(\mathbf{x}_j) \left(1 + \frac{1}{\tau^2 \|\mathbf{P}_k \mathbf{T}(\mathbf{x}_j)\|^2}\right)^{1/2},\end{aligned}$$

abhängig davon, ob die $\mathbf{V}_j$ orthonormale Spalten besitzen. $\mathbf{P}_k$ bezeichnet hier den orthogonalen Projektor auf den Ansatzraum $\mathcal{Z}_k$. Nach Voraussetzung 6.2.1 gilt für den größten Singulärwert $\sigma_1(\mathbf{x}_j)$ von $\mathbf{DF}(\mathbf{x})$ für alle $\mathbf{x} \in \Omega_k^I$: $\sigma_1(\mathbf{x}_j) \leq c_1$. Die oberen Schranken hängen also nur über den Wert $\|\mathbf{P}_k \mathbf{T}(\mathbf{x}_j)\|$ von k ab, da der Wert τ ebenfalls nur vom größten und kleinsten Singulärwert von $\mathbf{DF}(\mathbf{x}_j)$ abhängt.

Für größer werdende k wird der Abstand zwischen $\mathbf{x}_j$ und $\mathbf{c}(s_j)$ beliebig klein und wegen der Stetigkeit des Tangentialfeldes gilt dies auch für $\mathbf{T}(\mathbf{x}_j)$ und $\mathbf{T}(\mathbf{c}(s_j))$. Aufgrund der hierarchischen Struktur der $\mathcal{Z}_k$ nähert sich $\|\mathbf{P}_k \mathbf{T}(\mathbf{x}_j)\|^2$ für alle j beliebig Nahe der 1 an. Es existiert also ein $\tilde{k} \in \mathbb{N}$, sodass für alle $k \geq \tilde{k}$ und $j = 1, \ldots, d_k$

$$\|\mathbf{P}_k \mathbf{T}(\mathbf{x}_j)\|^2 \geq \frac{1}{2}$$

gilt. Somit existiert also auch in diesem Fall ein von k unabhängiges $\hat{c}_0$ mit

$$(\hat{\sigma}_{m_k}^j)^{-1} \leq \hat{c}_0.$$

Für $\hat{\sigma}_1^j$ ergibt sich

$$\begin{aligned}\hat{\sigma}_1^j &= \max_{\|\hat{\mathbf{y}}\|=1} \|\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}_j)\hat{\mathbf{y}}\| = \max_{\|\hat{\mathbf{y}}\|=1} \|\mathbf{V}_j^T \mathbf{D}\mathbf{F}(\mathbf{x}_j)\mathbf{Z}_k\hat{\mathbf{y}}\| \\ &\leq \max_{\|\mathbf{u}\|=1} \|\mathbf{V}_j^T \mathbf{D}\mathbf{F}(\mathbf{x}_j)\mathbf{u}\| \leq \|\mathbf{V}_j\|\sigma_1(\mathbf{x}_j) =: \hat{c}_1,\end{aligned}$$

wobei die Spalten von $\mathbf{Z}_k$ eine Orthonormalbasis von $\mathcal{Z}_k$ bilden. Abhängig vom verwendeten Verfahren zum Aufbau der $\mathbf{V}_j$ gilt nun entweder $\|\mathbf{V}_j\| = 1$ oder $\|\mathbf{V}_j\| \leq \sigma_1(\mathbf{x}_j)$. In beiden Fällen ist $\hat{c}_1$ unabhängig von k und der Satz somit bewiesen. □

Bemerkung 6.4.12. *Die Abschätzung für den größten Singulärwert $\hat{\sigma}_1^j$ kann wegen Bedingung (ii) aus Definition 6.4.1 auch unabhängig vom Punkt $\hat{\mathbf{x}}_j$ getroffen werden. Allgemein existiert also eine von k unabhängige Konstante $\hat{c}_1 > 0$, sodass für den größten Singulärwert $\hat{\sigma}_1$ von $\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}})$ mit $\hat{\mathbf{x}} \in \Omega_k^I$ gilt:*

$$\hat{\sigma}_1 \leq \hat{c}_1.$$

Für die inexakte Interpolation können die Konstanten c_0 und c_1 aus Satz 2.2.3 somit ab einem bestimmten $\tilde{k}$ unabhängig von der Anzahl d_k der Interpolationsknoten gewählt werden. Im nächsten Schritt wird ein zu 6.4.5 analoges Lemma bewiesen um wieder die Existenz einer k-unabhängigen Lipschitzkonstante $\hat{L}$ zu garantieren. Da die Aussagen unabhängig von $\mathcal{Z}_k$ getroffen werden, können große Teile des Beweises von Lemma 6.4.5 direkt verwendet werden.

Lemma 6.4.13. *Sei $\mathbf{F}$ nach Voraussetzung 6.2.1, sowie eine Folge zulässiger inexakter Interpolationen $\{\Omega_k^I, X_k^I, W_k^I, \mathcal{Z}_k\}$ nach Definition 6.4.9 gegeben. Zudem seien Matrizen $\mathbf{V}_j, j = 1, \ldots, d_k$ mittels eines der beiden Verfahren aus Kapitel 5 aufgebaut. Dann existiert ein $\hat{L} > 0$ unabhängig von k, sodass für alle $\hat{\mathbf{x}} \in \hat{\Omega}_k^j := \{\hat{\mathbf{x}} \in \mathbb{R}^{m+1} : \ \mathbf{Z}\hat{\mathbf{x}} \in \Omega_k^j \cap \Omega_k^I\}$.*

$$\|\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}) - \mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}_j)\| \leq \hat{L}\|\hat{\mathbf{x}} - \hat{\mathbf{x}}_j\|$$

gilt.

Beweis. Der Beweis läuft analog zu dem des Lemmas 6.4.5 mit dem Unterschied, dass der Term

$$B_2 = \left\| \sum_{i=1}^{d_k} \left(\mathbf{V}_i^T \mathbf{F}(\mathbf{x}_j) \mathbf{Z}_k^T (\nabla w_i(\mathbf{x}) - \nabla w_i(\mathbf{x}_j))^T \right) \right\|,$$

da $\mathbf{F}(\mathbf{x}_j) \neq \mathbf{0}$ gilt, nicht Null wird. Aus Voraussetzung 6.2.1 folgt zunächst, dass für alle $\mathbf{x}, \mathbf{y} \in \Omega$ die Ungleichung

$$\|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y})\| \leq c_1\|\mathbf{x} - \mathbf{y}\|$$

gilt. Aus Bedingung (iii) und (v) von Definition 6.4.9 folgt somit

$$\begin{aligned} B_2 &\leq \sum_{l=-1}^{1} \|\mathbf{F}(\mathbf{x}_j)\| \|\nabla w_j(\mathbf{x}) - \nabla w_j(\mathbf{x}_j)\| \\ &\leq 3\|\mathbf{F}(\mathbf{x}_j) - \mathbf{F}(\mathbf{c}_j)\| \|\nabla w_j(\mathbf{x}) - \nabla w_j(\mathbf{x}_j)\| \\ &\leq 3c_1 \|\mathbf{x}_j - \mathbf{c}_j\| \|\nabla^2 w_j(\mathbf{x}_*)\| \|\mathbf{x} - \mathbf{x}_j\| \leq 3c_1 c_X c_w^2 \hat{\|}\mathbf{x} - \mathbf{x}_j\|. \end{aligned}$$

Es lässt sich also ein $\hat{L} > 0$ finden, sodass unabhängig von k

$$\|\mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}) - \mathbf{D}\hat{\mathbf{F}}_I(\hat{\mathbf{x}}_j)\| \leq \hat{L}\|\hat{\mathbf{x}} - \hat{\mathbf{x}}_j\|, \ j = 0, \ldots, d_k$$

gilt. □

Die verbliebende zu untersuchende Größe ist die Konstante c_2. Dazu wird nun das folgende zu Lemma 6.4.6 ähnliche Lemma bewiesen.

Lemma 6.4.14. *Sei* $\mathbf{F}$ *nach Voraussetzung 6.2.1 und* $\{\Omega_k^I, X_k^I, W_k^I, \mathcal{Z}_k\}$ *eine Folge zulässiger inexakter Interpolationen nach Definition 6.4.9.*

Sei $\hat{\mathbf{r}}_j := (\hat{\mathbf{x}}_{j+1} - \hat{\mathbf{x}}_j)/\|\hat{\mathbf{x}}_{j+1} - \hat{\mathbf{x}}_j\|, \ j = 1, \ldots, d_k - 1$ *definiert, sowie mit* $\hat{\mathbf{T}}(\hat{\mathbf{x}}_j)$ *der Tangentialvektor der interpolierten Reduktion in* $\hat{\mathbf{x}}_j$ *bezeichnet.*

Dann existiert ein $\tilde{k} \in \mathbb{N}$*, sodass für alle* $k \geq \tilde{k}$ *eine Konstante* $\hat{c}_2 > 0$ *unabhängig von* k *existiert mit*

$$|\langle \hat{\mathbf{T}}(\hat{\mathbf{x}}_j), \hat{\mathbf{r}}_j \rangle| \geq \hat{c}_2, \ j = 1, \ldots, d_k.$$

Beweis. Seien die Vektoren $\mathbf{r}_j := (\mathbf{x}_{j+1} - \mathbf{x}_j)/\|\mathbf{x}_{j+1} - \mathbf{x}_j\|, \ j = 1, \ldots, d_k - 1$ definiert. Der Beweis verläuft zunächst analog zu dem des Lemmas 6.4.6. Die folgenden Aussagen werden ohne Einschränkung für $\mathbf{x}_1$ und $\mathbf{r}_1$, bzw. $\hat{\mathbf{x}}_1$ und $\hat{\mathbf{r}}_1$ hergeleitet. Aus $X_k \subset \mathcal{Z}_k$ folgt zunächst $\mathbf{r}_1 \in \mathcal{Z}_k$. Sei nun mit $\mathbf{P}$ der Projektor auf $\operatorname{span}\{\mathbf{Z}_k\hat{\mathbf{T}}(\hat{\mathbf{x}}_1), \mathbf{r}_1\}$ bezeichnet, wobei $\mathbf{Z}_k$ eine Matrix darstellt, deren Spalten eine Orthonormalbasis von $\mathcal{Z}_k$ bilden.

Für den Winkel α_1 zwischen $\mathbf{Z}_k\hat{\mathbf{T}}(\mathbf{x}_1)$ und $\mathbf{P}\mathbf{T}(\mathbf{x}_1)$ ergibt sich

$$\begin{aligned} \cos(\alpha_1) &= \frac{|\langle \mathbf{Z}_k\hat{\mathbf{T}}(\hat{\mathbf{x}}_1), \mathbf{P}\mathbf{T}(\mathbf{x}_1)\rangle|}{\|\mathbf{P}\mathbf{T}(\mathbf{x}_1)\|} = \frac{|\langle \mathbf{Z}_k\hat{\mathbf{T}}(\hat{\mathbf{x}}_1), \mathbf{T}(\mathbf{x}_1)\rangle|}{\|\mathbf{P}\mathbf{T}(\mathbf{x}_1)\|} \\ &\geq \frac{|\langle \mathbf{Z}_k\hat{\mathbf{T}}(\hat{\mathbf{x}}_1), \mathbf{T}(\mathbf{x}_1)\rangle|}{\|\mathbf{P}_k\mathbf{T}(\mathbf{x}_1)\|}, \end{aligned}$$

wobei $\mathbf{P}_k$ den orthogonalen Projektor auf $\mathcal{Z}_k$ bezeichnet. Nach Voraussetzung 6.2.1 und Lemma 5.2.3 existiert ein von den Singulärwerten von $\mathbf{DF}$ abhängiges (und damit k-unabhängiges) $q > 0$ mit

$$|\langle \mathbf{Z}_k\hat{\mathbf{T}}(\hat{\mathbf{x}}_j), \mathbf{T}(\mathbf{x}_j)\rangle| \geq \sqrt{\frac{q^2}{1+q^2}}\|\mathbf{P}_k\mathbf{T}(\mathbf{x}_j)\|.$$

Somit ergibt sich mit $\tau := \sqrt{q^2/(1+q^2)}$

$$\cos(\alpha_1) \geq \tau.$$

Da $0 < \tau < 1$ gilt, folgt daraus $\alpha_1 < \pi/2$.

Sei mit α_2 der Winkel zwischen $\mathbf{PT}(\mathbf{x}_1)$ und $\mathbf{r}_1$ bezeichnet. Für diesen ergibt sich

$$\cos(\alpha_2) = \frac{|\langle \mathbf{PT}(\mathbf{x}_1), \mathbf{r}_1 \rangle|}{\|\mathbf{PT}(\mathbf{x}_1)\|} = \frac{|\langle \mathbf{T}(\mathbf{x}_1), \mathbf{r}_1 \rangle|}{\|\mathbf{PT}(\mathbf{x}_1)\|} \geq |\langle \mathbf{T}(\mathbf{x}_1), \mathbf{r}_1 \rangle|.$$

Man betrachtet nun den Term $|\langle \mathbf{T}(\mathbf{x}_1), \mathbf{r}_1 \rangle|$. Für diesen gilt

$$\begin{aligned} |\langle \mathbf{T}(\mathbf{x}_1), \mathbf{r}_1 \rangle| &\geq \frac{1}{2}(\|\mathbf{T}(\mathbf{x}_1)\|^2 + \|\mathbf{r}_1\|^2 - \|\mathbf{T}(\mathbf{x}_1) - \mathbf{r}_1\|^2) \\ &= 1 - \frac{1}{2}\|\mathbf{T}(\mathbf{x}_1) - \mathbf{r}_1\| \end{aligned}$$

Mittels Dreiecksungleichung ergibt sich für den Ausdruck $\|\mathbf{T}(\mathbf{x}_1) - \mathbf{r}_1\|$

$$\|\mathbf{T}(\mathbf{x}_1) - \mathbf{r}_1\| \leq \|\mathbf{T}(\mathbf{x}_1) - \mathbf{T}(\mathbf{c}(s_1))\| + \|\mathbf{T}(\mathbf{c}(s_1)) - \mathbf{r}_1^{\mathbf{c}}\| + \|\mathbf{r}_1^{\mathbf{c}} - \mathbf{r}_1\|. \tag{6.9}$$

Der Vektor $\mathbf{r_c}$ ergibt sich aus $\mathbf{r}_1^{\mathbf{c}} = (\mathbf{c}(s_2) - \mathbf{c}(s_1))/\|\mathbf{c}(s_2) - \mathbf{c}(s_1)\|$. Aus Bedingung (vii) von Definition 6.4.9 folgt nun die Existenz einer Konstante $c_X > 0$ mit

$$\|\mathbf{x}_j - \mathbf{c}(s_j)\| \leq c_X h_k^2, \;\; j = 1, \ldots, d_k.$$

Alle drei Summanden werden nun separat betrachtet. Zunächst sei dafür festgehalten, dass sich für den Abstand zwischen $\mathbf{c}(s_2)$ und $\mathbf{c}(s_1)$

$$\|\mathbf{c}(s_2) - \mathbf{c}(s_1)\| \leq |s_2 - s_1| \leq h_k$$

ergibt. Dieser sinkt also mit einer Größenordnung h_k.

Da der Abstand zwischen $\mathbf{x}_j$ und $\mathbf{c}(s_j)$ schneller (nämlich quadratisch in h_k) gegen 0 geht, folgt $\mathbf{r}_1 \to \mathbf{r}_1^{\mathbf{c}}$ für $k \to \infty$. Aus den gleichen Gründen ergibt sich $\mathbf{r}_1^{\mathbf{c}} \to \mathbf{T}(\mathbf{c}(s_1))$. Für den übrigen Summanden aus (6.9) erhält man für $k \to \infty$ und der Stetigkeit des Tangentialfeldes

$$\|\mathbf{T}(\mathbf{x}_1) - \mathbf{T}(\mathbf{c}(s_1))\| \to 0.$$

Insgesamt nähert sich der Wert $\cos(\alpha_2) \geq |\langle \mathbf{T}(\mathbf{x}_1), \mathbf{r}_1 \rangle|$ also mit größer werdendem k der 1 an. Somit existiert ein $\tilde{k} \in \mathbb{N}$, sodass für $k \geq \tilde{k}$ und den Winkel α_2

$$\arccos(\tau) + \alpha_2 \leq \arccos\left(\frac{\tau}{2}\right)$$

gilt. Der maximale Wert, den der Winkel zwischen $\hat{\mathbf{T}}(\hat{\mathbf{x}}_1)$ und $\hat{\mathbf{r}}_1$ annehmen kann, ist $\alpha_1 + \alpha_2$. Mit $\hat{c}_2 := \tau/2$ ergibt sich so

$$|\langle \hat{\mathbf{T}}(\hat{\mathbf{x}}_1), \hat{\mathbf{r}}_1 \rangle| \geq \cos(\alpha_1 + \alpha_2) \geq \cos\left(\arccos\left(\frac{\tau}{2}\right)\right) = \frac{\tau}{2} = \hat{c}_2.$$

□

Bevor die Hauptaussage bewiesen werden kann, wird noch ein zusätzliches Hilfslemma benötigt, um den Funktionswert von $\mathbf{F}$ zwischen den Interpolationsknoten abzuschätzen.

Lemma 6.4.15. *Sei $\mathbf{F}$ nach Voraussetzung 6.2.1 und $\{\Omega_k^I, X_k^I, W_k^I, \mathcal{Z}_k\}$ eine Folge zulässiger inexakter Interpolationen. Sei $\mathbf{x}^*$ ein Punkt auf der Strecke zwischen $\mathbf{x}_j$ und $\mathbf{x}_{j+1}$, dann existiert eine k-unabhängige Konstante $c^* > 0$ mit*

$$\mathbf{F}(\mathbf{x}^*) \leq c^* h_k^2.$$

Beweis. Ohne Einschränkung werden die beiden Punkte $\mathbf{x}_1$ und $\mathbf{x}_2$ betrachtet. Ist $\mathbf{P}_k$ der Projektor auf den Raum $\mathcal{Z}_k$, dann gilt für die Ableitung von $\mathbf{c}$ in s_1

$$\mathbf{c}'(s_1) = \frac{\mathbf{c}(s_2) - \mathbf{c}(s_1)}{s_2 - s_1} + \mathcal{O}(|s_2 - s_1|)$$

und somit existiert eine Konstante $c_d > 0$ mit

$$\begin{aligned}\|\mathbf{P}\mathbf{c}'(s_1) - \mathbf{c}'(s_1)\| \leq \frac{\|\mathbf{P}\mathbf{c}(s_2) - \mathbf{c}(s_2)\|}{s_2 - s_1} + \frac{\|\mathbf{P}\mathbf{c}(s_1) - \mathbf{c}(s_1)\|}{s_2 - s_1} \\ + \mathcal{O}(|s_2 - s_1|) \leq 2c_X c_h^{-1} h_k + \mathcal{O}(|s_2 - s_1|) \\ \leq c_d h_k.\end{aligned}$$

Der Punkt $\mathbf{x}^*$ liegt auf der linearen Interpolierenden $\mathbf{I}(s)$ von $\mathbf{P}\mathbf{c}(s)$ mit $\mathbf{I}(s_1) = \mathbf{c}(s_1)$, $\mathbf{I}(s_2) = \mathbf{c}(s_2)$ und $\mathbf{I}(s^*) = \mathbf{x}^*$. Es gilt

$$\begin{aligned}\|\mathbf{F}(\mathbf{x}^*)\| = \|\mathbf{F}(\mathbf{x}^*) - \mathbf{F}(\mathbf{c}(s^*))\| \leq c_1 \|\mathbf{x}^* - \mathbf{c}(s^*)\| \\ \leq c_1 \|\mathbf{x}^* - \mathbf{P}\mathbf{c}(s^*)\| + c_1 \|\mathbf{P}\mathbf{c}(s^*) - \mathbf{c}(s^*)\|.\end{aligned}$$

Für den ersten Summanden ergibt sich

$$\|\mathbf{x}^* - \mathbf{P}\mathbf{c}(s^*)\| = \|\mathbf{I}(s^*) - \mathbf{P}\mathbf{c}(s^*)\| \leq c_{\mathbf{c}} h_k^2.$$

Für den zweiten Term erhält man mit $\mathbf{c}(s^*) = \mathbf{x}_1 + \mathbf{c}'(s_1)(s^* - s_1) + \mathcal{O}((s^* - s_1)^2)$ und einer Konstanten $c_P > 0$

$$\begin{aligned}\|\mathbf{P}\mathbf{c}(s^*) - \mathbf{c}(s^*)\| \leq \|\mathbf{P}\mathbf{x}_1 - \mathbf{x}_1\| + \|\mathbf{P}\mathbf{c}'(s_1) - \mathbf{c}'(s_1)\|(s^* - s_1) \\ + \mathcal{O}((s^* - s_1)^2) \leq c_X h_k^2 + c_d h_k^2 + \mathcal{O}(h_k^2) \\ \leq c_P h_k^2\end{aligned}$$

Somit ergibt sich

$$\|\mathbf{F}(\mathbf{x}^*)\| \leq c_1(c_\mathbf{c} + c_P)h_k^2.$$

□

Es folgt nun die Hauptaussage analog zu Satz 6.4.7, dass also bei genügend großem k das inexakt interpolierte Problem eine Lösungskurve besitzt. Diese verläuft nicht durch die Interpolationsknoten (da diese keine Lösung von $\mathbf{F}(\mathbf{x}) = \mathbf{0}$) darstellen, aber durch alle Träger der zu den Knoten gehörenden Gewichtsfunktionen.

Satz 6.4.16. *Sei* $\mathbf{F}$ *nach Voraussetzung 6.2.1, sowie eine Folge von zulässigen inexakten Interpolationen* $\{\Omega_k^I, X_k^I, W_k^I, \mathcal{Z}_k\}$ *nach Definition 6.4.9 gegeben und die Matrizen* $\mathbf{V}_j$ *nach einem der Verfahren aus Kapitel 5 aufgebaut.*

Dann existiert ein $\tilde{k} \in \mathbb{N}$*, sodass das interpolierte Problem*

$$\hat{\mathbf{F}}_I(\hat{\mathbf{x}}) = \mathbf{0}$$

für $k \geq \tilde{k}$ *eine Lösungskurve* $\hat{\mathbf{c}}$ *besitzt, sodass* $\mathbf{Z}_k\hat{\mathbf{c}}$ *durch alle Träger* Ω_k^j*,* $j = 1, \ldots, d_k$ *verläuft.*

Beweis. Ohne Einschränkung werden die Punkte $\hat{\mathbf{x}}_1$ und $\hat{\mathbf{x}}_2$ betrachtet. Sei $\mathbf{Z}_k$ eine Matrix, deren Spalten eine Orthonormalbasis von $\mathcal{Z}_k$ bilden. Es wird nun gezeigt, dass für jeden Punkt $\mathbf{x}^* = \mathbf{Z}_k\hat{\mathbf{x}}^*$ auf der Verbindungsstrecke zwischen $\hat{\mathbf{x}}_1$ und $\hat{\mathbf{x}}_2$ die Bedingunge von Satz 2.2.3 gelten. Der Beweis dazu verläuft analog zu dem des Satzes 6.4.7 unter Verwendung der Lemmata 6.4.11, 6.4.13, 6.4.14, sowie Bemerkung 6.4.12.

Es existieren eine Konstante M^* unabhängig von k, sowie $\beta^* := \tau h_k$, sodass mit $\alpha^* = \delta^* := \tau h_k/(2\sqrt{2}M^*)$ unter der Bedingung

$$\|\hat{\mathbf{F}}_I(\hat{\mathbf{x}}^*)\| \leq \delta^*$$

eine Lösungskurve von $\hat{\mathbf{F}}_I(\hat{\mathbf{x}}) = \mathbf{0}$ in der Nähe von $\hat{\mathbf{x}}^*$ existiert. Es gilt wieder $\|\hat{\mathbf{F}}_I(\hat{\mathbf{x}}^*)\| \leq \hat{c}_1\|\mathbf{F}(\mathbf{x}^*)\|$ und nach Lemma 6.4.15 existiert ein $c^* > 0$ unabhängig von k, sodass

$$\|\mathbf{F}(\mathbf{x}^*)\| \leq c^* h_k^2$$

erfüllt ist. Ist dann $\tilde{k}$ so groß, dass für $k \geq \tilde{k}$

$$h_k \leq \frac{\tau}{4\sqrt{2}M^*\hat{c}_1c^*}$$

gilt, dann ist die Bedingung $\|\hat{\mathbf{F}}_I(\hat{\mathbf{x}}^*)\| \leq \delta^*$ erfüllt und es existiert eine Lösungskurve

$$\hat{\mathbf{c}}^*(s) = \hat{\mathbf{x}}^* + s\hat{\mathbf{r}}_1 + \hat{\mathbf{Y}}\hat{\mathbf{g}}^*(s)$$

mit einer eindeutigen stetig differenzierbaren Funktion
$\hat{\mathbf{g}}^* : B(0;\alpha^*) \to B(\mathbf{0};\tau h_k)$ und $\hat{\mathbf{F}}_I(\hat{\mathbf{c}}^*(s)) = \mathbf{0}$ und $\hat{\mathbf{r}}_j = (\hat{\mathbf{x}}_{j+1} - \hat{\mathbf{x}}_j)/\|\hat{\mathbf{x}}_{j+1} - \hat{\mathbf{x}}_j\|$. Mit der im Beweis von Satz 6.4.7 verwendeten Argumentation lässt sich garantieren, dass diese Lösungskurven zu ein und der selben Funktion $\hat{\mathbf{c}}$ gehören, die sich als

$$\hat{\mathbf{c}} : \begin{cases} [-\alpha^* + s_1, s_2 + \alpha^*] & \to \mathbb{R}^{m+1}, \\ s & \mapsto \hat{\mathbf{x}}_1 + s\hat{\mathbf{r}}_1 + \hat{\mathbf{Y}}\hat{\mathbf{g}}(s) \end{cases}$$

schreiben lässt (wobei $s_1 = 0$ und $s_2 = \|\hat{\mathbf{x}}_2 - \hat{\mathbf{x}}_1\|$ gilt). Für $\mathbf{Z}_k\hat{\mathbf{c}}$ gilt

$$\|\mathbf{Z}_k\hat{\mathbf{c}}(s_1) - \mathbf{x}_1\| = \|\hat{\mathbf{c}}(s_1) - \hat{\mathbf{x}}_1\| \leq \|\hat{\mathbf{Y}}\hat{\mathbf{g}}(s_1)\| \leq \tau h_k$$

Aufgrund von Bedingung (vi) von Definition 6.4.9 gilt $B(\mathbf{x}_1;\tau h_k) \subset \Omega_1^k$ und somit schneidet die Kurve $\mathbf{Z}_k\hat{\mathbf{c}}$ den Träger Ω_1^k. Die gleiche Aussage lässt sich für den Träger der zu $\mathbf{x}_2$ gehörenden Gewichtsfunktion treffen.

Es wurde also bewiesen, dass zwischen allen Punktepaaren $\hat{\mathbf{x}}_j$ und $\hat{\mathbf{x}}_{j+1}$ eine Lösungskurve $\hat{\mathbf{c}}_j$ verläuft, die sich bezüglich des Verbindungsvektors parametrisieren lässt und beide Träger Ω_k^j und Ω_k^{j+1} schneidet. Es bleibt zu zeigen, dass diese Einzelkurven zu einer Gesamtkurve gehören.

Seien dazu ohne Einschränkung die Punkte $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2$ und $\hat{\mathbf{x}}_3$ betrachtet. Zwischen $\hat{\mathbf{x}}_1$ und $\hat{\mathbf{x}}_2$ sowie zischen $\hat{\mathbf{x}}_2$ und $\hat{\mathbf{x}}_3$ existieren jeweils eine Lösungskurve $\hat{\mathbf{c}}_1(s) = \hat{\mathbf{x}}_1 + s\hat{\mathbf{r}}_1 + \hat{\mathbf{Y}}_1\hat{\mathbf{g}}_1(s)$ und $\hat{\mathbf{c}}_2(t) = \hat{\mathbf{x}}_2 + t\hat{\mathbf{r}}_2 + \hat{\mathbf{Y}}_2\hat{\mathbf{g}}_2(t)$. Die Funktion $\hat{\mathbf{g}}_1 : B(0;\alpha^*) \to B(\mathbf{0};\beta^*)$ ist eindeutig in ihrem Urbild- Bildbereich. Das heißt, existiert ein Punkt $(\hat{\mathbf{y}}_1, t_1)$ mit $\hat{\mathbf{F}}_I(\hat{\mathbf{x}}_2 + t_1\hat{\mathbf{r}}_2 + \hat{\mathbf{Y}}_2\hat{\mathbf{y}}_1) = \mathbf{0}$ und $t_1 \in B(0;\alpha^*)$, sowie $\hat{\mathbf{y}}_1 \in B(\mathbf{0};\beta^*)$, gehört dieser Punkt zur Lösungskurve $\hat{\mathbf{c}}_2$.

Sei jetzt s_2 der Wert für den $\hat{\mathbf{x}}_1 + s_2\hat{\mathbf{r}}_1 = \hat{\mathbf{x}}_2$ gilt. Der Punkt $\hat{\mathbf{c}}_1(s_2) = \hat{\mathbf{x}}_1 + s_2\hat{\mathbf{r}}_1 + \hat{\mathbf{Y}}_1\hat{\mathbf{g}}_0(s_2)$ besitzt bezüglich $\hat{\mathbf{x}}_2, \mathbf{r}_2$ und $\hat{\mathbf{Y}}_2$ die Koordinaten $(t_1, \hat{\mathbf{y}}_1)$ mit

$$\hat{\mathbf{c}}_1(s_2) = \hat{\mathbf{x}}_2 + t_1\hat{\mathbf{r}}_2 + \hat{\mathbf{Y}}_2\hat{\mathbf{y}}_1.$$

Für den Wert t_1 ergibt sich

$$\begin{aligned} |t_1| &= |\hat{\mathbf{r}}_2^T(\hat{\mathbf{x}}_2 - t_1\hat{\mathbf{r}}_2 + \hat{\mathbf{Y}}_2\hat{\mathbf{y}}_1 - \hat{\mathbf{x}}_2)| = |\hat{\mathbf{r}}_2^T(\hat{\mathbf{c}}_1(s_2))| \\ &= |\hat{\mathbf{r}}_2^T(\hat{\mathbf{x}}_1 + s_2\hat{\mathbf{r}}_1 + \hat{\mathbf{Y}}_1\hat{\mathbf{g}}_0(s_2) - \hat{\mathbf{x}}_2)| = |\hat{\mathbf{r}}_2^T(\mathbf{Y}_1\mathbf{g}_0(s_2))| \\ &\leq \|\hat{\mathbf{r}}_2^T\mathbf{Y}_1\|\|\mathbf{g}_0(s_2)\| \leq \|\hat{\mathbf{r}}_2^T\mathbf{Y}_1\|\tau h_k. \end{aligned}$$

Wegen der Stetigkeit der Kurve existiert nun ein $\tilde{k}$, sodass für $k \geq \tilde{k}$ die Abschätzung $\|\hat{\mathbf{r}}_2^T\mathbf{Y}_1\| \leq 1/(2\sqrt{2}M^*)$ gilt und damit

$$|t_1| \leq \frac{\tau h_k}{2\sqrt{2}M^*} = \alpha^*$$

erfüllt ist. Für den Wert $\hat{\mathbf{y}}^1$ ergibt sich direkt

$$\begin{aligned}\|\hat{\mathbf{y}}_1\| &= \|\hat{\mathbf{Y}}_2^T(\hat{\mathbf{x}}_2 + t_1\hat{\mathbf{r}}_2 + \hat{\mathbf{Y}}_2\hat{\mathbf{y}}_1 - \hat{\mathbf{x}}_2)\| = \|\hat{\mathbf{Y}}_2^T(\hat{\mathbf{c}}_1(s_2))\| \\ &= \|\hat{\mathbf{Y}}_2^T(\hat{\mathbf{x}}_1 + s_2\hat{\mathbf{r}}_1 + \hat{\mathbf{Y}}_1\hat{\mathbf{g}}_0(s_2) - \hat{\mathbf{x}}_2)\| = \|\hat{\mathbf{Y}}_2^T\hat{\mathbf{Y}}_1\|\|\hat{\mathbf{g}}_0(s_2)\| \\ &\leq \|\hat{\mathbf{g}}_0(s_2)\| \leq \tau h_k = \beta^*.\end{aligned}$$

Somit gilt wegen der Eindeutigkeit von $\hat{\mathbf{g}}_1$ der Zusammenhang $\hat{\mathbf{g}}_1(t_1) = \hat{\mathbf{y}}_1$ und daher $\hat{\mathbf{c}}_1(s_2) = \hat{\mathbf{c}}_2(t_1)$. Die Kurven sind also verbunden. □

Kapitel 7

Zweiparametrige Systeme und numerische Untersuchungen

Bisher wurden Methoden vorgestellt, die Lösungen einparametriger nichtlineare Gleichungen mittels Basisreduktion zu approximieren. Diese Methoden sollen verwendet werden, um Parameterstudien für einen zweiten Parameter durchzuführen. Das Ausgangsproblem für eine Funktion $\mathbf{F} \in C^1(Y \times \Lambda \times D, \mathbb{R}^n)$ mit $Y \subset \mathbb{R}^n$ und $\Lambda, D \subset \mathbb{R}$ hat hier die Gestalt

$$\mathbf{F}(\mathbf{u}, \lambda, \mu) = \mathbf{0}.$$

Wie zuvor werden die Variablen $\mathbf{u}$ und λ mittels $\mathbf{x} := (\mathbf{u}^T, \lambda)^T$ zu einem Vektor zusammengefasst und das Problem mit $\mathbf{F} \in C^1(\Omega \times D, \mathbb{R}^n)$ mit $\Omega \subset \mathbb{R}^{n+1}$ zu

$$\mathbf{F}(\mathbf{x}, \mu) = \mathbf{0} \tag{7.1}$$

umformuliert. Man geht nun davon aus, dass für einen Parameter μ_0 eine Lösungskurve $\mathbf{c}_0 : S \to \Omega$ mit $S \subset \mathbb{R}$ existiert, sodass für alle $s \in S$

$$\mathbf{F}(\mathbf{c}_0(s), \mu_0) = \mathbf{0}$$

gilt. Für diese Lösungskurve wird eine interpolationsbasierte Reduktion nach Kapitel 6 aufgebaut. Ziel ist es, diese Reduktion für die Approximation der Lösung von (7.1) für Parameter μ in $B(\mu_0; \gamma)$ zu verwenden.

Zunächst werden Bedingungen angegeben, unter denen das volldimensionale System (7.1) in der Nähe von μ_0 eine Lösungskurve besitzt. Ausgehend von einer diskreten Menge von Punkten auf der Lösungskurve $\mathbf{c}_0$ findet man senkrecht zum Tangentialfeld in diesen Punkten Lösungen der Gleichung $\mathbf{F}(\mathbf{x}, \mu^*) = \mathbf{0}$ für einen festen nicht zu weit von μ_0 entfernten Parameter μ^*. Die durch diese Punkte verlaufenden lokalen Lösungskurven gehören zu einer gemeinsamen Kurve, wenn die Punkte nicht zu weit auseinander liegen und μ^* nicht zu weit von μ_0 entfernt ist. Der Beweis verläuft dabei ähnlich zu dem des Satzes (6.4.7), da die Ausgangssituation - eine bekannte Menge von Lösungspunkten, die beliebig nah beeinander liegen - die gleiche ist.

Satz 7.0.17. *Sei* $\mathbf{F} \in C^1(\Omega \times D, \mathbb{R}^n)$, *sodass* $\mathbf{F}(\cdot,\cdot,\mu_0)$ *Voraussetzung 6.2.1 erfüllt. Seien weiterhin* $\tilde{c}_0, \tilde{c}_1 > 0$ *gegeben, sodass für den kleinsten und größten Singulärwert* $\sigma_n(\mathbf{x},\mu)$ *bzw.* $\sigma_1(\mathbf{x},\mu)$ *von* $\mathbf{D_x F}(\mathbf{x},\mu)$ *für alle* $(\mathbf{x},\mu) \in \Omega \times D$

$$\sigma_n^{-1}(\mathbf{x},\mu) \leq \tilde{c}_0, \text{ sowie } \sigma_1(\mathbf{x},\mu) \leq \tilde{c}_1$$

gilt. Es existiere des Weiteren ein $\tilde{L} > 0$, *sodass für alle* $(\mathbf{x},\mu) \in \Omega \times D$

$$\|\mathbf{DF}(\mathbf{x},\mu) - \mathbf{DF}(\bar{\mathbf{x}},\bar{\mu})\| \leq \tilde{L}\|(\mathbf{x}^T,\mu)^T - (\bar{\mathbf{x}}^T,\bar{\mu})^T\|$$

gilt.

Dann existiert eine Konstante $\gamma > 0$, *nur abhängig von* $\tilde{c}_0$, $\tilde{c}_1$ *und* $\tilde{L}$ *und ein Interwall* S_μ, *sodass für alle* $\mu \in B(\mu_0;\gamma)$ *eine Lösungskurve* $\mathbf{c}_\mu : S_\mu \to \mathbb{R}^{n+1}$ *mit*

$$\mathbf{F}(\mathbf{c}_\mu(s),\mu) = \mathbf{0}$$

existiert. Weiterhin gilt für $\mu \to \mu_0$: $S_\mu \to S$.

Beweis. Zur Veranschaulichung sind wichtige im Beweis auftretende Größen in Abbildung 7.1 dargestellt.

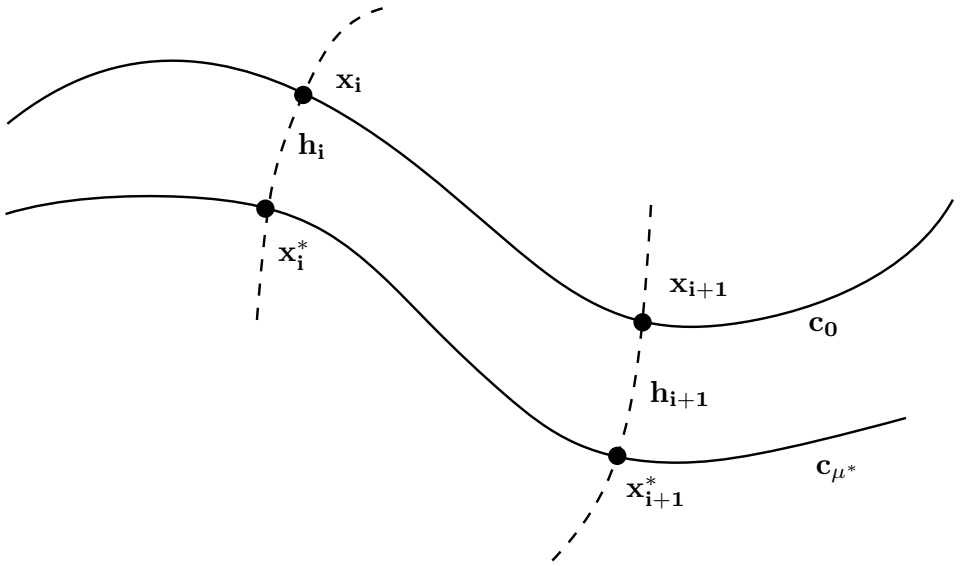

Abbildung 7.1: Skizze zum Beweis des Satzes 7.0.17

Nach Voraussetzung 6.2.1 besitzt das Problem $\mathbf{F}(\mathbf{x},\mu_0) = \mathbf{0}$ eine Lösungskurve $\mathbf{c} \in C^1(S,\mathbb{R}^{n+1})$ mit $\mathbf{F}(\mathbf{c}(s),\mu_0) = \mathbf{0}$ für alle $s \in S \subset \mathbb{R}$. Auf dieser Kurve seien N Punkte $X = \{\mathbf{x}_i = \mathbf{c}_0(s_i),\ i = 1,\ldots,N\}$ mit $h_N := \max\{|s_{i+1} - s_i|,\ i =$

$1, \ldots, N-1\}$ gewählt. Es soll dabei $h_N \to 0$ für $N \to \infty$ gelten. Für jeden Punkt $\mathbf{x}_i$ sei die Funktion

$$\mathbf{H}_i : \begin{cases} \mathbb{R}^{n+1} \times D & \to \mathbb{R}^{n+1}, \\ (\mathbf{x}, \mu) & \mapsto \begin{pmatrix} \mathbf{F}(\mathbf{x}, \mu) \\ \mathbf{T}_0(\mathbf{x}_i)^T(\mathbf{x} - \mathbf{x}_i) \end{pmatrix} \end{cases}$$

definiert, wobei $\mathbf{T}_0(\mathbf{x}_i)$ der Tangentialvektor von $\mathbf{D}_\mathbf{x}\mathbf{F}(\mathbf{x}_i, \mu_0)$ ist. Es gilt dann $\mathbf{H}_i(\mathbf{x}_i, \mu_0) = \mathbf{0}$. Für den kleinsten Singulärwert $\sigma_{n+1}^{\mathbf{H}}$ von $\mathbf{D}_\mathbf{x}\mathbf{H}_i(\mathbf{x}_i, \mu_0)$ ergibt sich

$$\begin{aligned} (\sigma_{n+1}^{\mathbf{H}})^2 &= \min_{\|\mathbf{u}\|=1} \|\mathbf{D}_\mathbf{x}\mathbf{H}_i(\mathbf{x}_i, \mu_0)\mathbf{u}\|^2 = \min_{\|\mathbf{u}\|=1} \left\| \begin{pmatrix} \mathbf{D}_\mathbf{x}\mathbf{F}(\mathbf{x}_i, \mu_0)\mathbf{u} \\ \mathbf{T}_i^T\mathbf{u} \end{pmatrix} \right\|^2 \\ &= \min_{\|\mathbf{u}\|=1} (\|\mathbf{D}_\mathbf{x}\mathbf{F}(\mathbf{x}_i, \mu_0)\mathbf{u}\|^2 + |\mathbf{T}_i^T\mathbf{u}|^2) \\ &\geq \min_{\|\mathbf{u}\|=1} \|\mathbf{D}_\mathbf{x}\mathbf{F}(\mathbf{x}_i, \mu_0)\mathbf{u}\|^2 + \min_{\|\mathbf{u}\|=1} |\mathbf{T}_i^T\mathbf{u}|^2 = \sigma_n(\mathbf{x}_i, \mu_0)^2 \end{aligned}$$

woraus dann

$$\|\mathbf{D}_\mathbf{x}\mathbf{H}_i(\mathbf{x}_i, \mu_0)^{-1}\| = (\sigma_{n+1}^{\mathbf{H}})^{-1} \leq \sigma_n(\mathbf{x}_i, \mu_0)^{-1} \leq \tilde{c}_0$$

folgt. Des Weiteren gilt

$$\|\mathbf{D}_\mu\mathbf{H}_i(\mathbf{x}_i, \mu_0)\| = \|\mathbf{D}_\mu\mathbf{F}(\mathbf{x}_i, \mu_0)\| \leq \tilde{c}_1.$$

Für $M > 0$ und $\beta := 1/(2M\tilde{L})$ gilt für alle $(\mathbf{x}_i, \mu_0) \in B(\mathbf{x}_i, \mu_0; \beta)$

$$\|\mathbf{D}\mathbf{H}_i(\mathbf{x}, \mu) - \mathbf{D}\mathbf{H}_i(\mathbf{x}_i, \mu_0)\| = \|\mathbf{D}\mathbf{F}(\mathbf{x}, \mu) - \mathbf{D}\mathbf{F}(\mathbf{x}_i, \mu_0)\| \leq \frac{1}{2M}$$

Setzt man nun $M := \sqrt{2}\max\{\tilde{c}_0, 1 + \tilde{c}_0\tilde{c}_1\}$ und $\alpha := \beta/(2M)$ existiert nach Korollar 2.1.4 eine Funktion $\mathbf{h}_i : B(\mu_0; \alpha) \to B(\mathbf{x}_i; \beta)$, sodass für alle $\mu \in B(\mu_0; \alpha)$

$$\mathbf{H}_i(\mathbf{h}_i(\mu), \mu) = \mathbf{0}$$

gilt. Man geht hier davon aus, dass stets $B(\mathbf{x}_i; \mu_0; \beta) \subset \Omega \times D$ gilt, da falls das nicht der Fall ist, einfach kleine Kugeln verwendet werden können, da die Abschätzungen in diesen immer noch gelten. Man beachte, dass wegen $\mathbf{T}_0(\mathbf{x}_i)^T(\mathbf{h}_i(\mu) - \mathbf{x}_i) = \mathbf{0}$, diese Kurve im Orthogonalraum von $\mathbf{T}_0(\mathbf{x}_i)$ und somit ausgehend von $\mathbf{x}_i$ senkrecht zur Lösungskurve $\mathbf{c}_0$ verläuft. Für $\mathbf{h}$ gilt weiterhin

$$\|\mathbf{h}(\mu) - \mathbf{h}_i(\mu_0)\| \leq 2M\|\mu - \mu_0\|.$$

Man beachte, dass die Größen α und β unabhängig vom Punkt $\mathbf{x}_i$ sind. Sei nun $\mu^* \in B(\mu_0;\gamma) \subset B(\mu_0;\alpha)$ und $\mathbf{x}_i^* := \mathbf{h}_i(\mu^*)$. Für den kleinsten und größten Singulärwert $\sigma_n(\mathbf{x}_i^*,\mu^*)$ und $\sigma_1(\mathbf{x}_i^*,\mu^*)$ von $\mathbf{D_x F}(\mathbf{x}_i^*,\mu^*)$ gilt

$$\sigma_n(\mathbf{x}_i^*,\mu^*)^{-1} \leq \tilde{c}_0, \text{ und } \sigma_1(\mathbf{x}_i^*,\mu^*) \leq \tilde{c}_1.$$

Sei nun der Punkt $\mathbf{x}_{i+1}^* = h_{i+1}(\mu^*)$ betrachtet. Für den Abstand der beiden Punkte gilt

$$\begin{aligned}\|\mathbf{x}_{i+1}^* - \mathbf{x}_i^*\| &\leq \|\mathbf{x}_{i+1}^* - \mathbf{x}_{i+1}\| + \|\mathbf{x}_{i+1} - \mathbf{x}_i\| + \|\mathbf{x}_i^* - \mathbf{x}_i\| \\ &\leq 2M|\mu^* - \mu_0| + |s_{i+1} - s_i| + 2M|\mu^* - \mu_0| \\ &\leq 4M\gamma + h_N.\end{aligned}$$

Das Tangentialfeld $\mathbf{T}(\mathbf{x},\mu)$ der Matrix $\mathbf{D_x F}(\mathbf{x},\mu)$ ist stetig bezüglich $(\mathbf{x},\mu)$, somit gilt für genügend kleine h_N und γ für den Vektor $\mathbf{r}_i^* = (\mathbf{x}_{i+1}^* - \mathbf{x}_i^*)/\|\mathbf{x}_{i+1}^* - \mathbf{x}_i^*\|$ die Abschätzung

$$|\langle \mathbf{T}(\mathbf{x}_i^*,\mu^*), \mathbf{r}_i^* \rangle| \geq \frac{1}{2}.$$

Sei $M^* := \sqrt{2}\max(2\tilde{c}_0, 1 + 2\tilde{c}_0\tilde{c}_1)$ und $\beta^* = 1/(2M\tilde{L})$, dann gilt

$$\|\mathbf{D_x F}(\mathbf{x},\mu^*) - \mathbf{D_x F}(\mathbf{x}_i^*,\mu^*)\| \leq \frac{1}{2M^*}.$$

Sei weiterhin $\alpha^* := \beta^*/(2M^*)$, dann existiert nach Satz 2.2.3 und Bemerkung 2.2.4 eine Funktion $\mathbf{g}_i^* : B(0;\alpha^*) \to B(\mathbf{0};\beta^*)$, sodass für die C^1-Funktion

$$\mathbf{c}_i^* : \begin{cases} B(0;\alpha^*) & \to \mathbb{R}^{n+1} \\ s & \mapsto \mathbf{x}_i^* + s\mathbf{r}_i^* + \mathbf{Y}_i^*\mathbf{g}_i^*(s) \end{cases},$$

$\mathbf{F}(\mathbf{c}_i^*(s),\mu^*) = \mathbf{0}$ gilt. Hierbei enthalten die Spalten von $\mathbf{Y}_i^*$ eine Orthonormalbasis von $R(\mathbf{r}_i^*)^\perp$. Der Punkt $\mathbf{x}_{i+1}^*$ hat bezüglich $\hat{\mathbf{r}}_i^*$ und $\mathbf{Y}_i^*$ die Koordinaten $\mathbf{x}_{i+1}^* = \mathbf{x}_i^* + s_{i+1}\mathbf{r}_i^* + \mathbf{Y}_i^*\mathbf{y}_{i+1}$. Da $\mathbf{g}_i^*$ eindeutig ist, gehört der Punkt $\mathbf{x}_{i+1}^*$ unter der Bedingung $s_{i+1} \in B(0;\alpha^*)$ und $\mathbf{y}_{i+1} \in B(\mathbf{0};\beta^*)$ zur Lösungskurve $\mathbf{c}_i^*$. Wegen $\mathbf{y}_{i+1} = \mathbf{0}$ ist die zweite Bedingung sofort erfüllt. Für s_1 ergibt sich

$$s_1 = \|\mathbf{x}_{i+1}^* - \mathbf{x}_i^*\| \leq 4M\gamma + h_N.$$

Für genügend kleine h_N und γ ist $s_1 \leq \alpha^*$ also erfüllt, da die Größe α^* fest gewählt werden kann und nicht von der Wahl der Punkte $\mathbf{x}_i$ auf der Lösungskurve $\mathbf{c}_0$ abhängt. Ist N also groß genug und μ^* nicht zu weit von μ_0 entfernt, existiert ein Intervall $S_{\mu^*} \in \mathbb{R}$ und eine Kurve $\mathbf{c}_{\mu^*} : S \to \mathbb{R}^{n+1}$, die durch alle $\mathbf{x}_i^*$ verläuft und für die

$$\mathbf{F}(\mathbf{c}_{\mu^*}(s),\mu^*) = \mathbf{0}$$

gilt. □

Unter gewissen Regularitätsbedingungen kann also garantiert werden, dass für jeden Parameter μ in einer Umgebung von μ_0 eine Lösungskurve von (7.1) existiert.

Ziel der Arbeit ist es, mit Hilfe der für einen festen Parameter μ_0 aufgebauten interpolationsbasierten Reduktion Parameterstudien durchzuführen. Der numerische Gesamtaufwand wird dadurch stark verringert, da der Aufbau des Ansatzraums $\mathcal{Z}$ und der Testräume $\mathcal{V}_i$ den kostspieligen Anteil der Berechnungen darstellen und nur einmal durchgeführt werden müssen. Der große Vorteil besteht darin, dass im Gegensatz zu den globaeln RB-Methoden (vergleiche Kapitel 3.2) Umkehrpunkte bezüglich λ, die bei vielen nichtlinearen Gleichungen auftreten, zugelassen sind. Im Folgenden werden numerische Beispiele präsentiert, die zeigen, dass Parameterstudien mittels interpolationsbasierter Reduktion in der Tat möglich sind, und zu guten Ergebnissen führen.

7.1 Das verallgemeinerte Bratu-Problem

Als erstes numerisches Beispiel wird das verallgemeinerte Bratu-Problem betrachtet (vergleiche dazu [28]). Dieses ist durch das Randwertprobem

$$\begin{cases} -\Delta u = \lambda \exp\left(\frac{u}{\mu u+1}\right), & \text{in } \Omega = (0,1)^2 \\ u = 0, & \text{auf } \partial\Omega. \end{cases}$$

gegeben. In Kapitel 6.1 wurde bereits der Fall $\mu = 0$ betrachtet. Diskretisiert man das Problem mittels Finiter-Differenzen-Methode und einem äquidistanten Gitter mit Schrittweite h, so ergibt sich das nichtlineare Gleichungssystem

$$\mathbf{G} : \begin{cases} \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \supseteq Y \times \Lambda \times D & \to \mathbb{R}^n, \\ (\mathbf{u}^T, \lambda, \mu)^T & \mapsto \mathbf{D}\mathbf{u} - \lambda \mathbf{f}(\mathbf{u}, \mu) \end{cases},$$

mit den offenen Mengen Λ und Y. Der Vektor $\mathbf{u} = (u_1, \ldots, u_n)^T$ enthält dabei die Funktionswerte von u in den einzelnen Knotenpunkten des Gitters. Die Matrix $\mathbf{D} \in \mathbb{R}^{n,n}$ repräsentiert den negativen diskretisierten Laplace-Operator. Die Matrix hängt von der Art und Weise ab, wie das Gitter numeriert ist. Benutzt man zum Beispiel eine Zuweisung wie in [38] ist $\mathbf{D}$ eine Blockdiagonalmatrix der Form

$$\mathbf{D} = \frac{1}{h^2} \begin{pmatrix} \mathbf{B} & \mathbf{I} & & & \\ \mathbf{I} & \mathbf{B} & \mathbf{I} & & \\ & \ddots & \ddots & \ddots & \\ & & \mathbf{I} & \mathbf{B} & \mathbf{I} \\ & & & \mathbf{I} & \mathbf{B} \end{pmatrix} \text{ mit } \mathbf{B} = \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{pmatrix}.$$

Der Vektor $\mathbf{f}$ ist dann gegeben als

$$\mathbf{f}(\mathbf{u}, \mu) := \left(\exp\left(\frac{u_1}{\mu u_1 + 1} \right), \ldots, \exp\left(\frac{u_n}{\mu u_n + 1} \right) \right)^T.$$

Man fast nun die Variablen $\mathbf{u}$ und λ als $\mathbf{x} := (\mathbf{u}^T, \lambda)^T$ zusammen, setzt $\Omega := \Lambda \times Y$ und betrachtet für $\mathbf{F} : \Omega \times D \to \mathbb{R}^n$ das Problem

$$\mathbf{F}(\mathbf{x}, \mu) := \mathbf{G}(\mathbf{u}, \lambda, \mu) = \mathbf{0}.$$

Zunächst wird eine Interpolation mit inexakten Knoten nach Kapitel 6.4.2 für den Fall $\mu = 0$ betrachtet. Die Berechnung der volldimensionalen Lösungskurve mit $n = 400$ erfolgt durch die Anwendung des Matlab-Lösers `ode45` auf die Differentialgleichung

$$\mathbf{c}_0'(s) = \mathbf{T}(\mathbf{c}_0(s))$$
$$\mathbf{c}_0(0) = \mathbf{0}.$$

In Abbildung 7.2 ist der Verlauf der ersten Komponente der Lösung $\mathbf{u}$ bezüglich des Parameters λ dargestellt.

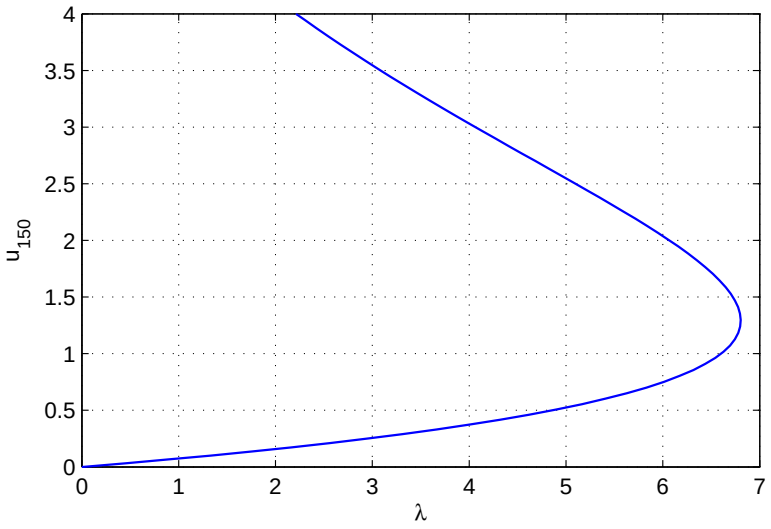

Abbildung 7.2: Erste Komponente der Lösung von $\mathbf{G}(\mathbf{u}, \lambda, 0) = \mathbf{0}$ in Abhängigkeit von λ

Die diskretisierte Lösungskurve besteht aus 100 Snapshots $\mathbf{x}_1, \ldots, \mathbf{x}_{100}$, mit deren Hilfe für verschiedene $m \in \mathbb{N}$ ein POD-Basis der Dimension $m + 1$

aufgebaut wird (vergleiche dazu Kapitel 4.2). Um die Interpolationsknoten zu wählen, wird ein Algorithmus analog zu dem in Kapitel 4.1 behandelten Greedy-Algorithmus verwendet. Dabei werden für verschiedene $d \in \mathbb{N}$ eine Menge von d Punkten aus der Snapshotmenge ausgewählt und in den Ansatzraum projiziert. Für den Fälle $d = 3$ und $d = 10$ sind die Interpolationsknoten in Abbildung 7.3 dargestellt, wobei der POD-Ansatzraum hier die Dimension 5 hat. Man beachte, dass die Knoten keine Lösungen des Problems $\mathbf{F}(\mathbf{x}, 0) = \mathbf{0}$ darstellen, sondern Projektionen solcher Lösungen in den Ansatzraum. Da dieser den Raum, in dem sich die volldimensionale Lösungskurve bewegt, sehr gut approximiert, ist in der Abbildung kein Unterschied zwischen den ausgewählten Snapshots und den Interpolationsknoten, die deren Projektion darstellen, zu erkennen.

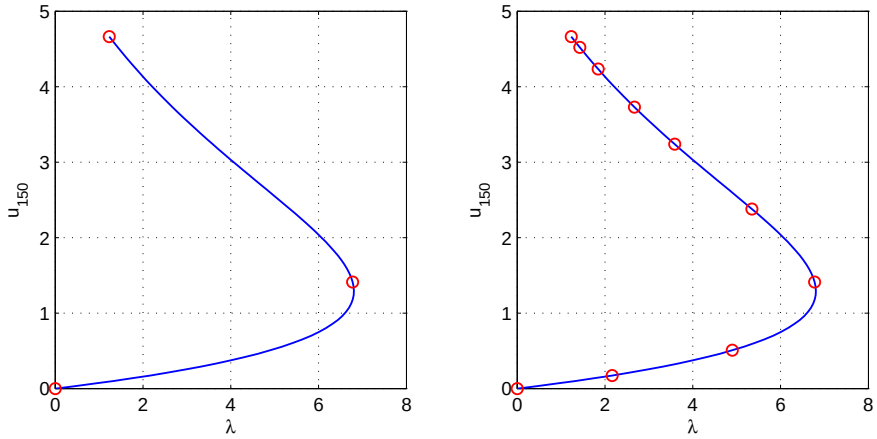

Abbildung 7.3: Auswahl von $d = 3$ und $d = 10$ Interpolationsknoten mittels Greedy-Algorithmus

Die Lösung des interpolierten Problems wird mittels Prädiktor-Korrektor-Verfahren berechnet (vergleiche dazu [2]. Da keine gemeinsame Parametrisierung der volldimensionalen Lösungskurve $\mathbf{c}$ und der Approximation $\mathbf{c}_R$ gegeben ist, kann der auftretende Fehler nicht einfach über die Differenz $\|\mathbf{c}(s) - \mathbf{c}_R(s)\|$ angegeben werden.

Zur Definition des Fehlers wird stattdessen das Newton-Verfahren im volldimensionalen Raum auf alle Punkte der Approximation angewendet wird.Auf diese Weise gelangt man zu einem Punkt auf der Ursprungskurve, der Nahe an dem betrachteten Punkt auf der approximierten Kurve liegt. Auf diese Weise wird hier ein Abstand zwischen den Punkten der Approximation und der volldimensionalen Lösungskurve definiert und über das Maximum dieser Abstände der Fehler der Approximation. Für steigende Dimension der Ansatzräume und

$d = 3$ Interpolationsknoten ist der Fehler der jeweiligen Approximation in Abbildung 7.4 dargestellt.

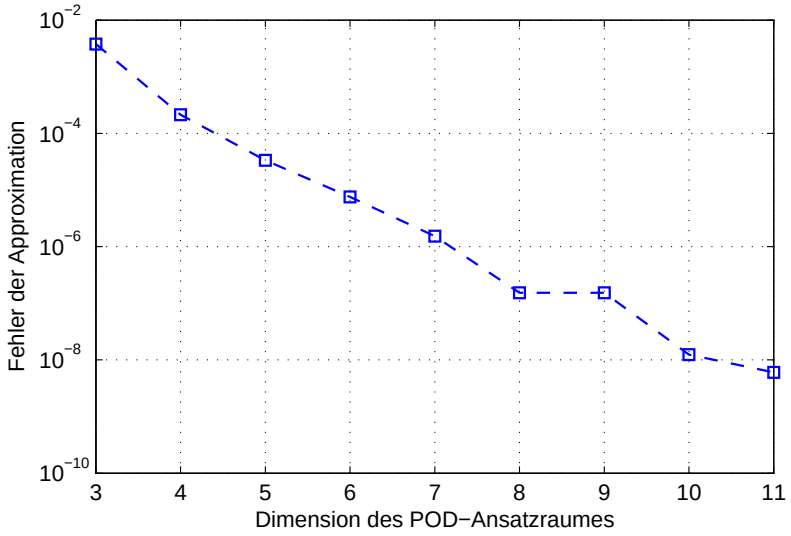

Abbildung 7.4: Fehler der Approximation bei $d = 3$ Interpolationsknoten und steigender Dimension des Ansatzraumes

Variiert man den zweiten Parameter μ über ein Intervall $[-1, 1]$ ergibt sich ein Kurvenverlauf wie in Abbildung 7.5. Der Umkehrpunkt bezüglich λ verschwindet also mit größer werdendem μ.

Es zeigt sich, dass die für den Fall $\mu = 0$ aufgebaute interpolationsbasierte Reduktion mit 3 Knoten und einem POD-Ansatzraum der Dimension 5 für den gesamten betrachteten Parameterbereich $[-1, 1]$ verwendet werden kann. Graphisch sind die Ergebnisse in Abbildung 7.6 für eine Auswahl von μ dargestellt und unterscheiden sich im Rahmen der Darstellungsgenauigkeit in der Abbildung nicht von den volldimensionalen Lösungskurven. Der für die μ auftretende Fehler findet sich in Abbildung 7.7.

Die für $\mu = 0$ einmal aufgebaute Reduktion lässt sich also in diesem Fall für Parameterstudien bezüglich μ verwenden. Da für die hier betrachteten Beispiele von 3 Interpolationsknoten ausgegangen wurde, sind die zu den einzelnen Knoten gehörenden Träger groß genug, um die zu approximierende Lösungskurve zu umschließen. Wird die Anzahl der verwendeten Knoten erhöht, und damit die Größe der zu den Gewichtsfunktionen gehörenden Träger verkleinert, kann der Fall auftreten, dass die zu approximierende Lösungskurve außerhalb der Vereinigung der Träger und damit auch außerhalb des Definitionsbereiches der interpolierten Funktion liegt. In diesem Fall kann eine Korrektur der Kno-

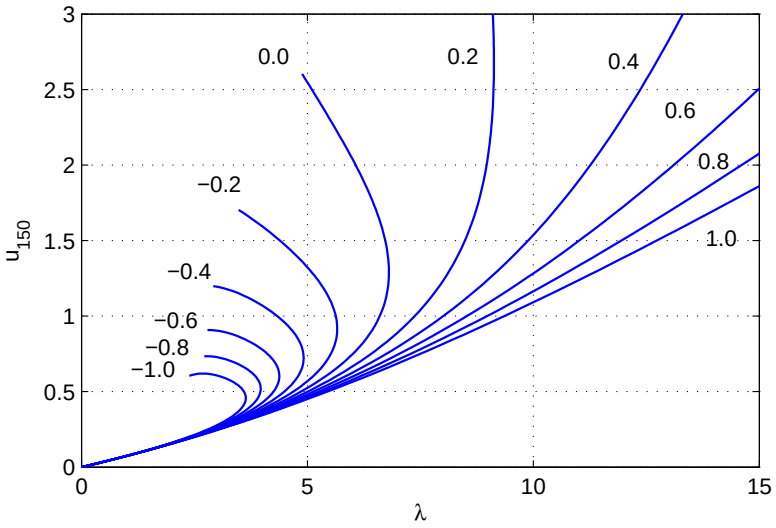

Abbildung 7.5: Verlauf der Lösung von $\mathbf{G}(\mathbf{u}, \lambda, \mu) = \mathbf{0}$ für $\mu \in [-1, 1]$.

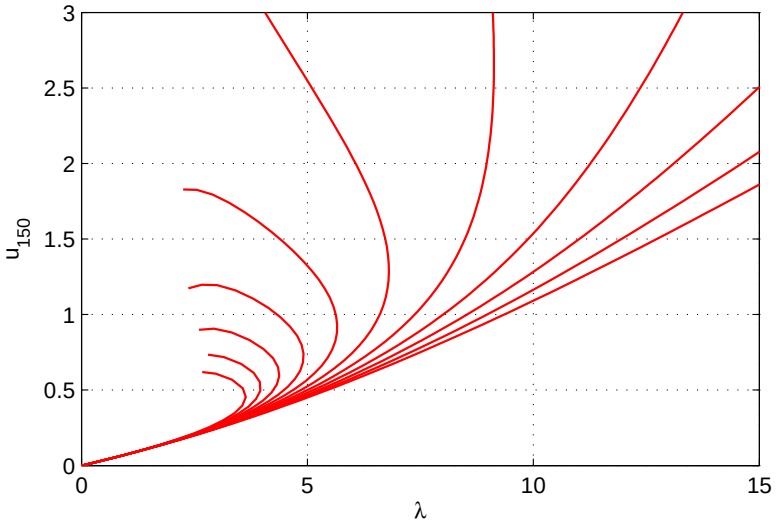

Abbildung 7.6: Lösung und Approximation von $\mathbf{G}(\mathbf{u}, \lambda, \mu) = \mathbf{0}$ für $\mu \in [-1, 1]$ mittels für $\mu = 0$ bestimmter interpolationsbasierter Reduktion

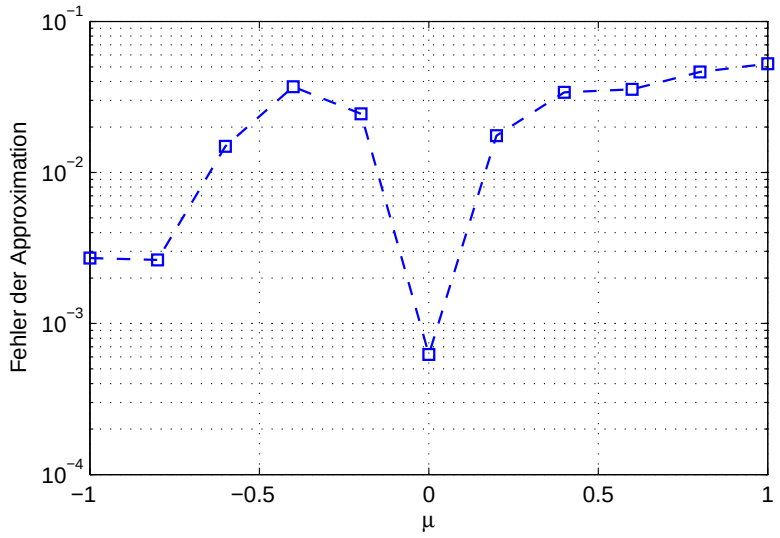

Abbildung 7.7: Fehler der Approximation bei $d = 3$ Interpolationsknoten für $\mu \in [-1, 1]$.

ten mittels des Newtonverfahrens erfolgen. Eine solche Korrektur ist für den Fall $\mu = 0.5$ in Abbildung 7.8 dargestellt.

Es kann möglich sein, die für den Fall $\mu = 0$ erzeugten Matrizen $\mathbf{V}_i$ weiter zu verwenden. Da aber zum Berechnen der korrigierten Knoten die Tangentialvektoren in der Nähe des neuen Knoten benötigt werden, können diese verwendet werden, um neue, besser zu den neuen Knoten passende, Matrizen $\mathbf{V}_i$ zu berechnen.

Bemerkung 7.1.1. *Die Lösungskurve des Bratu-Problems bewegt sich in nur wenigen Raumdimensionen. Daher ist ebenfalls eine Reduktion mit einem POD-Ansatzraum ohne zusätzliche Interpolation möglich. Der Testraum kann in diesem Fall einfach für einen beliebigen Punkt auf der Kurve berechnet werden. Das dies keinesfalls immer der Fall ist, zeigt das nächste Beispiel.*

7.2 Exotherme Reaktion

Das zweite Beispiel beschreibt eine exotherme Reaktion, die über das Randwertproblem

$$\begin{cases} -\Delta u = k_0(\mu - u)\exp\left(-\frac{\lambda}{1+u}\right) & \text{in } \Omega = (0,1)^2 \\ u = 0 & \text{auf } \partial\Omega \end{cases}$$

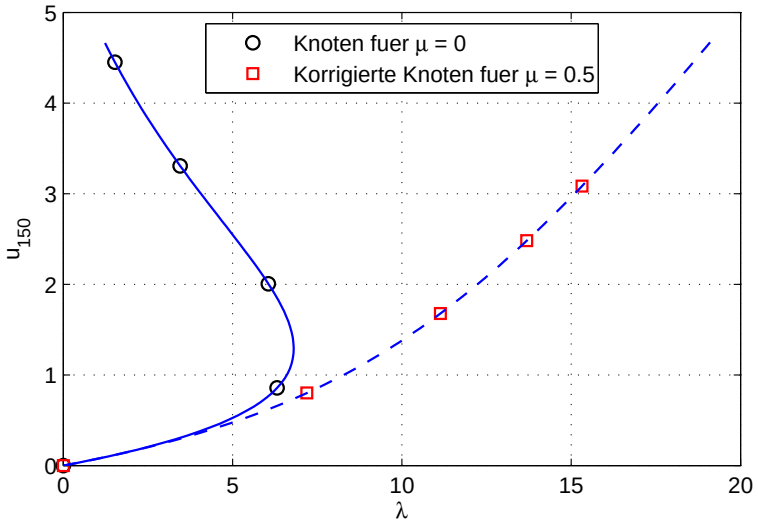

Abbildung 7.8: Korrektur der Interpolationsknoten

modelliert wird, wobei $k_0 = 10^7$ gewählt ist und die für den eindimensionalen Fall in [58] betrachtet wurde. Analog zum vorherigen Beispiel lässt sich dieses Problem diskretisieren und man erhält das nichtlineare Gleichungssystem

$$\mathbf{F}(\mathbf{x}, \mu) = \mathbf{0}$$

mit

$$\mathbf{F}(\mathbf{x}, \mu) = \mathbf{G}(\mathbf{u}, \lambda, \mu) := \mathbf{D}\mathbf{u} - k_0 \mathbf{f}(\mathbf{u}, \lambda, \mu),$$

wobei $\mathbf{D}$ die den (negativen) Laplace-Operator repräsentierende Matrix wie in Kapitel 7.1 bezeichnet. Die Funktion $\mathbf{f}$ ist definiert über

$$f_i(\mathbf{u}, \lambda, \mu) = (\mu - u_i) \exp\left(-\frac{\lambda}{1 + u_i}\right).$$

Für $\mu \in [0, 3.0]$ und $n = 400$ ist der Verlauf der Lösungskurven für die Komponente u_{150} in Abbildung 7.9 dargestellt. Die Wahl des Index 150 zur Betrachtung der Kurve ist deshalb sinnvoll, weil sich die Umkehrpunkte dort besonders gut erkennen lassen.

Anhand der hier betrachteten exothermen Reaktion lässt sich das Problem, das bei einer nicht interpolationsbasierten Reduktion mit festem Ansatz- und Testraum auftritt, gut verdeutlichen. Dafür wird der Fall $\mu = 1.2$ betrachtet. Über eine bezüglich s gleichverteilte Auswahl von 20 Knoten wird ein

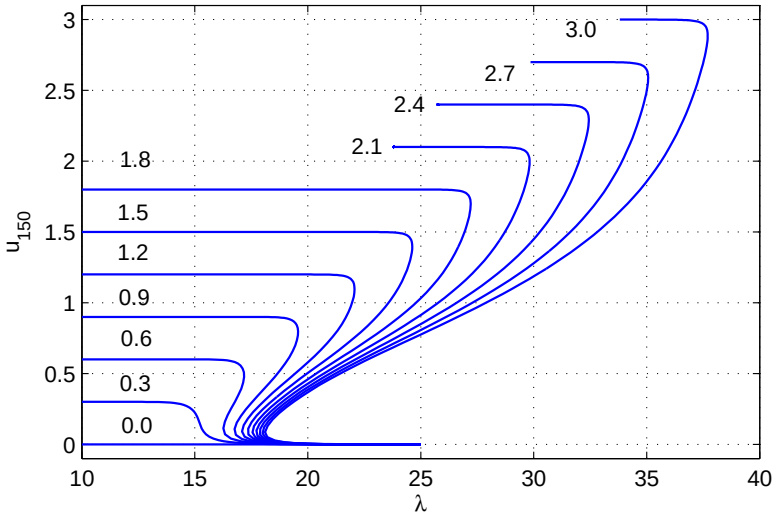

Abbildung 7.9: Lösungskurven für verschiedene $\mu \in [0, 2.1]$.

POD-Raum der Dimension 6 aufgebaut. Anstatt die Lösung einer interpolationsbasierten Reduktion zu berechnen, wird eine Reduktion mit einem festen Testraum, der bezüglich des Knotens 9 aufgebaut wurde erzeugt. Es zeigt sich, dass die Lösung der festen Reduktion in zwei separate Lösungskurven zerfällt. Dies ist im linken Teil der Abbildung 7.10 dargestellt.

Verwendet man stattdessen den interpolationsbasieren Ansatz, ergibt sich ein sehr viel besseres Ergebnis. Im rechten Teil der Abbildung 7.10 ist dieses Ergebnis dargestellt. Zur besseren Veranschaulichung wurde hier die Komponente u_{145} gewählt, da das Aufbrechen der Lösungskurve sich für den Index 145 besser nachvollziehen lässt, als für 150.

Es ist zu erkennen, dass auch die interpolationsbasierte Reduktion (obwohl sie zumindest nicht in Teilstücke zerfällt) keine besonders gute Approximation der volldimensionalen Kurve darstellt. Die liegt daran, dass die Dimension des Ansatzraumes noch zu niedrig ist, um dem im Vergleich zum vorherigen Beispiel komplexen Verlauf der Lösungskurve, gerecht zu werden. Daher wird von nun an ein POD-Ansatzraum der Dimension 16 für $\mu = 1.0$ verwendet um die Qualität der Parameterstudien bezüglich μ zu untersuchen.

Die Wahl der Knoten ist ausschlaggebend für die Interpolation und bei zu wenigen oder falsch verteilten Knoten können Probleme auftreten. Nutzt man zum Beispiel eine Menge von 4 bezüglich s gleichverteilten Knoten wie in Abbildung 7.11, zerfällt die Lösung des interpolierten Problems, wie in Abbildung 7.12 zu erkennen ist. Die numerischen Ergebnisse legen nahe, dass ein gestörter Bifurkationspunkt aufgetreten ist. Dies liegt vermutlich daran, dass

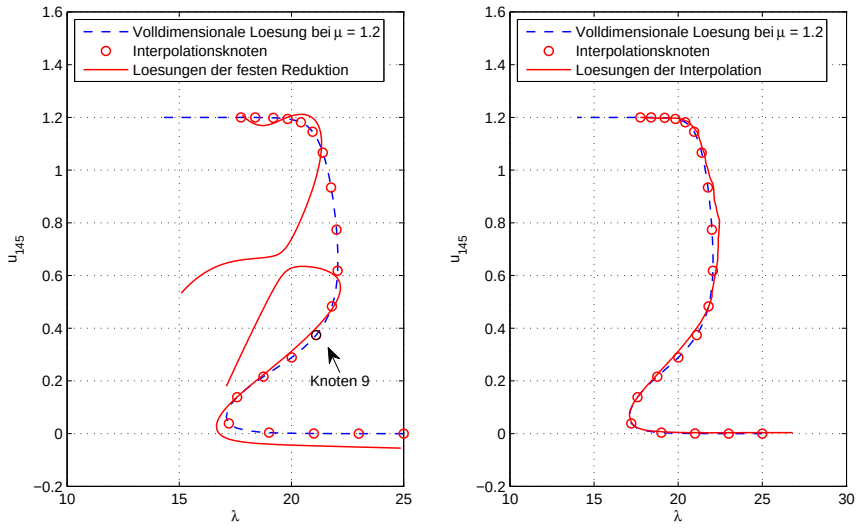

Abbildung 7.10: Vergleich zwischen Reduktion mit festem Testraum und interpolationsbasierter Reduktion

die Knoten, für die die Matrizen $\mathbf{V}_i$ aufgebaut werden, zu weit auseinander liegen bzw. Eigenschaft (v) von Definition 6.4.9 nicht erfüllt ist. Dadurch besitzt die Reduktion

$$\left(w_1(\mathbf{Z}\hat{\mathbf{x}})\mathbf{V}_1 + w_2(\mathbf{Z}\hat{\mathbf{x}})\mathbf{V}_2\right)^T \mathbf{F}(\mathbf{Z}\hat{\mathbf{x}}, 1.0) = \mathbf{0}$$

zusätzliche Lösungen, die offenbar zu einer solchen gestörten Bifurkation führen. Dieses Problem lässt sich durch Hinzunahme von zusätzlichen Knoten vermeiden, wie sich in Abbildung 7.13, in der 7 Knoten verwendet wurden, erkennen lässt.

Es ist anzunehmen, dass weniger die Menge und Entfernung der Knoten ausschlaggebend ist, als viel mehr die Lage auf der Kurve. Eine genauere Analyse, welche Punkte sich als Knoten besonders gut eignen, liegt bisher nicht vor. Nach Satz 6.4.16 lassen sich Bifurkationen aber in jedem Fall vermeiden, wenn der Abstand der Knoten nur gering genug ist.

Für $\mu = 1.0$ und 7 Interpolationsknoten wurde wie zuvor beim Bratu-Problem untersucht, ob sich die einmal aufgebaute interpolationsbasierte Reduktion zur Approximation der Lösungskurven für weitere μ eignen. Durch die im Vergleich zum in Kapitel 7.1 untersuchten Bratu-Problem höhere Anzahl von Interpolationsknoten sind die Möglichkeiten für Parameterstudien begrenzt, da die Träger der Gewichtsfunktionen kleiner sind. In Abbildung 7.14 sind die Lösungskurven der interpolationsbasierten Reduktion für einen Para-

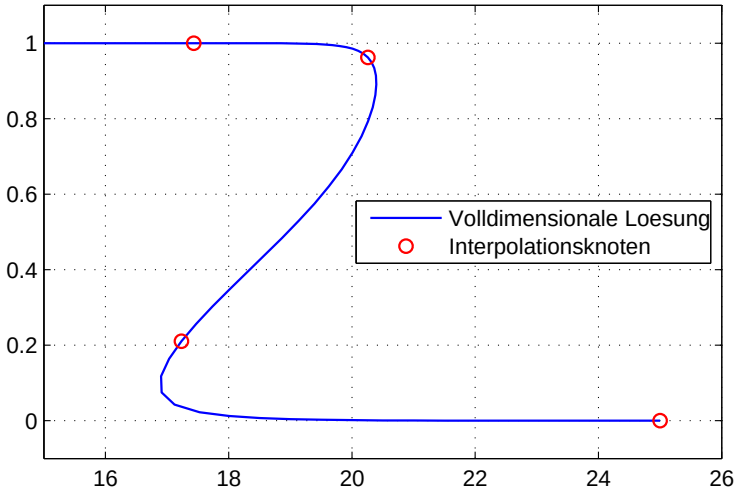

Abbildung 7.11: Lösungskurve für $\mu = 1.0$ mit 4 gleichverteilten Interpolationsknoten

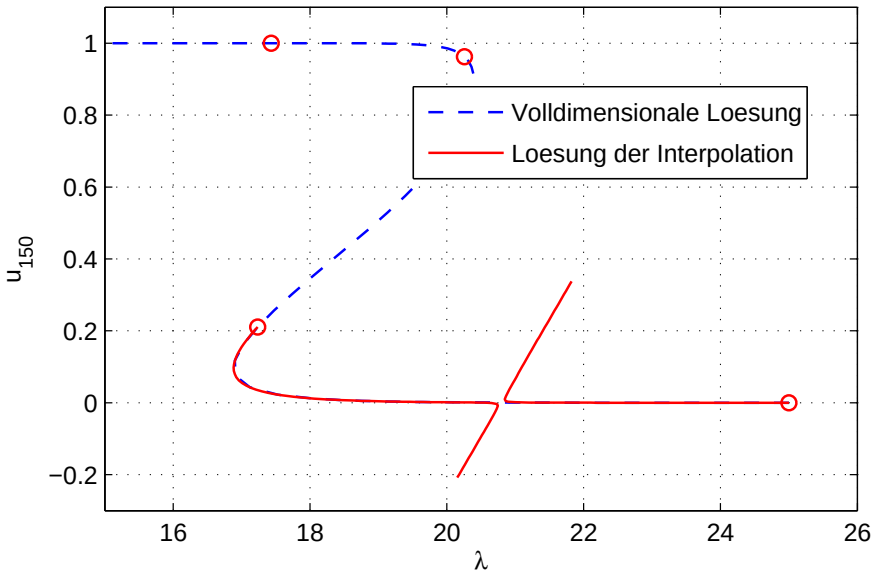

Abbildung 7.12: Gestörter Bifurkationspunkt

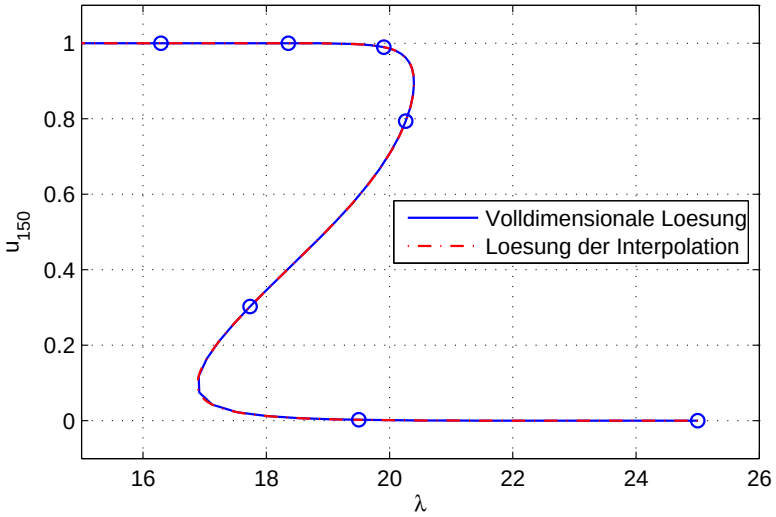

Abbildung 7.13: Interpolation mit 7 Knoten ohne Bifurkation

meterbereich $\mu \in [0.5, 1.5]$ abgebildet. Dort wo die Kurven stoppen, verlassen sie die Träger der Gewichtsfunktionen.

Die Fehler der jeweiligen Approximationen sind in Abbildung 7.15 dargestellt.

7.3 Bifurkationspunkte

In Kapitel 6 wurden Voraussetzungen benannt, unter denen sich die Existenz einer Lösung der interpolationsbasierten Reduktion theoretisch sichern lässt (vergleiche Voraussetzung 6.2.1). Diese Voraussetzungen schließen Kurven, die Bifurkationspunkte besitzen, aus, da in solchen Punkten der kleinste Singulärwert von **DF** Null wird und somit Bedingung (v) von Voraussetzung 6.2.1 nicht erfüllt ist. Dadurch lässt sich das grundlegende Werkzeug, der Satz über implizite Funktionen (Satz 2.1.2) nicht anwenden. Ein solcher Bifurkationspunkt existiert in der Lösungsmenge des in [37] betrachteten Randwertproblems

$$\begin{cases} -\Delta u = \lambda(u - u^2 - 3u^9 + u^{10}) & \text{in } \Omega = (0,1)^2, \\ u = 0 & \text{auf } \partial\Omega \end{cases}. \tag{7.2}$$

Wie bei den beiden vorherigen Beispielen wird das Problem wieder mittels Finite-Differenzen-Methode diskretisiert. Die erste Komponente der diskretisierten Lösung ist in Abbildung 7.16 dargestellt.

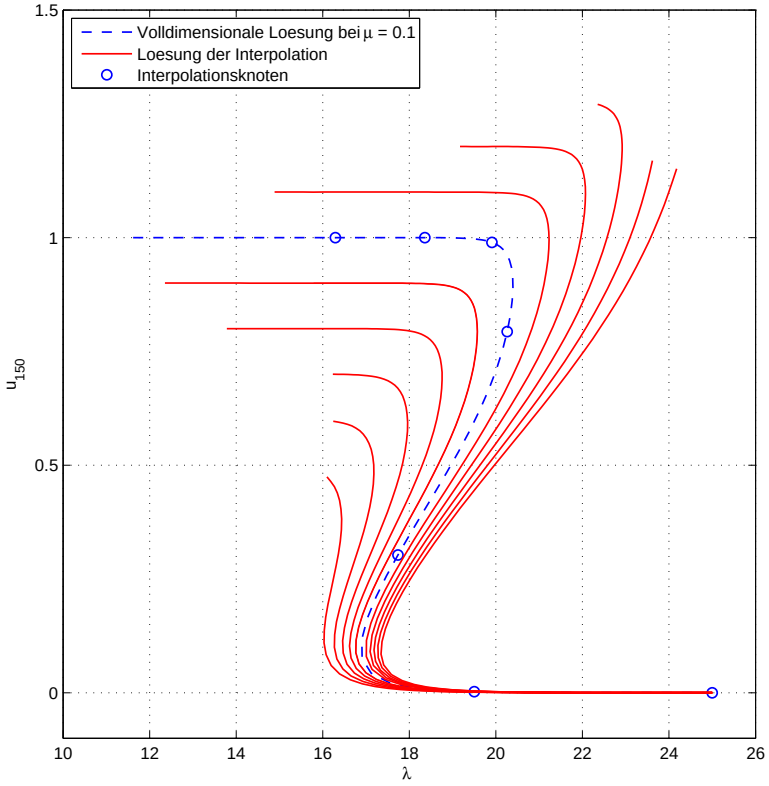

Abbildung 7.14: Lösungen der interpolationsbasierten Reduktion für $\mu \in [0.5, 1.5]$

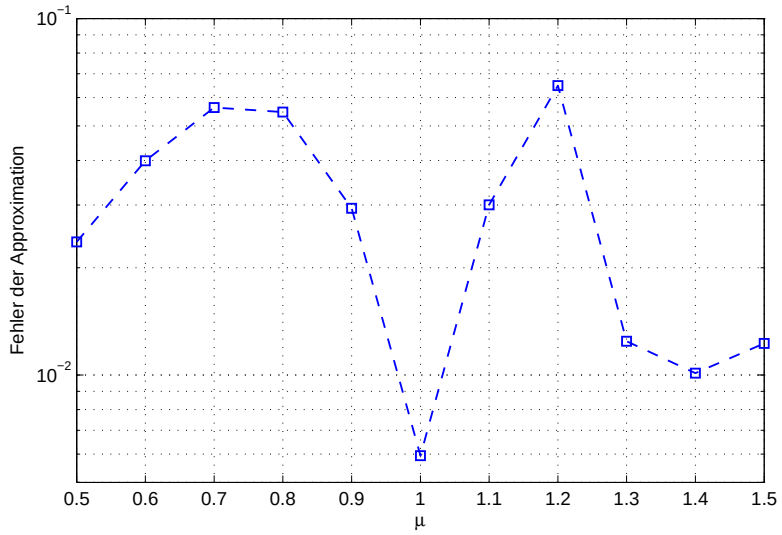

Abbildung 7.15: Fehler der Approximation bei $d = 7$ Interpolationsknoten für $\mu \in [0, 1.5]$

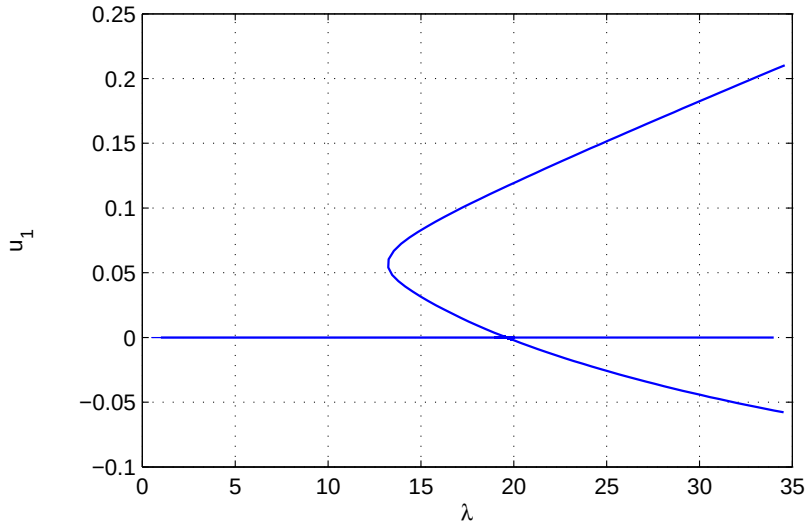

Abbildung 7.16: Lösungskurve von (7.2)

Mittels Prädiktor-Korrektor-Verfahren lassen sich unter Umständen nur Teile der Lösung erfassen (da solche Methoden in Bifurkationspunkten entweder nicht konvergieren, oder aber diese direkt überspringen). Anhand dieses Beispiels zeigt sich aber, dass die Kenntnis eines Teilstückes der Lösung bereits ausreichen kann, um eine sinnvolle interpolationsbasierte Reduktion zu erzeugen. Baut man mittels des in Abbildung 7.17 dargestellten oberen Teils der Lösungskurve einen POD-Ansatzraum der Größe 8 auf und wählt die 3 ebenfalls abgebildeten Knoten, so ergibt sich eine Reduktion, deren Lösungsgesamtheit in Abbildung 7.18 dargestellt ist. Die Verzweigung findet sich dabei immer im Ansatzraum wieder, da dieser als linearer Unterraum die Nulllösung enthält.

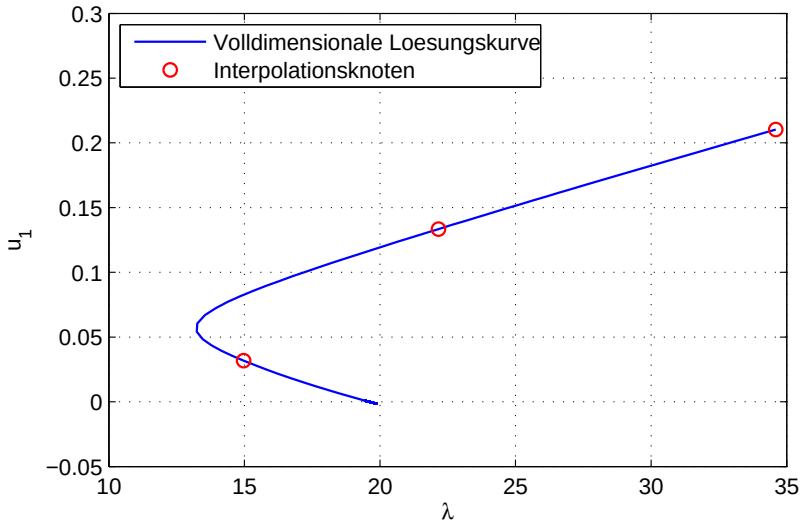

Abbildung 7.17: Oberer Teil der Lösungskurve und Interpolationsknoten

Es zeigt sich also, dass die restriktiven Forderungen die gestellt wurden, um die Existenz einer Lösung der interpolationsbasierten Reduktion zu sichern, in der Praxis nicht unbedingt erfüllt sein müssen, um gute Resultate zu liefern. Das Verfahren lässt sich somit auf manche Probleme anwenden, die bei dessen Entwicklung ursprünglich nicht in Betracht gezogen wurden.

Ergänzt man die ursprüngliche Differentialgleichung um einen zweiten Parameter μ zu

$$\begin{cases} -\Delta u = \lambda(u - u^2 - 3u^9 + u^{10} + \mu) & \text{in } \Omega = (0,1)^2, \\ u = 0 & \text{auf } \partial\Omega \end{cases}. \tag{7.3}$$

so ist $u \equiv 0$ für $\mu > 0$ keine Lösung mehr. Das diskretisierte Problem besitzt

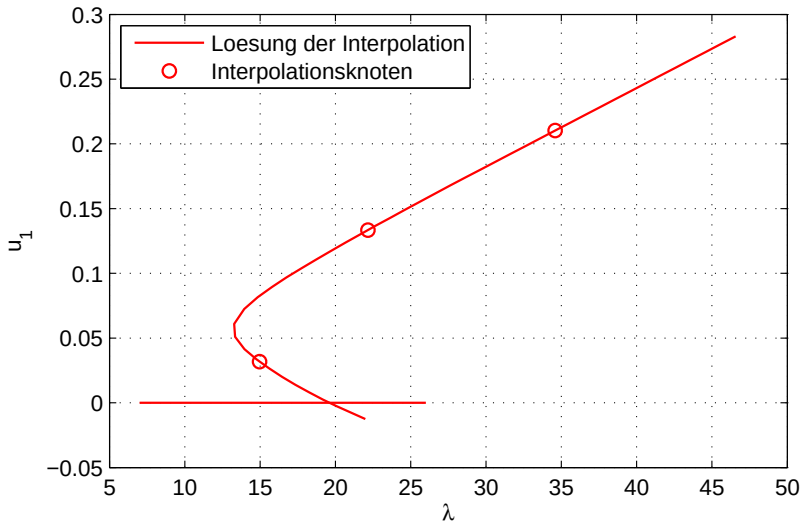

Abbildung 7.18: Approximation der Lösungskurve

nun keinen Bifurkationspunkt mehr und die Lösungsmenge zerfällt in zwei Teilkurven. Der Übergang von echten zu gestörten Bifurkationspunkten und das damit verbundene Aufbrechen der Lösungskurven wurde zum Beispiel in [55] untersucht. In Abbildung 7.19 sind die Lösungskurven um den ursprünglichen Birfuktationspunkt für eine Auswahl von Parametern μ dargestellt.

Wie bei den vorherigen Beispielen wird die für $\mu = 1.0$ aufgebaute interpolationsbasierte Reduktion mit 3 Knoten verwendet. Es zeigt sich, dass die Approximationen der linken Lösungskurve das Verhalten der volldimensionalen Lösungen gut approximieren, wie in Abbildung 7.20 dargestellt ist. Es ist jedoch ohne Anpassung der Interpolationsknoten unmöglich die rechten Lösungskurven zu approximieren, da die Träger der Gewichtsfunktionen nicht weit genug reichen.

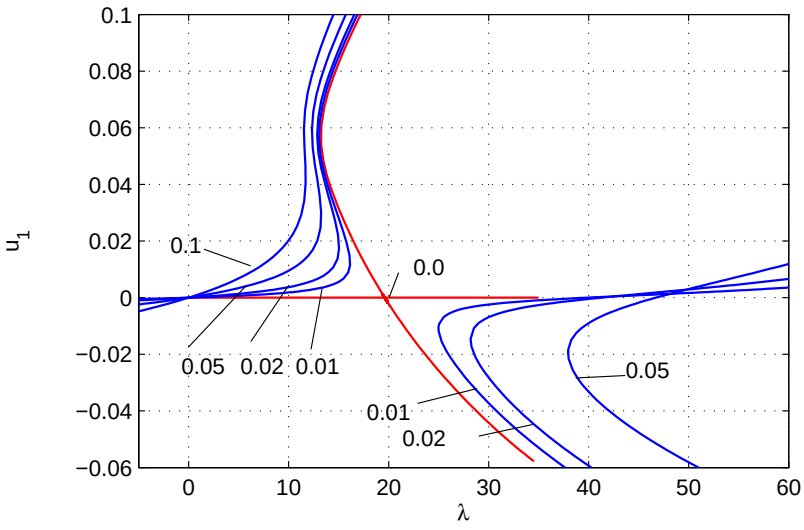

Abbildung 7.19: Lösungskurven für verschiedene μ

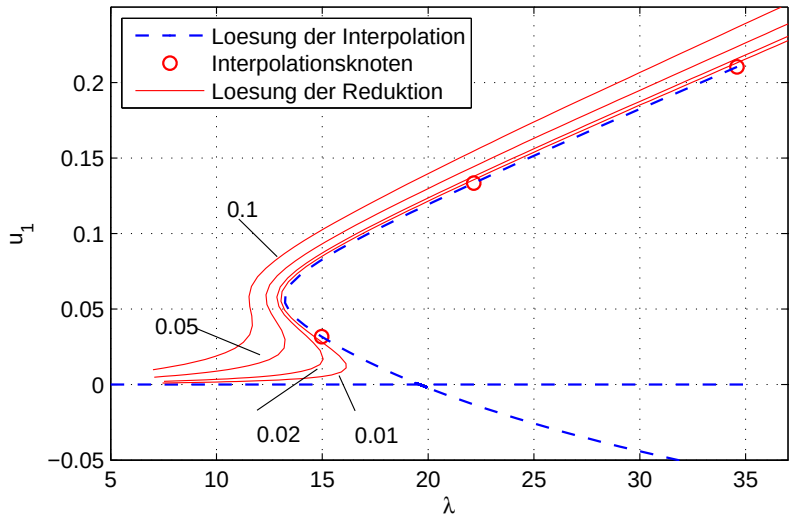

Abbildung 7.20: Lösungskurven der interpolationsbasierten Reduktion für verschiedene μ

Kapitel 8

Abschließende Betrachtungen

Im Folgenden wird auf die Methode der empirischen Interpolation eingegangen, die eine verbreitete Methode darstellt, den Rechenaufwand bei der Reduktion großer nichtlinearer Systeme, zu verringern. Es wird dargestellt, inwiefern diese sich mit der interpolationsbasierten Reduktion bzw. allgemein auf die in dieser Arbeit betrachteten Problemstellungen anwenden lässt. Zudem werden noch offene Fragen und Möglichkeiten zur Weiterentwicklung betrachtet.

8.1 Empirische Interpolation

Bei den globalen RB-Methoden ist die Verwendung der sogenannten empirischen Interpolation weit verbreitet. Vor allem bei Problemen, bei denen die Auswertung der zu reduzierenden Funktion kostenintensiv ist, hat sie sich als hilfreich zur Verringerung des Rechenaufwandes gezeigt. Erstmals findet sich die empirische Interpolation in [6] und wurde seit dem stetig weiter entwickelt, [76, 40, 24, 22, 66] und auf die verschiedendsten nichtlinearen Probleme angewendet, [69, 21, 18].

Es zeigt sich, dass sich die empirische Interpolation mit kleinen Abwandlungen auf die interpolationsbasierte Reduktion anwenden und sich so der Rechenaufwand noch einmal verringern lässt.

Die Grundidee besteht darin, den nichtlinearen Anteil das Ausgangsproblem selbst durch eine Approximation zu ersetzen. In dieser Arbeit beschränkt man sich auf die diskrete Variante des Verfahrens nach [14], die nichtlineare Probleme der Form

$$\mathbf{F}(\mathbf{u}(\lambda)) = \mathbf{A}\mathbf{u}(\lambda) + \mathbf{N}(\mathbf{u}(\lambda))$$

betrachtet, wobei $\mathbf{A}$ den linearen und $\mathbf{N}$ den nichtlinearen Teil des Systems bezeichnen. Es wird dabei, wie bei den Snapshot-Verfahren üblich, davon ausgegangen, dass sich die Lösung des parameterabhängigen nichtlinearen Systems bezüglich λ parametrisieren lässt.

Im Folgenden wird eine Version der empirischen Interpolation für die in dieser Arbeit betrachteten nichtlinearen Systeme der Form $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ mit $\mathbf{F} : \mathbb{R}^{n+1} \to \mathbb{R}^n$ hergeleitet.

Dazu sei zunächst $\mathbf{F}$ zerlegbar in einen linearen und einen nichtlinearen Anteil:

$$\mathbf{F}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{N}(\mathbf{x}).$$

Dies ist bei den in Kapitel 7 betrachteten Beispielen offenbar der Fall, da diese Struktur sich gewöhnlich bei diskretisierten partiellen Differentialgleichungen findet. Ist die Auswertung von $\mathbf{F}$ mit großem numerischen Aufwand verbunden, liegt dies am nichtlinearen Anteil $\mathbf{N}$. Ziel ist es, diesen zu approximieren, was in zwei Teilschritten geschieht.

Als erstes wird ein geeigneter Approximationsraum $\mathcal{Q}$ aufgebaut und als zweites eine Auswahl von k Indizes $\theta_j \in \{1, \ldots, n\}$ zur Reduktion der Anzahl der Gleichungen getroffen. Repräsentiert die Matrix $\mathbf{Q}$ den Raum $\mathcal{Q}$ und die Projektionsmatrix $\mathbf{P}$ die ausgewählten Indizes (indem ihre Spalten aus den zu den jeweiligen Indizes gehörenden Einheitsvektoren besteht), so ergibt sich die Interpolation über

$$\tilde{\mathbf{N}}(\mathbf{x}) = \mathbf{Q}(\mathbf{P}^T\mathbf{Q})^{-1}\mathbf{P}^T\mathbf{N}(\mathbf{x}).$$

Dabei reduziert sich die Rechenzeit dadurch, dass $\mathbf{P}^T\mathbf{N}(\mathbf{x})$ nur noch k Zeilen besitzt, die Funktion $\mathbf{N}$ also nur noch k-mal ausgewertet werden muss, während die Matrix $\mathbf{Q}(\mathbf{P}^T\mathbf{Q})^{-1}$ nur einmal berechnet wird. Mit Blick auf eine Online-Offline-Unterscheidung findet dieser Prozess also in der Offline-Phase statt.

Aufbau von $\mathbf{Q}$

Als Ausgangssituation dient wieder eine Menge von Snapshots $X = \{\mathbf{x}_i := \mathbf{x}(s_i),\ i = 1, \ldots, q\}$ einer Lösungskurve $\mathbf{c} : S \to \mathbb{R}^{n+1}$ mit $\mathbf{F}(\mathbf{c}(s)) = \mathbf{0}$. Für diese betrachtet man nun die Menge $N = \{\mathbf{n}_i := \mathbf{N}(\mathbf{x}_i),\ i = 1, \ldots, q\}$, die die Funktionswerte des nichtlinearen Anteils in den Snapshots enthält.

Als Startpunkt wird $p_1 := \arg\max_{j \in \{1,\ldots,m\}} \|\mathbf{n}_j\|_2$ gewählt und $\mathbf{Q}_1 = \mathbf{q}^1 = \mathbf{n}_{p_1}$ gesetzt. Im nächsten Schritt wird der Punkt aus N gesucht, der bezüglich der $\|\cdot\|_2$-Norm am weitesten von $R(\mathbf{Q}_1)$ entfernt liegt, also

$$p_2 = \arg \max_{j \in \{1,\ldots,q\}} \|\mathbf{n}_j - \mathbf{Q}_1(\mathbf{Q}_1^T\mathbf{Q}_1)^{-1}\mathbf{Q}_1\mathbf{n}_j\|_2. \tag{8.1}$$

Dann wird $\mathbf{q}_2 := \mathbf{n}_{p_2}$ und $\mathbf{Q}_2 := (\mathbf{q}_1, \mathbf{q}_2)$ gesetzt. Dieser Vorgang wird bis zu einer festgelegten Schrittzahl k fortgesetzt und $\mathbf{Q} := \mathbf{Q}_k = (\mathbf{q}_1, \ldots, \mathbf{q}_k)$ definiert.

Aufbau von $\mathbf{P}$

Im Raum $\mathcal{Q}$ wird eine Approximation des nichtlinearen Anteils $\mathbf{N}(s)$ gesucht,

also eine Funktion $\mathbf{g} : S \to \mathbb{R}^k$, sodass

$$\mathbf{N}(\mathbf{c}(s)) \approx \mathbf{Q}\mathbf{g}(s)$$

gilt. Der nichtlineare Anteil entlang der Lösungskurve soll somit möglichst gut approximiert wird. Da dieses System stets überbestimmt ist, werden k Indizes $\vartheta_j, j = 1, \ldots, k$ ausgewählt, für die gelten soll

$$\mathbf{N}_{\vartheta_j}(s) = (\mathbf{Q}\mathbf{g}(s))_{\vartheta_j}, j = 1, \ldots, k.$$

Dies lässt sich mit der Matrix $\mathbf{P} = [\mathbf{e}_{\vartheta_1}, \ldots, \mathbf{e}_{\vartheta_k}]$ (wobei $\mathbf{e}_i$ den i-ten Einheitsvektor beschreibt) dann schreiben als

$$\mathbf{P}^T\mathbf{N}(\mathbf{c}(s)) = \mathbf{P}^T\mathbf{Q}\mathbf{g}(s).$$

Zur Auswahl der ϑ_j wird der in [14] vorgestellte DEIM-Algorithmus verwendet. Zunächst setzt man $\vartheta_1 = \arg\max_{i=1,\ldots,n} |(\mathbf{q}_1)_i|$, sowie $\mathbf{Q}_1 = \mathbf{q}_1, \mathbf{P}_1 = \mathbf{e}_{\vartheta_1}$. Nun lösen wir das Gleichungssystem

$$(\mathbf{P}_1^T\mathbf{Q}_1)\mathbf{c}_1 = \mathbf{P}_1^T\mathbf{q}_1$$

für den Vektor $\mathbf{c}_1$ und definieren damit den Vektor $\mathbf{r}_1 := \mathbf{q}_1 - \mathbf{Q}_1\mathbf{c}_1$, der an der Stelle ϑ_1 den Wert 0 hat. Im nächsten Schritt wird dann
$\vartheta_2 = \arg\max_{i=1,\ldots,n} |(\mathbf{r}_1)_i|$ und $\mathbf{Q}_2 = (\mathbf{q}_1, \mathbf{q}_2), \mathbf{P}_2 = (\mathbf{e}_{\vartheta_1}, \mathbf{e}_{\vartheta_2})$ gesetzt. Das Verfahren setzt man nun fort, bis man k Indizes $\vartheta_1, \ldots, \vartheta_k$ erhalten hat um dann schließlich $\mathbf{P} := \mathbf{P}_k = (\mathbf{e}_{\vartheta_1}, \ldots, \mathbf{e}_{\vartheta_k})$ zu definieren.

Somit lässt sich nun eine Interpolation von $\mathbf{N}$ erzeugen, die entlang der Lösungskurve $\mathbf{c}$ in den Zeilen $\theta_1, \ldots, \theta_k$ mit $\mathbf{N}(\mathbf{c}(s))$ überein stimmt. Angewendet auf das Ausgangsproblem ergibt sich

$$\tilde{\mathbf{F}}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{Q}(\mathbf{P}^T\mathbf{Q})^{-1}\mathbf{P}^T\mathbf{N}(\mathbf{x}) = \mathbf{0}. \tag{8.2}$$

Es gilt nun nicht mehr $\mathbf{F}(\mathbf{c}(s)) = \mathbf{0}$, sondern $\mathbf{P}^T\mathbf{F}(\mathbf{c}(s)) = \mathbf{0}$. Durch die Auswahl der Indizes sind die nicht durch $\mathbf{P}$ repräsentierten Einträge von $\mathbf{F}(\mathbf{c}(s))$ jedoch klein.

Die Verknüpfung dieses Verfahrens mit der interpolationsbasierten Reduktion erscheint deshalb sinnvoll, weil beide eine Menge von Snapshots, also Lösungspunkten auf der Kurve $\mathbf{c}$, verwenden um die jeweiligen Approximationsräume aufzubauen. Des Weiteren wurden in Kapitel 6.4.2 bereits Existenzaussagen bei der Verwendung inexakter Interpolationsknoten gemacht, also für den Fall, dass die Funktionswerte in den verwendeten Knoten nicht $\mathbf{0}$ sind, aber beliebig klein werden können. Die Ausgangssituation ist also vergleichbar mit dem Ergebnis der empirischen Interpolation. Eine genauere Untersuchung bezüglich hinreichender Bedingungen für die Lösungsexistenz einer so reduzierten Funktion steht noch aus.

Bemerkung 8.1.1. *Der beim Aufbau des Approximationsraums $\mathcal{Q}$ verwendete Algorithmus entspricht im Wesentlichen dem Greedy-Algorithmus aus Kapitel 4.1. Alternativ kann auch das POD-Verfahren verwendet werden um anstelle einer Auswahl von k Vektoren der Menge N einen k-dimensionalen Raum zu erzeugen, der N möglichst gut approximiert. In der Ursprungsarbeit [6], in der die empirische Interpolation auf die noch nicht diskretisierte Ausgangsgleichung angewendet wurde, wurde für den Greedy-Algorithmus die L_∞-Norm verwendet. Da im diskreten Fall jedoch orthogonale Projektionen verwendet werden, ist es sinnvoll aus Konsistenzgründen stattdessen die $\|\cdot\|_2$ Norm zu verwenden. Zudem ist das Problem (8.1) für die Maximums-Norm nicht wohldefiniert.*

8.2 Weiterführende Betrachtungen

Zum Ende einer Arbeit führt die tiefe Vertrautheit mit einem Thema oft dazu, dass zahlreiche Erweiterungen auf der Hand zu liegen scheinen. Daher werden nun hier einige sich aufdrängende mögliche Fortsetzungen der Untersuchungen aufgelistet. Bevor dies geschieht, sei aber darauf verwiesen, dass die saubere Behandlung dieser weiterführenden Ideen bei genauer Betrachtung aufwändige Analysen nach sich ziehen.

- Zunächst wird ein Blick auf die für die interpolationsbasierte Reduktion verwendeten Funktionen geworfen. In dieser Arbeit sind die Träger dieser Funktionen Kugeln im $\mathbb{R}^{n+1}$, deren Mittelpunkt in den Interpolationsknoten liegt. Alternativ lassen sich die Träger auch als Polytope bestimmen (vergleiche [71]). Dabei kann die Ausdehnung des Trägers in die einzelnen Raumrichtungen separat festgelegt werden (und nicht über einen festen Radius). Dies hat den Vorteil, dass die $\mathbf{u}$- und λ-Komponente von $\mathbf{x} = (\mathbf{u}^T, \lambda)^T \in \mathbb{R}^{n+1}$ getrennt voneinander betrachtet werden können. Dies ist sinnvoll, da der Parameterraum Λ und der Lösungsraum Y physikalisch nichts miteinander zu tun haben und große Unterschiede in den Skalierungen aufweisen können.

- Eine tiefer gehende Analyse der Auswahl der Interpolationsknoten könnte die interpolationsbasierte Reduktion weiter verbessern. In dieser Arbeit wurden die Knoten entweder über einen Greedy-Algorithmus bestimmt oder gleichverteilt bezüglich des Parameters s gewählt. Im Beispiel aus Kapitel 7.2 wurde gezeigt, dass es manchmal zum Zerfallen der Approximation der Lösungskurve kommen kann, wenn zu wenige Interpolationsknoten verwendet werden. Es ist nicht auszuschließen, dass es Möglichkeiten der Auswahl der Interpolationsknoten gibt, die Eigenschaften der Lösungskurve berücksichtigt um das Zerfallen auch bei der

Verwendung weniger Knoten zu vermeiden. Da man generell daran interessiert ist, mit möglichst wenigen Knoten auszukommen, ist eine nähere Untersuchung, welche Punkte auf der Lösungskurve sich gut als Interpolationsknoten eignen, sinnvoll.

- In Kapitel 7 wurden zweiparametrige Systeme betrachtet und Beispiele gezeigt, bei denen für einen festen Parameter aufgebaute Reduktionen für Parameterstudien verwendet werden können. Untersuchungen, inwieweit sich die Reduktion für Parameterbereiche, für die sie nicht direkt weiter verwendet werden kann, kostengünstig anpassen lässt, könnten zu einer zusätzlichen Verringerung der Rechenzeit führen. Ziel sollte es sein, mit möglichst geringem numerischen Aufwand, den Ansatzraum, die Testräume und die Interpolationsknoten anzupassen.

- Schließlich bedarf es noch einer genaueren Untersuchung der Verknüpfung der interpolationsbasierten Reduktion mit der empirischen Interpolation aus Kapitel 8.1. Zwar wurde die Grundlage durch die Anpassung des Verfahrens an allgemeine nichtlineare Probleme geschaffen, ein Satz analog zu den Existenzsätzen aus Kapitel 6 existiert jedoch noch nicht. Während Berührungs- und Verbindungspunkte beider Ansätze ganz offensichtlich erscheinen, wäre für eine wirkliche Hybridmethode erhebliche Arbeit zu leisten. Auch wenn die feste Überzeugung besteht, dass solch eine Verbindung die Vorteile beider Methoden erben könnte, müssen diese Konstruktionen deshalb späteren Untersuchungen vorbehalten bleiben.

Literaturverzeichnis

[1] Abel, J. F. ; Shephard, M. S.: An algorithm for multipoint constraints in finite element analysis. In: *International Journal for Numerical Methods in Engineering* 14 (1979), Nr. 3, S. 464–467. – ISSN 1097–0207

[2] Allgower, E. ; Georg, K. : *Introduction to Numerical Continuation Methods*. Society for Industrial and Applied Mathematics, 2003

[3] Almroth, B. O. ; Brogan, F. A. ; Stern, P. : Automatic choice of global shape functions in structural analysis. In: *Aiaa Journal* 16 (1978), S. 525–528

[4] Antoulas, A. : *Approximation of Large-Scale Dynamical Systems*. Society for Industrial and Applied Mathematics, 2005

[5] Bank, R. E. ; Dupont, T. F. ; Yserentant, H. : The hierarchical basis multigrid method. In: *Numerische Mathematik* 52 (1988), Nr. 4, S. 427–458. – ISSN 0029–599X

[6] Barrault, M. ; Maday, Y. ; Nguyen, N. C. ; Patera, A. T.: An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations. In: *Comptes Rendus Mathematique* 339 (2004), Nr. 9, S. 667 – 672. – ISSN 1631–073X

[7] Barrett, A. ; Reddien, G. : On the Reduced Basis Method. In: *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik* 75 (1995), Nr. 7, S. 543–549. – ISSN 1521–4001

[8] Binev, P. ; Cohen, A. ; Dahmen, W. ; DeVore, R. ; Petrova, G. ; Wojtaszczyk, P. : Convergence Rates for Greedy Algorithms in Reduced Basis Methods. In: *SIAM Journal on Mathematical Analysis* 43 (2011), Nr. 3, S. 1457–1472

[9] Brezzi, F. ; Rappaz, J. ; Raviart, P. : Finite dimensional approximation of nonlinear problems. In: *Numerische Mathematik* 36 (1980), Nr. 1, S. 1–25

[10] Briggs, W. ; Henson, V. ; McCormick, S. : *A Multigrid Tutorial, Second Edition.* Society for Industrial and Applied Mathematics, 2000

[11] Bunse-Gerstner, A. ; Byers, R. ; Mehrmann, V. ; Nichols, N. K.: Numerical computation of an analytic singular value decomposition of a matrix valued function. In: *Numerische Mathematik* 60 (1991), Nr. 1, S. 1–39. – ISSN 0029–599X

[12] Cari, E. P. T. ; Theodoro, A. R. ; Miholaro, A. P. ; Bretas, N. G. ; Alberto, L. F. C.: Trajectory Sensitivity Method and Master-Slave Synchronization to Estimate Parameters of Nonlinear Systems. In: *Mathematical Problems in Engineering* (2009)

[13] Chan, T. F. ; Keller, H. B.: Arc-Length Continuation and Multigrid Techniques for Nonlinear Elliptic Eigenvalue Problems. In: *SIAM Journal on Scientific and Statistical Computing* 3 (1982), Nr. 2, S. 173–194

[14] Chaturantabut, S. ; Sorensen, D. : Nonlinear Model Reduction via Discrete Empirical Interpolation. In: *SIAM Journal on Scientific Computing* 32 (2010), Nr. 5, S. 2737–2764

[15] Christensen, E. A. ; Brons, M. ; Sorensen, J. N.: Evaluation of Proper Orthogonal Decomposition–Based Decomposition Techniques Applied to Parameter-Dependent Nonturbulent Flows. In: *SIAM Journal on Scientific Computing* 21 (1999), Nr. 4, S. 1419–1434

[16] Coleman, T. F. ; Sorensen, D. C.: A note on the computation of an orthonormal basis for the null space of a matrix. In: *Mathematical Programming* 29 (1984), Nr. 2, S. 234–242. – ISSN 0025–5610

[17] Dickson, K. I. ; Kelley, C. T. ; Ipsen, I. C. F. ; Kevrekidis, I. G.: Condition Estimates for Pseudo Arclength Continuation. In: *SIAM Journal on Numerical Analysis* 45 (2007), Nr. 1, S. 263–276

[18] Drohmann, M. ; Haasdonk, B. ; Ohlberger, M. : Reduced basis model reduction of parametrized two-phase flow in porous media. In: *Submitted to the Proceedings of Mathmod* (2012)

[19] Eckart, C. ; Young, G. : The approximation of one matrix by another of lower rank. In: *Psychometrika* 1 (1936), Nr. 3, S. 211–218

[20] Fink, J. P. ; Rheinboldt, W. C.: On the Error Behavior of the Reduced Basis Technique for Nonlinear Finite Element Approximations. In: *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik* 63 (1983), Nr. 1, S. 21–28. – ISSN 1521–4001

[21] Grepl, M. A.: *Bericht / Institut für Geometrie und Praktische Mathematik, RWTH Aachen*. Bd. 322: *Model order reduction of parametrized nonlinear reaction-diffusion systems*. Aachen : Institut für Geometrie und Praktische Mathematik, RWTH Aachen, 2011. – 45 S. : graph. Darst.

[22] Grepl, M. A. ; Maday, Y. ; Nguyen, N. C. ; Patera, A. T.: Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 41 (2007), Nr. 3, S. 575–605

[23] Grepl, M. A. ; Maday, Y. ; Nguyen, N. C. ; Patera, A. T.: Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 41 (2007), Nr. 3, S. 575–605

[24] Haasdonk, B. ; Ohlberger, M. ; Rozza, G. : A Reduced Basis Method for Evolution Schemes with Parameter-Dependent Explicit Operators. In: *ETNA, Electronic Transactions on Numerical Analysis* 32 (2008), S. 145–168

[25] Haasdonk, B. ; Ohlberger, M.: Reduced basis method for finite volume approximations of parametrized linear evolution equations. In: *ESAIM: M2AN* 42 (2008), Nr. 2, S. 277–302

[26] Hackbusch, W. : *Springer Series in Computational Mathematics*. Bd. 4: *Multi-Grid Methods and Applications*. Springer-Verlag Berlin Heidelberg, 1985

[27] Ito, K. ; Ravindran, S. : A Reduced-Order Method for Simulation and Control of Fluid Flows. 1998 (2). – Forschungsbericht. – 403 – 425 S.. – ISSN 0021–9991

[28] Kapania, R. : A pseudo-spectral solution of 2-parameter Bratu's equation. In: *Computational Mechanics* 6 (1990), Nr. 1, S. 55–63. – ISSN 0178–7675

[29] Kato, T. : *Pertubation Theory for Linear Operators*. Springer-Verlag Berlin Heidelberg, 1995

[30] Kelejian, H. H.: Aggregation and Disaggregation of Nonlinear Equations. In: *Evaluation of Econometric Models*. National Bureau of Economic Research, Inc, 1980, S. 135–152

[31] Kelley, C. T.: *Iterative Methods for Nonlinear Equations*. SIAM, 1995 (Frontiers in Applied Mathematics 16)

[32] Kitagawa, K. ; Nakamura, H. ; Yagawa, G. : Comparison between substructure method and domain decomposition method. In: Sloot, P. (Hrsg.) ; Bubak, M. (Hrsg.) ; Hertzberger, B. (Hrsg.): *High-Performance Computing and Networking* Bd. 1401. Springer Berlin Heidelberg, 1998, S. 358–367

[33] Kubicek, M. ; Mark, M. : *Computational Methods in Bifurcation Theory and Dissipative Structures.* Springer-Verlag, 1983

[34] Kunisch, K. ; Volkwein, S. : Galerkin Proper Orthogonal Decomposition Methods for a General Equation in Fluid Dynamics. In: *SIAM Journal on Numerical Analysis* 40 (2002), Nr. 2, S. 492–515

[35] Lass, O. ; Volkwein, S. : POD Galerkin Schemes for Nonlinear Elliptic-Parabolic Systems. In: *SIAM Journal on Scientific Computing* 35 (2013), Nr. 3, S. A1271–A1298

[36] Leung, A. Y. T.: A simple dynamic substructure method. In: *Earthquake Engineering and Structural Dynamics* 16 (1988), Nr. 6, S. 827–837. – ISSN 1096–9845

[37] Lions, P. L.: On the Existence of Positive Solutions of Semilinear Elliptic Equations. In: *SIAM Review* 24 (1982), Nr. 4, S. 441–467

[38] Mackens, W. : *Kondensation großer nichtlinearer Gleichungssysteme mit der Methode der reduzierten Basen.* Habilitationsschrift, Aachen, 1988

[39] Maday, Y. : Reduced Basis Method for the Rapid and Reliable Solution of Partial Differential Equations. In: *Proceedings of the International Conference of Mathematicians* Eurpoean Mathematical Society, 2006, S. 1–17

[40] Maier, I. ; Haasdonk, B. : A Dirichlet–Neumann reduced basis method for homogeneous domain decomposition problems. In: *Applied Numerical Mathematics* 78 (2014), S. 31 – 48

[41] Melenk, J. ; Babuška, I. : The partition of unity finite element method: Basic theory and applications. In: *Computer Methods in Applied Mechanics and Engineering* 139 (1996), Nr. 1–4, S. 289 – 314

[42] Noor, A. K.: Recent Advances and Applications of Reduction Methods. In: *Applied Mechanics Reviews* 47 (1994), S. 125–146

[43] Noor, A. K. ; Balch, C. D. ; Shibut, M. A.: Reduction methods for nonlinear steady-state thermal analysis. In: *International Journal for Numerical Methods in Engineering* 20 (1984), Nr. 7, S. 1323–1348

[44] Noor, A. K. ; Peters, J. M.: Reduced Basis Technique for Nonlinear Analysis of Structures. 1980. – Forschungsbericht. – 455–462 S.

[45] Ortega, J. ; Rheinboldt, W. : *Iterative Solution of Nonlinear Equations in Several Variables*. Society for Industrial and Applied Mathematics, 2000

[46] Parlett, B. : *The Symmetric Eigenvalue Problem*. Society for Industrial and Applied Mathematics, 1998

[47] Patera, A. T. ; Rozza, G. : *Reduced Basis Approximation and A Posteriori Error Estimation for Parametrized Partial Differential Equations*. Version 1.0, Copyright MIT 2006, to appear in (tentative rubric) MIT Pappalardo Graduate Monographs in Mechanical Engineering, 2006

[48] Pearson, K. : LIII. On lines and planes of closest fit to systems of points in space. In: *Philosophical Magazine Series 6* 2 (1901), Nr. 11, S. 559–572

[49] Porsching, T. A.: Estimation of the Error in the Reduced Basis Method Solution of Nonlinear Equations. In: *Mathematics of Computation* 45 (1985), Nr. 172, S. 487–496

[50] Prud'homme, C. ; Rovas, D. V. ; Veroy, K. ; Machiels, L. ; Maday, Y. ; Patera, A. T. ; Turinici, G. : Reliable Real-Time Solution of Parametrized Partial Differential Equations: Reduced-Basis Output Bound Methods. In: *Journal of Fluids Engineering-transactions of The Asme* 124 (2002)

[51] Quarteroni, A. ; Rozza, G. : Numerical solution of parametrized Navier–Stokes equations by reduced basis methods. In: *Numerical Methods for Partial Differential Equations* 23 (2007), Nr. 4, S. 923–948

[52] Quarteroni, A. M. ; Valli, A. : *Numerical Approximation of Partial Differential Equations*. 1st ed. 1994. 2nd printing. Springer Publishing Company, Incorporated, 2008

[53] Rathinam, M. ; Petzold, L. R.: A New Look at Proper Orthogonal Decomposition. In: *SIAM Journal on Numerical Analysis* 41 (2003), Nr. 5, S. 1893–1925

[54] Raudenbush, S. : *HLM 6: Hierarchical Linear and Nonlinear Modeling*. Scientific Software Int. Incorporated, 2004

[55] Reiss, E. L.: Imperfect Bifurcation. In: Rabinowitz, P. H. (Hrsg.): *Proceedings of an Advanced Seminar Conducted by The Mathematics Research Center*, Academic Press, Inc., 1976, S. 37–72

[56] Rewienski, M. J.: *A Trajectory Piecewise-Linear Approach to Model Order Reduction Nonlinear Dynamical Systems*, Massachussets Institute of Technology, Diss., 2003

[57] Rheinboldt, W. C.: *Numerical Analysis of Parametrized Nonlinear Equations*. New York, NY, USA : Wiley-Interscience, 1986

[58] Rheinboldt, W. C.: On the computation of multi-dimensional solution manifolds of parametrized equations. In: *Numerische Mathematik* 53 (1988), Nr. 1-2, S. 165–181. – ISSN 0029–599X

[59] Rheinboldt, W. C.: On the Theory and Error Estimation of the Reduced Basis Method for Multi-parameter Problems. In: *Nonlinear Anal.* 21 (1993), Nr. 11, S. 849–858. – ISSN 0362–546X

[60] Rheinboldt, W. C.: Solution Fields of Nonlinear Equations and Continuation Methods. In: *SIAM Journal on Numerical Analysis* 17 (1980), Nr. 2, S. 221–237

[61] Rheinboldt, W. C.: Numerical continuation methods: a perspective. In: *Journal of Computational and Applied Mathematics* 124 (2000), Nr. 1–2, S. 229 – 244. – ISSN 0377–0427. – Numerical Analysis 2000. Vol. IV: Optimization and Nonlinear Equations

[62] Rozza, G. : An introduction to reduced basis method for parametrized PDEs. In: *Applied and Industrial Mathematics in Italy* Bd. III, WorldScientific, 2009 (Advances in Mathematics for Applied Sciences)

[63] Rozza, G. ; Huynh, D. ; Patera, A. : Reduced Basis Approximation and a Posteriori Error Estimation for Affinely Parametrized Elliptic Coercive Partial Differential Equations. In: *Archives of Computational Methods in Engineering* 15 (2008), Nr. 3, S. 229–275. – ISSN 1134–3060

[64] Schmidt, J. W.: Zur Konvergenz von kubischen Interpolationssplines. In: *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik* 58 (1978), Nr. 2. – ISSN 1521–4001

[65] Schweitzer, P. J. ; Kindle, K. W.: An iterative aggregation-disaggregation algorithm for solving linear equations. 1986. – ISSN 0096–3003, S. 313 – 353

[66] Smetana, K. ; Ohlberger, M. : Hierarchical model reduction of nonlinear partial differential equations based on the adaptive empirical projection method and reduced basis techniques. In: *ArXiv e-prints* (2014)

[67] Spence, A. ; Graham, I. G.: Numerical Methods for Bifurcation Problems. In: *The Graduate Student's Guide to Numerical Analysis '98* Bd. 26. Springer Berlin Heidelberg, 1999, S. 177–216

[68] Syam, M. I. ; Siyyam, H. I.: Numerical differentiation of implicitly defined curves. In: *Journal of Computational and Applied Mathematics* 108 (1999), Nr. 1–2, S. 131 – 144

[69] Tonn, T. ; Urban, K. : A reduced-basis method for solving parameter-dependent convection-diffusion problems around rigid bodies. In: *ECCOMAS CFD 2006: Proceedings of the European Conference on Computational Fluid Dynamics, Egmond aan Zee, The Netherlands, September 5-8, 2006* Delft University of Technology; European Community on Computational Methods in Applied Sciences (ECCOMAS), 2006

[70] Toselli, A. ; Widlund, O. : *Springer Series in Computational Mathematics*. Bd. 34: *Domain Decomposition Methods - Algorithms and Theory*. Springer Berlin Heidelberg, 2005

[71] Tu, L. W.: *An Introduction to Manifolds*. Springer-Verlag New York, 2011

[72] Veroy, K. ; Prud'homme, C. ; Rovas, D. V. ; Patera, A. T.: A posteriori error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations. In: *Proceedings of the 16th AIAA computational fluid dynamics conference* 3847 (2003), S. 23–26

[73] Vladimir, T. : *Greedy Approximation*. Cambridge University Press, 2011

[74] Volkwein, S. : Model reduction using proper orthogonal decomposition. In: *Lecture Notes, Institute of Mathematics and Scientific Computing, University of Graz* (2011)

[75] Weng, S. ; Xia, Y. ; Xu, Y.-L. ; Zhu, H.-P. : Substructure based approach to finite element model updating. In: *Computers and Structures* 89 (2011), Nr. 9–10, S. 772 – 782. – ISSN 0045–7949

[76] Wirtz, D. ; Sorensen, D. C. ; Haasdonk, B. : A Posteriori Error Estimation for DEIM Reduced Nonlinear Dynamical Systems. In: *SIAM Journal on Scientific Computing* 36 (2014), Nr. 2, S. A311–A338

Lebenslauf

31. Juli 1984	Geboren in Magdeburg
September 1992 - August 1995	Grundschule Oskar-Linke, Magdeburg
September 1995 - März 2004	Gymnasium Otto-von-Guericke, Magdeburg
September 2004 - Juni 2005	Zivildienst am Altstädtischen Krankenhaus, Magdeburg
Oktober 2005 - November 2010	Studium der Mathematik an der Otto-von-Guericke-Universität, Magdeburg
November 2009 - November 2010	Praktikum bei der Robert Bosch GmbH, Stuttgart verbunden mit der Anfertigung der Diplomarbeit
November 2010 - August 2015	Wissenschaftlicher Mitarbeiter an der Technischen Universität Hamburg-Harburg
August 2015	Emigration nach Großbritannien
Dezember 2015 - heute	Actuarial Consultant bei Barnett Waddingham LLP