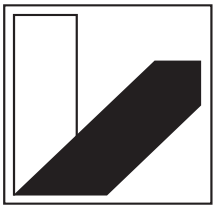

UNIVERSITÄT
BAYREUTH

Diskrete Approximation von hoher Ordnung für lineare Kontrollsysteme

Diplomarbeit

von

Hannes Buchholzer

FAKULTÄT FÜR MATHEMATIK UND PHYSIK

MATHEMATISCHES INSTITUT

Datum: 1. Oktober 2006

Aufgabenstellung / Betreuung:
Prof. Dr. F. Lempio
Dr. R. Baier

Erster Gutachter:
Prof. Dr. F. Lempio

Danksagung

An dieser Stelle möchte ich die Gelegenheit ergreifen mich bei einigen Personen zu bedanken.

Einen besonderen Dank möchte ich an Herrn Baier richten. Er hat mir die mengenwertige numerische Analysis vorgestellt und mein Interesse dafür geweckt. Außerdem hat er mich während der Erstellung dieser Arbeit außerordentlich engagiert und kompetent betreut. Die Zusammenarbeit mit ihm hat mit viel Freude gemacht und hat sich für mich als sehr fruchtbar und hilfreich erwiesen.

Desweiteren möchte ich mich bei Herrn Lempio bedanken. Er hat mich während meines Studiums öfter gut beraten (fast als einziger Professor). Ohne seine Hilfe hätte ich mein Studium vielleicht nicht beenden können. Er hat mich auch im Rahmen dieser Arbeit sehr kompetent betreut. Außerdem hat er durch seine Vorlesungen "Numerische Mathematik" und "Mathematische Methoden des Operations Research" die Grundlagen für diese Arbeit gelegt und mein Interesse an der numerischen Mathematik geweckt. Im Rahmen dieser Veranstaltungen hat er sich viel Zeit für die Betreuung der Studenten genommen.

Ein weiterer Dank geht an Professor Simader, der mir ein wenig Hilfestellung im Bezug auf Sobolev-Räume und deren Einbettungen gegeben hat.

Ein besonderer Dank geht auch an meinen lieben Vater, der mich während des gesamten Studiums unterstützt hat und mir so erst dass Studium ermöglicht hat.

Schließlich möchte ich mich noch ganz besonders bei meinem besten Freund bedanken, der mir in dieser Zeit fachlich wie privat sehr beistand.

Inhaltsverzeichnis

Inhaltsverzeichnis iv

Tabellenverzeichnis vii

Abbildungsverzeichnis ix

Einleitung 1

Liste der Symbole 7

1. Hilfsmittel 11
- 1.1. Mengenoperationen und konvexe Mengen . . . 11
- 1.2. Der Hausdorff-Abstand von Mengen . . . 14
- 1.3. Die Stützfunktion und Ihre Eigenschaften . . . 16
- 1.4. Die Stützpunktmenge . . . 18
- 1.5. Mengenwertige Abbildungen und das Aumann-Integral . . . 20
- 1.6. Matrixnormen . . . 23
- 1.7. Das Lebesgue-Integral und Sobolev-Räume . . . 25
- 1.8. Absolutstetige Funktionen . . . 29

2. Einschrittverfahren und Kontrollprobleme 35
- 2.1. Grundlegende Definitionen und Notationen . . . 35
- 2.2. Einschrittverfahren . . . 38
- 2.3. Kontrollprobleme . . . 44

3. Lineare Kontrollprobleme 51
- 3.1. Die Theorie von Ferretti . . . 51
  - 3.1.1. Entwicklung der diskreten und analytischen Lösung . . . 52
  - 3.1.2. Modifizierte RK-Verfahren und Konsistenz . . . 59
- 3.2. Ergänzungen und neue Verfahren . . . 66
- 3.3. Approximation der erreichbaren Menge . . . 76

4. Nichtlineare Kontrollprobleme 79
- 4.1. Theorie von Ferretti . . . 79
- 4.2. Nachteile dieses Zugangs . . . 88
  - 4.2.1. Steigerung der Konsistenzordnung . . . 88

4.2.2. Mengenwertige Verfahren 91

5. Algorithmische und mathematische Umsetzung 93
5.1. Mengenarithmetik mit Stützpunkten 94
5.2. Eigenschaften der Auswahlmenge 96
5.3. Berechnung durch das Hausdorff-Momentenproblem 101
5.3.1. Das Hausdorff-Momentenproblem 101
5.3.2. Stützpunktberechnung durch das Hausdorff-Momentenproblem 105
5.4. Rückführung auf das Aumann-Integral 112
5.4.1. Stützpunkte für einen Quadersteuerbereich 113
5.4.2. Polynomnullstellen 118
5.4.3. Stützpunkte für Kugelsteuerbereich 122

6. Anwendungen und Beispiele 129
6.1. Einführung und Grundlagen 129
6.1.1. Hard- und Software 129
6.1.2. Die implementierten Verfahren 130
6.1.3. Verschiedene Resultate 131
6.2. Beispiele 136

Zusammenfassung und Ausblick 157

A. Inhalt der CDROM und Installation der Software 161
A.1. Inhalt der CDROM 161
A.2. Bemerkungen über die Installation der Software 162

B. Benutzung des selbsterstellten Programmpaketes 165
B.1. Scilab-Schnittstelle 165
B.2. C++ Quellcode 166

Tabellenverzeichnis

6.1. Fehler der Klasse-1-Verfahren 2. Ordnung in Beispiel 6.2.1 137
6.2. Fehler der Klasse-1-Verfahren 3. und 4. Ordnung in Beispiel 6.2.1 . . . 137
6.3. Fehler der Klasse-2-Verfahren 2. Ordnung in Beispiel 6.2.1 138
6.4. Fehler der Klasse-2-Verfahren 3. und 4. Ordnung in Beispiel 6.2.1 . . . 138
6.5. Fehler der Klasse-3-Verfahren 2. Ordnung in Beispiel 6.2.1 139
6.6. Fehler der Klasse-3-Verfahren 3. und 4. Ordnung in Beispiel 6.2.1 . . . 139
6.7. Konvergenzord. der Ferretti-Verfahren der 1. Klasse in Beispiel 6.2.2 . 141
6.8. Konvergenzord. der Ferretti-Verfahren der 2. Klasse in Beispiel 6.2.2 . 142
6.9. Konvergenz der Ferretti-Verfahren aus der 1. Klasse in Beispiel 6.2.3 . 144
6.10. Fehler der Ferretti-Verfahren aus der 1. Klasse in Beispiel 6.2.3 145
6.11. Relativer Zeitvergleich für Beispiel 6.2.4 (1) 148
6.12. Relativer Zeitvergleich für Beispiel 6.2.4 (2) 149
6.13. Absoluter Zeitvergleich für Beispiel 6.2.4 149
6.14. Konvergenz der Ferretti-Verfahren aus der 1. Klasse in Beispiel 6.2.5 . 151
6.15. Konvergenz der Ferretti-Verfahren aus der 1. Klasse in Beispiel 6.2.6 . 154

Abbildungsverzeichnis

3.1. Diskrete Kontrollbereiche für zwei RK-Verfahren 64
5.1. Die Hausdorff-Menge $\mathcal{H}_2$ 104
5.2. Zwei Geraden mit Normalenvektoren durch den Rand von $\mathcal{H}_2$ 109
5.3. Die Menge $\mathcal{H}_2$ und die Halbebene $\mathcal{H}_{\mathfrak{n}}^z$ für $\mathfrak{n} = \mathfrak{n}(f'(z))$ 110
6.1. Die erreichbare Menge aus Beispiel 6.2.1 136
6.2. Die erreichbare Menge aus Beispiel 6.2.2 142
6.3. Approximationen der erreichbare Menge aus Beispiel 6.2.3 145
6.4. Approximationen der erreichbare Menge aus Beispiel 6.2.3 146
6.5. Die erreichbare Menge von Beispiel 6.2.4 147
6.6. Die erreichbare Menge aus Beispiel 6.2.5 152
6.7. Die erreichbare Menge aus Beispiel 6.2.5 (Fortsetzung) 153
6.8. Die erreichbare Menge aus Beispiel 6.2.6 155
6.9. Die erreichbare Menge aus Beispiel 6.2.6 (Fortsetzung) 156

Einleitung

In vorliegenden Arbeit wird eine Familie von numerischen Verfahren zur Approximation der erreichbaren Menge eines linearen Kontrollproblems entwickelt. Dabei ist das entsprechende Kontrollsystem linear im Zustand und auch in der Kontrolle, die hier nur als integrierbar vorausgesetzt wird mit Werten in einem kompakten Steuerbereich. Die erreichbare Menge beschreibt nun ausgehend von einem Startzustand (oder mehreren Startzuständen) zu einer Startzeit t_0 alle Zustände in die sich das System zur Zeit t unter allen zulässigen Kontrollen bewegen kann. Man erhält also einen Eindruck, welche Zustände möglich sind, unter den zulässigen Steuerungen, und welche Zustände auf keinen Fall erreicht werden können.

Mathematisch hat ein solch Kontrollproblem folgende Gestalt

$$x'(t) = Ax(t) + Bu(t) \quad , \quad x(t_0) = x_0\,,\, t \in I$$

wobei x_0 der Startzustand oder die Anfangsbedingung, u eine integrierbare Kontrollfunktion ist, welche ihre Werte im Kontrollbereich U haben muss und I ist ein reelles Intervall. Diese Begriffe kann man an folgendem trivialen Beispiel illustrieren. Ein Autofahrer möchte wissen, welche Entfernung er in einer bestimmten Zeit t_f von seinem Heimatort aus zurücklegen kann, wobei die Fahrgeschwindigkeit zwischen einer maximalen Geschwindigkeit v_{max} und einer minimalen Geschwindigkeit v_{min} liegen muss. Das zugehörige Kontrollproblem hat dann folgende Gestalt. Die Zustandsfunktion $t \mapsto x(t)$ beschreibt die Entfernung vom Heimatort zum Zeitpunkt t. Das Intervall I ist das Zeitintervall $[0, t_f]$. Die Steuerfunktion $t \mapsto u(t)$ liefert die Geschwindigkeit des Autos zum Zeitpunkt t. Und weiter ist $U = [v_{min}, v_{min}]$ der Steuerbereich, der die Bedingung an die Geschwindigkeit enthält. Außerdem ist $A = 0$ und $B = 1$(aber im allgemeinen sind A und B Matrizen). Die Anfangsbedingung dieses Problems ist $x_0 = 0$, denn Autofahrer startet in seinem Heimatort, also bei Entfernung 0. Die erreichbare Menge $\mathcal{R}(t_f, 0)$ beschreibt nun das Intervall der Entfernungen, welche der Autofahrer unter den zulässigen Bedingungen erreichen kann. Mit Kontrollsystemen, insbesondere mit nichtlinearen, kann man natürlich auch noch viel kompliziertere Probleme modellieren.

Gibt man nun in einem Kontrollproblem eine beliebige Steuerung vor, so erhält man nichts anderes als eine gewöhnliche Differentialgleichung. Zur Lösung solcher Anfangswertprobleme stehen jedoch zahlreiche effektive numerische Verfahren zur Verfügung. Doch sie alle haben hier Probleme, da die Kontrolle nicht einmal stetig ist.

Dies ist der Ausgangspunkt dieser Arbeit. Alle hier vorgestellten Verfahren beruhen

auf dem gleichen Konzept. Es wird eine Kontrolle vorgegeben und die entsprechende Trajektorie in eine Summe entwickelt, die Mehrfachintegrale enthält. Diese Entwicklung der Lösung beruht auf ganz einfachen Mitteln. Anschließend werden aufgrund dieser Entwicklung die Kontrollvektoren eines Runge-Kutta-Verfahrens entsprechend modifiziert, sodass die Konvergenzordnung dieser Verfahren wie im glatten Fall erhalten bleibt. Man kann diese Methode auch als ein Art Auswahlstrategie interpretieren, denn die Kontrollvektoren dürfen hierbei nicht beliebig aus dem Kontrollbereich U gewählt werden, sondern müssen zusammen in einer bestimmten Teilmenge von $\mathcal{U}_{m,p} \subset U^p$ enthalten sein, wobei p die Anzahl der Kontrollvektoren ist. Auf diesem punktwertigen Ansatz aufbauend kann man dann ganz leicht die erreichbare Menge von gleicher Ordnung approximieren.

Diese Idee sowie wesentlich Teile dieser Arbeit gehen auf einen Artikel Ferrettis [13] zurück. Im Grunde ist diese Arbeit eigentlich ein Ausarbeitung dieses Artikels. Dabei stand von Anfang an die numerische Realisierung dieser mengenwertiger Verfahren im Vordergrund. Durch die praktische Realisierung dieser Verfahren und den damit verbundenen numerischen Tests, haben sich auch einige fruchtbare Erkenntnisse ergeben, die über den Artikel Ferrettis hinausgehen bzw. Licht auf die vorhandenen Ergebnisse werfen. So konnte über diesen Artikel hinaus folgendes gezeigt werden:

- Ferretti behandelt als Steuerbereich nur das Einheitsintervall, d.h. $U = [0, 1]$; dies konnte erweitert werden auf beliebige achsenparallele Quader und Kugeln; der entwickelte Ansatz kann leicht für andere geometrische Objekte angepasst werden.

- Bei Ferretti ergibt sich eine implizite Ordnungsbeschränkung von Ordnung 4; diese Beschränkung konnte vollständig beseitigt werden. Beliebige Konvergenzordnungen sind nun möglich.

- Es gibt für jede Konvergenzordnung im Grunde nur ein Verfahren, d.h. zwei Runge-Kutta Verfahren der gleichen Ordnung liefern theoretisch wie praktisch immer die gleiche Approximation. Dieses Grundverfahren ist zwar ebenfalls ein Einschrittverfahren aber kein explizites Runge-Kutta Verfahren mehr. Der Zugang über Runge-Kutta-Verfahren, den Ferretti gewählt hat, ist also unnötig.

Außerdem waren zur praktischen Durchführung dieser mengenwertiger Verfahren eine Reihe von Überlegungen notwendig, die Ferretti in seinem Artikel nicht angesprochen hat.

Eine weitere wichtige Frage war von Anfang an, in wie weit sich die Ergebnisse für lineare Kontrollprobleme auf den nichtlinearen Fall übertragen lassen. Ferretti hat dazu in seinem Artikel gezeigt, dass sich mit dieser Methode Trajektorien von Ordnung 2 approximieren lassen. Meine Untersuchungen haben ergeben, dass sich diese Ordnung weder steigern lässt, noch lässt sich dieses Ergebnis nutzen, um die erreichbare Menge des nichtlinearen Systems durch ein effektives Verfahren zu approximieren. In diesem schwierigen Bereich konnte also kein Fortschritt entwickelt werden.

Man kann dem obigen linearen Kontrollsystem folgende Differentialinklusion zuordnen

$$x'(t) \in Ax(t) + BU \quad , \quad x(t_0) = x_0.$$

Es ist bekannt, dass die erreichbare Menge beider Probleme identisch sind. Differentialinklusion kommen beispielsweise vor, wenn bei einem optimalen Steuerungsproblem die Steuerung (kleinen) Ungenauigkeiten unterworfen ist. Dann ersetzt man jeden Wert der Steuerung durch eine kleine Kugel, die diesen Wert als Mittelpunkt hat.

Robert Baier hat in [4] mengenwertige Quadraturverfahren vorgestellt mit denen man die erreichbare Menge von Differentialinklusionen approximieren kann. Dabei hängt die Konvergenzordnung der Verfahren von der Glattheit der Stützfunktion ab. Es hat sich gezeigt, dass gerade bei Quadern als Steuerbereich die Glattheit der Stützfunktion oft nur Konvergenzordnung 2 zulässt. Die Verfahren, die in dieser Arbeit vorgestellt werden, können auch bei solchen Fällen beliebig hohe Konvergenzordnungen erreichen.

Neben [13] als Hauptliteratur, wurde noch zahlreiche Literatur verwendet, da die behandelte Thematik zahlreiche Bereiche berührt. Es werden Ergebnisse aus der konvexen Analysis, der Theorie gewöhnlicher Differentialgleichungen unter den schwachen Carathéodory Voraussetzungen, über Einschrittverfahren, über mengenwertige Funktionen, über das Aumann-Integral, der Theorie absolutstetiger Funktionen und aus weiteren Bereichen der Mathematik benötigt. Von der benutzten Literatur haben besonders die Arbeit von R. Baier [4] und das sehr gute Skript von R. Baier und M. Gerdts [16] einen starken und fruchtbaren Einfluss auf die Gestaltung dieser Arbeit gehabt. Manche Abschnitte sind stark an diesen Büchern angelehnt.

Damit diese Arbeit trotz der Vielfalt der benötigten Resultate möglichst in sich abgeschlossen ist, beinhaltet Kapitel 1 eine Zusammenstellung der meisten benötigten Hilfsmittel. Um den Umfang dieser Arbeit nicht unnötig aufzublähen, wurden dabei die Beweise soweit wie möglich zitiert. In mehren Abschnitten werden jeweils mathematische Objekte eingeführt und dann einige wichtige Eigenschaften dieser Objekte notiert. Dabei geht es um Mengenoperationen, konvexe Mengen, die Stützfunktion, die Stützpunktmenge, mengenwertige Abbildungen, das Aumann-Integral, Sobolev-Räume und deren Einbettungen, Matrixnormen und absolutstetige Funktionen.

In Kapitel 2 werden die theoretischen Grundlagen für die Arbeit gelegt. Im ersten Abschnitt werden Notationen festgelegt und einige weitere kleinere Definitionen getätigt. Außerdem werden die gewöhnlichen Differentialgleichungen eingeführt. Im nächsten Abschnitt werden dann Einschrittverfahren zur ihrer Lösung vorgestellt. Es wird die Verfahrensfunktion eines Einschrittverfahrens definiert, desweiteren werden fundamentale Begriffe wie Konsistenz, Stabilität und Konvergenz vorgestellt. Außerdem werden einige wichtige Resultate in diesem Zusammenhang gebracht. Der dritte und letzte Abschnitt behandelt schließlich die Kontrollprobleme. Es wird das lineare und nichtlineare Kontrollproblem eingeführt. Und es werden Eindeutigkeits- und Existenzaussagen für den linearen und nichtlinearen Fall getätigt. Außerdem wird

die erreichbare Menge definiert und in Verbindung mit der zugeordneten linearen Differentialinklusion gebracht.

Als nächstes werden in Kapitel 3 die linearen Kontrollprobleme ausführlich behandelt. Der erste Abschnitt ist im wesentlichen eine Ausarbeitung des Artikels von Ferretti für den linearen Fall. Dabei wird diese Theorie ziemlich ausführlich und ein wenig anders dargestellt als Ferretti dies getan hat. Das Hauptresultat ist ein Konsistenzbeweis für die modifizierten Runge-Kutta-Verfahren, die in diesem Abschnitt eingeführt werden. Im nächsten Abschnitt werden Ergänzungen zu dem was Ferretti in seinem Artikel gebracht hat vorgestellt. Es wird ein neues Einschritt-Verfahren, das ich Ferretti-Verfahren genannt habe, zur Lösung dieses linearen Kontrollproblems definiert. Für dieses Verfahren wird Konsistenz gezeigt und für beide Verfahrensarten wird Konvergenz gezeigt. Außerdem wird bewiesen, dass alle modifizierten Runge-Kutta Verfahren mit diesem Verfahren identisch sind. Im dritten Abschnitt werden alle diese Verfahren zu mengenwertigen Verfahren erweitert um damit die erreichbare Menge zu approximieren. Dann wird die Konvergenz der Approximation gegen die erreichbare Menge bewiesen.

In Kapitel 4 wird das nichtlineare Kontrollproblem behandelt. Zunächst wird im ersten Abschnitt die Theorie von Ferretti aus [13] dargestellt. Das bedeutet es werden die Methoden des linearen Falles auf den nichtlinearen Fall übertragen. Damit kann man für die modifizierten Runge-Kutta-Verfahren, welche hier genauso definiert werden wie im linearen Fall, Konsistenzordnung 2 zeigen. Im nächsten Abschnitt wird auf Nachteile dieses Zugangs eingegangen. Es wird gezeigt, warum eine höhere Konsistenzordnung im Rahmen dieses Zugangs kaum möglich ist. Und es wird dargelegt, dass man diese punktwertigen Verfahren hier nicht zu effektiven mengenwertigen Verfahren ausbauen kann. Die so gewonnenen mengenwertigen Verfahren haben alle exponentiellen Aufwand.

Das Kapitel 5 geht auf die numerische Umsetzung der mengenwertigen Verfahren zur Berechnung der erreichbaren Menge eines linearen Kontrollproblems ein. Im ersten Abschnitt wird gezeigt, wie man mit Stützpunkten aus den mengenwertigen Verfahren punktwertige Verfahren machen kann, die dem Rechner zugänglich sind. In den folgenden Abschnitten geht es vorallem darum Stützpunkte an die Auswahlmenge möglichst effektiv zu berechnen. Dazu werden verschiedene Ansätze diskutiert und Algorithmen erarbeitet.

Schließlich wird in Kapitel 6 das numerische Verhalten der vorgestellten mengenwertigen Verfahren veranschaulicht. Der erste Abschnitt enthält einige Informationen über das selbsterstellte Softwarepaket zur Berechnung der erreichbaren Menge eines linearen Kontrollproblems. Außerdem wird einige theoretische Vorarbeit für die Beispiele getan. Es wird gezeigt, wie man anhand numerischer Resultate die Konvergenzordnung eines Verfahrens schätzen kann, wie man Punkte aus der Einheitssphäre systematisch wählen kann und wie groß der Fehler ist, wenn man eine Menge durch die konvexe Hülle von endlich vielen ihrer Stützpunkte approximiert. Im nächsten und letzten Abschnitt wird in mehreren Beispielen das numerische Verhalten der im-

plementierten Verfahren demonstriert. Dabei sind die Beispiele so ausgewählt, dass jeweils ein oder zwei besondere Aspekte der Verfahren exemplarisch belegt werden. Außerdem werden einige Besonderheiten der Verfahren aufgezeigt.

Liste der Symbole

$[a,b], (a,b)$	abgeschlossenes bzw. offenes reelles Intervall
$[a,b), (a,b]$	rechts bzw. links halboffenes reelles Intervall

$\mathrm{AC}(I)^k$	Menge aller vektorwertigen absolutstetigen Funktionen $\mathfrak{f}: I \longrightarrow \mathbb{R}^k$ auf dem Intervall $I \subset \mathbb{R}$
$\mathrm{AC}(I)$	Menge aller absolutstetigen Funktionen $f: I \longrightarrow \mathbb{R}$ auf dem Intervall $I \subset \mathbb{R}$
$L^p(\Omega)^k$	Menge aller messbarer Abbildungen $\mathfrak{f}: \Omega \longrightarrow \mathbb{R}^k$ auf der Menge $\Omega \subset \mathbb{R}^l$ mit $\int_\Omega \|\mathfrak{f}\|_p^p \, d\lambda < \infty$. Dabei ist $L^p(\Omega) := L^p(\Omega)^1$
$L^p(\Omega; U)$	wie vorher, doch gilt hier zusätzlich $\mathfrak{f}(\Omega) \subset U$ mit $U \subset \mathbb{R}^k$
$W^{m,p}(\Omega)^k$	Sobolev-Raum aller Abbildungeng $\mathfrak{f} \in L^p(\Omega)^k$, sodass für den Multiindex $\|\alpha\| \leq m$ auch die schwachen Ableitungen $D^{(\alpha)}\mathfrak{f} \in L^p(\Omega)$ sind. Dabei ist $W^{m,p}(\Omega) := W^{m,p}(\Omega)^1$; siehe Abschnitt 1.7
$W^{m,p}(\Omega; U)$	wie vorher, doch erfüllt hier ein Element $\mathfrak{f}$ zusätzlich $\mathfrak{f}(\Omega) \subset U$ mit $U \subset \mathbb{R}^k$
$\mathcal{C}^p(\Omega)^k$	Raum der p-mal stetig differenzierbaren Abbildungen $\mathfrak{f}: \Omega \longrightarrow \mathbb{R}^k$. Dabei ist $\mathcal{C}^p(\Omega) := \mathcal{C}^p(\Omega)^1$
$\mathcal{C}^p(\Omega; U)$	wie vorher, doch erfüllt hier ein Element $\mathfrak{f}$ zusätzlich $\mathfrak{f}(\Omega) \subset U$ mit $U \subset \mathbb{R}^k$
$\mathcal{C}_B^p(\Omega)^k$	wie $\mathcal{C}^p(\Omega)^k$ nur ist hier jede Abbildung einschließlich aller ihrer partieller Ableitungen bis zur p-ten Ordnung beschränkt
$\mathcal{C}_B^p(\Omega; U)$	wie vorher, doch erfüllt hier ein Element $\mathfrak{f}$ zusätzlich $\mathfrak{f}(\Omega) \subset U$ mit $U \subset \mathbb{R}^k$
$\mathcal{C}^\infty(\Omega)^k$	Raum der beliebig oft stetig differenzierbaren Abbildungen $\mathfrak{f}: \Omega \longrightarrow \mathbb{R}^k$. Dabei ist $\mathcal{C}^p(\Omega) := \mathcal{C}^p(\Omega)^1$

$\langle \mathfrak{v}, \mathfrak{w} \rangle$	euklidisches Skalarprodukt der Vektoren $\mathfrak{v}, \mathfrak{w} \in \mathbb{R}^k$
$\mathfrak{v}^T, \mathfrak{M}^T$	Transponierte des Vektors $\mathfrak{v} \in \mathbb{R}^k$ bzw. der Matrix $\mathfrak{M} \in \mathbb{R}^{k \times l}$
$\|\mathfrak{v}\|_p$	p-Norm eines Vektors $\mathfrak{v} \in \mathbb{R}^k$ mit $p \in [1, \infty]$; siehe Seite 24
$\|\mathfrak{M}\|_Z$	Zeilensummennorm einer Matrix $\mathfrak{M} \in \mathbb{R}^{k \times l}$; siehe Seite 25
$\|\mathfrak{M}\|_{Sp}$	Spektralnorm einer Matrix $\mathfrak{M} \in \mathbb{R}^{k \times l}$; siehe Seite 25
$\|U\|_p$	p-Norm einer Menge $U \subset \mathbb{R}^k$ mit $p \in [1, \infty]$; siehe Seite 36
$\|\mathfrak{f}\|_{L^p(\Omega)^k}$	L^p-Norm einer Funktion $\mathfrak{f} \in L^p(\Omega)^k$; siehe Abschnitt 1.7
$\|\mathfrak{f}\|_{W^{m,p}(\Omega)^k}$	Sobolev-Norm einer Funktion $\mathfrak{f} \in W^{m,p}(\Omega)^k$; siehe Abschnitt 1.7
$\|\mathfrak{f}\|_{W^{m,p}}$	wie vorher; Kurzform
$\|\mathfrak{f}\|_U$	das Supremum einer Funktion $\mathfrak{f} : \Omega \longrightarrow \mathbb{R}^k$, mit $U \subset \Omega \subset \mathbb{R}^l$ bzgl. der Maximumsnorm. D.h. $\sup_{\mathfrak{u} \in U} \|\mathfrak{f}(\mathfrak{u})\|_\infty$
$B_r(\mathfrak{m})$	abgeschlossene euklidische Kugel im $\mathbb{R}^k$ mit Radius $r \geq 0$ und Mittelpunkt $\mathfrak{m} \in \mathbb{R}^k$
S_{k-1}	euklidische Einheitssphäre im $\mathbb{R}^k$; identisch mit dem Rand der euklidischen Einheitskugel $\partial B_1(0)$
$\emptyset$	leere Menge
$\mathbb{N}$	Menge der natürlichen Zahlen
$\mathbb{R}, \mathbb{R}^n$	Körper der reellen Zahlen bzw. Vektorraum der reellen Spaltenvektoren
$\text{int}(U)$	das Innere einer Menge $U \subset \mathbb{R}^k$
$\overline{U}$	der Abschluss einer Menge $U \subset \mathbb{R}^k$
∂U	der Rand einer Menge $U \subset \mathbb{R}^k$
$\text{co}(U)$	die konvexe Hülle einer Menge $U \subset \mathbb{R}^k$; siehe Seite 13
$\overline{\text{co}}(U)$	die abgeschlossene konvexe Hülle einer Menge $U \subset \mathbb{R}^k$; siehe Seite 13

$\delta^*(\mathfrak{l}, U)$	die Stützfunktion der Menge $U \subset \mathbb{R}^k$ in Richtung $\mathfrak{l} \in \mathbb{R}^k$; siehe Seite 16
$Y(\mathfrak{l}, U)$	die Stützpunktmenge der Menge $U \subset \mathbb{R}^k$ in Richtung $\mathfrak{l} \in \mathbb{R}^k$; siehe Seite 18
$y(\mathfrak{l}, U)$	ein Stützpunkt der Menge $U \subset \mathbb{R}^k$ in Richtung $\mathfrak{l} \in \mathbb{R}^k$; siehe Seite 18
$\sup_{\mathfrak{x} \in \Omega} f(\mathfrak{x})$	das Supremum der Funktion $f : \mathbb{R}^k \longrightarrow \mathbb{R}$ auf der Menge $\Omega \subset \mathbb{R}^k$
$\max_{\mathfrak{x} \in \Omega} f(\mathfrak{x})$	das Maximum der Funktion $f : \mathbb{R}^k \longrightarrow \mathbb{R}$ auf der Menge $\Omega \subset \mathbb{R}^k$
$\inf_{\mathfrak{x} \in \Omega} f(\mathfrak{x})$	das Infimum der Funktion $f : \mathbb{R}^k \longrightarrow \mathbb{R}$ auf der Menge $\Omega \subset \mathbb{R}^k$
$\min_{\mathfrak{x} \in \Omega} f(\mathfrak{x})$	das Minimum der Funktion $f : \mathbb{R}^k \longrightarrow \mathbb{R}$ auf der Menge $\Omega \subset \mathbb{R}^k$
$\mathrm{dist}(\mathfrak{x}, U)$	die Distanz des Punktes $\mathfrak{x} \in \mathbb{R}^k$ zur Menge $U \subset \mathbb{R}^k$; siehe Seite 14
$\mathrm{d}(U, V)$	der einseitige Hausdorff-Abstand der Menge $U \subset \mathbb{R}^k$ von der Menge $V \subset \mathbb{R}^k$; siehe Seite 14
$\mathrm{d_H}(U, V)$	der Hausdorff-Abstand der Mengen $U \subset \mathbb{R}^k$ und $V \subset \mathbb{R}^k$; siehe Seite 14

$\mathfrak{f}', f'$	die Ableitung einer Funktion $\mathfrak{f} : \mathbb{R} \longrightarrow \mathbb{R}^k$ bzw. $f : \mathbb{R} \longrightarrow \mathbb{R}$
$\mathfrak{f}\,\vert_D$	die Einschränkung der Funktion $\mathfrak{f} : V \longrightarrow \mathbb{R}^k$ auf die Menge $D \subset V$
$\mathfrak{J}_\mathfrak{a}$	die Jacobi-Matrix einer Abbildung $\mathfrak{a} : \mathbb{R}^l \longrightarrow \mathbb{R}^k$
$\int_a^b \mathfrak{f}(t)\,dt$	das Lebesgue-Integral einer vektorwertigen Funktion $\mathfrak{f} : \mathbb{R} \longrightarrow \mathbb{R}^k$ (wird komponentenweise gebildet) über das Intervall $[a, b]$
$\int_a^b f(t)\,dt$	das Lebesgue-Integral einer Funktion $f : \mathbb{R} \longrightarrow \mathbb{R}$ über das Intervall $[a, b]$
$\lceil x \rceil$	aufrunden auf die kleinste natürliche Zahl n mit $x \leq n$.
$\forall' t \in \Omega$	dieser Quantor bezeichnet fast alle $t \in \Omega$, d.h. alle $t \in \Omega - N$, wobei N eine Nullmenge ist.

$\mathfrak{E}_m$	die $m \times m$-Einheitsmatrix
$\mathrm{span}(\mathfrak{v}_1, \ldots, \mathfrak{v}_k)$	die lineare Hülle der Vektoren $\mathfrak{v}_1, \ldots, \mathfrak{v}_k \in \mathbb{R}^l$

$\mathcal{R}(t, x_0)$	die erreichbare Menge eines vorgegebenen linearen Kontrollproblems zu Zeitpunkt t ausgehend vom Anfangswert x_0 zur Zeit $t = 0$
$\mathcal{R}_h(t, x_0)$	eine Approximation an die obige erreichbare Menge durch ein vorgegebenes mengenwertiges Verfahren mit Schrittweite h berechnet
$\Gamma_{m,p}$	die definierende Matrix eines modifizierten RK-Verfahrens; siehe Seite 59 bzw. 60
$\Delta_{m,p}$	die Diagonalmatrix mit Fakultäten als Einträgen; siehe siehe Seite 59 bzw. 60
$\mathcal{I}_{m,p}$	die Auswahlmenge; siehe Seite 62
$\mathcal{M}_{m,p}$	die Momentenmenge; siehe Seite 98
$\mathcal{U}_{m,p}$	der diskrete Kontrollbereich, definiert als $\Gamma_{m,p}^{-1}\Delta_{m,p}\mathcal{I}_{m,p}$; siehe Seite 63
$\Phi_h(t, x)$	die Verfahrensfunktion eines Einschrittverfahrens
$\hat{x}$	analytische (bzw. exakte) Lösung eines Anfangswert- bzw. Kontrollproblems
x_h	diskrete Approximation an $\hat{x}$, konstruiert mit einem Einschrittverfahren

Kapitel 1.

Hilfsmittel

In diesem Kapitel werden diverse Hilfsmittel zur Verfügung gestellt, die in den späteren Kapiteln gebraucht werden. Um das Verständnis zu erleichtern wird dabei mehr dargestellt als nötig wäre.

In Abschnitt 1.1 - 1.4 werden einige Definitionen und Resultate über Mengen zusammengetragen. Es werden Mengenoperationen betrachtet, konvexe Mengen eingeführt und Ergebnisse aus der konvexen Analysis zur Charakterisierung konvexer Mengen zusammengestellt.

Anschließend werden mengenwertige Abbildungen und das Integral über solche Abbildungen, nämlich das Aumann-Integral, eingeführt.

In Abschnitt 1.6 werden einige Matrixnormen vorgestellt, die wir später benötigen.

Im nächsten Abschnitt werden das Lebesgue-Integral und Sobolev-Räume für vektorwertige Abbildungen eingeführt. Außerdem stellen wir hier einige wichtige Resultate über Funktionen aus einem speziellen Sobolev-Raum zur Verfügung.

Im letzten Abschnitt führen wir absolutstetige Funktionen ein, und stellen einige wichtige Ergebnisse zusammen. Denn diese Klasse von Funktionen ist in dieser Arbeit besonders wichtig, da potentielle Lösungen für schwache Differentialgleichungen hier zu suchen sind.

1.1. Mengenoperationen und konvexe Mengen

Zuerst definieren wir grundlegende Mengenoperationen und stellen einige ihrer Eigenschaften zusammen. Danach führen wir konvexe Mengen ein und zitieren noch einige Ergebnisse in diesem Zusammenhang.

Definition 1.1.1. Seien $A, B \subset \mathbb{R}^k$ und $c \in \mathbb{R}$. Die *Summe zweier Mengen* wird dann definiert als

$$A + B := \{\mathfrak{a} + \mathfrak{b} \mid \mathfrak{a} \in A,\ \mathfrak{b} \in B\}$$

und die *Multiplikation einer Menge mit einem Skalar* wird definiert als

$$c \cdot A := \{c \cdot \mathfrak{a} \mid \mathfrak{a} \in A\} \ .$$

Desweiteren ist die *Mengensubtraktion* definiert als

$$A - B := A + ((-1)B) \, .$$

Für einen Vektor $\mathfrak{v} \in \mathbb{R}^k$ definiert man

$$\mathfrak{v} + A := \{\mathfrak{v}\} + A$$

die Addition mit einer Menge.

Im Folgenden wird mit $B_1(0) \subset \mathbb{R}^k$ die abgeschlossene euklidische Einheitskugel bezeichnet und $\partial B_1(0) = S_{k-1}$ bezeichnet ihren Rand.
Diese beiden eingeführten Mengenoperationen erfüllen folgende Rechenregeln:

Proposition 1.1.2 (Rechenregeln)**.** *Seien $A, B, C \subset \mathbb{R}^k$ und $d, e \in \mathbb{R}$. Dann gilt:*

(i) $A + B = B + A$ *(Kommutativität bzgl. der Addition)*

(ii) $(A + B) + C = A + (B + C)$ *(Assoziativität bzgl. der Addition)*

(iii) $A + \{0_{\mathbb{R}^k}\} = A$ *($\{0_{\mathbb{R}^k}\}$ ist neutrales Element bzgl. der Addition)*

(iv) $d \cdot (eA) = (de)A$ *(Assoziativität bzgl. der skalaren Multiplikation)*

(v) $d(A + B) = dA + dB$ *(1. Distributivgesetz)*

(vi) $1 \cdot A = A$ *(1 ist neutrales Element bzgl. der skalaren Multiplikation)*

(vii) $(d + e)A \subset dA + eA$ *(2. Distributivgesetz als Inklusion)*

Beweis. zu (i)-(iii) siehe [16, Proposition 3.81].
zu (iv)-(vii) siehe [16, Proposition 3.88 bzw. 3.91]. □

Leider ist Potenzmenge von $\mathbb{R}^k$ mit der Mengenaddition keine Gruppe sondern nur eine Halbgruppe, da es nicht zu jeder Menge ein Inverses gibt. Dies soll folgendes Beispiel illustrieren:

Beispiel 1.1.3. Betrachte die abgeschlossenen Kugel $B_1(0) \subset \mathbb{R}^k$. Diese ist symmetrisch zum Ursprung. Und deswegen gilt: $-B_1(0) = B_1(0)$. Und damit folgt:

$$B_1(0) - B_1(0) = B_1(0) + (-B_1(0)) = B_1(0) + B_1(0) = B_2(0) \supsetneq \{0_{\mathbb{R}^k}\} \, .$$

Damit zeigt sich, dass die natürliche mengenwertige Erweiterung der punktwertigen Subtraktion im allgemeinen kein Inverses liefert. Bleibt noch zu zeigen, dass überhaupt kein Inverses zu $B_1(0)$ existiert.
Angenommen $A \subset \mathbb{R}^k$ sei ein Inverses zu $B_1(0)$. Dann gilt:

$$B_1(0) + A = \{\mathfrak{b} + \mathfrak{a} \mid \mathfrak{b} \in B_1(0), \mathfrak{a} \in A\} = \{0_{\mathbb{R}^k}\} \, .$$

Damit folgt: Zu $\mathfrak{b} \in B_1(0)$ muss $-\mathfrak{b} \in A$ sein. Damit folgt sofort $A \supset B_1(0)$ wegen der Punktsymmetrie von $B_1(0)$ zum Ursprung. Damit folgt trivialerweise: $B_1(0) + A \supset B_1(0) + B_1(0) = B_2(0)$ (siehe oben). Es existiert also kein Inverses zu $B_1(0)$.

Daraus sieht man, dass die Subtraktion Schwierigkeiten macht. Dies wirkt sich auch auf mengenwertige, numerische Verfahren aus. Außerdem ist damit auch klar, dass die Potenzmenge von $\mathbb{R}^k$ auch kein reeller Vektorraum sein kann, da sie nicht einmal ein Gruppe bzgl. der Addition ist.
Im folgenden definieren wir noch die lineare Transformation einer Menge.

Definition 1.1.4 (Multiplikation einer Menge mit einer Matrix)**.** Sei $\mathfrak{M} \in \mathbb{R}^{l\times k}$ eine Matrix und $A \subset \mathbb{R}^k$ eine beliebige Menge. Dann definiert man die *Multiplikation von der Matrix $\mathfrak{M}$ mit der Menge A* als:

$$\mathfrak{M} \cdot A := \{\mathfrak{M} \cdot \mathfrak{a} \mid \mathfrak{a} \in A\}$$

d.h. als das Bild von A unter der linearen Abbildung $\mathfrak{x} \mapsto \mathfrak{M}\mathfrak{x}$.

Im weiteren stellen wir noch einige Definitionen und Ergebnisse bzgl. konvexer Mengen bereit.

Definition 1.1.5. Eine Menge $K \subset \mathbb{R}^k$ heißt *konvex*, wenn für zwei beliebige Punkte $\mathfrak{k}_1, \mathfrak{k}_2 \in K$ auch deren Verbindungsstrecke in K liegt, d.h.

$$\lambda\mathfrak{k}_1 + (1-t)\mathfrak{k}_2 \in K \quad \forall t \in [0,1] \ .$$

Wenn darüber hinaus noch gilt: $\mathrm{int}(K) \neq \emptyset$ und für $\mathfrak{k}_1, \mathfrak{k}_2 \in \partial K$ liegt das relative Innere der Verbindungsstrecke im Inneren von K, d.h.

$$t\mathfrak{k}_1 + (1-t)\mathfrak{k}_2 \in \mathrm{int}(K) \quad \forall t \in (0,1)\,,$$

dann heißt K sogar *strikt konvex*.
Die *konvexe Hülle* einer Menge $A \subset \mathbb{R}^k$ wird mit $\mathrm{co}(A)$ bezeichnet und ist definiert als

$$\mathrm{co}(A) := \bigcap_{A\subset K \text{ konvex}} K \ .$$

Die *abgeschlossene konvexe Hülle* einer Menge $A \subset \mathbb{R}^k$ wird mit $\overline{\mathrm{co}}(A)$ bezeichnet und ist definiert als

$$\overline{\mathrm{co}}(A) := \bigcap_{A\subset K \text{ konvex und abgeschlossen}} K \ .$$

Bemerkung 1.1.6. Hier noch einige Ergebnisse zu konvexen Mengen:

(i) Ist $K \subset \mathbb{R}^k$ konvex so sind $\mathrm{int}(K)$, das Innere von K, und $\overline{K}$, der Abschluss von K, ebenfalls konvex.

(ii) Die Vereinigung von konvexen Mengen ist im allgemeinen nicht konvex. Der beliebige, auch abzählbare, Durchschnitt von konvexen Mengen ist wieder konvex.

Proposition 1.1.7. *Seine $K, K_1, K_2 \subset \mathbb{R}^k$ konvexe Mengen. Dann gilt:*

(i) $K_1 + K_2$ ist konvex.

(ii) Für $t \in \mathbb{R}$ ist tK konvex.

(iii) Für $\mathfrak{M} \in \mathbb{R}^{l\times k}$ ist $\mathfrak{M}K$ konvex.

Beweis. zu (i) siehe [16, Proposition 3.86(iv)].
zu (iii) siehe [17, Proposition 3.102(iv)].
(ii) ist ein Spezialfall von (iii) mit $\mathfrak{M} = t \cdot \mathfrak{E}_k$ und $l := k$. □

1.2. Der Hausdorff-Abstand von Mengen

Ein weiteres wichtiges Hilfsmittel ist der Hausdorff-Abstand von zwei Mengen, der es ermöglicht vom Abstand bzw. von der Entfernung zweier Mengen zu sprechen.

Definition 1.2.1. Seien $A, B \subset \mathbb{R}^k$ nichtleere Mengen und $\mathfrak{x} \in \mathbb{R}^k$. Dann ist der *Abstand des Punktes $\mathfrak{x}$ zur Menge A* definiert durch

$$\operatorname{dist}(\mathfrak{x}, A) := \inf_{\mathfrak{a}\in A} \|\mathfrak{x} - \mathfrak{a}\|_2 \ .$$

Dabei wird $\operatorname{dist}(\cdot, A)$ auch die *Distanzfunktion der Menge A* genannt.
Der *einseitige Hausdorff-Abstand* der Mengen A und B ist dann definiert als

$$\mathrm{d}(A, B) := \sup_{\mathfrak{a}\in A} \operatorname{dist}(\mathfrak{a}, B) \in [0, \infty] \, .$$

Und schließlich wird der *Hausdorff-Abstand* der Mengen A und B so definiert:

$$\mathrm{d_H}(A, B) = \max \{\mathrm{d}(A, B), \mathrm{d}(B, A)\} \in [0, \infty] \, .$$

Um diese recht abstrakte Definition besser zu verstehen, erfolgt hier noch eine andere Charakterisierung dieser Abstandsbegriffe.

Lemma 1.2.2. *Sein $A, B \subset \mathbb{R}^k$ nichtleere Mengen. Dann folgt:*

(i) $\mathrm{d}(A, B) = \inf \{\epsilon > 0 \mid A \subset B + \epsilon B_1(0)\}$

(ii) $\mathrm{d_H}(A, B) = \inf \{\epsilon > 0 \mid A \subset B + \epsilon B_1(0) \ \wedge \ B \subset A + \epsilon B_1(0)\}$

Beweis. Zum Beweis siehe [16, Lemma 135]. □

Wenn wir also für zwei nichtleere Mengen $A, B \subset \mathbb{R}^k$ den einseitigen Hausdorff-Abstand $\mathrm{d}(A,B) \leq r$ haben, bedeutet dies, das die Menge A innerhalb einer r-Umgebung von B liegt. Die folgende Proposition stellt einige Eigenschaften dieser beider Abstandsbegriffe zusammen.

Satz 1.2.3. *Sein $A, B, C \subset \mathbb{R}^k$ nichtleere und abgeschlossene Mengen. Weiter sei $r \in \mathbb{R}$. Dann gelten folgende Resultate:*

(i) $\mathrm{d}(A,B) = 0 \iff A \subset B$

(ii) $\mathrm{d_H}(A,B) = 0 \iff A = B$

(iii) $\mathrm{d}(A,B) \leq r \iff A \subset B + rB_1(0)$

(iv) $\mathrm{d_H}(A,C) \leq \mathrm{d_H}(A,B) + \mathrm{d_H}(B,C)$

Beweis. (i) $\mathrm{d}(A,B) = 0 \implies \mathrm{dist}(\mathfrak{a},B) = 0 \quad \forall \mathfrak{a} \in A$. Aber nach [16, Lemma 3.142(i)] folgt aus $\mathrm{dist}(\mathfrak{a},B) = 0$, dass $\mathfrak{a} \in B$ ist. Also ist insgesamt $A \subset B$.

Ist umgekehrt $A \subset B$, so ist $\mathrm{dist}(\mathfrak{a},B) = 0 \quad \forall \mathfrak{a} \in A$ und damit $\mathrm{d}(A,B) = 0$.

(ii) Nach der Definition des Hausdorff-Abstandes wendet man zweimal (i) an.

(iv) Mit [16, Lemma 3.142(ii)] gilt $\mathrm{dist}(\mathfrak{a},C) \leq \mathrm{dist}(\mathfrak{a},B) + \mathrm{d}(B,C) \quad \forall \mathfrak{a} \in A$. Deswegen kann man auf beiden Seiten zum Supremum übergehen und erhält $\mathrm{d}(A,C) \leq \mathrm{d}(A,B) + \mathrm{d}(B,C)$. Analog erhält man auch $\mathrm{d}(C,A) \leq \mathrm{d}(C,B) + \mathrm{d}(B,A)$. Nach Definition das Hausdorff-Abstandes folgt dann $\mathrm{d_H}(A,C) \leq \mathrm{d_H}(A,B) + \mathrm{d_H}(B,C)$.

(iii) Es sei $\mathrm{d}(A,B) \leq r$. Dann folgt $A \subset B + rB_1(0)$ mit Lemma 1.2.2(i). Ist umgekehrt $A \subset B + rB_1(0)$ folgt ebenfalls mit Lemma 1.2.2(i), dass $\mathrm{d}(A,B) \leq r$ sein muss. □

Unmittelbar aus der Definition ersichtlich ist, dass der Hausdorff-Abstand symmetrisch ist. Diese Symmetrieeigenschaft macht ihn zu einem Kandidaten für eine Metrik. Der folgende Satz unterstreicht die Bedeutung und Nützlichkeit dieses Abstandsbegriffs.

Theorem 1.2.4. *Die beiden Mengen*

$$\begin{aligned} \mathcal{C}\left(\mathbb{R}^k\right) &:= \left\{A \subset \mathbb{R}^k \mid A \text{ nichtleer und kompakt}\right\} \\ \mathcal{CC}\left(\mathbb{R}^k\right) &:= \left\{K \in \mathcal{C}\left(\mathbb{R}^k\right) \mid K \text{ konvex}\right\} \end{aligned}$$

bilden zusammen mit dem Hausdorff-Abstand einen metrischen Raum.

Beweis. Zum Beweis siehe [16, Theorem 3.162]. □

1.3. Die Stützfunktion und Ihre Eigenschaften

Im Folgenden werden wir ein wichtiges Hilfmittel aus der konvexen Analysis, die Stützfunktion betrachten. Mit ihrer Hilfe kann man konvexe Mengen beschreiben. Auch ist sie nützlich um konvexe Mengen im Rechner darzustellen.

Definition 1.3.1. Sei $A \subset \mathbb{R}^k$ eine Menge. Dann wird die *Stützfunktion von* A definiert durch

$$\delta^*(\cdot, A) : \mathbb{R}^k \longrightarrow \mathbb{R} \cup \{\pm\infty\} \quad , \quad \mathfrak{l} \longmapsto \begin{cases} \sup\limits_{\mathfrak{a} \in A} \langle \mathfrak{l}, \mathfrak{a} \rangle & \text{falls } A \neq \emptyset \\ -\infty & \text{falls } A = \emptyset \end{cases}$$

Im Folgenden sollen einige wichtige Eigenschaften der Stützfunktion notiert werden. Besonders wichtig für die Realisierung numerischer Verfahren, die mit konvexen Mengen arbeiten ist der ersten Punkt des folgenden Satzes. Dieser erlaubt die Rekonstruktion einer abgeschlossenen konvexen Menge aus der vollen Kenntnis ihrer Stützfunktion. Außerdem werden einige hilfreiche Eigenschaften der Stützfunktion bewiesen.

Satz 1.3.2. *Seien* $A,\ A_1,\ A_2 \subset \mathbb{R}^k$ *nichtleere Mengen und* $\mathfrak{l}, \mathfrak{l}_1, \mathfrak{l}_2 \in \mathbb{R}^k$. *Dann gilt:*

(i) *Für eine abgeschlossene, konvexe Menge* $K \subset \mathbb{R}^k$ *gilt:*
$K = \bigcap\limits_{\mathfrak{l} \in S_{k-1}} \left\{ \mathfrak{x} \in \mathbb{R}^k \mid \langle \mathfrak{l}, \mathfrak{x} \rangle \leq \delta^*(\mathfrak{l}, K) \right\}$.

(ii) $\delta^*(\cdot, rA) = r\delta^*(\cdot, A)$ *für* $r \geq 0$

(iii) $\delta^*(\cdot, A_1 + A_2) = \delta^*(\cdot, A_1) + \delta^*(\cdot, A_2)$

(iv) $\delta^*(r\mathfrak{l}, A) = r\delta^*(\mathfrak{l}, A)$ *für* $r \geq 0$

(v) $\delta^*(\mathfrak{l}_1 + \mathfrak{l}_2, A) \leq \delta^*(\mathfrak{l}_1, A) + \delta^*(\mathfrak{l}_2, A)$

(vi) $\delta^*(\mathfrak{l}, \mathfrak{M}A) = \delta^*\left(\mathfrak{M}^T\mathfrak{l}, A\right)$ *für* $\mathfrak{M} \in \mathbb{R}^{l \times k}$

Beweis. Zum Beweis von (ii)-(v) siehe [4, Satz 0.1.3.3.].
(i) Diese Aussage findet sich in [16, Proposition 3.61].
(vi) Es gilt $\langle \mathfrak{l}, \mathfrak{M}\mathfrak{a} \rangle = \langle \mathfrak{M}\mathfrak{a}, \mathfrak{l} \rangle = (\mathfrak{M}\mathfrak{a})^T \mathfrak{l} = \mathfrak{a}^T\mathfrak{M}^T\mathfrak{l} = \mathfrak{a}^T\left(\mathfrak{M}^T\mathfrak{l}\right) = \left\langle \mathfrak{a}, \mathfrak{M}^T\mathfrak{l} \right\rangle = \left\langle \mathfrak{M}^T\mathfrak{l}, \mathfrak{a} \right\rangle$ für $\mathfrak{a} \in A$. Zusammen mit der Definition der Stützfunktion folgt die Behauptung. □

Einige weitere Eigenschaften, die sich auf die Stützfunktion als Funktion beziehen.

Satz 1.3.3. *Sei* $A \subset \mathbb{R}^k$ *nichtleer. Dann gilt:*

(i) $\delta^*(\cdot, A)$ *ist konvex.*

(ii) Es sei A zusätzlich beschränkt, d.h. es gibt ein $r > 0$ mit $A \subset B_r(0)$. In diesem Fall ist $\delta^*(\cdot, A)$ Lipschitz-stetig mit Lipschitz-Konstante r .

Beweis. Zu (i) siehe [17, Proposition 3.109.] und zu (ii) siehe [16, Proposition 3.114.]. □

Einen sehr wichtigen Zusammenhang zwischen der Stützfunktion und dem Hausdorff-Abstand zeigt der nächste Satz auf. Er liefert die Grundlage zur Berechnung bzw. zur Approximation des Hausdorff-Abstandes zweier konvexer Mengen.

Satz 1.3.4. *Seien $K_1, K_2 \subset \mathbb{R}^k$ nichtleere, kompakte und konvexe Mengen. Dann gilt:*
$d_H(K_1, K_2) = \sup_{\mathfrak{l} \in B_1(0)} |\delta^*(\mathfrak{l}, K_1) - \delta^*(\mathfrak{l}, K_2)| = \sup_{\mathfrak{l} \in S_{k-1}} |\delta^*(\mathfrak{l}, K_1) - \delta^*(\mathfrak{l}, K_2)|$

Beweis. Nach [16, Proposition 3.156] gilt $\mathrm{d_H}(K_1, K_2) = \sup_{\mathfrak{l} \in B_1(0)} |\delta^*(\mathfrak{l}, K_1) - \delta^*(\mathfrak{l}, K_2)|$.
Für $\mathfrak{l} \in B_1(0)$ gilt wegen der positiven Homogenität aus Satz 1.3.2(iv)

$$\begin{aligned} |\delta^*(\mathfrak{l}, K_1) - \delta^*(\mathfrak{l}, K_2)| &= \|\mathfrak{l}\|_2 \left| \delta^*\left(\frac{\mathfrak{l}}{\|\mathfrak{l}\|_2}, K_1\right) - \delta^*\left(\frac{\mathfrak{l}}{\|\mathfrak{l}\|_2}, K_2\right) \right| \\ &\leq \left| \delta^*\left(\frac{\mathfrak{l}}{\|\mathfrak{l}\|_2}, K_1\right) - \delta^*\left(\frac{\mathfrak{l}}{\|\mathfrak{l}\|_2}, K_2\right) \right| . \end{aligned}$$

Damit folgt dann die zweite Gleichung. □

Schließlich soll noch ein Ergebnis angeführt werden, das den engen Zusammenhang zwischen der Stützfunktion und Mengerelationen aufzeigt.

Satz 1.3.5. *Seien $K_1, K_2 \subset \mathbb{R}^k$ nichtleer und kompakt. Dann gilt:*

$$K_1 \subset K_2 \Longleftrightarrow \delta^*(\mathfrak{l}, K_1) \leq \delta^*(\mathfrak{l}, K_2) \quad \forall \mathfrak{l} \in S_{k-1}$$

Beweis. Zum Beweis siehe [16, Proposition 3.120.]. □

Zum Schluss soll noch die Stützfunktion in einigen konkreten und wichtigen Fällen berechnet werde.

Beispiel 1.3.6. Sei $\mathfrak{l} \in \mathbb{R}^k$. Wir geben für einige spezielle Mengen $A \subset \mathbb{R}^k$ die Stützfunktion an. Diese Beispiele finden sich in [4, Tabelle 3.1, S176].

(i) $A = co\{\mathfrak{x}^1, \mathfrak{x}^2\}$ eine Strecke im $\mathbb{R}^k$. Dann ist $\delta^*(\mathfrak{l}, A) = \max\{\langle \mathfrak{l}, \mathfrak{x}^1 \rangle, \langle \mathfrak{l}, \mathfrak{x}^2 \rangle\}$.

(ii) $A = [-1, 1]^k$ die Maximumskugel im $\mathbb{R}^k$. Dann ist $\delta^*(\mathfrak{l}, A) = \|\mathfrak{l}\|_1$.

(iii) $A = B_r(\mathfrak{m})$ eine Kugel im $\mathbb{R}^k$. In diesem Fall ist $\delta^*(\mathfrak{l}, A) = \langle \mathfrak{l}, \mathfrak{m} \rangle + r \|\mathfrak{l}\|_2$.

(iv) $A = M_1 \times M_2$ ein Kreuzprodukt mit $M_i \subset \mathbb{R}^{k_i} \quad (i = 1, 2)$ und $k_1 + k_2 = k$. Weiter sei $\mathfrak{l} = \begin{pmatrix} \mathfrak{l}^1 \\ \mathfrak{l}^2 \end{pmatrix}$ mit $\mathfrak{l}^i \in \mathbb{R}^{k_i}$ $(i = 1, 2)$. Dann ist $\delta^*(\mathfrak{l}, A) = \delta^*(\mathfrak{l}^1, M_1) + \delta^*(\mathfrak{l}^2, M_2)$.

1.4. Die Stützpunktmenge

In diesem Abschnitt soll ein mit der Stützfunktion eng verwandtes Mittel der konvexen Analysis vorgestellt werden um konvexe Mengen zu charakterisieren.

Definition 1.4.1. Sei $A \subset \mathbb{R}^k$ nichtleer, und $\mathfrak{l} \in \mathbb{R}^k$. Dann wird die Stützpunktmenge in Richtung $\mathfrak{l}$ an A definiert als:

$$Y(\mathfrak{l}, A) := \{\mathfrak{a} \in A \mid \langle \mathfrak{l}, \mathfrak{a} \rangle = \delta^* (\mathfrak{l}, A)\} .$$

Mit $y(\mathfrak{l}, A)$ bezeichnen wir ein Element aus $Y(\mathfrak{l}, A)$ und nennen ein solches Element Stützpunkt von A in Richtung $\mathfrak{l}$.

Bemerkung 1.4.2.

(i) Ein Stützpunkt $y(\mathfrak{l}, A)$ ist also ein Element von A in dem das Supremum der Stützfunktion angenommen wird, d.h. $\langle \mathfrak{l}, y(\mathfrak{l}, A) \rangle = \delta^* (\mathfrak{l}, A) = \sup_{\mathfrak{a} \in A} \langle \mathfrak{l}, \mathfrak{a} \rangle$. Dieser Stützpunkt muss nicht eindeutig sein, d.h. $Y(\mathfrak{l}, A)$ kann mehr als ein Element enthalten.

(ii) In obiger Definition kann die Stützpunktmenge auch leer sein. Die Stützpunktmenge ist nur für konvexe und kompakte Mengen wirklich sinnvoll wie wir später noch sehen werden.

(iii) Die Definition der Stützpunktmenge ist unabhängig von der Norm von $\mathfrak{l}$, d.h. für $\mathfrak{l} \in \mathbb{R}^k, \mathfrak{l} \neq 0_{\mathbb{R}^k}$ gilt: $Y(\mathfrak{l}, A) = Y(r\mathfrak{l}, A) \qquad \forall r > 0$

(iv) Anschaulich kann man sich einen Stützpunkt so vorstellen:
Aus dem "Unendlichen" schiebt man eine Hyperebene, die durch $\mathfrak{l} \in \mathbb{R}^k$ charakterisiert wird, in Gegenrichtung an die Menge A heran. Ein Berührpunkt dieser Hyperebene mit der Menge A ist dann ein Stützpunkt $y(\mathfrak{l}, A)$. Im konvexen Fall ist dieser Stützpunkt oft eindeutig. Im strikt konvexen Fall ist er immer eindeutig.

Als nächstes folgt ein wichtiges Ergebnis, das die Rekonstruktion einer konvexen Menge aus ihren Stützpunkten erlaubt. Damit wird der duale Zugang mit Stützpunkten zur Beschreibung einer konvexen Menge gerechtfertigt.

Satz 1.4.3. *Sei $K \subset \mathbb{R}^k$ eine kompakte, konvexe und nichtleere Menge. Dann gilt:*

(i) $\partial K = \bigcup_{\mathfrak{l} \in S_{k-1}} Y(\mathfrak{l}, K).$

(ii) $K = co(\partial K) = \mathrm{co} \left(\bigcup_{\mathfrak{l} \in S_{k-1}} Y(\mathfrak{l}, K) \right)$

(iii) $K = \overline{\text{co}}\{y(\mathfrak{l}, K) \mid \mathfrak{l} \in S_{k-1}\}$, *wobei* $y(\mathfrak{l}, K) \in Y(\mathfrak{l}, K)$ *eine beliebige Auswahl an Stützpunkten ist.*
Ist die Menge $\{y(\mathfrak{l}, K) \mid \mathfrak{l} \in S_{k-1}\}$ *abgeschlossen, so genügt es die konvexe Hülle zu bilden.*

Beweis. (i) Mit [16, Lemma 3.52] folgt, dass für $\mathfrak{x} \in \partial K$ gilt: $\exists \mathfrak{l} \in S_{k-1}$ mit $\langle \mathfrak{l}, \mathfrak{x} \rangle = \delta^*(\mathfrak{l}, K)$. Das bedeutet $\mathfrak{x} \in Y(\mathfrak{l}, K)$. Ist andererseits $\mathfrak{x} \in Y(\mathfrak{l}, K)$, dann gilt nach Definition: $\langle \mathfrak{l}, \mathfrak{x} \rangle = \sup_{\mathfrak{k} \in K} \langle \mathfrak{l}, \mathfrak{k} \rangle$ und $\mathfrak{x} \in K$. Wäre $\mathfrak{x} \in \text{int}(K)$ so gäbe es ein $\epsilon > 0$ mit $B_\epsilon(\mathfrak{x}) \subset K$. Und in dieser Kugel gäbe es ein Element $\mathfrak{z}$ mit $\langle \mathfrak{l}, \mathfrak{z} \rangle > \langle \mathfrak{l}, \mathfrak{x} \rangle$. Dies wäre ein Widerspruch dazu, dass das Supremum in $\mathfrak{x}$ angenommen wird. Also ist $\mathfrak{x} \in \partial K$.

(ii) Mit dem Theorem von Krein-Milman (vgl. [26, sec. 32, Theorem D, p. 84]) folgt die Darstellung von K als konvexe Hülle ihrer Extrempunkte. Nach [18, Proposition 3.25] ist aber die Menge der Extrempunkte Teilmenge des Randes von K. Also ist $\text{co}(\partial K)$ eine Obermenge von der konvexen Hülle der Extrempunkte von K, denn die konvexe Hülle erhält die Inklusion. Also gilt $K \subset \text{co}(\partial K)$. Andererseits ist K kompakt und deswegen ist $\partial K \subset K$. Aber da K auch konvex ist, gilt nach Definition 1.1.5 der konvexen Hülle $\text{co}(\partial K) \subset K$, denn K gehört zu den Mengen, über die der Schnitt gebildet wird. Also insgesamt gilt $K = \text{co}(\partial K)$. Mit (i) folgt die zweite Gleichung.

(iii) siehe [4, Satz 3.1.3.3.(ii)]. □

Im Folgenden wollen wir noch einige Rechenregeln für den Umgang mit der Stützpunktmenge bereitstellen. Diese Regeln erleichtern die Berechnung von Stützpunktmengen.

Proposition 1.4.4. *Sein* $K, K_1, K_2 \subset \mathbb{R}^k$ *konvex. Weiter sei* $A \in \mathbb{R}^{l \times k}$ *. Dann gilt:*

(i) $Y(\mathfrak{l}, K_1 + K_2) = Y(\mathfrak{l}, K_1) + Y(\mathfrak{l}, K_2) \qquad \left(\mathfrak{l} \in \mathbb{R}^k, \mathfrak{l} \neq 0_{\mathbb{R}^k}\right)$

(ii) $Y(\mathfrak{l}, A \cdot K) = A \cdot Y\left(A^T \mathfrak{l}, K\right) \qquad \left(\mathfrak{l} \in \mathbb{R}^l, \mathfrak{l} \neq 0_{\mathbb{R}^l}\right)$

(iii) $Y(\mathfrak{l}, r \cdot K) = r \cdot Y(\mathfrak{l}, K) \qquad \left(r \geq 0, \mathfrak{l} \in \mathbb{R}^k, \mathfrak{l} \neq 0_{\mathbb{R}^k}\right)$

Beweis. Der Beweis findet sich in [4, Satz 3.1.3.4], wobei die Eigenschaften der Stützfunktion auf denen dieser Beweis beruht auch in diesem Skript in Satz 1.3.2 zu finden sind. □

Jetzt sollen noch die Stützpunktmengen in einigen wichtigen Spezialfällen berechnet werden. Alle Stützpunktmengen, die in dieser Arbeit gebraucht werden, können mit Hilfe der obigen Rechenregeln auf diese Spezialfälle zurückgeführt werde.

Beispiel 1.4.5. Sei $\mathfrak{l} \in S_{k-1}$ und $K \subset \mathbb{R}^k$ kompakt und konvex. Im Fall der Eindeutigkeit, geben wir den Stützpunkt an, sonst die Stützpunktmenge an die Menge K in Richtung $\mathfrak{l}$. Einige dieser Beispiele finden sich in [4, Tabelle 3.2 , S. 186].

(i) $K = \{\mathfrak{k}\}$, $\mathfrak{k} \in \mathbb{R}^k$ ein einzelner Punkt. Dann haben wir trivialerweise: $y(\mathfrak{l}, K) = \mathfrak{k}$.

(ii) $K = B_r(\mathfrak{m})$ die euklidische Einheitskugel mit Radius $r > 0$ und Mittelpunkt $\mathfrak{m} \in \mathbb{R}^k$. Dann ist $y(\mathfrak{l}, K) = \mathfrak{m} + r \cdot \mathfrak{l}$. Denn

$$\delta^*(\mathfrak{l}, K) = \sup_{\mathfrak{k} \in K} \langle \mathfrak{l}, \mathfrak{k} \rangle = \langle \mathfrak{l}, \mathfrak{m} \rangle + r \cdot \sup_{\mathfrak{s} \in s_{k-1}} \langle \mathfrak{l}, \mathfrak{s} \rangle = \langle \mathfrak{l}, \mathfrak{m} \rangle + r \left\| \mathfrak{l} \right\|^2 ,$$

wobei man die letzte Gleichung mit der Cauchy-Schwarzschen Ungleichung begründen kann. D.h. das Supremum wird eindeutig in $\mathfrak{k} = \mathfrak{m} + r\mathfrak{l}$ angenommen. Der Stützpunkt ist eindeutig.

(iii) $K = \operatorname{co}\{\mathfrak{u}^1, \mathfrak{u}^2\}$ eine Strecke mit den Endpunkten $\mathfrak{u}^1, \mathfrak{u}^2 \in \mathbb{R}^k$. Dann haben wir: $y(\mathfrak{l}, K) = \begin{cases} \mathfrak{u}^2 & \text{falls } \langle \mathfrak{l}, \mathfrak{u}^2 - \mathfrak{u}^1 \rangle > 0 \\ \mathfrak{u}^1 & \text{falls } \langle \mathfrak{l}, \mathfrak{u}^2 - \mathfrak{u}^1 \rangle < 0 \end{cases}$. Falls $\langle \mathfrak{l}, \mathfrak{u}^2 - \mathfrak{u}^1 \rangle = 0$ ist, ist der Stützpunkt an diese Menge nicht mehr eindeutig. In diesem Fall haben wir: $Y(\mathfrak{l}, K) = K$.

(iv) $K = [a, b] \subset \mathbb{R}$ ein reelles Intervall. Dann ist $k = 1$ und $\mathfrak{l} = l = \pm 1$. Dies ist ein Spezialfall von dem vorigen Beispiel. Es gilt $y(l, K) = \begin{cases} b & \text{falls } l > 0 \\ a & \text{falls } l < 0 \end{cases}$. Falls $l = 0$ ist, ist der Stützpunkt nicht mehr eindeutig. Die Stützpunktmenge ist dann $Y(0, K) = K$.

(v) Seien $M_1 \subset \mathbb{R}^{k_1}$ und $M_2 \subset \mathbb{R}^{k_2}$ mit $k_1 + k_2 = k$. Wir wollen den Stützpunkt an $K = M_1 \times M_2$ berechnen. Dazu zerlegen wir $\mathfrak{l}$ in $\mathfrak{l}^1 \in \mathbb{R}^{k_1}$ und $\mathfrak{l}^2 \in \mathbb{R}^{k_2}$, d.h. $\mathfrak{l} = \begin{pmatrix} \mathfrak{l}^1 \\ \mathfrak{l}^2 \end{pmatrix}$. Dann gilt $y(\mathfrak{l}, K) = y\left(\begin{pmatrix} \mathfrak{l}^1 \\ \mathfrak{l}^2 \end{pmatrix}, M_1 \times M_2\right) = \begin{pmatrix} y(\mathfrak{l}^1, M_1) \\ y(\mathfrak{l}^2, M_2) \end{pmatrix}$. Falls die Stützpunkte nicht eindeutig sind, haben wir $Y\left(\begin{pmatrix} \mathfrak{l}^1 \\ \mathfrak{l}^2 \end{pmatrix}, M_1 \times M_2\right) = Y(\mathfrak{l}^1, M_1) \times Y(\mathfrak{l}^2, M_2)$.

1.5. Mengenwertige Abbildungen und das Aumann-Integral

Dieser Abschnitt schließt inhaltlich an die vorigen Abschnitte dieses Kapitels an und ist eine kurze Zusammenstellung über das Aumann-Integral und mengenwertige Abbildungen. Die Darstellung lehnt sich an [16] von R. Baier und M. Gerdts an. Dort findet man auch nähere Details und Verweise auf entsprechende Literatur.

Zunächst sollen die grundlegenden Definitionen bereitgestellt werden.

Definition 1.5.1. Sei $I \subset \mathbb{R}$ ein Intervall. Eine *mengenwertige Abbildung* F ist eine Abbildung, die jedem $t \in I$ eine Menge $F(t) \subset \mathbb{R}^k$ zuordnet. Dafür schreiben wir kurz $F : I \Longrightarrow \mathbb{R}^k$.

Sie heißt *messbar*, wenn für alle offenen Mengen $O \subset \mathbb{R}^k$ das Urbild

$$F^{-1}(O) := \{t \in I \mid F(t) \cap O \neq \emptyset\}$$

Lebesgue-messbar ist.

Im weiteren seien alle Bilder nichtleer, d.h. $\forall t \in I : \quad F(t) \neq \emptyset$.

Dann heißt $F(\cdot)$ *stetig* in $\hat{t} \in I$, wenn es für alle $\epsilon > 0$ ein $\delta > 0$ gibt, sodass für alle $t \in I$ mit $\left|\hat{t} - t\right| < \delta$ gilt $\mathrm{d_H}\left(F\left(\hat{t}\right), F(t)\right) < \epsilon$. Man bezeichnet $F(\cdot)$ als stetig, wenn es in jedem $t \in I$ stetig ist.

Weiter heißt $F(\cdot)$ *beschränkt*, wenn es eine Konstante $C > 0$ gibt, mit $\|F(t)\|_2 \leq C$ für alle $t \in I$. Und es heißt *integrierbar beschränkt*, wenn es eine Funktion $k \in L^1(I)$ gibt, sodass $\|F(t)\|_2 \leq k(t)$ für fast alle $t \in I$ gilt.

Wichtig für diese Arbeit ist die folgende

Proposition 1.5.2. *Sei $K \subset \mathbb{R}^l$ kompakt, konvex und nichtleer und $I \subset \mathbb{R}$ ein kompaktes Intervall. Weiter sei $\mathfrak{A} : I \longrightarrow \mathbb{R}^{k\times l}$ eine stetige matrixwertige Abbildung. Dann ist die mengenwertige Abbildung $F : I \Longrightarrow \mathbb{R}^k$, $F(t) := \mathfrak{A}(t)K$ stetig, messbar und integrierbar beschränkt.*

Beweis. a) Stetigkeit. Sei $\hat{t} \in I$ und $\epsilon > 0$. Da $\mathfrak{A}(\cdot)$ stetig ist, gibt es ein $\delta > 0$, sodass für alle $t \in I$ mit $\left|\hat{t} - t\right| < \delta$ gilt: $\left\|\mathfrak{A}\left(\hat{t}\right) - \mathfrak{A}(t)\right\|_{Sp} < \frac{\epsilon}{\|K\|_2}$. Dabei ist mit $\|\cdot\|_{Sp}$ die Spektralnorm für Matrizen gemeint. Sie wird in Beispiel 1.6.4 eingeführt. Dann folgt mit [16, Proposition 3.153(ii)] für alle $t \in I$ mit $\left|\hat{t} - t\right| < \delta$: $\mathrm{d_H}\left(F\left(\hat{t}\right), F(t)\right) \leq \left\|\mathfrak{A}\left(\hat{t}\right) - \mathfrak{A}(t)\right\|_{Sp} \cdot \|U\|_2 < \epsilon$. Da $\hat{t} \in I$ beliebig war, ist $F(\cdot)$ stetig auf I.

b) Messbarkeit. Sei $V := \mathbb{Q}^l \cap K$. Dann ist V abzählbar und dicht in K, d.h. der Abschluss von V ist K. Dann ist auch $\mathfrak{A}(t) \cdot V$ abzählbar und dicht in $\mathfrak{A}(t) \cdot K$. Sei nun $(\mathfrak{v}_i)_{i\in\mathbb{N}}$ eine Abzählung von V. Dann definieren wir die vektorwertigen Funktionen $\mathfrak{f}_i(t) := \mathfrak{A}(t) \cdot \mathfrak{v}_i$ für alle $i \in \mathbb{N}$. Für alle $t \in I$ und $i \in \mathbb{N}$ ist dann offensichtlich $\mathfrak{f}_i(t) \in F(t)$. Insgesamt haben wir also

$$F(t) = \overline{\cup_{i\in\mathbb{N}} \{\mathfrak{f}_i(t)\}} \quad \text{für alle } t \in I\,.$$

Mit [2, Theorem 8.1.4] folgt die Messbarkeit von $F(t)$, da alle $\mathfrak{f}_i(\cdot)$ stetig und deswegen auch messbar sind. Außerdem ist $F(t)$ für alle $t \in I$ nichtleer und abgeschlossen, da das Bild eines Kompaktums unter einer stetigen Abbildung wieder kompakt ist.

c) Integrierbare Beschränktheit. Für ein beliebiges $\mathfrak{k} \in K$ und $t \in I$ gilt $\|\mathfrak{A}(t) \cdot \mathfrak{k}\|_2 \leq \|\mathfrak{A}(t)\|_{Sp} \cdot \|\mathfrak{k}\|_2$. Damit folgt $\|\mathfrak{A}(t) \cdot K\|_2 \leq \|\mathfrak{A}(t)\|_{Sp} \|K\|_2$. Es ist $\|\mathfrak{A}(t)\|_{Sp}$ stetig und auf dem kompakten Intervall I auch beschränkt. Also ist $k(t) := \|\mathfrak{A}(t)\|_{Sp} \|K\|_2$ integrierbar über I. Und wir haben $\|F(t)\|_2 \leq k(t)$ für alle $t \in I$. □

Damit können wir nun das Aumann-Integral einführen.

Definition 1.5.3. Sei $I \subset \mathbb{R}$ ein Intervall und $F : I \rightrightarrows \mathbb{R}^k$ eine mengenwertige Abbildung. Eine Funktion $\mathfrak{f} : I \longrightarrow \mathbb{R}^k$ heißt integrierbare Auswahl von $F(\cdot)$, wenn $\mathfrak{f} \in L^1(I)^k$ ist und $\mathfrak{f}(t) \in F(t)$ gilt für fast alle $t \in I$.
Die Menge

$$\int_I F(t)\,dt := \left\{ \int_I \mathfrak{f}(t)\,dt \mid \mathfrak{f} \text{ integrierbare Auswahl von } F(\cdot) \right\}$$

heißt das (Aumann-)Integral von $F(\cdot)$.

Schließlich folgt noch eine wichtige Eigenschaft des Aumann-Integrals.

Proposition 1.5.4. *Sei $I \subset \mathbb{R}$ ein kompaktes Intervall und $F : I \rightrightarrows \mathbb{R}^k$ eine messbare und integrierbar beschränkte mengenwertige Abbildung mit nichtleeren und abgeschlossenen Bildern. Dann ist das Aumann-Integral*

$$\int_I F(t)\,dt$$

konvex, kompakt und nichtleer.
Außerdem ist $t \mapsto \delta^*(\mathfrak{l}, F(t))$ für jeden Vektor $\mathfrak{l} \in \mathbb{R}^k$ messbar und es gilt

$$\delta^*\left(\mathfrak{l}, \int_I F(t)\,dt\right) = \int_I \delta^*(\mathfrak{l}, F(t))\,dt\,.$$

Weiter ist auch $t \mapsto Y(\mathfrak{l}, F(t))$ jeden Vektor $\mathfrak{l} \in \mathbb{R}^k$ eine messbare, integrierbar beschränkt mengenwertige Abbildung mit abgeschlossenen Bilder und es gilt

$$Y\left(\mathfrak{l}, \int_I F(t)\,dt\right) = \int_I Y(\mathfrak{l}, F(t))\,dt\,.$$

Beweis. Die Aussage über das Aumann-Integral findet man in [3]. Die Messbarkeit der Stützfunktion folgt aus [2, Theorem 8.2.14]. Die Vertauschbarkeit des Integrals mit Stützfunktion folgt aus [2, Proposition 8.6.2, 3.].
Nun kommen wir zu den Aussagen über die Stützpunktmenge. Nach Definition der Stützpunktmenge ist $Y(\mathfrak{l}, F(t))$ Teilmenge von $F(t)$. Deswegen ist $Y(\mathfrak{l}, F(t))$ auch für jeden Vektor $\mathfrak{l} \in \mathbb{R}^k$ integrierbar beschränkt. Es ist $Y(\mathfrak{l}, F(t))$ der Schnitt von der abgeschlossenen Menge $F(t)$ mit der (abgeschlossenen) Hyperebene

$$\{\mathfrak{x} \in \mathbb{R}^n \mid \langle \mathfrak{l}, \mathfrak{x} \rangle = \delta^*(\mathfrak{l}, F(t))\}\,.$$

Deswegen ist $Y(\mathfrak{l}, F(t))$ stets abgeschlossen.
Jetzt zeigen wir: Für einen beliebigen Vektor $\mathfrak{l} \in \mathbb{R}^k$ ist die mengenwertige Abbildung $t \mapsto Y(\mathfrak{l}, F(t))$ messbar. Dazu definieren wir die Funktion $f : I \times \mathbb{R}^k \longrightarrow \mathbb{R}$ durch

$f(t,\mathfrak{x}) := -\langle \mathfrak{l}, \mathfrak{x}\rangle$. Dann ist $t \mapsto f(t,\mathfrak{x})$ für alle $\mathfrak{x} \in \mathbb{R}^k$ eine konstante Funktion und damit stetig und messbar. Ebenso ist $\mathfrak{x} \mapsto f(t,\mathfrak{x})$ für alle $t \in I$ stetig. Dies bedeutet f ist eine Carathéodory-Funktion. Schließlich definieren wir noch die mengenwertige Abbildung $R : I \rightrightarrows \mathbb{R}^k$ durch $R(t) := \left\{\mathfrak{x} \in F(t) \mid f(t,\mathfrak{x}) = \inf_{\mathfrak{y}\in F(t)} f(t,\mathfrak{y})\right\}$. Dann ist $Y(\mathfrak{l}, F(t)) = R(t)$ für alle $t \in I$, denn $\sup_{\mathfrak{y}\in F(t)} \langle \mathfrak{l}, \mathfrak{y}\rangle = -\inf_{\mathfrak{y}\in F(t)} -\langle \mathfrak{l}, \mathfrak{y}\rangle$. Mit [2, Theorem 8.2.11 (Marginal Map)] folgt dann, dass $t \mapsto Y(\mathfrak{l}, F(t))$ messbar ist.
Schließlich zeigen wir noch die Gleichung $Y\left(\mathfrak{l}, \int_I F(t)\,dt\right) = \int_I Y(\mathfrak{l}, F(t))\,dt$.
" $\subset$ " : Sei $\mathfrak{x} \in Y\left(\mathfrak{l}, \int_I F(t)\,dt\right)$. Dies bedeutet $\mathfrak{x} \in \int_I F(t)\,dt$ und $\langle \mathfrak{l}, \mathfrak{x}\rangle = \delta^*\left(\mathfrak{l}, \int_I F(t)\,dt\right)$. Also gibt es nach Definition des Aumann-Integrals ein $\mathfrak{f} \in L^1(I)$ mit $\mathfrak{f}(t) \in F(t)$ für $\forall' t \in I$ mit $\mathfrak{x} = \int_I \mathfrak{f}(t)\,dt$. Da das Skalarprodukt linear ist, gilt:

$$\langle \mathfrak{l}, \mathfrak{x}\rangle = \left\langle \mathfrak{l}, \int_I \mathfrak{f}(t)\,dt \right\rangle = \int_I \langle \mathfrak{l}, \mathfrak{f}(t)\rangle\,dt$$

Mit der Aussage für die Stützfunktion haben wir also

$$\int_I \langle \mathfrak{l}, \mathfrak{f}(t)\rangle\,dt = \delta^*\left(\mathfrak{l}, \int_I F(t)\,dt\right) = \int_I \delta^*(\mathfrak{l}, F(t))\,dt\,.$$

Also gilt

$$\int_I \langle \mathfrak{l}, \mathfrak{f}(t)\rangle - \delta^*(\mathfrak{l}, F(t))\,dt = 0\,.$$

Es ist aber nach Definition der Stützfunktion $\langle \mathfrak{l}, \mathfrak{f}(t)\rangle \leq \delta^*(\mathfrak{l}, F(t))$. Deswegen muss also $\langle \mathfrak{l}, \mathfrak{f}(t)\rangle = \delta^*(\mathfrak{l}, F(t))$ für fast alle $t \in I$ gelten. Deswegen ist $\mathfrak{f}(t) \in Y(\mathfrak{l}, F(t))$ für fast alle $t \in I$. Dies bedeutet, dass $\mathfrak{f}(\cdot)$ eine integrierbare Auswahl von $Y(\mathfrak{l}, F(\cdot))$ ist. Also ist $\mathfrak{x} = \int_I \mathfrak{f}(t)\,dt \in \int_I Y(\mathfrak{l}, F(t))\,dt$.
" $\supset$ " : Sei $\mathfrak{x} \in \int_I Y(\mathfrak{l}, F(t))\,dt$. Dies bedeutet, dass es eine integrierbare Auswahl $\mathfrak{f} \in L^1(I)$ mit $\mathfrak{f}(t) \in Y(\mathfrak{l}, F(t))$ für $\forall' t \in I$ mit $\mathfrak{x} = \int_I \mathfrak{f}(t)\,dt$. Aber $\mathfrak{f}(t) \in Y(\mathfrak{l}, F(t))$ bedeutet, dass $\mathfrak{f}(t) \in F(t)$ ist und $\langle \mathfrak{l}, \mathfrak{f}(t)\rangle = \delta^*(\mathfrak{l}, F(t))$ ist. Damit ist $\mathfrak{f}(t)$ auch eine integrierbare Auswahl von $F(\cdot)$. Also ist $\mathfrak{x} \in \int_I F(t)\,dt$. Wegen der Linearität des Skalarproduktes und mit obiger Aussage über die Stützfunktion gilt

$$\begin{aligned}\langle \mathfrak{l}, \mathfrak{x}\rangle &= \left\langle \mathfrak{l}, \int_I \mathfrak{f}(t)\,dt \right\rangle = \int_I \langle \mathfrak{l}, \mathfrak{f}(t)\rangle\,dt \\ &= \int_I \delta^*(\mathfrak{l}, F(t))\,dt = \delta^*\left(\mathfrak{l}, \int_I F(t)\,dt\right)\,.\end{aligned}$$

Dies bedeutet, dass $\mathfrak{x} \in Y\left(\mathfrak{l}, \int_I F(t)\,dt\right)$ ist. □

1.6. Matrixnormen

In diesem Abschnitt sollen kurz Matrixnormen eingeführt werden. Außerdem werden einige spezielle Matrixnormen vorgestellt. Weitere Details findet man in [22, Abschnitt 2.2] und [12, § 9]. Im Folgenden sei $\|\cdot\|$ eine beliebige Vektor bzw. Matrixnorm.

Wir beginnen mit der Definition einer Matrixnorm.

Definition 1.6.1 (Matrixnorm)**.** Eine *Matrixnorm* ist eine Abbildung $\|\cdot\| : \mathbb{R}^{l\times k} \longrightarrow [0,\infty)$ die für alle $\mathfrak{A},\mathfrak{B} \in \mathbb{R}^{l\times k}$ und $r \in \mathbb{R}$ folgende Eigenschaften hat

(i) $\|\mathfrak{A}\| = 0 \iff \mathfrak{A} = 0_{\mathbb{R}^{l\times k}}$

(ii) $\|r\mathfrak{A}\| = |r|\,\|\mathfrak{A}\|$

(iii) $\|\mathfrak{A}+\mathfrak{B}\| \leq \|\mathfrak{A}\| + \|\mathfrak{B}\|$.

Sie heißt *submultiplikativ*, wenn für alle $\mathfrak{C} \in \mathbb{R}^{k\times j}$ gilt

$$\|\mathfrak{A}\cdot\mathfrak{C}\| \leq \|\mathfrak{A}\|\cdot\|\mathfrak{C}\| \ .$$

Sie heißt *verträglich* mit den Vektornormen $\|\cdot\|_a$ auf $\mathbb{R}^k$ und $\|\cdot\|_b$ auf $\mathbb{R}^l$, wenn für alle $\mathfrak{v} \in \mathbb{R}^k$ gilt

$$\|\mathfrak{A}\mathfrak{v}\|_b \leq \|\mathfrak{A}\|\cdot\|\mathfrak{v}\|_a \ .$$

Die wichtigsten und bekanntesten Matrixnormen sind die so genannten lub-Normen. Eine *lub-Norm* ist eine Abbildung $\mathrm{lub}(\cdot) : \mathbb{R}^{l\times k} \longrightarrow [0,\infty)$, die für eine Matrix $\mathfrak{A} \in \mathbb{R}^{l\times k}$ definiert ist mittels

$$\mathrm{lub}(\mathfrak{A}) := \sup_{\substack{\mathfrak{v}\in\mathbb{R}^k \\ \mathfrak{v}\neq 0}} \frac{\|\mathfrak{A}\mathfrak{v}\|}{\|\mathfrak{v}\|} \ .$$

Diese Abbildung hat folgende Eigenschaften:

Satz 1.6.2. *Eine lub-Norm auf $\mathbb{R}^{l\times k}$ ist eine submultiplikative Matrixnorm. Sie ist verträglich mit den Vektornormen $\|\cdot\|$ auf $\mathbb{R}^l$ und $\mathbb{R}^k$ durch die sie induziert wurde.*

Beweis. siehe [12, Abschnitt 9.3, S. 136] □

Da im Zusammenhang mit Matrixnormen auch die Vektornormen von Bedeutung sind, sollen im folgenden Beispiel die bekannten p-Normen für Vektoren vorgestellt werden.

Beispiel 1.6.3. Sei $\mathfrak{v} = (v_1,\ldots,v_k)^T \in \mathbb{R}^k$. Dann definiert man für $p \in [1,\infty]$ die p-Normen als

$$\|\mathfrak{v}\|_p = \left(\sum_{i=1}^{k} |v_i|^p\right)^{\frac{1}{p}} \text{ für } p \in [1,\infty) \text{ und } \|\mathfrak{v}\|_\infty = \max_{i=1,\ldots,k} |v_i| \ .$$

Die ∞-Norm nennt man auch Maximumsnorm.

Um zu sehen, dass diese Abbildungen für $p < \infty$ die Dreiecksungleichung erfüllen, braucht man die Minkowski-Ungleichung, welche man beispielsweise in [20, Abschnitt 9.8, S. 162] findet. Die anderen Eigenschaften einer Norm sind für diese Abbildungen leicht nachzuweisen.

Schließlich werden im nächsten Beispiel noch einige wichtige lub-Normen vorgestellt.

Beispiel 1.6.4. Sei $\mathfrak{A} = (a_{i,j}) \in \mathbb{R}^{l\times k}$. Es folgen einige lub-Normen zusammen mit den Vektornormen, die sie induzieren. Die zugehörigen Beweise findet man in [22, Abschnitt 2.2].

(i) Zeilensummennorm

$$\|\mathfrak{A}\|_Z = \max_{i=1,\dots,l} \sum_{j=1}^{k} |a_{i,j}| \ .$$

Sie ist verträglich mit der Maximumsnorm $\|\cdot\|_\infty$.

(ii) Spaltensummennorm

$$\|\mathfrak{A}\|_S = \max_{j=1,\dots,k} \sum_{i=1}^{l} |a_{i,j}| \ .$$

Sie ist verträglich mit der 1-Norm $\|\cdot\|_1$.

(iii) Spektralnorm

$$\|\mathfrak{A}\|_{Sp} = \sup_{\substack{\mathfrak{v}\in\mathbb{R}^k \\ \mathfrak{v}\neq 0}} \sqrt{\frac{\mathfrak{v}^T\mathfrak{A}^T\mathfrak{A}\mathfrak{v}}{\mathfrak{v}^T\mathfrak{v}}} = \sqrt{\lambda_{max}\left(\mathfrak{A}^T\mathfrak{A}\right)}\,.$$

Dabei ist $\lambda_{max}\left(\mathfrak{A}^T\mathfrak{A}\right)$ der betragsgrößte Eignewert von $\mathfrak{A}^T\mathfrak{A}$. Sie ist verträglich mit der euklidischen Norm $\|\cdot\|_2$.

1.7. Das Lebesgue-Integral und Sobolev-Räume

In diesem Abschnitt sollen einige Definitionen und Resultate über das Lebesgue-Integral und schwach differenzierbare Funktionen zusammengestellt werden.

Alle in diesem Abschnitt, wie auch in der ganzen Arbeit vorkommende Integrale sind Lebesgue-Integrale. Vorausgesetzt wird die Kenntnis der Lebesgue-Integrationstheorie für reellwertige Funktionen, d.h. die Definition des Lebesgue-Integrals, der L^p-Räume und der L^p- Normen und Ergebnisse wie die Minkowski- und Hölder-Ungleichung etc. . Diese Theorie wird beispielsweise in [11] oder besser [31] ausführlich dargestellt.

Mit $L^p(\Omega)$ $(1 \leq p \leq \infty)$ bezeichnen wir den L^p- Raum der Funktionen $f : \Omega \longrightarrow \mathbb{R}$ mit $\Omega \subset \mathbb{R}^l$. Die dazugehörige Norm wird mit $\|\cdot\|_{L^p(\Omega)}$ bezeichnet.

Es folgt eine kurze Einführung des Lebesgue-Integrals und der L^p-Räume für vektorwertige Funktionen.

Gegeben sei eine messbare Funktion $\mathfrak{f} = \begin{pmatrix} f_1 \\ \vdots \\ f_k \end{pmatrix} : \Omega \longrightarrow \mathbb{R}^k$ mit $\Omega \subset \mathbb{R}^l$ messbar. Dann sind auch alle $f_i : \Omega \longrightarrow \mathbb{R}$ messbar und umgekehrt (siehe [7, Kapitel III, §22, Bemerkung 2]). Das Lebesgue-Integral von $\mathfrak{f}$ sei definiert als

$$\int_\Omega \mathfrak{f}\, d\lambda = \begin{pmatrix} \int_\Omega f_1\, d\lambda \\ \vdots \\ \int_\Omega f_k\, d\lambda \end{pmatrix}.$$

Statt dem Symbol $\int_\Omega d\lambda$ werden wir auch $\int_\Omega dt$ verwenden. Die dazugehörigen L^p-Räume sind für $1 \leq p < \infty$ definiert als

$$L^p(\Omega)^k := \left\{ \mathfrak{f} = \begin{pmatrix} f_1 \\ \vdots \\ f_k \end{pmatrix} : \Omega \longrightarrow \mathbb{R}^k \mid \mathfrak{f} \text{ meßbar und } \int_\Omega \|\mathfrak{f}(t)\|_p^p\, dt < \infty \right\}$$

bzw. für $p = \infty$ als

$$L^\infty(\Omega)^k := \left\{ \mathfrak{f} = \begin{pmatrix} f_1 \\ \vdots \\ f_k \end{pmatrix} : \Omega \longrightarrow \mathbb{R}^k \mid \mathfrak{f} \text{ meßbar und } \max_{i=1,\ldots,k} \|f_i\|_{L^\infty(\Omega)} < \infty \right\}.$$

Für $\mathfrak{f} \in L^p(\Omega)^k$ führen wir die Norm

$$\|\mathfrak{f}\|_{L^p(\Omega)^k} := \left(\int_\Omega \|\mathfrak{f}(t)\|_p^p\, dt \right)^{\frac{1}{p}} = \left(\sum_{i=1}^k \|f_i\|_{L^p(\Omega)}^p \right)^{\frac{1}{p}} \quad (1 \leq p < \infty)$$

und

$$\|\mathfrak{f}\|_{L^\infty(\Omega)^k} := \max_{i=1,\ldots,k} \|f_i\|_{L^\infty(\Omega)}$$

ein. Wenn die Bedeutung aus dem Zusammenhang klar ist, schreiben wir auch kurz $\|\mathfrak{f}\|_{L^p}$ statt $\|\mathfrak{f}\|_{L^p(\Omega)^k}$. Man kann leicht zeigen, dass dies tatsächlich Normen sind. Es sind dann $\left(L^p(\Omega)^k, \|\cdot\|_{L^p(\Omega)^k} \right)$ Banach-Räume für $1 \leq p \leq \infty$. Offensichtlich gilt:

$$\mathfrak{f} \in L^p(\Omega)^k \iff f_i \in L^p(\Omega) \quad (i = 1, \ldots, k).$$

Als nächstes sollen die Sobolev-Räume kurz eingeführt werden. Dazu sei die Kenntnis der schwachen Ableitung vorausgesetzt. Im Weiteren sei $\alpha = (\alpha_1, \ldots, \alpha_l) \in (\mathbb{N}_0)^l$ ein Multiindex mit $|\alpha| := \alpha_1 + \ldots + \alpha_l$. Für eine Funktion $f : \Omega \subset \mathbb{R}^l \longrightarrow \mathbb{R}$ sei

$$D^\alpha f := \frac{\partial^{\alpha_1} \ldots \partial^{\alpha_l}}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_l}} f$$

die gewöhnliche partielle Ableitung von f und

$$D^{(\alpha)} f := \left(\frac{\partial^{\alpha_1} \cdots \partial^{\alpha_l}}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_l}} f \right)_w$$

die entsprechende schwache (partielle) Ableitung von f.
Dann definieren wir die Sobolev-Räume als

$$W^{m,p}(\Omega)^k := \left\{ \mathfrak{f} \in L^p(\Omega)^k \mid \text{für } i = 1, \ldots, k \text{ und } 1 \leq |\alpha| \leq m \text{ existiert } D^{(\alpha)} f_i \in L^p(\Omega) \right\}$$

und ihre Norm als

$$\|\mathfrak{f}\|_{W^{m,p}(\Omega)^k} := \left(\sum_{0 \leq |\alpha| \leq m} \left\| D^{(\alpha)} \mathfrak{f} \right\|_{L^p(\Omega)^k} \right) \quad (1 \leq p < \infty)$$

bzw.

$$\|\mathfrak{f}\|_{W^{m,\infty}(\Omega)^k} := \max_{0 \leq |\alpha| \leq m} \left\| D^{(\alpha)} \mathfrak{f} \right\|_{L^\infty(\Omega)^k} .$$

Für $k = 1$ schreiben wir kurz $W^{m,p}(\Omega)$ statt $W^{m,p}(\Omega)^k$ und wenn keine Verwechslungsgefahr vorliegt auch $\|\mathfrak{f}\|_{W^{m,p}}$ für $\|\mathfrak{f}\|_{W^{m,p}(\Omega)^k}$. Man sieht leicht, dass für $1 \leq p \leq \infty$ gilt

$$\mathfrak{f} \in W^{m,p}(\Omega)^k \quad \Longleftrightarrow \quad f_i \in W^{m,p}(\Omega) \quad (i = 1, \ldots, k)$$

und

$$\mathfrak{f} \in W^{m,p}(\Omega)^k \quad \Longrightarrow \quad \|\mathfrak{f}\|_{W^{m,p}(\Omega)^k} < \infty .$$

Falls für $\mathfrak{f} \in W^{m,p}(\Omega)^k$ bzw. $\mathfrak{f} \in L^p(\Omega)^k$ gilt $\mathfrak{f}(\Omega) \subset \Sigma \subset \mathbb{R}^k$ schreiben wir meist kurz $\mathfrak{f} \in W^{m,p}(\Omega; \Sigma)$ bzw. $\mathfrak{f} \in L^p(\Omega; \Sigma)$.
Auch die Sobolev-Räume sind zusammen mit den entsprechenden Normen Banach-Räume.
Damit sind wir mit den Einführungen und Definitionen fertig und es soll als nächstes ein wichtiger Einbettungssatz folgen.

Satz 1.7.1. *Sei $\Omega = B_r(\mathfrak{m}) \subset \mathbb{R}^l$ eine beliebige Kugel. Weiter sei $\mathfrak{f} \in W^{2,\infty}(\Omega)^k$. Dann ist $\mathfrak{f}$ stetig differenzierbar und $\mathfrak{f}$ und jede partielle Ableitung sind beschränkt, d.h.*

$$\mathfrak{f} \in C_B^1(\Omega)^k .$$

Beweis. Seien $\mathfrak{m} \in \mathbb{R}^l$ und $r > 0$ beliebig vorgegeben. Für $\mathfrak{f} = (f_1, \ldots, f_k)^T$ gilt $\mathfrak{f} \in W^{2,\infty}(\Omega)^k \quad \Longleftrightarrow \quad f_i \in W^{2,\infty}(\Omega) \quad (i = 1, \ldots, k)$, wie oben bemerkt wurde. Sei also $g := f_i \quad (1 \leq i \leq k)$, dann ist $\|g\|_{L^\infty(\Omega)} < \infty$ und $\left\| D^{(\alpha)} g \right\|_{L^\infty(\Omega)} < \infty$ für $1 \leq |\alpha| \leq 2$.

D.h. g und alle schwache Ableitungen sind wesentlich beschränkt. Da Ω endliches Maß hat, gilt $g \in L^p(\Omega)$ und $D^{(\alpha)}g \in L^p(\Omega)$ $(1 \leq |\alpha| \leq 2)$ für $1 \leq p \leq \infty$. Und damit ist $g \in W^{2,p}(\Omega)$ für $1 \leq p \leq \infty$. Wir wollen nun [1, Kapitel V, Theorem 5.4] anwenden. Dazu bemerken wir zunächst, dass die Kugel Ω die Kegel-Eigenschaft hat (Englisch: "cone property"). Denn eine Kugel ist konvex, damit auch sternförmig, und hat dann mit [9, Abschnitt 3.2, Lemma 3] die Kegel-Eigenschaft (bei Burenkov heißt sie "cone condition"). Dann liefert dieses Theorem, dass $g \in \mathcal{C}_B^1(\Omega)$ ist. Also sind alle $f_i \in \mathcal{C}_B^1(\Omega)$. Nach [21, Abschnitt 3.1, Reduktionslemma] ist dann $\mathfrak{f} \in \mathcal{C}^1(\Omega)^k$. Da $D^\alpha f_i$ beschränkt ist für $|\alpha| = 0, 1$ und $i = 1, \ldots, k$ ist $D^\alpha \mathfrak{f}$ in der Maximumsnorm beschränkt für $|\alpha| = 0, 1$. Also gilt sogar $\mathfrak{f} \in \mathcal{C}_B^1(\Omega)^k$. □

Da die Sobolev-Räume eigentlich Äquivalenzklassen enthalten muss dieser Satz so interpretiert werden, dass eine Äquivalenzklasse $[\mathfrak{f}] \in W^{2,\infty}(\mathbb{R}^n)^k$ einen Repräsentanten $\mathfrak{f} \in [\mathfrak{f}]$ enthält, der stetig differenzierbar ist. Im weiteren wollen wir diese Äquivalenzklassen mit ihren (eindeutigen) stetig differenzierbaren Repräsentanten identifizieren und wie Funktionen behandeln. Deswegen schreiben wir auch $\mathfrak{f}$ statt $[\mathfrak{f}]$.
Zum Schluss wollen wir noch die Taylor-Entwicklung zweiter Ordnung für Abbildungen aus den Sobolev-Räumen zeigen.

Satz 1.7.2. *Sei $\Omega = B_r(\mathfrak{m}) \subset \mathbb{R}^l$ eine beliebige Kugel, und $\mathfrak{f} \in W^{2,\infty}(\Omega)^k$. Dann gilt für alle $\mathfrak{x} \in \Omega$ und $\xi \in \mathbb{R}^l$ mit $\mathfrak{x} + \xi \in \Omega$*

$$\mathfrak{f}(\mathfrak{x} + \xi) = \mathfrak{f}(\mathfrak{x}) + \mathfrak{J}_\mathfrak{f}(\mathfrak{x}) \cdot \xi + \mathfrak{r}_\mathfrak{x}(\xi), \tag{1.1}$$

wobei $\mathfrak{r}_\mathfrak{x} : \mathbb{R}^l \longrightarrow \mathbb{R}^k$ eine Abbildung ist, die $\|\mathfrak{r}_\mathfrak{x}(\xi)\|_\infty \leq \frac{l^2}{2} \|\mathfrak{f}\|_{W^{2,\infty}(\Omega)} \cdot \|\xi\|_\infty^2$ erfüllt für $\|\xi\|_\infty \to 0$. D.h. $\|\mathfrak{r}_\mathfrak{x}(\xi)\|_\infty = \mathcal{O}\left(\|\xi\|_\infty^2\right)$ für $\|\xi\|_\infty \to 0$.
Dabei ist $\mathfrak{J}_\mathfrak{f}(\mathfrak{x}) \in \mathbb{R}^{k \times l}$ die Jacobi-Matrix von $\mathfrak{f}$.

Beweis. Der Beweis basiert auf einigen mächtigen Resultaten aus [9], welche im Folgenden zusammengestellt werden sollen.
Bereits aus dem Beweis des letzten Satzes wissen wir, dass Ω die Kegel-Bedingung erfüllt (Englisch "cone condition"). Dann liefert [9, Abschnitt 2.3, Theorem 1] für $g \in W^{2,\infty}(\Omega)$ die Existenz einer Folge $(\varphi_n) \in \mathcal{C}^\infty(\Omega) \cap W^{2,\infty}(\Omega)$ $(n \in \mathbb{N})$ für die gilt $\varphi_n \longrightarrow g$ in der $W^{1,\infty}$-Norm und $\|\varphi_n\|_{W^{2,\infty}(\Omega)} \longrightarrow \|g\|_{W^{2,\infty}(\Omega)}$ für $n \to \infty$. Der Schnitt $\mathcal{C}^\infty(\Omega) \cap W^{2,\infty}(\Omega)$ (der eigentlich leer ist) ist so zu verstehen, dass man Äquivalenzklassen und deren stetig Repräsentanten identifiziert. Nach dem vorigen Satz ist $g \in \mathcal{C}_B^1(\Omega)$. Da das wesentliche Supremum gleich ist dem Supremum für eine stetige Funktion haben wir also

$$\|\varphi_n - g\|_\Omega \longrightarrow 0 \quad \text{und} \quad \|\partial_i \varphi_n - \partial_i g\|_\Omega \longrightarrow 0 \qquad \text{für} \quad n \to \infty \, (i = 1, \ldots, l).$$

Daraus folgt dann auch die punktweise Konvergenz.

Sei nun $\mathfrak{x} \in \Omega$ und $\xi \in \mathbb{R}^l$ mit $\mathfrak{x} + \xi \in \Omega$. Dann gilt für alle $n \in \mathbb{N}$ die Taylor-Formel

$$\varphi_n(\mathfrak{x}+\xi) = \varphi_n(\mathfrak{x}) + \nabla\varphi_n(\mathfrak{x})^T \cdot \xi + \frac{1}{2}\sum_{i,j=1}^{l} \xi_i\xi_j \int_0^1 \partial_i\partial_j\varphi_n\left(\mathfrak{x}+t\xi\right)\, dt\,.$$

Damit folgt für $n \in \mathbb{N}$ die Abschätzung

$$\begin{aligned}\left|\varphi_n(\mathfrak{x}+\xi) - \varphi_n(\mathfrak{x}) - \nabla\varphi_n(\mathfrak{x})^T\cdot\xi\right| &\leq \frac{1}{2}\sum_{i,j=1}^{l}\|\xi\|_\infty^2 \int_0^1 \left|\partial_i\partial_j\varphi_n\left(\mathfrak{x}+t\xi\right)\right| dt \\ &\leq \frac{1}{2}\sum_{i,j=1}^{l}\|\xi\|_\infty^2 \int_0^1 \|\varphi_n\|_{W^{2,\infty}(\Omega)}\, dt \\ &= \frac{l^2}{2}\|\xi\|_\infty^2\,\|\varphi_n\|_{W^{2,\infty}(\Omega)}\,.\end{aligned}$$

Sei nun $\tilde{r}_\mathfrak{x}(\xi) := g(\mathfrak{x}+\xi) - g(\mathfrak{x}) - \nabla g(\mathfrak{x})^T\cdot\xi$. Wegen der punktweisen Konvergenz und weil der Betrag stetig ist, gilt

$$\left|\varphi_n(\mathfrak{x}+\xi) - \varphi_n(\mathfrak{x}) - \nabla\varphi_n(\mathfrak{x})^T\cdot\xi\right| \longrightarrow |\tilde{r}_\mathfrak{x}(\xi)| \quad \text{für} \quad n\to\infty\,.$$

Weiter haben wir wegen der Konvergenz der $W^{2,\infty}$-Normen

$$\frac{l^2}{2}\|\xi\|_\infty\,\|\varphi_n\|_{W^{2,\infty}(\Omega)}^2 \longrightarrow \frac{l^2}{2}\|\xi\|_\infty^2\,\|g\|_{W^{2,\infty}(\Omega)} \quad \text{für} \quad n\to\infty\,.$$

Also gilt $|\tilde{r}_\mathfrak{x}(\xi)| \leq \frac{l^2}{2}\|\xi\|_\infty^2\,\|g\|_{W^{2,\infty}(\Omega)}$.
Sei nun $\mathfrak{f} = (f_1,\ldots,f_k)^T \in W^{2,\infty}(\Omega)^k$ und $\mathfrak{r}_\mathfrak{x}(\xi) := \mathfrak{f}(\mathfrak{x}+\xi) - \mathfrak{f}(\mathfrak{x}) - \mathfrak{J}_\mathfrak{f}(\mathfrak{x})\cdot\xi$. Da für alle $i = 1,\ldots,k$ gilt $f_i \in W^{2,\infty}(\Omega)$ und $\mathfrak{J}_\mathfrak{f}(\mathfrak{x}) = (\nabla f_1(\mathfrak{x}),\ldots,\nabla f_k(\mathfrak{x}))^T$ ist, gilt für jede Komponente $\left|(\mathfrak{r}_\mathfrak{x}(\xi))_i\right| \leq \frac{l^2}{2}\|\xi\|_\infty^2\,\|f_i\|_{W^{2,\infty}(\Omega)}$. Also haben wir insgesamt

$$\|\mathfrak{r}_\mathfrak{x}(\xi)\|_\infty \leq \frac{l^2}{2}\|\xi\|_\infty^2\,\|\mathfrak{f}\|_{W^{2,\infty}(\Omega)^k}\,.$$

Trivialerweise gilt natürlich $\mathfrak{f}(\mathfrak{x}+\xi) = \mathfrak{f}(\mathfrak{x}) + \mathfrak{J}_\mathfrak{f}(\mathfrak{x})\cdot\xi + \mathfrak{r}_\mathfrak{x}(\xi)$. □

1.8. Absolutstetige Funktionen

In diesem Abschnitt sollen einige Ergebnisse über absolutstetige Funktionen, die später gebraucht werden, zusammengestellt werden. In ganzen Abschnitt sei $I = [a,b]$ ein reelles Intervall. Doch zunächst die

Definition 1.8.1. Sei $\mathfrak{f} : I \longrightarrow \mathbb{R}^k$ eine vektorwertige Funktion. Dann heißt $\mathfrak{f}$ *absolutstetig*, wenn es zu jedem $\epsilon > 0$ ein $\delta > 0$ existiert, sodass für jedes System von $l \in \mathbb{N}$ paarweise disjunkten Intervallen $(a_i, b_i)_{i=1,\dots,l} \subset [a, b]$ mit der Gesamtlänge

$$\sum_{i=1}^{l} b_i - a_i < \delta$$

gilt

$$\sum_{i=1}^{l} \|\mathfrak{f}(b_i) - \mathfrak{f}(a_i)\|_2 < \epsilon \,.$$

Die Menge der absolutstetigen Funktionen auf dem Intervall I soll mit $\mathrm{AC}(I)^k$ bezeichnet werden, wobei k die Dimension des Wertebereiches ist. Für $k = 1$ schreibt man kurz $\mathrm{AC}(I)$.
Eine Abbildung $\mathfrak{h} : \Omega \longrightarrow \mathbb{R}^k$ mit $\Omega \subset \mathbb{R}^s$ heißt *Lipschitz-stetig* mit einer Lipschitzkonstanten $L > 0$, wenn

$$\|\mathfrak{h}(\mathfrak{x}_1) - \mathfrak{h}(\mathfrak{x}_2)\|_2 \leq L \|\mathfrak{x}_1 - \mathfrak{x}_2\|_2 \qquad \forall \mathfrak{x}_1, \mathfrak{x}_2 \in \Omega \,.$$

Die folgenden Resultate finden sich in [25, Kapitel IX], wobei sie dort nur für den Fall $k = 1$ formuliert sind. Ersetzt man aber den Betrag durch eine Norm, so sind die Beweise genauso gültig. Alternativ kann man sich leicht klar machen, dass eine vektorwertige Funktion genau dann absolutstetig ist, wenn ihre Komponenten absolutstetig sind.

Satz 1.8.2. *Seien $\mathfrak{f}, \mathfrak{g} \in AC(I)^k$ und $r \in \mathbb{R}$. Dann gilt:*

(i) Die Summe $\mathfrak{f} + \mathfrak{g}$ und die Differenz $\mathfrak{f} - \mathfrak{g}$ ist ebenfalls absolutstetig.

(ii) Die Funktion $r\mathfrak{f}$ ist auch absolutstetig.

(iii) Für $f \in AC(I)$ ist $f \cdot \mathfrak{g} \in AC(I)^k$.

Mit (i) und (ii) ist $AC(I)^k$ ein reeller Vektorraum.

Beweis. (i) siehe [25, Kapitel IX, §1, Satz 1].
(ii) Es sei $\sum_{i=1}^{l} \|r\mathfrak{f}(b_i) - r\mathfrak{f}(a_i)\|_2 = |r| \sum_{i=1}^{l} \|\mathfrak{f}(b_i) - \mathfrak{f}(a_i)\|_2 < \epsilon$. Für $r \neq 0$ folgt damit $\sum_{i=1}^{l} \|\mathfrak{f}(b_i) - \mathfrak{f}(a_i)\|_2 < \frac{\epsilon}{|r|}$. Die Existenz eines passenden δ folgt aus der Absolutstetigkeit von $\mathfrak{f}$. Für $r = 0$ ist die Nullfunktion trivialerweise absolutstetig.
(iii) Wende [25, Kapitel IX, §1, Satz 1] auf jede Komponente von $\mathfrak{g}$ an. Dann ist jede Komponente von $f\mathfrak{g}$ absolutstetig. Mit der ∞-Norm folgert man leicht, dass die ganze Funktion absolutstetig ist. Wegen der Äquivalenz der Normen auf dem $\mathbb{R}^k$, ist $f\mathfrak{g}$ auch absolutstetig bezüglich der 2-Norm. □

Das folgende Ergebnis bringt Stetigkeitsbegriffe zusammen.

Satz 1.8.3. *Für eine Funktion* $\mathfrak{f} : I \longrightarrow \mathbb{R}^k$ *gilt*

(i) Ist $\mathfrak{f}$ *Lipschitz-stetig, dann ist* $\mathfrak{f}$ *auch absolutstetig.*

(ii) Ist $\mathfrak{f}$ *absolutstetig, dann ist* $\mathfrak{f}$ *auch stetig.*

Beweis. (ii) Setze in der Definition für absolutstetige Funktionen $l = 1$.
(i) Es sei $\epsilon > 0$ vorgegeben. Dann setzte $\delta := \frac{\epsilon}{L}$, wobei L die Lipschitz-Konstante von $\mathfrak{f}$ ist. Es sei ein beliebiges Intervallsystem wie in obiger Definition gegeben. Dann gilt für $\sum_{i=1}^{l} b_i - a_i = \sum_{i=1}^{l} |b_i - a_i| < \delta = \frac{\epsilon}{L}$

$$\sum_{i=1}^{l} \|\mathfrak{f}(b_i) - \mathfrak{f}(a_i)\|_2 \leq \sum_{i=1}^{l} L\,|b_i - a_i| \leq L\delta = \epsilon\,.$$

Damit ist $\mathfrak{f}$ absolutstetig. □

Satz 1.8.4. *Es seien reelle Zahlen* $a = c_1 < c_2 < \ldots < c_{s+1} = b$ *gegeben mit* $s \in \mathbb{N}$. *Weiter sei* $\mathfrak{f} : [a,b] \longrightarrow \mathbb{R}^k$ *stetig und* $\mathfrak{f}\,|_{[c_i,c_{i+1}]}$ *absolutstetig für* $i = 1, \ldots, s$. *Dann ist* $\mathfrak{f}$ *auf ganz* $[a,b]$ *absolutstetig.*

Beweis. Es reicht diese Aussage, für $s = 2$ zu zeigen. Induktiv kann man dann auf beliebiges s schließen. Sei also $s = 2$ und $\epsilon > 0$ vorgegeben. Für $\mathfrak{f}\,|_{[c_i,c_{i+1}]}$ gibt es ein $\delta_i > 0$, sodass für jedes System von $l_i \in \mathbb{N}$ paarweise disjunkten Intervallen $(a_j, b_j)_{j=1,\ldots,l_i} \subset [c_i, c_{i+1}]$ mit der Gesamtlänge $\sum_{j=1}^{l_i} b_j - a_j < \delta_i$ gilt $\sum_{j=1}^{l_i} \|\mathfrak{f}(b_j) - \mathfrak{f}(a_j)\| < \frac{\epsilon}{2}$ $(i = 1, 2)$. Setze nun $\delta := \min\{\delta_1, \delta_2\}$.
Es sei ein beliebiges System von $l \in \mathbb{N}$ paarweise disjunkten Intervallen $(a_j, b_j)_{j=1,\ldots,l} \subset [a,b]$ mit der Gesamtlänge $\sum_{j=1}^{l} b_j - a_j < \delta$ vorgegeben. Die Indizes seien so gewählt, dass diese Intervalle aufsteigend geordnet sind, d.h. $a_j < a_{j+1}$.
Fall 1: c_2 ist in keinem dieser Intervalle enthalten. Dann gibt es ein $j_0 \in \{0, \ldots, m\}$, sodass $(a_j, b_j) \subset [a, c_2]$ für $j = 1, \ldots, j_0$ und $(a_j, b_j) \subset [c_2, b]$ für $j = j_0 + 1, \ldots, m$ ist. Dann gilt $\sum_{j=1}^{l} \|\mathfrak{f}(b_j) - \mathfrak{f}(a_j)\| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$, da beide Systeme eine Gesamtlänge $< \delta$ haben.
Fall 2: Es gibt ein $j_0 \in \{1, \ldots, m\}$, sodass $c_2 \in (a_{j_0}, b_{j_0})$ ist. Da die Intervalle disjunkt sind, kann c_2 nur in einem Intervall enthalten sein. Wir ersetzen nun das Intervall (a_{j_0}, b_{j_0}) durch die Intervalle $(a_{j_0}, c_2)\,, (c_2, b_{j_0})$. Auf das so entstandene Intervallsystem wenden wir nun Fall 1 an und erhalten mit Dreiecksungleichung

$$\begin{aligned}\sum_{j=1}^{l} \|\mathfrak{f}(b_j) - \mathfrak{f}(a_j)\| &\leq \sum_{\substack{j=1\\ j\neq j_0}}^{l} \|\mathfrak{f}(b_j) - \mathfrak{f}(a_j)\| + \|\mathfrak{f}(c_2) - \mathfrak{f}(a_{j_0})\| + \|\mathfrak{f}(b_{j_0}) - \mathfrak{f}(c_2)\| \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon\,.\end{aligned}$$

□

Und zum Schluss folgt noch ein ganz wichtiges Ergebnis, das oft als Hauptsatz der Differential- und Integralrechnung für die Lebesgue-Integrationstheorie bezeichnet wird.

Theorem 1.8.5. *Sei $\mathfrak{f} \in AC(I)^k$ dann existiert die Ableitung von $\mathfrak{f}$ fast überall auf I. Weiter bezeichne $\mathfrak{f}'$ diese Ableitung, wobei $\mathfrak{f}'|_N := 0$ gesetzt wird auf der Nullmenge N auf der die Ableitung von $\mathfrak{f}$ nicht existiert. Dann gilt weiter*
$\mathfrak{f}' \in L^1\left(I;\mathbb{R}^k\right)$ *und* $\mathfrak{f}(t) = \mathfrak{f}(a) + \int_a^t \mathfrak{f}'(\tau)\,d\tau$ *für alle* $t \in I$.

Beweis. siehe [25, Kapitel IX, §2 Folgerung von Satz 1 und §4 Satz3]. □

Es gilt auch die Umkehrung dieses Theorems:

Satz 1.8.6. *Sei $\mathfrak{f} : I \longrightarrow \mathbb{R}^k$ eine Funktion. Wenn es eine Funktion $\mathfrak{g} \in L^1\left(I;\mathbb{R}^k\right)$ gibt und eine Konstante $c \in \mathbb{R}$ mit*

$$\mathfrak{f}(t) = c + \int_a^t \mathfrak{g}(\tau)\,d\tau \qquad \forall t \in I$$

dann ist $\mathfrak{f}$ absolutstetig und es ist $\mathfrak{f}' = \mathfrak{g}$ fast überall.

Beweis. siehe [25, Kapitel IX, §4 Satz 1+2]. □

Zum Schluss soll noch ein Kriterium bereitgestellt werden um leicht zu entscheiden, wann eine Funktion absolutstetig ist.

Satz 1.8.7. *Sei $\mathfrak{f} : [a,b] \longrightarrow \mathbb{R}^k$ eine stetige Funktion und $s \in \mathbb{N}$.*

(i) $\mathfrak{f}$ sei bis auf endlich viele Stellen c_i $(i = 1,\ldots,s)$ stetig differenzierbar ist. Außerdem sei die Ableitung, dort wo sie existiert beschränkt.

(ii) $\mathfrak{f}$ sei auf ganz $[a,b]$ stetig differenzierbar.

Falls (i) oder (ii) gilt, so ist $\mathfrak{f}$ absolutstetig.

Beweis. (i) Es reicht diesen Satz für $s = 1$ zu zeigen. Induktiv kann man dann auf beliebiges $s \in \mathbb{N}$ schließen. Sei $\mathfrak{g}(t) := \mathfrak{f}'(t)$ für $t \neq c_1$ und $\mathfrak{g}(c_1) := 0$. Dann ist $\mathfrak{g}$ auf $[a,b]$ messbar und beschränkt, da es bis auf $t = c_1$ stetig ist. Weiter sind für beliebige $r,t \in [a,b]$ die Funktionen $t \mapsto \int_r^t \mathfrak{g}(\tau)\,d\tau$ und $r \mapsto \int_r^t \mathfrak{g}(\tau)\,d\tau$ stetig. Nach dem HDI (=Hauptsatz der Differential- und Integralrechnung) gilt für alle $t \in [a,c_1)$

$$\mathfrak{f}(t) = \mathfrak{f}(a) + \int_a^t \mathfrak{g}(\tau)\;d\tau\,.$$

Wegen Stetigkeit beider Seiten gilt ebenfalls $\mathfrak{f}(c_1) = \mathfrak{f}(a) + \int_a^{c_1} \mathfrak{g}(\tau)\;d\tau$. Sei $t \in (c_1,b]$ beliebig. Dann gilt ebenfalls nach dem HDI für alle r mit $c_1 < r < t$

$$\mathfrak{f}(t) - \mathfrak{f}(r) = \int_r^t \mathfrak{g}(\tau)\;d\tau\,.$$

Wieder wegen Stetigkeit beider Seiten gilt $\mathfrak{f}(t) - \mathfrak{f}(c_1) = \int_{c_1}^{t} \mathfrak{g}(\tau)\, d\tau$. Nach vorher ist dies äquivalent mit

$$\mathfrak{f}(t) = \mathfrak{f}(a) + \int_a^{c_1} \mathfrak{g}(\tau)\, d\tau + \int_{c_1}^{t} \mathfrak{g}(\tau)\, d\tau = \mathfrak{f}(a) + \int_a^{t} \mathfrak{g}(\tau)\, d\tau\,.$$

Weil $[a, b]$ endliches Maß hat und $\mathfrak{g}$ beschränkt ist, ist $\mathfrak{g} \in L^1\left([a, b]; \mathbb{R}^k\right)$. Mit Satz 1.8.6 folgt, dass $\mathfrak{f}$ absolutstetig ist.

(ii) Da $\mathfrak{f}'$ stetig ist, ist es auf dem kompakten Intervall $[a, b]$ beschränkt. Mit (i) folgt dann die Behauptung. □

Kapitel 2.

Einschrittverfahren und Kontrollprobleme

In diesem Kapitel sollen einige theoretische Grundlagen für zentrale Inhalte dieser Arbeit gelegt werden. Im ersten Abschnitt werden einige Notationen, die für die ganze Arbeit gelten, festgelegt und Definitionen getätigt. Außerdem werden einige ganz grundlegende Ergebnisse zitiert. Im zweiten Abschnitt werden Einschrittverfahren definiert und einige wichtige Eigenschaften notiert. Darüber hinaus werden einige Begriffe im Umfeld von Einschrittverfahren dargelegt. Schließlich wird noch die wichtige Klasse der Runge-Kutta-Verfahren vorgestellt.

Im letzten Abschnitt werden Kontrollprobleme von einem mehr theoretischen Standpunkt aus eingeführt. Desweiteren werden zwei spezielle Kontrollprobleme vorgestellt. Auch wird auf Existenz und Eindeutigkeit der Lösungen dieser Probleme eingegangen. Schließlich wird noch erläutert wie die Einschrittverfahren des vorigen Abschnittes zur Lösung der Kontrollprobleme verwendet werden können. Ein Kontrollproblem kann man als ein Anfangswertproblem auffassen, das nicht mehr nur vom Zustand $x(t) \in \mathbb{R}^n$ und eventuell von der Zeit $t \in \mathbb{R}$ abhängig ist, sonder auch von Kontrollwerten $\mathfrak{u}(t) \in \mathbb{R}^m$. Man kann es auch als eine Schar von Anfangswertproblemen betrachten, wobei der Parameter der Schar die Kontrollfunktion ist, d.h. für jede Kontrollfunktion haben wir ein eigenes Anfangswertproblem. In dieser Arbeit wird das Kontrollproblem vor allem von diesem Standpunkt aus untersucht.

2.1. Grundlegende Definitionen und Notationen

In diesem Abschnitt sollen grundlegende Definitionen und Schreibweisen kurz und bündig eingeführt werden. Dabei werden wir uns auf wichtige bzw. nicht alltägliche Definitionen und Schreibweisen beschränken. Alle anderen Symbole und Schreibweisen findet man in der Liste der Symbole auf Seite 7.

Vektoren und vektorwertige Funktionen bzw. Abbildungen aus dem $\mathbb{R}^k$ sollen Spaltenvektoren sein und werden häufig als kleine Frakturbuchstaben notiert, z.B. $\mathfrak{v} \in \mathbb{R}^k$ bzw. $\mathfrak{f} : \Omega \longrightarrow \mathbb{R}^k$. Deren Komponenten werden meist als entsprechende lateini-

sche Kleinbuchstaben geschrieben, z.B. $\mathfrak{v} = (v_1, \ldots, v_k)^T$(manchmal auch als indizierte Frakturbuchstaben). Matrizen werden fast immer als große Frakturbuchstaben manchmal auch als große griechische Buchstaben notiert, z.B. $\mathfrak{M} \in \mathbb{R}^{k,l}$ bzw. $\Gamma \in \mathbb{R}^{k,l}$. Die Einträge der Matrizen werden gewöhnlich mit entsprechenden lateinischen Kleinbuchstaben bezeichnet, z.B. $\mathfrak{M} = (m_{i,j})_{\substack{i=1,\ldots,k \\ j=1,\ldots,l}}$. Mengen, die für die in dieser Arbeit vorgestellte Theorie eine besondere Bedeutung haben werden kalligraphisch dargestellt, z.B. $\mathcal{U}$. Andere Mengen werden einfach als lateinische manchmal auch als griechische Großbuchstaben geschrieben, z.B. I bzw. Ω. Reelle und natürliche Zahlen werden meist als lateinische Kleinbuchstaben, falls es sich um Konstanten handelt auch als lateinische Großbuchstaben geschrieben, z.B. n, r bzw. C . Diese Schreibweisen sollen nur als Hilfe dienen. Es gilt: Ausnahmen bestätigen die Regel.

Es werden die Matrix-Normen benutzt, die in Beispiel 1.6.4 vorgestellt wurden. Dies sind alles lub-Normen und haben die Eigenschaften, welche in Abschnitt 1.6 gezeigt wurden. Als Vektornormen werden die gewöhnlichen p-Normen benutzt, die in Beispiel 1.6.3 vorgestellt wurden. Dabei nutzen wir oft, dass die Zeilensummennorm mit der ∞-Norm verträglich ist, und manchmal auch, dass die Spektralnorm mit der 2-Norm verträglich ist (siehe Abschnitt 1.6). Für diese Matrix- und Vektornormen gilt der folgende

Satz 2.1.1. *Seien $\|\cdot\|_a$ und $\|\cdot\|_b$ zwei Normen auf dem endlichdimensionalen Vektorraum X. Dann gibt es zwei Konstanten $0 < M_1 < M_2$ mit*

$$M_1 \|\mathfrak{x}\|_a \leq \|\mathfrak{x}\|_b \leq M_2 \|\mathfrak{x}\|_a \qquad \forall \mathfrak{x} \in X\,.$$

Man sagt dann: Auf einem endlichdimensionalen Vektorraum sind je zwei Normen äquivalent.

Beweis. siehe [21, Unterabschnitt 1.2.III.] □

Da wir nur mit endlichdimensionalen Vektorräumen arbeiten, können wir uns also die jeweils bequemste Norm heraussuchen. Desweiteren benutzen wir auch noch eine ganz andere Norm.

Definition 2.1.2. Sei $U \subset \mathbb{R}^k$ nichtleer. Dann definiert man für $p \in [1, \infty]$

$$\|U\|_p := \sup_{\mathfrak{u} \in U} \|\mathfrak{u}\|_p \in [0, \infty]$$

die p-Norm von U.

Diese Norm ist genau dann endlich, wenn die entsprechende Menge beschränkt ist. Es ist leicht zu sehen, dass diese Mengennormen tatsächlich Normen sind. Dies ist jedoch für diese Arbeit nicht wichtig.

Schließlich nutzen wir noch die Supremumsnorm für Abbildungen. Für eine Abbildung $\mathfrak{f} : V \longrightarrow \mathbb{R}^k$ mit $V \subset \mathbb{R}^l$ ist sie definiert als

$$\|\mathfrak{f}\|_V := \sup_{\mathfrak{v} \in V} \|\mathfrak{f}(\mathfrak{v})\|_\infty \,.$$

Dabei ist zu beachten, dass sie in dieser Arbeit immer bezüglich der Maximumsnorm gebildet wird.
Die Ableitung einer vektorwertigen Funktion $\mathfrak{f} : \mathbb{R} \longrightarrow \mathbb{R}^l$ wird mit $\mathfrak{f}'$ bezeichnet und komponentenweise gebildet. Desweiteren wird der Gradient einer Funktion $g : \mathbb{R}^k \longrightarrow \mathbb{R}$ in $\mathfrak{x} = (x_1, \ldots, x_k)^T$ mit

$$\nabla g(\mathfrak{x}) = \left(\frac{\partial g}{\partial x_1}(\mathfrak{x}), \ldots, \frac{\partial g}{\partial x_k}(\mathfrak{x}) \right)^T$$

und die Jacobi-Matrix einer Abbildung $\mathfrak{h} = \begin{pmatrix} h_1 \\ \vdots \\ h_l \end{pmatrix} : \mathbb{R}^k \longrightarrow \mathbb{R}^l$ mit

$$\mathfrak{J}_{\mathfrak{h}}(\mathfrak{x}) = \begin{pmatrix} \frac{\partial h_1}{\partial x_1}(\mathfrak{x}) & \cdots & \frac{\partial h_1}{\partial x_k}(\mathfrak{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial h_l}{\partial x_1}(\mathfrak{x}) & \cdots & \frac{\partial h_l}{\partial x_k}(\mathfrak{x}) \end{pmatrix}$$

bezeichnet.
Im Bezug auf die Integration gelten die Definitionen, Aussagen und Bezeichnungen aus Abschnitt 1.7. Insbesondere sind alle vorkommenden Integrale Lebesgue-Integrale und das Integral einer vektorwertigen Funktion oder Abbildung wird komponentenweise gebildet.
Nachdem wir alle Bezeichnungen festgelegt haben, wollen wir nun in den Inhalt einsteigen.
Zunächst führen wir eine wichtige Problemklasse ein, die in vielen Bereichen Anwendungen hat. Es handelt sich um die Anfangswertprobleme. Wir formulieren sie hier unter Standardvoraussetzungen. Denn dann sind sie durch die Einschrittverfahren des nächsten Abschnittes lösbar. Später bei den Kontrollproblemen werden wir dann schwächere Voraussetzungen vorgeben.

Problem 2.1.3. Ein *Anfangswertproblem* im $\mathbb{R}^n$ ist gegeben durch eine *Anfangsbedingung* $x_0 \in \mathbb{R}^n$ und eine *gewöhnliche Differentialgleichung* im $\mathbb{R}^n$

$$x'(t) = \mathfrak{f}(t, x(t)) \qquad \forall t \in [t_0, t_f] \tag{2.1}$$
$$x(t_0) = x_0\,, \tag{2.2}$$

wobei $\mathfrak{f} : [t_0, t_f] \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig ist.
Eine Funktion $\hat{x} : [t_0, t_f] \longrightarrow \mathbb{R}^n$ heißt *Lösung des Anfangswertproblems*, wenn sie stetig differenzierbar ist, $\hat{x}(t_0) = x_0$ gilt und die obige Gleichung (2.1) für alle $t \in [t_0, t_f]$ erfüllt.

Um numerische Lösungsverfahren für dieses Problem zu definieren ist die theoretische Existenz und Eindeutigkeit einer Lösung wichtig. Diese Ergebnisse liefert der folgende Satz.

Satz 2.1.4. *Sei* $\mathfrak{f}$ *wie oben definiert und zusätzlich noch Lipschitz-stetig bzgl.* x*, d.h. es gibt eine Konstante* $L > 0$ *mit*

$$\|\mathfrak{f}(t,y) - \mathfrak{f}(t,z)\| \leq L \|y - z\| \qquad \forall\, t \in [t_0, t_f]\ ,\ \forall y, z \in \mathbb{R}^n\ .$$

Dann existiert auf ganz $[t_0, t_f]$ *eine eindeutige Lösung des obigen Anfangswertproblems.*

Beweis. Zum Beweis siehe [29, § 10, Satz VI.] . □

2.2. Einschrittverfahren

Einschrittverfahren sind Differentialgleichungslöser. Es handelt sich dabei um Algorithmen, die mit Hilfe einer Verfahrensfunktion auf einem diskreten Gitter eine Approximation an die exakte Lösung des Anfangsproblems konstruieren. Dieser Abschnitt soll nicht eine vollständige Theorie der Einschrittverfahren darstellen, sonder nur kurz und knapp Begriffe und Ergebnisse wiederholen bzw. bereitstellen. Die Darstellung ist dabei an [16, Kapitel 5] angelehnt

Wir betrachten hier das Anfangswertproblem wie es in Problem 2.1.3 definiert ist. Zunächst unterteilen wir das Zeitintervall $[t_0, t_f]$ mit Hilfe der Schrittweite

$$0 < h < 1 \tag{2.3}$$

in ein äquidistantes Gitter

$$\mathbb{G}_h := \{t_0, t_1, \ldots, t_N = t_f \mid t_i = i \cdot h\}\ .$$

Dabei ist $N + 1$ die Anzahl der Gitterpunkte bzw. N die Anzahl der Teilintervalle (t_i, t_{i+1}). Offenbar gilt $N = \frac{t_f - t_0}{h}$. Für den Rest dieser Arbeit beziehen sich h und $\mathbb{G}_h$ auf diese Definitionen.

Unter einer Gitterfunktion verstehen wir eine Funktion

$$x_h : \mathbb{G}_h \longrightarrow \mathbb{R}^n\ .$$

Nur haben wir alle Begriffe beisammen um ein allgemeines Einschrittverfahren definieren zu können.

Definition 2.2.1. Unter einem (allgemeinen) *Einschrittverfahren* verstehen wir einen Algorithmus, der zu dem Anfangswertproblem von oben mit Hilfe einer *Verfahrensfunktion*

$$\begin{array}{rcl} \Phi : [t_0, t_f] \times \mathbb{R}^n \times (0,1) & \longrightarrow & \mathbb{R}^n \\ (t, x, h) & \longmapsto & \Phi(t, x, h) \end{array}$$

eine Gitterfunktion x_h auf folgenden Weise konstruiert:

$$x_h(t_0) = x_0 \tag{2.4}$$
$$x_h(t_{i+1}) = x_h(t_i) + h \cdot \Phi(t_i, x_h(t_i), h) \tag{2.5}$$

Bemerkung 2.2.2.

(i) Ziel ist natürlich, dass die so konstruierte Gitterfunktion, die exakte Lösung des Anfangsproblems $\hat{x}$ an den Gitterpunkten möglichst gut approximiert.

(ii) Meistens hängt Verfahrensfunktion Φ von der definierenden Funktion $\mathfrak{f}$ des Anfangsproblems ab.

(iii) Aus Gleichung (2.5) folgt $\Phi(t_i, x_h(t_i), h) = \frac{x_h(t_{i+1}) - x_h(t_i)}{h}$. Damit ist Φ der Differenzenquotient der konstruierten Gitterfunktion in zwei benachbarten Gitterpunkten. Die Verfahrensfunktion berechnet also den Differenzenquotienten, der zum nächsten Gitterpunkt führt.

Im Folgenden werden wir die Gitterfunktion x_h die vom Einschrittverfahren konstruiert wird auch Verfahrenslösung nennen.
Eine sehr wichtige Klasse von Einschrittverfahren stellen die Runge-Kutta-Verfahren dar. Diese sollen hier vorgestellt werden.

Beispiel 2.2.3. Sei $s \in \mathbb{N}$. Ein s-stufiges allgemeines Runge-Kutta-Verfahren mit Koeffizienten $c_i,\ b_j \in \mathbb{R},\ a_{i,j} \in \mathbb{R} \quad (i,j = 1,\dots,s)$ definiert die Verfahrensfunktion mittels der Hilfsfunktionen $K_i\ (i = 1,\dots,s)$

$$K_i(t,x,h) = \mathfrak{f}\left(t + c_i \cdot h,\ x + h\sum_{j=1}^{s} a_{i,j} K_j(t,x,h)\right) \quad \text{für } i = 1,\dots,s \tag{2.6}$$
$$\Phi(t,x,h) = \sum_{I=1}^{s} b_i K_i(t,x,h), \tag{2.7}$$

dabei müssen die Koeffizienten den Bedingungen $\sum_{i=1}^{s} b_i = 1$ und $\sum_{j=1}^{s} a_{i,j} = c_i\ (i = 1,\dots,s)$ genügen. Mit Hilfe der Verfahrensfunktion wird dann eine Gitterfunktion durch die Gleichungen (2.4) und (2.5) berechnet.
Die Koeffizienten des Verfahrens fassen wir wie folgt zusammen:

$$\mathfrak{b} = (b_1,\dots,b_s)^T \in \mathbb{R}^s \quad , \quad \mathfrak{c} = (c_1,\dots,c_s)^T \in \mathbb{R}^s$$

sowie

$$\mathfrak{A} = \begin{pmatrix} a_{1,1} & \cdots & a_{1,s} \\ \vdots & \ddots & \vdots \\ a_{s,1} & \cdots & a_{s,s} \end{pmatrix} \in \mathbb{R}^{s\times s}.$$

Falls $\mathfrak{A}$ eine untere Dreiecksmatrix mit verschwindender Hauptdiagonale ist, d.h. $a_{i,j} = 0$ für $i \leq j$, spricht man von einem expliziten Runge-Kutta Verfahren, da zur Berechnung der Verfahrensfunktion $\Phi(t, x, h)$ die Gleichung (2.6) explizit gelöst werden kann. Denn um K_i auszuwerten gehen auf der rechten Seite in diesem Fall nur $K_1, \ldots, K_{i-1}$ ein. Ist $\mathfrak{A}$ keine untere Dreiecksmatrix so spricht man von einem impliziten Runge-Kutta Verfahren, da die erwähnte Gleichung nun implizit gelöst werden muss.

Im Allgemeinen werden die Koeffizienten des Verfahrens im so genannten Butcher-Schema

$\mathfrak{c}$	$\mathfrak{A}$
1	$\mathfrak{b}^T$

notiert.

Im folgenden wollen wir einige Vorarbeit leisten um dann Konvergenz für Einschrittverfahren definieren zu können. Dazu definieren wir folgende Supremumsnorm für Gitterfunktionen x_h

$$\|x_h\|_{\mathbb{G}_h} := \max_{t \in \mathbb{G}_h} \|x_h(t)\|_\infty \ ,$$

die der gewöhnlichen Supremumsnorm auf der Menge $\mathbb{G}_h$ entspricht. Man kann sich leicht davon überzeugen, dass dies tatsächlich eine Norm ist. Denn die Normeigenschaften der Maximumsnorm vererben sich, da das Maximum nur über eine endliche Menge gebildet wird.

Der Einschränkungsoperator ist gegeben durch

$$\Delta_h : \{x : [t_0, t_f] \to \mathbb{R}^n\} \longrightarrow \{x_h : \mathbb{G}_h \to \mathbb{R}^n\} \quad , \quad \Delta_h(x) = x|_{\mathbb{G}_h} \ .$$

Schließlich soll die exakte Lösung des Anfangswertproblems (2.1), (2.2) mit $\hat{x}$ bezeichnet werden. Nun sind wir bereit folgenden Konvergenzbegriff einzuführen:

Definition 2.2.4. Die folgende Gitterfunktion

$$e_h : \mathbb{G}_h \longrightarrow \mathbb{R}^n \quad , \quad e_h := x_h - \Delta_h(\hat{x}) \ ,$$

wird *Gitterfehler* bzw. *globaler Fehler* genannt wird.

Ein Einschrittverfahren, mit dem die diskreten Lösungen x_h für das Anfangswertproblem (2.1),(2.2) konstruiert werden, heißt *konvergent*, wenn

$$\lim_{h \to 0} \|e_h\|_{\mathbb{G}_h} = 0$$

gilt. Es heißt *konvergent von Ordnung* p, wenn zusätzlich

$$\|e_h\|_{\mathbb{G}_h} = \mathcal{O}(h^p) \qquad \text{für } h \to 0$$

gilt.

Im nächsten Beispiel sollen einige explizite Runge-Kutta-Verfahren vorgestellt werden, die auch für die weitere Arbeit von Bedeutung sind.

Beispiel 2.2.5. Zunächst wollen wir zweistufige explizite Runge-Kutta Verfahren betrachten. Das **Heun-Verfahren** ist definiert durch das Butcher-Array

$$\begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline 1 & \frac{1}{2} & \frac{1}{2} \end{array}$$

und hat als Verfahrensfunktion

$$\Phi(t,x,h) = \frac{1}{2}\left[\mathfrak{f}(t,x) + \mathfrak{f}(t+h,\, x+h\mathfrak{f}(t,x))\right] .$$

Es hat Konvergenzordnung 2 .
Ebenfalls von Konvergenzordnung 2 ist das **verbesserte Euler-Verfahren**. Es wird definiert durch das Butcher-Array

$$\begin{array}{c|cc} 0 & & \\ \frac{1}{2} & \frac{1}{2} & \\ \hline 1 & 0 & 1 \end{array}$$

und seine Verfahrensfunktion ist

$$\Phi(t,x,h) = \mathfrak{f}\left(t+\frac{h}{2},\, x+\frac{h}{2}\mathfrak{f}(t,x)\right) .$$

Als nächstes betrachten wir ein dreistufiges Verfahren, das die Konvergenzordnung 3 besitzt. Das **Heun-3-Verfahren** wird definiert durch das Butcher-Array:

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{3} & \frac{1}{3} & & \\ \frac{2}{3} & 0 & \frac{2}{3} & \\ \hline 1 & \frac{1}{4} & 0 & \frac{3}{4} \end{array}$$

Und zum Schluss soll hier noch das **klassische Runge-Kutta-Verfahren** vorgestellt werden. Es hat 4 Stufen und auch Konvergenzordnung 4. Das definierende Butcher-Schema sieht so aus:

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ 1 & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline 1 & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

Die Verfahrensfunktionen für diese letzten beiden Beispiele sind schon zu komplex um sie hier anzugeben.
Die Aussagen über die Konvergenzordnung, gelten nur für genügend glatte rechte Seiten $\mathfrak{f}$.

Als nächstes wollen wir die Konsistenz von Einschrittverfahren einführen. Zusammen mit der Stabilität hat man dann ein Werkzeug um die Konvergenz eines Verfahrens zu beweisen. Da die Konsistenz eine lokale Eigenschaft ist, müssen wir das Verhalten eines Verfahrens zu jedem Zeitpunkt und für jeden Zustand untersuchen.
Es sei y die Lösung des Anfangswertproblems

$$\begin{aligned} x\left(t^*\right) &= x^* \\ x'(t) &= \mathfrak{f}\left(t, x(t)\right) \end{aligned}$$

für beliebige

$$t^* \in [t_0, t_f] \quad \text{und} \quad x^* \in \mathbb{R}^n \,.$$

Es handelt sich dabei um das gleiche Anfangswertproblem wie in Problem 2.1.3 nur mit anderer Anfangsbedingung (die hier ihren Namen nicht mehr verdient hat).

Definition 2.2.6. Es sei Φ die Verfahrensfunktion eines Einschrittverfahrens. Der *lokale Diskretisierungsfehler* dieses Verfahrens in $(t^*,\, x^*)$ ist definiert durch

$$\epsilon_h : [t_0, t_f] \times \mathbb{R}^n \longrightarrow \mathbb{R}^n \quad , \quad \epsilon_h\left(t^*,\, x^*\right) := \frac{y\left(t^* + h\right) - x^*}{h} - \Phi\left(t^*, x^*,\, h\right) \,.$$

Das Einschrittverfahren heißt *konsistent*, wenn

$$\lim_{h \to 0} \left\| \epsilon_h\left(t^*,\, x^*\right) \right\| = 0 \qquad \forall t^* \in [t_0, t_f] \;,\, \forall x^* \in \mathbb{R}^n \,.$$

Es heißt *konsistent von Ordnung* p , wenn

$$\epsilon_h\left(t^*,\, x^*\right) = \mathcal{O}\left(h^p\right) \qquad \forall t^* \in [t_0, t_f] \;,\, \forall x^* \in \mathbb{R}^n \,.$$

Bemerkung 2.2.7. Man kann den lokalen Diskretisierungsfehler verschieden interpretieren. Hier sind zwei gängige Möglichkeiten:

(i) Es ist $\frac{y(t^*+h)-x^*}{h}$ der rechtsseitige Differenzenquotient der exakten Lösung in t^* und $t^* + h$. Wegen $\Phi\left(t^*,\, x^*,\, h\right) = \frac{x_h(t^*+h)-x^*}{h}$ ist dies die der Differenzenquotient der approximierten Lösung in t^* und t^*+h. Es ist klar, dass je besser diese beiden Werte übereinstimmen, desto besser ist die approximierte Lösung $x_h\left(t^* + h\right)$ im ersten Gitterpunkt.

(ii) Wir wollen den lokalen Diskretisierungsfehler ein wenig anders darstellen

$$\begin{aligned}\epsilon_h(t^*, x^*) &= \frac{y(t^*+h) - x^*}{h} - \Phi(t^*,\, x^*,\, h) = \frac{y(t^*+h) - (x^* + h\Phi(t^*, x^*, h))}{h} \\ &= \frac{y(t^*+h) - x_h(t^*+h)}{h}\,.\end{aligned}$$

Deswegen nennt man $\epsilon_h(t^*, x^*)$ auch lokalen Fehler per Einheit.

(iii) Es reicht Konsistenz in einem Streifen um die exakte Lösung zu fordern.

Konsistenz ist nur eine lokale Eigenschaft und reicht nicht aus um Konvergenz sicherzustellen. Aber zusammen mit Stabilität kann Konvergenz sichergestellt werden. Stabilität hat zu tun mit dem Verhalten der Verfahrenslösung auf gestörte Eingabedaten und gestörte Differenzenquotienten.

Für eine beliebige Gitterfunktion $y_h : \mathbb{G}_h \longrightarrow \mathbb{R}^n$ definieren wird den *Defekt* von y_h als eine Gitterfunktion $\delta_h : \mathbb{G}_h \longrightarrow \mathbb{R}^n$ durch

$$\begin{aligned}\delta_h(t_0) &= y_h(t_0) - x_0 \\ \delta_h(t_i) &= \frac{y_h(t_i) - y_h(t_{i-1})}{h} - \Phi(t_{i-1},\, y_h(t_{i-1}),\, h)\,.\end{aligned}$$

Für eine diskrete Lösung x_h eines Einschrittverfahrens mit Verfahrensfunktion Φ, ist der Defekt identisch Null. Die Gitternorm des Defektes für eine Gitterfunktion ist ein Maß dafür, wie stark die Gitterfunktion vom Anfangswert und von dem Differenzenquotienten des Verfahrens Φ abweicht. Damit kommen wir zur Definition der Stabilität

Definition 2.2.8. Sei $(x_h)_h$ eine Sequenz von Lösungen eines Einschrittverfahrens mit $h \to 0$. Weiter sei $(y_h)_h$ ein Sequenz von Gitterfunktionen $y_h : \mathbb{G}_h \longrightarrow \mathbb{R}^n$ mit Defekt δ_h. Ein Einschrittverfahren heißt *stabil*, wenn es zwei Konstanten $S, R > 0$ (unabhängig von h) gibt, sodass für fast alle h gilt:

$$\|\delta_h\|_{\mathbb{G}_h} < R \qquad \Longrightarrow \qquad \|x_h - y_h\|_{\mathbb{G}_h} < S \cdot \|\delta_h\|_{\mathbb{G}_h}\,.$$

Dabei heißt R Stabilitätsschwelle und S Stabilitätsschranke.

Vereinfacht sagt diese Definition aus: Ist eine Gitterfunktion in t_0 nahe an dem Anfangswert und ihr Differenzenquotient nahe an der Verfahrensfunktion, dann ist die Gitterfunktion nahe an der Verfahrenslösung. Die Stabilität eines Einschrittverfahrens ist wichtig, da ein Rechner nicht mit exakter Arithmetik arbeitet. Unter Umständen ist auch der Anfangswert nicht exakt.

Das folgende Resultat macht es einfacher Stabilität zu Überprüfen.

Proposition 2.2.9. *Die Verfahrensfunktion Φ des Einschrittverfahrens sei Lipschitz-stetig bzgl. x für ausreichend kleine Schrittweiten, d.h. $\exists\, 0 < h_0 < 1, L > 0$ sodass gilt*

$$\|\Phi(t,\, x,\, h) - \Phi(t,\, y,\, h)\| \leq L\,\|x - y\| \qquad \textit{für alle}\, x, y \in \mathbb{R}^n\,,\; t \in [t_0, t_f]\; \textit{und}\, h \in (0,\, h_0]\,.$$

Dann ist das Einschrittverfahren stabil.

Beweis. siehe [16, Proposition 5.21]. □

Für Runge-Kutta-Verfahren ist die Verfahrensfunktion dann Lipschitz-stetig, wenn die rechte Seite $\mathfrak{f}$ des Anfangswertproblems Lipschitz-stetig ist.
Nun können wir das Hauptresultat dieses Abschnitts formulieren:

Theorem 2.2.10 (Konvergenzsatz)**.** *Ist ein Einschrittverfahren für ein Anfangswertproblem konsistent und stabil so ist es auch konvergent.*
Dabei entspricht die Konsistenzordnung der Konvergenzordnung des Einschrittverfahrens.

Beweis. siehe [16, Theorem 5.20]. □

2.3. Kontrollprobleme

In diesem Abschnitt sollen Kontrollprobleme vorgestellt werden und zwar nur in einem Umfang und in der Art wie es für die weitere Arbeit erforderlich ist.

Problem 2.3.1 (Kontrollproblem)**.** Sei $I = [0, t_f]$ $(t_f > 0)$, $\Omega \subset \mathbb{R}^n$ und $U \subset \mathbb{R}^m$ kompakt und nichtleer. Weiter sei $\mathfrak{f} : \Omega \times U \longrightarrow \mathbb{R}^n$ stetig und $x_0 \in \Omega$.
Das Kontrollproblem (KP) besteht nun darin, zu einer vorgegebenen Steuer- oder Kontrollfunktion $\mathfrak{u} \in L^1(I; U)$ eine dazugehörige absolutstetige Funktion $\hat{x} : I \longrightarrow \mathbb{R}^n$ zu finden, die folgende Gleichungen erfüllt:

$$x'(t) = \mathfrak{f}(x(t), \mathfrak{u}(t)) \qquad \text{für fast alle}\, t \in I \tag{2.8}$$
$$x(0) = x_0\,. \tag{2.9}$$

Dann heißt $\hat{x}$ Lösung des Kontrollproblems zur Kontrollfunktion $\mathfrak{u}$ oder kurz Lösung des Kontrollproblems und x_0 heißt Anfangswert . Oft wird $\mathfrak{u}$ auch kurz Kontrolle oder Steuerung genannt. Weiter heißt I Zeitintervall, Ω Zustandsraum und U Kontrollbereich.

Bemerkung 2.3.2.

(i) Das hier vorgestellte Kontrollproblem ist autonom, d.h. ist $\hat{x}$ die Lösung von (KP) zu einer Steuerung $\mathfrak{u}$ auf einem Zeitintervall $I = [t_0, t_f]$, dann ist $\hat{x}(\cdot + t_0)$ Lösung von (KP) zur Kontrolle $\mathfrak{u}(\cdot + t_0)$ auf dem Zeitintervall $I = [0, t_f - t_0]$. Deshalb bedeutet es keine Einschränkung $t_0 = 0$ zu setzten.

(ii) Da $\mathfrak{u}(t) \in U$ für alle $t \in I$ gilt und U als kompakt vorausgesetzt wurde, folgt automatisch, dass $\mathfrak{u} \in L^\infty(I; \mathbb{R}^m)$ ist.

Oft ist es interessant und wichtig zu wissen wohin es möglich ist eine Lösung von (KP) überall hinzusteuern, etwa wenn ein Kontrollproblem die Steuerung eines Roboters beschreibt. Dazu ist folgende Definition nützlich.

Definition 2.3.3. Es sei das vorige Kontrollproblem 2.3.1 gegeben mit allen dort definierten Variablen. Unter der erreichbaren Menge für (KP) versteht man folgende Menge

$$\mathcal{R}(t, x_0) = \{y \in \mathbb{R}^n \mid \exists \mathfrak{u} \text{ Steuerung und eine dazugehörige Lösung } \hat{x} \text{ von (KP)} \\ \text{mit } \hat{x}(0) = x_0 \text{ und } \hat{x}(t) = y\}$$

dabei ist x_0 die Anfangsbedingung des Kontrollproblems und $t \in I$.

Eine besonders wichtige Klasse stellen die linearen Kontrollprobleme dar, die hier kurz vorgestellt werden sollen.

Problem 2.3.4 (Lineares Kontrollproblem)**.** Sei $\mathfrak{A} \in \mathbb{R}^{n\times n}$ und $\mathfrak{B} \in \mathbb{R}^{n\times m}$. Wenn in Problem 2.3.1 gilt

$$\mathfrak{f}(x, \mathfrak{u}) = \mathfrak{A}x + \mathfrak{B}\mathfrak{u} \quad \text{und} \quad \Omega = \mathbb{R}^n$$

so spricht man von einem linearen Kontrollproblem (mit konstanten Koeffizienten). Es soll hier kurz mit (LKP) bezeichnet werden.

Dieses lineare Kontrollproblem wird Hauptgegenstand dieser Arbeit sein. Für diese Problemklasse existieren starke Existenz und Eindeutigkeitsaussagen, die hier vorgestellt werden.

Satz 2.3.5. *Für das linearen Kontrollproblem 2.3.4 sei eine Steuerung $\mathfrak{u}$ gegeben. Dann existiert eine eindeutige Lösung $\hat{x}(\cdot)$ welche folgende Gestalt hat*

$$\hat{x}(t) = e^{\mathfrak{A}t}x_0 + \int_0^t e^{\mathfrak{A}(t-\tau)}\mathfrak{B}\mathfrak{u}(\tau)\, d\tau \qquad t \in I.$$

Außerdem ist $\hat{x}$ gleichmäßig beschränkt, genauer gilt

$$\|\hat{x}\|_I \leq e^{\|\mathfrak{A}\|_Z t_f} \left(\|x_0\|_\infty + t_f \|\mathfrak{B}\|_Z \|U\|_\infty\right)$$

für alle möglichen Lösungen zu allen möglichen Steuerungen. Dabei ist $\|\hat{x}\|_I = \sup_{t\in I} \|\hat{x}(t)\|_\infty$ und $\|U\|_\infty = \sup_{\mathfrak{u}\in U} \|\mathfrak{u}\|_\infty$.

Beweis. Gegeben sei eine Steuerung $\mathfrak{u}$ für (LKP). Zunächst soll gezeigt werde, dass $\hat{x}(t)$ absolutstetig ist. Die Funktion $\mathfrak{f}(t) := e^{\mathfrak{A}t}x_0$ ist stetig differenzierbar auf I und deswegen nach Satz 1.8.7(ii) auch absolutstetig auf I. Sei $\mathfrak{g}(t) := \int_0^t e^{\mathfrak{A}(t-\tau)}\mathfrak{B}\mathfrak{u}(\tau)\, d\tau$.

Dann ist der Integrand von $\mathfrak{g}$ integrierbar auf I und damit ist $\mathfrak{g}$ absolutstetig auf I nach Satz 1.8.6. Insgesamt ist wegen Satz 1.8.2 $\hat{x}(t) = \mathfrak{f}(t) + \mathfrak{g}(t)$ absolutstetig auf I. Diese Darstellung der Lösung nennt man auch Variation-der-Konstanten-Formel.

Als nächstes rechnen wir mit Satz 1.8.6 nach, dass $\hat{x}$ eine Lösung ist. Es gilt für fast alle $t \in I$

$$\hat{x}'(t) = \mathfrak{A}e^{\mathfrak{A}t}x_0 + \left(e^{\mathfrak{A}t} \cdot \int_0^t e^{-\mathfrak{A}\tau}\mathfrak{B}\mathfrak{u}(\tau)\,d\tau\right)'$$

und mit der Produktregel erhalten wir damit

$$\begin{aligned}
\hat{x}'(t) &= \mathfrak{A}e^{\mathfrak{A}t}x_0 + \mathfrak{A}e^{\mathfrak{A}t} \cdot \int_0^t e^{-\mathfrak{A}\tau}\mathfrak{B}\mathfrak{u}(\tau)\,d\tau + e^{\mathfrak{A}t} \cdot e^{-\mathfrak{A}t}\mathfrak{B}\mathfrak{u}(t) \\
&= \mathfrak{A}\left(e^{\mathfrak{A}t}x_0 + e^{\mathfrak{A}t} \cdot \int_0^t e^{-\mathfrak{A}\tau}\mathfrak{B}\mathfrak{u}(\tau)\,d\tau\right) + \mathfrak{B}\mathfrak{u}(t) \\
&= \mathfrak{A}\hat{x}(t) + \mathfrak{B}\mathfrak{u}(t)
\end{aligned}$$

für fast alle $t \in I$.

Zum Schluss zeigen wir noch, dass die Lösungen $\hat{x}$ in allen Steuerungen gleichmäßig beschränkt sind auf I. Für $k \in \mathbb{N}$ gilt mit den Rechenregeln für Matrixnormen (siehe Abschnitt 1.6 auf Seite 23)

$$\left\|\sum_{i=0}^{k} \frac{(\mathfrak{A}s)^i}{i!}\right\|_Z \leq \sum_{i=0}^{k} \left\|\frac{(\mathfrak{A}s)^i}{i!}\right\|_Z \leq \sum_{i=0}^{k} \frac{\|\mathfrak{A}s\|_Z^i}{i!} = \sum_{i=0}^{k} \frac{\|\mathfrak{A}\|_Z^i\,|s|^i}{i!}\,.$$

Geht man dann zum Limes über so gilt $\left\|e^{\mathfrak{A}s}\right\|_Z \leq e^{\|\mathfrak{A}\|_Z|s|}$ $(s \in \mathbb{R})$. Zusammen mit der Variation-der-Konstanten-Formel haben wir für $t \in I = [0, t_f]$

$$\begin{aligned}
\|\hat{x}(t)\|_\infty &\leq \left\|e^{\mathfrak{A}t}x_0\right\|_\infty + \left\|\int_0^t e^{\mathfrak{A}(t-\tau)}\mathfrak{B}\mathfrak{u}(\tau)\,d\tau\right\|_\infty \\
&\leq e^{\|\mathfrak{A}\|_Z|t|}\,\|x_0\|_\infty + \int_0^t \left\|e^{\mathfrak{A}(t-\tau)}\right\|_Z \|\mathfrak{B}\|_Z\,\|\mathfrak{u}(\tau)\|_\infty\,d\tau \\
&\leq e^{\|\mathfrak{A}\|_Z t_f}\,\|x_0\|_\infty + \int_0^t e^{\|\mathfrak{A}\|_Z t_f}\,\|\mathfrak{B}\|_Z\,\|U\|_\infty\,d\tau \\
&\leq e^{\|\mathfrak{A}\|_Z t_f}\,\|x_0\|_\infty + t_f e^{\|\mathfrak{A}\|_Z t_f}\,\|\mathfrak{B}\|_Z\,\|U\|_\infty \\
&= e^{\|\mathfrak{A}\|_Z t_f}\,(\|x_0\|_\infty + t_f\,\|\mathfrak{B}\|_Z\,\|U\|_\infty)\,,
\end{aligned}$$

wobei wir die Rechenregeln für Normen intensiv genutzt haben und die Tatsache, dass die Exponentialfunktion strenge monoton steigend ist. □

An dieser Stelle wollen wir einen Problemtyp einführen, der sich nur auf den ersten Blick von dem vorigen linearen Kontrollproblem unterscheidet.

Problem 2.3.6 (Lineare Differentialinklusion)**.** Es seien $\mathfrak{A}, \mathfrak{B}, U, x_0$ und I wie in Problem 2.3.1 bzw. 2.3.4. Eine absolutstetige Funktion $\hat{x} : I \longrightarrow \mathbb{R}^n$ heißt Lösung der linearen Differentialinklusion (LDI), wenn sie

$$\begin{aligned} x'(t) &\in \mathfrak{A}x(t) + \mathfrak{B}U \\ x(0) &= x_0 \end{aligned}$$

erfüllt für fast alle $t \in I$. Weiter heißt x_0 Anfangsbedingung von (LDI).

Auch für diesen Problemtyp wollen wir die erreichbare Menge definieren.

Definition 2.3.7. Es sei (LDI) aus Problem 2.3.6 gegeben. Die *erreichbare Menge* für dieses Problem ist definiert als

$$\mathcal{R}(t, x_0) = \{y \in \mathbb{R}^n \mid \exists \text{ Lösung } \hat{x} \text{ von (LDI) mit } \hat{x}(0) = x_0 \text{ und } \hat{x}(t) = y\} ,$$

wobei x_0 die Anfangsbedingung von (LDI) ist.

Der folgende Satz zeigt, dass das lineare Kontrollproblem äquivalent zu der linearen Differentialinklusion ist.

Satz 2.3.8. *Es seien $\mathfrak{A}, \mathfrak{B}, U, x_0$ und I wie in Problem 2.3.1 bzw. 2.3.4. Ist $\hat{x}$ eine Lösung von (LKP) in Problem 2.3.4 zu einer zulässigen Steuerung $\mathfrak{u}$, so löst $\hat{x}$ auch (LDI) in Problem 2.3.6.*

Ist umgekehrt $\hat{x}$ eine Lösung von (LDI), so existiert eine Funktion $\mathfrak{u} \in L^1([0, t_f]; U)$, sodass $\hat{x}$ eine Lösung von (LKP) zur Steuerung $\mathfrak{u}$ ist.

Es stimmen also die erreichbaren Mengen dieser beider Problemtypen überein.

Beweis. siehe [14, Theorem 27]. □

Aufgrund dieses Satzes können wir uns darauf beschränken die erreichbare Menge eines linearen Kontrollproblems zu untersuchen bzw. zu approximieren und erhalten gleichzeitig Ergebnisse für die erreichbare Menge der zugeordneten linearen Differentialinklusion.

Schließlich sollen noch einige Eigenschaften der erreichbaren Menge aufgelistet werden.

Satz 2.3.9. *Sei $t \in [0, t_f]$. Die erreichbare Menge $\mathcal{R}(t, x_0)$ von (LDI) bzw. (LKP) ist kompakt, konvex und nichtleer.*

Beweis. Die Anfangsmenge $\{x_0\}$ ist kompakt, konvex und nichtleer. Die Matrixfunktion $\mathfrak{A} : [0, t_f] \longrightarrow \mathbb{R}^{n \times n}$, $t \mapsto \mathfrak{A}$ ist stetig und integrierbar. Die mengenwertige Abbildung $F : [0, t_f] \Longrightarrow \mathbb{R}^n$, $t \mapsto \mathfrak{B}U$ ist messbar nach Proposition 1.5.2 und hat nichtleere und abgeschlossenen Bilder, da $\mathfrak{B}U$ kompakt ist (Bilder kompakter Mengen unter stetigen Abbildungen sind kompakt). Dann liefert [4, Satz 2.1.4] die Behauptung. □

Im weiteren soll noch kurz eine weitere Klasse von Kontrollproblemen vorgestellt werden.

Problem 2.3.10. Sei $\mathfrak{a} \in W^{2,\infty}(\mathbb{R}^n;\mathbb{R}^n)$ ein Vektorfeld. Unter einem nichtlinearen Kontrollproblem (NKP) verstehen wir hier ein Kontrollproblem aus Problem 2.3.1 mit

$$\mathfrak{f}(x,\mathfrak{u}) = \mathfrak{a}(x) + \mathfrak{B}\mathfrak{u} \quad \text{und} \quad \Omega = B_R(x_0)\ ,$$

wobei $R := 2n \cdot \|x_0\|_\infty + 2n \cdot \max\{t_f, 1\} \cdot (\|\mathfrak{a}\|_{W^{2,\infty}} + \|\mathfrak{B}\|_Z \|U\|_\infty)$ ist. Nochmal ausgeschrieben:
$\hat{x} : I \longrightarrow \mathbb{R}^n$ heißt Lösung von (NKP) zu einer Steuerfunktion $\mathfrak{u} \in L^1(I;U)$, wenn $\hat{x}$ absolutstetig ist und die Gleichungen

$$\begin{aligned} x'(t) &= \mathfrak{a}(x(t)) + \mathfrak{B}\mathfrak{u}(t) \\ x(0) &= x_0 \end{aligned} \tag{2.10}$$

erfüllt, wobei alle Bezeichnungen wie in Problem 2.3.1 sind.

Bemerkung 2.3.11.

(i) In Problem 2.3.1 wurde $\mathfrak{f}(x,\mathfrak{u})$ als stetig vorausgesetzt. Es ist aber $\mathfrak{a}$ stetig differenzierbar auf jeder beliebigen Kugel wie Satz 1.7.1 zeigt. Da die Elemente der L^p- Räume wie auch der Sobolev-Räume Äquivalenzklassen sind, ist es genauer zu sagen, dass die Äquivalenzklasse $\mathfrak{a}$ einen stetig differenzierbaren Repräsentanten enthält. Im Folgenden sollen diese Äquivalenzklassen mit ihren (eindeutigen) stetig differenzierbaren Repräsentanten identifiziert werden.

(ii) Leider gilt nicht, dass $\mathfrak{a}$ auf ganz $\mathbb{R}^n$ stetig differenzierbar ist. Deswegen wurde das obige nichtlineare Kontrollproblem nur auf einer Kugel $\Omega = B_R(x_0)$ um die Anfangsbedingung definiert. Die Definition des Radius R sorgt aber dafür, dass jede Lösung innerhalb von Ω bleibt (vgl. den nächsten Satz).

(iii) Eigentlich ist das obige Problem nur ein spezielles nichtlineares Kontrollproblem. Aber da in dieser Arbeit keine andere Klasse von nichtlinearen Kontrollproblemen betrachtet wird, besteht keine Verwechslungsgefahr.

Auch für diesen Problemtyp ist die Definition 2.3.3 der erreichbaren Menge gültig.
Wir wollen nun auch für (NKP) Existenz und Eindeutigkeit zeigen.

Satz 2.3.12. *Es sei ein nichtlineares Kontrollproblem (NKP) und ein Steuerungsfunktion* $\mathfrak{u}$ *aus Problem 2.3.10 geben.*
Dann existiert eine eindeutige Lösung $\hat{x}$ *von (NKP). Außerdem ist die Lösung beschränkt, genauer gilt*

$$\|\hat{x}\|_\infty \leq \|x_0\|_\infty + t_0 \cdot (\|\mathfrak{a}\|_{W^{2,\infty}} + \|\mathfrak{B}\|_Z \cdot \|U\|_\infty)\,.$$

Und diese Lösung bleibt immer im Zustandsraum $\Omega = B_R(x_0)$.

Beweis. Es sei $\mathfrak{u}$ eine fest vorgegebene Steuerung. Angenommen es gibt dazu ein Lösung $\hat{x}$. Da $\hat{x}$ absolutstetig ist, gilt nach Theorem 1.8.5 $\hat{x}(t) = x_0 + \int_0^t \hat{x}'(\tau)\, d\tau$. Für $\hat{x}'$ können wir nun die rechte Seite von (2.10) einsetzen und bilden die Maximumsnorm

$$\|\hat{x}(t)\|_\infty = \left\| x_0 + \int_0^t \mathfrak{a}\left(\hat{x}(\tau)\right) + \mathfrak{B}\mathfrak{u}(\tau)\, d\tau \right\|_\infty .$$

Nun wenden wir einige Rechenregeln für Normen an und erhalten

$$\|\hat{x}(t)\|_\infty \leq \|x_0\|_\infty + \int_0^t \|\mathfrak{a}\left(\hat{x}(\tau)\right)\|_\infty + \|\mathfrak{B}\|_Z \|\mathfrak{u}(\tau)\|_\infty \, d\tau \,.$$

Da U als kompakt vorausgesetzt wurde existiert $\|U\|_\infty = \sup_{u\in U} \|\mathfrak{u}\|_\infty$ und damit ist $\|\mathfrak{u}(t)\|_\infty \leq \|U\|_\infty$ überall auf $I = [0, t_f]$. Weiter gilt $\|\mathfrak{a}(x)\|_\infty \leq \|\mathfrak{a}\|_{W^{2,\infty}}$ fast überall. Damit haben wir

$$\begin{aligned} \|\hat{x}(t)\|_\infty &\leq \|x_0\|_\infty + \int_0^t \|\mathfrak{a}\|_{W^{2,\infty}} + \|\mathfrak{B}\|_Z \|U\|_\infty \, d\tau \\ &= \|x_0\|_\infty + t\left(\|\mathfrak{a}\|_{W^{2,\infty}} + \|\mathfrak{B}\|_Z \|U\|_\infty\right) \\ &\leq \|x_0\|_\infty + t_f\left(\|\mathfrak{a}\|_{W^{2,\infty}} + \|\mathfrak{B}\|_Z \|U\|_\infty\right) . \end{aligned}$$

Also ist $\|\hat{x}(t)\|_2 \leq \sqrt{n}\|x_0\|_\infty + \sqrt{n}t_f\left(\|\mathfrak{a}\|_{W^{2,\infty}} + \|\mathfrak{B}\|_Z \|U\|_\infty\right) \quad (t \in I)$, denn es gilt $\|v\|_2 \leq \sqrt{n}\|v\|_\infty \quad \forall v \in \mathbb{R}^n$. Damit ist $\|\hat{x}(t)\|_2 \leq \frac{R}{2} \quad (t \in I)$, wobei R wie in Problem 2.3.10 definiert ist. Deshalb verlässt $\hat{x}(\cdot)$ den Zustandsraum $\Omega = B_R(x_0)$ nicht sondern bleibt sogar in $B_{\frac{R}{2}}(x_0)$. Wir können nun $\mathfrak{a}\,|_\Omega$ vermöge Satz 1.7.1 als stetig differenzierbar ansehen.

Jetzt wollen wir die Existenz und Eindeutigkeit einer Lösung zeigen. Dazu wollen wir zu unserer vorgegebenen Kontrollfunktion dieses nichtlineare Kontrollproblem als Anfangswertproblem auffassen mit der rechten Seite $\mathfrak{f}(t,x) := \mathfrak{a}(x) + \mathfrak{B}\mathfrak{u}(t)$. Dann erfüllt $\mathfrak{f}$ die so genannte Carathéodory-Bedingung, d.h. $\mathfrak{f}$ ist bzgl. x stetig und bzgl. t integrierbar. Nun definieren wir noch $\beta(t) := \|\mathfrak{a}\|_{W^{2,\infty}} + \|\mathfrak{B}\|_Z \|\mathfrak{u}(t)\|_\infty$ und damit gilt

$$\|\mathfrak{f}(t,x)\|_\infty \leq \beta(t) \quad \forall t \in I\,,\, \forall x \in \Omega\,,$$

denn $\mathfrak{a}$ ist stetig auf Ω. Es ist dann $\beta(\cdot)$ integrierbar auf I, da ja I kompakt ist.

Jetzt definieren wir noch eine weitere Funktion $\alpha(t) := n\|\mathfrak{a}\|_{W^{2,\infty}}$. Damit gilt

$$\left\| \frac{\partial \mathfrak{f}}{\partial x}(t,x) \right\|_Z = \|\mathfrak{J}_\mathfrak{a}(x)\|_Z = \max_{i=1,\dots,n} \sum_{j=1}^n \left| \frac{\partial \mathfrak{a}_i}{\partial x_j}(x) \right| \leq \alpha(t) \qquad \forall t \in I\,,\, \forall x \in \Omega\,,$$

denn $\mathfrak{a}$ ist stetig differenzierbar auf Ω. Natürlich ist $\alpha(\cdot)$ ebenfalls integrierbar über I. Mit [27, Proposition C.3.4] angewendet für $\alpha(\cdot)$ erhalten wir die lokale Lipschitz-Bedingung für $\mathfrak{f}$ bzgl x. Dann liefert [27, Theorem 54] mit $\beta(\cdot)$ die Existenz einer eindeutigen Lösung $\hat{x}(\cdot)$ auf einem Intervall $J = [0, t_1] \subset I = [0, t_f]$. Da wir garantieren können, dass $\hat{x}(\cdot)$ das Kompaktum $B_{\frac{R}{2}}(x_0) \subset \Omega$ nicht verlässt, liefert schließlich [27, Proposition C.3.6], dass $J = I$ ist. Also existiert die eindeutige Lösung $\hat{x}(\cdot)$ auf ganz I. □

Damit haben wir alle benötigten Resultate und Definitionen bereitgestellt. Es soll nun noch eine Bemerkung über numerische Lösungsmöglichkeiten des Kontrollproblems folgen.

Bemerkung 2.3.13. Im obigen Beweis haben wir (NKP) aus Problem 2.3.10 für eine feste Kontrollfunktion einfach als Anfangswertproblem aufgefasst um Existenz und Eindeutigkeit zu zeigen. Für Anfangswertprobleme stehen aber Einschrittverfahren zur numerischen Lösung bereit, wie wir im vorigen Abschnitt gesehen haben.

Dies wollen wir nun auf das allgemeine Kontrollproblem 2.3.1 anwenden. Dazu sei eine Steuerfunktion $u(\cdot)$ fest vorgegeben. Gesucht wird nun die Lösung $\hat{x}$ zu dieser vorgegebenen Steuerung. Wir definieren $\tilde{f}(t, x(t)) := f(x(t), u(t))$, wobei f die rechte Seite von Gleichung (2.8) des Kontrollproblems ist. Dann ist $\hat{x}(\cdot)$ die Lösung des Anfangswertproblems

$$\begin{aligned} x'(t) &= \tilde{f}(t, x(t)) \qquad \forall' t \in I \\ x(0) &= x_0 \,, \end{aligned}$$

wobei sich natürlich an der Anfangsbedingung nichts ändert. Analytisch kann man, zumindest für unsere beiden Spezialfälle, auch für dieses Anfangswertproblem Existenz und Eindeutigkeit einer Lösung zeigen, da $\tilde{f}$ die so genannte Carathéodory-Bedingung erfüllt.

Auf dieses Anfangswertproblem können nun die Einschrittverfahren aus dem vorigen Abschnitt angewandt werden. Allerdings erfüllt nun $\tilde{f}$ nicht die Glattheitsvoraussetzungen um Konsistenz und Konvergenz der Verfahren garantieren zu können, da $\tilde{f}$ bezüglich (t, x) nicht einmal stetig ist. Insbesondere gehen die entsprechenden Konvergenz und Konsistenzordnungen verloren, die bei entsprechend glatter rechter Seite vorliegen.

Genau dieser Verlust der Konvergenzordnung ist die Thematik dieser Arbeit. Im folgenden soll gezeigt werden wie man Runge-Kutta-Verfahren verändern muss um wenigstens für das lineare Kontrollproblem die klassische Konvergenzordnung zu erhalten.

Kapitel 3.

Lineare Kontrollprobleme

In diesem Kapitel wollen wir numerische Verfahren zu Lösung des linearen Kontrollproblems (LKP), welches in Abschnitt 2.3 eingeführt wurde, vorstellen und untersuchen. Dabei werden wir dieses Problem aus der Sicht eines Anfangswertproblems betrachten, d.h. zu einer vorgegebenen Steuerfunktion $\mathfrak{u} \in L^1(I; U)$ suchen wir eine Lösung des Anfangswertproblems

$$x'(t) = \mathfrak{A}x(t) + \mathfrak{B}\mathfrak{u}(t) \qquad \forall' t \in I \tag{3.1}$$

$$x(0) = x_0\,, \tag{3.2}$$

wobei $U \subset \mathbb{R}^m$ kompakt, $I = [0, t_f]$ ein kompaktes Intervall, $\mathfrak{A} \in \mathbb{R}^{n\times n}$, $\mathfrak{B} \in \mathbb{R}^{n\times m}$ und $x_0 \in \mathbb{R}^n$ die Anfangsbedingung ist. Die Schwierigkeit bei der numerischen Lösung besteht in der schwachen Voraussetzung an die Steuerfunktion $\mathfrak{u}$. Gewöhnliche Verfahren zu Lösung von Anfangswertproblemen haben damit Schwierigkeiten.

Trotz dieses punktwertigen Zugangs ist dennoch das Ziel dieses Kapitels mengenwertige Verfahren zur Approximation der erreichbaren Menge von (LKP) zu entwickeln

Im ersten Abschnitt werden wir die Theorie von Ferretti, die er in [13] vorgestellt hat untersuchen. Dabei geht es darum, für Runge-Kutta-Verfahren eine Auswahlstrategie für die diskreten Kontrollvektoren zu bestimmen, sodass die gewöhnliche Konvergenzordnung der Verfahren erhalten bleibt.

Im zweiten Abschnitt soll diese Theorie ergänzt werden. Wir werden dabei andere Verfahren kennenlernen, die natürlicher zu den im ersten Abschnitt vorgestellten Methoden passen. Das wird auch zu einem tieferen Verständnis der Theorie von Ferretti beitragen.

Im letzten Abschnitt rückt die erreichbare Menge des linearen Kontrollproblems in den Mittelpunkt. Und die vorher punktwertigen Verfahren sollen dann zu mengenwertigen Verfahren ausgebaut werden um die erreichbare Menge zu bestimmen.

3.1. Die Theorie von Ferretti

In diesem Abschnitt wollen wir Runge-Kutta-Verfahren zu Lösung von (LKP) studieren, welche wir auch kurz mit RK-Verfahren bezeichnen. Wir betrachten ausschließ-

lich explizite Schemata, auch wenn dies nicht extra erwähnt wird.

Zunächst wollen wir ausgehen von einem gewöhnlichen Runge-Kutta-Verfahren für das obige Kontrollproblem. Dies dient dazu eine Gitterfunktion $x_h : \mathbb{G}_h \longrightarrow \mathbb{R}^n$ zu konstruieren mit Schrittweite $h = \frac{t_f}{N}$ und $N \in \mathbb{N}$ der Anzahl Iterationen des Verfahrens. Angepasst für das lineare Kontrollproblem (3.1),(3.2) schreiben wir hier nochmal ein allgemeines explizites $p-$stufiges Runge-Kutta-Schema (siehe Definition 2.2.1 und Beispiel 2.2.3) ausführlich als

$$x_h(0) = x_0 \tag{3.3}$$

$$x_h((k+1)h) = x_h(kh) + h\sum_{i=1}^{p} b_i K_i \qquad (k = 0, \ldots, N-1) \tag{3.4}$$

mit

$$K_1 = \mathfrak{A}x_h(kh) + \mathfrak{B}\mathfrak{u}_{1,k} \tag{3.5}$$

$$K_i = \mathfrak{A}\left(x_h(kh) + h\sum_{j=1}^{i-1} a_{i,j}K_j\right) + \mathfrak{B}\mathfrak{u}_{i,k}\,. \tag{3.6}$$

Dabei werden die Koeffizienten $a_{i,j}$, b_i und c_i durch das Butcher-Array definiert (siehe Beispiel 2.2.3) . Für ein gewöhnliches Runge-Kutta-Verfahren sind dabei die Koeffizienten $\mathfrak{u}_{i,k} = \mathfrak{u}(kh + c_i h) \in \mathbb{R}^m$ Diskretisierungen der Kontrollfunktion. Wie schon erwähnt, kann aber so keine Konvergenz erreicht werden (abgesehen davon macht eine punktwertige Auswertung einer L^1-Funktion keinen Sinn). Wir wollen sie hier in diesem Schema deswegen zunächst nur als unbestimmte Variablen betrachten. Sie sollen im Folgenden als diskrete Kontroll- bzw. Steuervektoren bezeichnet werden. Für $k = 0$ schreiben wir auch einfach $\mathfrak{u}_i$ statt $\mathfrak{u}_{i,0}$. Die Bezeichnung "ein Runge-Kutta-Verfahren der Ordnung s" soll bedeuten, das das zu Grunde liegende Runge-Kutta-Verfahren unter genügenden Glattheitsvoraussetzungen an die rechte Seite eines Anfangswertproblems von Ordnung s konvergiert.

3.1.1. Entwicklung der diskreten und analytischen Lösung

Wir wissen bereits aus Abschnitt 2.3, dass eine eindeutige Lösung von (LKP) stets existiert. Deswegen wollen wir eine Entwicklung dieser Lösung erarbeiten. Dabei betrachten wir (LKP) lokal auf dem Intervall $[0, h]$ und geben auch eine lokale Anfangsbedingung $\bar{x}_0$ vor, die in einem Kompaktum variieren darf. Ziel ist es dann diese lokalen Ergebnisse für eine globale Lösung zu nutzen.

Satz 3.1.1. *Es sei $G \subset \mathbb{R}^n$ ein Kompaktum und die lokale Anfangsbedingung $\bar{x}_0 \in G$. Zu einer vorgegebenen Steuerfunktion $\mathfrak{u} \in L^1([0, h]; U)$ genügt die exakte Lösung $\hat{x}(\cdot)$ des lokalen*

(LKP) auf $[0,h]$ *folgender Entwicklung*

$$\hat{x}(h) = \sum_{i=0}^{p} \frac{h^i}{i!}\mathfrak{A}^i \cdot \bar{x}_0 + \mathfrak{B}\int_0^h \mathfrak{u}(s_1)\, ds_1 + \mathfrak{A}\mathfrak{B}\int_0^h\int_0^{s_1} \mathfrak{u}(s_2)\, ds_2\, ds_1 + \ldots + \mathfrak{A}^{p-1}\mathfrak{B}\int_0^h \cdots \int_0^{s_{p-1}} \mathfrak{u}(s_p)\, ds_p \ldots ds_1 + \mathfrak{r}(h)\,,$$

wobei $h \in I$ *ist.*

Für den Restterm gilt

$$\|\mathfrak{r}(h)\|_\infty \leq \left[\left\|\mathfrak{A}^{p+1}\right\|_Z \cdot e^{\|\mathfrak{A}\|_Z t_f}\left(\|G\|_\infty + t_f \|\mathfrak{B}\|_Z \|U\|_\infty\right) + \|\mathfrak{A}^p\mathfrak{B}\|_Z \cdot \|U\|_\infty\right] \cdot h^{p+1}.$$

Wir haben also $\mathfrak{r}(h) = \mathcal{O}(h^{p+1})$ *mit einer Konstante, die unabhängig von der Steuerfunktion, von der Schrittweite* h *und auch unabhängig in der Anfangsbedingung ist, die allerdings auf* G *beschränkt ist.*

Beweis. Nach Definition einer Lösung des Kontrollproblems aus Problem 2.3.1 ist $\hat{x}$ absolutstetig. Wie schon erwähnt existiert eine eindeutige Lösung. Also gilt

$$\hat{x}(s) = \hat{x}(0) + \int_0^s \hat{x}'(t)\, dt \qquad \forall s \in I\,. \tag{3.7}$$

Damit zeigen wir zunächst durch Induktion nach $p \in \mathbb{N}_0$ die Behauptung

$$\hat{x}(h) = \sum_{i=0}^{p} \frac{h^i}{i!}\mathfrak{A}^i \cdot \bar{x}_0 + \sum_{i=0}^{p} \mathfrak{A}^i\mathfrak{B}\int_0^h \cdots \int_0^{s_i} \mathfrak{u}(s_{i+1})\, ds_{i+1} \ldots ds_1 + \mathfrak{A}^{p+1}\int_0^h \cdots \int_0^{s_p} \hat{x}(s_{p+1})\, ds_{p+1} \ldots ds_1\,. \tag{3.8}$$

Der Induktionsanfang für $p = 0$ ist zusammen mit (3.1) schon die obige Integraldarstellung der Lösung (3.7) mit $s = h$.

Induktionsschluss: $p \rightsquigarrow (p+1)$

Nach Induktionsvoraussetzung gilt für p die obige Gleichung (3.8). Darin ersetzen

wir den Integranden $\hat{x}\left(s_{p+1}\right)$ mit Gleichung (3.7) und erhalten

$$\begin{aligned}
\hat{x}\left(h\right) &= \sum_{i=0}^{p} \frac{h^i}{i!}\mathfrak{A}^i \cdot \bar{x}_0 + \sum_{i=0}^{p} \mathfrak{A}^i\mathfrak{B} \int_0^h \cdots \int_0^{s_i} \mathfrak{u}\left(s_{i+1}\right)\, ds_{i+1} \ldots ds_1 \\
&\quad + \mathfrak{A}^{p+1} \int_0^h \cdots \int_0^{s_p} \hat{x}\left(0\right) + \int_0^{s_{p+1}} \hat{x}'\left(t\right)\, dt\, ds_{p+1} \ldots ds_1 \\
&\overset{(*)}{=} \sum_{i=0}^{p+1} \frac{h^i}{i!}\mathfrak{A}^i \cdot \bar{x}_0 + \sum_{i=0}^{p} \mathfrak{A}^i\mathfrak{B} \int_0^h \cdots \int_0^{s_i} \mathfrak{u}\left(s_{i+1}\right)\, ds_{i+1} \ldots ds_1 \\
&\quad + \mathfrak{A}^{p+1} \int_0^h \cdots \int_0^{s_{p+1}} \mathfrak{A}\hat{x}\left(t\right) + \mathfrak{B}\mathfrak{u}(t)\, dt\, ds_{p+1} \ldots ds_1 \\
&= \sum_{i=0}^{p+1} \frac{h^i}{i!}\mathfrak{A}^i \cdot \bar{x}_0 + \sum_{i=0}^{p+1} \mathfrak{A}^i\mathfrak{B} \int_0^h \cdots \int_0^{s_i} \mathfrak{u}\left(s_{i+1}\right)\, ds_{i+1} \ldots ds_1 \\
&\quad + \mathfrak{A}^{p+2} \int_0^h \cdots \int_0^{s_{p+2}} \hat{x}\left(s_{p+2}\right)\, ds_{p+2} \ldots ds_1 \,,
\end{aligned}$$

wobei wir in $(*)$ die lokale Anfangsbedingung $\hat{x}(0) = \bar{x}_0$ und für $\hat{x}'$ die rechte Seite der Differentialgleichung von (LKP) eingesetzt haben. Damit ist obige Behauptung bewiesen.

Da der Steuerbereich U als kompakt vorausgesetzt wurde, ist $\|U\|_\infty < \infty$. Damit gilt

$$\begin{aligned}
&\left\| \mathfrak{A}^p\mathfrak{B} \int_0^h \cdots \int_0^{s_p} \mathfrak{u}\left(s_{p+1}\right)\, ds_{p+1} \ldots ds_1 \right\|_\infty \\
&\qquad \leq \|\mathfrak{A}^p\mathfrak{B}\|_Z \int_0^h \cdots \int_0^{s_p} \|\mathfrak{u}\left(s_{p+1}\right)\|_\infty\, ds_{p+1} \ldots ds_1 \\
&\qquad \leq \|\mathfrak{A}^p\mathfrak{B}\|_Z \cdot \|U\|_\infty \cdot h^{p+1} \,.
\end{aligned}$$

Weiter gilt für die Lösung $\hat{x}$ nach Satz 2.3.5

$$\|\hat{x}\|_{[0,h]} \leq e^{\|\mathfrak{A}\|_Z t_f} \left(\|\bar{x}_0\|_\infty + t_f \|\mathfrak{B}\|_Z \|U\|_\infty\right) \,,$$

denn $h \leq t_f$, da $h \in I = [0, t_f]$. Also ist sie in allen möglichen Steuerfunktionen gleichmäßig beschränkt.

Da die lokale Anfangsbedingung $\bar{x}_0$ auf ein Kompaktum $G \subset \mathbb{R}^n$ beschränkt ist, gilt $\|\bar{x}_0\|_\infty \leq \|G\|_\infty$. Damit kann man die Lösung abschätzen durch

$$\|\hat{x}\|_{[0,h]} \leq e^{\|\mathfrak{A}\|_Z t_f} \left(\|G\|_\infty + t_f \|\mathfrak{B}\|_Z \|U\|_\infty\right) \,,$$

und sie ist auch gleichmäßig beschränkt in den Anfangsbedingungen aus G.

Damit haben wir

$$\begin{aligned}
&\left\| \mathfrak{A}^{p+1} \int_0^h \cdots \int_0^{s_p} \hat{x}\left(s_{p+1}\right)\, ds_{p+1} \ldots ds_1 \right\|_\infty \\
&\qquad \leq \left\|\mathfrak{A}^{p+1}\right\|_Z \cdot e^{\|\mathfrak{A}\|_Z t_f} \left(\|G\|_\infty + t_f \|\mathfrak{B}\|_Z \|U\|_\infty\right) \cdot h^{p+1} \,,
\end{aligned}$$

ähnlich wie vorher.
Mit der Dreiecksungleichung für Normen erhalten wir dann

$$\left\| \mathfrak{A}^p \mathfrak{B} \int_0^h \cdots \int_0^{s_p} \mathfrak{u}\left(s_{p+1}\right)\, ds_{p+1} \ldots ds_1 + \mathfrak{A}^{p+1} \int_0^h \cdots \int_0^{s_p} \hat{x}\left(s_{p+1}\right)\, ds_{p+1} \ldots ds_1 \right\|_\infty$$
$$\leq \left[\left\| \mathfrak{A}^{p+1} \right\|_Z \cdot e^{\|\mathfrak{A}\|_Z t_f} \left(\|G\|_\infty + t_f \|\mathfrak{B}\|_Z \|U\|_\infty \right) + \|\mathfrak{A}^p \mathfrak{B}\|_Z \cdot \|U\|_\infty \right] \cdot h^{p+1}.$$

Also definieren wir die Restfunktion $\mathfrak{r} : I \to \mathbb{R}^n$ durch

$$\mathfrak{r}\left(h\right) := \mathfrak{A}^p \mathfrak{B} \int_0^h \cdots \int_0^{s_p} \mathfrak{u}\left(s_{p+1}\right)\, ds_{p+1} \ldots ds_1 + \mathfrak{A}^{p+1} \int_0^h \cdots \int_0^{s_p} \hat{x}\left(s_{p+1}\right)\, ds_{p+1} \ldots ds_1 \,.$$

Offensichtlich gilt für diese Funktion die behauptete Abschätzung. □

Vor dem nächsten Ergebnis benötigen wir noch folgendes

Lemma 3.1.2. *Es sei ein Polynom $p : \mathbb{R} \longrightarrow \mathbb{R}$, $p(y) = \sum_{i=0}^m \lambda_i y^i$ gegeben mit $\lambda_i \in \mathbb{R}$. Es gebe ein $\hat{y} > 0$, sodass für $0 < y < \hat{y}$ gilt:*

$$|p(y)| \leq C \cdot y^{m+1} \,,$$

mit einer Konstanten $C > 0$.
Dann ist p die Nullfunktion, d.h. $\lambda_i = 0$ $(i = 0, \ldots, m)$.

Beweis. Angenommen p ist nicht das Nullpolynom.
Dann gibt es minimales $0 \leq i_0 \leq m$ mit $\lambda_{i_0} \neq 0$. Wir nehmen o.E. an $\lambda_{i_0} > 0$. Falls nicht ersetzte im Folgenden p durch $-p$.
Wir definieren dann

$$\tilde{p}(y) := \frac{p(y)}{y^{i_0}} = \sum_{i=0}^{m-i_0} \lambda_{i_0+i} y^i \,.$$

Dann ist $|p(y)| \leq C \cdot y^{m+1}$ äquivalent mit $|\tilde{p}(y)| \leq C \cdot y^{m+1-i_0}$ für $0 < y < \hat{y}$. Da Polynomfunktionen stetig sind, gilt $\lim_{y \to 0} \tilde{p}(y) = \lambda_{i_0}$. Trivialerweise gilt auch $\lim_{y \to 0} C \cdot y^{m+1-i_0} = 0$, denn $m + 1 - i_0 \geq 1$. Also gibt es ein $y_1 \in (0, \hat{y})$ mit $p\left(y_1\right) > \frac{3}{4}\lambda_{i_0}$ und $C \cdot y_1^{m+1-i_0} < \frac{1}{4}\lambda_{i_0}$. Dies ist ein Widerspruch zur Voraussetzung. □

Als nächstes wollen wir auch für die diskrete Lösung x_h eine Entwicklung erarbeiten. Für ein konkretes Verfahren erhält man diese Entwicklung einfach durch ausrechnen d.h. in dem man die Gleichungen (3.3)-(3.6) konkret nach $x_h(h)$ auflöst. Da wie hier nur explizite Schemata betrachten ist dies möglich.

Satz 3.1.3. *Die diskrete Lösung x_h von dem lokalen (LKP) mit Anfangsbedingung $\bar{x}_0$ sei mit einem p-stufigen Runge-Kutta-Schema der Ordnung p konstruiert. Dann hat sie die Darstellung*

$$x_h(h) = \sum_{i=0}^{p} \frac{h^i}{i!}\mathfrak{A}^i \cdot \bar{x}_0 + \sum_{i=1}^{p} \frac{h^i}{i!}\mathfrak{A}^{i-1}\mathfrak{B} \sum_{j=1}^{p} \gamma_{i,j}\mathfrak{u}_j \,.$$

Dabei erfüllen die Koeffizienten $\gamma_{i,j}$ die Gleichungen $\sum_{j=1}^{p} \gamma_{i,j} = 1$ $(i = 1, \ldots, p)$ und es ist $\gamma_{i,j} = 0$ für $p - i < j - 1$. Sie ergeben sich aus den Koeffizienten im Butcher-Array des Verfahrens.
Falls $b_p > 0$ und $a_{i+1,i} > 0$ $(i = 1, \ldots, p-1)$ ist, gilt zusätzlich $\gamma_{i,p-i+1} > 0$ für $i = 1, \ldots, p$.

Beweis. Wir zeigen den Beweis in mehreren Schritten. Dies ist ein technisch schwieriger Beweis wegen der vielen Indizes.
Schritt 1 : Behauptung 1: Die K_i in (3.5)+(3.6) haben im ersten Zeitschritt für $i = 1, \ldots, p$ die Gestalt

$$K_i = \sum_{j=0}^{i-1} \mathfrak{A}^{j+1} h^j \alpha_{i,j} \cdot \bar{x}_0 + \sum_{j=1}^{i-1} h^j \mathfrak{A}^j \mathfrak{B} \sum_{k=1}^{i-j} \beta_{i,j,k}\mathfrak{u}_k + \mathfrak{B}\mathfrak{u}_i$$

mit Koeffizienten $\alpha_{i,j}, \beta_{i,j,k} \in \mathbb{R}$. Dabei setzten wir $\beta_{1,0,1} := 1$, wobei dies der Koeffizient von $\mathfrak{u}_1$ in K_1 sein soll.
Für $i = 1$ ist Behauptung 1 richtig mit $\alpha_{1,0} = 1$ und $\beta_{1,0,1} = 1$. Die Induktionsvoraussetzung (I.V.) gelte für alle K_j mit $j \leq i$.
Induktionsschritt $i \rightsquigarrow i + 1$

$$\begin{aligned}
K_{i+1} &= \mathfrak{A}\left(\bar{x}_0 + h\sum_{j=1}^{i} a_{i+1,j}K_j\right) + \mathfrak{B}\mathfrak{u}_{i+1} \\
&= \mathfrak{A}\bar{x}_0 + \sum_{j=1}^{i} \mathfrak{A}h a_{i+1,j}K_j + \mathfrak{B}\mathfrak{u}_{i+1} \\
&\overset{\text{I.V.}}{=} \mathfrak{A}\bar{x}_0 + \sum_{j=1}^{i} a_{i+1,j} \sum_{k=0}^{j-1} \mathfrak{A}^{k+2} h^{k+1} \alpha_{j,k} \cdot \bar{x}_0 + \\
&\quad + \sum_{j=1}^{i} a_{i+1,j} \left(\sum_{l=1}^{j-1} \mathfrak{A}^{l+1}\mathfrak{B} h^{l+1} \sum_{k=1}^{j-l} \beta_{j,l,k}\mathfrak{u}_k + \mathfrak{A}\mathfrak{B}h\mathfrak{u}_j \right) + \mathfrak{B}\mathfrak{u}_{i+1} \\
&= \sum_{k=0}^{i} \mathfrak{A}^{k+1} h^k \alpha_{i+1,k} \cdot \bar{x}_0 + \sum_{l=1}^{i} h^l \mathfrak{A}^l \mathfrak{B} \sum_{j=1}^{i+1-l} \beta_{i+1,l,j}\mathfrak{u}_j + \mathfrak{B}\mathfrak{u}_{i+1} \,,
\end{aligned}$$

wobei wir im letzten Schritt die Summationsreihenfolge vertauscht haben. Die zweite Summe steckt in den neuen Koeffizienten

$$\alpha_{i+1,0} := 1 \quad , \quad \alpha_{i+1,k} := a_{i+1,k} \sum_{j=k}^{i} \alpha_{j,k-1} \quad (k = 1, \ldots, i+1)$$

und

$$\beta_{i+1,1,j} := a_{i+1,j} \quad (j = 1, \ldots, i),$$
$$\beta_{i+1,l,j} := \sum_{k=l-1+j}^{i} a_{i+1,k} \cdot \beta_{k,l-1,j} \qquad \begin{pmatrix} l = 2, \ldots, i+1 \\ j = 1, \ldots, i+1-l \end{pmatrix} . \quad (3.9)$$

Schritt 2 : Mit dieser Formel für die K_i rechnen wir $x_h(h)$ aus mittels (3.4).

$$\begin{aligned} x_h(h) &= \bar{x}_0 + h \sum_{i=1}^{p} b_i K_i \\ &= \bar{x}_0 + \sum_{i=1}^{p} b_i \sum_{j=0}^{i-1} \mathfrak{A}^{j+1} h^{j+1} \alpha_{i,j} \cdot \bar{x}_0 \\ &\quad + \sum_{i=1}^{p} b_i \sum_{j=1}^{i-1} h^{j+1} \mathfrak{A}^j \mathfrak{B} \sum_{k=1}^{i-j} \beta_{i,j,k} \mathfrak{u}_k + \sum_{i=1}^{p} b_i \mathfrak{B} \mathfrak{u}_i \\ &= \sum_{j=0}^{p} \mathfrak{A}^j h^j \tilde{\alpha}_j \cdot \bar{x}_0 + \sum_{j=1}^{p} h^j \mathfrak{A}^{j-1} \mathfrak{B} \sum_{k=0}^{p-j} \delta_{j,k} \mathfrak{u}_k \,, \end{aligned}$$

wobei wieder wie vorher die Summationsreihenfolge im letzten Schritt vertauscht wurde. Die zweite Summe ist wieder in den neuen Koeffizienten

$$\tilde{\alpha}_0 = 1 \quad , \quad \tilde{\alpha}_j = \sum_{i=j}^{p} b_i \alpha_{i,j-1} \qquad (j = 1, \ldots, p)$$

und

$$\delta_{1,k} = b_k \quad (k = 1, \ldots, p), \quad \delta_{j,k} = \sum_{l=k+j-1}^{p} b_l \beta_{l,j-1,k} \qquad \begin{pmatrix} j = & 2, \ldots, p \\ k = & 1, \ldots, p-j+1 \end{pmatrix} \quad (3.10)$$

versteckt. Wir wollen die letzte Gleichung noch ein wenig anders schreiben. Dazu setzten wir $d_j = \sum_{k=1}^{p-j+1} \delta_{j,k}$ $(j = 1, \ldots, p)$ und

$$\gamma_{j,k} := \frac{1}{d_j} \delta_{j,k} \begin{pmatrix} j = & 1, \ldots, p \\ k = & 1, \ldots, p-j+1 \end{pmatrix} , \; \gamma_{j,k} := 0 \begin{pmatrix} j = & 2, \ldots, p \\ k = & p-j+2, \ldots, p \end{pmatrix} . \quad (3.11)$$

Es ist dann $\sum_{k=1}^{p-j+1} \gamma_{j,k} = 1$ $(j = 1, \ldots, p)$ und für die diskrete Lösung haben wir die Darstellung

$$x_h(h) = \sum_{j=0}^{p} \mathfrak{A}^j h^j \tilde{\alpha}_j \cdot \bar{x}_0 + \sum_{j=1}^{p} h^j \mathfrak{A}^{j-1} \mathfrak{B} d_j \sum_{k=0}^{p-1} \gamma_{j,k} \mathfrak{u}_k \,.$$

Schritt 3 : Jetzt müssen wir noch zeigen, dass $\tilde{\alpha}_j = \frac{1}{j!}$ und $d_j = \frac{1}{j!}$. Dazu bemerken wir, dass die Darstellung, die wir zeigen wollen, auch gelten muss für die gewöhnliche Form des RK-Verfahrens, in der die $\mathfrak{u}_k = \mathfrak{u}(c_k h)$ Diskretisierungen der Steuerfunktion sind. Denn dieser Satz betrifft ja nur die direkte algebraische Darstellung von $x_h(h)$ gewonnen aus den Gleichungen (3.3)-(3.6). Im folgenden seien die $\mathfrak{u}_k$ solche Diskretisierungen, außerdem sei die Dimension der Kontrollen $m = 1$, denn dieser Darstellung gilt ja für beliebige $m \in \mathbb{N}$. Wir geben eine konstante Steuerfunktion $\mathfrak{u}(t) = \bar{u}$ für $t \in [0, h]$ mit $\bar{u} \in \mathbb{R}$ vor. Dies ist eine beliebig glatte Steuerung. Damit ist die rechte Seite des Anfangswertproblems (3.1)-(3.2) beliebig oft differenzierbar. Dann wissen wir, dass das RK-Verfahren Konsistenzordnung p hat. Deshalb muss zusammen mit der Entwicklung für die analytische Lösung aus Satz 3.1.1 gelten

$$\hat{x}(h) - x_h(h) = \sum_{i=0}^{p} h^i \mathfrak{A}^i \bar{x}_0 \left(\frac{1}{i!} - \tilde{\alpha}_i \right) + \sum_{i=1}^{p} \mathfrak{A}^{i-1} \mathfrak{B} \left(\int_0^h \cdots \int_0^{s_{i-1}} \mathfrak{u}(s_i)\, ds_i \ldots ds_1 - h^i d_i \sum_{k=1}^{p} \gamma_{i,k} \mathfrak{u}_k \right) = \mathcal{O}\left(h^{p+1}\right) . \quad (3.12)$$

Jetzt setzen wir $\bar{u} = 0$. Dann fällt die ganze zweite Summe weg. Und wir erhalten

$$\left\| \sum_{i=0}^{p} h^i \mathfrak{A}^i \bar{x}_0 \left(\frac{1}{i!} - \tilde{\alpha}_i \right) \right\|_\infty \leq C \cdot h^{p+1} ,$$

wobei $h \in \left(0, \hat{h}\right)$, $\hat{h} > 0$ eine maximale Schrittweite und $C > 0$ eine Konstante ist . Im Allgemeinen ist $\mathfrak{A}^i \bar{x}_0 \neq 0_{\mathbb{R}^n}$ (diese Darstellung muss ja für alle Matrizen $\mathfrak{A}$ gelten). Jetzt wenden wir auf ein beliebige Komponente dieses vektorwertigen Polynoms das Lemma 3.1.2 an und erhalten $\tilde{\alpha}_i = \frac{1}{i!}$ $(i = 0, \ldots, p)$.
Nun sei $\bar{u} = 1$. Aus (3.12) erhalten wir mit den neuen Werten für $\tilde{\alpha}_i$

$$\left\| \sum_{i=1}^{p} \mathfrak{A}^{i-1} \mathfrak{B} \left(\frac{h^i}{i!} \cdot 1 - h^i d_i \cdot 1 \right) \right\|_\infty \leq C \cdot h^{p+1} \qquad h \in (0, \hat{h}).$$

Auch hier ist $\mathfrak{A}^{i-1}\mathfrak{B}$ im Allgemeinen nicht der Nullvektor ($\mathfrak{B}$ ist hier ein Vektor, da $m = 1$ gesetzt ist). Mit dem gleichen Argument wie vorher erhalten wir $d_i = \frac{1}{i!}$.
Schritt 4 : Jetzt wolle wir noch den Zusatz zeigen. Es sei jetzt $b_p > 0$ und $a_{i+1,i} > 0$ $(i = 1, \ldots, p)$.
Wir zeigen: In Schritt 1 ist $\beta_{i,j,i-j} > 0$ $(j = 0, \ldots, i-1)$. Für $i = 1$ ist $\beta_{1,0,1} = 1$. Für $k \leq i$ sei die Behauptung richtig. Wir zeigen den Induktionsschritt $i \rightsquigarrow (i+1)$. Wenn man in (3.9) $j = i + 1 - l$ setzt, erhält man

$$\beta_{i+1,1,i} = a_{i+1,i} \quad , \quad \beta_{i+1,l,i+1-l} = a_{i+1,i} \cdot \beta_{i,l-1,i+1-l} \quad (l = 2, \ldots, i+1) .$$

Es ist stets $\beta_{i+1,i+1,0} = 1$, denn dies ist der Koeffizient von $\mathfrak{Bu}_i$ in K_i.
Zusammen mit der Induktionsannahme für $\beta_{i,j,i-j}$ und der Voraussetzung für die $a_{i+1,i}$ folgt die Behauptung.
Nun setzen wir in (3.10) $k = p$ bzw. $k = p - j + 1$ und erhalten

$$\delta_{1,p} = b_p \quad , \quad \delta_{j,p-j+1} = b_p \beta_{p,j-1,p-j+1} \qquad (j = 2, \ldots, p) \ .$$

Wegen der obigen Behauptung und da $b_p > 0$ ist nach Voraussetzung, sind auch die $\delta_{i,p-i+1} > 0$ $(i = 1, \ldots, p)$. Setzt man nun in (3.11) $k = p - j + 1$ und berücksichtigt, dass $d_j = j!$ ist, so erhält man, dass die $\gamma_{j,p-j+1} > 0$ sind für $j = 1, \ldots, p$.
Damit ist der Satz bewiesen. □

Bemerkung 3.1.4. In dem Satz ist vorausgesetzt, dass die Ordnungszahl mit der Stufenzahl des Runge-Kutta-Verfahrens übereinstimmen muss. Diese Voraussetzung, welche Ferretti in seinem Artikel [13] nicht erwähnt, schränkt die Klasse der Runge-Kutta-Verfahren doch erheblich ein. Insbesondere gibt es kein Runge-Kutta-Verfahren der Ordnung $s \geq 5$, das diese Bedingung erfüllt. Diese Aussage findet man in [10, Theorem 370B, p. 259]. Zum Beispiel hat ein Runge-Kutta-Verfahren der Ordnung 5 mindestens 6 Stufen.

3.1.2. Modifizierte RK-Verfahren und Konsistenz

An dieser Stelle benötigen wir einige Definitionen. Wir wollen die Koeffizienten des letzten Satzes als Matrix darstellen. Und wir benötigen eine weitere Matrix aus technischen Gründen. Wir definieren

$$\Gamma_p := \begin{pmatrix} \gamma_{1,1} & \cdots & \gamma_{1,p} \\ \vdots & \ddots & \vdots \\ \gamma_{p,1} & \cdots & \gamma_{p,p} \end{pmatrix} \quad \text{und} \quad \Delta_p := \begin{pmatrix} 1! & 0 & \cdots & 0 \\ 0 & 2! & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & p! \end{pmatrix} ,$$

welche beide reelle $p \times p$-Matrizen sind. Die Matrix Γ_p nennen wir *definierende Matrix* eines RK-Verfahrens, denn durch ihre Einträge wird mittels Satz 3.1.3 ein RK-Verfahren für (LKP) definiert
Bevor wir weiter machen, soll zur Verdeutlichung die definierende Matrix des Heun-Verfahrens berechnet werden.

Beispiel 3.1.5. In diesem Beispiel sei $m = 1$ (die Dimension des Kontrollbereichs). Wir betrachten das Heun-Schema. Das Butcher Array dafür sieht so aus:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline 1 & \frac{1}{2} & \frac{1}{2} \end{array}$$

Es hat zwei Stufen. Damit ist $p = 2$. Mit Hilfe von Beispiel 2.2.3 können wir damit folgende Koeffizienten zuordnen:

$$\mathfrak{b}^T = (b_1, b_2) = \left(\frac{1}{2}, \frac{1}{2}\right) \quad , \quad \mathfrak{c}^T = (c_1, c_2) = (0, 1)$$

und

$$a_{2,1} = 1 \quad , \quad a_{1,1} = a_{1,2} = a_{2,2} = 0 \, .$$

Wir wollen die definierende Matrix Γ_2 für das Heun-Verfahren berechnen. Wir setzten also die gerade aus dem Butcher-Array bestimmten Koeffizienten in die definierenden Gleichungen (3.3)-(3.6) ein und erhalten für den ersten Zeitschritt, d.h. $k = 0$

$$x_h(h) = x_h(0) + h\left(b_1 K_1 + b_2 K_2\right) = \bar{x}_0 + h\left(\frac{1}{2}K_1 + \frac{1}{2}K_2\right) \, ,$$

wobei

$$\begin{aligned} K_1 &= \mathfrak{A} x_h(0) + \mathfrak{B}\mathfrak{u}_1 = \mathfrak{A}\bar{x}_0 + \mathfrak{B}\mathfrak{u}_1 \\ K_2 &= \mathfrak{A}\left(x_h\left(0\right) + h\sum_{j=1}^{2-1} a_{i,j} K_j\right) + \mathfrak{B}\mathfrak{u}_2 = \mathfrak{A}\bar{x}_0 + h\mathfrak{A}K_1 + \mathfrak{B}\mathfrak{u}_2 \end{aligned}$$

ist. Damit ergibt sich

$$K_2 = \mathfrak{A}\bar{x}_0 + h\mathfrak{A}^2\bar{x}_0 + h\mathfrak{A}\mathfrak{B}\mathfrak{u}_1 + \mathfrak{B}\mathfrak{u}_2$$

und insgesamt

$$x_h(h) = \bar{x}_0 + h\mathfrak{A}\bar{x}_0 + \frac{h^2}{2}\mathfrak{A}^2\bar{x}_0 + h\mathfrak{B}\left(\frac{1}{2}\mathfrak{u}_1 + \frac{1}{2}\mathfrak{u}_2\right) + \frac{h^2}{2}\mathfrak{A}\mathfrak{B}\mathfrak{u}_1 \, .$$

Vergleichen wir dies mit der Darstellung aus Satz 3.1.3 so erhalten wir

$$\gamma_{1,1} = \gamma_{1,2} = \frac{1}{2} \quad , \quad \gamma_{2,1} = 1 \quad \text{und} \quad \gamma_{2,2} = 0 \, .$$

Also ist

$$\Gamma_2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{pmatrix}$$

die definierende Matrix für das Heun-Verfahren.

Wir brauchen im Folgenden noch vergrößerte Versionen der Matrizen Γ_p und Δ_p. Dabei wird jeder Eintrag ersetzt durch eine $m \times m$ Diagonalmatrix mit diesem Eintrag in der Diagonale. Sei

$$\Gamma_{m,p} := \begin{pmatrix} \gamma_{1,1}\mathfrak{E}_m & \cdots & \gamma_{1,p}\mathfrak{E}_m \\ \vdots & \ddots & \vdots \\ \gamma_{p,1}\mathfrak{E}_m & \cdots & \gamma_{p,p}\mathfrak{E}_m \end{pmatrix} \quad \text{und} \quad \Delta_{m,p} := \begin{pmatrix} 1!\mathfrak{E}_m & 0 & \cdots & 0 \\ 0 & 2!\mathfrak{E}_m & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & p!\mathfrak{E}_m \end{pmatrix} \, ,$$

wobei $\mathfrak{E}_m$ die $m \times m$-Einheitsmatrix ist. Die Matrizen $\Gamma_{m,p}$ und $\Delta_{m,p}$ sind damit $mp \times mp$-Matrizen.

Jetzt sind wir in der Lage Satz 3.1.3 in Matrixschreibweise darzustellen:

$$x_h(h) = \sum_{i=0}^{p} \frac{h^i}{i!}\mathfrak{A}^i \cdot \bar{x}_0 + \left[h\mathfrak{B} \mid h^2\mathfrak{A}\mathfrak{B} \mid h^3\mathfrak{A}^2\mathfrak{B} \mid \ldots \mid h^p\mathfrak{A}^{p-1}\mathfrak{B}\right] \Delta_{m,p}^{-1}\Gamma_{m,p} \begin{pmatrix} \mathfrak{u}_1 \\ \vdots \\ \mathfrak{u}_p \end{pmatrix} \tag{3.13}$$

Dabei ist $[\mathfrak{B} \mid \mathfrak{A}\mathfrak{B} \mid \mathfrak{A}^2\mathfrak{B} \mid \ldots \mid \mathfrak{A}^{p-1}\mathfrak{B}]$ die $n \times mp$-Matrix die durch Aneinanderreihung der Spalten der einzelnen Matrizen entsteht. Und $\left(\mathfrak{u}_1^T, \ldots, \mathfrak{u}_p^T\right)^T$ ist ein Spaltenvektor mit mp Komponenten. Dabei sorgt die Struktur der vergrößerten Matrizen $\Gamma_{m,p}$ und $\Delta_{m,p}$ dafür, dass die Komponenten von jedem Kontrollvektor $\mathfrak{u}_i$ je gleich behandelt werden. Mit anderen Worten wird so die skalare Multiplikation mit einem Kontrollvektor $\gamma_{i,j} \cdot \mathfrak{u}_j$ dargestellt.

Es sei nochmal erinnert, dass n die Dimension des Kontrollproblems ist, m die Dimension des Kontrollbereichs und der Kontrollvektoren und p ist die Stufenzahl des Verfahrens (und deren Ordnung). Diese Bezeichnungen behalten für den Rest dieser Arbeit ihre Gültigkeit, außer es wird etwas anderes angegeben.

Da wir die Inversen der Matrizen $\Gamma_{m,p}$ benötigen, müssen wir deren Invertierbarkeit beweisen.

Lemma 3.1.6.

(i) Sei $\mathfrak{M} \in \mathbb{R}^{k,k}$ eine Matrix mit $\mathfrak{M} = (m_{i,j})_{i,j=1,\ldots,k}$ und $\mathfrak{M}_l := (\mathfrak{E}_l \cdot m_{i,j})_{i,j=1,\ldots,k}$ die vergrößerte Version. Dann gilt:

$$\mathfrak{M} \quad \text{invertierbar} \quad \Longrightarrow \quad \mathfrak{M}_l \quad \text{invertierbar} \qquad \forall l \in \mathbb{N}$$

und

$$\mathfrak{M}_l \quad \text{invertierbar für ein } l \in \mathbb{N} \quad \Longrightarrow \quad \mathfrak{M} \quad \text{invertierbar}\,.$$

(ii) Die definierende Matrix Γ_p eines RK-Verfahrens ist invertierbar, falls die Koeffizienten im Butcher-Array $b_p > 0$ und $a_{i+1,i} > 0$ $(i = 1, \ldots, p-1)$ erfüllen. Insbesondere ist dann auch $\Gamma_{m,p}$ invertierbar $\quad \forall m \in \mathbb{N}$.

Beweis.
zu (i) : Es sei $(\mathfrak{f}_i)_{i=1,\ldots,k}$ die kanonische Basis im $\mathbb{R}^k$ und $(\mathfrak{e}_i)_{i=1,\ldots,lk}$ die kanonische Basis im $\mathbb{R}^{lk}$. Dann definieren wir für $j = 1, \ldots, m$ die lineare Abbildung $\sigma_j : \mathbb{R}^k \longrightarrow \mathbb{R}^{lk}$ durch

$$\sigma_j(\mathfrak{f}_i) = \mathfrak{e}_{j+l(i-1)} \quad (i = 1, \ldots, k)\,.$$

Weiter bezeichnen wir mit $U_j := \operatorname{span}\left(\mathfrak{e}_j, \mathfrak{e}_{j+l}, \ldots, \mathfrak{e}_{j+l(k-1)}\right)$ das Bild von σ_j. Offensichtlich ist σ_j injektiv und es gilt $\mathbb{R}^{lk} = \bigoplus_{j=1,\ldots l} U_j$.

Im Folgenden bezeichnen wir mit $\mathfrak{m}^i$ die i-te Spalte von M und mit $\bar{\mathfrak{m}}^i$ die i-te Spalten von $\mathfrak{M}_l$. Dann gilt $\sigma_j\left(\mathfrak{m}^i\right) = \bar{\mathfrak{m}}^{j+(i-1)l}$.

a) Sei $l \in \mathbb{N}$ beliebig und $\mathfrak{M}$ invertierbar. Wir zeigen: Dann ist $\mathfrak{M}_l$ invertierbar.

Weil $\mathfrak{M}$ invertierbar ist, sind die Spalten von $\mathfrak{M}$ linear unabhängig und bilden eine Basis des $\mathbb{R}^k$. Deren Bilder unter σ_j, nämlich $\bar{\mathfrak{m}}^j, \bar{\mathfrak{m}}^{j+l}, \ldots, \bar{\mathfrak{m}}^{j+(k-1)l}$, sind linear unabhängig, weil σ_j injektiv ist. Also bilden diese Vektoren eine Basis des Unterraumes U_j. Da $\mathbb{R}^{lk}$ direkte Summe der Unterräume U_j für $j = 1, \ldots, l$ ist, bilden die Spalten von $\mathfrak{M}_l$ eine Basis des $\mathbb{R}^{lk}$. Damit ist $\mathfrak{M}_l$ invertierbar.

b) Sei $l \in \mathbb{N}$, sodass $\mathfrak{M}_l$ invertierbar ist. Wir zeigen: Dann ist $\mathfrak{M}$ invertierbar.

Sei $j \in \{1, \ldots, l\}$. Da $\mathfrak{M}_l$ invertierbar ist, sind die Spalten $\bar{\mathfrak{m}}^j, \bar{\mathfrak{m}}^{j+l}, \ldots, \bar{\mathfrak{m}}^{j+(k-1)l}$ linear unabhängig. Also sind auch deren Urbilder unter σ_j, nämlich $\mathfrak{m}^1, \ldots, \mathfrak{m}^k$, linear unabhängig, da σ_j injektiv ist. Dies sind aber schon alle Spalten von $\mathfrak{M}$. Also ist $\mathfrak{M}$ invertierbar.

zu (ii): Sei $b_p > 0$ und $a_{i+1,i} > 0$ $(i = 1, \ldots, p-1)$. Nach Satz 3.1.3 ist dann $\gamma_{i,p-i+1} > 0$ für $i = 1, \ldots, p$ und $\gamma_{i,j} = 0$ für $p - i < j - 1$. Dann hat Γ_p die Gestalt

$$\Gamma_p = \begin{pmatrix} \gamma_{1,1} & \cdots & \gamma_{1,p} \\ \vdots & \iddots & 0 \\ \gamma_{p,1} & 0 & 0 \end{pmatrix}$$

mit stets positiver Diagonale. Damit sind die Spalten linear unabhängig und Γ_p ist invertierbar. □

Bemerkung 3.1.7. Die Bedingungen an die Koeffizienten eines p-stufigen expliziten RK-Verfahrens im Butcher-Array, nämlich

$$b_p > 0 \quad \text{und} \quad a_{i+1,i} > 0 \ (i = 1, \ldots, p-1) \,,$$

sind überhaupt nicht restriktiv. Alle expliziten Runge-Kutta Verfahren mit höchstens 4 Stufen, die ich in der Literatur gefunden habe, erfüllen sie. Insbesondere erfüllen auch alle RK-Verfahren aus Beispiel 2.2.5 diese Bedingungen.

Wie bereits in Bemerkung 3.1.4 erwähnt wurde, kommen RK-Verfahren mit mehr als 4 Stufen nicht in Frage.

Nun definieren wir die *Auswahlmenge*

$$\mathcal{I}_{m,p} := \Bigg\{ \left(\zeta_1^T, \ldots, \zeta_p^T\right)^T \in \mathbb{R}^{mp} \mid \zeta_1 = \int_0^1 \mathfrak{u}\left(s_1\right) \, ds_1 \,, \ldots, \\ \zeta_p = \int_0^1 \cdots \int_0^{s_{p-1}} \mathfrak{u}\left(s_p\right) \, ds_p \ldots ds_1 \text{ für } \mathfrak{u} \in L^1\left([0,1]\,; U\right) \Bigg\} \,.$$

Dabei werden die Vektoren ζ_i zu einem langen Spaltenvektor vereinigt. Damit können wir nun folgende wichtige Definition tätigen.

Definition 3.1.8. Es sei für ein p-stufiges RK-Verfahren der Ordnung p die definierende Matrix Γ_p nach Satz 3.1.3 gegeben. Für die Koeffizienten im Butcher-Array des RK-Verfahrens gelte $b_p > 0$ und $a_{i+1,i} > 0 \quad (i = 1, \dots, p-1)$. Dann ist $\Gamma_{m,p}$ invertierbar und wir definieren den *diskreten Kontrollbereich* als die Menge $\mathcal{U}_{m,p} := \Gamma_{m,p}^{-1} \Delta_{m,p} \mathcal{I}_{m,p}$.

Es sei eine Kontrollfunktion $\mathfrak{u} \in L^1(I;U)$ für (LKP) vorgegeben. Und es sei $h = \frac{t_f}{N}$ die Schrittweite, wobei N die Anzahl der Iterationen ist, vorgegeben. Wir definieren das folgende Einschrittverfahren zur Konstruktion einer diskreten Lösung x_h von (LKP) mittels seiner Verfahrensfunktion

$$\Phi_h(x, \mathfrak{u}_1, \dots, \mathfrak{u}_p) := \sum_{i=1}^{p} \frac{h^{i-1}}{i!} \mathfrak{A}^i \cdot x + \sum_{i=1}^{p} h^{i-1} \mathfrak{A}^{i-1} \mathfrak{B} \cdot \sum_{j=1}^{p} \gamma_{i,j} \mathfrak{u}_j \,,$$

dabei sind die Kontrollvektoren $\left(\mathfrak{u}_1^T, \dots, \mathfrak{u}_p^T\right)^T \in \mathcal{U}_{m,p}$, und nennen es *modifiziertes RK-Schema der Ordnung p*. Falls die Kontrollvektoren nicht beliebig aus $\mathcal{U}_{m,p}$ sind, sondern folgender Gleichung

$$\left(\mathfrak{u}_1^T, \dots, \mathfrak{u}_p^T\right)^T = \left(\bar{\mathfrak{u}}_1(t)^T, \dots, \bar{\mathfrak{u}}_p(t)^T\right)^T := \Gamma_{m,p}^{-1} \Delta_{m,p} \left(\zeta_1(t)^T, \dots, \zeta_p(t)^T\right)^T$$

mit $\zeta_i(t) = \int_0^1 \cdots \int_0^{s_{i-1}} \mathfrak{u}(t + s_i h)\, ds_i \dots ds_1$ für $i = 1, \dots, p$ und $t \in [0, t_f - h]$ genügen, so heiße das Einschrittverfahren dann *modifiziertes RK-Verfahren der Ordnung p*. Wir schreiben dann die Verfahrensfunktion als

$$\Phi_h(t, x) := \Phi_h(x; \bar{\mathfrak{u}}_1(t), \dots, \bar{\mathfrak{u}}_p(t)) \,.$$

Weiter schreiben wir $\bar{\mathfrak{u}}_{i,k}$ statt $\bar{\mathfrak{u}}_i(kh)$ und kurz $\bar{\mathfrak{u}}_i$ statt $\bar{\mathfrak{u}}_{i,0}$.

Bemerkung 3.1.9.

(i) Die Kontrollvektoren der mod. RK-Verfahren sind in $\mathcal{U}_{m,p}$ für $t \in [0, t_f - h]$. Denn für $\mathfrak{u} \in L^1(I;U)$ und $t \in [0, t_f - h]$, ist $\mathfrak{u}(t + h\cdot) \in L^1([0,1];U)$, weil $[t, t+h] \subset I = [0, t_f]$ ist. Also ist dann $\left(\zeta_1(t)^T, \dots, \zeta_p(t)^T\right)^T \in \mathcal{I}_{m,p}$. Nach Definition der Menge $\mathcal{U}_{m,p}$ sind dann die Kontrollvektoren $\left(\bar{\mathfrak{u}}_1(t)^T, \dots, \bar{\mathfrak{u}}_p(t)^T\right)^T \in \mathcal{U}_{m,p}$.

(ii) Der diskrete Kontrollbereich $\mathcal{U}_{m,p}$ hängt von dem konkreten RK-Verfahren ab, das dem modifizierten RK-Verfahren zugrunde liegt, da hier die Matrix $\Gamma_{m,p}$ eingeht.

(iii) Das modifizierte RK-Schema ist im Gegensatz zum modifizierten RK-Verfahren eigentlich kein richtiges Einschrittverfahren, da die Kontrollvektoren dadurch nicht festgelegt werden. Es ist vielmehr eine Art Vorstufen eines Einschrittverfahrens.

In Abbildung 3.1 sieht man den diskreten Kontrollbereich $\mathcal{U}_{1,2}$ für das mod. Heun-Verfahren und für das mod. verbesserte Euler-Verfahren. Es ist deutlich zu erkennen, dass diese beiden Mengen sich unterscheiden.

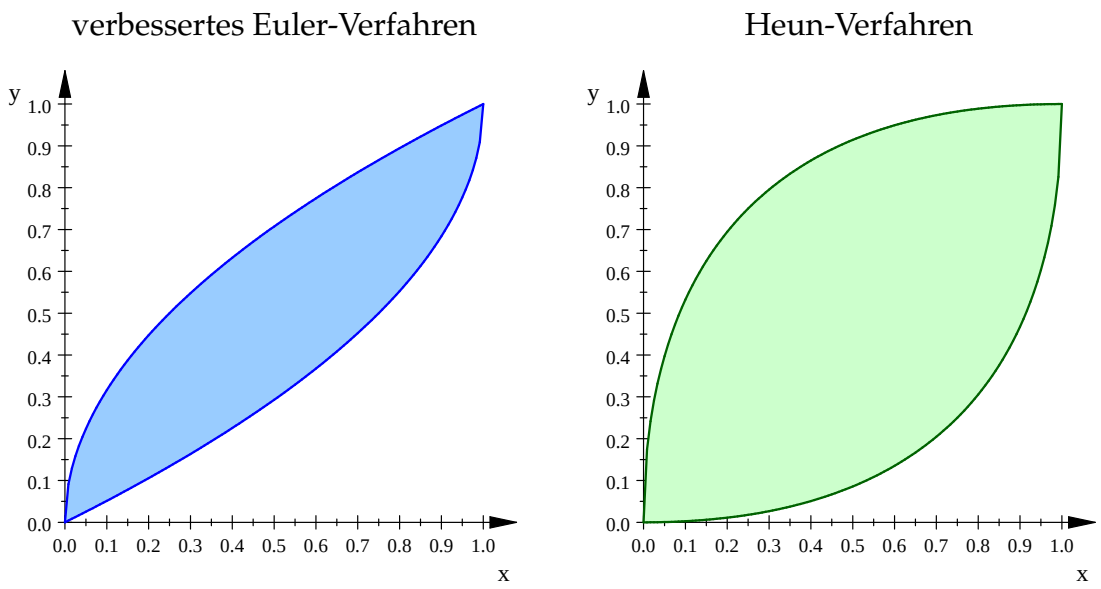

Abbildung 3.1.: Diskrete Kontrollbereiche für das modifizierte Heun-Verfahren und für das modifizierte, verbesserte Euler-Verfahren

Damit können wir das Hauptresultat dieses Abschnitts formulieren.

Theorem 3.1.10. *Sei $h \in I$. Für das lokale lineare Kontrollproblem (LKP) auf dem Intervall $[0,h]$ (d.h. $N=1$) sei ein modifiziertes RK-Schema der Ordnung p gegeben. Weiter sei $G \subset \mathbb{R}^n$ ein Kompaktum und die lokale Anfangsbedingung $\bar{x}_0 \in G$. Die diskrete Lösung x_h werde mittels dieses Schemas berechnet. Weiter seien dabei die Koeffizienten $b_p > 0$ und $a_{i+1,i} > 0 \quad (i = 1, \ldots, p-1)$. Dann gilt:*

(A1) *Es seien eine Kontrollfunktion $\mathfrak{u} \in L^1([0,h]; U)$ und die dazugehörige Lösung $\hat{x}$ von (LKP) vorgegeben. Dann gibt es Kontrollvektoren $\left(\mathfrak{u}_1^T, \ldots, \mathfrak{u}_p^T\right)^T \in \mathcal{U}_{m,p}$, welche die diskrete Lösung $x_h(h)$ bestimmen, sodass (A3) gilt. Die Kontrollvektoren können dabei gemäß dem modifizierten RK-Verfahren gewählt werden.*

(A2) *Es seien Kontrollvektoren $\left(\mathfrak{u}_1^T, \ldots, \mathfrak{u}_p^T\right)^T \in \mathcal{U}_{m,p}$ und die dazugehörige diskrete Lösung $x_h(h)$ vorgegeben. Dann gibt es eine Kontrollfunktion $\mathfrak{u} \in L^1([0,h]; U)$, welche die exakte Lösung $\hat{x}$ bestimmt, sodass (A3) gilt.*

(A3) *Es gilt*

$$\begin{aligned}&\|\hat{x}(h) - x_h(h)\|_\infty \leq \\ &\quad \left[\left\|\mathfrak{A}^{p+1}\right\|_Z \cdot e^{\|\mathfrak{A}\|_Z t_f} \left(\|G\|_\infty + t_f \|\mathfrak{B}\|_Z \|U\|_\infty\right) + \|\mathfrak{A}^p \mathfrak{B}\|_Z \cdot \|U\|_\infty\right] \cdot h^{p+1}\end{aligned}$$

für $h \to 0$, wobei die Konstante unabhängig ist von der diskreten und exakten Lösung und deren Kontrollen, von der Schrittweite h und auch von der Wahl der lokalen Anfangsbedingung $\bar{x}_0$ aus dem Kompaktum G.

Insbesondere besagt (A1), dass ein modifiziertes Runge-Kutta-Verfahren von Ordnung p konsistent ist von Ordnung p.

Beweis. Wegen der Bedingungen an die Koeffizienten b_p und $a_{i+1,i}$ des Runge-Kutta-Verfahrens sind die Matrizen Γ_p bzw. $\Gamma_{m,p}$ nach Lemma 3.1.6 invertierbar.
$(A1) \implies (A3)$
Sei $\mathfrak{u} \in L^1([0,h]; U)$ gegeben. Dann ist die Funktion $\mathfrak{u}(h\cdot) \in L^1([0,1]; U)$ und wir definieren

$$\zeta_i := \int_0^1 \cdots \int_0^{s_{i-1}} \mathfrak{u}(hs_i)\, ds_i \dots ds_1 \qquad (i = 1, \dots, p)\,. \tag{3.14}$$

Dann ist offensichtlich $\left(\zeta_1^T, \dots, \zeta_p^T\right)^T \in \mathcal{I}_{m,p}$. Da $\Gamma_{m,p}$ invertierbar ist, setzen wir die Kontrollvektoren mittels

$$\begin{pmatrix} \mathfrak{u}_1 \\ \vdots \\ \mathfrak{u}_p \end{pmatrix} := \Gamma_{m,p}^{-1} \Delta_{m,p} \begin{pmatrix} \zeta_1 \\ \vdots \\ \zeta_p \end{pmatrix}. \tag{3.15}$$

Damit ist $\left(\mathfrak{u}_1^T, \dots, \mathfrak{u}_p^T\right)^T \in \mathcal{U}_{m,p}$ nach Definition. Nach Konstruktion entspricht der Vektor $\left(\mathfrak{u}_1^T, \dots, \mathfrak{u}_p^T\right)^T$ dem Kontrollvektor $\left(\bar{\mathfrak{u}}_1^T, \dots, \bar{\mathfrak{u}}_p^T\right)^T$ des modifizierten RK-Verfahrens.
Wendet man die Substitutionsregel mehrfach auf (3.14) an so erhält man

$$\zeta_i = \frac{1}{h^i} \int_0^h \cdots \int_0^{s_{i-1}} \mathfrak{u}(s_i)\, ds_i \dots ds_1 \qquad (i = 1, \dots, p)\,.$$

Damit folgt mit einer kleinen Umformung von (3.15)

$$\frac{h^i}{i!} \sum_{j=1}^{p} \gamma_{i,j} \mathfrak{u}_j = \int_0^h \cdots \int_0^{s_{i-1}} \mathfrak{u}(s_i)\, ds_i \dots ds_1 \qquad (i = 1, \dots, p)\,. \tag{3.16}$$

Mit diesen Kontrollvektoren konstruieren wir die diskrete Lösung $x_h(h)$. Mittels der Entwicklungen in Satz 3.1.1 und der Darstellung aus Satz 3.1.3 erhalten wir

$$\begin{aligned} \hat{x}(h) - x_h(h) =& \sum_{i=0}^{p} \frac{h^i}{i!} \mathfrak{A}^i \cdot \bar{x}_0 - \sum_{i=0}^{p} \frac{h^i}{i!} \mathfrak{A}^i \cdot x_1 + \mathfrak{r}(h) + \\ &+ \sum_{i=1}^{p} \mathfrak{A}^{i-1} \mathfrak{B} \left[\int_0^h \cdots \int_0^{s_{i-1}} \mathfrak{u}(s_i)\, ds_i \dots ds_1 - \frac{h^i}{i!} \sum_{j=0}^{p-1} \gamma_{i,j} \mathfrak{u}_j \right] \\ =& \mathfrak{r}(h)\,, \end{aligned} \tag{3.17}$$

wobei $\mathfrak{r}(h)$ der Restterm aus Satz 3.1.1 ist, für den auch die dortige Abschätzung gilt.
$(A2) \implies (A3)$
Sei $\left(\mathfrak{u}_1^T, \ldots, \mathfrak{u}_p^T\right)^T \in \mathcal{U}_{m,p}$ gegeben. Dann gibt es nach Definition von $\mathcal{U}_{m,p}$ einen Vektor $\left(\zeta_1^T, \ldots, \zeta_p^T\right)^T \in \mathcal{I}_{m,p}$ mit

$$\Gamma_{m,p} \begin{pmatrix} \mathfrak{u}_1 \\ \vdots \\ \mathfrak{u}_p \end{pmatrix} = \Delta_{m,p} \begin{pmatrix} \zeta_1 \\ \vdots \\ \zeta_p \end{pmatrix} \quad \text{bzw} \quad \begin{pmatrix} \mathfrak{u}_1 \\ \vdots \\ \mathfrak{u}_p \end{pmatrix} = \Gamma_{m,p}^{-1} \Delta_{m,p} \begin{pmatrix} \zeta_1 \\ \vdots \\ \zeta_p \end{pmatrix},$$

da $\Gamma_{m,p}$ invertierbar ist.
Wiederum nach Definition von $\mathcal{I}_{m,p}$ gibt es eine Funktion $\tilde{\mathfrak{u}} \in L^1\left([0,1]; U\right)$ mit

$$\zeta_i = \int_0^1 \cdots \int_0^{s_{i-1}} \tilde{\mathfrak{u}}\left(s_i\right) \, ds_i \ldots ds_1 = \frac{1}{h^i} \int_0^h \cdots \int_0^{s_{i-1}} \tilde{\mathfrak{u}}\left(\frac{1}{h} s_i\right) \, ds_i \ldots ds_1 \qquad (i = 1, \ldots, p), \tag{3.18}$$

wobei die zweite Darstellung mit Hilfe der Substitutionregel erfolgt.
Nun setzten wir $\mathfrak{u}(t) := \tilde{\mathfrak{u}}\left(\frac{1}{h} \cdot t\right)$. Dann ist $\mathfrak{u} \in L^1\left([0,h]; U\right)$. Mit diesem $\mathfrak{u}$ gilt nun die Darstellung (3.16) und schließlich auch die Gleichungen (3.17). □

3.2. Ergänzungen und neue Verfahren

In diesem Abschnitt soll die Theorie von Ferretti ergänzt werden. In seinem Artikel liefert er keinen Konvergenzbeweis für die modifizierten Runge-Kutta-Verfahren. Dies soll hier nachgeholt werden. Außerdem wollen wir die Einschränkung beseitigen, dass sich mit diesem Zugang maximal die Konvergenzordnung 4 erreichen lässt (vgl. Bemerkung 3.1.4). Dies geschieht durch Einführung neuer Verfahren.

Wir wollen mit der Definition dieser Verfahren anfangen, dann können wir die Konvergenz für beide Verfahrensarten simultan zeigen.

Definition 3.2.1. Es sei eine Kontrollfunktion $\mathfrak{u} \in L^1\left(I; U\right)$ für (LKP) vorgegeben. Wie bei allen Einschrittverfahren gelte für die Schrittweite $h = \frac{t_f}{N}$, wobei $N \in \mathbb{N}$ die Anzahl der Iterationen des Verfahrens ist.

Das *Ferretti-Schema der Ordnung* p zu der Konstruktion einer diskreten Lösung $x_h : \mathbb{G}_h \longrightarrow \mathbb{R}^n$ für (LKP) ist gegeben durch die Verfahrensfunktion

$$\Phi_h\left(x, \mathfrak{v}_1, \ldots, \mathfrak{v}_p\right) := \sum_{i=1}^{p} \frac{h^{i-1}}{i!} \mathfrak{A}^i \cdot x + \sum_{i=1}^{p} h^{i-1} \mathfrak{A}^{i-1} \mathfrak{B} \cdot \mathfrak{v}_i \,,$$

dabei sind die Kontrollvektoren $(\mathfrak{v}_1^T, \ldots, \mathfrak{v}_p^T)^T \in \mathcal{I}_{m,p}$.
Dann wird x_h mit der Verfahrensfunktion gemäß dem allgemeinen Einschrittverfahren konstruiert (vgl. Definition 2.2.1). Falls die Kontrollvektoren nicht beliebig aus $\mathcal{I}_{m,p}$ sind, sondern folgende Gleichungen

$$\mathfrak{v}_i = \bar{\mathfrak{v}}_i(t) := \int_0^1 \cdots \int_0^{s_{i-1}} \mathfrak{u}\,(s_i h + t)\; ds_i \ldots ds_1 \qquad i = 1, \ldots, p\ ,\ t \in [0, t_f - h] \quad (3.19)$$

erfüllen, so heißt dieses Einschrittverfahren dann *Ferretti-Verfahren der Ordnung* p. Dafür schreiben wir die Verfahrensfunktion kurz als

$$\Phi_h\,(t, x) := \Phi_h\,(x, \bar{\mathfrak{v}}_1(t), \ldots, \bar{\mathfrak{v}}_p(t))\ .$$

Zur Abkürzung schreiben wir auch $\bar{\mathfrak{v}}_{i,k}$ statt $\bar{\mathfrak{v}}_i(kh)$ und auch $\bar{\mathfrak{v}}_i$ statt $\bar{\mathfrak{v}}_{i,0}$.

Bemerkung 3.2.2.

(i) Zum Unterschied zu den Kontrollvektoren für die Runge-Kutta-Verfahren schreiben wir hier $\mathfrak{v}_i$ statt $\mathfrak{u}_i$ bzw. $\bar{\mathfrak{v}}_{i,k}$ statt $\bar{\mathfrak{u}}_{i,k}$. Wir haben gesehen, dass die $\bar{\mathfrak{u}}_{i,k} \in \mathcal{U}_{m,p}$ sind für alle i, k (siehe Bemerkung 3.1.9). Es bleibt zu zeigen, dass die Kontrollvektoren $(\bar{\mathfrak{v}}_{1,k}^T, \ldots, \bar{\mathfrak{v}}_{p,k}^T)^T$ der Ferretti-Verfahren tatsächlich in $\mathcal{I}_{m,p}$ sind für $k = 0, \ldots, N-1$. Dies folgt mit dem selben Argument wie in Bemerkung 3.1.9 .

(ii) In seinem Artikel [13] hat Ferretti diese Verfahren nicht betrachtet. Es scheint mir dennoch angemessen sie nach ihm zu benennen, da sie auf der Entwicklung der Lösung aus Satz 3.1.1 auf Seite 52 beruhen, die er in seiner Arbeit vorgestellt hat.

(iii) Auch das Ferretti-Schema ist im Gegensatz zum Ferretti-Verfahren eigentlich kein Einschrittverfahren, da die Kontrollvektoren nicht festgelegt werden.

Damit können wir schon Konsistenz für diese Verfahren zeigen. Wir formulieren das Analogon zu Theorem 3.1.10 für die Ferretti-Verfahren.

Theorem 3.2.3. *Sei $h \in I$. Für das lokale lineare Kontrollproblem (LKP) auf dem Intervall $[0, h]$ (d.h. $N = 1$) sei das Ferretti-Schema der Ordnung p gegeben. Weiter sei $G \subset \mathbb{R}^n$ ein Kompaktum und die lokale Anfangsbedingung $\bar{x}_0 \in G$. Die diskrete Lösung x_h werde mittels dieses Schemas berechnet. Dann gilt:*

(A1) *Es seien eine Kontrollfunktion $\mathfrak{u} \in L^1\,([0, h]\,; U)$ und die dazugehörige Lösung $\hat{x}$ von (LKP) vorgegeben. Wählt man die Kontrollvektoren $\left(\mathfrak{v}_1^T, \ldots, \mathfrak{v}_p^T\right)^T \in \mathcal{I}_{m,p}$ gemäß dem Ferretti-Verfahren, dann gilt für das dadurch festgelegte x_h die Aussage (A3).*

(A2) *Es seien Kontrollvektoren* $\left(\mathfrak{v}_1^T, \ldots, \mathfrak{v}_p^T\right)^T \in \mathcal{I}_{m,p}$ *und die mittels des Ferretti-Schemas konstruierte diskrete Lösung* $x_h(h)$ *vorgegeben. Dann gibt es eine Kontrollfunktion* $\mathfrak{u} \in L^1([0,h];U)$*, welche die exakte Lösung* $\hat{x}$ *bestimmt, sodass (A3) gilt.*

(A3) *Es gilt*

$$\|\hat{x}(h) - x_h(h)\|_\infty \leq \left[\left\|\mathfrak{A}^{p+1}\right\|_Z \cdot e^{\|\mathfrak{A}\|_Z t_f}\left(\|G\|_\infty + t_f \|\mathfrak{B}\|_Z \|U\|_\infty\right) + \|\mathfrak{A}^p\mathfrak{B}\|_Z \cdot \|U\|_\infty\right] \cdot h^{p+1}$$

für $h \to 0$*, wobei die Konstante unabhängig ist von der diskreten und exakten Lösung und deren Kontrollen, von der Schrittweite* h *und auch von der Wahl der Anfangsbedingung* $\bar{x}_0$ *aus dem Kompaktum* G*.*

Insbesondere besagt (A1), dass ein Ferretti-Verfahren der Ordnung p *konsistent ist mit Konsistenzordnung* p*.*

Beweis.
$(A1) \implies (A3)$
Es sei $\mathfrak{u} \in L^1([0,h];U)$ und damit $\hat{x}$ gegeben. Weiter sei $\mathfrak{v}_i := \bar{\mathfrak{v}}_i(0) \quad (i = 1, \ldots, p)$ für diese vorgegebene Steuerfunktion $\mathfrak{u}$ gemäß (3.19) berechnet. Nach Bemerkung 3.2.2(i) ist dann $\left(\mathfrak{v}_1^T, \ldots, \mathfrak{v}_p^T\right)^T \in \mathcal{I}_{m,p}$. Wendet man nun iteriert die Substitutionsregel (=Transformationsformel) auf $\mathfrak{v}_i$ an, so erhält man

$$\mathfrak{v}_i = \int_0^1 \cdots \int_0^{s_{i-1}} \mathfrak{u}(s_i h)\, ds_i \ldots ds_1 = \frac{1}{h^i} \int_0^h \cdots \int_0^{s_{i-1}} \mathfrak{u}(s_i)\, ds_i \ldots ds_1 \qquad i = 1, \ldots, p\,.$$

Mittels der Entwicklung der Lösung $\hat{x}$ in Satz 3.1.1 und der Konstruktion von x_h nach dem Ferretti-Verfahren hat man damit

$$\begin{aligned}\hat{x}(h) - x_h(h) =& \sum_{i=0}^{p} \frac{h^i}{i!}\mathfrak{A}^i \cdot \bar{x}_0 - \sum_{i=0}^{p} \frac{h^i}{i!}\mathfrak{A}^i \cdot \bar{x}_0 + \mathcal{O}\left(h^{p+1}\right) \\ &+ \sum_{i=1}^{p} \mathfrak{A}^{i-1}\mathfrak{B}\left[\int_0^h \cdots \int_0^{s_{i-1}} \mathfrak{u}(s_i)\, ds_i \ldots ds_1 - h^i \mathfrak{v}_i\right] \qquad (3.20) \\ =& \mathfrak{r}(h)\,,\end{aligned}$$

wobei $\mathfrak{r}(h)$ der Restterm aus Satz 3.1.1 ist, für den auch die dortige Abschätzung gilt.
$(A2) \implies (A3)$
Es sei $\left(\mathfrak{v}_1^T, \ldots, \mathfrak{v}_p^T\right)^T \in \mathcal{I}_{m,p}$ und x_h gegeben. Mit derselben Argumentation wie im Beweis von Satz 3.1.10 folgt die Existenz eines $\mathfrak{u} \in L^1([0,h];U)$, sodass

$$\mathfrak{v}_i = \frac{1}{h^i} \int_0^h \cdots \int_0^{s_{i-1}} \mathfrak{u}(s_i)\, ds_i \ldots ds_1 \qquad (i = 1, \ldots, p)$$

gilt (siehe Gleichung 3.18). Für die zu $\mathfrak{u}$ gehörende analytische Lösung $\hat{x}$ gilt damit (3.20). Damit ist der Satz bewiesen. □

Nun können wir den Konvergenzsatz für beide Verfahrensarten zeigen. Dabei wird, wie auch bei den Konsistenzaussagen, auch die andere Richtung gezeigt, bei der man die diskrete Lösung x_h vorgibt. Dies ist im nächsten Abschnitt für die Erweiterung auf mengenwertigen Verfahren nützlich.

Theorem 3.2.4. *Für das lineare Kontrollproblem (LKP) auf $I = [0, t_f]$ mit Anfangsbedingung x_0 sei ein p-stufiges RK-Schema der Ordnung p oder das Ferretti-Schema der Ordnung p gegeben. D.h. die diskrete Lösung x_h wird mit einem der beiden Schemata konstruiert. Bei dem RK-Schema seien dabei die Koeffizienten $b_{p-1} > 0$ und $a_{i+1,i} > 0 \quad (i = 1, \dots, p-2)$. Dann gilt:*

(A) *Sei $\mathfrak{u} \in L^1(I; U)$ und damit auch die Lösung für (LKP) $\hat{x}$ vorgegeben. Dann erzeugt*

- *(i) das modifizierte RK-Verfahren eine Folge von Kontrollvektoren $\left(\bar{\mathfrak{u}}_{1,k}^T, \dots, \bar{\mathfrak{u}}_{p,k}^T\right)^T \in \mathcal{U}_{m,p}$ $(k = 0, \dots, N-1)$ welche die diskrete Lösung x_h festlegen, sodass (C) gilt.*
- *(ii) das Ferretti Verfahren eine Folge von Kontrollvektoren $\left(\bar{\mathfrak{v}}_{1,k}^T, \dots, \bar{\mathfrak{v}}_{p,k}^T\right)^T \in \mathcal{I}_{m,p}$ $(k = 0, \dots, N-1)$ welche die diskrete Lösung x_h festlegen, sodass (C) gilt.*

(B) *Es sei eine folge von Kontrollvektoren*

- *(i) $\left(\mathfrak{u}_{1,k}^T, \dots, \mathfrak{u}_{p,k}^T\right)^T \in \mathcal{U}_{m,p}$ $(k = 0, \dots, N-1)$ zusammen mit der dazu nach dem RK-Schema konstruierten diskreten Lösung x_h gegeben.*
- *(ii) $\left(\mathfrak{v}_{1,k}^T, \dots, \mathfrak{v}_{p,k}^T\right)^T \in \mathcal{I}_{m,p}$ $(k = 0, \dots, N-1)$ zusammen mit der dazu nach dem Ferretti-Schema konstruierten diskreten Lösung x_h gegeben.*

Dann gibt es eine Steuerfunktion $\mathfrak{u} \in L^1(I; U)$, welche die analytische Lösung $\hat{x}$ bestimmt, sodass (C) gilt.

(C) *Es gilt:*

$$\|\Delta_h(\hat{x}) - x_h\|_{\mathbb{G}_h} \leq M \frac{e^{Lt_f} - 1}{L} \cdot h^p \qquad \text{für } h \to 0\,,$$

wobei $M := \|\mathfrak{A}^{p+1}\|_Z \cdot e^{\|\mathfrak{A}\|_Z t_f} \left(\|x_0\|_\infty + R + t_f \|\mathfrak{B}\|_Z \|U\|_\infty\right) + \|\mathfrak{A}^p \mathfrak{B}\|_Z \cdot \|U\|_\infty$, $L := \left\|\sum_{i=1}^p \frac{h^{i-1}}{i!} \mathfrak{A}^i\right\|_Z$ und $R := e^{\|A\|_Z t_f} \left(\|x_0\|_\infty + t_f \|B\|_Z \|U\|_\infty\right)$ ist.
Die obige Konstante ist also in der Schrittweite h nach oben und unten beschränkt und von der diskreten und exakten Lösung und deren Kontrollen unabhängig.
Dabei sind die Bezeichnungen für den Einschränkungsoperator Δ_h und die Supremumsnorm für Gitterfunktionen $\|\cdot\|_{\mathbb{G}_h}$ aus Abschnitt 2.2 entnommen.

Beweis.
$(A) \quad \Longrightarrow \quad (C)$
Mit Definition 3.1.8 hat man als Verfahrensfunktion für die modifizieren RK-Verfahren

$$\Phi_h(t,x) = \sum_{i=1}^{p} \frac{h^{i-1}}{i!}\mathfrak{A}^i \cdot x + \sum_{i=1}^{p} \frac{h^{i-1}}{i!}\mathfrak{A}^{i-1}\mathfrak{B} \sum_{j=0}^{p-1} \gamma_{i,j}\bar{\mathfrak{u}}_j(t)$$

und mit Definition 3.2.1 für das Ferretti-Verfahren

$$\Phi_h(t,x) = \sum_{i=1}^{p} \frac{h^{i-1}}{i!}\mathfrak{A}^i \cdot x + \sum_{i=1}^{p} h^{i-1}\mathfrak{A}^{i-1}\mathfrak{B} \cdot \bar{\mathfrak{v}}_i(t)\,.$$

Für beide Verfahrensfunktionen gilt

$$\|\Phi_h(t,x_1) - \Phi_h(t,x_2)\|_\infty \leq \left\| \sum_{i=1}^{p} \frac{h^{i-1}}{i!}\mathfrak{A}^i \right\|_Z \|x_1 - x_2\|_\infty \,.$$

Also sind sie Lipschitz-stetig mit Lipschitz-Konstante $L = \left\|\sum_{i=1}^{p} \frac{h^{i-1}}{i!}\mathfrak{A}^i\right\|_Z$.
Da die Schrittweite $h \in (0,1)$ ist (siehe Gleichung 2.3), gilt

$$\|\mathfrak{A}\|_Z \leq L \leq \sum_{i=1}^{p} \frac{\|\mathfrak{A}\|_Z^i}{i!}\,.$$

Da L stetig von h abhängt, ist der Ausdruck $\frac{e^{Lt_f}-1}{L}$ ebenfalls stetig in h und damit auch nach oben und unten beschränkt. Außerdem ist sowohl der Zähler als auch der Nenner monoton steigend, sodass der ganze Bruch keinen großen Schwankungen unterworfen sein dürfte.
Die Konsistenz von Ordnung p für beide Verfahrensarten haben wir in Theorem 3.1.10 und 3.2.3 gezeigt. Mit Proposition 2.2.9 und Satz 2.2.10 folgt die Konvergenz von Ordnung p für beide Verfahren.
Die Folge der Kontrollvektoren $\left(\bar{\mathfrak{u}}_{1,k}^T, \ldots, \bar{\mathfrak{u}}_{p,k}^T\right)^T \in \mathcal{U}_{m,p}$ bzw. $\left(\bar{\mathfrak{v}}_{1,k}^T, \ldots, \bar{\mathfrak{v}}_{p,k}^T\right)^T \in \mathcal{I}_{m,p}$ für $k = 0, \ldots, N-1$ wird aus der vorgegebenen Steuerfunktion $\mathfrak{u}$, die die exakte Lösung $\hat{x}$ bestimmt, erzeugt gemäß der Definition des modifizierten RK-Verfahrens bzw. des Ferretti-Verfahrens.
Im Folgenden wollen wir noch die Konstante der Konvergenzordnung genau bestimmen. Dazu müssen wir aus Theorem 3.1.10 bzw. 3.2.3 das Kompaktum G so bestimmen, dass

$$\|\hat{x}(ih) - \tilde{x}_h(ih)\|_\infty \leq \left[\left\|\mathfrak{A}^{p+1}\right\|_Z \cdot e^{\|\mathfrak{A}\|_Z t_f}\left(\|G\|_\infty + t_f \|\mathfrak{B}\|_Z \|U\|_\infty\right) + \|\mathfrak{A}^p\mathfrak{B}\|_Z \cdot \|U\|_\infty\right] \cdot h^{p+1}$$

gilt für $i = 1, \ldots, N$, wobei $\hat{x}$ die analytische Lösung von (LKP) zur Anfangsbedingung x_0 und Steuerfunktion $\mathfrak{u}$ ist und $\tilde{x}_h(ih) := \hat{x}((i-1)h) + h\Phi_h((i-1)h, \hat{x}((i-1)h))$ ist. Es ist also $\tilde{x}_h(ih)$ die lokale diskrete Lösung von (LKP) zum lokalen Anfangswert $\hat{x}((i-1)h)$. Dabei muss G so gewählt werde, dass die lokale Anfangsbedingung $\hat{x}((i-1)h)$ für $i = 1, \ldots, N$ stets in G ist. Da aber $\hat{x}$ die globale Lösung von (LKP) ist, gilt nach Satz 2.3.5 die Abschätzung

$$\|\hat{x}((i-1)h)\|_\infty \leq e^{\|A\|_Z t_f} \left(\|x_0\|_\infty + t_f \|B\|_Z \|U\|_\infty\right),$$

für $i = 1, \ldots N$. Also wählen wir G als den Würfel

$$G := [(x_0)_1 - R, (x_0)_1 + R] \times [(x_0)_2 - R, (x_0)_2 + R] \times \cdots \times [(x_0)_n - R, (x_0)_n + R],$$

wobei $(x_0)_i$ die i-te Komponente der Anfangsbedingung bezeichnen soll und $R = e^{\|A\|_Z t_f}\left(\|x_0\|_\infty + t_f \|B\|_Z \|U\|_\infty\right)$ ist.
Wegen $\hat{x}(ih) - \tilde{x}_h(ih) = h \cdot \left(\frac{\hat{x}(ih) - \hat{x}((i-1)h)}{h} - \Phi_h((i-1)h, \hat{x}((i-1)h))\right)$ gilt dann für $i = 0, \ldots, N$

$$\left\| \frac{\hat{x}(ih) - \hat{x}((i-1)h)}{h} - \Phi_h((i-1)h, \hat{x}((i-1)h)) \right\|_\infty \leq$$
$$\left[\left\|\mathfrak{A}^{p+1}\right\|_Z \cdot e^{\|\mathfrak{A}\|_Z t_f}\left(\|x_0\|_\infty + R + t_f \|\mathfrak{B}\|_Z \|U\|_\infty\right) + \|\mathfrak{A}^p \mathfrak{B}\|_Z \cdot \|U\|_\infty\right] \cdot h^p,$$

denn $\|G\|_\infty = \|x_0\|_\infty + R$.
Nach [28, Satz 7.2.2.3] gilt dann

$$\|\hat{x}(ih) - x_h(ih)\|_\infty \leq h^p \cdot M \frac{e^{Lih} - 1}{L} \quad \text{für} \quad i = 1, \ldots, N,$$

mit der Konstante

$$M = \left\|\mathfrak{A}^{p+1}\right\|_Z \cdot e^{\|\mathfrak{A}\|_Z t_f}\left(\|x_0\|_\infty + R + t_f \|\mathfrak{B}\|_Z \|U\|_\infty\right) + \|\mathfrak{A}^p \mathfrak{B}\|_Z \cdot \|U\|_\infty .$$

Also haben wir insgesamt

$$\|\Delta_h(\hat{x}) - x_h\|_{\mathbb{G}_h} \leq h^p \cdot M \frac{e^{Lt_f} - 1}{L}.$$

$(B) \quad \Longrightarrow \quad (C)$
<u>modifiziertes RK-Verfahren:</u>

Im Folgenden sei $\Phi_h(\cdot)$ die Verfahrensfunktion des mod. RK-Schemas. Weiter seien $\left(\mathfrak{u}_{1,k}^T, \ldots, \mathfrak{u}_{p,k}^T\right)^T \in \mathcal{U}_{m,p}$ $(k = 0, \ldots, N-1)$ zusammen mit der diskreten Lösung $x_h(\cdot)$ von (LKP) gegeben.

Für $k = 0, \ldots, N-1$ konstruieren wir induktiv absolutstetige Funktionen $x^k : [0,h] \longrightarrow \mathbb{R}^n$, die dann unmittelbar mit der gesuchten analytischen Lösung $\hat{x}(\cdot)$ in Zusammenhang stehen. Die Kontrollvektoren $\left(\mathfrak{u}_{1,k}^T, \ldots, \mathfrak{u}_{p,k}^T\right)^T$ bestimmen eine lokale diskrete Lösung von (LKP) auf $[0,h]$ mit Anfangswert $\bar{x}_0 = x_0\,(k=0)$ bzw. $\bar{x}_0 = x^{k-1}(h)\,(k \geq 1)$, die wir hier mit $\tilde{x}_h^k\,(\cdot)$ bezeichnen. Nach Satz 3.1.10 gibt es dazu eine Kontrollfunktion $\mathfrak{u}^k \in L^1\left([0,h];U\right)$, welche eine analytische Lösung $x^k : [0,h] \longrightarrow \mathbb{R}^n$ des lokalen (LKP) festlegt, sodass

$$\left\|x^k(h) - \tilde{x}_h^k\,(ih)\right\|_\infty \leq \left[\left\|\mathfrak{A}^{p+1}\right\|_Z \cdot e^{\|\mathfrak{A}\|_Z t_f}\left(\|G\|_\infty + t_f\,\|\mathfrak{B}\|_Z\,\|U\|_\infty\right) + \|\mathfrak{A}^p\mathfrak{B}\|_Z \cdot \|U\|_\infty\right] h^{p+1} \quad (3.21)$$

ist. Dabei ist $\tilde{x}_h^k\,(ih) = \bar{x}_0 + h\Phi_h\left(\bar{x}_0; \mathfrak{u}_{1,k}, \ldots, \mathfrak{u}_{p,k}\right)$.

Außerdem muss hier das Kompaktum G so gewählt werden, dass alle lokalen Anfangsbedingungen x_0 und $x^{k-1}(h)\,(k = 1, \ldots, N-1)$ in G sind. Es sind aber diese lokalen Lösungen x^k nach Satz 2.3.5 jeweils beschränkt, deswegen ist es möglich ein solches Kompaktum zu finden.

Ferretti-Verfahren:

Seien $\left(\mathfrak{v}_{1,k}^T, \ldots, \mathfrak{v}_{p,k}^T\right)^T \in \mathcal{I}_{m,p}\;(k = 0, \ldots, N-1)$ zusammen mit $x_h(\cdot)$ gegeben. Im Folgenden sei $\Phi_h\,(\cdot)$ die Verfahrensfunktion des Ferretti-Schemas. Das Folgende geht völlig analog zu dem Vorigen, deswegen soll es jetzt kürzer notiert werden.

Für $k = 0, \ldots, N-1$ konstruieren wir wieder induktiv absolutstetige Funktionen $x^k : [0,h] \longrightarrow \mathbb{R}^n$. Die Kontrollvektoren $\left(\mathfrak{v}_{1,k}^T, \ldots, \mathfrak{v}_{p,k}^T\right)^T$ bestimmen eine lokale diskrete Lösung von (LKP) auf $[0,h]$ mit Anfangswert $\bar{x}_0 = x_0\,(k=0)$ bzw. $\bar{x}_0 = x^{k-1}(h)\,(k \geq 1)$, die wir hier mit $\tilde{x}_h^k\,(\cdot)$ bezeichnen. Nach Satz 3.1.10 gibt es dazu eine Kontrollfunktion $\mathfrak{u}^k \in L^1\left([0,h];U\right)$, welche eine analytische Lösung $x^k : [0,h] \longrightarrow \mathbb{R}^n$ des lokalen (LKP) festlegt, sodass

$$\left\|x^k(h) - \tilde{x}_h^k\,(ih)\right\|_\infty \leq \left[\left\|\mathfrak{A}^{p+1}\right\|_Z \cdot e^{\|\mathfrak{A}\|_Z t_f}\left(\|G\|_\infty + t_f\,\|\mathfrak{B}\|_Z\,\|U\|_\infty\right) + \|\mathfrak{A}^p\mathfrak{B}\|_Z \cdot \|U\|_\infty\right] h^{p+1} \quad (3.22)$$

ist.

Wieder muss hier das Kompaktum G so gewählt werden, dass alle lokalen Anfangsbedingungen x_0 und $x^{k-1}(h)\,(k = 1, \ldots, N-1)$ in G sind.

Beide Verfahrensarten:

Das Folgende gilt für beide Verfahrensarten. Deswegen ist jetzt Φ_h eine der beiden Verfahrensfunktionen und die lokalen analytischen Lösungen x^k seien bezüglich einer der beiden Verfahrenarten konstruiert.

Jetzt definieren wir

$$\hat{x}(t) := x^k(t - kh) \text{ für } t \in [kh, (k+1)h]\;,\; k = 0, \ldots, N-1\,.$$

Dann ist $\hat{x}$ auf ganz I definiert, offenbar stetig und es ist $\hat{x}(0) = x_0$. Weiter ist $\hat{x}$ nach Satz 1.8.4 absolutstetig, denn die Einschränkungen auf $[kh, (k+1)h]$ sind absolutstetig.
Analog definieren wir

$$\mathfrak{u}(t) := \mathfrak{u}^k(t - kh) \text{ für } t \in [kh,\ (k+1)h]\ ,\ k = 0, \ldots, N-1\,.$$

Dann ist $\mathfrak{u}$ ebenfalls auf ganz I definiert. Weiter ist $\mathfrak{u}(t) \in U\ \forall t \in I$. Für eine Borel-Menge $A \subset U$ gilt dann $\mathfrak{u}^{-1}(A) = \bigcup_{k=0}^{N} \left(\mathfrak{u}^k\right)^{-1}(A)$. Also muss $\mathfrak{u}$ messbar sein, da ja die $\mathfrak{u}^k$ messbar sind. Insgesamt ist $\mathfrak{u} \in L^1(I;U)$, denn, da I ein kompaktes Intervall ist und U ebenfalls kompakt ist, muss $\|\mathfrak{u}\|_{L^1(I;U)} < \infty$ sein.
Nach Definition ist klar, dass $\hat{x}$ eine Lösung von (LKP) zur Kontrollfunktion $\mathfrak{u}$ mit Anfangswert x_0 auf dem Intervall I ist.
Nach Satz 2.3.5 ist $\hat{x}$ unabhängig von der Kontrollfunktion $\mathfrak{u}$ beschränkt. Wir können also wie in der vorigen Richtung und auch mit der gleichen Begründung das Kompaktum G wählen als

$$G := [(x_0)_1 - R, (x_0)_1 + R] \times [(x_0)_2 - R, (x_0)_2 + R] \times \cdots \times [(x_0)_n - R, (x_0)_n + R]\,,$$

wobei R genau wie vorher ist. Dann gilt für $k = 1, \ldots, N$

$$\hat{x}(kh) = x^{k-1}(h) \in G \quad \text{und} \quad x_0 \in G.$$

Es ist dann M wieder die Konstante der lokalen Abschätzung (3.21) bzw. (3.22) für dieses speziellen Kompaktum G.
Es bleibt noch zu zeigen, dass dieses $\hat{x}$ die Aussage (C) erfüllt. Es gilt

$$\begin{aligned}
&\|\hat{x}((i+1)h) - x_h((i+1)h)\|_\infty \\
\leq\ &\|\hat{x}((i+1)h) - [\hat{x}(ih) + h\Phi_h(\hat{x}(ih)\,; \mathfrak{u}_{0,i}, \ldots, \mathfrak{u}_{p-1,i})]\|_\infty \\
&+ \|[\hat{x}(ih) + h\Phi_h(\hat{x}(ih)\,; \mathfrak{u}_{0,i}, \ldots, \mathfrak{u}_{p-1,i})] - x_h((i+1)h)\|_\infty
\end{aligned}$$

Für den ersten Summanden der rechten Seite erinnern wir uns an die Definition von $\hat{x}$ und erhalten mittels (3.21) bzw. (3.22)

$$\begin{aligned}
&\|\hat{x}((i+1)h) - [\hat{x}(ih) + h\Phi(\hat{x}(ih)\,; \mathfrak{u}_{0,i}, \ldots, \mathfrak{u}_{p-1,i})]\|_\infty \\
=\ &\left\|x^i(h) - \left[x^{i-1}(h) + h\Phi\left(x^{i-1}(h)\,; \mathfrak{u}_{0,i}, \ldots, \mathfrak{u}_{p-1,i}\right)\right]\right\|_\infty \\
\leq\ &M \cdot h^{p+1} \qquad (i = 0, \ldots, N-1)\,.
\end{aligned}$$

In der ersten Richtung des Beweises haben wir bereits gesehen das $\Phi_h(t,x)$ Lipschitz-stetig bzgl. x ist. Genauso zeigt man, dass die Verfahrensfunktion $\Phi_h(x; \mathfrak{u}_1, \ldots, \mathfrak{u}_p)$ des mod. RK-Schemas bzw. Ferretti-Schemas ebenfalls Lipschitz-stetig bzgl. x ist mit

Lipschitzkonstante L. Dann kann man den zweiten Summanden so abschätzen:

$$\begin{aligned}
&\left\|\left[\hat{x}\left(ih\right)+h\Phi_h\left(\hat{x}\left(ih\right);\mathfrak{u}_{0,i},\ldots,\mathfrak{u}_{p-1,i}\right)\right]-x_h\left((i+1)h\right)\right\|_\infty \\
=&\left\|\left[\hat{x}\left(ih\right)+h\Phi_h\left(\hat{x}\left(ih\right);\mathfrak{u}_{0,i},\ldots,\mathfrak{u}_{p-1,i}\right)\right]-\left[x_h\left(ih\right)+\Phi_h\left(x_h\left(ih\right);\mathfrak{u}_{0,i},\ldots,\mathfrak{u}_{p-1,i}\right)\right]\right\|_\infty \\
\leq&\left\|\hat{x}\left(ih\right)-x_h\left(ih\right)\right\|_\infty+h\left\|\Phi_h\left(\hat{x}\left(ih\right);\mathfrak{u}_{0,i},\ldots,\mathfrak{u}_{p-1,i}\right)-\Phi_h\left(x_h\left(ih\right);\mathfrak{u}_{0,i},\ldots,\mathfrak{u}_{p-1,i}\right)\right\|_\infty \\
\leq&\left(1+hL\right)\left\|\hat{x}\left(ih\right)-x_h\left(ih\right)\right\|_\infty
\end{aligned}$$

Insgesamt haben wir dann

$$\left\|\hat{x}\left((i+1)h\right)-x_h\left((i+1)h\right)\right\|_\infty\leq\left(1+hL\right)\left\|\hat{x}(ih)-x_h(ih)\right\|_\infty+Mh^{p+1}\,.$$

Daraus folgt nach [28, Hilfssatz 7.2.2.2.]

$$\begin{aligned}
\left\|\hat{x}\left(kh\right)-x_h\left(kh\right)\right\|_\infty &\leq e^{hL}\left\|\hat{x}(0)-x_h(0)\right\|_\infty+\frac{e^{khL}-1}{hL}M\cdot h^{p+1} \\
&= e^{hL}\left\|x_0-x_0\right\|_\infty+\frac{e^{khL}-1}{L}M\cdot h^{p} \\
&\leq \frac{e^{t_fL}-1}{L}M\cdot h^{p}\,.
\end{aligned}$$

Die letzte Ungleichung gilt wegen $kh\in I=[0,t_f]$ für $k=0,\ldots,N$.
Also haben wir insgesamt

$$\left\|\Delta_h\left(\hat{x}\right)-x_h\right\|_{\mathbb{G}_h}\leq h^p\cdot M\frac{e^{Lt_f}-1}{L}\,.$$

Dabei ist die Konstante $M\frac{e^{Lt_f}-1}{L}$ unabhängig von den vorgegebenen Kontrollvektoren und der dazu bestimmten Kontrollfunktion. Ebenso ist sie dann natürlich unabhängig von der diskreten Lösung x_h und der analytischen Lösung $\hat{x}$. Außerdem ist sie noch oben und unter beschränkt in der Schrittweite h.
Damit ist alles gezeigt. □

Die numerischen Tests der mengenwertigen Verfahren aus dem nächsten Abschnitt haben das nun folgende Resultat angeregt. Es handelt sich dabei um einen wichtigen Zusammenhang zwischen den Ferretti-Verfahren und den modifizierten Runge-Kutta-Verfahren.

Satz 3.2.5. *Für das lineare Kontrollproblem (LKP) auf $I=[0,t_f]$ mit Anfangsbedingung x_0 sei eine Kontrollfunktion $\mathfrak{u}\in L^1\left(I;U\right)$ gegeben. Weiter sei die diskrete Lösung x_h^R mit einem modifiziertes RK-Verfahren der Ordnung p und die diskrete Lösung x_h^F mit dem Ferretti-Verfahren der Ordnung p konstruiert. Dann gilt:*

$$x_h^R=x_h^F\,.$$

Für jede Konvergenzordnung gibt es im Grunde nur ein Verfahren. Unter allen Verfahren der gleichen Konvergenzordnung ist das Ferretti-Verfahren ausgezeichnet.

Beweis. Für (LKP) sei $\mathfrak{u} \in L^1(I;U)$ vorgegeben. Es reicht, den ersten Zeitschritt zu betrachten, weil alle diese Verfahren Einschrittverfahren sind, also nur den letzten Schritt bei der Berechnung der nächsten Approximation verwerten.
Das Ferretti-Verfahren der Ordnung p berechnet für diese Steuerfunktion $\mathfrak{u}$ Kontrollvektoren $\left(\bar{\mathfrak{v}}_1^T, \ldots, \bar{\mathfrak{v}}_p^T\right)^T \in \mathcal{I}_{m,p}$ und ein modifiziertes RK-Verfahren der Ordnung p erzeugt die Kontrollvektoren $\left(\bar{\mathfrak{u}}_1^T, \ldots, \bar{\mathfrak{u}}_p^T\right)^T \in \mathcal{U}_{m,p}$. Nach Definition dieser Kontrollvektoren gilt offensichtlich der Zusammenhang

$$\begin{pmatrix} \bar{\mathfrak{u}}_1 \\ \vdots \\ \bar{\mathfrak{u}}_p \end{pmatrix} = \Gamma_{m,p}^{-1} \Delta_{m,p} \begin{pmatrix} \bar{\mathfrak{v}}_1 \\ \vdots \\ \bar{\mathfrak{v}}_p \end{pmatrix} . \tag{3.23}$$

Die nach dem mod. RK-Verfahren berechnete diskrete Lösung hat nach Definition 3.1.8 der Verfahrensfunktion und (3.13) auf Seite 61 folgende Darstellung

$$\begin{aligned} x_h^R(h) &= \sum_{i=0}^{p} \frac{h^i}{i!} \mathfrak{A}^i \cdot x_0 + \left[h\mathfrak{B} \mid h^2 \mathfrak{A}\mathfrak{B} \mid \ldots \mid h^p \mathfrak{A}^{p-1}\mathfrak{B} \right] \Delta_{m,p}^{-1} \Gamma_{m,p} \begin{pmatrix} \bar{\mathfrak{u}}_1 \\ \vdots \\ \bar{\mathfrak{u}}_p \end{pmatrix} \\ &\overset{(3.23)}{=} \sum_{i=0}^{p} \frac{h^i}{i!} \mathfrak{A}^i \cdot x_0 + \left[h\mathfrak{B} \mid h^2 \mathfrak{A}\mathfrak{B} \mid \ldots \mid h^p \mathfrak{A}^{p-1}\mathfrak{B} \right] \begin{pmatrix} \bar{\mathfrak{v}}_1 \\ \vdots \\ \bar{\mathfrak{v}}_p \end{pmatrix} \\ &= \sum_{i=0}^{p} \frac{h^i}{i!} \mathfrak{A}^i \cdot x_0 + \sum_{i=1}^{p} h^i \mathfrak{A}^{i-1} \mathfrak{B} \cdot \bar{\mathfrak{v}}_i \\ &= x_h^F(h) \end{aligned}$$

Die dritte Gleichung (von oben) ist genau die definierende Gleichung des Ferretti-Verfahrens. Und die zweite Gleichung ist eine Matrixdarstellung dieses Verfahrens. An der ersten und zweiten Gleichung sieht man, dass man ein modifiziertes RK-Verfahren als lineare Transformation des Ferretti-Verfahrens der gleichen Ordnung verstehen kann. Die Kontrollvektoren sind dann so definiert, dass diese Transformation wieder rückgängig gemacht wird.
Das bedeutet, dass alle modifizierten RK-Verfahren der Ordnung p und das Ferretti-Verfahren der Ordnung p, zu einer vorgegebenen Kontrollfunktion stets die gleiche diskrete Lösung x_h liefern. Die Verfahren sind also identisch. □

Bemerkung 3.2.6.

(i) Im Grunde ist die Idee für das Ferretti-Verfahren ganz trivial. Man nimmt einfach die Entwicklung aus Satz 3.1.1 und schneidet den Restterm ab. Dadurch hat man eine lokale Approximation an die exakte Lösung, die von der Ordnung des Restterms ist.

(ii) Eigentlich ist es unnötig komplizierend und überflüssig die Entwicklung der exakten Lösung aus Satz 3.1.1 mit der Darstellung eines Runge-Kutta Verfahren aus Satz 3.1.3 zu vergleichen und die Kontrollvektoren abzugleichen. Im Grunde gibt es zu jeder Ordnung nur ein Verfahren, nämlich das entsprechende Ferretti-Verfahren. Jedes andere Verfahren ist lediglich eine lineare Transformation dieses Verfahrens, wobei aber dann die diskreten Kontrollen so gewählt werden müssen, dass diese lineare Transformation wieder rückgängig gemacht wird.

(iii) Im Gegensatz zu den modifizierten RK-Verfahren gilt für die Ferretti-Verfahren keine Ordnungsbeschränkung.

(iv) Die Äquivalenz dieser beiden Verfahrensarten ist mit der Darstellung der Verfahren in Matrixschreibweise leicht zu sehen. In dem Artikel von Ferretti [13] ist dies viel schwerer, da der ganze Betrachtungsweise auf Runge-Kutta-Verfahren ausgerichtet ist.

(v) Das Ferretti-Verfahren kann nicht als explizites RK-Verfahren interpretiert werden. Denn dann müsste man die definierende Matrix $\Gamma_p = \Delta_p$ wählen können, sodass dann $\Delta_{m,p}^{-1}\Gamma_{m,p} = \mathfrak{E}_{mp}$ wäre (siehe den letzten Beweis). Jedoch in der ersten Zeile von Γ_p stehen immer die Gewichte des jeweiligen RK-Verfahrens, d.h. $\gamma_{1,j} = b_j$. Also müsste $b_1 = 1$ und $b_j = 0$ für $j = 2, \dots, p$ sein. Es müsste also ein explizites RK-Verfahren geben, dass von der Ordnung p konvergiert, aber dazu nur die erste Stufe verwendet. Das ist nicht möglich wenn $p \geq 2$ ist. (siehe [10, Theorem 370A, p. 258]).

3.3. Approximation der erreichbaren Menge

Ziel diese Abschnittes ist es die punktwertigen Verfahren, die in den beiden vorigen Abschnitten gewonnen wurden, zu mengenwertigen Verfahren auszubauen um die erreichbare Menge eines linearen Kontrollproblems zu approximieren. Wie wir in Abschnitt 2.3 (Satz 2.3.8) gesehen haben ist sie identisch mit der erreichbaren Menge der zugeordneten linearen Differentialinklusion. In der Literatur wird häufig von der linearen Differentialinklusion ausgegangen, jedoch war das hier wegen der Entwicklung der punktwertigen Verfahren nicht möglich bzw. nicht angemessen.

Die mengenwertigen Verfahren gewinnt man nun einfach dadurch, dass man die diskreten Kontrollvektoren der Verfahren ersetzt durch die diskrete Kontrollmenge $\mathcal{U}_{m,p}$ bzw. die Auswahlmenge $\mathcal{I}_{m,p}$. Um diese Verfahren definieren zu können müssen wir die Matrixschreibweisen der punktwertigen Verfahren gebrauchen. Es sei daran erinnert, dass $\mathcal{U}_{m,p}$ und $\mathcal{I}_{m,p}$ Spaltenvektoren enthalten.

Weiter möge sich der Leser an die Bezeichnungen aus Kapitel 2 für die erreichbare Menge $\mathcal{R}(t, x_0)$ von (LKP) und für das äquidistante Gitter $\mathbb{G}_h = \{0, h, \dots, Nh = t_f\}$ erinnern (bei dem linearen Kontrollproblem ist $t_0 = 0$).

Definition 3.3.1. Wie bei den punktwertigen Einschrittverfahren gelte für die Schrittweite $h = \frac{t_f}{N}$, wobei $N \in \mathbb{N}$ die Anzahl der Iterationen des Verfahrens ist.
Die folgenden mengenwertigen Einschrittverfahren dienen zur Konstruktion einer mengenwertigen Gitterfunktion $\mathcal{R}_h : \mathbb{G}_h \Longrightarrow \mathbb{R}^n$, wobei $\mathcal{R}_h(ih, x_0)$ *diskrete erreichbaren Menge* (zum Zeitpunkt ih) heiße. Sie soll die erreichbare Menge $\mathcal{R}(ih, x_0)$ von dem linearen Kontrollproblem bzw. der linearen Differentialinklusion approximieren.
Diese mengenwertigen Einschrittverfahren sind gegeben durch

$$\begin{aligned} \mathcal{R}_h(0, x_0) &= \{x_0\} \\ \mathcal{R}_h((i+1)h, x_0) &= \mathfrak{C} \cdot \mathcal{R}_h(ih, x_0) + \mathfrak{D} \cdot \mathcal{U} \qquad (i = 0, \ldots, N-1) \end{aligned}$$

mit der $n \times n$-Matrix $\mathfrak{C} := \sum_{i=0}^{p} \frac{h^i}{i!} \mathfrak{A}^i$, einer $n \times mp$-Matrix $\mathfrak{D}$ und der Menge $\mathcal{U} \subset \mathbb{R}^n$.
Es sei $\Gamma_{m,p}$ die definierende Matrix eines mod. RK-Verfahrens der Ordnung p. Dann ist für das dazugehörige *mengenwertige RK-Verfahren* die Matrix

$$\mathfrak{D} := \left[h\mathfrak{B} \mid h^2\mathfrak{A}\mathfrak{B} \mid \ldots \mid h^p\mathfrak{A}^{p-1}\mathfrak{B}\right] \Delta_{m,p}^{-1} \Gamma_{m,p}$$

und die Menge $\mathcal{U} := \mathcal{U}_{m,p}$.
Und für das *mengenwertige Ferretti-Verfahren* der Ordnung p ist die Matrix

$$\mathfrak{D} := \left[h\mathfrak{B} \mid h^2\mathfrak{A}\mathfrak{B} \mid \ldots \mid h^p\mathfrak{A}^{p-1}\mathfrak{B}\right]$$

und die Menge $\mathcal{U} := \mathcal{I}_{m,p}$.

Mit den punktwertigen Ergebnissen der vorigen Abschnitte können wir nun unmittelbar Konvergenz der diskreten erreichbaren Menge gegen die (analytische) erreichbare Menge zeigen.

Theorem 3.3.2. *Für das lineare Kontrollproblem (LKP) auf $I = [0, t_f]$ mit Anfangsbedingung x_0 sei ein mengenwertiges RK-Verfahren der Ordnung p oder das mengenwertige Ferretti-Verfahren der Ordnung p gegeben. D.h. die diskrete erreichbare Menge $\mathcal{R}_h(\cdot, x_0)$ wird mit einem der beiden Verfahren konstruiert. Dann gilt für alle $t \in \mathbb{G}_h$*

$$d_H\left(\mathcal{R}_h(t, x_0), \mathcal{R}(t, x_0)\right) \leq \sqrt{n} M \frac{e^{Lt_f} - 1}{L} \cdot h^p ,$$

wobei M und L die Konstanten aus Theorem 3.2.4 sind, welche nach oben und unten beschränkt sind in der Schrittweite h.

Beweis. Der Hausdorff-Abstand ist definiert als (siehe Abschnitt 1.2)

$$\mathrm{d}_H\left(\mathcal{R}_h(t, x_0), \mathcal{R}(t, x_0)\right) = \max\left\{\mathrm{d}\left(\mathcal{R}_h(t, x_0), \mathcal{R}(t, x_0)\right), \mathrm{d}\left(\mathcal{R}(t, x_0), \mathcal{R}_h(t, x_0)\right)\right\} .$$

Und nach Definition des einseitigen Hausdorff-Abstandes gilt für $t = kh \quad (k = 0, \ldots, N)$

$$\begin{aligned} \mathrm{d}\left(\mathcal{R}_h\left(t, x_0\right), \mathcal{R}\left(t, x_0\right)\right) &= \sup_{x_h(t) \in \mathcal{R}_h(t, x_0)} \operatorname{dist}\left(x_h\left(t\right), \mathcal{R}\left(t, x_0\right)\right) \\ &= \sup_{x_h(t) \in \mathcal{R}_h(t, x_0)} \inf_{\hat{x}(t) \in \mathcal{R}(t, x_0)} \left\| x_h\left(t\right) - \hat{x}\left(t\right) \right\|_2 \\ &\leq \sup_{x_h(t) \in \mathcal{R}_h(t, x_0)} \inf_{\hat{x}(t) \in \mathcal{R}(t, x_0)} \sqrt{n} \left\| x_h\left(t\right) - \hat{x}\left(t\right) \right\|_\infty \\ &\leq \sup_{x_h(t) \in \mathcal{R}_h(t, x_0)} \sqrt{n} M \frac{e^{Lt_f} - 1}{L} \cdot h^p = \sqrt{n} M \frac{e^{Lt_f} - 1}{L} \cdot h^p \,, \end{aligned}$$

wobei die zweite Ungleichung nach Theorem 3.2.4(B) gilt. Dabei sind die Konstanten L und M unabhängig von der diskreten Lösung x_h und der analytischen Lösung $\hat{x}$ (siehe Theorem 3.2.4). Sie gelten auch für jeden Zeitpunkt t des Gitters $\mathbb{G}_h = \{kh \mid k = 0, \ldots N\}$.

Nach dem gleichen Theorem, aber mit Teil (A), folgt für den anderen einseitigen Hausdorff-Abstand

$$\begin{aligned} d\left(\mathcal{R}\left(t, x_0\right), \mathcal{R}_h\left(t, x_0\right)\right) &= \sup_{\hat{x}(t) \in \mathcal{R}(t, x_0)} dist\left(\hat{x}\left(t\right), \mathcal{R}_h\left(t, x_0\right)\right) \\ &= \sup_{\hat{x}(t) \in \mathcal{R}(t, x_0)} \inf_{x_h(t) \in \mathcal{R}_h(t, x_0)} \left\| \hat{x}\left(t\right) - x_h\left(t\right) \right\|_2 \\ &\leq \sup_{\hat{x}(t) \in \mathcal{R}(t, x_0)} \sqrt{n} M \frac{e^{Lt_f} - 1}{L} \cdot h^p = \sqrt{n} M \frac{e^{Lt_f} - 1}{L} \cdot h^p \,, \end{aligned}$$

für alle $t \in \mathbb{G}_h$.

Insgesamt gilt dann für alle $t \in \mathbb{G}_h$

$$d\left(\mathcal{R}_h\left(t, x_0\right), \mathcal{R}\left(t, x_0\right)\right) \leq \sqrt{n} M \frac{e^{Lt_f} - 1}{L} \cdot h^p \,.$$

□

Kapitel 4.

Nichtlineare Kontrollprobleme

Dieses Kapitel stellt einen Einschub dar. Es soll untersucht werden, ob sich die Konzepte von Kapitel 3 auf das nichtlineare Kontrollproblem (NKP), welches in Abschnitt 2.3 vorgestellt wurde, übertragen lassen. Dabei werden wir dieses Problem wieder als ein Anfangswertproblem betrachten. Wir suchen also zu einer vorgegebenen Steuerfunktion $\mathfrak{u} \in L^1(I;U)$ eine Lösung des Anfangswertproblems

$$x'(t) = \mathfrak{a}(x(t)) + \mathfrak{B}\mathfrak{u}(t) \qquad \forall' t \in I \tag{4.1}$$

$$x(0) = x_0\,, \tag{4.2}$$

wobei $U \subset \mathbb{R}^m$ der kompakte Steuerbereich, $\mathfrak{a} \in W^{2,\infty}(\mathbb{R}^n)^n$ ein Vektorfeld, $I = [0, t_f]$ das Zeitintervall, $\mathfrak{B} \in \mathbb{R}^{n\times m}$ eine Matrix und $x_0 \in \mathbb{R}^n$ die Anfangsbedingung ist.

Im ersten Abschnitt werden wir eine lokale Approximation der Lösung von Ordnung 3 herleiten und diese in Verbindung mit expliziten Runge-Kutta Verfahren der Ordnung 2 bringen. Bei dem so entwickelten Verfahren wird eine spezielle Auswahl der Kontrollvektoren ein große Rolle spielen. Dieser Abschnitt stellt im wesentlichen eine Ausarbeitung von Kapitel 2 des Artikels von Ferretti [13] dar.

Im zweiten Abschnitt soll darüber hinaus noch gezeigt werden warum eine höhere lokale Approximation d.h. von Ordnung > 3 mit dieser hier vorgestellten Methode kaum möglich ist und warum die Übertragung der gewonnen punktwertigen Approximationen in mengenwertige Verfahren auf ineffiziente Verfahren führt. Es wird also auf die Nachteile dieses Zugangs für das nichtlineare Kontrollproblem eingegangen.

4.1. Theorie von Ferretti

In diesem Abschnitt werden die Konzepte aus Kapitel 3 auf den nichtlinearen Fall angewandt.

Zunächst werden wir eine Approximation der Lösung von (4.1) und (4.2) herleiten, die wir später weiter verwenden. Dabei werden an des Vektorfeld $\mathfrak{a}$, an die Kontrollfunktion $\mathfrak{u}$ und an die Lösung x die schwachen Voraussetzungen aus Problem 2.3.10 gestellt. Deswegen ist auch die Lösung nur in der Klasse der absolutstetigen Funktionen zu suchen.

Satz 4.1.1. *Zu einer vorgegebenen Steuerfunktion $\mathfrak{u} \in L^1(I;U)$ genügt die exakte Lösung $\hat{x}(\cdot)$ von (NKP) im ersten Zeitschritt folgender Entwicklung*

$$\hat{x}(h) = x_0 + h \cdot \mathfrak{a}(x_0) + \frac{h^2}{2} \cdot \mathfrak{J}_{\mathfrak{a}}(x_0) \cdot \mathfrak{a}(x_0) + \mathfrak{B} \int_0^h \mathfrak{u}(s)\, ds + \mathfrak{J}_{\mathfrak{a}}(x_0) \cdot \mathfrak{B} \int_0^h \int_0^s \mathfrak{u}(r)\, dr\, ds + \mathfrak{r}(h) \quad (4.3)$$

für $h \in I$. Dabei ist $\mathfrak{r} : [0,\infty) \longrightarrow \mathbb{R}^n$ eine Restfunktion, die der Abschätzung

$$\|\mathfrak{r}(h)\|_\infty \leq \left(\frac{n^2}{6} \|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n} M^2 + \frac{n}{6} \|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n} M \right) \cdot h^3$$

mit einer Konstanten $M := \|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n} + \|\mathfrak{B}\|_Z \|U\|_\infty$ für $h \to 0$ genügt.

Beweis. Bevor wir die eigentliche Formel zeigen können, müssen wir etwas Vorarbeit leisten.
Da $\hat{x}(\cdot)$ als absolutstetig vorausgesetzt ist (siehe Problem 2.3.10), gilt nach Theorem 1.8.5 folgende Darstellung

$$\hat{x}(h) = \hat{x}(0) + \int_0^h \hat{x}'(t)\, dt = x_0 + \int_0^h \hat{x}'(t)\, dt\,. \quad (4.4)$$

Nach Bemerkung 2.3.11 ist $\mathfrak{a}$ zumindest auf dem Zustandsraum Ω stetig differenzierbar. Deswegen gilt

$$\|\mathfrak{a}(\mathfrak{v})\|_\infty \leq \|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n} \quad \forall \mathfrak{v} \in \Omega\,.$$

Insbesondere gilt auch $\|\mathfrak{a}(\hat{x}(t))\|_\infty \leq \|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n} \quad \forall t \in I$, da $\hat{x}$ den Zustandsraum Ω nach Satz 2.3.12 nicht verlässt.
Wegen $\mathfrak{u}(t) \in U$ und weil U kompakt ist, gilt $\|\mathfrak{u}(t)\|_\infty \leq \|U\|_\infty < \infty\ \forall t \in I$, und damit ist auch $\hat{x}'$ durch eine Konstante $M > 0$ beschränkt, die sich so ergibt:

$$\|\hat{x}'(t)\|_\infty \overset{(4.1)}{=} \|\mathfrak{a}(\hat{x}(t)) + \mathfrak{B}\mathfrak{u}(t)\|_\infty \quad (4.5)$$

$$\leq \|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n} + \|\mathfrak{B}\|_Z \|U\|_\infty =: M \quad (4.6)$$

Damit folgt nun unmittelbar

$$\left\| \int_0^h \hat{x}'(t)\, dt \right\|_\infty \leq h \cdot M \quad (4.7)$$

Desweiteren benötigen wir eine Taylorentwicklung 0. Ordnung von dem Vektorfeld $\mathfrak{a}$ im Zustandsraum Ω. Da $\mathfrak{a}$ hier stetig differenzierbar ist, sieht die Taylorentwicklung nach [21, Abschnitt 2.4, Satz] so aus

$$\mathfrak{a}\left(x_0 + \int_0^h \hat{x}'(t)\, dt \right) = \mathfrak{a}(x_0) + \mathfrak{r}_1(h)\ , \quad (4.8)$$

wobei $\mathfrak{r}_1 : [0,\infty) \longrightarrow \mathbb{R}^n$ eine stetige Restfunktion ist, definiert durch

$$\mathfrak{r}_1(h) := \mathfrak{a}\left(x_0 + \int_0^h \hat{x}'(t)\,dt\right) - \mathfrak{a}(x_0) = \mathfrak{J}_{\mathfrak{a}}(\xi)\int_0^h \hat{x}'(t)\,dt$$

mit einer Zwischenstelle $\xi \in \Omega$. Mit (4.7) und weil $\|\mathfrak{J}_{\mathfrak{a}}(\mathfrak{v})\|_Z \leq n\,\|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n} \quad \forall \mathfrak{v} \in \Omega$ (siehe Beispiel 1.6.4) ist, erhalten wir daraus

$$\|\mathfrak{r}_1(h)\|_\infty \leq \left(n\,\|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n}\,M\right)\cdot h \quad \forall h \in [0,1]. \tag{4.9}$$

Für eine Taylorentwicklung 1. Ordnung greifen wir auf Satz 1.7.2 zurück und erhalten

$$\mathfrak{a}\left(x_0 + \int_0^h \hat{x}'(t)\,dt\right) = \mathfrak{a}(x_0) + \mathfrak{J}_{\mathfrak{a}}(x_0)\cdot\int_0^h \hat{x}'(t)\,dt + \mathfrak{r}_2(h)\,, \tag{4.10}$$

wobei $\mathfrak{r}_2 : [0,\infty) \longrightarrow \mathbb{R}^n$ eine stetige Restfunktion ist, definiert durch

$$\mathfrak{r}_2(h) := \mathfrak{a}\left(x_0 + \int_0^h \hat{x}'(t)\,dt\right) - \mathfrak{a}(x_0) - \mathfrak{J}_{\mathfrak{a}}(x_0)\cdot\int_0^h \hat{x}'(t)\,dt\,. \tag{4.11}$$

Mit Satz 1.7.2 und (4.7) ist diese Restfunktion beschränkt durch

$$\|\mathfrak{r}_2(h)\|_\infty \leq \left(\frac{n^2}{2}\,\|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n}\,M^2\right)\cdot h^2\,.$$

Damit können wir nun das eigentliche Resultat zeigen. Wir beginnen mit (4.4)

$$\begin{aligned}
\hat{x}(h) &= x_0 + \int_0^h \hat{x}'(t)\,dt \overset{(4.1)}{=} x_0 + \int_0^h \mathfrak{a}(\hat{x}(t)) + \mathfrak{B}\mathfrak{u}(t)\,dt \\
&\overset{(4.4)}{=} x_0 + \int_0^h \mathfrak{a}\left(x_0 + \int_0^h \hat{x}'(t)\,dt\right)\,dt + \mathfrak{B}\int_0^h \mathfrak{u}(t)\,dt \\
&\overset{(4.10)}{=} x_0 + \int_0^h \mathfrak{a}(x_0) + \mathfrak{J}_{\mathfrak{a}}(x_0)\cdot\left(\int_0^t \hat{x}'(s)\,ds\right) + \mathfrak{r}_2(t)\;dt + \mathfrak{B}\int_0^h \mathfrak{u}(t)\,dt
\end{aligned}$$

und fahren fort mit

$$
\begin{aligned}
& x_0 + \int_0^h \mathfrak{a}(x_0) + \mathfrak{J}_{\mathfrak{a}}(x_0) \cdot \left(\int_0^t \hat{x}'(s)\, ds \right) + \mathfrak{r}_2(t)\ dt + \mathfrak{B} \int_0^h \mathfrak{u}(t)\, dt \qquad (4.12) \\
\overset{(4.1)}{=}\ & x_0 + \mathfrak{a}(x_0)\, h + \mathfrak{J}_{\mathfrak{a}}(x_0) \cdot \int_0^h \int_0^t \mathfrak{a}(\hat{x}(s)) + \mathfrak{B}u(s)\, ds\, dt \\
& \qquad + \mathfrak{B} \int_0^h \mathfrak{u}(t)\, dt + \int_0^h \mathfrak{r}_2(t)\ dt \\
\overset{(4.4)}{=}\ & x_0 + \mathfrak{a}(x_0)\, h + \mathfrak{J}_{\mathfrak{a}}(x_0) \cdot \int_0^h \int_0^t \mathfrak{a}\left(x_0 + \int_0^s \hat{x}'(r) dr \right) ds\, dt \\
& \qquad + \mathfrak{B} \int_0^h \mathfrak{u}(t)\, dt + \mathfrak{B} \int_0^h \int_0^t \mathfrak{u}(s)\, ds\, dt + \int_0^h \mathfrak{r}_2(t)\ dt \\
\overset{(4.8)}{=}\ & x_0 + \mathfrak{a}(x_0)\, h + \mathfrak{J}_{\mathfrak{a}}(x_0) \cdot \int_0^h \int_0^t \mathfrak{a}(x_0) + \mathfrak{r}_1(s)\ ds\, dt \\
& \qquad + \mathfrak{B} \int_0^h \mathfrak{u}(t)\, dt + \mathfrak{B} \int_0^h \int_0^t \mathfrak{u}(s)\, ds\, dt + \int_0^h \mathfrak{r}_2(t)\ dt \\
=\ & x_0 + \mathfrak{a}(x_0)\, h + \mathfrak{J}_{\mathfrak{a}}(x_0) \cdot \mathfrak{a}(x_0) \frac{h^2}{2} \\
& \qquad + \mathfrak{B} \int_0^h \mathfrak{u}(t)\, dt + \mathfrak{B} \int_0^h \int_0^t \mathfrak{u}(s)\, ds\, dt + \mathfrak{r}(h)\ , \qquad (4.13)
\end{aligned}
$$

wobei $\mathfrak{r} : [0, \infty) \longrightarrow \mathbb{R}^n$, $\mathfrak{r}(h) := \int_0^h \mathfrak{r}_2(t)\ dt + \int_0^h \int_0^t \mathfrak{r}_1(s)\ ds\, dt$ die finale Restfunktion ist. Mit (4.11) und (4.9) kann diese Funktion so

$$
\|\mathfrak{r}(h)\|_\infty \leq \left(\frac{n^2}{6} \|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n} M^2 + \frac{n}{6} \|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n} M \right) \cdot h^3
$$

abgeschätzt werden.
Damit ist alles gezeigt. □

Bemerkung 4.1.2.

(i) Dieses Resultat kann auch mit Hilfe der Picard-Iteration oder mit Hilfe von Volterra- oder Fliß-Entwicklungen gezeigt werden.

(ii) In dem Doppelintegral $\int_0^h \int_0^s \mathfrak{u}(r)\, dr\, ds$ wird über die obere Grenze des inneren Integrals integriert. Die beiden Integrale sind also nicht vertauschbar, da das äußere Integral von dem inneren abhängt.

(iii) Dieser Satz 4.1.1 zeigt eine lokale Entwicklung der Lösung von dem Anfangswertproblem (4.1)+(4.2) von Ordnung 3. Aber dies ist nur eine lokale Entwicklung, womit sich nur Verfahren mit Konvergenzordnung 2 gewinnen lassen.

Im Folgenden soll ein zweistufiges explizites Runge-Kutta Verfahren auf unser Anfangswertproblem (4.1),(4.2) anwendet werden, obwohl die rechte Seite von (4.1) nicht einmal stetig ist und somit keine Konvergenz des RK-Verfahrens sichergestellt werden kann. Wie schon in Beispiel 2.2.3 erwähnt, wird ein solches Verfahren über das Butcher-Array definiert, welches hier so aussieht:

$$\begin{array}{c|cc} 0 & & \\ c_2 & a_{2,1} & \\ \hline 1 & b_1 & b_2 \end{array}$$

In Beispiel 2.2.5 findet man dazu konkrete Verfahren. Dabei müssen die Koeffizienten die Gleichungen

$$c_2 = a_{2,1} \quad \text{und} \quad 1 = b_1 + b_2 \tag{4.14}$$

erfüllen.

In Beispiel 2.2.3 setzen wir nun $f(t,x) = \mathfrak{a}(x) + B\mathfrak{u}(t)$ und erhalten als allgemeine Verfahrensfunktion

$$\Phi(t,x,h) = b_1(\mathfrak{a}(x) + \mathfrak{B}\mathfrak{u}(t)) + b_2(\mathfrak{B}\mathfrak{u}(t+c_2h) + \mathfrak{a}(x + a_{2,1}h \cdot (\mathfrak{a}(x) + \mathfrak{B}\mathfrak{u}(t)))) .$$

Daraus resultiert nach Definition 2.2.1 folgende Gitterfunktion

$$\begin{aligned} x_h((i+1)h) &= x_h(ih) + hb_1(\mathfrak{a}(x_h(ih)) + \mathfrak{B}\mathfrak{u}(ih)) + hb_2\mathfrak{B}\mathfrak{u}(ih + c_2h) \\ &\quad + hb_2\mathfrak{a}(x_h(ih) + a_{2,1}h \cdot (\mathfrak{a}(x_h(ih)) + \mathfrak{B}\mathfrak{u}(ih))) \qquad (i = 0, \dots, N-1) \\ x_h(0) &= x_0 . \end{aligned}$$

Damit diese explizite RK-Verfahren in einem glatten Fall auch von 2. Ordnung konvergieren, müssen die Koeffizienten noch eine weitere Bedingung erfüllen. (Diese Konvergenz gilt hier nicht, da die rechte Seite i.A. nicht einmal stetig ist). Nach [10, sec. 32, p. 156] muss die Gleichung

$$b_2c_2 = b_2a_{2,1} = \frac{1}{2} \tag{4.15}$$

gelten. Wie man leicht sehen kann, erfüllen das Heun-Verfahren und das verbesserte Euler-Verfahren aus Beispiel 2.2.5 diese Bedingung.

Wie schon erwähnt, wird die Gitterfunktion x_h im Allgemeinen nicht gegen die Lösung $\hat{x}$ konvergieren, da die Kontrollfunktion $\mathfrak{u}$ nicht einmal stetig ist. Außerdem macht eine punktwertige Auswertung einer L^1-Funktion keinen Sinn. Wie im linearen Fall ersetzten wir in obiger Gleichung die Diskretisierungen der Kontrollfunktion $\mathfrak{u}(ih)$ und $\mathfrak{u}(ih + c_2h)$ durch die Variablen $\mathfrak{u}_{1,i}$ und $\mathfrak{u}_{2,i}$. Für den ersten Zeitschritt, d.h. $i = 0$, schreiben wir kurz $\mathfrak{u}_1$und $\mathfrak{u}_2$. Diese Größen nennen wir Kontrollvektoren. Sie sind Vektoren aus dem $\mathbb{R}^m$. Damit erhalten wir

$$\begin{aligned} x_h((i+1)h) &= x_h(ih) + hb_1(\mathfrak{a}(x_h(ih)) + \mathfrak{B}\mathfrak{u}_{1,i}) + hb_2\mathfrak{B}\mathfrak{u}_{2,i} \\ &\quad + hb_2\mathfrak{a}(x_h(ih) + a_{2,1}h \cdot (\mathfrak{a}(x_h(ih)) + \mathfrak{B}\mathfrak{u}_{1,i})) \qquad (4.16) \\ x_h(0) &= x_0 , \qquad (4.17) \end{aligned}$$

wobei i von 0 bis $N-1$ läuft.

Ziel ist es nun durch geschickte Wahl dieser Kontrollvektoren, was man eine Auswahlstrategie nennen kann, die Konvergenz von x_h gegen $\hat{x}$, wie sie in Definition 2.2.4 erklärt wurde, zu sichern.

An dieser Stelle sei noch einmal erinnert an die Auswahlmenge $\mathcal{I}_{m,p}$, den diskreten Kontrollbereich $\mathcal{U}_{m,p}$ und die Matrizen $\Gamma_{m,p}$ und $\Delta_{m,p}$, welche alle in Unterabschnitt 3.1.2 eingeführt wurden. Dabei gilt die Beziehung

$$\mathcal{U}_{m,p} = \Gamma_{m,p}^{-1} \Delta_{m,p} \mathcal{I}_{m,p} \,.$$

Die Matrix $\Gamma_{m,2}$ für das obige explizite RK-Schema hat zusammen mit den Bedingungen (4.14) und (4.15) die Gestalt

$$\Gamma_{m,2} = \begin{pmatrix} b_1 \mathfrak{E}_m & b_2 \mathfrak{E}_m \\ 1\mathfrak{E}_m & 0\mathfrak{E}_m \end{pmatrix} .$$

Und damit gilt für den diskreten Kontrollbereich

$$\mathcal{U}_{m,2} = \begin{pmatrix} b_1 \mathfrak{E}_m & b_2 \mathfrak{E}_m \\ \frac{1}{2}\mathfrak{E}_m & 0\mathfrak{E}_m \end{pmatrix}^{-1} \mathcal{I}_{m,2} \,.$$

Der folgende Satz zeigt nun, welche Auswahlstrategie gewählt werden muss, damit die Verfahren konsistent sind von Ordnung 2. Dabei soll die Formulierung "ein Runge-Kutta-Schema der Ordnung 2" bedeuten, dass das zu Grunde liegende RK-Verfahren bei genügender Glattheit der rechten Seite die Konvergenzordnung 2 besitzt.

Theorem 4.1.3. *Für das nichtlineare Kontrollproblem (NKP) auf dem Intervall $[0,h]$ sei ein explizites Runge-Kutta Schema der Ordnung 2 mit 2 Stufen gegeben. Dabei sei das Gewicht $b_2 \neq 0$. Die diskrete Lösung x_h werde mittels diesem Schema konstruiert. Dann gilt:*

(A1) *Zu einer vorgegebenen Kontrollfunktion $\mathfrak{u} \in L^1([0,h];U)$ existiert ein Kontrollvektor $\begin{pmatrix} \mathfrak{u}_1 \\ \mathfrak{u}_2 \end{pmatrix} \in \mathcal{U}_{m,2}$, sodass (A3) gilt.*

(A2) *Zu einem vorgegebenen Kontrollvektor $\begin{pmatrix} \mathfrak{u}_1 \\ \mathfrak{u}_2 \end{pmatrix} \in \mathcal{U}_{m,2}$ existiert eine Kontrollfunktion $\mathfrak{u} \in L^1([0,h];U)$, sodass (A3) gilt.*

(A3) *Es gilt*

$$\|\hat{x}(h) - x_h(h)\|_\infty \leq C \cdot h^3 \qquad \text{für } h \to 0 \,,$$

wobei

$$C := \frac{n^2}{6} \|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n} M^2 + \frac{n}{6} \|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n} M + \frac{a_{2,1} n^2}{4} \|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n} \tilde{M}^2$$

ist mit $M = \|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n} + \|\mathfrak{B}\|_Z \|U\|_\infty$ *und* $\tilde{M} = \|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n} + \|\mathfrak{B}\|_Z \|\mathcal{U}_{m,2}\|_\infty$.

Beweis. Zuerst schreiben wir noch einmal die definierende Gleichung (4.16) der diskreten Lösung x_h für den ersten Zeitschritt d.h. $i = 0$, wobei wir zur Vereinfachung $x_h(0)$ durch die Anfangsbedingung x_0 ersetzen:

$$x_h(h) = x_0 + h\left[b_1\left(\mathfrak{a}(x_0) + \mathfrak{B}\mathfrak{u}_1\right) + b_2\left(\mathfrak{a}\left(x_0 + a_{2,1}h \cdot \left(\mathfrak{a}(x_0) + \mathfrak{B}\mathfrak{u}_1\right)\right) + \mathfrak{B}\mathfrak{u}_2\right)\right] .$$

Im Vorgriff auf Proposition 5.2.3 benutzten wir, dass $\mathcal{U}_{m,2}$ beschränkt ist. Dies wäre an dieser Stelle umständlich zu zeigen. Es ist dann $\|\mathcal{U}_{m,2}\|_\infty < \infty$ und es gilt

$$\|\mathfrak{a}(x_0) + \mathfrak{B}\mathfrak{u}_1\|_\infty \leq \|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n} + \|\mathfrak{B}\|_Z \|\mathcal{U}_{m,2}\|_\infty .$$

Dann ist auch $\xi := a_{2,1}h \cdot \left(\mathfrak{a}(x_0) + \mathfrak{B}\mathfrak{u}_1\right)$ beschränkt, denn es ist ja $h < 1$ vorausgesetzt (siehe Gleichung 2.3 auf Seite 38). Also gilt für den Radius

$$r := \sqrt{n} a_{2,1} h \left(\|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n} + \|B\|_Z \|\mathcal{U}_{m,2}\|_\infty\right) ,$$

dass stets $x_0 + \xi \in B_r(x_0)$ ist (die Kugel wird bezüglich der 2-Norm gebildet).
Mit Hilfe von Satz 1.7.2 führen wir auf $B_r(x_0)$ eine Taylorentwicklung 1. Ordnung für $\mathfrak{a}$ um den Punkt x_0 durch und erhalten

$$\mathfrak{a}\left(x_0 + a_{2,1}h \cdot \left(\mathfrak{a}(x_0) + \mathfrak{B}\mathfrak{u}_1\right)\right) = \mathfrak{a}(x_0) + a_{2,1}h \cdot \mathfrak{J}_\mathfrak{a}(x_0)\left(\mathfrak{a}(x_0) + \mathfrak{B}\mathfrak{u}_1\right) + \tilde{\mathfrak{r}}(h) ,$$

mit einer stetigen Restfunktion $\tilde{\mathfrak{r}}[0,\infty) \longrightarrow \mathbb{R}^n$ definiert durch

$$\tilde{\mathfrak{r}}(h) := \mathfrak{a}\left(x_0 + a_{2,1}h \cdot \left(\mathfrak{a}(x_0) + \mathfrak{B}\mathfrak{u}_1\right)\right) - \mathfrak{a}(x_0) + a_{2,1}h \cdot \mathfrak{J}_\mathfrak{a}(x_0)\left(\mathfrak{a}(x_0) + \mathfrak{B}\mathfrak{u}_1\right) .$$

Nach Satz 1.7.2 mit $\xi = a_{2,1}h \cdot \left(\mathfrak{a}(x_0) + \mathfrak{B}\mathfrak{u}_1\right)$ ist diese Restfunktion beschränkt durch

$$\|\tilde{\mathfrak{r}}(h)\|_\infty \leq \left(\frac{a_{2,1}^2 n^2}{2} \|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n} \tilde{M}^2\right) \cdot h^2 ,$$

wobei hier $\tilde{M} := \|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n} + \|B\|_Z \|\mathcal{U}_{m,2}\|_\infty$ ist.
Setzt man diese Entwicklung in obige Gleichung ein und wendet die Bedingungen (4.15) an, so erhält man

$$x_h(h) = x_0 + h(b_1 + b_2)\,\mathfrak{a}(x_0) + h\mathfrak{B}(b_1\mathfrak{u}_1 + b_2\mathfrak{u}_2) + \frac{h^2}{2} \cdot \mathfrak{J}_\mathfrak{a}(x_0) \cdot \mathfrak{B}\mathfrak{u}_1 + \\ + \frac{h^2}{2} \cdot \mathfrak{J}_\mathfrak{a}(x_0) \cdot \mathfrak{a}(x_0) + hb_2\tilde{\mathfrak{r}}(h) , \quad (4.18)$$

wobei $\|hb_2\tilde{\mathfrak{r}}(h)\|_\infty \leq \left(\frac{a_{2,1}n^2}{4}\|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n}\tilde{M}^2\right)\cdot h^3$ ist ebenfalls mit (4.15).
Nach dieser Vorarbeit wollen wir die eigentlichen Aussagen zeigen:
$(\mathbf{A1}) \implies (\mathbf{A3})$
Sei $\mathfrak{u} \in L^1([0,h];U)$ eine beliebige Kontrollfunktion. Wir berechnen die Integrale

$$\zeta_1 := \frac{1}{h}\int_0^h \mathfrak{u}(t)dt \quad \text{und} \quad \zeta_2 := \frac{1}{h^2}\int_0^h\int_0^t \mathfrak{u}(s)\,dt\,ds\,.$$

Die Matrix $\begin{pmatrix} b_1 & b_2 \\ \frac{1}{2} & 0 \end{pmatrix}$ ist wegen $b_2 \neq 0$ invertierbar. Deswegen ist nach Lemma 3.1.6(i) auch die vergrößerte Version $\begin{pmatrix} b_1\mathfrak{E}_m & b_2\mathfrak{E}_m \\ \frac{1}{2}\mathfrak{E}_m & 0\mathfrak{E}_m \end{pmatrix}$ invertierbar. Deswegen können wir $\mathfrak{u}_1, \mathfrak{u}_2$ setzten als

$$\begin{pmatrix} \mathfrak{u}_1 \\ \mathfrak{u}_2 \end{pmatrix} := \begin{pmatrix} b_1\mathfrak{E}_m & b_2\mathfrak{E}_m \\ \frac{1}{2}\mathfrak{E}_m & 0\mathfrak{E}_m \end{pmatrix}^{-1}\begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix}. \tag{4.19}$$

Mit einfacher bzw. zweifacher Anwendung der Substitutionsregel folgt:

$$\zeta_1 = \int_0^1 \tilde{\mathfrak{u}}(t)\,dt \quad \text{und} \quad \zeta_2 = \int_0^1\int_0^t \tilde{\mathfrak{u}}(s)\,dt\,ds\,, \tag{4.20}$$

wobei $\tilde{\mathfrak{u}}(\cdot) := \mathfrak{u}(h\cdot)$ ist. Da $\tilde{\mathfrak{u}}(\cdot) \in L^1([0,1];U)$ ist, ist damit auch $\begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} \in \mathcal{I}_{m,2}$ nach Definition von $\mathcal{I}_{m,2}$. Weiter ist dann auch $\begin{pmatrix} \mathfrak{u}_1 \\ \mathfrak{u}_2 \end{pmatrix} \in \mathcal{U}_{m,2}$ ebenfalls nach Definition von $\mathcal{U}_{m,2}$.
Aus (4.19) folgen unmittelbar die Gleichungen

$$b_1\mathfrak{u}_1 + b_2\mathfrak{u}_2 = \frac{1}{h}\int_0^h \mathfrak{u}(t)dt \tag{4.21}$$

$$\frac{1}{2}\mathfrak{u}_1 = \frac{1}{h^2}\int_0^h\int_0^t \mathfrak{u}(s)\,dt\,ds\,. \tag{4.22}$$

Diese setzen wir in Gleichung (4.18) ein und bilden die Differenz mit der Entwicklung (4.3) von $\hat{x}(h)$ aus Satz 4.1.1 und erhalten mit Dreiecksungleichung:

$$\|\hat{x}(h) - x_h(h)\|_\infty = \|\mathfrak{r}(h) - hb_2\tilde{\mathfrak{r}}(h)\|_\infty \leq C\cdot h^3\,,$$

wobei $\mathfrak{r}(h)$ und die Abschätzung dafür aus Satz 4.3 stammen und

$$C = \frac{n^2}{6}\|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n}M^2 + \frac{n}{6}\|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n}M + \frac{a_{2,1}n^2}{4}\|\mathfrak{a}\|_{W^{2,\infty}(\mathbb{R}^n)^n}\tilde{M}^2$$

ist.

$(\mathbf{A2}) \implies (\mathbf{A3})$

Sei $\begin{pmatrix} \mathfrak{u}_1 \\ \mathfrak{u}_2 \end{pmatrix} \in \mathcal{U}_{m,2}$ vorgegeben. Nach Definition von $\mathcal{U}_{m,2}$ gibt es $\begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} \in \mathcal{I}_{m,2}$ mit $\begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} = \begin{pmatrix} b_1\mathfrak{E}_m & b_2\mathfrak{E}_m \\ \frac{1}{2}\mathfrak{E}_m & 0\mathfrak{E}_m \end{pmatrix} \begin{pmatrix} \mathfrak{u}_1 \\ \mathfrak{u}_2 \end{pmatrix}$.

Und nach Definition von $\mathcal{I}_{m,2}$ gibt es ein $\tilde{\mathfrak{u}} \in L^1([0,1];U)$ mit

$$\zeta_1 = \int_0^1 \tilde{\mathfrak{u}}(t)dt \quad \text{und} \quad \zeta_2 = \int_0^1 \int_0^t \tilde{\mathfrak{u}}(s)\,dt\,ds\,.$$

Mit der Trafo-Formel folgt:

$$\zeta_1 = \frac{1}{h}\int_0^h \mathfrak{u}(t)dt \quad \text{und} \quad \zeta_2 = \frac{1}{h^2}\int_0^h \int_0^t \mathfrak{u}(s)\,dt\,ds\,,$$

wobei $\mathfrak{u}(t) := \tilde{\mathfrak{u}}(\frac{1}{h}\cdot t)$ gesetzt ist und deswegen ist $\mathfrak{u} \in L^1([0,h];U)$.

Bleibt noch zu zeigen, dass mit $\mathfrak{u}$ als gesuchter Kontrollfunktion (A3) erfüllt ist.

Es sei nun $\hat{x}$ die exakte Lösung von (NKP) bzgl. der Steuerfunktion $\mathfrak{u}$. Wie in der vorherigen Richtung gelten wieder die Gleichungen (4.21) und (4.22). Der Rest geht wie bei der vorherigen Richtung ab diesen beiden Gleichungen. □

Bemerkung 4.1.4.

- Man kann die vorgestellten Verfahren auch als modifizierte Runge-Kutta-Verfahren ansehen mit einer speziellen Auswahlstrategie. Bei einem herkömmlichen Runge-Kutta Verfahren wäre $\mathfrak{u}_1 = \mathfrak{u}(0)$ und $\mathfrak{u}_2 = \mathfrak{u}(c_1 h)$. Hier aber ist $\mathfrak{u}_1 = \int_0^1 \mathfrak{u}(ht)dt$ und $\mathfrak{u}_2 = \int_0^1 \int_0^t \mathfrak{u}(hs)\,ds\,dt$. Also hängen auch hier die diskreten Kontrollen von der Kontrollfunktion $\mathfrak{u}$ ab.

- Dieser Satz zeigt die Konsistenz für die vorgestellten modifizierten Runge-Kutta-Verfahren. Dabei ist die Konsistenzordnung aber nur 2, da die gewöhnliche Darstellung für Konsistenz der lokale Diskretisierungsfehler $\frac{\hat{x}(h)-x_h(h)}{h}$ ist.

- Da $\mathfrak{a}$ auf dem Zustandsraum Ω stetig differenzierbar ist mit beschränkten partiellen Ableitungen, ist es auch Lipschitz-stetig auf Ω. Damit kann man zeigen, dass die Verfahrensfunktion Φ Lipschitz-stetig ist. Mit Proposition 2.2.9 folgt die Stabilität dieser modifizierten RK-Verfahren. Und mit Theorem 2.2.10 folgt die Konvergenz von Ordnung 2 dieser Verfahren.

4.2. Nachteile dieses Zugangs

Da der Schwerpunkt dieser Arbeit nicht bei diesen im vorigen Abschnitt gewonnen Verfahren liegt, soll auf einen formalen Konvergenzbeweis verzichtet werden.

Bisher konnte dieser Zugang von Ferretti vom linearen auf den nichtlinearen Fall übertragen werden, zumindest für Konvergenzordnung 2. In diesem Abschnitt will ich auf die Nachteile dieses Zugangs eingehen.

Im ersten Unterabschnitt soll gezeigt werden, warum sich die Konsistenzordnung dieser modifizierten RK-Verfahren kaum steigern lässt. Und im zweiten Unterabschnitt soll gezeigt werden, warum sich die bisher gewonnenen Verfahren nicht in effektive Verfahren zur Berechnung der erreichbaren Menge umsetzten lassen wie im linearen Fall.

4.2.1. Steigerung der Konsistenzordnung

Um die Konsistenzordnung zu steigern, brauchen wir vor allem eine Entwicklung höherer Ordnung von der exakten Lösung $\hat{x}$. Denn RK-Verfahren höherer Ordnung stehen ja zur Verfügung. Ich will daher hier aufzeigen mit welchen Schwierigkeiten es verbunden ist eine lokale Entwicklung von $\hat{x}$ von Ordnung 4 zu erreichen. Dabei sollen nur die bisherigen Mittel, also die Taylorentwicklung und die Darstellung der Lösung als $\hat{x}(t) = \hat{x}(0) + \int_0^t \hat{x}'(s)\,ds$, verwendet werden (andere und bessere Mittel sind mir auch nicht bekannt).

Zur Vereinfachung setzten wir $n, m = 1$ voraus, d.h. alle beteiligten Größen des Kontrollproblems sind nun Skalare. Deswegen schreiben wir auch a statt $\mathfrak{a}$, u statt $\mathfrak{u}$ und b statt $\mathfrak{B}$. Außerdem soll $a \in \mathcal{C}_B^3(\mathbb{R};\mathbb{R})$ sein. Es gibt also ein $M > 0$, sodass

$$\left|a^{(i)}(t)\right| < M \qquad \forall t \in \mathbb{R}\,,\ i = 0,\dots,3$$

ist.

Wir betrachten also noch einmal den Beweis des Satzes 4.1.1 . Die Taylorentwicklung (4.10) wird in Zeile (4.12) eingesetzt und führt dann auf einen $\mathcal{O}(h^3)$ Term. Deshalb müssen wir an dieser Stelle ein Taylorentwicklung der Ordnung 2 einsetzten. Deswegen soll hier diese quadratische Taylorentwicklung bereitgestellt werden:

$$a\left(x_0 + \int_0^h \hat{x}'(t)\,dt\right) = a(x_0) + a'(x_0)\cdot\int_0^h \hat{x}'(t)\,dt + \frac{a''(x_0)}{2}\cdot\left(\int_0^h \hat{x}'(t)\,dt\right)^2 + \mathcal{O}\left(h^3\right)\,. \quad (4.23)$$

Der Restterm $\mathcal{O}(h^3)$ ergibt sich, da $a^{(3)}$ beschränkt ist und da $\int_0^h \hat{x}'(t)\,dt = \mathcal{O}(h)$ ist (vgl. (4.7)).

Wir gehen jetzt genau so vor wie in diesem Beweis und starten wieder mit (4.4)

$$\begin{aligned}\hat{x}(h) &= x_0 + \int_0^h \hat{x}'(t)\,dt \overset{(4.1)}{=} x_0 + \int_0^h a\left(\hat{x}(t)\right) + bu(t)\,dt \\ &\overset{(4.4)}{=} x_0 + \int_0^h a\left(x_0 + \int_0^t \hat{x}'(s)\,ds\right)\,dt + b\int_0^h u(t)\,dt\end{aligned}$$

An dieser Stelle müssen wir, wie oben erwähnt, jetzt für $a\left(x_0 + \int_0^t \hat{x}'(t)\,dt\right)$ die quadratische Taylorentwicklung einsetzten. Dann folgt:

$$\begin{aligned}\hat{x}'(h) &= x_0 + \int_0^h a\left(x_0\right) + a'\left(x_0\right)\int_0^t \hat{x}'(s)\,ds + \frac{a''\left(x_0\right)}{2}\cdot\left(\int_0^t \hat{x}'(s)\,ds\right)^2 + \mathcal{O}\left(t^3\right)\,dt \\ &\quad + b\int_0^h u(t)\,dt \\ &\overset{(4.1)}{=} x_0 + a\left(x_0\right)h + a'\left(x_0\right)\int_0^h\int_0^t a\left(\hat{x}(s)\right)\,ds\,dt + a'\left(x_0\right)b\int_0^h\int_0^t u(s)\,ds\,dt \\ &\quad + \frac{a''\left(x_0\right)}{2}\int_0^h\left(\int_0^t a\left(\hat{x}(s)\right) + bu(s)\,ds\right)^2 dt + b\int_0^h u(t)\,dt + \mathcal{O}\left(h^4\right) \qquad (4.24)\end{aligned}$$

Ab jetzt wollen wir die beiden Terme mit $a\left(\hat{x}(s)\right)$ gesondert betrachten. Wir fahren fort mit dem ersten Term:

$$\begin{aligned}&\int_0^h\int_0^t a\left(\hat{x}(s)\right)\,ds\,dt \\ \overset{(4.4)}{=}\ &\int_0^h\int_0^t a\left(x_0 + \int_0^s \hat{x}'(r)dr\right)\,ds\,dt \\ \overset{(4.10)}{=}\ &\int_0^h\int_0^t a\left(x_0\right) + a'\left(x_0\right)\int_0^s \hat{x}'(r)dr + \mathcal{O}\left(s^2\right)\,ds\,dt \\ \overset{(4.1)}{=}\ &a\left(x_0\right)\frac{h^2}{2} + a'\left(x_0\right)\int_0^h\int_0^t\int_0^s a\left(\hat{x}(r)\right) + bu(r)\,dr\,ds\,dt + \mathcal{O}\left(h^4\right) \\ \overset{(4.4)+(4.8)}{=}\ &a\left(x_0\right)\frac{h^2}{2} + a'\left(x_0\right)\int_0^h\int_0^t\int_0^s a\left(x_0\right) + \mathcal{O}\left(r\right)\,dr\,ds\,dt \\ &\qquad + a'\left(x_0\right)b\int_0^h\int_0^t\int_0^s u(r)\,dr\,ds\,dt + \mathcal{O}\left(h^4\right) \\ =\ &a\left(x_0\right)\frac{h^2}{2} + a'\left(x_0\right)a\left(x_0\right)\frac{h^3}{6} + a'\left(x_0\right)b\int_0^h\int_0^t\int_0^s u(r)\,dr\,ds\,dt + \mathcal{O}\left(h^4\right)\end{aligned}$$

Jetzt behandeln wir den zweiten Term. Um die folgende komplizierte Rechnung möglichst einfach zu machen, müssen Terme, bei denen abzusehen ist, dass sie zu $\mathcal{O}\left(h^4\right)$

werden, möglichst frühzeitig durch das entsprechende Landau-Symbol ersetzt werden. Ich gebe nur Zwischenschritte wieder:

$$\begin{aligned}
&\int_0^h \left(\int_0^t a\left(\hat{x}(s)\right) + bu(s)\, ds \right)^2 dt \\
= &\int_0^h \left(\int_0^t a\left(x_0\right) + a'\left(x_0\right) \int_0^s \hat{x}'(r) dr + \mathcal{O}\left(s^2\right) + bu(s)\, ds \right)^2 dt \\
= &\int_0^h \left(ta\left(x_0\right) + a'\left(x_0\right) \int_0^t \int_0^s a\left(x_0\right) + \mathcal{O}(r) + bu(r)\, dr + \mathcal{O}\left(s^2\right) + bu(s)\, ds \right)^2 dt \\
= &\int_0^h \left(ta\left(x_0\right) + \frac{t^2}{2} a\left(x_0\right) a'\left(x_0\right) + \int_0^t \int_0^s bu(r)\, dr\, ds + \int_0^t bu(s)\, ds + \mathcal{O}\left(t^3\right) \right)^2 dt \\
= &\frac{h^3}{3} a\left(x_0\right)^2 + 2a\left(x_0\right) \int_0^h t \int_0^t bu(s)\, ds\, dt + \int_0^h \left(\int_0^t bu(s)\, ds \right)^2 dt + \mathcal{O}\left(h^4\right)
\end{aligned}$$

Setzt man diese Ergebnisse in (4.24) ein, so erhält man

$$\begin{aligned}
\hat{x}'(h) = \; & x_0 + h \cdot a\left(x_0\right) + a'\left(x_0\right) b \int_0^h \int_0^t u(s)\, ds\, dt + \frac{h^2}{2} a'\left(x_0\right) a\left(x_0\right) \\
& + \frac{h^3}{6} a\left(x_0\right)^2 a''\left(x_0\right) + a'\left(x_0\right)^2 b \int_0^h \int_0^t \int_0^s u(r)\, dr\, ds\, dt + b \int_0^h u(t)\, dt \\
& + a\left(x_0\right) a''\left(x_0\right) b \int_0^h t \int_0^t u(s)\, ds\, dt + \frac{1}{2} a''\left(x_0\right) b^2 \int_0^h \left(\int_0^t u(s)\, ds \right)^2 dt \\
& + \frac{h^3}{6} a\left(x_0\right) a'\left(x_0\right)^2 + \mathcal{O}\left(h^4\right)
\end{aligned}$$

Damit haben wir eine lokale Approximation der Lösung von Ordnung 4 erreicht. Man sieht schon, dass die Formel erheblich komplizierter ist und der Aufwand zu ihrer Berechnung erheblich angestiegen ist im Vergleich zur Entwicklung aus Satz 4.1.1. Die Hürde der komplizierten Berechnung könnte man aber auch bei höheren Entwicklungen mit Computeralgebrasystemen überwinden.

Ein größerer Nachteil ist das Integral $\int_0^h \left(\int_0^t u(s)\, ds \right)^2 dt$, welches in dieser Entwicklung vorkommt. Denn dieses Integral kann man nicht als Mehrfachintegral darstellen, wie sie in der Menge $\mathcal{I}_{m,p}$ vorkommen. Damit reiht sich diese Entwicklung nicht in die Theorie ein, wie wir sie für das lineare Kontrollproblem in Kapitel 3 kennengelernt haben. Will man diese Entwicklung für punktwertige Verfahren nutzen, so muss man dieses Integral ausrechnen, was mit erheblichem Aufwand verbunden ist. Dagegen macht des Integral $\int_0^h t \int_0^t u(s)\, ds\, dt$ keine Probleme, weil man dies mittels partieller Integration als Differenz zweier Mehrfachintegrale darstellen kann.

Für Entwicklungen von noch höherer Ordnung sind entsprechend noch mehr unangenehme Integrale zu erwarten.

4.2.2. Mengenwertige Verfahren

In diesem Unterabschnitt soll auf die Nachteile eingegangen werden, die entstehen, wenn man die Verfahren aus dem vorigen Abschnitt oder auch die Entwicklung aus dem vorigen Unterabschnitt in mengenwertige Verfahren umsetzen will.

Wir benutzten die Schreibweise von Abschnitt 3.3 und stellen die diskreten erreichbaren Mengen der modifizierten RK-Verfahren aus dem vorigen Abschnitt dar für die ersten vier Zeitschritte:

$$\begin{aligned}
\mathcal{R}_h(0,x_0) &= \{x_0\} \\
\mathcal{R}_h(h,x_0) &= \Bigg\{ x_0 + hb_1\mathfrak{a}(x_0) + h\mathfrak{B}(b_1\mathfrak{u}_1 + b_2\mathfrak{u}_2) \\
&\qquad + hb_2\mathfrak{a}(x_0 + a_{2,1}h(\mathfrak{a}(x_0) + \mathfrak{B}\mathfrak{u}_1)) \mid \begin{pmatrix}\mathfrak{u}_1\\ \mathfrak{u}_2\end{pmatrix} \in \mathcal{U}_{m,p} \Bigg\} \\
\mathcal{R}_h(2h,x_0) &= \bigcup_{\mathfrak{v}\in\mathcal{R}_h(h,x_0)} \mathcal{T}_{\mathfrak{v}} \\
\mathcal{R}_h(3h,x_0) &= \bigcup_{\mathfrak{v}\in\mathcal{R}_h(2h,x_0)} \mathcal{T}_{\mathfrak{v}}
\end{aligned}$$

mit

$$\mathcal{T}_{\mathfrak{v}} = \left\{ \mathfrak{v} + hb_1\mathfrak{a}(\mathfrak{v}) + h\mathfrak{B}(b_1\mathfrak{u}_1 + b_2\mathfrak{u}_2) + hb_2\mathfrak{a}(\mathfrak{v} + a_{2,1}h(\mathfrak{a}(\mathfrak{v}) + \mathfrak{B}\mathfrak{u}_1)) \mid \begin{pmatrix}\mathfrak{u}_1\\ \mathfrak{u}_2\end{pmatrix} \in \mathcal{U}_{m,p} \right\}$$

Man sieht schon bei der Menge $\mathcal{R}_h(h,x_0)$, dass es nicht mehr möglich ist die diskrete Kontrollmenge $\mathcal{U}_{m,p}$ als Summand einer Minkowski-Summe zu schreiben, und zwar wegen dem Summanden $hb_2\mathfrak{a}(x_0 + a_{2,1}h\mathfrak{a}(x_0 + \mathfrak{B}\mathfrak{u}_1))$. Dies ist schon ein großer Nachteil, da nun auch nicht mehr sichergestellt ist, dass $\mathcal{R}_h(h,x_0)$ konvex ist. Denn nun können die Methoden der konvexen Analysis, die in Kapitel 5 beschrieben werden nicht mehr angewandt werden.

Ein noch größerer Nachteil ist bei der Menge $\mathcal{R}_h(2h,x_0)$ zu sehen. Denn hier muss eine im Allgemeinen überabzählbare Vereinigung der Mengen $\mathcal{T}_{\mathfrak{v}}$ gebildet werden. Selbst wenn jede Menge der Vereinigung konvex wäre, so wäre die gesamte Vereinigung im Allgemeinen nicht mehr konvex. Damit können wiederum die Methoden der konvexen Analysis nicht angewandt werden. Man müsste endlich viele $\mathfrak{v} \in \mathcal{R}_h(h,x_0)$ auswählen und dafür dann die Mengen $\mathcal{T}_{\mathfrak{v}}$ berechnen. Anschließend müsste man die Vereinigung dieser $\mathcal{T}_{\mathfrak{v}}$ bilden oder mit jeder einzelnen Menge im nächsten Zeitschritt weiterrechnen, d.h. die Vereinigung $\mathcal{R}_h(2h,x_0)$ wird für alle diese $\mathcal{T}_{\mathfrak{v}}$ gebildet statt für $\mathcal{R}_h(2h,x_0)$, usw. Dies ist in jedem Fall ein exponentieller Aufwand.

Der Nachteil, den wir für $\mathcal{R}_h(h,x_0)$ festgestellt haben, lässt sich beseitigen, wenn man direkt die Entwicklung (4.3) aus Satz 4.1.1 in ein mengenwertiges Verfahren umsetzt,

ähnlich wie wir das auch bei den Ferretti-Verfahren in Kapitel 3 gemacht haben. Dann hätte man als erreichbare Menge

$$\begin{aligned}\mathcal{R}_h(h, x_0) &= \Bigg\{ x_0 + h\mathfrak{a}(x_0) + \frac{h^2}{2}\mathfrak{J}_{\mathfrak{a}}(x_0)\,\mathfrak{a}(x_0) + h\mathfrak{B}\mathfrak{v}_1 + h^2\mathfrak{J}_{\mathfrak{a}}(x_0)\,\mathfrak{B}\mathfrak{v}_2 \\ &\qquad\qquad \Big| \begin{pmatrix} \mathfrak{v}_1 \\ \mathfrak{v}_2 \end{pmatrix} \in \mathcal{I}_{m,p} \Bigg\} \\ &= \left\{ x_0 + h\mathfrak{a}(x_0) + \frac{h^2}{2}\mathfrak{J}_{\mathfrak{a}}(x_0)\,\mathfrak{a}(x_0) \right\} + \left[h\mathfrak{B} \mid h^2\mathfrak{J}_{\mathfrak{a}}(x_0)\,\mathfrak{B} \right] \mathcal{I}_{m,p}\,,\end{aligned}$$

wobei die Schrittweite h in den Summanden $h\mathfrak{B}\mathfrak{v}_1$ und $h^2\mathfrak{J}_{\mathfrak{a}}(x_0)\,\mathfrak{B}\mathfrak{v}_2$ dadurch hineinkommt, dass die Integrale auf das Einheitsintervall transformiert werdem müssen (vgl. (4.20)). Diese Menge ist nun konvex. Außerdem erleichtert die Darstellung als Minkowski-Summe die numerische Berechnung mit den Mitteln von Kapitel 5. Jedoch das Problem mit der Vereinigung und dem exponentiellen Aufwand bei den folgenden diskreten erreichbaren Mengen kann auch für dieses Verfahren nicht vermieden werden. Die weiteren diskreten erreichbaren Mengen sind dann im Allgemeinen auch nicht mehr konvex.

Mit den gleichen Schwierigkeiten hat auch ein mengenwertiges Verfahren zu kämpfen, welches auf der Entwicklung aus dem vorherigen Unterabschnitt beruht. Nur kommt hier noch hinzu, dass das Integral $\int_0^h \left(\int_0^t u(s)\,ds \right)^2 dt$ nicht in der Menge $\mathcal{I}_{m,p}$ enthalten ist. Man müsste also eine neue Menge definieren, die aber dann wiederum schwerer zu berechnen wäre. Alle Methoden zur effektiven Berechnung von $\mathcal{I}_{m,p}$ aus Kapitel 5 wären nicht mehr anwendbar.

Insgesamt führt dieser Ansatz für das nichtlineare Kontrollproblem auf höchst ineffektive Verfahren. Deswegen betrachten wir das nichtlineare Kontrollproblem nicht weiter.

Kapitel 5.

Algorithmische und mathematische Umsetzung

In diesem Kapitel wollen wir uns näher mit der konkreten Umsetzung der in Kapitel 3 gewonnen mengenwertigen Verfahren befassen. Dabei sollen vor allem die mathematischen Aspekte näher behandelt werden, aber auch ein wenig die algorithmischen . Zunächst gehen wir von dem mengenwertigen Ferretti-Verfahren der Ordnung p aus, welches zur Approximation der erreichbaren Menge des linearen Kontrollproblems dient. Zur Erinnerung soll es hier nochmal geschrieben werden

$$\begin{aligned} \mathcal{R}_h\left(0, x_0\right) &= \{x_0\} \\ \mathcal{R}_h\left((i+1)h, x_0\right) &= \mathfrak{C} \cdot \mathcal{R}_h\left(ih, x_0\right) + \mathfrak{D} \cdot \mathcal{I}_{m,p} \qquad (i = 0, \ldots, N-1)\,, \end{aligned} \tag{5.1}$$

wobei $\mathfrak{C}$ und $\mathfrak{D}$ wie in Definition 3.3.1 sind. Weiter ist $x_0 \in \mathbb{R}^n$ die Anfangsbedingung des lineare Kontrollproblems. In diesem Kapitel setzen wir den Kontrollbereich U, der ja die Menge $\mathcal{I}_{m,p}$ bestimmt, als kompakt, konvex und nichtleer voraus. Es brauchen die mengenwertigen RK-Verfahren nicht extra betrachtet werden, da wir ohnehin gesehen haben, dass sie mit dem Ferretti-Verfahren der jeweiligen Ordnung übereinstimmen.

Um diese mengenwertigen Verfahren im Rechner realisieren zu können, muss geklärt werden, wie Mengen im Rechner repräsentiert werden sollen. Wir brauchen eine algebraische Darstellung von Mengen. Dafür stehen verschiedene Möglichkeiten zur Verfügung. Besonders gut eignet sich der duale Zugang mit Stützfunktionen oder Stützpunktmengen.

Im ersten Abschnitt soll gezeigt werden, wie die mengenwertige Iteration mit Hilfe von Stützpunkten im Rechner umgesetzt werden kann. Hier wird sich herausstellen, dass die Hauptarbeit dabei die Stützpunktberechnung für die Auswahlmenge $\mathcal{I}_{m,p}$ ist.

Im nächsten Abschnitt sollen einige Eigenschaften der Auswahlmenge $\mathcal{I}_{m,p}$ herausgearbeitet werden. Insbesondere eine lineare Transformation dieser Menge in die so genannte Momentenmenge. Damit kann die Stützpunktberechnung für die Auswahlmenge auf die Stützpunktberechnung für die Momentenmenge zurückgeführt werden.

Im zweiten Abschnitt wird erörtert wie die Stützpunkte für die Momentenmenge mittels dem Hausdorff-Momentenproblem berechnet werden können.
Und schließlich behandelt der letzte Abschnitt die Stützpunktberechnung für die Momentenmenge mittels der Darstellung dieser Menge als Aumann-Integral.

5.1. Mengenarithmetik mit Stützpunkten

In diesem Abschnitt werden wir intensiv die Ergebnisse aus dem Abschnitt 1.4 über die Stützpunktmenge nutzen, insbesondere die Rechenregeln.
Nach Satz 1.4.3 gilt für eine kompakte, konvexe und nichtleere Menge $K \subset \mathbb{R}^n$

$$K = \operatorname{co}\left(\bigcup_{\mathfrak{l} \in S_{n-1}} Y\left(\mathfrak{l}, K\right) \right) .$$

Man kann also aus der Kenntnis aller Stützpunktmengen die Originalmenge zurückgewinnen. Natürlich ist es in der Praxis nicht möglich die Stützpunktmengen aller Einheitsvektoren zu berechnen. Deshalb berechnet man sie nur in endlich vielen Richtungen und approximiert damit die Originalmenge.
Dies Konzept soll auf das obige mengenwertige Verfahren (5.1) angewandt werden. Die diskreten erreichbaren Mengen $\mathcal{R}_h\left(ih, x_0\right)$ sollen für alle $i = 0, \ldots, N$ hier zunächst als konvex, kompakt und nichtleer angenommen werden. Dies wird in Abschnitt 5.2 verifiziert.
Wir berechnen die Stützpunktmengen in einer beliebigen Richtung $\mathfrak{l} \in S_{n-1}$ und erhalten:

$$\begin{aligned} Y\left(\mathfrak{l}, \mathcal{R}_h\left(0, x_0\right)\right) &= \{x_0\} \\ Y\left(\mathfrak{l}, \mathcal{R}_h\left((i+1)h, x_0\right)\right) &= Y\left(\mathfrak{l}, \mathfrak{C} \cdot \mathcal{R}_h\left(ih, x_0\right) + \mathfrak{D} \cdot \mathcal{I}_{m,p}\right) \\ &= \mathfrak{C} \cdot Y\left(\mathfrak{C}^T\mathfrak{l}, \mathcal{R}_h\left(ih, x_0\right)\right) + \mathfrak{D} \cdot Y\left(\mathfrak{D}^T\mathfrak{l}, \mathcal{I}_{m,p}\right) , \end{aligned}$$

für $i = 0, \ldots, N-1$.
Um also die Stützpunktmenge $Y\left(\mathfrak{l}, \mathcal{R}_h\left(Nh, x_0\right)\right)$ an die erreichbare Menge zu berechnen, braucht man die Stützpunktmengen $Y\left(\mathfrak{C}^T\mathfrak{l}, \mathcal{R}_h\left((N-1)h, x_0\right)\right)$ und $Y\left(\mathfrak{D}^T\mathfrak{l}, \mathcal{I}_{m,p}\right)$. Wiederum um die Stützpunktmenge $Y\left(\mathfrak{C}^T\mathfrak{l}, \mathcal{R}_h\left((N-1)h, x_0\right)\right)$ zu berechnen braucht man die Stützpunktmengen $Y\left(\mathfrak{C}^T\left(\mathfrak{C}^T\mathfrak{l}\right), \mathcal{R}_h\left((N-2)h, x_0\right)\right)$ und $Y\left(\mathfrak{C}^T\left(\mathfrak{D}^T\mathfrak{l}\right), \mathcal{I}_{m,p}\right)$, usw. Mit dieser Information wollen wir einen Algorithmus formulieren um die Stützpunktmenge an die erreichbare Menge $\mathcal{R}_h\left(Nh, x_0\right)$ in Richtung $\mathfrak{l} \in S_{n-1}$ zu berechnen.

(i) Starte mit

$$\begin{aligned} \tilde{\mathfrak{l}} &\leftarrow \left(\mathfrak{C}^T\right)^N \cdot \mathfrak{l} \\ \tilde{\mathcal{R}} &\leftarrow Y\left(\tilde{\mathfrak{l}}, \mathcal{R}_h\left(0, x_0\right)\right) = \{x_0\} \end{aligned}$$

(ii) Für $k = 1, \ldots, N$ berechne

$$\begin{aligned} \tilde{\mathfrak{l}} &\leftarrow \left(\mathfrak{C}^T\right)^{N-k} \cdot \mathfrak{l} \\ \tilde{\mathcal{R}} &\leftarrow \mathfrak{C} \cdot \tilde{\mathcal{R}} + \mathfrak{D} \cdot Y\left(\mathfrak{D}^T \tilde{\mathfrak{l}}, \mathcal{I}_{m,p}\right) \end{aligned}$$

Zum Schluss ist dann $\tilde{\mathcal{R}} = Y\left(\mathfrak{l}, \mathcal{R}_h\left(Nh, x_0\right)\right)$, also die Stützpunktmenge an die erreichbare Menge in Richtung $\mathfrak{l}$.

Bemerkung 5.1.1.

(i) Da bei uns die Anfangsmenge stets einelementig ist, d.h. $\mathcal{R}_h\left(0, x_0\right) = \{x_0\}$, ist auch $Y\left(\tilde{\mathfrak{l}}, \mathcal{R}_h\left(0, x_0\right)\right) = \{x_0\} \quad \forall \tilde{\mathfrak{l}} \in \mathbb{R}^n$.

(ii) Die tatsächliche Implementierung der Algorithmen erfolgt nicht mit Stützpunktmengen, sondern mit deren Elementen den Stützpunkten. Dies führt zu einem geringeren Aufwand und hat folgende Rechtfertigung. In den allermeisten Fällen ist die Stützpunktmenge einelementig d.h. der Stützpunkt ist eindeutig. Falls der Stützpunkt nicht eindeutig ist, wird einer heuristisch ausgewählt (bei einer Strecke ist dies oft der Mittelpunkt). Dann ist das Endergebnis wieder in $\tilde{\mathcal{R}} = Y\left(\mathfrak{l}, \mathcal{R}_h\left(Nh, x_0\right)\right)$ enthalten, weil die Minkowski-Summe und die lineare Transformation von Mengen elementweise definiert sind. Und schließlich rechtfertigt Satz 1.4.3 Aussage 3 auch dieses Vorgehen.

Wir wollen den obigen Algorithmus noch ein wenig näher betrachten. Am Anfang braucht man $\left(\mathfrak{C}^T\right)^N$ um $\tilde{\mathfrak{l}}$ zu berechnen. Und $\tilde{\mathfrak{l}}$ braucht man nur um $Y\left(\mathfrak{D}^T \tilde{\mathfrak{l}}, \mathcal{I}_{m,p}\right)$ zu berechnen. Um die direkte Berechnung der Matrizen $\left(\mathfrak{C}^T\right)^N$ und $\left(\mathfrak{C}^T\right)^{N-k}$ zu vermeiden, kann man die Stützpunktmengen an $\mathcal{I}_{m,p}$ in umgekehrter Reihenfolge so berechnen:

Algorithmus 5.1.2. *Es sei eine Richtung $\mathfrak{l} \in S_{n-1}$ vorgegeben.*
Starte mit

$$\begin{aligned} \mathfrak{r} &\leftarrow \mathfrak{l} \\ \tilde{\mathcal{I}}_N &\leftarrow Y\left(\mathfrak{D}^T \mathfrak{r}, \mathcal{I}_{m,p}\right) \end{aligned}$$

Für $k = 1, \ldots, N$ berechne

$$\begin{aligned} \mathfrak{r} &\leftarrow \mathfrak{C}^T \cdot \mathfrak{r} \\ \tilde{\mathcal{I}}_{N-k} &\leftarrow Y\left(\mathfrak{D}^T \mathfrak{r}, \mathcal{I}_{m,p}\right) . \end{aligned}$$

Damit erhalten wir insgesamt folgenden Algorithmus:

Algorithmus 5.1.3. *Zu einer Richtung* $\mathfrak{l} \in S_{n-1}$ *berechnet man die Stützpunktmenge an* $\mathcal{R}_h(Nh, x_0)$ *folgendermaßen:*

Vorarbeit: Berechne die $\tilde{\mathcal{I}}_i$ $(i = 0, \ldots, N)$ *gemäß Algorithmus 5.1.2 in Richtung* $\mathfrak{l}$.

Starte mit

$$\tilde{\mathcal{R}} \leftarrow Y(\mathfrak{l}, \mathcal{R}_h(0, x_0)) = \{x_0\} \ .$$

Für $k = 1, \ldots, N$ *berechne nacheinander*

$$\tilde{\mathcal{R}} \leftarrow \mathfrak{C} \cdot \tilde{\mathcal{R}} + \mathfrak{D} \cdot \tilde{\mathcal{I}}_k \ .$$

Zum Schluss ist dann $\tilde{\mathcal{R}} = Y(\mathfrak{l}, \mathcal{R}_h(Nh, x_0))$ *die Stützpunktmenge an die diskrete erreichbare Menge in Richtung* $\mathfrak{l}$.

Um diesen Algorithmus durchführen zu können muss nun geklärt werden, wie die Stützpunkte an die Menge $\mathcal{I}_{m,p}$ in Algorithmus 5.1.2 berechnet werden können. Alles weitere in diesem Kapitel hat zum Ziel dafür ein effizientes Verfahren zu entwickeln. Es wird sich auch zeigen, das der Hauptaufwand von Algorithmus 5.1.3 in dieser Stützpunktberechnung liegt. Da dieser Algorithmus in möglichst vielen Richtungen $\mathfrak{l} \in S_{n-1}$ durchgeführt werden muss, ist es besonders wichtig für diese Stützpunktberechnung ein möglichst effizientes Verfahren einzusetzen.

5.2. Eigenschaften der Auswahlmenge

In diesem Abschnitt soll die Auswahlmenge als lineare Transformation der Momentenmenge dargestellt werden, welche wir in diesem Abschnitt einführen. Die Stützpunktberechnung für die Auswahlmenge, kann damit auf die Stützpunktberechnung für die Momentenmenge zurückgeführt werden. Außerdem sollen einige Eigenschaften über die Auswahlmenge $\mathcal{I}_{m,p}$ bereitgestellt werden.

Hier soll noch einmal an die Definition von $\mathcal{I}_{m,p}$ erinnert werden:

$$\mathcal{I}_{m,p} = \Bigg\{ (\zeta_1, \ldots, \zeta_p)^T \in \mathbb{R}^{mp} \mid \zeta_i^T = \int_0^1 \ldots \int_0^{s_{i-1}} \mathfrak{u}(s_i)\, ds_i \ldots ds_1 \quad (i = 1, \ldots, p) \\ , \mathfrak{u} \in L^1([0,1]; U) \Bigg\} \ ,$$

wobei U der Kontrollbereich ist. Da der Kontrollbereich als kompakte, konvexe und nichtleere Teilmenge des $\mathbb{R}^m$ vorausgesetzt ist, gilt automatisch $\mathcal{I}_{m,p} \subset \mathbb{R}^{pm}$.

Da der Umgang mit Mehrfachintegralen ziemlich schwer ist, wollen wir zunächst $\mathcal{I}_{m,p}$ anders darstellen. Diese neue Darstellung liefert das nächste Lemma.

Lemma 5.2.1. *Sei* $\mathfrak{u} \in L^1([0,a];\mathbb{R}^m)$ *und* $a > 0$. *Dann gilt für* $i \geq 2$

$$\begin{aligned}\int_0^a s^{i-1}\mathfrak{u}(s)\,ds = a^{i-1}\int_0^a \mathfrak{u}(s_1)\,ds_1 - (i-1)\,a^{i-2}\int_0^1\int_0^{s_1}\mathfrak{u}(s_2)\,ds_2\,ds_1 \\ + (i-1)(i-2)\,a^{i-3}\int_0^1\int_0^{s_1}\int_0^{s_2}\mathfrak{u}(s_3)\,ds_3\,ds_2\,ds_1 \\ - \ldots + (-1)^{i-1}(i-1)!\int_0^1\ldots\int_0^{s_{i-1}}\mathfrak{u}(s_i)\,ds_i\ldots ds_1\end{aligned}$$

Beweis. Sei $\mathfrak{u} \in L^1([0,a];\mathbb{R}^m)$ gegeben, wobei wir hier $\mathfrak{u}$ mit einem beliebigen Repräsentanten der Äquivalenzklasse identifizieren. Es ist $\mathfrak{v}(x) := \int_0^x \mathfrak{u}(t)\,dt$ nach Satz 1.8.6 absolutstetig. Deswegen ist $\mathfrak{v}(x)$ fast überall differenzierbar und eine Stammfunktion von $\mathfrak{u}$, d.h. es gilt für fast alle $x \in [0,a]$: $\mathfrak{v}'(x) = \mathfrak{u}(x)$. (siehe [25, Kapitel IX, §2 + §4]).

Es ist $x \mapsto x^i$ auf $[0,a]$ Lipschitz-stetig, denn die Ableitung ist auf $[0,a]$ beschränkt. Dann ist nach Satz 1.8.3(i) auch $x \mapsto x^i$ für $i \in \mathbb{N}$ absolutstetig.

Also gilt mit [25, Kapitel IX, §7, Satz 5] die Formel der partiellen Integration für x^i und jede Komponente von $\mathfrak{v}(x)$, und damit auch für die ganze vektorwertige Funktion.

Wir zeigen obige Formel mit Induktion. Dabei verwenden wir die partielle Integration.

Induktionsanfang $i = 2$:

$\int_0^a \underbrace{s^1}_{f}\underbrace{\mathfrak{u}(s)}_{g'}\,ds = \left[s\cdot\int_0^s \mathfrak{u}(t)\,dt\right]_0^a - \int_0^a 1\cdot\int_0^s \mathfrak{u}(t)\,dt\,ds = a\int_0^a \mathfrak{u}(t)\,dt - \int_0^a\int_0^s \mathfrak{u}(t)\,dt\,ds$

Induktionsvoraussetzung: Die Behauptung sei bewiesen für $j \leq (i-1)$.

Induktionsschluss: $(i-1) \rightsquigarrow i$.

$$\begin{aligned}\int_0^a \underbrace{s^{i-1}}_{f}\underbrace{\mathfrak{u}(s)}_{g'}\,ds &= \left[s^{i-1}\cdot\int_0^s \mathfrak{u}(t)\,dt\right]_0^a - \int_0^a (i-1)s^{i-2}\cdot\int_0^s \mathfrak{u}(t)\,dt\,ds \\ &= a^{i-1}\int_0^a \mathfrak{u}(t)\,dt - (i-1)\int_0^a s^{i-2}\cdot\int_0^s \mathfrak{u}(t)\,dt\,ds \\ &= a^{i-1}\int_0^a \mathfrak{u}(s)\,ds - (i-1)\int_0^a s^{i-2}\cdot\mathfrak{v}(s)\,ds \qquad (5.2)\end{aligned}$$

Aus $\mathfrak{v} \in \mathrm{AC}([0,a];\mathbb{R}^m)$ folgt $\mathfrak{v} \in L^1([0,a];\mathbb{R}^m)$. Also ist die Induktionsvoraussetzung für $\mathfrak{v}$ anwendbar, sie lautet:

$$\begin{aligned}\int_0^a s^{i-2}\cdot\mathfrak{v}(s)\,ds = a^{i-2}\int_0^a \mathfrak{v}(s_1)\,ds_1 - (i-2)a^{i-3}\int_0^a\int_0^{s_1}\mathfrak{v}(s_2)\,ds_2\,ds_1 \\ + \cdots + (-1)^{i-2}(i-2)!a^0\int_0^a\ldots\int_0^{s_{i-2}}\mathfrak{v}(s_{i-1})\,ds_{i-1}\ldots ds_1\end{aligned}$$

Diese Formel setzen wir nun in (5.2) ein und ersetzen dabei $\mathfrak{v}$ durch seine Definition. Dann erhalten wir

$$\begin{aligned}\int_0^a s^{i-1}\mathfrak{u}(s)\,ds = a^{i-1}\int_0^a \mathfrak{u}(s_1)\;ds_1 - (i-1)a^{i-2}\int_0^a\int_0^{s_1}\mathfrak{u}(s_2)\;ds_2\,ds_1 \\ + (i-1)(i-2)\,a^{i-3}\int_0^a\int_0^{s_1}\int_0^{s_2}\mathfrak{u}(s_3)\;ds_3\,ds_2\,ds_1 \\ - \ldots + (-1)^{i-1}(i-1)!a^0\int_0^a\ldots\int_0^{s_{i-1}}\mathfrak{u}(s_i)\;ds_i\ldots ds_1\end{aligned}$$

also genau das gewünschte. □

An dieser Stelle wollen wir nun die *Momentenmenge* $\mathcal{M}_{m,p}\subset\mathbb{R}^{pm}$ definieren als

$$\mathcal{M}_{m,p} := \left\{ \left(\mu_0^T,\ldots,\mu_{p-1}^T\right)^T \in \mathbb{R}^{pm} \mid \mu_i^T = \int_0^1 t^i\mathfrak{u}(t)\;dt \quad (i=0,\ldots,p-1) \right. \\ \left. ,\ \mathfrak{u}\in L^1\left([0,1];U\right)\right\}.$$

Im Folgenden seien die μ_j und ζ_k wie in den Mengen $\mathcal{M}_{m,p}$ und $\mathcal{I}_{m,p}$. Dann kann man obiges Lemma für $a=1$ so darstellen:

$$\mu_{i-1} = \zeta_1 - (i-1)\zeta_2 + (i-1)(i-2)\zeta_3 - \ldots + (-1)^{i-1}(i-1)!\zeta_i\,. \tag{5.3}$$

Wir wollen nun dafür eine Matrixschreibweise anstreben und deswegen definieren wir die untere Dreiecksmatrix

$$\mathfrak{T}_{m,p} := \begin{pmatrix} t_{1,1}\cdot\mathfrak{E}_m & 0\cdot\mathfrak{E}_m & 0\cdot\mathfrak{E}_m \\ \vdots & \ddots & 0\cdot\mathfrak{E}_m \\ t_{p,1}\cdot\mathfrak{E}_m & \cdots & t_{p,p}\cdot\mathfrak{E}_m \end{pmatrix} \in \mathbb{R}^{pm\times pm}$$

mit $t_{i,j} = \frac{(-1)^{j-1}(i-1)!}{(i-j)!}$ für $j\le i$. Jetzt können wir die Transformation der μ_j in die ζ_k in Matrixschreibweise darstellen

$$\begin{pmatrix}\mu_0\\ \vdots \\ \mu_{p-1}\end{pmatrix} = \mathfrak{T}_{m,p}\begin{pmatrix}\zeta_1\\ \vdots\\ \zeta_p\end{pmatrix},$$

dabei sorgen die Einheitsmatrizen in der Definition von $\mathfrak{T}_{m,p}$ dafür, dass alle Komponenten der ζ_k jeweils gleich behandelt werden, wie man das auch in Gleichung (5.3) sehen kann. Für $\mathfrak{T}_{1,p}$ schreiben wir kurz $\mathfrak{T}_p$.

Die Matrix $\mathfrak{T}_{m,p}$ ist eine untere Dreiecksmatrix mit nichtverschwindender Diagonale und daher invertierbar. Deswegen gilt ebenso

$$\mathfrak{T}_{m,p}^{-1} \begin{pmatrix} \mu_0 \\ \vdots \\ \mu_{p-1} \end{pmatrix} = \begin{pmatrix} \zeta_1 \\ \vdots \\ \zeta_p \end{pmatrix} .$$

Also haben wir für $\mathcal{I}_{m,p}$ folgende Darstellung :

$$\mathcal{I}_{m,p} = \mathfrak{T}_{m,p}^{-1} \cdot \mathcal{M}_{m,p}$$

Und für die diskrete Kontrollmenge aus Kapitel 3 haben wir dann die Darstellung

$$\mathcal{U}_{m,p} = \left(\Gamma_{m,p}^{-1} \cdot \Delta_{m,p} \cdot \mathfrak{T}_{m,p}^{-1}\right) \cdot \mathcal{M}_{m,p} .$$

Wir wollen nun noch eine kompaktere Schreibweise für die Menge $\mathcal{M}_{m,p}$ anstreben. Dazu definieren wir die matrixwertige Funktion

$$\mathfrak{Q}_p : [0,1] \longrightarrow \mathbb{R}^{mp \times m} \ , \ \mathfrak{Q}_p(t) := \begin{pmatrix} 1 \cdot \mathfrak{E}_m \\ t \cdot \mathfrak{E}_m \\ \vdots \\ t^{p-1} \cdot \mathfrak{E}_m \end{pmatrix} . \tag{5.4}$$

Diese ist offenbar stetig. Damit kann man ein Element $\left(\mu_0^T, \ldots, \mu_{p-1}^T\right)^T \in \mathcal{M}_{m,p}$ schreiben als

$$\left(\mu_0^T, \ldots, \mu_{p-1}^T\right)^T = \int_0^1 \mathfrak{Q}_p(t) \cdot \mathfrak{u}(t)\, dt$$

mit einem geeigneten $\mathfrak{u} \in L^1\left([0,1]; U\right)$. Weiter definieren wir folgende mengenwertige Abbildung

$$F : [0,1] \Longrightarrow \mathbb{R}^{mp} \ , \ F(t) := \mathfrak{Q}_p(t) \cdot U \ .$$

Dann ist $F(\cdot)$ nach Proposition 1.5.2 stetig, messbar und integrierbar beschränkt. Die integrierbaren Auswahlen haben die Gestalt $\mathfrak{Q}_p(t) \cdot \mathfrak{u}(t)$ mit $\mathfrak{u} \in L^1\left([0,1]; U\right)$. Damit können wir die Menge $\mathcal{M}_{m,p}$ als Aumann-Integral

$$\mathcal{M}_{m,p} = \int_0^1 \mathfrak{Q}_p(t) \cdot U\, dt$$

schreiben. Zum weiteren Vorgehen brauchen wir noch ein kleines

Lemma 5.2.2. *Seien $K, L \subset \mathbb{R}^k$ zwei Kompakta und $\mathfrak{M} \in \mathbb{R}^{l,k}$ eine Matrix $(l, k \in \mathbb{N})$. Dann sind auch*

$$\mathfrak{M} \cdot K \qquad \textit{und} \qquad K + L$$

kompakt.

Beweis. Das Bild eines Kompaktums unter einer stetigen Abbildung ist kompakt ([21, Satz auf S.31]). Mit der Abbildung $\mathfrak{x} \mapsto \mathfrak{M} \cdot \mathfrak{x}$, welche stetig ist, folgt die Kompaktheit von $\mathfrak{M} \cdot K$. Nach [21, Folgerung auf S. 32] ist die Menge $K \times L$ kompakt. Da die Addition reeller Vektoren $(\mathfrak{y}, \mathfrak{x}) \mapsto \mathfrak{x} + \mathfrak{y}$ eine stetige Abbildung ist, ist die Menge $K + L$ kompakt. □

Mit diesem Lemma und der Darstellung von $\mathcal{M}_{m,p}$ als Aumann-Integral können wir einige wichtige Eigenschaften der Mengen $\mathcal{M}_{m,p}$, $\mathcal{I}_{m,p}$ und $\mathcal{U}_{m,p}$ zeigen.

Proposition 5.2.3. *Es sei der Kontrollbereich U kompakt, konvex und nichtleer.*
Dann sind die Mengen $\mathcal{M}_{m,p}$, $\mathcal{I}_{m,p}$ und $\mathcal{U}_{m,p}$ ebenfalls kompakt, konvex und nichtleer für $m, p \geq 1$.

Beweis. Aufgrund des vorigen Lemmas sind die Mengen $\mathfrak{Q}_p(t) \cdot U$ für alle $t \in [0,1]$ kompakt, also abgeschlossen und beschränkt. Da U nichtleer ist, ist auch $\mathfrak{Q}_p(t) \cdot U$ stets nichtleer. Wir wissen bereits, dass die mengenwertige Abbildung $F(\cdot)$ stetig, messbar und integrierbar beschränkt ist. Damit können wir Proposition 1.5.4 anwenden und erhalten, dass $\mathcal{M}_{m,p}$ kompakt, konvex und nichtleer ist.

Nun ist aber $\mathcal{I}_{m,p} = \mathfrak{T}_{m,p}^{-1} \cdot \mathcal{M}_{m,p}$ also ist $\mathcal{I}_{m,p}$ offenbar nichtleer, da $\mathfrak{T}_{m,p}^{-1}$ eine bijektive Abbildung ist. Nach dem vorigen Lemma ist $\mathcal{I}_{m,p}$ kompakt. Und nach Proposition 1.1.7 ist $\mathcal{I}_{m,p}$ auch konvex. Also ist $\mathcal{I}_{m,p}$ kompakt, konvex und nichtleer.

Weiter ist $\mathcal{U}_{m,p} = \left(\Gamma_{m,p}^{-1} \cdot \Delta_{m,p}\right) \cdot \mathcal{I}_{m,p}$ und $\Gamma_{m,p}^{-1} \cdot \Delta_{m,p}$ ist wieder eine bijektive Abbildung. Mit den gleichen Argumenten wie vorher folgt, dass auch $\mathcal{U}_{m,p}$ kompakt, konvex und nichtleer ist. □

In Abschnitt 5.1 haben wir vorausgesetzt, dass die Mengen $\mathcal{R}_h\left(ih, x_0\right)\ (i = 0, \ldots, N)$ kompakt, konvex und nichtleer sind. Dies soll hier verifiziert werden. Es ist die Menge $\mathcal{R}_h\left(0, x_0\right)$ einelementig und deswegen kompakt, konvex und offenbar nichtleer.

Weiter ist mit Lemma 5.2.2 und obiger Proposition die Menge $\mathfrak{D} \cdot \mathcal{I}_{m,p}$ aus (5.1) konvex und kompakt. Mittels weiterer Anwendung dieses Lemmas und der Proposition 1.1.7 erhält man, dass $\mathfrak{C} \cdot \mathcal{R}_h\left(h, x_0\right)$ und $\mathfrak{C} \cdot \mathcal{R}_h\left(h, x_0\right) + \mathfrak{D} \cdot \mathcal{I}_{m,p}$ aus (5.1) kompakt und konvex sind. Durch induktive Anwendung dieses Argumentes sind dann alle $\mathcal{R}_h\left(ih, x_0\right)\ (i = 0, \ldots, N)$ kompakt und konvex. Da $\mathcal{R}_h\left(0, x_0\right)$ und $\mathcal{I}_{m,p}$ nichtleer sind, sind auch alle $\mathcal{R}_h\left(ih, x_0\right)\ (i = 0, \ldots, N)$ nichtleer.

Damit ist die Darstellung der diskreten erreichbaren Mengen $\mathcal{R}_h\left(ih, x_0\right)$ mittels ihrer Stützpunktmengen gerechtfertigt, denn Satz 1.4.3 ist für sie anwendbar.

Mit der Darstellung $\mathcal{I}_{m,p} = \mathfrak{T}_{m,p}^{-1} \cdot \mathcal{M}_{m,p}$ kann die Berechnung der Stützpunktmenge an $\mathcal{I}_{m,p}$ auf $\mathcal{M}_{m,p}$ zurückgeführt werden. In Algorithmus 5.1.2 hat man dann mit den Rechenregeln für Stützpunktmengen $Y\left(\mathfrak{D}^T \mathfrak{r},\, \mathcal{I}_{m,p}\right) = \mathfrak{T}_{m,p}^{-1} \cdot Y\left(\left(\mathfrak{T}_{m,p}^{-1}\right)^T \mathfrak{D}^T \mathfrak{r},\, \mathcal{M}_{m,p}\right)$.
Mit der Matrix $\mathfrak{Z}_{m,p} := \mathfrak{D}\left(\mathfrak{T}_{m,p}^{-1}\right)$ als Abkürzung schreibt sich das als $Y\left(\mathfrak{D}^T \mathfrak{r},\, \mathcal{I}_{m,p}\right) = \mathfrak{T}_{m,p}^{-1} \cdot Y\left(\mathfrak{Z}_{m,p}^T \mathfrak{r},\, \mathcal{M}_{m,p}\right)$.

Es muss also jetzt noch geklärt werden, wie die Stützpunkte an die Menge $\mathcal{M}_{m,p}$ möglichst effizient berechnet werden können. Die nächsten beiden Abschnitte stellen dafür zwei verschiedene Ansätze vor.

5.3. Berechnung durch das Hausdorff-Momentenproblem

In diesem Abschnitt werden wir zunächst dem Artikel Ferretti's folgen und mit Hilfe des endlichen Hausdorff-Problems die Menge $\mathcal{M}_{m,p}$ nun nicht als Aumann-Integral sondern algebraisch darstellen. Dies geschieht im ersten Unterabschnitt. In den nächsten beiden Unterabschnitten werden zwei Ansätze vorgestellt wie man mit dieser algebraischen Darstellung von $\mathcal{M}_{m,p}$ dann Stützpunkte berechnen kann. Dies geht dann wieder über Ferretti's Arbeit hinaus. Beide Ansätze habe ich bei der Implementierung der mengenwertigen Verfahren nicht benutzt, da sie zu unflexibel und kompliziert sind. Außerdem sind sie beide nicht effektiver als die Methode, die in Abschnitt 5.4 vorgestellt wird.

Wir setzen in diesem Abschnitt den Kontrollbereich U als das reelle Einheitsintervall voraussetzen, d.h. $U = [0, 1]$, außer U wird explizit anders angegeben . Entsprechend sind dann auch die Mengen $\mathcal{M}_{m,p}$, $\mathcal{I}_{m,p}$ und $\mathcal{U}_{m,p}$ Teilmengen des $\mathbb{R}^p$, d.h. $m = 1$.

5.3.1. Das Hausdorff-Momentenproblem

Definition 5.3.1. Es sei eine Sequenz von reellen Zahl $(\mu_k)_{k=0,\dots,p-1}$ gegeben. Das *endliche* (oder unvollständige) *Hausdorff-Momentenproblem* besteht nun darin, zu entscheiden, ob es ein $u \in L^1([0, 1]; [0, 1])$ gibt, sodass diese Folge als Sequenz von Momenten

$$\mu_k = \int_0^1 t^k u(t)\, dt$$

dargestellt werden kann.

Bemerkung 5.3.2.

(i) Das Attribut endlich bzw. unvollständig bezieht sich darauf, das hier nur eine endliche Folge reeller Zahlen betrachtet wird. Für eine unendliche Folge reeller Zahlen heißt dieses Problem dann nur Hausdorff-Momentenproblem.

(ii) Dieses Problem wurde zuerst von Hausdorff in [19] vorgestellt und auch gelöst. Allerdings benötigt er selbst für den endlichen Fall eine unendliche Anzahl von Ungleichungen.

Wir werden hier die bessere Lösung von Ghizzetti aus [18] vorstellen wie sie Ferretti in seinem Artikel [13] darstellt. Dazu müssen wir zunächst etwas umfangreichere Definitionen vornehmen.

Es sei eine reelle Folge $(\mu_k)_{k=0,\ldots,p-1}$ gegeben. Dann werden die Variablen σ_k, δ_k und σ_k^* definiert durch

$$\begin{aligned}
\sigma_0 &= \mu_0 \\
2\sigma_1 &= -\mu_0\sigma_0 + 2\mu_1 \\
3\sigma_2 &= -\mu_0\sigma_1 - 2\mu_1\sigma_0 + 3\mu_2 \\
&\vdots \\
(k+1)\sigma_k &= -\mu_0\sigma_{k-1} - 2\mu_1\sigma_{k-2} - \ldots - k\mu_{k-1}\sigma_0 + (k+1)\mu_k
\end{aligned} \tag{5.5}$$

und

$$\begin{aligned}
\delta_0 &= \mu_0 \\
2\delta_1 &= \mu_0\delta_0 + 2\mu_1 \\
3\delta_2 &= \mu_0\delta_1 + 2\mu_1\delta_0 + 3\mu_2 \\
&\vdots \\
(k+1)\delta_k &= \mu_0\delta_{k-1} + 2\mu_1\delta_{k-2} + \ldots + k\mu_{k-1}\delta_0 + (k+1)\mu_k
\end{aligned} \tag{5.6}$$

und schließlich

$$\begin{aligned}
\sigma_0^* &= 1 - \delta_0 \\
\sigma_1^* &= \delta_0 - \delta_1 \\
\sigma_2^* &= \delta_1 - \delta_2 \\
&\vdots \\
\sigma_k^* &= \delta_k - \delta_{k-1}\,.
\end{aligned} \tag{5.7}$$

Desweiteren definieren wir die rationalen Funktionen Φ_k, $\Psi_k \quad (k = 1, \ldots, p-1)$ für gerade Indizes

$$\Phi_{2n}\left(\mu_0, \mu_1, \ldots, \mu_{2n-1}\right) := \frac{1}{2n+1}\left(\mu_0\sigma_{2n-1} + 2\mu_1\sigma_{2n-2} + \ldots + 2n\mu_{2n-1}\sigma_0\right) - \frac{\begin{vmatrix} \sigma_0 & \cdots & \sigma_{n-1} & \sigma_n \\ \vdots & & \vdots & \vdots \\ \sigma_{n-1} & \cdots & \sigma_{2n-2} & \sigma_{2n-1} \\ \sigma_n & \cdots & \sigma_{2n-1} & 0 \end{vmatrix}}{\begin{vmatrix} \sigma_0 & \cdots & \sigma_{n-1} \\ \vdots & & \vdots \\ \sigma_{n-1} & \cdots & \sigma_{2n-2} \end{vmatrix}}$$

und

$$\Psi_{2n}\left(\mu_0,\mu_1,\ldots,\mu_{2n-1}\right) := \left(1-\frac{\mu_0}{2n+1}\right)\delta_{2n-1}-\frac{2\mu_1}{2n+1}\delta_{2n-2}-\frac{3\mu_2}{2n+1}\delta_{2n-3} - \cdots - \frac{2n\mu_{2n-1}}{2n+1}\delta_0 + \frac{\begin{vmatrix} \sigma_0^* & \cdots & \sigma_{n-1}^* & \sigma_n^* \\ \vdots & & \vdots & \vdots \\ \sigma_{n-1}^* & \cdots & \sigma_{2n-2}^* & \sigma_{2n-1}^* \\ \sigma_n^* & \cdots & \sigma_{2n-1}^* & 0 \end{vmatrix}}{\begin{vmatrix} \sigma_0^* & \cdots & \sigma_{n-1}^* \\ \vdots & & \vdots \\ \sigma_{n-1}^* & \cdots & \sigma_{2n-2}^* \end{vmatrix}}$$

und für ungerade Indizes

$$\Phi_{2n-1}\left(\mu_0,\mu_1,\ldots,\mu_{2n-2}\right) := \frac{1}{2n}\left(\mu_0\sigma_{2n-2}+2\mu_1\sigma_{2n-1}+\ldots+(2n-1)\mu_{2n-2}\sigma_0\right) - \frac{\begin{vmatrix} \sigma_1 & \cdots & \sigma_{n-1} & \sigma_n \\ \vdots & & \vdots & \vdots \\ \sigma_{n-1} & \cdots & \sigma_{2n-3} & \sigma_{2n-2} \\ \sigma_n & \cdots & \sigma_{2n-2} & 0 \end{vmatrix}}{\begin{vmatrix} \sigma_1 & \cdots & \sigma_{n-1} \\ \vdots & & \vdots \\ \sigma_{n-1} & \cdots & \sigma_{2n-3} \end{vmatrix}}$$

und

$$\Psi_{2n-1}\left(\mu_0,\mu_1,\ldots,\mu_{2n-2}\right) := -\frac{1}{2n}\left(\mu_0\delta_{2n-2}+2\mu_1\delta_{2n-1}+\ldots+(2n-1)\mu_{2n-2}\delta_0\right) - \frac{\begin{vmatrix} 1 & 1 & \cdots & 1 & 1 \\ \delta_0 & \delta_1 & \cdots & \delta_{n-1} & \delta_n \\ \vdots & \vdots & & \vdots & \vdots \\ \delta_{n-2} & \delta_{n-1} & \cdots & \delta_{2n-3} & \delta_{2n-1} \\ \delta_{n-1} & \delta_n & \cdots & \delta_{2n-2} & 0 \end{vmatrix}}{\begin{vmatrix} 1 & 1 & \cdots & 1 \\ \delta_0 & \delta_1 & \cdots & \delta_{n-1} \\ \vdots & \vdots & & \vdots \\ \delta_{n-2} & \delta_{n-1} & \cdots & \delta_{2n-3} \end{vmatrix}}.$$

Damit ist für eine vorgegebene reelle Folge $(\mu_k)_{k=0,\dots,p-1}$ das Momentenproblem genau dann lösbar, wenn $\mu_0 \in [0,1]$ ist und die übrigen μ_k folgende Ungleichungen erfüllen

$$\Phi_k(\mu_0,\dots,\mu_{k-1}) \le \mu_k \le \Psi_k(\mu_0,\dots,\mu_{k-1}) \quad k = 1,\dots,p-1\,. \tag{5.8}$$

Damit definieren wir die *Hausdorff-Menge* als

$$\mathcal{H}_p := \left\{ (\mu_0,\dots,\mu_{p-1})^T \in \mathbb{R}^p \mid \mu_0 \in [0,1]\,,\ \mu_k \text{ erfüllt (5.8) für } k = 1,\dots,p-1 \right\}.$$

Aufgrund der Definition von $\mathcal{M}_{m,p}$ gilt dann $\mathcal{M}_{1,p} = \mathcal{H}_p$, wenn $U = [0,1]$ ist. Die Funktionen Φ_k und Ψ_k sind rationale Funktionen. Also hat man mit $\mathcal{H}_p$ eine algebraische Darstellung von $\mathcal{M}_{1,p}$, zumindest wenn $U = [0,1]$ ist.

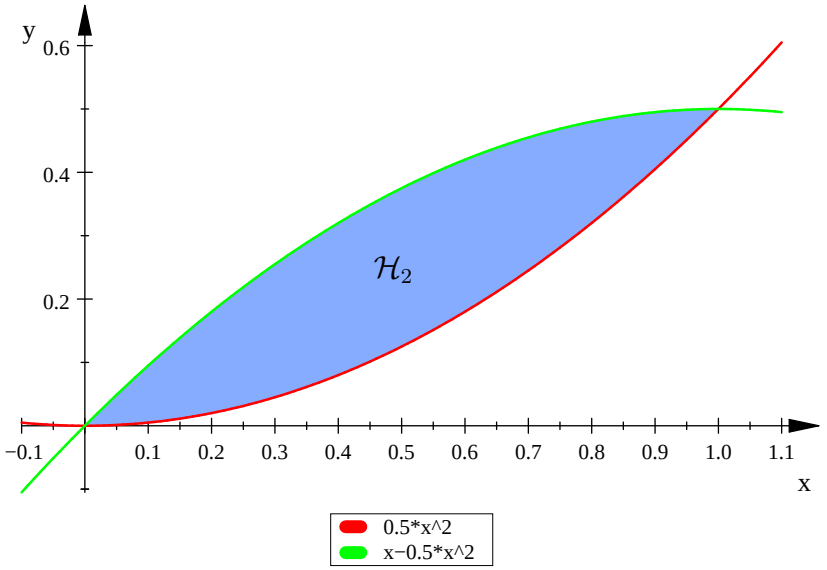

Abbildung 5.1.: Die Hausdorff-Menge $\mathcal{H}_2$

Beispiel 5.3.3. Für $k = 1, 2$ soll dieses Ergebnis hier konkret dargestellt werden. Es ist $\Phi_1(\mu_0) = \frac{1}{2}\mu_0^2$ und $\Psi_1(\mu_0) = -\frac{1}{2}\mu_0^2 + \mu_0$. Damit ist

$$\mathcal{H}_2 = \left\{ \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix} \in \mathbb{R}^2 \mid 0 \le \mu_0 \le 1\,,\ \frac{1}{2}\mu_0^2 \le \mu_1 \le -\frac{1}{2}\mu_0^2 + \mu_0 \right\}.$$

Diese Menge ist in Abbildung 5.1 dargestellt.
Weiter ist

$$\Phi_2(\mu_0,\mu_1) = \frac{1}{12\mu_0}\left(\mu_0^4 + 12\mu_1^2\right)$$

und

$$\Psi_2(\mu_0,\mu_1) = \frac{1}{12(\mu_0-1)}\left(\mu_0^4 - 4\mu_0^3 + 6\mu_0^2 + 12\mu_1^2 + 12\mu_1\right) .$$

Damit gilt dann

$$\mathcal{H}_3 = \left\{(\mu_0,\mu_1,\mu_2)^T \in \mathbb{R}^3 \mid (\mu_0,\mu_1)^T \in \mathcal{H}_2,\ \Phi_2(\mu_0,\mu_1) \le \mu_2 \le \Psi_2(\mu_0,\mu_1)\right\} .$$

Für ein beliebig es reelles Intervall $[a,b]$ gilt $[a,b] = (b-a)[0,1] + a$ bzw. $[0,1] = \frac{1}{b-a}[a,b] - \frac{a}{b-a}$. Für $U = [a,b]$ folgt dann $\mathcal{M}_{1,p} = (b-a)\mathcal{H}_p + a\left(1,\frac{1}{2},\ldots,\frac{1}{p}\right)^T$. Denn für $u \in L^1([0,1];[a,b])$ ist $\tilde{u}(t) := \frac{1}{b-a}u(t) - \frac{a}{b-a}$ in $L^1([0,1];[0,1])$ und es gilt $u(t) = (b-a)\tilde{u}(t) + a$. Damit folgt dann

$$\int_0^1 t^k u(t)\,dt = (b-a)\int_0^1 t^k\tilde{u}(t)\,dt + a\cdot\frac{1}{k+1} .$$

Ist der Kontrollbereich U ein Quader $[a_1,b_1]\times\cdots\times[a_m,b_m] \subset \mathbb{R}^m$ dann gilt

$$\left(\mu_0^T,\ldots,\mu_{p-1}^T\right)^T \in \mathcal{M}_{m,p} \iff$$
$$(\mu_{i,0},\ldots,\mu_{i,p-1})^T \in ((b_i-a_i)\mathcal{H}_p + a_i\cdot\mathfrak{v}_p) \quad (i=1,\ldots,m),$$

wobei $\mu_{i,k}$ die i-te Komponente des Vektors μ_k ist und $\mathfrak{v}_p = \left(1,\frac{1}{2},\ldots,\frac{1}{p}\right)^T$ ist. Auf diese Weise kann die Menge $\mathcal{M}_{m,p}$ für einen derartigen Quader U auf die Menge $\mathcal{H}_p$ zurückgeführt werden. Andere Steuerbereiche wie Kugel, Ellipsen oder andere geometrische Objekte lassen sich mit diesem Ansatz nicht behandeln.

5.3.2. Stützpunktberechnung durch das Hausdorff-Momentenproblem

In diesem Unterabschnitt sollen kurz zwei Konzepte vorgestellt werden, wie die Stützpunktmenge für $\mathcal{H}_p$ berechnet werden kann. Beide Ansätze beruhen darauf, dass ein Stützpunkt Lösung eines Optimierungsproblems ist. Die Begriffe und Sätze aus der Optimierung die wir hier brauchen, kann man in [15, Abschnitt 2.2] finden.
Es sei $\mathfrak{l} \in \mathbb{R}^p$, $\mathfrak{l} \neq 0_{\mathbb{R}^p}$. Dann gilt nach Definition

$$Y(\mathfrak{l},\mathcal{H}_p) = \left\{\mathfrak{a}\in\mathcal{H}_p \mid \langle\mathfrak{l},\mathfrak{a}\rangle = \sup_{\mathfrak{v}\in\mathcal{H}_p}\langle\mathfrak{l},\mathfrak{v}\rangle\right\} .$$

Die Stützpunkte, sind also genau die Punkte, für die das Supremum angenommen wird. Wir wissen bereits, dass $\mathcal{M}_{m,p}$ kompakt, konvex und nichtleer ist. Da $\mathcal{H}_p = \mathcal{M}_{1,p}$ für $U = [0,1]$ ist, ist auch $\mathcal{H}_p$ kompakt, konvex und nichtleer. Wir können also das

Supremum durch ein Maximum ersetzten. Es ist dann ein Stützpunkt $y(\mathfrak{l}, \mathcal{H}_p) =: \bar{\mathfrak{x}}$ Lösung des folgenden Optimierungsproblems

$$\begin{array}{rcccl} & & \max_{\mathfrak{x}\in\mathbb{R}^p} \mathfrak{l}^T\mathfrak{x} & & s.t. \\ 0 & \leq & \mathfrak{x}_1 & \leq & 1 \\ \Phi_1(\mathfrak{x}_1) & \leq & \mathfrak{x}_2 & \leq & \Psi_1(\mathfrak{x}_1) \\ & & \vdots & & \\ \Phi_{p-1}(\mathfrak{x}_1,\ldots,\mathfrak{x}_{p-1}) & \leq & \mathfrak{x}_p & \leq & \Psi_{p-1}(\mathfrak{x}_1,\ldots,\mathfrak{x}_{p-1}) \,, \end{array}$$

dabei entsprechen die Nebenbedingungen der Forderung, dass $\mathfrak{x}$ in $\mathcal{H}_p$ sein muss.

Es sollen nun zwei Möglichkeiten vorgestellt werden, wie dieses Optimierungsproblem gelöst werden kann.

5.3.2.1. Die Karush-Kuhn-Tucker Bedingungen

Die Karush-Kuhn-Tucker (KKT) Bedingungen sind ein notwendiges Kriterium für ein lokales Maximum (allgemein Extremum), vorausgesetzt eine so genannte constraint qualification ist erfüllt. Jeder Punkt der diesen Bedingungen genügt heißt KKT-Punkt. Bei einem konvexen Optimierungsproblem sind die KKT-Bedingungen sogar hinreichend und jeder KKT-Punkt ist ein globales Maximum (Extremum). Da das obige Optimierungsproblem konvex ist, können wir mit den KKT-Bedingungen Maxima finden. Man kann sich leicht überlegen, dass die LICQ hier stets erfüllt sein muss (nur Ungleichungsnebenbedingungen, pro Zeile kann nur eine Restriktion aktiv sein, pro Zeile kommt eine Variable hinzu). In diesem konkreten Fall muss man noch einiges an Zusatzinformationen hineinstecken um mit den KKT-Bedingungen eindeutig Maxima bestimmen zu können.

Im folgend wollen wir dies für $p = 2$ ausarbeiten. Wir geben einen Richtungsvektor $\mathfrak{l} = (l_1, l_2)^T \in S_1$ vor, für den wir den Stützpunkt berechnen wollen. In diesem Fall sieht das Optimierungsproblem so aus

$$\begin{array}{rcccl} & & \max\limits_{\mathfrak{x}\in\mathbb{R}^2} \mathfrak{l}^T\mathfrak{x} & & s.t. \\ 0 & \leq & \mathfrak{x}_1 & \leq & 1 \\ \frac{1}{2}\mathfrak{x}_1^2 & \leq & \mathfrak{x}_2 & \leq & -\frac{1}{2}\mathfrak{x}_1^2 + \mathfrak{x}_1 \,. \end{array} \tag{5.9}$$

Beide Funktionen $g_1(\mathfrak{x}) := \frac{1}{2}\mathfrak{x}_1^2 - \mathfrak{x}_2$ und $g_2(\mathfrak{x}) := \frac{1}{2}\mathfrak{x}_1^2 - \mathfrak{x}_1 - \mathfrak{x}_2$ haben die positiv semidefinite Matrix $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ als Hessematrix. Nach [21, Konvexitätskriterium, S.74] sind also beide Funktionen auf ganz $\mathbb{R}^2$ konvex. Es handelt sich also wirklich um ein konvexes Problem. Man kann leicht sehen, dass die LICQ stets erfüllt ist. Um eindeutige KKT-Bedingungen zu bekommen müssen wir ein wenig Arbeit hineinstecken. Zunächst

wissen wir, dass wegen der linearen Zielfunktion, das Maximum auf dem Rand von $\mathcal{H}_2$ liegen muss. Andernfalls gäbe es eine ϵ- Kugel um das Maximum, die Teilmenge von $\mathcal{H}_2$ wäre, und die Zielfunktion könnte noch vergrößert werden. Also gilt für das Maximum $g_1(\bar{\mathfrak{x}}) = 0$ oder $g_2(\bar{\mathfrak{x}}) = 0$. Falls $l_2 \leq 0$ ist, muss für das Maximum $g_1(\bar{\mathfrak{x}}) = 0$ gelten, da dies der untere Rand von $\mathcal{H}_2$ ist (siehe Abbildung 5.1). Oder mathematisch ausgedrückt: Es ist

$$\mathfrak{l}^T \begin{pmatrix} \mathfrak{x}_1 \\ \frac{1}{2}\mathfrak{x}_1^2 \end{pmatrix} \geq \mathfrak{l}^T \begin{pmatrix} \mathfrak{x}_1 \\ -\frac{1}{2}\mathfrak{x}_1^2 + \mathfrak{x}_1 \end{pmatrix} \qquad \text{für } \mathfrak{x}_1 \in [0,1] \text{ und } l_2 \leq 0\,.$$

Falls $l_2 > 0$ ist, gilt $g_2(\bar{\mathfrak{x}}) = 0$. Im folgenden betrachten wir den Fall $l_2 \leq 0$. Jetzt sieht unser Optimierungsproblem so aus

$$\begin{array}{rcll} \max_{\mathfrak{x}\in\mathbb{R}^2} \mathfrak{l}^T\mathfrak{x} & & s.t. \\ \mathfrak{x}_1 - 1 & \leq & 0 \\ -\mathfrak{x}_1 & \leq & 0 \\ g_1(\mathfrak{x}) & = & 0\,. \end{array}$$

Dafür lauten die KKT-Bedingungen:

$$\begin{array}{lcll} \text{I)} & -\mathfrak{l} + \lambda_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} -1 \\ 0 \end{pmatrix} + \mu \begin{pmatrix} \bar{\mathfrak{x}}_1 \\ -1 \end{pmatrix} & = 0 & \\ \text{II)} & \bar{\mathfrak{x}}_2 = \frac{1}{2}\bar{\mathfrak{x}}_1^2 & & 0 \leq \bar{\mathfrak{x}}_1 \leq 1 \\ \text{III)} & \lambda_1(\bar{\mathfrak{x}}_1 - 1) = 0,\ \lambda_1 \geq 0 & & \lambda_2\bar{\mathfrak{x}}_1 = 0,\ \lambda_2 \geq 0\,. \end{array}$$

Aus I) ergibt sich sofort

$$\text{IV)} \quad l_1 = \lambda_1 - \lambda_2 + l_2\mathfrak{x}_1\,.$$

1. Fall: $\lambda_1 = \lambda_2 = 0$

Dann folgt aus IV) sofort $\bar{\mathfrak{x}}_1 = \frac{l_1}{l_2}$. Falls $l_2 = 0$ oder $\bar{\mathfrak{x}}_1 \notin [0,1]$ muss Fall 2 angewendet werden (man kennt ja λ_1 und λ_2 nicht).

2. Fall: $\lambda_1 > 0$ oder $\lambda_2 > 0$

Dann folgt aus III): entweder ist $\bar{\mathfrak{x}}_1 = 1$ oder $\bar{\mathfrak{x}}_1 = 0$. Mit II) ist dann $\bar{\mathfrak{x}} = (0,0)^T$ oder $\bar{\mathfrak{x}} = \left(1, \frac{1}{2}\right)^T$. Dies lässt sich am besten durch konkretes Ausrechnen der Zielfunktion ermitteln.

Man sieht schon, dass es selbst in diesem einfachen Fall, nicht leicht ist konkrete Formeln zu bekommen. Ohne Zusatzinformationen kommt man nicht zum Ziel. Für wachsendes p sind noch wesentlich mehr Schwierigkeiten und Fallunterscheidungen zu erwarten. Außerdem verliert man für $p > 3$ die Anschauung als Hilfmittel. Desweiteren hat sich gezeigt, dass Nullstellen von Polynomen vom Grad $p - 1$ zu berechnen sind. All diese Nachteile haben mich bewogen diesen Zugang nicht zu wählen.

5.3.2.2. Das äußere Einheitsnormalenfeld

Der Begriff "Einheitsnormalenfeld" stammt aus der Integrationstheorie für Hyperflächen, er wird in [21, Abschnitt 12.1] erläutert. Im Folgenden soll diese Idee für $p = 2$ vorgestellt werden, d.h. für $\mathcal{H}_2$. Jedoch soll nur das Konzept anhand dieses Spezialfalls erläutert werden.

Wir geben wieder eine Richtung $\mathfrak{l} = (l_1, l_2)^T \in S_1$ vor, in der der Stützpunkt an $\mathcal{H}_2$ berechnet werden soll. Wiederum gehen wir vom Optimierungsproblem (5.9) aus und setzten ohne Einschränkung voraus, dass $l_2 < 0$ ist. Denn der Fall $l_2 = 0$ lässt sich leicht extra behandeln und der Fall $l_2 > 0$ geht analog. Wir wissen schon, dass der gesuchte Stützpunkt auf der unteren Kurve $\mathcal{T}$, definiert durch $\mathcal{T} := \{\mathfrak{x} \in \mathbb{R}^2 \mid g_1(\mathfrak{x}) = 0\}$, liegen muss. Aber $\mathcal{T}$ ist Graph der Funktion $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \frac{1}{2}x^2$. Diese Funktion ist streng konvex, denn $f''(x) = 1$ (siehe [20, Abschnitt 9.7, Folgerung 1]).

Das folgende Lemma wollen wir auf f anwenden. Es wird jedoch für Funktionen mit mehreren reellen Veränderlichen formuliert, um anzudeuten, dass dieser Zugang auch auf höhere Dimensionen übertragen werden kann, d.h. für $p > 2$.

Lemma 5.3.4. *Sei $e : \mathbb{R}^k \to \mathbb{R}$ streng konvex, $e \in \mathcal{C}^2(\mathbb{R}^k)$ und $\mathfrak{z} \in \mathbb{R}^k$ ein beliebiger Vektor. Dann verläuft die Linearisierung bzw. Tangente in $\mathfrak{z}$ stets unterhalb der Funktion, d.h.*

$$e(\mathfrak{x}) > e(\mathfrak{z}) + \nabla e(\mathfrak{z})^T (\mathfrak{x} - \mathfrak{z}) \qquad \textit{für} \quad \mathfrak{x} \neq \mathfrak{z}\,.$$

Für jeden anderen Vektor $\mathfrak{v} \in \mathbb{R}^k$ gibt es eine Richtung $\mathfrak{h} \in \mathbb{R}^k$, sodass für $t \in (0, 1)$

$$e(\mathfrak{x}) < e(\mathfrak{z}) + \mathfrak{v}^T (t\mathfrak{h})$$

ist.

Beweis. Zum Beweis für den ersten Teil siehe [26, Abschnitt 42, Theorem A].

Nun zum zweiten Teil. Es ist $\nabla e(\mathfrak{z}) - \mathfrak{v} \neq 0_{\mathbb{R}^k}$. Also gibt es einen Vektor $\hat{\mathfrak{h}} \in \mathbb{R}^k$ mit $(\nabla e(\mathfrak{z}) - \mathfrak{v})^T \hat{\mathfrak{h}} < 0$. Also gilt

$$e(\mathfrak{z}) + \nabla e(\mathfrak{z})^T \left(t\hat{\mathfrak{h}}\right) < e(\mathfrak{z}) + \mathfrak{v}^T \left(t\hat{\mathfrak{h}}\right) \quad \text{für } t > 0.$$

Nach dem Taylor'schen Satz (siehe [20, Abschnitt 14.1, Folgerung 2]) gilt aber

$$e\left(\mathfrak{z} + t\hat{\mathfrak{h}}\right) = e(\mathfrak{z}) + \nabla e(\mathfrak{z})^T \left(t\hat{\mathfrak{h}}\right) + \mathcal{O}\left(t^2\right)\,.$$

Also gibt es ein $t_0 > 0$, sodass

$$e\left(\mathfrak{z} + t\hat{\mathfrak{h}}\right) < e(\mathfrak{z}) + \mathfrak{v}^T \left(t\hat{\mathfrak{h}}\right) \quad \text{für } t \in (0, t_0)\,.$$

Setzte nun $\mathfrak{h} := \frac{1}{t_0}\hat{\mathfrak{h}}$. □

Dieses Lemma sagt aus, dass die Linearisierung unserer streng konvexen und zweimal stetig differenzierbaren Funktion f in einem Punkt z stets unterhalb der Funktion verläuft (außer in z selbst). Für jede andere lineare Funktion die ebenfalls durch $(z, f(z))^T$ geht trifft dies nicht zu.
Sei $z \in (0,1)$ und $a \in \mathbb{R}$. Dann definieren wir eine Schar affiner Funktionen durch den Punkt $(z, f(z))^T$ mittels

$$g_{z,a} : \mathbb{R} \to \mathbb{R}\,, \quad g_{z,a}(x) := f(z) + a(x - z)\ .$$

Jeder dieser affinen Funktionen kann man eindeutig den äußeren Einheitsnormalenvektor zuordnen

$$\mathfrak{n} = \mathfrak{n}(a) := \frac{1}{\sqrt{1+a^2}} \begin{pmatrix} a \\ -1 \end{pmatrix}\ .$$

Das ist genau der Einheitsvektor, der senkrecht auf der Gerade, definiert durch $g_{z,a}(x) - z = 0$, steht und aus der Menge $\mathcal{H}_2$ heraus zeigt.

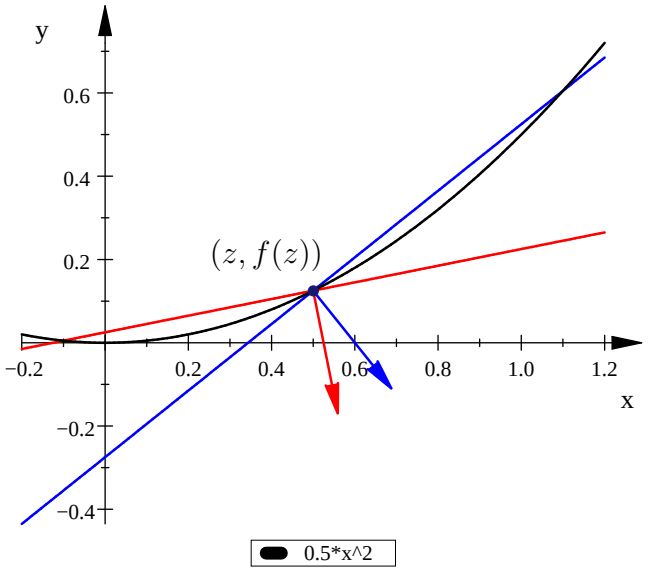

Abbildung 5.2.: Verschiedene affine Geraden mit ihren Normalenvektoren durch den Punkt $(0.5, 0.125)$ des unteren Randes von $\mathcal{H}_2$

In Abbildung 5.2 sieht man für $z = 0.5$, $a = 0.2$ (rot) und $a = 0.8$ (blau) zwei solcher affiner Funktionen zusammen mit ihren äußeren Normalen. Beide Geraden sind keine Tangenten und verlaufen deshalb nicht immer unterhalb von f.
Für $a = f'(z)$ ist $g_{z,f'(z)}$ gerade die Tangente an den Graphen von f und $\mathfrak{n} = \mathfrak{n}(f'(z))$ ist ihr Normalenvektor. Um in Abbildung 5.2 die Tangente an f zu bekommen müsste der Parameter $a = f'(0.5) = 0.5$ sein.

Mit $\mathfrak{n} = \mathfrak{n}(a)$ kann man die Halbebene unterhalb von $g_{z,a}$ definieren als

$$\mathcal{H}_{\mathfrak{n}}^{z} := \left\{ \mathfrak{v} \in \mathbb{R}^2 \mid \mathfrak{n}^T \left(\mathfrak{v} - \begin{pmatrix} z \\ f(z) \end{pmatrix} \right) > 0 \right\} .$$

Nach dem vorigen Lemma ist $\mathcal{H}_2 \cap \mathcal{H}_{\mathfrak{n}}^{z}$ genau dann leer, wenn

$$\mathfrak{n} = \mathfrak{n}(f'(z))$$

ist, denn dann verläuft $g_{z,f'(z)}$ stets unterhalb von der Hausdorff-Menge $\mathcal{H}_2$ (siehe Abbildung 5.3).

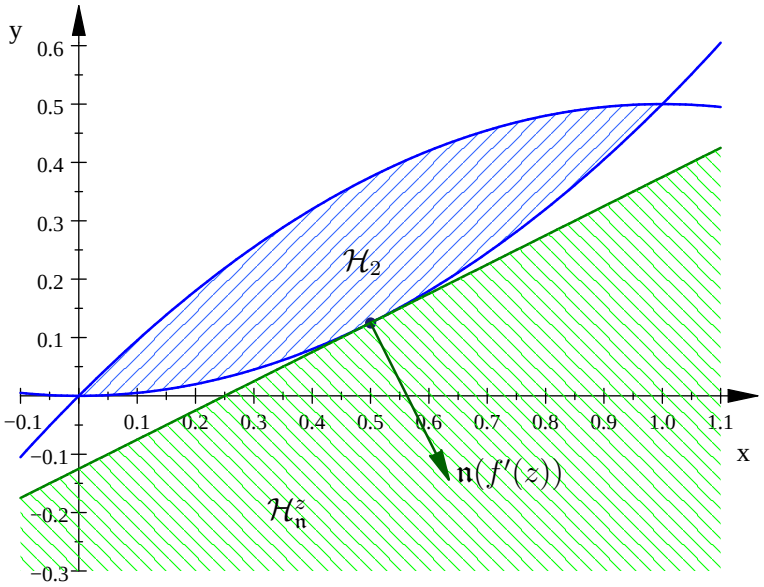

Abbildung 5.3.: Die Menge $\mathcal{H}_2$ und die Halbebene $\mathcal{H}_{\mathfrak{n}}^{z}$ für $\mathfrak{n} = \mathfrak{n}(f'(z))$

In einem Maximum $\mathfrak{m} = (z, f(z))^T$ des Optimierungsproblems (5.9) muss notwendig gelten, dass $\mathfrak{l} = \mathfrak{n}(f'(z))$ ist, d.h. $\mathfrak{l}$ ist ein äußerer Normalenvektor an den Graphen von f im Punkt $\mathfrak{m}$.

Angenommen in dem Maximum $\mathfrak{m}$ gilt $\mathfrak{l} \neq \mathfrak{n}(f'(z))$, dann muss man zwei Fälle unterscheiden.

1. Fall: $\mathfrak{l} = \begin{pmatrix} \pm 1 \\ 0 \end{pmatrix}$ (Diese beiden Punkte sind nicht im Bildbereich von $\mathfrak{n}(\cdot)$). Dieser Fall ist wegen der Voraussetzung $l_2 < 0$ ausgeschlossen.

2. Fall: Es gibt ein $b \neq f'(z)$ mit $\mathfrak{l} = \mathfrak{n}(b)$. Dann ist nach vorher $\mathcal{H}_2 \cap \mathcal{H}_{\mathfrak{n}(b)}^{z} \neq \emptyset$. Für einen beliebigen Vektor $\mathfrak{v} \in \mathcal{H}_2 \cap \mathcal{H}_{\mathfrak{n}(b)}^{z}$ gilt nach Definition

$$\mathfrak{l}^T \mathfrak{v} > \mathfrak{l}^T \mathfrak{m} .$$

Also kann $\mathfrak{m}$ kein Maximum von (5.9) sein. Widerspruch zur Voraussetzung.

Da f streng konvex ist, steigt f' streng monoton. Also ist die Abbildung

$$\mathfrak{f} : \mathbb{R} \to S_1 - \left\{ \begin{pmatrix} \pm 1 \\ 0 \end{pmatrix} \right\}, \; \mathfrak{f}(x) := \frac{1}{\sqrt{1+f'(x)^2}} \begin{pmatrix} f'(x) \\ -1 \end{pmatrix} = \frac{1}{\sqrt{1+x^2}} \begin{pmatrix} x \\ -1 \end{pmatrix},$$

die einem x den äußeren Einheitsnormalenvektor von f zuordnet bijektiv. Das bedeutet, dass die notwendige Optimalitätsbedingung $\mathfrak{l} = \mathfrak{n}(f'(z))$ für $z \in (0,1)$ auch hinreichend ist (in $(0, f(0))^T$ und $(1, f(1))^T$ hat $\mathcal{H}_2$ Ecken, die gesondert betrachtet werden müssen). Also kann man zu dem gegebenen äußeren Normalenvektor $\mathfrak{l}$ den Abszissenwert z des zugehörigen Stützpunktes berechnen, indem man den Ansatz macht

$$\mathfrak{l} = \mathfrak{f}(z).$$

Daraus resultiert das Gleichungssystem

$$\begin{aligned} l_1 &= \frac{1}{\sqrt{1+z^2}} z \\ l_2 &= \frac{-1}{\sqrt{1+z^2}}. \end{aligned}$$

Und man erhält unmittelbar als Lösung

$$z = -\frac{l_1}{l_2},$$

da $l_2 < 0$ ist (d.h. $l_2 \neq 0$). Hieran sieht man, dass das Ergebnis z unabhängig von der Länge des Vektors $\mathfrak{l}$ ist. Dieser muss also nicht unbedingt ein Einheitsvektor sein.

Jetzt muss noch überprüft werden, ob z zulässig ist.

Gilt $0 < z < 1$, so haben wir ein richtiges Ergebnis. Der Stützpunkt bzw. das Maximum von (5.9) ist dann

$$\mathfrak{x} = \begin{pmatrix} z \\ f(z) \end{pmatrix} = \left(-\frac{l_1}{l_2}, \frac{l_1^2}{2l_2^2} \right)^T.$$

Ist dagegen $z \geq 1$ oder $z \leq 0$, so ist zwar $\mathfrak{l}$ der äußere Normalenvektor an den Graphen von f im Punkt $(z, f(z))^T$, jedoch gehört dieser Teil des Graphen evtl. nicht mehr zu $\mathcal{H}_2$. In diesem Fall kommen als Stützpunkte nur die Ecken

$$(0,0)^T \text{ oder } \left(1, \frac{1}{2}\right)^T$$

in Frage. Die richtige Ecke ermittelt man am besten durch ausrechnen von $\mathfrak{l}^T \left(1, \frac{1}{2}\right)^T$. Dies Ergebnis kann man dann mit $\mathfrak{l}^T (0,0)^T = 0$ vergleichen.

Die hier vorgestellte Methode führt auf sehr ähnliche Formeln wie der vorige Ansatz. Man kann sie auch auf höhere Dimensionen ausdehnen, was schon in Lemma 5.3.4 angedeutet ist. Mit wachsendem p wird es auch mit dieser Methode immer schwieriger Stützpunkte an $\mathcal{H}_p$ zu berechnen. Außerdem muss man wie auch vorher Nullstellen von Polynomen vom Grad $p-1$ berechnen.

5.4. Rückführung auf das Aumann-Integral

In diesem Abschnitt wollen wir die Stützpunktmenge $Y(\mathfrak{l}, \mathcal{M}_{m,p})$ für eine beliebige Richtung $\mathfrak{l} \in \mathbb{R}^{mp}, \mathfrak{l} \neq 0$, mittels der Darstellung von $\mathcal{M}_{m,p}$ als Aumann-Integral berechnen.

Im ersten Unterabschnitt entwickeln wir damit ein Verfahren für Quadersteuermengen. Ein wesentlicher und sehr aufwendiger Teil davon ist die Berechnung von Polynomnullstellen.

Im nächsten Unterabschnitt soll deswegen das Problem der Nullstellenbestimmung von Polynomen diskutiert werden. Es werden analytische Formeln angegeben und ein zweiteiliges numerisches Verfahren vorgestellt.

Im letzten Unterabschnitt soll mittels dieser Darstellung von $\mathcal{M}_{m,p}$ ein Verfahren entwickelt werden um Stützpunkte dieser Menge zu berechnen, wenn der Steuerbereich eine Kugel ist. Hier spielt dann auch die mengenwertige numerische Integration hinein.

Mit den Bezeichnungen aus Abschnitt 5.2 haben wir

$$\mathcal{M}_{m,p} = \int_0^1 \mathfrak{Q}_p(t) \cdot U \, dt \,, \tag{5.10}$$

wobei U kompakt, konvex und nichtleer ist. Diese Darstellung ist erheblich flexibler als diejenige durch das Hausdorff-Momenten-Problem, da dort U ein Quader sein muss, dessen Kanten parallel sind zu der kanonischen Basis. Mit diesem Ansatz kann man nicht nur Steuermengen behandeln, die achsenparallele Quader oder Kugeln sind, sonder auch solche, die Ellipsen oder Polytope sind. Man kann auch darüber hinaus beliebe Steuermengen behandeln, für deren Stützpunktberechnung eine Formel existiert.

Mit dieser Darstellung von $\mathcal{M}_{m,p}$ gilt

$$Y(\mathfrak{l}, \mathcal{M}_{m,p}) = Y\left(\mathfrak{l}, \int_0^1 \mathfrak{Q}_p(t) \cdot U \, dt\right) .$$

Darauf wenden wir Proposition 1.5.4 an, denn die Voraussetzungen dafür sind erfüllt, wie schon in Abschnitt 5.2 erwähnt wurde. Wir erhalten dann

$$Y\left(\mathfrak{l}, \int_0^1 \mathfrak{Q}_p(t) \cdot U \, dt\right) = \int_0^1 Y\left(\mathfrak{l}, \mathfrak{Q}_p(t) \cdot U\right) \, dt = \int_0^1 \mathfrak{Q}_p(t) \cdot Y\left(\mathfrak{Q}_p(t)^T \mathfrak{l}, U\right) \, dt \,,$$

wobei wir auch die Rechenregeln für Stützpunktmengen genutzt haben. Insgesamt haben wir also

$$Y\left(\mathfrak{l}, \mathcal{M}_{m,p}\right) = \int_0^1 \mathfrak{Q}_p(t) \cdot Y\left(\mathfrak{Q}_p(t)^T \mathfrak{l}, U\right)\, dt\,. \tag{5.11}$$

5.4.1. Stützpunkte für einen Quadersteuerbereich

Im Folgenden werden wir konkret herausarbeiten, wie die Stützpunkte $y\left(\mathfrak{l}, \mathcal{M}_{m,p}\right)$, d.h. die Elemente von $Y\left(\mathfrak{l}, \mathcal{M}_{m,p}\right)$, berechnet werden können für einen achsenparallelen Quader als Steuerbereich.

Sei also $U = [a_1, b_1] \times \cdots \times [a_m, b_m] \subset \mathbb{R}^m$. Weiter sei $\mathfrak{e}_i = (e_{i,1}, \ldots, e_{i,mp})^T \in \mathbb{R}^{mp}$ derjenige Einheitsvektor mit $e_{i,j} = \delta_{i,j}$, wobei hier $\delta_{i,j}$ das Kronecker-Symbol bezeichnet. Außerdem bezeichnen wir mit $\mathfrak{q}_i(t) \quad (i = 1, \ldots, m)$ die i-te Spalte der Matrixfunktion $\mathfrak{Q}_p(t)$. Diese Abbildung wurde definiert als $\mathfrak{Q}_p(t) = \left(1 \cdot \mathfrak{E}_m, t \cdot \mathfrak{E}_m, \ldots, t^{p-1} \cdot \mathfrak{E}_m\right)^T$ deswegen besitzen ihre Spaltenvektoren die Darstellung $\mathfrak{q}_i(t) = \sum_{j=0}^{p-1} t^j \cdot \mathfrak{e}_{i+jm}$.

Sei nun $\mathfrak{r} = (r_1, \ldots, r_m)^T \in \mathbb{R}^m$ ein Vektor mit $\mathfrak{r} \neq 0_{\mathbb{R}^m}$. Mit Hilfe von Beispiel 1.4.5(v) gilt für die Stützpunktmenge $Y\left(\mathfrak{r}, U\right) = Y\left(r_1, [a_1, b_1]\right) \times \cdots \times Y\left(r_m, [a_m, b_m]\right)$.

Für $\mathfrak{l} = (l_1, \ldots, l_{mp})^T \in \mathbb{R}^{mp}$ ist die i-te Komponente $(1 \leq i \leq m)$ von $\mathfrak{Q}_p(t)^T \cdot \mathfrak{l}$ genau das Polynom $p_i(t) := \mathfrak{q}_i(t)^T \cdot \mathfrak{l} = \sum_{j=0}^{p-1} t^j \cdot l_{i+jm}$. Also ist $Y\left(\mathfrak{Q}_p(t)^T \mathfrak{l}, U\right) = Y\left(p_1(t), [a_1, b_1]\right) \times \cdots \times Y\left(p_m(t), [a_m, b_m]\right)$. Deswegen haben wir insgesamt

$$\int_0^1 \mathfrak{Q}_p(t) \cdot Y\left(\mathfrak{Q}_p(t)^T \mathfrak{l}, U\right)\, dt = \int_0^1 \sum_{i=1}^m \mathfrak{q}_i(t) \cdot Y\left(p_i(t), [a_i, b_i]\right)\, dt\,.$$

Nach [16, Lemma 3.123] sind $Y\left(r_i, [a_i, b_i]\right)$ und $Y\left(\mathfrak{r}, U\right)$ abgeschlossen. Da $\mathfrak{q}_i(t)\,[a_i, b_i]$ nach Proposition 1.5.2 messbar und integrierbar beschränkt ist, hat nach Proposition 1.5.4 auch $t \mapsto Y\left(\mathfrak{l}, \mathfrak{q}_i(t)\,[a_i, b_i]\right) = \mathfrak{q}_i(t) \cdot Y\left(p_i(t), [a_i, b_i]\right)$ diese Eigenschaften. Damit folgt mit [16, Corollary 4.47] insgesamt

$$Y\left(\mathfrak{l}, \mathcal{M}_{m,p}\right) = \int_0^1 \sum_{i=1}^m \mathfrak{q}_i(t) \cdot Y\left(p_i(t), [a_i, b_i]\right)\, dt = \sum_{i=1}^m \int_0^1 \mathfrak{q}_i(t) \cdot Y\left(p_i(t), [a_i, b_i]\right)\, dt\,.$$

Im Folgenden sei $1 \leq i \leq m$ fest gewählt. Es soll nun das Aumann-Integral $\int_0^1 \mathfrak{q}_i(t) \cdot Y\left(p_i(t), [a_i, b_i]\right)\, dt$ berechnet werden. Wir wenden Beispiel 1.4.5(vi) an und erhalten

$$Y\left(p_i(t), [a_i, b_i]\right) = \begin{cases} \{b_i\} & \text{für } p_i(t) > 0 \\ \{a_i\} & \text{für } p_i(t) < 0 \end{cases} \quad \text{und} \quad Y\left(p_i(t), [a_i, b_i]\right) = [a_i, b_i] \text{ falls } p_i(t) = 0 \tag{5.12}$$

ist. Falls das Polynom p_i nicht die Nullfunktion ist, ist der Stützpunkt $y\left(p_i(t), [a_i, b_i]\right)$ fast überall eindeutig. Diese beiden Fälle müssen wir unterscheiden. Ein Polynom ist genau dann das Nullpolynom, wenn alle Koeffizienten verschwinden.

1. Fall: p_i ist nicht das Nullpolynom, d.h. $\exists\, j \in \{0,\dots,p-1\}$ mit $l_{i+jm} \neq 0$.
Nach Definition von p_i ist dies also nicht die Nullfunktion. Dann ist die Menge $N_i := \{t \in [0,1] \mid p_i(t) = 0\}$ endlich und damit ein Nullmenge. Also ist der Stützwert $y\left(p_i(t), [a_i, b_i]\right)$ auf $[0,1] - N_i$ eindeutig.
Jetzt definieren wir die Funktion $u : [0,1] \to \mathbb{R}$, $u(t) = \begin{cases} b_i & \text{für } p_i(t) > 0 \\ a_i & \text{für } p_i(t) \leq 0 \end{cases}$. Diese Funktion ist stückweise konstant und damit messbar (siehe [25, p. 97]). Also ist u eine messbare Auswahl (siehe Definition 1.5.3) der mengenwertigen Abbildung $t \mapsto Y\left(p_i(t), [a_i, b_i]\right)$. Jede andere messbare Auswahl muss wegen (5.12) auf $[0,1] - N_i$ mit u übereinstimmen. Es ist dann $\mathfrak{q}_i \cdot u$ (punktweise definiert) ebenfalls eine messbare Auswahl von $t \mapsto \mathfrak{q}_i(t) \cdot Y\left(p_i(t), [a_i, b_i]\right)$, die auf $[0,1] - N_i$ eindeutig ist. Also besteht das Aumann-Integral $\int_0^1 \mathfrak{q}_i(t) \cdot Y\left(p_i(t), [a_i, b_i]\right)\, dt$ lediglich aus dem Element $\int_0^1 \mathfrak{q}_i(t) \cdot u(t)\, dt$. Nach Definition von $\mathfrak{q}_i$ kann dieses vektorwertige Integral geschrieben werden als $\sum_{j=0}^{p-1} \mathfrak{e}_{i+jm} \cdot \int_0^1 t^j \cdot u(t)\, dt$.
Es seien nun $t_1 < \ldots < t_r$ die Nullstellen mit ungerader Vielfachheit von p_i in $(0,1)$. Es ist also $r \leq p-1$. Weiter seien $t_0 := 0$ und $t_{r+1} := 1$. Dann kann man aufgrund der Definition von u schreiben

$$\int_0^1 t^j u(t)\, dt = \sum_{k=0}^{r} \int_{t_k}^{t_{k+1}} t^j c_k\, dt = \sum_{k=0}^{r} \frac{c_k}{j+1} \left(t_{k+1}^{j+1} - t_k^{j+1}\right) \qquad \text{für } 0 \leq j \leq p-1\,, \tag{5.13}$$

mit Koeffizienten $c_k = \begin{cases} b_i & \text{für } p_i(t) > 0 \text{ in } (t_k, t_{k+1}) \\ a_i & \text{für } p_i(t) < 0 \text{ in } (t_k, t_{k+1}) \end{cases}$. Jede Nullstelle a mit ungerader Vielfachheit s eines Polynoms muss zu einer Vorzeichenänderung des Polynoms führen, da der Faktor $(x-a)^s$ sein Vorzeichen ändert (Analog führt jede Nullstelle gerader Vielfachheit zu keiner Vorzeichenänderung). Deswegen kann man die c_k auch so charakterisieren: Falls $p_i(t) > 0$ für $t \in (0, t_1)$ setzen wir $c_k = b_i$ wenn $k \equiv 0\,(mod\ 2)$ und sonst $c_k = a_i$ $(k = 0,\dots,r)$. Falls aber $p_i(t) < 0$ für $t \in (0, t_1)$ machen wir es genau so nur vertauschen wir a_i und b_i.
Insgesamt kann man dann das einzige Element von $\int_0^1 \mathfrak{q}_i(t) \cdot Y\left(p_i(t), [a_i, b_i]\right)\, dt \subset \mathbb{R}^{mp}$ so berechnen

$$\sum_{j=0}^{p-1} \mathfrak{e}_{i+jm} \sum_{k=0}^{r} \frac{c_k}{j+1} \left(t_{k+1}^{j+1} - t_k^{j+1}\right) = \sum_{k=0}^{r} c_k \sum_{j=0}^{p-1} \mathfrak{e}_{i+jm} \cdot \frac{1}{j+1} \left(t_{k+1}^{j+1} - t_k^{j+1}\right)\,. \tag{5.14}$$

2. Fall: p_i ist das Nullpolynom, d.h. $l_{i+jm} = 0 \quad \forall\, j \in \{0,\dots,p-1\}$.
Dann ist $Y\left(p_i(t), [a_i, b_i]\right) = [a_i, b_i]$. In diesem Fall begnügen wir uns damit ein Element des Aumann-Integrals $\int_0^1 \mathfrak{q}_i(t) \cdot Y\left(p_i(t), [a_i, b_i]\right)$ zu berechnen. Die Funktion $t \mapsto \mathfrak{q}_i(t) \cdot \frac{a_i + b_i}{2}$ ist offensichtlich eine messbare Auswahl von dem mengenwertigen Integranden $\mathfrak{q}_i(t) \cdot Y\left(p_i(t), [a_i, b_i]\right)$. Wie vorher nutzen wir die Darstellung von $\mathfrak{q}_i$ als Summen von Einheitsvektoren und berechnet das Integral

$\int_0^1 \mathfrak{q}_i(t) \cdot \frac{a_i+b_i}{2}\, dt$ als

$$\int_0^1 \mathfrak{q}_i(t) \cdot \frac{a_i+b_i}{2}\, dt = \frac{a_i+b_i}{2} \sum_{j=0}^{p-1} \mathfrak{e}_{i+jm} \cdot \frac{1}{j+1}\,.$$

Auf diese Weise hat man ganz leicht ein Element von $\int_0^1 \mathfrak{q}_i(t) \cdot Y\left(p_i(t), [a_i, b_i]\right)$ berechnet. An dieser Stelle ist die Fallunterscheidung beendet.

Bemerkung 5.4.1. Wir betrachten noch einmal den 1. Fall. Es seien $t_{a-1} < t_a < t_{a+1}$ aufeinanderfolgende Nullstellen von p_i in $[0,1]$ und t_a habe eine gerade Vielfachheit. Dann ändert sich das Vorzeichen von p_i in t_a nicht, d.h. p_i hat in (t_{a-1}, t_a) und (t_a, t_{a+1}) das gleiche Vorzeichen. Also hat nach Definition auch u in $(t_{a-1}, t_{a+1}) - \{t_a\}$ den gleichen Wert. Da aber $\{t_a\}$ eine Nullmenge ist, müssen wir in (5.13) Nullstellen mit gerader Vielfachheit nicht berücksichtigen und damit auch nicht in (5.14) .

Es soll hier noch herausgearbeitet werden wie man im 1.Fall algebraisch entscheiden kann, ob $p_i(t)$ in $(0, t_1)$ größer, kleiner oder gleich 0 ist. Dazu benötigen wir folgende

Definition 5.4.2 (und Satz)**.** Seien $\mathfrak{x} = (x_1, \ldots, x_k)^T, \mathfrak{y} = (y_1, \ldots, y_k)^T \in \mathbb{R}^k$. Dann ist die so genannte lexikographische Ordnung auf dem $\mathbb{R}^k$ gegeben durch:

$$\mathfrak{x} \preceq \mathfrak{y}$$

genau dann, wenn

$$\mathfrak{x} = \mathfrak{y}$$

oder es gibt ein $i_0 \in \{1, \ldots, k\}$ mit

$$x_i = y_i \;\; (i = 1, \ldots, i_0 - 1) \qquad \text{und} \qquad x_{i_0} < y_{i_0}\,. \tag{5.15}$$

Falls $\mathfrak{x} \preceq \mathfrak{y}$ ist, aber $\mathfrak{x} \neq \mathfrak{y}$ schreibt wir kurz $\mathfrak{x} \prec \mathfrak{y}$.
Durch diese Ordnung wird der $\mathbb{R}^k$ zu einem total geordneten Vektorraum.

Beweis. Es ist um der Vollständigkeit willen zu verifizieren, dass diese Relation tatsächlich eine totale Ordnung ist. Seien dazu $\mathfrak{x}, \mathfrak{y}, \mathfrak{z} \in \mathbb{R}^k$. Dazu muss man zeigen, dass sie reflexiv, transitiv und antisymmetrisch ist und dass zwei beliebige Vektoren vergleichbar sind. Die zuletzt genannte Eigenschaft gilt offenbar. Die Gültigkeit der Reflexivität ist ebenfalls offensichtlich.

Antisymmetrie: Es gelte $\mathfrak{x} \preceq \mathfrak{y}$ und $\mathfrak{y} \preceq \mathfrak{x}$. Zu zeigen ist, dass dann $\mathfrak{x} = \mathfrak{y}$ sein muss. Gäbe es ein $i_0 \in \{1, \ldots, k\}$ mit (5.15) $x_i = y_i \;\; (i = 1, \ldots, i_0 - 1)$ und $x_{i_0} < y_{i_0}$ so könnte unmöglich $\mathfrak{y} \preceq \mathfrak{x}$ gelten. Damit muss $\mathfrak{x} = \mathfrak{y}$ sein.

Reflexivität: Es gelte $\mathfrak{x} \preceq \mathfrak{y}$ und $\mathfrak{y} \preceq \mathfrak{z}$. Es ist zu zeigen, dass $\mathfrak{x} \preceq \mathfrak{z}$ gilt. Wir können o.E. annehmen, dass $\mathfrak{x} \neq \mathfrak{y}$ und $\mathfrak{y} \neq \mathfrak{z}$ ist, sonst ist die Aussage trivial. Seien $i_0, i_1 \in \{1, \ldots, k\}$ sodass (5.15) für $\mathfrak{x}, \mathfrak{y}$ bzw. $\mathfrak{y}, \mathfrak{z}$ für i_0 bzw. i_1 gilt. Falls $i_0 \leq i_1$ ist, gilt $x_i = z_i \; (i = 1, \ldots, i_0 - 1)$ und $x_{i_0} < y_{i_0} \leq z_{i_0}$. Und im Gegenfall gilt $x_i = z_i \; (i = 1, \ldots, i_1 - 1)$ und $x_{i_1} \leq y_{i_1} < z_{i_1}$. Also ist in beiden Fällen $\mathfrak{x} \preceq \mathfrak{z}$. □

Mit dieser Definition können wir dann formulieren:

Lemma 5.4.3. *Sei* $\mathfrak{a} = (a_1, \ldots, a_k)^T \in \mathbb{R}^k$ *und* $p_\mathfrak{a} : \mathbb{R} \to \mathbb{R}$ *ein Polynom definiert durch* $p_\mathfrak{a}(x) := \sum_{i=1}^{k} a_i x^{i-1}$. *Weiter sei* x_1 *die kleinste Nullstelle von* p *in* $(0, 1]$ *oder falls diese nicht existiert sei* $x_1 = 1$. *Dann gilt für alle* $x \in (0, x_1)$*:*

$$p_\mathfrak{a}(x) \left\{ \begin{array}{c} > \\ = \\ < \end{array} \right\} 0 \quad \Longleftrightarrow \quad \mathfrak{a} \left\{ \begin{array}{c} \succ \\ = \\ \prec \end{array} \right\} 0_{\mathbb{R}^k},$$

d.h. es gilt jeweils die Ungleichung bzw. Gleichung in der gleichen Komponente.

Beweis. Wir zeigen $p_\mathfrak{a}(x) > 0 \iff \mathfrak{a} \succ 0_{\mathbb{R}^n}$.
"$\Longrightarrow$". Sei $i_0 \in \{1, \ldots, n\}$ der kleinste Index mit $a_i = 0$ $(i < i_0)$ und $a_{i_0} \neq 0$. Wir zeigen $a_{i_0} > 0$. Dann definieren wir $q(x) := \frac{1}{x^{i_0-1}} \cdot p_\mathfrak{a}(x) = \sum_{i=i_0}^{n} a_i x^{i-i_0}$. Dann ist auch $q(x) > 0$ in $(0, x_1)$. Weil $\sum_{i=i_0+1}^{n} a_i x^{i-i_0} \longrightarrow 0$ für $x \to 0$, gibt es ein ϵ mit $x_1 > \epsilon > 0$, sodass $\sum_{i=i_0+1}^{n} a_i x^{i-i_0} < \frac{|a_{i_0}|}{2}$ für $0 \leq x < \epsilon$. Aber auch für diese x gilt $0 < q(x) \leq a_{i_0} + \left|\sum_{i=i_0+1}^{n} a_i x^{i-i_0}\right| \leq a_{i_0} + \frac{|a_{i_0}|}{2}$. Also muss $a_{i_0} > 0$ sein. Nach Definition ist dann $\mathfrak{a} \succ 0_{\mathbb{R}^n}$.
"$\Longleftarrow$". Sei $i_0 \in \{1, \ldots, n\}$ der kleinste Index mit $a_i = 0$ $(i < i_0)$ und $a_{i_0} \neq 0$. Wir zeigen $p_\mathfrak{a}(x) > 0$ in $(0, x_1)$. Wir nutzten die Bezeichungen von vorigen Richtung. Dann gilt für $x \in [0, \epsilon)$: $\sum_{i=i_0+1}^{n} a_i x^{i-i_0} > \frac{-a_{i_0}}{2}$. Also ist $q(x) > \frac{a_{i_0}}{2} > 0$ für $x \in [0, \epsilon)$. Also ist auch $p_\mathfrak{a}(x) = q(x) \cdot x^{i_0-1} > 0$ für $x \in [0, \epsilon)$. Und da $x_1 > \epsilon$ die erste Nullstelle in $(0, 1]$ ist, ist $p_\mathfrak{a}(x)$ positiv auf ganz $(0, x_1)$.
Jetzt zeigen wir $p_\mathfrak{a}(x) = 0 \iff \mathfrak{a} = 0_{\mathbb{R}^n}$.
"$\Longrightarrow$". Sei $p_\mathfrak{a}(x) = 0$. Angenommen $\mathfrak{a} \neq 0_{\mathbb{R}^n}$. Dann gibt es einen größten Index $i_0 \in \{1, \ldots, n\}$ mit $a_i = 0$ $(i > i_0)$ und $a_{i_0} \neq 0$. Dann ist das Polynom, das $p_\mathfrak{a}$ zugrunde liegt, nicht das Nullpolynom und es gilt $\lim_{x\to\infty} |p_\mathfrak{a}(x)| = \infty$. Also kann $p_\mathfrak{a}$ nicht die Nullfunktion sein. Widerspruch.
Die andere Richtung ist trivial.
Es gelten die Äquivalenzen $p_\mathfrak{a}(x) < 0 \;\Leftrightarrow\; -p_\mathfrak{a}(x) = p_{-\mathfrak{a}}(x) > 0$ und $\mathfrak{a} \prec 0_{\mathbb{R}^n} \;\Leftrightarrow\; -\mathfrak{a} \succ 0_{\mathbb{R}^n}$. Also gilt mit dem obigen Beweisteil auch die Äquivalenz $p_\mathfrak{a}(x) < 0 \quad \forall x \in (0, x_1) \iff \mathfrak{a} \prec 0_{\mathbb{R}^n}$. □

Mit diesem Lemma und dem zuvor Erarbeiteten können wir nun einen Algorithmus formulieren um für diesen speziellen Steuerbereich einen Stützpunkt an die Menge $\mathcal{M}_{m,p}$ zu berechnen.

Algorithmus 5.4.4. *Sei* $U = [a_1, b_1] \times \cdots \times [a_m, b_m] \subset \mathbb{R}^m$ *mit der dazugehörigen Menge* $\mathcal{M}_{m,p}$ *gegeben. Weiter sei* $\mathfrak{l} = (l_1, \ldots, l_{mp})^T \in \mathbb{R}^{mp}$, $\mathfrak{l} \neq 0_{\mathbb{R}^{mp}}$, *gegeben. Außerdem seien die reellwertigen Polynome* $p_i(t) := \sum_{j=0}^{p-1} t^j \cdot l_{i+jm}$ $(i = 1, \ldots, m)$ *zusammen mit den Zahlen*

$0 = t_{i,0} < t_{i,1} < \ldots < t_{i,r+1} = 1$ *gegeben, wobei dies für die Indizes* $i = 1, \ldots, m$ *genau die Nullstellen mit ungerader Vielfachheit von* p_i *in* $(0,1)$ *sein sollen. Falls* p_i *das Nullpolynom ist, soll* $r := 0$ *sein. Mit* $\mathfrak{l}^i := \left(l_{i+0}, l_{i+m}, \ldots, l_{i+(p-1)m}\right)^T$ *bezeichnen wir den Koeffizientenvektor von* p_i*. Dann kann man einen Stützpunkt an* $\mathcal{M}_{m,p}$ *in Richtung* $\mathfrak{l}$ *so berechnen*

$$y\left(\mathfrak{l}, \mathcal{M}_{m,p}\right) = \sum_{i=1}^{m} \sum_{k=0}^{r} c_{i,k} \sum_{j=0}^{p-1} \mathfrak{e}_{i+jm} \cdot \frac{1}{j+1} \left(t_{i,k+1}^{j+1} - t_{i,k}^{j+1}\right) , \tag{5.16}$$

wobei die reellen Zahlen $c_{i,k}$ *folgende Bedeutung haben:*
Für jedes $i = 1, \ldots, m$ *bestimmen sich die* $c_{i,k}$ $(k = 0, \ldots, r)$ *nach folgender Fallunterscheidung*

$\mathfrak{l}^i \succ 0_{\mathbb{R}^p}$ *dann ist* $c_{i,k} = b_i$ *falls* k *gerade ist und* $c_{i,k} = a_i$ *falls* k *ungerade ist,*

$\mathfrak{l}^i = 0_{\mathbb{R}^p}$ *dann ist* $c_{i,0} = \frac{a_i + b_i}{2}$ *(und* $p_i = 0$*),*

$\mathfrak{l}^i \prec 0_{\mathbb{R}^p}$ *dann ist* $c_{i,k} = a_i$ *falls* k *gerade ist und* $c_{i,l} = b_i$ *falls* k *ungerade ist.*

Falls $\mathfrak{l}^i \neq 0_{\mathbb{R}^p}$ *für alle* $i = 1, \ldots, m$ *so ist der berechnete Stützpunkt eindeutig, d.h. die Stützpunktmenge* $Y\left(\mathfrak{l}, \mathcal{M}_{m,p}\right)$ *enthält nur diesen Stützpunkt.*

Bemerkung 5.4.5.

(i) Diese Berechnung von $y\left(\mathfrak{l}, \mathcal{M}_{m,p}\right)$ sieht zunächst komplexer aus, als sie tatsächlich ist. Falls der Vektor $\mathfrak{l}$ in irgendeinem Sinn gleichverteilt über die Sphäre S_{n-1} gewählt wird, haben die dadurch definierten Polynome p_i nur selten Nullstellen in $(0,1)$. Und falls doch, dann haben sie meist nur eine. Die Softwaretests bestätigen dies. Für den Fall das p_i keine Nullstellen in $(0,1)$ hat kann man den entsprechenden Teil in der Summe vorberechnen.
Außerdem können die Potenzen $t_{i,k}^{j+1}$ schrittweise aufmultipliziert werden und müssen nicht direkt berechnet werden.

(ii) Zusammen mit den Algorithmen 5.1.2 und 5.1.3 hat man damit für Quadersteuermengen einen Algorithmus um diese mengenwertigen Ferretti-Verfahren zu implementieren.

(iii) Die versteckte Hauptarbeit in der Berechnung von $y\left(\mathfrak{l}, \mathcal{M}_{m,p}\right)$ steckt in der Berechnung der Nullstellen der Polynome p_i. Dabei haben diese Polynome den Grad $p-1$. Je größer also p ist, welches auch die Konvergenzordnung der mengenwertigen Verfahren ist, desto schwieriger bzw. aufwendiger wird es die Nullstellen zu berechnen.

Deswegen soll im Folgenden auf die Nullstellenberechnung eingegangen werden.

5.4.2. Polynomnullstellen

In diesem Unterabschnitt soll darauf eingegangen werden, wie wir für Algorithmus 5.4.4 die Nullstellen eines Polynoms im Intervall $(0, 1)$ effektiv berechnen können.

Für Polynome vom Grad ≤ 4 existieren Formeln für die direkte Berechnung aller Nullstellen. Daher ist es in diesem Fall am besten alle Nullstellen zu berechnen und dann diejenigen die in $(0, 1)$ liegen auszusondern.

Für Polynome vom Grad ≥ 5 existieren solche Formeln nicht mehr. Deswegen müssen die Nullstellen numerisch approximiert werden.

In der Software zu dieser Arbeit habe ich mengenwertige Ferretti-Verfahren der Ordnung $2, 3, 4$ und 8 realisiert. Für ein Verfahren der Ordnung p müssen Nullstellen eines Polynoms vom Grad $p-1$ berechnet werden. Deswegen wollen wir hier die Nullstellenformeln für Polynome vom Grad 2 und 3 betrachten. Für lineare Polynome ist die Nullstellenberechnung trivial und für Polynome vom Grad 7 muss sie numerisch erfolgen.

Sei also ein Polynom f_k gegeben, definiert durch $f_k(x) := a_k x^k + a_{k-1}x^{k-1} + \ldots + a_0$, wobei alle Koeffizienten reell sein sollen. Dann berechnen sich die Nullstellen x_i für den quadratischen Fall mittels der Formel

$$x_1 = \frac{-a_1 - \sqrt{a_1^2 - 4a_2a_0}}{2a_2} \quad , \quad x_2 = \frac{-a_1 + \sqrt{a_1^2 - 4a_2a_0}}{2a_2} \, .$$

Mittels der Formel von Cardano, welche in [8, Abschnitt 6.2, Satz 1] dargestellt wird, kann man die Nullstellen des kubischen Polynoms f_3 berechnen. Dazu wird das Polynom zunächst normiert, indem alle Koeffizienten durch den Leitkoeffizient a_3 geteilt werden. Anschließend führen wir die Substitution $x = y - \frac{a_2}{3 \cdot a_3}$ durch und erhalten das Polynom

$$\hat{f}_3(y) := y^3 + 3ry + 2s \, ,$$

wobei $r := \frac{a_1}{3a_3} - \frac{a_2^2}{9a_3^2}$ und $s := \frac{a_0}{2a_3} + \frac{a_2^3}{27a_3^3} - \frac{a_2a_1}{6a_3^2}$ ist. Das Lösungsverhalten von $\hat{f}_3$ wird durch die Diskriminante $D = s^2 + r^3$ bestimmt.

1. Fall $D > 0$ **:** Es gibt nur eine reelle Nullstelle

$$y_1 = \sqrt[3]{-s + \sqrt{D}} + \sqrt[3]{-s - \sqrt{D}} \, .$$

2. Fall $D = 0$ **:** Es gibt zwei reelle Nullstellen, von denen eine zweifach ist

$$y_1 = 2\sqrt[3]{-s} \quad , \quad y_{2,3} = -\sqrt[3]{-s} \, .$$

Falls $s = 0$ ist haben wir sogar die dreifache Nullstelle $y_{1,2,3} = 0$.

3. Fall $D < 0$ **:** Es gibt drei reelle Nullstellen. Dieser wichtige Fall, wird in der Literatur als "casus irreducibilis" bezeichnet. Er hat in der Geschichte der Mathematik wesentlich zur Einführung der komplexen Zahlen beigetragen. Hier jedoch soll dieser Fall mit Hilfe der trigonometrischen Funktionen dargestellt werden (siehe [30, Unterabschnitt 2.1.6.2]), da dies günstiger für die Berechnung im Rechner ist. Dazu definieren wir $R := \text{sign}\,(s)\,\sqrt{-r}$ und $\beta := \frac{1}{3}\arccos\frac{s}{R^3}$ und erhalten damit

$$y_1 = -2R\cos\beta \quad , \quad y_{2,3} = 2R\cos\left(\beta \pm \frac{\pi}{3}\right) .$$

Um die Nullstellen von dem Ausgangspolynom f_3 zu erhalten muss man nun noch die Substitution rückgängig machen.

Im Folgenden wollen wir uns mit der numerischen Nullstellenbestimmung befassen, welche in zwei Phasen erfolgt. Dafür sei q ein Polynom mit reellen Koeffizienten von beliebigem Grad.

In der ersten Phase wird mit der Sturm'schen Regel bestimmt ob q in $(0,1)$ Nullstellen hat. Wenn ja werden diese mit der Sturm'schen Regel und Intervallhalbierung sukzessive von einander getrennt, d.h. es werden Teilintervalle gefunden, von denen jedes nur eine Nullstelle enthält.

In der zweiten Phase wird das Sekantenverfahren (Regula Falsi) auf diese Teilintervalle angewandt um die Nullstellen genau zu bestimmen.

Zunächst soll die Sturm'sche Regel erläutert werden. Wir definieren eine Folge von Polynomen $q_0 := q,\ q_1 := -q', \ldots,\ q_m$ welche gebildet werden durch Polynomdivision mit Rest:

$$\begin{aligned} q_{i-1} &= o_i q_i - q_{i+1}\ ,\ \text{Grad}\,(q_i) > \text{Grad}\,(q_{i+1}) \qquad (i = 1,\ldots,m-1)\ , \\ q_{m-1} &= o_m q_m \qquad\quad ,\ q_m \neq 0\ . \end{aligned} \tag{5.17}$$

Dies ist der euklidischen Algorithmus (bis auf ein $-$ beim Rest), deswegen ist q_m der ggT von q und q' (bis auf eine Einheit). Falls q eine Nullstelle $a \in \mathbb{R}$ mit Vielfachheit $k > 1$ hat, gilt $(x-a)^k \mid q$ und wegen der Produktregel und (5.17) auch $(x-a)^{k-1} \mid q_i\ (i = 1,\ldots,m)$. Also hat q_m die gleiche Nullstelle mit Vielfachheit $k-1$. Falls aber q nur einfache Nullstellen hat, muss q_m ein konstantes Polynom $\neq 0$ sein. Dann ist die obige Folge eine so genannte Sturm'sche Kette (vgl. [23, Definition 5.5.7 und Lemma 5.5.9]). Im Folgenden bezeichnen wir mit $w\,(c)\ (c \in \mathbb{R})$ die Anzahl der Vorzeichenwechsel in der Folge $q_0\,(c)\,, q_1\,(c)\,, \ldots, q_m\,(c)$ nach Streichung aller Nullen. Der folgende Satz zeigt, wie diese Polynomfolge zur groben Lokalisierung von Nullstellen von q eingesetzt werden kann.

Satz 5.4.6 (Regel von Sturm)**.** *Das Polynom q habe nur einfache Nullstellen. Dann ist die Anzahl der reellen Nullstellen von q im Intervall $[a,b)$ gleich $w(b) - w(a)$.*

Falls q auch mehrfache Nullstellen hat und $q(a)q(b) \neq 0$ ist, gilt darüber hinaus:

Die Anzahl der Nullstellen in (a,b) ist ebenfalls $w(b) - w(a)$ (ohne Vielfachheiten).

Beweis. Falls q nur einfache Nullstellen hat, findet sich ein Beweis in [23, Satz 5.5.8 und Lemma 5.5.9].

Jetzt habe q auch mehrfache Nullstellen. Wie im zitierten Beweis untersuchen wir

$$w(z-h), w(z), w(z+h)$$

für ein festes $z \in \mathbb{R}$ und ein hinreichend kleines $h > 0$. In einem offenen Intervall $J \subset (a,b)$ mit $q_i(s) \neq 0$ $(i = 0, \ldots, m\,;\, s \in J)$ gilt offenbar $w(s) = const.$ Deswegen sei nun wenigstens ein $q_i(z) = 0$.

1. Fall: $q_j(z) = 0$ für ein $j \in \{0, 1, \ldots, m\}$ und $q_m(z) \neq 0$.
Aus (5.17) für $i = j$ folgt sofort, dass $q_{j-1}(z)q_{j+1}(z) \leq 0$ ist. Wäre nun $q_{j-1}(z) = 0$ oder $q_{j+1}(z) = 0$, so erhielte man durch mehrfache Anwendung von (5.17) sofort

$$q_i(z) = 0 \quad , \quad i = 0, \ldots, m$$

was aber ein Widerspruch zu $q_m(z) \neq 0$ ist. Deshalb ist $q_{j-1}(z)q_{j+1}(z) < 0$. Es sind also die Argumente des Beweises von [23, Satz 5.5.8] für z lokal anwendbar und wir erhalten

$$w(z+h) = w(z) + 1 = w(z-h) + 1$$

für ein hinreichend kleines $h > 0$.

2. Fall: $q_j(z) = 0$ für ein $j \in \{0, 1, \ldots, m\}$ und $q_m(z) = 0$.
Wiederholte Anwendung von (5.17) liefert, dass $q_i(z) = 0$ $(i = 0, \ldots, m)$ ist.
Für zwei Polynome g, h soll im Folgenden $g^k \parallel h$ bedeuten, dass $g^k \mid h$ aber $g^{k+1} \nmid h$.
Wie oben schon erwähnt, gibt es ein $k > 1$ sodass $(x-z)^k \parallel q_0 = q$. Dann gibt es ein $\tilde{q}_0$ mit $q_0 = (x-z)^k \tilde{q}_0$ und $(x-z) \nmid \tilde{q}_0$. Nach Produktregel ist $q_1 = k\,(x-z)^{k-1}\,\tilde{q}_0 + (x-z)^k\,(\tilde{q}_0)'$. Wegen $(x-z) \nmid \tilde{q}_0$ gilt $(x-z)^{k-1} \parallel q_1$. Würde nun $(x-z)^k \mid q_2$ gelten so würde man mittels mehrmaliger Anwendung von (5.17) erhalten $(x-z)^k \mid q_m$ und q_m wäre nicht der ggT von p_0 und p_1. Widerspruch!
Also gilt wegen (5.17) für $i = 1$ nur $(x-z)^{k-1} \parallel q_2$. Mittels (5.17) erhält man dann sukzessive $(x-z)^{k-1} \parallel q_i \quad (i = 1, \ldots, m)$. Für zwei benachbarte Polynome q_i, q_{i+1} $(1 \leq i \leq m-1)$ gilt für hinreichend kleines h: Falls q_i in $(z-h, z+h)$ das Vorzeichen wechselt, so auch q_{i+1}. Und falls q_i in $(z-h, z+h)$ das Vorzeichen nicht wechselt, so auch q_{i+1} nicht. Also ist die Anzahl der Vorzeichenwechsel in den Folgen $q_i(z-h)$, $q_{i+1}(z-h)$ und $q_i(z+h)$, $q_{i+1}(z+h)$ gleich.
Aus dem Wissen, dass $q_1 = -q_0'$ ist, erhält man folgende Tabelle über das Vorzeichenverhalten von q_0, q_1:

	$z-h$	z	$z+h$	$z-h$	z	$z+h$	$z-h$	z	$z+h$	$z-h$	z	$z+h$
q_0	$+$	0	$+$	$-$	0	$-$	$+$	0	$-$	$-$	0	$+$
q_1	$+$	0	$-$	$-$	0	$+$	$+$	0	$+$	$-$	0	$-$

Also haben wir insgesamt $w(z+h) = w(z-h)+1$. Da $q_0(a)q_0(b) \neq 0$ ist, nach Voraussetzung, ist $z \in (a,b)$. Also kann $h > 0$ so klein gewählt werden, dass $z+h$ und $z-h$ in (a,b) sind.

Insgesamt ist also $w(b) - w(a)$ die Anzahl der reellen Nullstellen in (a,b) ohne Vielfachheiten gezählt. □

Wir können also diesen Satz anwenden um herauszufinden ob und gegebenfalls wieviele Nullstellen q in $(0,1)$ hat. Zusätzlich erhalten wir über q_m die Information ob mehrfach Nullstellen vorliegen. Anschließend isolieren wir durch gezielte Intervallhalbierung und Anwendung dieser Regel die Nullstellen. Mit dem folgenden Algorithmus, dem Sekantenverfahren (Regula Falsi) erfolgt nun die lokale Bestimmung der Nullstelle.

Algorithmus 5.4.7. *(Sekantenverfahren) Seien eine Funktion $f : [a,b] \to \mathbb{R}$ und Startwerte $x_0, x_1 \in [a,b]$ vorgegeben. Dann wird das Sekantenverfahren durch die Iterationsvorschrift*

$$x_{i+1} = \frac{x_i f(x_{i-1}) - x_{i-1} f(x_i)}{f(x_{i-1}) - f(x_i)},$$

für $i = 1, 2, \ldots$ definiert.

Speziell für Polynome formuliert, beschreibt der nächste Satz das Konvergenzverhalten des Sekantenverfahrens.

Satz 5.4.8. *Sei q ein beliebiges Polynom mit reellen Koeffizienten und $\hat{x}$ eine einfache Nullstelle von q. Dann gibt es eine Umgebung von $\hat{x}$ so, dass das Sekantenverfahren für beliebige Startwerte aus dieser Umgebung konvergiert. Die Konvergenzordnung beträgt $\frac{1}{2}\left(1+\sqrt{5}\right) \approx 1.618$.*

Beweis. Jedes Polynom als Funktion gesehen ist eine $\mathcal{C}^\infty(\mathbb{R})$-Funktion. Da $\hat{x}$ eine einfache Nullstelle von q ist kann es keine Nullstelle von q' sein (siehe [24, Kapitel 7, Satz 12]). Dann liefert [23, Satz 5.5.5] die Behauptung. □

Falls q auch mehrfache Nullstellen hat, was sehr selten vorkommt, kann das Sekantenverfahren nicht angewandt werden. Angenommen wir haben mit der Sturm'schen Regel ein Intervall $(c,d) \subset (0,1)$ ermittelt in der eine Nullstelle liegt, aber eventuell auch mit mehrfacher Vielfachheit. Falls dann $q(c)q(d) < 0$ ist, kann die Primitivform der Regula Falsi angewandt werden, welche man in [23, Abschnitt 5.5.2] beschrieben findet, um die Nullstelle zu finden. Falls aber $q(c)q(d) > 0$ ist, so liegt eine Nullstelle mit gerader Vielfachheit vor. An solchen Nullstellen sind wir nicht interessiert, denn in Algorithmus 5.4.4 werden sie nicht benötigt (siehe auch Bemerkung 5.4.1).

5.4.3. Stützpunkte für Kugelsteuerbereich

In diesem Unterabschnitt werden wir herausarbeiten, wie die Stützpunkte $y\left(\mathfrak{l}, \mathcal{M}_{m,p}\right)$, d.h. die Elemente von $Y\left(\mathfrak{l}, \mathcal{M}_{m,p}\right)$, berechnet werden können für eine Kugel als Steuerbereich, d.h. $U = B_r\left(\mathfrak{m}\right)$ mit Mittelpunkt $\mathfrak{m} \in \mathbb{R}^m$ und Radius $r > 0$.

Im Folgenden benutzten wir die Bezeichnungen $\mathfrak{Q}_p(t)$ und $p_i(t)$ aus Unterabschnitt 5.4. Weiter sei ein Richtungsvektor $\mathfrak{l} \in \mathbb{R}^{mp}$, $\mathfrak{l} \neq 0_{\mathbb{R}^{mp}}$, gegeben. Dazu definieren wir die vektorwertige Funktion $\mathfrak{p} : [0,1] \to \mathbb{R}^{mp}$, $\mathfrak{p}\left(t\right) := \mathfrak{Q}_p(t)^T \cdot \mathfrak{l} = \left(p_1(t), \ldots, p_m(t)\right)^T$. Da $\mathfrak{l} \neq 0_{\mathbb{R}^{mp}}$ vorausgesetzt ist, können nicht alle p_i das Nullpolynom sein, und damit ist $\mathfrak{p}\left(t\right)$ fast überall nicht der Nullvektor. Nach Beispiel 1.4.5 ist der Stützpunkt

$$y\left(\mathfrak{p}(t),\, B_r\left(\mathfrak{m}\right)\right) \;=\; \mathfrak{m} + r\frac{\mathfrak{p}(t)}{\left\|\mathfrak{p}(t)\right\|_2} \quad \text{falls } \mathfrak{p}(t) \neq 0_{\mathbb{R}^m}\,,$$

d.h. er ist fast überall eindeutig. Mit dem gleichen Argument wie in Unterabschnitt 5.4.1 in Fall 1 folgt dann, dass die Stützpunktmenge

$$Y\left(\mathfrak{l}, \mathcal{M}_{m,p}\right) = \int_0^1 \mathfrak{Q}_p(t) \cdot Y\left(\mathfrak{p}(t),\, B_r\left(\mathfrak{m}\right)\right)\, dt$$

aus (5.11) ebenfalls nur ein Element besitzt. Dieses Element berechnet sich als

$$y\left(\mathfrak{l}, \mathcal{M}_{m,p}\right) = \int_0^1 \mathfrak{Q}_p(t) \cdot \mathfrak{m}\, dt + r\int_0^1 \mathfrak{Q}_p(t)\frac{\mathfrak{p}(t)}{\left\|\mathfrak{p}(t)\right\|_2}\, dt\,, \tag{5.18}$$

wenn $\mathfrak{p}$ in $[0,1]$ nicht verschwindet.

Das erste Integral ist unabhängig von der Richtung $\mathfrak{l}$ und kann deshalb leicht im Voraus ausgewertet werden (es sei daran erinnert, dass wir den Stützpunkt in vielen Richtungen brauchen). Es hat in jeder Komponente einfache Polynome als Integranden und kann daher analytisch berechnet werden. Der Integrand des zweiten Integrals ist

$$\mathfrak{Q}_p(t)\frac{\mathfrak{p}(t)}{\left\|\mathfrak{p}(t)\right\|_2} = \begin{pmatrix} \mathfrak{p}(t) \\ t \cdot \mathfrak{p}(t) \\ \vdots \\ t^{p-1} \cdot \mathfrak{p}(t) \end{pmatrix} \frac{1}{\sqrt{p_1(t)^2 + \ldots + p_m(t)^2}}\,.$$

Dazu lässt sich im Allgemeinen keine Stammfunktion bestimmen. Damit lässt sich der Stützpunkt $y\left(\mathfrak{l}, \mathcal{M}_{m,p}\right)$ nicht mehr analytisch berechnen.

Zur numerischen Integration ist es von Vorteil wenn das Integrationsintervall möglichst klein ist. Damit wollen wir uns zuerst befassen.

5.4.3.1. Verkleinerung des Integrationsintervalls

Es soll hier angestrebt werden das Integrationsintervall von $[0,1]$ auf $[0,h]$ zu verkleinern, wobei $h > 0$ die Schrittweite ist, da dann weniger Iterationen für die numerische Integration gebraucht werden. Diese Intervallverkleinerung ist möglich, erfordert aber eine Neubetrachtung der Mengen $\mathcal{I}_{m,p}$ und $\mathcal{M}_{m,p}$.
Zunächst definieren wir die Menge

$$\hat{\mathcal{I}}_{m,p} := \Bigg\{ \left(\zeta_1^T, \ldots, \zeta_p^T\right)^T \in \mathbb{R}^{mp} \mid \zeta_1 = \int_0^h \mathfrak{u}(s_1)\, ds_1\, , \ldots , \\ \zeta_p = \int_0^h \cdots \int_0^{s_{p-1}} \mathfrak{u}(s_p)\, ds_p \ldots ds_1 \text{ für } \mathfrak{u} \in L^1([0,h]; U) \Bigg\} .$$

Weiter sei $\left(\xi_1^T, \ldots, \xi_p^T\right)^T \in \mathcal{I}_{m,p}$. Dann gibt es nach Definition von $\mathcal{I}_{m,p}$ eine Funktion $\mathfrak{u} \in L^1([0,1]; U)$ sodass gilt

$$\xi_i = \int_0^1 \ldots \int_0^{s_{i-1}} \mathfrak{u}(s_i)\, ds_i \ldots ds_1 \qquad (i = 1, \ldots, p) .$$

Mit mehrfacher Anwendung der Substitutionsregel erhält man dann

$$\xi_i = \frac{1}{h^i} \int_0^h \ldots \int_0^{s_{i-1}} \tilde{\mathfrak{u}}(s_i)\, ds_i \ldots ds_1 \qquad (i = 1, \ldots, p) ,$$

wobei $\tilde{\mathfrak{u}}(s) := \mathfrak{u}\left(s \cdot \frac{1}{h}\right)$ ist. Also ist $\tilde{\mathfrak{u}} \in L^1([0,h]; U)$.
Mit der Abkürzung $\nu_i := \int_0^h \ldots \int_0^{s_{i-1}} \tilde{\mathfrak{u}}(s_i)\, ds_i \ldots ds_1$ ist also $\left(\nu_1^T, \ldots, \nu_p^T\right)^T \in \hat{\mathcal{I}}_{m,p}$. Also gilt $\mathcal{I}_{m,p} \subset \mathfrak{H}_{m,p}^{-1} \cdot \hat{\mathcal{I}}_{m,p}$, wobei

$$\mathfrak{H}_{m,p} := \begin{pmatrix} h^1 \cdot \mathfrak{E}_m & & 0 \\ & \ddots & \\ 0 & & h^p \cdot \mathfrak{E}_m \end{pmatrix}$$

ist.
Sei umgekehrt $\left(\nu_1^T, \ldots, \nu_p^T\right)^T \in \hat{\mathcal{I}}_{m,p}$ gegeben. Dann gibt es ein $\tilde{\mathfrak{u}} \in L^1([0,h]; U)$ mit

$$\nu_i = \int_0^h \ldots \int_0^{s_{i-1}} \tilde{\mathfrak{u}}(s_i)\, ds_i \ldots ds_1 = h^i \int_0^1 \ldots \int_0^{s_{i-1}} \mathfrak{u}(s_i)\, ds_i \ldots ds_1 \qquad (i = 1, \ldots, p),$$

wobei $\mathfrak{u}(s) := \tilde{\mathfrak{u}}(s \cdot h)$ ist. Also ist $\mathfrak{u} \in L^1([0,1]; U)$. Es sei nun $\xi_i , i \in \{1, \ldots, p\}$, wie oben, aber mit der gerade konstruierten Funktion $\mathfrak{u}$ als Integrand. Dann ist der Vektor $\left(\xi_1^T, \ldots, \xi_p^T\right)^T \in \mathcal{I}_{m,p}$ und es gilt $\hat{\mathcal{I}}_{m,p} \subset \mathfrak{H}_{m,p} \cdot \mathcal{I}_{m,p}$.

Insgesamt haben wir also

$$\mathcal{I}_{m,p} = \mathfrak{H}_{m,p}^{-1} \cdot \hat{\mathcal{I}}_{m,p} \,.$$

Als nächstes definieren wir die Menge

$$\hat{\mathcal{M}}_{m,p} := \left\{ \left(\mu_0^T, \ldots, \mu_{p-1}^T\right)^T \in \mathbb{R}^{pm} \mid \mu_i = \int_0^h t^i \mathfrak{u}\,(t)\; dt \;\; (i = 0, \ldots, p-1), \right.$$
$$\left. \mathfrak{u} \in L^1\left([0,h]; U\right) \right\} \,.$$

Nach Lemma 5.2.1, angewendet für $a = h$, besteht zwischen den ζ_i aus $\hat{\mathcal{I}}_{m,p}$ und den μ_i aus $\hat{\mathcal{M}}_{m,p}$ folgender Zusammenhang:

$$\mu_{i-1} = h^{i-1}\zeta_1 - h^{i-2}(i-1)\zeta_2 + h^{i-3}(i-1)(i-2)\zeta_3 - \ldots + h^0\,(-1)^{i-1}\,(i-1)!\zeta_i \,,$$

wobei den ζ_i und μ_i die gleiche Funktion $\mathfrak{u}$ zugrunde liegt. Um diesen Zusammenhang in Matrixschreibweise darzustellen, definieren wir die Matrix

$$\hat{\mathfrak{T}}_{m,p} := \begin{pmatrix} \hat{t}_{1,1} \cdot \mathfrak{E}_m & 0 \cdot \mathfrak{E}_m & 0 \cdot \mathfrak{E}_m \\ \vdots & \ddots & 0 \cdot \mathfrak{E}_m \\ \hat{t}_{p,1} \cdot \mathfrak{E}_m & \cdots & \hat{t}_{p,p} \cdot \mathfrak{E}_m \end{pmatrix} \in \mathbb{R}^{pm \times pm}$$

mit $\hat{t}_{i,j} = \frac{(-1)^{j-1}(i-1)!}{(i-j)!} h^{i-j}$ für $j \leq i$. Ähnlich wie in Abschnitt 5.2 hat man dann den Zusammenhang:

$$\hat{\mathcal{I}}_{m,p} = \hat{\mathfrak{T}}_{m,p}^{-1} \cdot \hat{\mathcal{M}}_{m,p} \,,$$

wobei aber hier alle Größen von der Schrittweite h abhängen. Die Matrix $\hat{\mathfrak{T}}_{m,p}$ ist ebenso wie $\mathfrak{T}_{m,p}$ eine untere Dreiecksmatrix und damit invertierbar. Wir haben also insgesamt

$$\mathcal{I}_{m,p} = \left(H_{m,p}^{-1} \cdot \hat{\mathfrak{T}}_{m,p}^{-1}\right) \cdot \hat{\mathcal{M}}_{m,p} \,. \tag{5.19}$$

Damit ergibt sich für die Ferretti-Verfahren aus (5.1) folgende mengenwertige Iteration

$$\mathcal{R}_h\left((i+1)h, x_0\right) = \mathfrak{C} \cdot \mathcal{R}_h\left(ih, x_0\right) + \mathfrak{D} \cdot \hat{\mathcal{M}}_{m,p} \qquad (i = 0, \ldots, N-1) \,,$$

wobei

$$\begin{aligned} \mathfrak{D} &= \left[h\mathfrak{B} \mid h^2\mathfrak{A}\mathfrak{B} \mid h^3\mathfrak{A}^2\mathfrak{B} \mid \ldots \mid h^p\mathfrak{A}^{p-1}\mathfrak{B}\right] \mathfrak{H}_{m,p}^{-1} \cdot \hat{\mathfrak{T}}_{m,p}^{-1} \\ &= \left[\mathfrak{B} \mid \mathfrak{A}\mathfrak{B} \mid \mathfrak{A}^2\mathfrak{B} \mid \ldots \mid \mathfrak{A}^{p-1}\mathfrak{B}\right] \cdot \hat{\mathfrak{T}}_{m,p}^{-1} \end{aligned}$$

ist. Wir können uns also auch die Berechnung der Matrix $\mathfrak{H}_{m,p}^{-1}$ sparen.

Die Menge $\hat{\mathcal{M}}_{m,p}$ kann ebenso wie $\mathcal{M}_{m,p}$ als Aumann-Integral dargestellt werden mit dem gleichen Integranden, nur wird eben über das kürzere Intervall $[0,h]$ integriert. Wegen (5.19) und den Rechenregeln für Stützpunktmengen kann man die Stützpunktberechnung statt an $\mathcal{M}_{m,p}$ an $\hat{\mathcal{M}}_{m,p}$ durchführen. Es gilt dann fast unverändert die Formel (5.18) von vorhin

$$y\left(\mathfrak{l},\hat{\mathcal{M}}_{m,p}\right)=\int_0^h \mathfrak{Q}_p(t)\cdot\mathfrak{m}\,dt+\int_0^h \begin{pmatrix}\mathfrak{p}(t)\\ t\cdot\mathfrak{p}(t)\\ \vdots\\ t^{p-1}\cdot\mathfrak{p}(t)\end{pmatrix}\frac{1}{\sqrt{p_1(t)^2+\ldots+p_m(t)^2}}\,dt\,, \quad (5.20)$$

wenn $\mathfrak{p}$ in $[0,1]$ nicht verschwindet. Zur Berechnung dieser Integrale gilt das gleiche wie vorher. Das erste Integral kann man leicht analytisch vorausberechnen. Das zweite Integral kann im allgemeinen nur numerisch integriert werden.

Wir werden uns daher im Folgenden mit numerischen Integration befassen.

5.4.3.2. Numerische Integration

Zur numerischen Integration müssen wir unseren Blickwinkel weg von der punktwertigen Umsetzung der Verfahren durch Stützpunkte hin zum mengenwertigen Hintergrund lenken.

Es ist $\hat{\mathcal{M}}_{m,p}=\int_0^h \mathfrak{Q}_p(t)\cdot B_r\left(\mathfrak{m}\right)\,dt$ ein Aumann-Integral. Mit Hilfe von Stützpunkten, kann man es punktwertig ausrechnen. Aber dies hat nur etwas mit der Umsetzung zu tun, jedoch die Theorie für die numerische Integration muss sich am Aumann-Integral orientieren. Insbesondere muss die Konvergenzordnung der mengenwertigen numerischen Quadraturverfahren, mit der Konvergenzordnung der mengenwertigen Ferretti-Verfahrens übereinstimmen.

Die Konvergenzordnung der mengenwertigen Quadraturverfahren wiederum hängt von der Glattheit der Stützfunktion des Integranden ab (siehe [4, Kapitel 1]).

Deshalb berechnen wir nun die Stützfunktion des Integranden $\mathfrak{Q}_p(t)\cdot B_r\left(\mathfrak{m}\right)$ für eine Richtung $\mathfrak{l}\in S_{mp-1}$ mit Hilfe der Rechenregeln aus Satz 1.3.6:

$$\delta^*\left(\mathfrak{l},\mathfrak{Q}_p(t)\cdot B_r\left(\mathfrak{m}\right)\right)=\delta^*\left(\mathfrak{Q}_p(t)^T\mathfrak{l},B_r\left(\mathfrak{m}\right)\right)\,.$$

Mit Hilfe von Beispiel 1.3.6 erhalten wir

$$\begin{aligned}\delta^*\left(\mathfrak{l},\mathfrak{Q}_p(t)\cdot B_r\left(\mathfrak{m}\right)\right) &= \left\langle\mathfrak{Q}_p(t)^T\mathfrak{l},\mathfrak{m}\right\rangle+r\cdot\left\|\mathfrak{Q}_p(t)^T\mathfrak{l}\right\|_2\\ &= \mathfrak{p}\left(t\right)^T\mathfrak{m}+r\cdot\sqrt{p_1^2\left(t\right)+\cdots+p_m^2\left(t\right)}\,.\end{aligned}$$

Wir werden zeigen, dass diese Stützfunktion absolutstetig ist und eine Ableitung von beschränkter L^1-Variation hat. Die Begriffe beschränkte (=endliche) Variation und L^1-Variation findet man [4, Definition 0.2.1 + 0.2.4].

Der erste Summand $\mathfrak{p}(t)^T \mathfrak{m}$ ist ein Polynom. Er ist absolutstetig und hat auf $[0,1]$ eine Ableitung von endlicher Variation (vgl. Satz 1.8.7(ii) und [25, Kap. IX, §2, Satz 1]). Wir wollen den zweiten Summanden $f(t) := r \cdot \sqrt{p_1^2(t) + \cdots + p_m^2(t)}$ untersuchen. Mit dem Hauptsatz der Algebra können wir die reellen Nullstellen eines Polynoms abspalten. Wir erhalten

$$p_1^2(t) + \cdots + p_m^2(t) = \tilde{p}(t)^2 \cdot q(t),$$

wobei $\tilde{p}(t) := \prod_{i=1}^{l} (t - a_i)^{k_i}$ alle reellen Nullstellen von $p_1^2(t) + \cdots + p_m^2(t)$ enthält und $q(t)$ der Rest ist. Diese reellen Nullstellen müssen gerade Vielfachheit haben, deswegen geht $\tilde{p}$ quadratisch ein. Das Restpolynom $q(t)$ hat keine reellen Nullstellen und ist stets positiv. Es ist stetig und nimmt deswegen auf $[0,1]$ sein Minimum $s > 0$ an. Damit haben wir

$$f(t) = r\,|\tilde{p}(t)|\,\sqrt{q(t)}\,.$$

Da $q(t) > s$ ist in $[0,1]$, ist $f(\cdot)$ in $[0,1] - \tilde{N}$ stetig differenzierbar mit $\tilde{N} := \{a_1, \ldots, a_l\}$. Die Ableitung für $t \in [0,1]$ ist (für $t \in \tilde{N}$ setzen wir sie so)

$$f'(t) = r \cdot \left(\operatorname{sign}(\tilde{p}(t))\,\tilde{p}'(t)\sqrt{q(t)} + |\tilde{p}(t)|\,\frac{q'(t)}{2\sqrt{q(t)}} \right).$$

Sie ist auf $[0,1]$ beschränkt. Mit Satz 1.8.7(i) ist f also absolutstetig. Man kann des Intervall $[0,1]$ für jede der Funktionen $\operatorname{sign}(\tilde{p}(\cdot))$, $\tilde{p}'(\cdot)$, $\sqrt{q(\cdot)}$, $|\tilde{p}(\cdot)|$ und $q'(\cdot)$ so zerlegen, dass sie auf jedem Teilintervall monoton sind. Deswegen ist $f'(\cdot)$ von beschränkter Variation (vgl. [25, Kap. IX, §3, Folg. 2 + Satz 3 und 4]). Damit ist $\delta^*(\mathfrak{l}, \mathfrak{Q}_p(t) \cdot B_r(\mathfrak{m}))$ auf $[0,1]$ für alle $\mathfrak{l}$ absolutstetig (vgl. Satz 1.8.2) . Und die Ableitung

$$(\delta^*)'(\mathfrak{l}, \mathfrak{Q}_p(t) \cdot B_r(\mathfrak{m})) = \mathfrak{p}'(t)^T \mathfrak{m} + r \cdot \left(\operatorname{sign}(\tilde{p}(t))\,\tilde{p}'(t)\sqrt{q(t)} + |\tilde{p}(t)|\,\frac{q'(t)}{2\sqrt{q(t)}} \right),$$

ist von beschränkter L^1-Variation (vgl. [25, Kap. VIII, §3, Satz 3] und [4, Bemerkung 0.2.6.]). Da wir $h < 1$ vorausgesetzt haben (siehe Kapitel 2, Gleichung (2.3)), gilt dies auch auf dem Intervall $[0,h]$.

Nach [4, Korollar 1.3.7 (iv)], angewendet für $\nu = 0$, kann man also für die k-te abgeschlossene Newton-Cotes-Formel mit $k \geq 1$ nur Konvergenzordnung 2 erwarten. Ein Quadraturverfahren um das mengenwertige Integral $\int_0^h \mathfrak{Q}_p(t) \cdot B_r(\mathfrak{m})\,dt$ für $k = 1$ zu berechnen ist nach [4, Beispiel 1.1.3] die Trapezregel

$$\mathcal{T}^1(\mathfrak{Q}_p(\cdot) \cdot B_r(\mathfrak{m})) := \frac{h}{2}\left(\mathfrak{Q}_p(0) \cdot B_r(\mathfrak{m}) + \mathfrak{Q}_p(h) \cdot B_r(\mathfrak{m})\right),$$

bzw. die iterierte Trapezregel

$$\mathcal{T}^N(\mathfrak{Q}_p(\cdot) \cdot B_r(\mathfrak{m})) := \frac{\hat{h}}{2}\left(\mathfrak{Q}_p(0) \cdot B_r(\mathfrak{m}) + 2\sum_{i=1}^{N-1} \mathfrak{Q}_p\left(i \cdot \hat{h}\right) \cdot B_r(\mathfrak{m}) + \mathfrak{Q}_p(h) \cdot B_r(\mathfrak{m})\right),$$

wobei $\hat{h} = \frac{h}{N}$ ist.
Es gilt also

$$d_H\left(\int_0^h \mathfrak{Q}_p(t)\cdot B_r(\mathfrak{m})\,dt\,,\,\mathcal{T}^N\left(\mathfrak{Q}_p(\cdot)\cdot B_r(\mathfrak{m})\right)\right) = \mathcal{O}\left(\hat{h}^2\right).$$

Dabei ist h die Schrittweite des mengenwertigen Ferretti-Verfahrens. Falls man für Ferretti-Verfahren der Ordnung > 2 eine höhere Konvergenzordnung der numerischen Integration braucht, muss man die iterierte Trapezregel verwenden. Dabei wird das Integrationsintervall $[0,h]$ in $\hat{N}$ äquidistante Abschnitte unterteilt mit der Schrittweite $\hat{h} = \frac{h}{\hat{N}}$. Wenn man für das Ferretti-Verfahren einen Integrationsfehler der Ordnung h^k braucht, dann muss man $\hat{N} \geq h^{1-\frac{k}{2}}$ wählen. Denn dann folgt

$$\hat{h}^2 = \left(\frac{h}{\hat{N}}\right)^2 \leq h^k.$$

Um den Aufwand für die Integration möglichst gering zu halten, wählen wir $\hat{N}$ dabei möglichst klein. Dies geschieht, indem wir $h^{1-\frac{k}{2}}$ einfach auf die nächste natürliche Zahl aufrunden, d.h. wir setzten

$$\hat{N} = \left\lceil h^{1-\frac{k}{2}} \right\rceil. \tag{5.21}$$

Diese mengenwertige Trapezregel realisieren wir im Rechner mittels Stützpunkten. Deswegen berechnen wir jetzt die Stützpunktmenge dieser Trapezregel für $\mathfrak{l} \in \mathbb{R}^{mp}$

$$\begin{aligned} Y\left(\mathfrak{l},\mathcal{T}^N\left(\mathfrak{Q}_p(\cdot)\cdot B_r(\mathfrak{m})\right)\right) &= \frac{\hat{h}}{2}\left[\mathfrak{Q}_p(0)\cdot Y\left(\mathfrak{p}(0),B_r(\mathfrak{m})\right)+\mathfrak{Q}_p(h)\cdot Y\left(\mathfrak{p}(h),B_r(\mathfrak{m})\right)\right] \\ &\quad +\hat{h}\sum_{i=1}^{N-1}\mathfrak{Q}_p\left(i\hat{h}\right)\cdot Y\left(\mathfrak{p}\left(i\hat{h}\right),B_r(\mathfrak{m})\right). \end{aligned}$$

Setzt man nun die Stützpunktformel (siehe Beispiel 1.4.5) für die Kugel ein, so erhält man

$$\begin{aligned} y\left(\mathfrak{l},\mathcal{T}^N\left(\mathfrak{Q}_p(\cdot)\cdot B_r(\mathfrak{m})\right)\right) &= \frac{\hat{h}}{2}\left[\mathfrak{Q}_p(0)\left(\mathfrak{m}+r\cdot\frac{\mathfrak{p}(0)}{\|\mathfrak{p}(0)\|_2}\right)+\mathfrak{Q}_p(h)\left(\mathfrak{m}+r\cdot\frac{\mathfrak{p}(h)}{\|\mathfrak{p}(h)\|_2}\right)\right] \\ &\quad +\hat{h}\sum_{i=1}^{N-1}\mathfrak{Q}_p\left(i\hat{h}\right)\left(\mathfrak{m}+r\cdot\frac{\mathfrak{p}\left(i\hat{h}\right)}{\left\|\mathfrak{p}\left(i\hat{h}\right)\right\|_2}\right), \end{aligned}$$

falls aber zufällig $\mathfrak{p}\left(j\hat{h}\right) = 0_{\mathbb{R}^{mp}}$ für ein $0 \leq j \leq N$, dann ist $Y\left(\mathfrak{p}\left(j\hat{h}\right),B_r(\mathfrak{m})\right) = B_r(\mathfrak{m})$ und wir wählen einfach einen Stützpunkt aus. Denn wegen Satz 1.4.3 3) reicht

es nur einen Stützpunkt aus der Stützpunktmenge zu berechnen. Dieser Fall kommt jedoch in der numerischen Praxis nur sehr selten vor, da die Menge auf der $\mathfrak{p}(t)$ verschwindet ein Nullmenge ist.

Viel wahrscheinlicher und problematischer ist der Fall, dass

$$\left\| \mathfrak{p}\left(j\hat{h}\right) \right\|_2 \leq 1000 \cdot eps$$

ist, wobei eps die Maschinengenauigkeit ist. Zwar können die Komponenten es Vektors $\frac{\mathfrak{p}(jh)}{\left\|\mathfrak{p}(j\hat{h})\right\|_2}$ nicht explodieren, sondern sind viel mehr alle ≤ 1. Aber es kann zu starken Rundungsfehlern kommen, da eventuell schon die erste oder zweite signifikante Stelle gerundet ist. Jedoch ist auch dieses Problem bei den numerischen Tests nicht aufgetreten bzw. hat sich nicht negativ auf die Ergebnisse ausgewirkt.

Deswegen habe ich bei der Implementation dieses Problem einfach ignoriert. Auf Wunsch kann jedoch der Anwender die Software so einstellen, dass Warnungen ausgegeben werden, wenn $\left\| \mathfrak{p}\left(j\hat{h}\right) \right\|_2$ eine bestimmte (einstellbare) Schwelle unterschreitet.

Kapitel 6.

Anwendungen und Beispiele

Seit der Themenausgabe für diese Arbeit lag der Schwerpunkt auf der Umsetzung der Ergebnisse von Ferretti aus [13] in konkrete Verfahren und auf der Implementierung und dem Test dieser Verfahren auf dem Rechner. Dieses Vorgehen hat sich als sehr fruchtbar für diese Arbeit erwiesen, was sich besonders im vorigen Kapitel niedergeschlagen hat.

In diesem Kapitel nun soll näher auf die numerische Umsetzung der vorgestellten Verfahren auf dem Rechner eingegangen werden. Es wird die Theorie anhand von Beispielen ergänzt und illustriert. Außerdem sollen noch einige Besonderheiten der implementierten Verfahren vorgestellt werden.

Im ersten Abschnitt werden theoretische Grundlagen zum Verständnis der Beispiele gelegt. Außerdem folgen einige Bemerkungen über die Implementierung und Tests der Verfahren.

Im zweiten Abschnitt werden dann anhand von mehreren Beispielen einzelne Aspekte der vorgestellten numerischen Verfahren demonstriert oder Besonderheiten dieser Verfahren aufgezeigt.

6.1. Einführung und Grundlagen

6.1.1. Hard- und Software

Das selbsterstellte Programmpaket zur Berechnung der erreichbaren Menge eines linearen Kontrollsystems wurde in C++ programmiert. Dabei wurde für den Umgang mit Matrizen und Vektoren die effiziente Bibliothek `uBlas` verwendet, die Teil des Projektes `Boost` ist. Zur Vereinfachten Benutzung wurde eine Schnittstellte zu dem Programm `Scilab` eingefügt, welches ähnliche wie `Matlab` eine graphische Benutzeroberfläche besitzt und eine Sprache für - sowie eine Sammlung von numerischen Algorithmen bietet. Die numerischen Test wurden auf einem gewöhnlichen PC mit einem AMD Duron Prozessor mit 1200 MHz Taktrate und 256 MB Arbeitsspeicher durchgeführt. Die erstellte Software stellt also keine besonderen Ansprüche an die Hardware eines Rechners. Der Speicherplatzbedarf für die benötigten Daten ist gering und hängt im Wesentlichen von der Anzahl der Iterationen ab. Zusätzlich zu

≤ 1000 Byte Datenspeicher (grob geschätzt), die immer gebraucht werden, werden ca. $10 \cdot n \cdot N$ Byte benötigt, wobei N die Anzahl der Iterationen ist und n die Dimension des Problems ist.

6.1.2. Die implementierten Verfahren

Insgesamt wurden drei Verfahrensklassen implementiert. Es handelt sich dabei immer um die mengenwertigen Verfahren, die in Kapitel 3 entwickelt wurden. Die Verfahren dieser drei Klassen unterscheiden sich nur in der Art wie sie Stützpunkte an die Momentenmenge berechnen und welche Kontrollmengen sie bearbeiten können.

In der ersten Verfahrensklasse muss der Steuerbereich U ein Quader sein. Dabei wurde die Stützpunktberechnung an die Momentenmenge $\mathcal{M}_{m,p}$ mit Algorithmus 5.4.4 durchgeführt, wobei die Berechnung der Polynomnullstellen den meisten Aufwand verursacht. Dies entspricht einer analytische Berechnung des Aumann-Integrals der Momentenmenge $\mathcal{M}_{m,p}$, deswegen wird auf diese Verfahren mit dem Zusatz "mit analytischer Integration bzw. Stützpunktberechnung" Bezug genommen. Es wurden die folgende mengenwertige RK-Verfahren implementiert:

- Mit Konvergenzordnung 2: das Heun-Verfahren (auch Euler-Cauchy-Verfahren genannt) und das verbesserte Euler-Verfahren.
- Mit Konvergenzordnung 3: das Heun3-Verfahren.
- Mit Konvergenzordnung 4: das klassische Runge-Kutta-Verfahren.

Außerdem wurden die mengenwertigen Ferretti-Verfahren der Ordnung 2,3 und 4 programmiert, welche ebenfalls zu dieser Verfahrensklasse gehören. Um zu verdeutlichen, dass die Ferretti-Verfahren keiner Ordnungsbeschränkung unterliegen, wurde auch das Ferretti-Verfahren der Ordnung 8 implementiert. Hierfür musste allerdings die Nullstellenbestimmung numerisch erfolgen, wie dies in Unterabschnitt 5.4.2 gezeigt wurde. Ein modifiziertes RK-Verfahren mit einer größeren Ordnung als 4 konnte nicht implementiert werden, weil ein solches Verfahren, wie in Bemerkung 3.1.4 erläutert, nicht existiert.

Zum Vergleich wurden alle diese Verfahren, bis auf das Ferretti-Verfahren der Ordnung 8, auch mit numerischer Integration implementiert. D.h. das Aumann-Integral der Momentenmenge $\mathcal{M}_{m,p}$ wird mit mengenwertigen Quadraturformeln approximiert ganz analog wie dies in Unterabschnitt 5.4.3 für eine Kugelsteuermenge gemacht wurde. Diese Verfahren bilden die zweite Verfahrensklasse und werden im Folgenden auch als Verfahren mit numerischer Integration bzw. Stützpunktberechnung bezeichnet. Genau wie bei der Kugelsteuermenge kann mangels Glattheit der Stützfunktion hierfür nur ein Verfahren der Ordnung 2 eingesetzt werden. Es wird die iterierte Trapezregel mit einer zweiten, kleineren Schrittweite eingesetzt, welche

so gewählt wird, dass die Konvergenzordnung des mengenwertigen Verfahrens erhalten bleibt.

Bei der dritten Verfahrensklasse muss der Steuerbereich eine Kugel sein. Alle Verfahren aus der zweiten Klasse wurden hier implementiert. Dabei erfolgt die Berechnung des Aumann-Integrals wieder durch eines mengenwertiges Quadraturverfahren. Wie schon in Unterabschnitt 5.4.3 ausführlich erläutert wurde, kommt dafür nur ein Verfahren der Ordnung 2 in Frage. Die Verfahren dieser Klasse werden auch als Verfahren mit Kugelsteuerbereich bezeichnet.

Es hat sich gezeigt, dass zwei Verfahren der gleichen Klasse bei gleicher Konvergenzordnung immer exakt die gleichen Ergebnisse liefern. Zum Beispiel berechnen das Heun-Verfahren und das verbesserte Eulerverfahren mit analytischer Integration bei gleichen Richtungen immer exakt die gleichen Stützpunkte für die erreichbare Menge. Dies war der Anstoß für Satz 3.2.5 aus Kapitel 3, welches besagt das alle modifizierten RK-Verfahren einer Ordnung mit dem Ferretti-Verfahren derselben Ordnung übereinstimmen. In Beispiel 6.2.1 wird dies exemplarisch dargestellt.

Da diese mengenwertigen Verfahren im Rechner mittels Stützpunkten umgesetzt wurden, führt für jede Richtung, in der der Stützpunkt an die erreichbare Menge berechnet werden soll, das entsprechende punktwertiges Verfahren die Berechnung durch (siehe Abschnitt 5.1).

Generell kann man sagen, dass mit Verfahren höherer Ordnung eine größere Genauigkeit erzielt werden kann als mit Verfahren geringerer Ordnung, selbst bei gröberer Schrittweite. Bei den Verfahren aus der ersten Klasse können sich höher konvergente Verfahren bei gleicher Genauigkeit auch im Bezug auf Rechenzeit durchsetzen. Für die Verfahren der anderen beiden Klassen gilt diese Aussage auch, jedoch ist der Abstand dabei nicht mehr so groß. Denn bei diesen Verfahren erhöht sich mit zunehmender Genauigkeit der Aufwand für die numerische Integration stark. Und zwar je höher die Konvergenzordnung des Verfahrens ist, desto schneller erhöht sich der Aufwand.

Alle Plots in diesem Kapitel wurde alle mit `Matlab` erstellt.

6.1.3. Verschiedene Resultate

Im Folgenden sollen noch einige Ergebnisse aus [4, Abschnitt 3.1] zusammengestellt werden, die im Bezug auf die folgenden Beispiele wichtig oder wenigstens interessant sind.

6.1.3.1. Approximation der diskreten erreichbaren Menge

Die Umsetzung der vorgestellten mengenwertigen Verfahren mit Hilfe von Stützpunkten basiert auf Satz 1.4.3. Danach gilt für eine nichtleere, kompakte und konvexe

Menge $K \subset \mathbb{R}^n$:

$$K = \overline{\text{co}} \left\{ y\left(\mathfrak{l}, K\right) \mid \mathfrak{l} \in S_{k-1} \right\}$$

Nun können wir im Rechner die Stützpunkte $y\left(\mathfrak{l}, S_{n-1}\right)$ nur für endlich viele Richtungen $\mathfrak{l} \in S_{n-1}$ berechnen. Da die Rechenzeit linear mit der Anzahl der Stützpunkte steigt, kann diese Zahl auch nicht beliebig groß werden. Es wird also die diskrete erreichbare Menge $\mathcal{R}_h\left(t_f, x_0\right)$, welche die Approximation eines mengenwertigen Verfahrens für die erreichbare Menge ist, wiederum durch

$$\tilde{\mathcal{R}}_h\left(t_f, x_0\right) = \text{co}\left\{ y\left(\mathfrak{l}, \mathcal{R}_h\left(t_f, x_0\right)\right) \mid \mathfrak{l}^i \in S_{n-1}, i = 1, \ldots, M \right\}$$

approximiert. Es fragt sich nun wie groß dieser zweite Approximationsfehler ist. Diese Frage beantwortet der folgende Satz.

Satz 6.1.1. *Sei $K \subset \mathbb{R}^n$ eine kompakte und konvexe Menge und $\mathcal{G}_M := \{\mathfrak{l}^i \mid i = 1, \ldots, M\} \subset S_{n-1}$ mit $d_H\left(\mathcal{G}_M, S_{n-1}\right) < \epsilon$. Weiter sei*

$$K_M := co\left\{ y\left(\mathfrak{l}^i, K\right) \mid i = 1, \ldots, M \right\} .$$

Dann gilt

$$d_H\left(K, K_M\right) \leq 2 \cdot \|K\|_2 \cdot \epsilon \, .$$

Beweis. siehe [16, Proposition 6.11]. □

Dabei ist $\mathrm{d_H}\left(\mathcal{G}_M, S_{n-1}\right) = \mathrm{d}\left(S_{n-1}, \mathcal{G}_M\right)$ ein Maß dafür wie groß die Lücken der Richtungsvektoren $\mathfrak{l} \in \mathcal{G}_M$ in S_{n-1} sind. Deswegen ist es am besten die Richtungsvektoren gleichverteilt über die Sphäre S_{n-1} zu wählen, wenn man keine nähere Informationen über die erreichbare Menge hat.

6.1.3.2. Parametrisierung der Sphäre

Zur Wahl der Richtungsvektoren auf der Sphäre ist folgende bijektive Parametrisierung nützlich:

$$\begin{aligned} \varphi : \left((0,\pi)^{n-2} \times [0, 2\pi)\right) \cup \{0_{\mathbb{R}^{n-1}}, \pi \mathfrak{e}_1\} &\longrightarrow S_{n-1} \\ (\alpha_1, \alpha_2 \ldots, \alpha_{n-1}) &\longmapsto \begin{pmatrix} \cos\left(\alpha_1\right) \\ \sin\left(\alpha_1\right)\cos\left(\alpha_2\right) \\ \vdots \\ \prod_{i=1}^{n-2} \sin\left(\alpha_i\right) \cdot \cos\left(\alpha_{n-1}\right) \\ \prod_{i=1}^{n-1} \sin\left(\alpha_i\right) \end{pmatrix} , \end{aligned}$$

dabei ist $\mathfrak{e}_1 = (1, 0, \ldots, 0)^T \in \mathbb{R}^{n-1}$ der erste kanonische Einheitsvektor.

Seien nun $M = (M_1, \ldots, M_{n-1})$ ein Vektor mit $M_i \in \mathbb{N}\,(i = 1, \ldots, n-1)$. Die Menge

$$\mathcal{P}_M := \left(\bigoplus_{i=1}^{n-2} \left\{ \frac{k_i \pi}{M_i} \mid k_i = 1, \ldots, M_i - 1 \right\} \times \left\{ \frac{k_{n-1} 2\pi}{M_{n-1}} \mid k_{n-1} = 0, \ldots, M_{n-1} - 1 \right\} \right) \cup \{0_{\mathbb{R}^{n-1}}, \pi \mathfrak{e}_1\}$$

ist eine äquidistante Unterteilung der Definitionsmenge von φ. Jetzt setzen wir

$$\mathcal{G}_M := \varphi\left(\mathcal{P}_M\right)$$

und erhalten eine Menge von $\prod_{i=1}^{n-2} (M_i - 1) \cdot M_{n-1} + 2$ Punkten in S_{n-1}, die über die ganze Sphäre verteilt sind. Für $n = 2$ entspricht diese Parametrisierung den Polarkoordinaten:

$$\varphi : [0, 2\pi) \longrightarrow S_1 \quad , \quad \varphi(\alpha) = (\cos(\alpha), \sin(\alpha))^T$$

Dann ist $\mathcal{P}_M = \left\{ \frac{k2\pi}{M} \mid k = 0, \ldots, M-1 \right\}$. In diesem Fall sind die Vektoren des Bildes $\mathcal{G}_M = \varphi\left(\mathcal{P}_M\right)$ sogar geometrisch exakt gleichverteilt über S_1 bzgl. der 2-Norm. In den Beispielen im nächsten Abschnitt wurde diese Parametrisierung zusammen mit der Menge $\mathcal{P}_M$ verwendet um Richtungsvektoren zu erzeugen, die symmetrisch über die Sphäre verteilt sind.

6.1.3.3. Schätzung der Konvergenzordnung

Jetzt soll noch eine Heuristik, die sich in der Praxis gut bewährt hat, vorgestellt werden, mit der man die Konvergenzordnung eines mengenwertigen Verfahrens schätzen kann. Es sei $\mathcal{G}_M = \{\mathfrak{l}^i \mid i = 1, \ldots, M\} \subset S_{n-1}$ eine Menge von Richtungsvektoren, die zur Stützpunktberechnung verwendet werden.

Weiter sei $\mathcal{R} := \operatorname{co}\{\tilde{y}_i \mid i = 1, \ldots, M\}$ eine Referenzmenge, wobei die Vektoren $\tilde{y}_i = y\left(\mathfrak{l}^i, \mathcal{R}_h\left(t_f, x_0\right)\right)$ möglichst gute Approximationen der Stützpunkte der erreichbaren Menge sein sollen. Am Besten verwendet man für deren Berechnung das Verfahren höchster Ordnung, das für den jeweiligen Steuerbereich zur Verfügung steht, mit feiner Diskretisierung. Denn die Referenzmenge soll hier die exakte erreichbare Menge ersetzen.

Weiter seien $\mathcal{R}_k = \operatorname{co}\left\{y_i^k \mid i = 1, \ldots, M\right\}$, $k = 1, 2$, zwei Approximationen der erreichbaren Menge. Die Stützpunkte $y_i^k = y\left(\mathfrak{l}^i, \mathcal{R}_{h_k}\left(t_f, x_0\right)\right)$ seien mit dem gleichen Verfahren für die Schrittweiten $h_k = \frac{t_f}{N_k}$ $(k = 1, 2)$ erzeugt. Mit Hilfe von Satz 1.3.4 kann man den Hausdorff-Abstand zwischen der Referenzmenge $\mathcal{R}$ und den Approximationen $\mathcal{R}_k$ so

$$\mathrm{d_H}\left(\mathcal{R}, \mathcal{R}_k\right) \approx \max_{i=1,\ldots,M} \left| \delta^*\left(\mathfrak{l}^i, \mathcal{R}\right) - \delta^*\left(\mathfrak{l}^i, \mathcal{R}_k\right) \right|$$

annähern. Mit Hilfe der Stützpunkte, die wir schon berechnet haben, können wir diese Stützfunktionen ganz leicht berechnen. Es ist natürlich

$$y\left(\mathfrak{l}^i, \mathcal{R}_k\right) = y_i^k \quad \text{und} \quad y\left(\mathfrak{l}^i, \mathcal{R}\right) = \tilde{y}_i\,.$$

Damit lassen sich die Stützfunktionen, nach Definition der Stützfunktion und der Stützpunkte, berechnen als

$$\delta^*\left(\mathfrak{l}^i, \mathcal{R}_k\right) = \left(\mathfrak{l}^i\right)^T y_i^k \quad \text{und} \quad \delta^*\left(\mathfrak{l}^i, \mathcal{R}\right) = \left(\mathfrak{l}^i\right)^T \tilde{y}_i\,.$$

Also gilt

$$\mathrm{d_H}\left(\mathcal{R}, \mathcal{R}_k\right) \approx \max_{i=1,\dots,M} \left|\left(\mathfrak{l}^i\right)^T \left(y_i^k - \tilde{y}_i\right)\right|\,.$$

Um die Konvergenzordnung p zu schätzen machen wir nun den Ansatz

$$\mathrm{d_H}\left(\mathcal{R}, \mathcal{R}_k\right) \approx C \cdot \left(h_k\right)^p\,,$$

der sich von dem Konvergenzsatz 3.3.2 ableitet. Damit gilt dann

$$\frac{\max_{i=1,\dots,M} \left|\left(\mathfrak{l}^i\right)^T \left(y_i^1 - \tilde{y}_i\right)\right|}{\max_{i=1,\dots,M} \left|\left(\mathfrak{l}^i\right)^T \left(y_i^2 - \tilde{y}_i\right)\right|} \approx \frac{\mathrm{d_H}\left(\mathcal{R}, \mathcal{R}_1\right)}{\mathrm{d_H}\left(\mathcal{R}, \mathcal{R}_2\right)} \approx \left(\frac{N_2}{N_1}\right)^p\,.$$

Durch Anwendung der Logarithmusfunktion auf beiden Seiten erhalten wir als Schätzung für die Konvergenzordnung

$$p \approx \frac{\ln\left(\frac{\max_{i=1,\dots,M} \left|\left(\mathfrak{l}^i\right)^T \left(y_i^1 - \tilde{y}_i\right)\right|}{\max_{i=1,\dots,M} \left|\left(\mathfrak{l}^i\right)^T \left(y_i^2 - \tilde{y}_i\right)\right|}\right)}{\ln\left(\frac{N_2}{N_1}\right)}\,.$$

In Einzelfällen kann man die Stützpunkte an die erreichbare Menge auch analytisch berechnen. Dann konstruiert man die Referenzmenge mit diesen Stützpunkten.

Wenn im folgenden Abschnitt von dem Hausdorff-Abstand einer Approximation $\mathcal{R}_k$ zu Referenzmenge, also $\mathrm{d_H}\left(\mathcal{R}, \mathcal{R}_k\right)$, die Rede ist, so ist immer der approximierte Abstand

$$\max_{i=1,\dots,M} \left|\left(\mathfrak{l}^i\right)^T \left(y_i^k - \tilde{y}_i\right)\right| \approx \mathrm{d_H}\left(\mathcal{R}, \mathcal{R}_k\right) \tag{6.1}$$

gemeint.

6.1.3.4. Eine andere Möglichkeit zur Berechnung der erreichbaren Menge

Neben den in dieser Arbeit vorgestellten Verfahren, existiert noch (mindestens) ein weiterer Zugang um die erreichbare Menge eines linearen Kontrollsystems bzw. einer

linearen Differentialinklusion zu berechnen. Eine solcher Zugang soll hier vorgestellt werden. Es sei folgendes lineares Kontrollproblem gegeben:

$$\begin{aligned} x'(t) &= \mathfrak{A}x(t) + \mathfrak{B}\mathfrak{u}(t) \quad (\text{für } t \in I) \\ x(0) &= x_0 \\ \mathfrak{u}(t) &\in U \quad \forall t \in I\,, \end{aligned}$$

mit $\mathfrak{A} \in \mathbb{R}^{n\times n}$, $\mathfrak{B} \in \mathbb{R}^{n\times m}$, $I = [0, t_f]$ ein kompaktes Intervall, $U \subset \mathbb{R}^m$ ein kompakter Kontrollbereich und $\mathfrak{u} \in L^1(I)^m$. Die erreichbare Menge dieses Kontrollproblems stimmt nach Satz 2.3.5 mit der erreichbaren Menge der folgenden zugeordneten Differentialinklusion überein:

$$\begin{aligned} x'(t) &\in \mathfrak{A}x(t) + \mathfrak{B}U \quad (\text{für } t \in I) \\ x(0) &= x_0 \end{aligned}$$

Nach [4, Satz 2.1.3] kann sie dargestellt werden als

$$\mathcal{R}(t_f, x_0) = e^{\mathfrak{A}t_f} \cdot x_0 + \int_0^{t_f} e^{\mathfrak{A}(t_f-\tau)}\mathfrak{B}U\, d\tau\,, \tag{6.2}$$

wobei hier $e^{\mathfrak{A}(t_f-\tau)} = \Phi(t_f, \tau)$ das Fundamentalsystem der zugeordneten homogenen Differentialgleichung ist (vergleiche auch Satz 2.3.5). Diese Darstellung beruht auf der Variation-der-Konstanten-Formel.

Man kann also die erreichbare Menge approximieren, indem man das Aumann-Integral $\int_0^{t_f} e^{A(t_f-\tau)}BU\, d\tau$ durch ein mengenwertiges Quadraturverfahren approximiert. Dieser Zugang wurde von R. Baier in [4, Abschnitt 2.2] untersucht. Dabei ist für die Konvergenzordnung der Quadraturverfahren die Glattheit der Stützfunktion des Integranden $\tau \mapsto \delta^*\left(\mathfrak{l}, e^{A(t_f-\tau)}BU\right)$ entscheidend, und zwar gleichmäßig in allen Richtungen $\mathfrak{l} \in S_{n-1}$. Außerdem muss die Konvergenzordnung der entsprechenden punktwertigen Verfahren berücksichtigt werden. Da es Quadraturverfahren von beliebiger Ordnung gibt kann also nur die (fehlende) Glattheit der Stützfunktion die Konvergenzordnung beschränken. Nach [4, Korollar 1.3.7 + Korollar 1.4.4] gilt für entsprechende Newton-Cotes bzw. Gauß-Quadraturverfahren:

Ist die ν-te Ableitung von $\tau \mapsto \delta^*\left(\mathfrak{l}, e^{A(t_f-\tau)}BU\right)$ gleichmäßig für alle $\mathfrak{l} \in S_{n-1}$ absolutstetig, so kann man Konvergenzordnung $\nu+1$ erzielen. Ist zusätzlich die $(\nu+1)$-te Ableitung von beschränkter L^1-Variation für alle $\mathfrak{l} \in S_{n-1}$, so kann man Konvergenzordnung $\nu+2$ erreichen.

Wenn also im folgenden Abschnitt ein Beispiel als glatt bezeichnet wird, ist gemeint, dass die Stützfunktion des Integranden $\tau \mapsto \delta^*\left(\mathfrak{l}, e^{A(t_f-\tau)}BU\right)$ beliebig oft differenzierbar ist für alle $\mathfrak{l} \in S_{n-1}$. Dagegen ist sie bei einem nichtglatten Beispiel nur gleichmäßig absolutstetig mit einer Ableitung von beschränkter L^1-Variation.

6.2. Beispiele

In allen folgenden zweidimensionalen Beispielen wurden die Stützpunkte an die erreichbare Menge in 200 Richtungen berechnet. Die Richtungen wurde gemäß der Parametrisierung aus 6.1.3.2 gewählt. Für die dreidimensionalen Beispiele wurde ebenfalls diese Parametrisierung verwendet, indem beide Winkelintervalle der Definitionsmenge je 20 mal äquidistant unterteilt wurden, d.h. für die Menge $\mathcal{P}_M$ wurde $M = (20, 20)$ gewählt. Damit ergeben sich 382 Richtungsvektoren. Mit dieser Anzahl und Verteilung der Richtungen können meist optisch gute Approximationen berechnet werden.

Die hier besprochenen Beispiele finden sich in den Artikeln [5], [6] sowie in dem Skript [4]. Sie wurden aber zum Teil abgewandelt.

Im ersten Beispiel, soll exemplarisch nachgewiesen werden, dass es zu einer Konvergenzordnung tatsächlich nur ein Verfahren gibt, wie dies in Satz 3.2.5 theoretisch gezeigt wurde. Damit alle Verfahren aus allen drei Verfahrensklassen für die Berechnung herangezogen werden können, muss die Kontrollmenge U ein (eindimensionales) Intervall sein, denn dies lässt sich auch als eindimensionale Kugel interpretieren.

Beispiel 6.2.1. Gegeben sei das lineare Kontrollproblem

$$\begin{aligned} x'(t) &= \mathfrak{A}x(t) + \mathfrak{B}\mathfrak{u}(t) \quad (\text{für } t \in I := [0,1]) \\ x(0) &= x_0 \\ \mathfrak{u}(t) &\in U \quad \forall t \in I \end{aligned}$$

mit

$$\mathfrak{A} = \begin{pmatrix} 0 & 1 \\ -2 & -3 \end{pmatrix}, \ \mathfrak{B} = \begin{pmatrix} -1 \\ 1 \end{pmatrix},$$

der Steuermenge $U = [-1, 1] \subset \mathbb{R}$ und der Anfangsbedingung $x_0 = (0, 0)^T$.

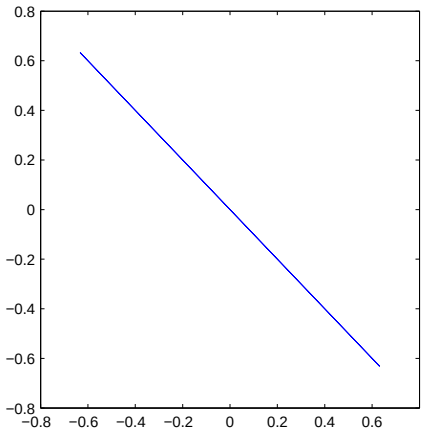

Abbildung 6.1.: Die erreichbare Menge aus Beispiel 6.2.1

Dabei kann man U auch schreiben als Kugel $B_1(0) \subset \mathbb{R}$. Es können also auch die Verfahren der dritten Klasse angewandt werden. In Abbildung 6.1 ist die erreichbare Menge dieses Beispiels dargestellt. Dies ist ein beliebig glattes Beispiel, weswegen die erreichbare Menge auch diese einfache Gestalt hat.

Die folgenden Tabellen enthalten den Hausdorff-Abstand zur Referenzmenge von verschiedenen Approximationen, die durch das Verfahren in der Kopfzeile mit einer Iterationsanzahl N aus der ersten Spalte erzeugt wurden.

Die Verfahren mit analytischer Integration und Quadersteuerbereich

N	verbessertes Euler-Verfahren	Heun-Verfahren	Ferretti-Verfahren der Ordnung 2
2	$3.21670778E-02$	$3.21670778E-02$	$3.21670778E-02$
4	$6.57551136E-03$	$6.57551136E-03$	$6.57551136E-03$
8	$1.49030237E-03$	$1.49030237E-03$	$1.49030237E-03$
16	$3.55105554E-04$	$3.55105554E-04$	$3.55105554E-04$
512	$3.31257474E-07$	$3.31257474E-07$	$3.31257474E-07$
1024	$8.27537118E-08$	$8.27537118E-08$	$8.27537118E-08$
2048	$2.06808552E-08$	$2.06808552E-08$	$2.06808552E-08$

Tabelle 6.1.: Hausdorff-Abstand der Verfahren zweiter Ordnung aus der ersten Klasse zur Referenzmenge in Beispiel 6.2.1

N	Heun3-Verfahren	Ferretti-Verf. der Ordnung 3	klassisches RK-Verfahren	Ferretti-Verf. der Ordnung 4
2	$4.0475924E-03$	$4.0475924E-03$	$4.1210609E-04$	$4.1210609E-04$
4	$4.1391899E-04$	$4.1391899E-04$	$2.0871296E-05$	$2.0871296E-05$
8	$4.6799531E-05$	$4.6799531E-05$	$1.1748586E-06$	$1.1748586E-06$
16	$5.5639652E-06$	$5.5639652E-06$	$6.9694040E-08$	$6.9694040E-08$
32	$6.7830039E-07$	$6.7830039E-07$	$4.2437850E-09$	$4.2437850E-09$
512	$1.6176175E-10$	$1.6176175E-10$	$7.1125188E-14$	$7.1125188E-14$
1024	$2.0204263E-11$	$2.0204263E-11$	$1.8291577E-14$	$1.8291567E-14$

Tabelle 6.2.: Hausdorff-Abstand der Verfahren dritter und vierter Ordnung aus der ersten Klasse zur Referenzmenge in Beispiel 6.2.1

In den Tabellen 6.1 und 6.2 sind die Verfahren mit analytischer Stützpunktberechnung für einen Quadersteuerbereich tabelliert. Man sieht, dass der Hausdorff-Abstand zur

Referenzmenge für die Verfahren gleicher Konvergenzordnung fast immer identisch ist.

Die Verfahren mit numerischer Integration und Quadersteuerbereich

N	verbessertes Euler-Verfahren	Heun-Verfahren	Ferretti-Verfahren der Ordnung 2
2	$3.21670778E-02$	$3.21670778E-02$	$3.21670778E-02$
4	$6.57551136E-03$	$6.57551136E-03$	$6.57551136E-03$
8	$1.49030237E-03$	$1.49030237E-03$	$1.49030237E-03$
16	$3.55105554E-04$	$3.55105554E-04$	$3.55105554E-04$
512	$3.31257474E-07$	$3.31257474E-07$	$3.31257474E-07$
1024	$8.27537114E-08$	$8.27537114E-08$	$8.27537114E-08$
2048	$2.06808548E-08$	$2.06808548E-08$	$2.06808548E-08$

Tabelle 6.3.: Hausdorff-Abstand der Verfahren zweiter Ordnung aus der zweiten Klasse zur Referenzmenge in Beispiel 6.2.1

N	Heun3-Verfahren	Ferretti-Verf. der Ordnung 3	klassisches RK-Verfahren	Ferretti-Verf. der Ordnung 4
2	$9.95549415E-03$	$9.95549415E-03$	$4.02601325E-03$	$4.02601325E-03$
4	$1.72916516E-03$	$1.72916516E-03$	$2.66910536E-04$	$2.66910536E-04$
8	$1.84380366E-04$	$1.84380366E-04$	$1.69638366E-05$	$1.69638366E-05$
16	$2.43256978E-05$	$2.43256978E-05$	$1.06627499E-06$	$1.06627499E-06$
32	$2.73087600E-06$	$2.73087600E-06$	$6.67896000E-08$	$6.67896000E-08$
512	$6.99490478E-10$	$6.99490478E-10$	$1.01365169E-12$	$1.01365169E-12$
1024	$8.96186794E-11$	$8.96186794E-11$	$4.88298755E-14$	$4.88298662E-14$

Tabelle 6.4.: Hausdorff-Abstand der Verfahren dritter und vierter Ordnung aus der zweiten Klasse zur Referenzmenge in Beispiel 6.2.1

In den Tabellen 6.3 und 6.4 sind die Verfahren mit numerischer Stützpunktberechnung für einen Quadersteuerbereich tabelliert (zweite Klasse). Auch hier sieht man, dass der Hausdorff-Abstand zur Referenzmenge für die Verfahren gleicher Konvergenzordnung fast immer identisch ist. Außerdem sind die Werte für die Verfahren 2. Ordnung ebenfalls fast identisch mit den Verfahren 2. Ordnung aus Tabelle 6.1. Dies könnte daran liegen, dass hier zur Stützpunktberechnung mit numerischer Integration die einfache Trapezregel verwendet wurde mit der Schrittweite des mengenwertigen Verfahrens. Es scheint, dass der Fehler der dabei gemacht wurde kleiner ist

als der Verfahrensfehler, weshalb er nicht ins Gewicht fällt. Bei den Verfahren höherer Ordnung musste die iterierte Trapezregel angewandt werden, mit einer zweiten Schrittweite, die kleiner gewählt werden musste, als die des gesamten Verfahrens. Es könnte aber auch an der einfachen Form der erreichbaren Menge liegen (siehe Abbildung 6.1 auf Seite 136).

Die Verfahren mit numerischer Integration und Kugelsteuerbereich

N	verbessertes Euler-Verfahren	Heun-Verfahren	Ferretti-Verfahren der Ordnung 2
2	$3.21670778E-02$	$3.21670778E-02$	$3.21670778E-02$
4	$6.57551136E-03$	$6.57551136E-03$	$6.57551136E-03$
8	$1.49030237E-03$	$1.49030237E-03$	$1.49030237E-03$
16	$3.55105554E-04$	$3.55105554E-04$	$3.55105554E-04$
512	$3.31257474E-07$	$3.31257474E-07$	$3.31257474E-07$
1024	$8.27537118E-08$	$8.27537118E-08$	$8.27537118E-08$
2048	$2.06808552E-08$	$2.06808552E-08$	$2.06808552E-08$

Tabelle 6.5.: Hausdorff-Abstand der Verfahren zweiter Ordnung aus der dritten Klasse zur Referenzmenge in Beispiel 6.2.1

N	Heun3-Verfahren	Ferretti-Verf. der Ordnung 3	klassisches RK-Verfahren	Ferretti-Verf. der Ordnung 4
2	$9.95549415E-03$	$9.95549415E-03$	$4.02601325E-03$	$4.02601325E-03$
4	$1.72916516E-03$	$1.72916516E-03$	$2.66910536E-04$	$2.66910536E-04$
8	$1.84380366E-04$	$1.84380366E-04$	$1.69638366E-05$	$1.69638366E-05$
16	$2.43256978E-05$	$2.43256978E-05$	$1.06627499E-06$	$1.06627499E-06$
32	$2.73087600E-06$	$2.73087600E-06$	$6.67895996E-08$	$6.67895996E-08$
512	$6.99490321E-10$	$6.99490321E-10$	$1.01325917E-12$	$1.01325917E-12$
1024	$8.96182869E-11$	$8.96182869E-11$	$4.82803431E-14$	$4.84373524E-14$

Tabelle 6.6.: Hausdorff-Abstand der Verfahren dritter und vierter Ordnung aus der dritten Klasse zur Referenzmenge in Beispiel 6.2.1

Und in den Tabellen 6.5 und 6.6 sind schließlich die Verfahren mit numerischer Stützpunktberechnung für einen Kugelsteuerbereich, d.h. die Verfahren der dritten Klasse, dargestellt. Wieder kann man sehen, dass die Werte für Verfahren der gleichen Ordnung fast immer identisch sind. Der größere Unterschied bei den Verfahren vierter Ordnung in der letzten Zeile, der aber schon im Bereich der Maschinengenauigkeit

liegt, ist wohl auf einen Rundungsfehler zurückzuführen. Auch kann man wieder sehen, dass die Werte für die Verfahren der zweiten Ordnung mit denen der Verfahren zweiter Ordnung aus den anderen Klassen sehr gut übereinstimmen. Dies hat wohl auch hier den gleichen Grund wie vorher.

In allen diesen Tabellen 6.1 bis 6.6 ist deutlich ersichtlich, dass alle Verfahren konvergieren. Außerdem kann man sehen, dass die Verfahren höherer Ordnung einen kleineren Fehler (Hausdorff-Abstand zur Referenzmenge) haben als Verfahren geringerer Ordnung der gleichen Klasse und bei gleicher Iterationsanzahl.

Aufgrund der Ergebnisse von Beispiel 6.2.1 und Satz 3.2.5 werden wir in den folgenden Beispielen nur noch die Ferretti-Verfahren aller Klassen betrachten. In dem folgenden Beispiel sollen die Konvergenzordnungen der Verfahren mit Quadersteuerbereich exemplarisch nachgewiesen werden.

Beispiel 6.2.2. Gegeben sei das lineare Kontrollproblem

$$\begin{aligned} x'(t) &= \mathfrak{A}x(t) + \mathfrak{B}\mathfrak{u}(t) \quad (\text{für } t \in I := [0,1]) \\ x(0) &= x_0 \\ \mathfrak{u}(t) &\in U \quad \forall t \in I \end{aligned}$$

mit

$$\mathfrak{A} = \begin{pmatrix} 0 & 1 \\ -2 & -3 \end{pmatrix}, \ \mathfrak{B} = \begin{pmatrix} -1 & 1 \\ 1 & -2 \end{pmatrix},$$

der Steuermenge $U = [-1,1]^2 \subset \mathbb{R}^2$ und der Anfangsbedingung $x_0 = (0,0)^T$. Dieses Beispiel ist eine Abwandlung des vorherigen. Aber diesmal ist der Steuerbereich ein echter Quader. Hier erhält man als Fundamentalmatrix dieses Problems

$$\Phi(t,\tau) = e^{A(t-\tau)} = \begin{pmatrix} 2e^{-t-\tau} - e^{-2t-2\tau} & e^{-t-\tau} - e^{-2t-2\tau} \\ 2e^{-t-\tau} + 2e^{-2t-2\tau} & -e^{-t-\tau} + 2e^{-2t-2\tau} \end{pmatrix}$$

und damit ergibt sich nach (6.2) als Darstellung der erreichbaren Menge das Aumann-Integral

$$\mathcal{R}\left(1,(0,0)^T\right) = \int_0^1 \begin{pmatrix} -e^{-1-\tau} & e^{-2-2\tau} \\ e^{-1-\tau} & -2e^{-2-2\tau} \end{pmatrix} [-1,1]^2 \, d\tau \,.$$

Dann ist nach Beispiel 1.3.6(ii) die Stützfunktion des Integranden für $\mathfrak{l} = (l_1,l_2)^T \in S_1$

$$\begin{aligned} \delta^*\left(\mathfrak{l}, \begin{pmatrix} -e^{-1-\tau} & e^{-2-2\tau} \\ e^{-1-\tau} & -2e^{-2-2\tau} \end{pmatrix} [-1,1]^2\right) &= \delta^*\left(\begin{pmatrix} -e^{-1-\tau} & e^{-1-\tau} \\ e^{-2-2\tau} & -2e^{-2-2\tau} \end{pmatrix} \mathfrak{l}, [-1,1]^2\right) \\ &= \left\| \begin{pmatrix} -e^{-1-\tau} & e^{-1-\tau} \\ e^{-2-2\tau} & -2e^{-2-2\tau} \end{pmatrix} \cdot \mathfrak{l} \right\|_1 \\ &= |l_2 - l_1| \, e^{-1-\tau} + |l_1 - 2l_2| \, e^{-2-2\tau} \end{aligned}$$

als Funktion von τ beliebig oft differenzierbar, und zwar gleichmäßig in t. Nach 6.1.3.4 handelt es sich also um ein glattes Beispiel. Das obige Aumann-Integral, und damit die erreichbare Menge, kann mit mengenwertigen Quadraturverfahren von beliebiger Ordnung approximiert werden.

Die in dieser Arbeit entwickelten Ferretti-Verfahren vermögen auch die erreichbare Menge von beliebiger Ordnung zu approximieren. Die folgenden Tabellen enthalten die geschätzte Konvergenzordnung (siehe 6.1.3.3) für die Ferretti-Verfahren mit Quadersteuerbereich, und zwar mit analytischer und numerischer Stützpunktberechnung.

	Ferretti-Verfahren der			
N	Ordnung 2	Ordnung 3	Ordnung 4	Ordnung 8
2	−1.30138147	4.20423998	3.83904365	9.31228905
4	2.63458358	3.49720577	4.60197148	8.65193644
8	2.27717346	3.26345978	4.29156059	8.32472713
16	2.12627179	3.13270146	4.14439220	8.14584987
32	2.06027541	3.06623315	4.07190431	———
64	2.02945635	3.03304003	4.03588080	———
128	2.01456231	3.01649996	4.01789246	———
256	2.00724026	3.00824432	4.00817888	———
512	2.00360998	3.00412255	3.97408291	———

Tabelle 6.7.: Konvergenzordnungen der Ferretti-Verfahren mit analytischer Integration in Beispiel 6.2.2

In Tabelle 6.7 sind die geschätzten Konvergenzordnungen für die Ferretti-Verfahren aus der ersten Klasse, d.h. mit analytischer Integration, aufgezeichnet. Eine negative Konvergenzordnung bedeutet, dass sich das Verfahren im Vergleich zur vorigen Schrittweite verschlechtert hat. In diesem Beispiel hat das Ferretti-Verfahren zufällig eine sehr gute Startnäherung für $N = 1$ produziert, die einen geringeren Hausdorff-Abstand zur Referenzmenge hat als die Lösung mit $N = 2$. Bei dem Ferretti-Verfahren der Ordnung 8 konnten ab $N = 32$ keine Schätzungen mehr gemacht werden, da hier der Hausdorff-Abstand der Lösungen zur Referenzmenge schon im Bereich der Maschinengenauigkeit ist. Man kann sehen, dass bei allen Verfahren die geschätzte Konvergenzordnung sehr gut mit der theoretisch erwarteten Konvergenzordnung übereinstimmt.

Einige Approximationen der erreichbare Menge für dieses Beispiel sind in Abbildung 6.2 dargestellt. Sie wurden mit dem Ferretti-Verfahren der Ordnung 2 mit analytischer (links) und numerischer (rechts) Integration berechnet. Es wurden die Näherungen jeweils für $N = 1, 2$ und 4 errechnet. In beiden Bildern ist die Approximation für $N = 4$ die bestmögliche im Rahmen der Zeichengenauigkeit und vertritt damit die exakte

erreichbare Menge. Man kann in den Bildern die Konvergenz der Verfahren erkennen und im linken Bild sieht man sogar die Konvergenzordnung visuell bestätigt. Alle Bilder sind wirklich mit 200 Richtungen errechnet worden. Die erreichbare Menge hat tatsächlich diese einfache Gestalt.

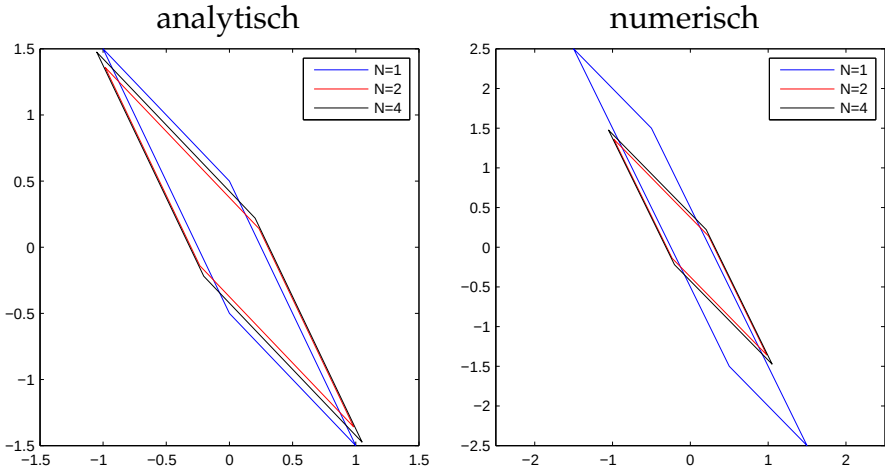

Abbildung 6.2.: Erreichbare Menge aus Beispiel 6.2.2 mit analytischer (links) und numerischer (rechts) Integration

	Ferretti-Verfahren der		
N	Ordnung 2	Ordnung 3	Ordnung 4
2	2.78175135	4.40010597	5.35119929
4	2.63458358	2.70517018	3.52222311
8	2.27717346	3.30582752	3.89212545
16	2.12627179	2.96675338	3.96600682
32	2.06027541	3.17400158	3.98715693
64	2.02945635	2.89594876	3.99453516
128	2.01456231	3.13953204	3.99750099
256	2.00724026	2.87801774	3.99891643
512	2.00360998	3.03877803	4.00317353

Tabelle 6.8.: Konvergenzordnungen der Ferretti-Verfahren mit numerischer Integration in Beispiel 6.2.2

In Tabelle 6.8 sieht man die geschätzten Konvergenzordnungen für die Ferretti-Verfahren aus der zweiten Klasse, d.h. mit numerischer Integration. Auch hier kann man sehen, dass bei allen Verfahren die geschätzte Konvergenzordnung sehr gut mit der

theoretisch erwarteten Konvergenzordnung übereinstimmt, auch wenn diese Werte schon mehr Unregelmäßigkeiten aufweisen als bei den Verfahren mit analytischer Integration. Dies liegt natürlich daran, dass durch die numerische Integration noch ein Fehler zum Verfahrensfehler hinzukommt, im Gegensatz zu den Verfahren mit analytischer Integration.
In Bemerkung 3.2.6 wurde erwähnt, dass die Ferretti-Verfahren keiner Ordnungsbeschränkung unterliegen. Dies war ja Motivation zur Entwicklung dieser Verfahren. Denn bei den mengenwertigen RK-Verfahren lässt sich nur Konvergenzordnung 4 erreichen (siehe Bemerkung 3.1.4). Durch die Ergebnisse für das Ferretti-Verfahren der Ordnung 8 ist dieser Sachverhalt hier praktisch bestätigt worden. Ein Ferretti-Verfahren der Ordnung 8 (oder zumindest > 4) mit numerischer Integration wurde nicht implementiert, da hier der Aufwand für die Integration schon so groß wird, dass es sich nicht mehr lohnt.

In dem nächsten Beispiel soll wieder vor allem die hohe Konvergenzordnung der Verfahren wieder für einen Quadersteuerbereich gezeigt werden. Aber diesmal handelt es sich um ein nichtglattes Beispiel.

Beispiel 6.2.3. Gegeben sei das lineare Kontrollproblem

$$\begin{aligned} x'(t) &= \mathfrak{A}x(t) + \mathfrak{B}\mathfrak{u}(t) \quad (\text{für } t \in I := [0, 2\pi]) \\ x(0) &= x_0 \\ \mathfrak{u}(t) &\in U \quad \forall t \in I \end{aligned}$$

mit

$$\mathfrak{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \ \mathfrak{B} = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

der Steuermenge $U = [-1, 1] \subset \mathbb{R}$ und der Anfangsbedingung $x_0 = (0, 0)^T$. Hier erhält man als Fundamentalmatrix dieses Problems

$$\Phi(t, \tau) = e^{A(t-\tau)} = \begin{pmatrix} \cos(t-\tau) & \sin(t-\tau) \\ -\sin(t-\tau) & \cos(t-\tau) \end{pmatrix}$$

und damit ergibt sich nach (6.2) als Darstellung der erreichbaren Menge das Aumann-Integral

$$\mathcal{R}\left(1, (0,0)^T\right) = \int_0^{2\pi} \begin{pmatrix} \sin(2\pi - \tau) \\ \cos(2\pi - \tau) \end{pmatrix} [-1, 1] \, d\tau \,.$$

Dann sieht nach Beispiel 1.3.6(ii) die Stützfunktion des Integranden für $\mathfrak{l} = (l_1, l_2)^T \in S_1$ so aus:

$$\begin{aligned} \delta^*\left(\mathfrak{l}, \begin{pmatrix} \sin(2\pi - \tau) \\ \cos(2\pi - \tau) \end{pmatrix} [-1, 1]\right) &= \delta^*\left(\left(\sin(2\pi - \tau), \cos(2\pi - \tau)\right)\mathfrak{l}, [-1, 1]\right) \\ &= |l_1 \sin(2\pi - \tau)| + |l_2 \cos(2\pi - \tau)| \end{aligned}$$

Diese Stützfunktion ist als Funktion von τ nach [4, Satz 1.6.13. (viii)] absolutstetig mit einer Ableitung von beschränkter L^1-Variation, und zwar gleichmäßig in $\mathfrak{l} \in S_1$. Also kann man, wie in 6.1.3.4 erwähnt, das obige Aumann-Integral, und damit die erreichbare Menge, mit mengenwertigen Quadraturverfahren nur von Ordnung 2 approximieren. Es handelt sich also um ein nichtglattes Beispiel. Dies liegt daran, dass $\mathfrak{B}$ keine Eigenvektoren der Matrix $\mathfrak{A}$ enthält wie in Beispiel 6.2.2.

Nun soll die erreichbare Menge dieses Kontrollproblems mit den Ferretti-Verfahren approximiert werden. Die folgenden Tabellen enthalten die geschätzte Konvergenzordnung (siehe 6.1.3.3) und den Hausdorff-Abstand zur Referenzmenge für die Approximationen, die von den Ferretti-Verfahren mit analytischer Integration und Quadersteuerbereich erzeugt wurden.

	Ferretti-Verfahren der			
N	Ordnung 2	Ordnung 3	Ordnung 4	Ordnung 8
2	−0.85446795	0.52855228	4.44235539	10.4439927
4	1.60624327	5.63304309	6.60617172	17.2867169
8	2.81627694	1.44311564	3.88703956	8.03687211
16	2.48261402	2.64624130	4.32132648	8.33424273
32	2.28173964	2.88656141	4.27455397	8.29501977
64	2.16665391	2.95501523	4.17757593	———
128	2.09295714	2.97971886	4.10184044	———
256	2.04959982	2.99085636	4.05475070	———
512	2.02565329	2.99562713	4.02872591	———

Tabelle 6.9.: Konvergenzordnungen der Ferretti-Verfahren mit analytischer Integration in Beispiel 6.2.3

In Tabelle 6.9 sind die geschätzten Konvergenzordnungen für die Ferretti-Verfahren aus der ersten Klasse, d.h. mit analytischer Integration, eingetragen. Man sieht das diese Werte ganz gut mit den theoretischen Konvergenzordnungen übereinstimmen, obwohl dieses Beispiel nicht mehr glatt ist. Im Vergleich zu Beispiel 6.2.2 sieht man aber, dass die Werte schon erheblich mehr Abweichungen aufweisen.

Trotz erwarteter Konvergenz müssen auch die Ferretti-Verfahren einen Tribut an die nichtglatte Natur dieses Beispiels zahlen. Dies kann man in Tabelle 6.10, in der der zugehörige Hausdorff-Abstand zur Referenzmenge für diese Verfahren aufgelistet ist, sehen. Es fällt auf, dass die Startwerte ziemlich schlecht sind. Auch ist die erreichte Genauigkeit bei gleicher Iterationszahl und gleicher Verfahrensklasse wesentlich schlechter, als bei einem glatten Kontrollproblem, wie man im Vergleich zu Beispiel 6.2.1 sehen kann.

Für die Ferretti-Verfahren mit numerischer Integration gilt im wesentlichen dasselbe. Nur kann man dort noch mehr Abweichungen der geschätzten Konvergenzordnung

von der theoretischen finden.

N	Ferretti-Verfahren der Ordnung 2	Ordnung 3	Ordnung 4	Ordnung 8
1	$1.6714714E+01$	$3.7229855E+01$	$1.2379365E+03$	$3.0122532E+05$
2	$3.0221738E+01$	$2.5809598E+01$	$5.6939829E+01$	$2.1624017E+02$
4	$9.9264060E+00$	$5.2007540E-01$	$5.8446742E-01$	$1.3524340E-03$
8	$1.4093154E+00$	$1.9126928E-01$	$3.9504340E-02$	$5.1496355E-06$
16	$2.5215461E-01$	$3.0552532E-02$	$1.9760409E-03$	$1.5955831E-08$
32	$5.1855502E-02$	$4.1314798E-03$	$1.0210028E-04$	$5.0800697E-11$
1024	$3.9788950E-05$	$1.3338822E-07$	$7.4674933E-11$	———
2048	$9.9019620E-06$	$1.6696355E-08$	$4.2939599E-12$	———

Tabelle 6.10.: Hausdorff-Abstand zur Referenzmenge für die Ferretti-Verfahren mit analytischer Integration

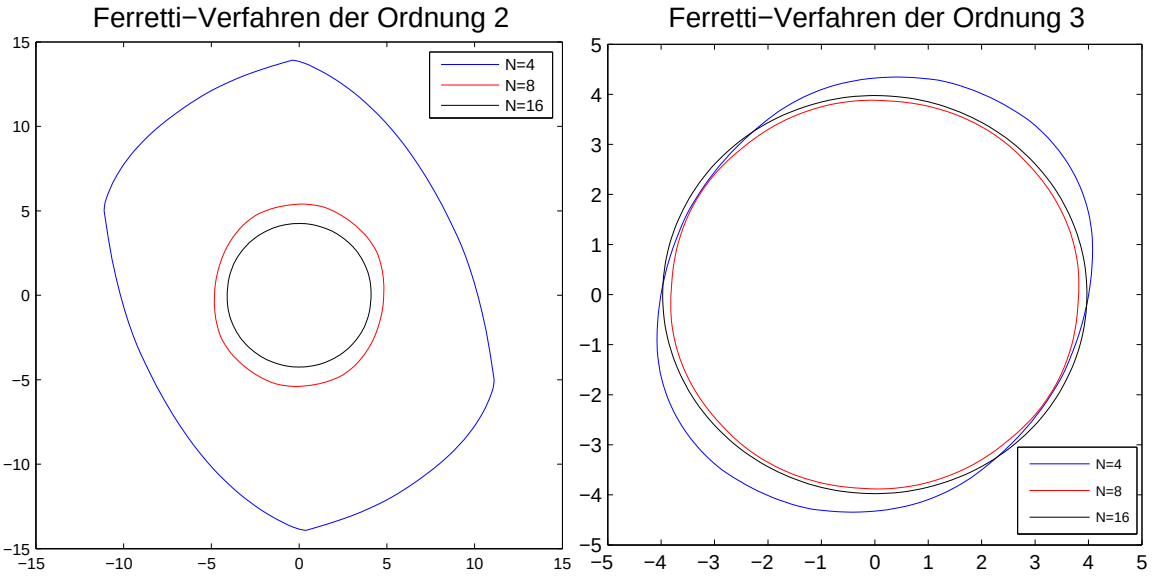

Abbildung 6.3.: Die erreichbare Menge aus Beispiel 6.2.3 approximiert mit den Ferretti-Verfahren der Ordnung 2 und 3 mit analytischer Integration

Die Ferretti-Verfahren der Ordnung 2 (links) und 3 (oben) liefern die Näherungen der erreichbaren Menge dieses Beispiels in Abbildung 6.3 für die Iterationszahlen $4, 8$ und 16. Dabei sind die Approximationen für $N = 16$ die bestmöglichen im Rahmen der Zeichengenauigkeit. Mit geschultem Auge kann man in diesen Abbildungen die

Konvergenzordnung der Verfahren erkennen. In Abbildung 6.4 sind Approximationen der erreichbaren Menge, errechnet mit dem Ferretti-Verfahren der Ordnung 4 für die Iterationszahlen 4 und 8, dargestellt. Die Näherung für $N = 16$ würde optisch mit der für $N = 4$ zusammenfallen.

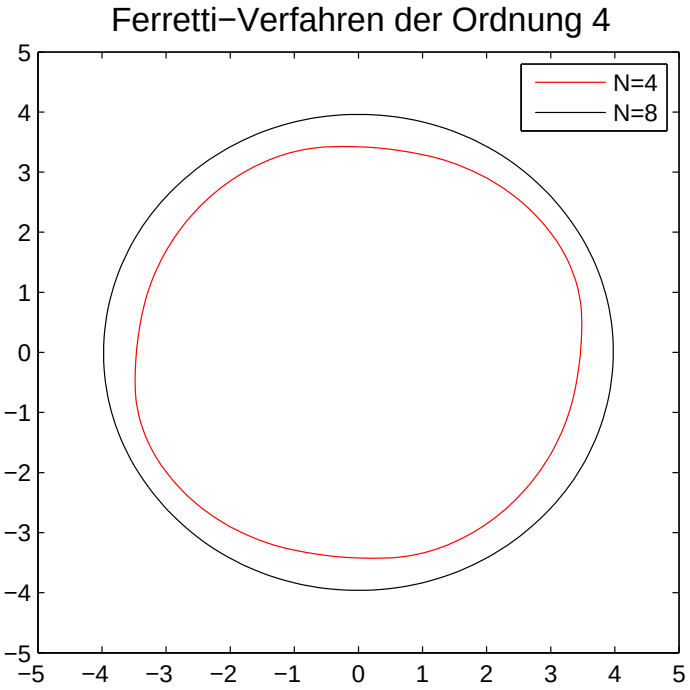

Abbildung 6.4.: Die erreichbare Menge aus Beispiel 6.2.3 approximiert mit dem Ferretti-Verfahren der Ordnung 4 mit analytischer Integration

Wie theoretisch vorhergesagt, konvergieren die Verfahren, obwohl dieses Beispiel nicht glatt ist. Dabei benötigen sie bei gleich feiner Diskretisierung ungefähr die gleiche Rechenzeit wie bei einem glatten Beispiel. Nimmt man aber die Genauigkeit als Maßstab, so benötigen sie um einen Faktor von 2 bis 3 mehr Rechenzeit als bei einem glatten Kontrollproblem. Der Lösungsansatz aus 6.1.3.4 der auf der Approximation der erreichbaren Menge mit mengenwertigen Quadraturverfahren beruht, lässt hier nur Konvergenzordnung 2 zu.

Die Verfahren für einen Quadersteuerbereich mit numerischen Integration wurden implementiert, um sie mit den Verfahren mit analytischer Integration vergleichen zu können. In dem folgenden längeren Beispiel werden die Rechenzeiten der Verfahren dieser beiden Klassen verglichen.

Beispiel 6.2.4. Gegeben sei das lineare Kontrollproblem

$$\begin{aligned} x'(t) &= \mathfrak{A}x(t) + \mathfrak{B}\mathfrak{u}(t) \quad (\text{für } t \in I := [0,1]) \\ x(0) &= x_0 \\ \mathfrak{u}(t) &\in U \quad \forall t \in I \end{aligned}$$

mit

$$\mathfrak{A} = \begin{pmatrix} 1 & 0 \\ 0 & -2 \end{pmatrix}, \ \mathfrak{B} = \begin{pmatrix} 1 \\ -3 \end{pmatrix},$$

der Steuermenge $U = [-1,1] \subset \mathbb{R}$ und der Anfangsbedingung $x_0 = (0,0)^T$. Es handelt sich hierbei um ein nichtglattes Beispiel. Da $\mathfrak{A}$ hier eine Diagonalmatrix ist, konnte die Referenzmenge hier exakt (bis auf Rundungsfehler) berechnet werden, mit Hilfe der Darstellung der erreichbaren Menge in (6.2) als Aumann-Integral. Deswegen wurde dieses Beispiel ausgewählt. Dies ermöglicht es auch für das Ferretti-Verfahren der Ordnung 8, mit dem sonst oft die Referenzmengen berechnet wurden, die minimale Iterationszahl zum Erreichen einer vorgegebenen Genauigkeit genau zu ermitteln.

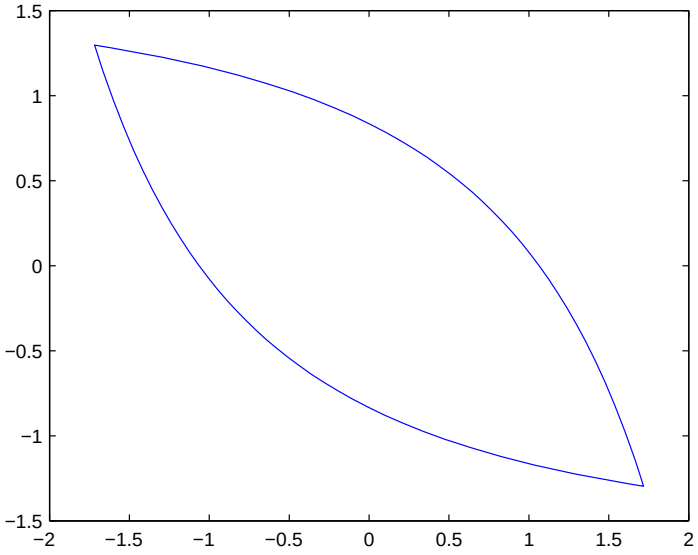

Abbildung 6.5.: Die erreichbare Menge von Beispiel 6.2.4

In Abbildung 6.5 ist die erreichbare Menge dieses Beispiels zu sehen.
Es folgen nun verschiedene Tabellen, in denen alle Ferretti-Verfahren mit Quadersteuerbereich hinsichtlich ihrer Rechenzeit verglichen wurden. Dabei wurde auf absolute Rechenzeiten verzichtet. Stattdessen sind die Einträge der Tabellen so genannte Zeitfaktoren. Dabei ist ein *Zeitfaktor* (=Zeitfak.) folgender Quotient:

$$\frac{\text{Zeit des jeweiligen Verfahrens für die vorgegebene Genauigkeit}}{\text{Zeit des Referenzverfahrens für die vorgegebene Genauigkeit}}$$

Er ist ein Maß dafür, um wieviel schneller bzw. langsamer das jeweilige Verfahren im Vergleich zum Referenzverfahren ist.

analytische Stützpunktberechnung

	Ordnung 2		Ordnung 3		Ordnung 4		Ordnung 8	
Genauigkeit	N	Zeitfak.	N	Zeitfak.	N	Zeitfak.	N	Zeitfak.
$\leq 1E-1$	3	0.49523	2	0.45714	2	0.67619	1	1.0
$\leq 1E-3$	23	1.98401	7	0.80669	4	0.76308	2	1.0
$\leq 1E-5$	230	19.244	28	2.9942	10	1.8023	2	1.0
$\leq 1E-7$	2298	109.26	127	7.5082	31	3.0164	4	1.0
$\leq 1E-9$	22951	827.93	584	24.138	96	6.4942	6	1.0
$\leq 1E-11$	—	—	2705	68.264	303	12.431	10	1.0
$\leq 1E-13$	—	—	—	—	973	24.174	17	1.0

Tabelle 6.11.: Relativer Zeitvergleich für die Ferretti-Verfahren mit analytischer Integration in Beispiel 6.2.4

Für die Ferretti-Verfahren mit analytischer Stützpunktberechnung wurden in Tabelle 6.11 die minimale Iterationsanzahl zu vorgegebener Genauigkeit und der dafür benötigte Zeitfaktor notiert. Das Referenzverfahren für den Zeitfaktor in dieser Tabelle ist das Ferretti-Verfahren der Ordnung 8 mit analytischer Integration. Es wird sehr deutlich, dass sich das Ferretti-Verfahren der Ordnung 8 mit zunehmender Genauigkeit immer mehr gegenüber den Verfahren geringerer Ordnung durchsetzt, sowohl im Bezug auf die benötigten Teilintervalle als auch bezüglich der benötigten Rechenzeit. Allerdings für geringere Genauigkeiten sind die anderen Verfahren schneller, da bei diesen Verfahren die Nullstellenberechnung für die analytische Integration mit Hilfe von Lösungsformeln erfolgt. Während dies bei dem Verfahren der Ordnung 8 numerisch geschehen muss, was natürlich aufwendiger ist (siehe Unterabschnitt 5.4.1). Auch kann mit diesem Verfahren eine höhere Genauigkeit erreicht werden, bevor das Verfahren in einen Bereich kommt, in dem Rundungsfehler das Ergebnis stark verfälschen. Um einen Eindruck für die Größenordnung der Rechenzeiten zu bekommen, soll hier angeführt werden, dass das Verfahren der Ordnung 8 für die beste Approximation 0.023 Sekunden benötigt hat.

numerische Stützpunktberechnung

Genauigkeit	Ordnung 2 N	Ordnung 2 Zeitfak.	Ordnung 3 N	Ordnung 3 Zeitfak.	Ordnung 4 N	Ordnung 4 Zeitfak.
$\leq 1E-1$	3	0.33333	2	0.45714	2	0.39048
$\leq 1E-3$	23	1.1628	10	1.0174	6	0.72674
$\leq 1E-5$	230	11.046	46	4.8837	17	2.6744
$\leq 1E-7$	2298	63.279	215	16.229	54	8.8524
$\leq 1E-9$	22974	514.37	1005	75.287	169	49.195
$\leq 1E-11$	—	—	4713	360.56	532	272.71
$\leq 1E-13$	—	—	—	—	1692	1724.0

Tabelle 6.12.: Relativer Zeitvergleich für die Ferretti-Verfahren mit numerischer Integration in Beispiel 6.2.4

Die Einträge in Tabelle 6.12 haben die gleiche Bedeutung wie in der vorigen Tabelle 6.11, nur wurden hier die Stützpunktberechnung numerisch durchgeführt. Auch hier ist das Referenzverfahren für den Zeitfaktor das Ferretti-Verfahren der Ordnung 8 mit analytischer Integration (!). Gegenüber den Ferretti-Verfahren der zweiten Klasse erweist sich das Referenzverfahren ebenfalls als überlegen im Bezug auf die Rechenzeit und die Anzahl der Teilintervalle. Dies gilt für sinkende Fehlerschranken in zunehmendem Maß besonders gegenüber den Verfahren der Ordnung 3 und 4.
Im Vergleich der Ergebnisse aus den Tabellen 6.11 und 6.12 fällt auf, dass die Verfahren der Ordnung 3 und 4 mit numerischer Stützpunktberechnung bei gleicher Genauigkeit mehr Teilintervalle benötigen als ihre direkten Konkurrenten. Dies liegt an dem approximativen Charakter der Stützpunktberechnung dieser Verfahren. Die beiden Verfahren der Ordnung 2 gleichen sich bezüglich der benötigten Teilintervalle.

Ferretti-Verfahren der Ordnung	Anzahl der Iterationen N: 1	4	16	64	256	1024	4096
2 analytisch	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2 numerisch	0.7447	0.6466	0.5937	0.5714	0.5664	0.5700	0.5955
3 analytisch	1.2553	1.2782	1.25	1.2275	1.2349	1.2428	1.3079
3 numerisch	1.0	1.0977	1.125	1.3016	1.6040	2.2867	3.5646
4 analytisch	1.8723	1.9699	2.0208	2.0053	2.0201	1.9545	—
4 numerisch	1.0638	1.1955	1.7187	3.8413	12.906	48.628	—
8 analytisch	4.3061	4.5940	4.5625	4.2381	—	—	—

Tabelle 6.13.: Absoluter Zeitvergleich aller Ferretti-Verfahren mit Quadersteuerbereich in Beispiel 6.2.4

Abschließend enthält Tabelle 6.13 einen direkten Zeitvergleich der Ferretti-Verfahren mit Quadersteuerbereich. Das Referenzverfahren für die Zeitfaktoren ist diesmal das analytischen Ferretti-Verfahren der Ordnung 2. Dabei ist diesmal die Iterationszahl in jeder Spalten konstant. Bei den analytischen Verfahren und bei dem numerischen Verfahren der Ordnung 2 bleibt der Zeitfaktor für wachsende Iterationszahlen in etwa gleich. Jedoch bei dem numerischen Verfahren der Ordnung 3 und noch mehr bei dem der Ordnung 4 verschlechtert sich dieser Faktor mit der Verfeinerung der Diskretisierung. Auch im direkten Vergleich mit den analytischen Verfahren der gleichen Ordnung schneiden diese beiden Verfahren mit zunehmender Iterationszahl immer schlechter ab. Besonders deutlich wird dies bei dem Ferretti-Verfahren der Ordnung 4. Dies liegt an der numerischen Integration mit der iterierten Trapezregel, die diese Verfahren benutzen. Dabei muss jedes Teilintervall erneut äquidistant unterteilt werden. Für das Verfahren der Ordnung 3 werden $\mathcal{O}\left(\sqrt{N}\right)$ Unterteilungen benötigt und für das der Ordnung 4 werden $\mathcal{O}(N)$ benötigt. Im Bezug auf die numerische Realisation gilt für diese Verfahren mit Quadraturintegration das gleiche, was für die Verfahren mit Kugelsteuermenge in Unterabschnitt 5.4.3 gesagt wurde. Das numerische Verfahren der Ordnung 2 benutzt zur Integration die einfache Trapezregel. Es scheint das dies im Vergleich zur iterierten Trapezregel nicht nur Vorteile bezüglich der Rechenzeit sondern auch bezüglich der Genauigkeit bringt. Denn bei diesem Verfahren ist der Integrationsfehler offenbar geringer als der Verfahrensfehler.

In seinem Artikel [13] hat Ferretti als Steuerbereich U stets das Intervall $[0,1]$ vorausgesetzt. Denn dies ermöglicht die Beschreibung der Momentenmenge $\mathcal{M}_{m,p}$ mit dem Hausdorff-Momentenproblem, wie dies in Abschnitt 5.3 geschehen ist. In diesem Abschnitt wurde auch gezeigt, dass man mit diesem Zugang unter Verwendung gewisser Tricks auch beliebige achsenparallele Quader behandeln kann. Jedoch mit dem in dieser Arbeit erwähnten Zugang über die Beschreibung von $\mathcal{M}_{m,p}$ als Aumann-Integral lassen sich Steuermengen von vielfältiger geometrischer Form behandeln, nur abhängig davon, ob und wie gut sich Stützpunkte für diese Menge berechnen lassen (in Abschnitt 5.4 wurde U nur als kompakt, konvex und nichtleer vorausgesetzt). In der vorliegenden Arbeit wurden Verfahren für Kugelsteuermengen implementiert. Das folgende Beispiel soll die Konvergenz der Ferretti-Verfahren mit einer echten Kugel als Steuermenge zeigen. Außerdem ist dies ein dreidimensionales Beispiel und zeigt, dass die implementierten Verfahren auch für höhere Dimensionen funktionieren.

Beispiel 6.2.5. Gegeben sei das lineare Kontrollproblem

$$\begin{aligned} x'(t) &= \mathfrak{A}x(t) + \mathfrak{B}\mathfrak{u}(t) \quad (\text{für } t \in I := [0, 1.7775]) \\ x(0) &= x_0 \\ \mathfrak{u}(t) &\in U \quad \forall t \in I \end{aligned}$$

mit

$$\mathfrak{A} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathfrak{B} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

der Steuermenge $U = B_1(0) \subset \mathbb{R}^3$ und der Anfangsbedingung $x_0 = (2, 0, 0)^T$. Es handelt sich um ein glattes Beispiel (vgl. [4, Beispiel 3.3.5.]).

In Tabelle 6.14 ist die geschätzte Konvergenzordnung der Ferretti-Verfahren aus der dritten Klasse bezüglich der Referenzmenge aufgezeichnet. Trotz einiger Schwankungen, die für die Verfahren mit numerischer Integration normal sind, wird die theoretische Konvergenzordnung bestätigt. Die letzten Näherungen mit 512 Iterationen haben in der Reihenfolge der aufsteigenden Konvergenzordnung einen approximierten Hausdorff-Abstand von $2.5948 \cdot 10^{-6}$, $6.2555 \cdot 10^{-9}$ und $2.1661 \cdot 10^{-11}$ zur Referenzmenge .

Was das Laufzeitverhalten dieser Verfahren angeht, so lässt sich qualitativ dasselbe aussagen wie für die Verfahren mit numerischer Integration für eine Quadermenge.

	Ferretti-Verfahren der		
N	Ordnung 2	Ordnung 3	Ordnung 4
2	2.16775853	3.98991432	3.98991432
4	2.06052708	2.00005916	3.16999393
8	1.99920859	3.16994049	3.47393812
16	1.99637013	2.83007662	4.00000063
32	1.99727530	2.64385641	3.85199886
64	1.99524551	2.97085368	3.92305142
128	1.99677308	2.72514011	3.96073746
256	1.99837443	3.06102889	3.98010263
512	1.99918421	2.77404219	3.98904359

Tabelle 6.14.: Konvergenzordnungen der Ferretti-Verfahren mit numerischer Integration für Beispiel 6.2.5

In den Abbildungen 6.6 und 6.7 ist die erreichbare Menge dieses Beispiels dargestellt. Damit man sich diese Menge leichter vorstellen kann, wurden 8 Bilder erstellt, die alle aus einer andere Blickrichtung gemacht wurden. Geht man die Bilder zeilenweise von links nach rechts durch (wie man liest), so wird vor allem die xy-Ebene um die geneigte z-Achse im Uhrzeigersinn gedreht. Dabei werden die Kanten des Gitters mit großem z-Wert mit warmen Farben gezeichnet (Orange, Rot), während Kanten mit kleinem z-Wert in kalten Farben (Türkis, Blau) gezeichnet werden.

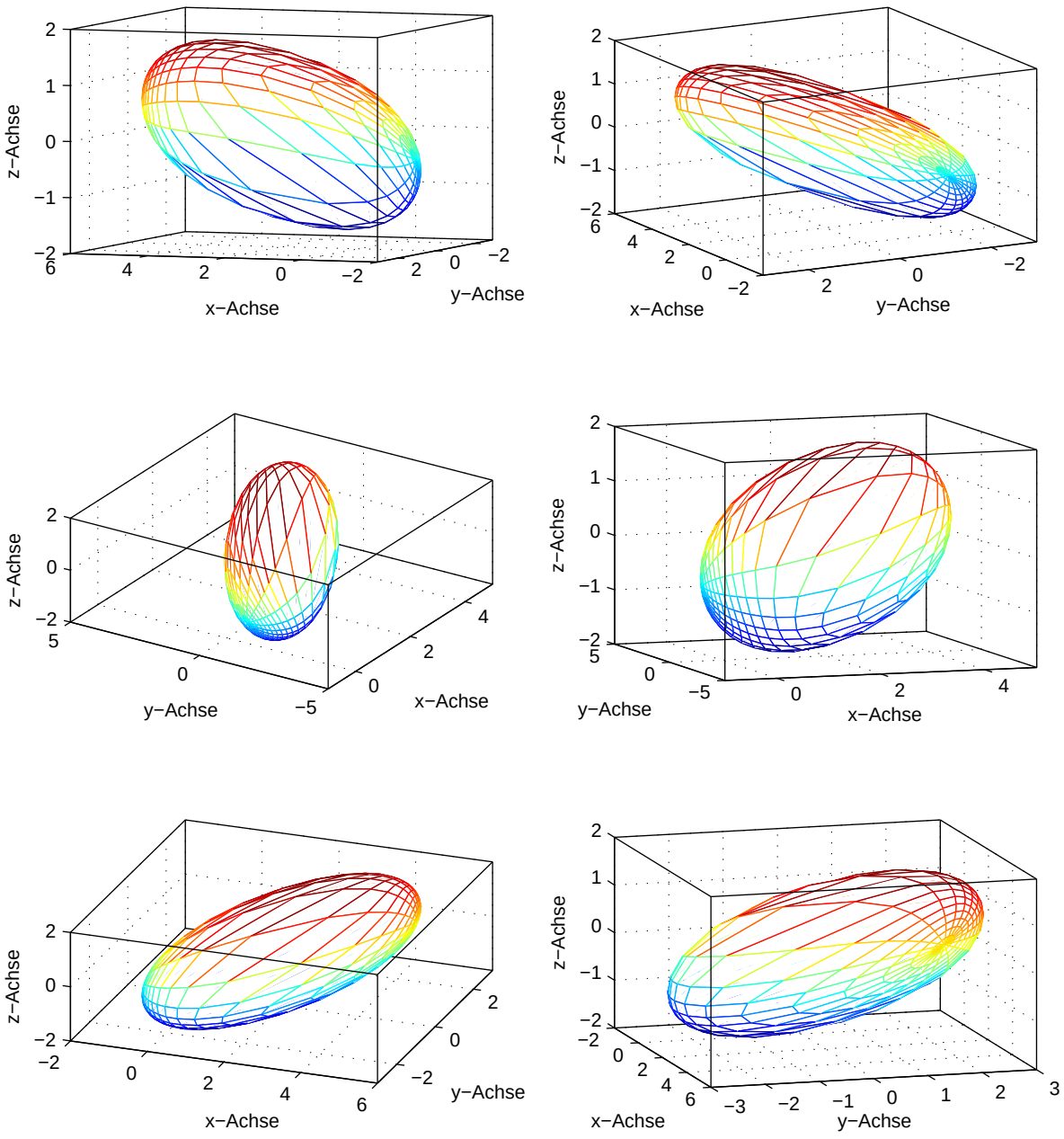

Abbildung 6.6.: Die erreichbare Menge aus Beispiel 6.2.5

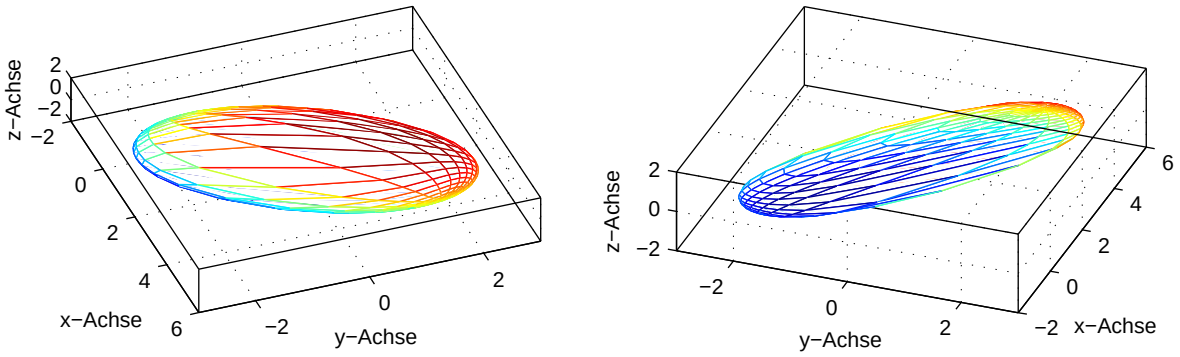

Abbildung 6.7.: Die erreichbare Menge aus Beispiel 6.2.5 (Fortsetzung)

Zum Abschluss folgt auch für die Verfahren mit Quadersteuerbereich ein dreidimensionales Beispiel um zu zeigen, dass die Verfahren nicht auf zweidimensionale Kontrollprobleme beschränkt sind. Die Betonung liegt wieder auf der Veranschaulichung der dreidimensionale erreichbaren Menge.

Beispiel 6.2.6. Gegeben sei das lineare Kontrollproblem

$$\begin{aligned} x'(t) &= \mathfrak{A}x(t) + \mathfrak{B}\mathfrak{u}(t) \quad (\text{für } t \in I := [0,1]) \\ x(0) &= x_0 \\ \mathfrak{u}(t) &\in U \quad \forall t \in I \end{aligned}$$

mit

$$\mathfrak{A} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -100 & -80 & -17 \end{pmatrix}, \ \mathfrak{B} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

der Steuermenge $U = [-1,1] \subset \mathbb{R}$ und der Anfangsbedingung $x_0 = (0,0,0)^T$. Dies ist ein nichtglattes Beispiel.

Dieses Beispiel war in allen Tests eins der schwierigsten Beispiele für die Software. Die Ursache dafür ist mir unbekannt. Zusätzliche Schwierigkeiten hatte das Ferretti-Verfahren der Ordnung 8 bei der numerischen Nullstellenbestimmung. Denn es treten Polynome auf, die durchgehens sehr große Koeffizienten haben ($> 10^7$). Dies führt zu Problemen mit dem Sekantenverfahren, denn obwohl die Nullstelle schon sehr genau bestimmt ist, ist die Auswertung an diese stelle immer noch größer als die vorgegebene Schranke. Also konvergiert das Sekantenverfahren nicht. Dies Problem konnte gelöst werden, indem die Schranke nach oben angepasst wurde.

N	Ferretti-Verfahren der Ordnung 2	Ordnung 3	Ordnung 4	Ordnung 8
2	-1.34134507	-6.61448765	4.01759970	6.13288877
4	2.38036051	4.73325262	6.92341180	1.47286993
8	6.39432037	5.47090244	6.80958616	5.21245475
16	1.89436534	4.08202795	3.85689541	8.57254837
32	2.12374474	3.21968426	4.09759948	7.92443433
64	2.06429052	3.01495086	4.00215855	8.16957062
128	2.09283980	3.09590412	4.09681694	—
256	2.04442513	3.04605473	4.04745243	—
512	2.01041695	3.00862062	4.00908206	—

Tabelle 6.15.: Konvergenzordnungen der Ferretti-Verfahren mit analytischer Integration für Beispiel 6.2.6

Die geschätzten Konvergenzordnungen der Ferretti-Verfahren mit analytischer Integration sind für dieses Beispiel in Tabelle 6.15 aufgetragen. Trotz den starken Schwankungen am Anfang konvergieren die Verfahren ab 16 Iterationen normal (negative Konvergenzordnungen bedeuten Verschlechterung). Auch das Verfahren der Ordnung 8 konvergiert dann normal. Der approximierte Hausdorff-Abstand zur Referenzmenge beträgt in der letzten Zeile, bei 512 Teilintervallen, $8.9343 \cdot 10^{-6}$, $5.0383 \cdot 10^{-8}$ und $2.0937 \cdot 10^{-10}$ bei aufsteigender Konvergenzordnung. Das Ferretti-Verfahren der Ordnung 8 hat bereits für $N = 64$ einen geschätzten Hausdorff-Abstand von $1.7779 \cdot 10^{-13}$. Bei den Verfahren mit numerischer Integration zeigt sich qualitativ dasselbe Bild. Dabei sind die Ergebnisse ab $N = 64$ den hier tabellierten sehr ähnlich, ebenso die für den geschätzten Hausdorff-Abstand.

In den Abbildungen 6.8 und 6.9 ist die erreichbare Menge für dieses Beispiel in insgesamt 10 Bildern dargestellt. Wie bei dem vorigen Beispiel ändert sich dabei die Perspektive, vor allem indem die xy-Ebene um die geneigte z-Achse im Uhrzeigersinn gedreht wird. Geht man die Bilder zeilenweise von links nach rechts durch (wie man liest), so nimmt der Drehwinkel jeweils zu. Der Farbcode entspricht dabei dem des letzten Beispiels. Große z-Werte werden in warmen Farben (Orange und Rot) und kleine z-Werte werden in kalten Farben (Türkis und Blau) dargestellt. Es handelt sich um eine komplizierte Menge, weswegen auch soviele Bilder gemacht wurden. Die Form ähnelt ein wenig der Form eines Ruderbootes, dessen obere Seite abgedeckt ist und sich ein wenig ausbeult.

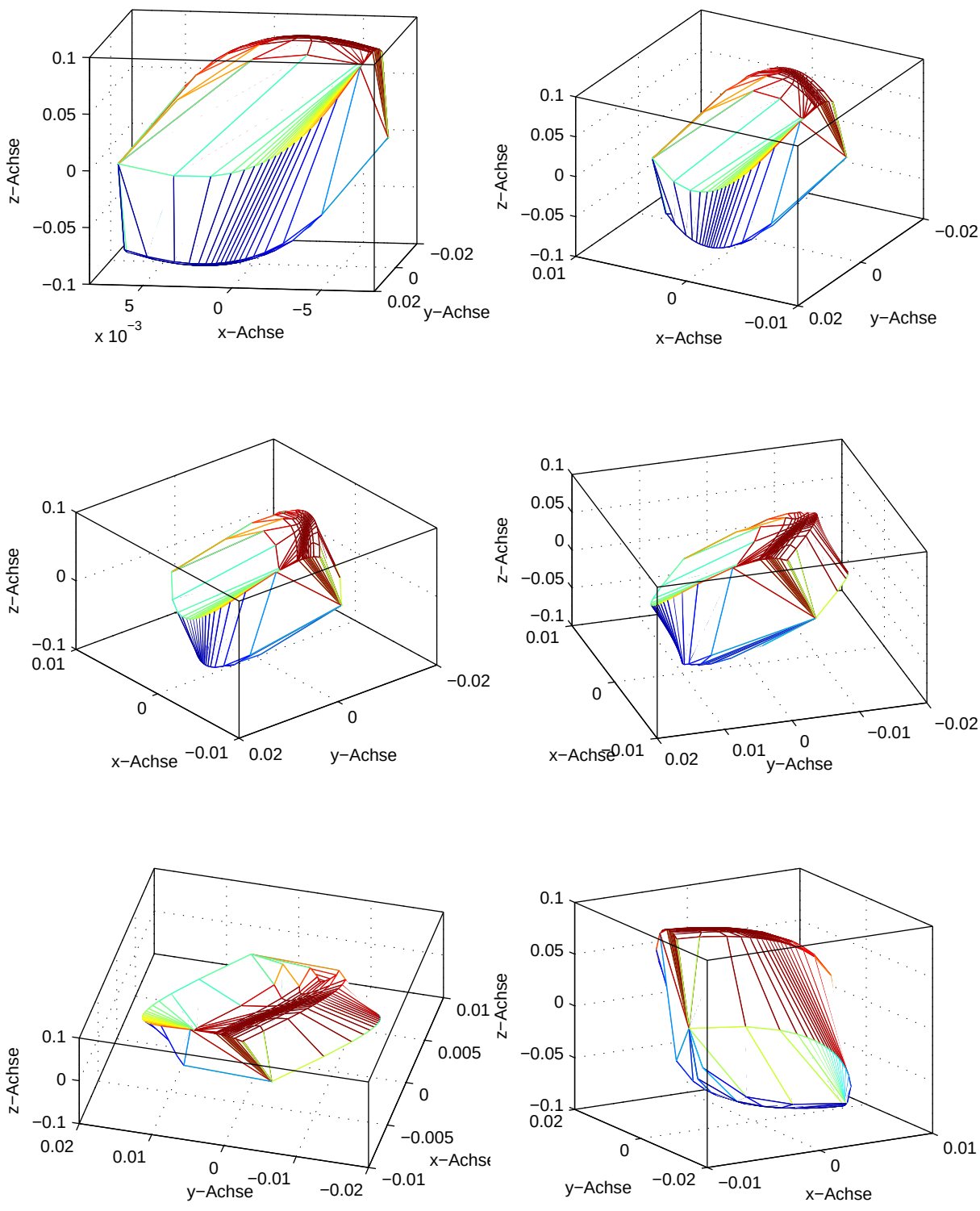

Abbildung 6.8.: Die erreichbare Menge aus Beispiel 6.2.6

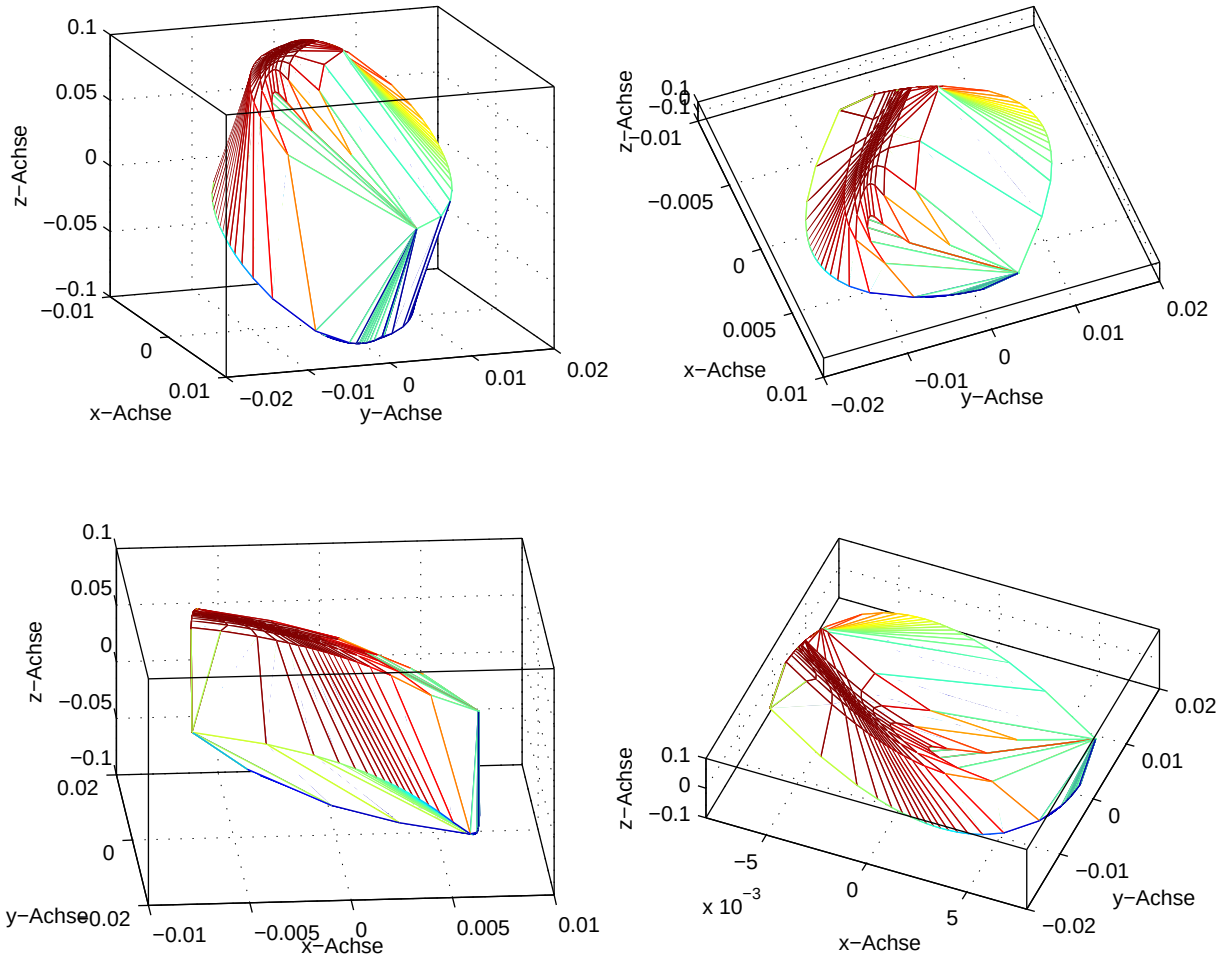

Abbildung 6.9.: Die erreichbare Menge aus Beispiel 6.2.6 (Fortsetzung)

Zusammenfassung und Ausblick

Auf der Theorie von Ferretti basierend ist es in dieser Arbeit gelungen für die Berechnung der erreichbaren Menge eines linearen Kontrollsystems bzw. einer linearen Differentialinklusion effiziente Verfahren zu entwickeln. Über die Theorie von Ferretti hinaus ist es gelungen Verfahren beliebiger Ordnung (mit analytischer Integration) für einen Quadersteuerbereich zu entwickeln. Außerdem konnte der Kontrollbereich von einem Einheitsintervall bei Ferretti zu einem belieben achsenparallelen Quader und zu beliebigen Kugeln erweitert werden. Auch für die Verfahren mit Kugelsteuerbereich besteht keine Beschränkung Konvergenzordnung, denn die Ergebnisse aus Kapitel 3 und aus Abschnitt 5.4 gelten allgemein für kompakte, konvexe und nichtleere Mengen. Jedoch werden diese Verfahren für Ordnungen > 4 wegen der numerischen Stützpunktberechnung schnell ineffektiv, wie man das in den Beispielen für die Verfahren geringerer schon sehen kann. Dasselbe gilt für die Verfahren mit numerischer Stützpunktberechnung und Quadersteuerbereich. Alle diese Verfahren konnten nicht nur theoretisch erarbeitet und deren Konvergenz bewiesen werden, sondern wurden auch praktisch implementiert und haben sich auch in der Praxis bewährt. Es konnte also die Theorie von Ferretti erfolgreich in numerische Verfahren auf dem Rechner umgesetzt werden. Im theoretischen Bereich konnte über die Arbeit von Ferretti hinaus noch gezeigt werden, dass es für eine Konvergenzordnung im Grunde nur ein Verfahren gibt, das hier Ferretti-Verfahren genannt wurde (dies hat nichts mit der mehrfachen Umsetzung dieses Verfahrens für verschiedene Stützpunktberechnungen und Kontrollbereiche zu tun).

Leider konnten die vorgestellten Konzepte für lineare Kontrollprobleme nicht auf den nichtlinearen Fall übertragen werden. Vielmehr hat sich gezeigt (siehe Kapitel 4), dass dieser Ansatz für nichtlineare Kontrollprobleme wenig Erfolg versprechend ist bzw. auf wenig effiziente Verfahren führt.

An dieser Stelle möchte ich noch einmal betonen wie wichtig und fruchtbar die praktische Umsetzung dieser Verfahren und deren Tests für diese Arbeit, auch und insbesondere für die Theorie, war. Hier haben sich vielfach Fragestellungen und Zusammenhänge ergeben, die interessant sind und auch teilweise über das Thema dieser Arbeit hinausgehen. Im Folgenden sollen einige Ergänzungs- und Erweiterungsmöglichkeiten für die entwickelten Verfahren aufgezählt werden, die aber noch näher untersucht werden müssten:

- die Erweiterung von Zeitunabhängigen Kontrollbereichen U auf zeitabhängige Kontrollbereiche $U(t)$, also mengenwertige Abbildungen, ist zumindest theore-

tisch möglich. Wie gut sich dies praktisch realisieren lässt müsste noch geklärt werden.

- ähnlich ist auch der Übergang von konstanten Matrizen $\mathfrak{B}$ auf Matrixfunktionen $t \mapsto \mathfrak{B}(t)$ zumindest theoretisch möglich. Auch hier ist nicht klar ob und wie gut sich dies umsetzen lässt.

- die numerische Stützpunktberechnung für Kugelsteuerbereiche könnte evtl. beschleunigt werden, indem man klärt ob dass Polynom $p_1(t)^2 + \ldots + p_m(t)^2$ in dem Integrationsintervall ϵ-Stellen ($1 > \epsilon > 0$) hat. Ist dies nicht der Fall, so ist die Funktion $t \mapsto \sqrt{p_1(t)^2 + \ldots + p_m(t)^2}$ beliebig oft Differenzierbar, da das Polynom von 0 weg beschränkt ist. Zur Integration könnte dann ein höher konvergentes Verfahren als die Trapezregel eingesetzt werden. Man spart sich dann eine weiter Unterteilung der Teilintervalle des mengenwertigen Verfahrens für die numerische Integration. Ob aber die mengenwertige Approximation durch die errechneten Stützpunkte von gewünschter Ordnung gegen die Momentenmenge $\mathcal{M}_{m,p}$ konvergiert, kann durch die Theorie der mengenwertigen Quadratur nicht bestätigt werden, und müsste noch untersucht werden.

Außerdem soll nun noch ein interessanter Zusammenhang aufgezeigt werden, der nicht so direkt mit den Verfahren dieser Arbeit in Verbindung steht, sondern auch allgemeiner von Interesse ist:
Es sei $\mathfrak{l} = (l_1, \ldots, l_k)^T \in S_{k-1}$ ein Koeffizientenvektor und $p(x) = \sum_{i=0}^{k} l_i x^i$ das dadurch definierte Polynom. Weiter sei $u : [0,1] \to \mathbb{R}$, $u(t) = \begin{cases} 1 & \text{für } p(t) > 0 \\ 0 & \text{für } p(t) \leq 0 \end{cases}$. Um den Stützpunkt in Richtung $\mathfrak{l}$ an die Momentenmenge $\mathcal{M}_{1,k}$, wobei der definierende Steuerbereich U für diese Menge hier $[0,1]$ sein soll, zu bestimmen muss man nach Unterabschnitt 5.4.1 das Integral

$$y(\mathfrak{l}, \mathcal{M}_{1,k}) = \int_0^1 \left(1, t, ^2 \ldots, t^{k-1}\right)^T u(t)\, dt$$

berechnen. Falls p in $[0,1]$ keine Nullstellen von ungerader Vielfachheit hat, ist $y(\mathfrak{l}, \mathcal{M}_{1,k})$ entweder gleich $\mathfrak{e}_1 := (0, \ldots, 0)^T$ oder $\mathfrak{e}_2 := \left(1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{k}\right)^T$. Da aber in diesem Fall $\mathcal{M}_{1,k} = \mathcal{H}_k$ ist, wobei $\mathcal{H}_k$ die Hausdorff-Menge aus Unterabschnitt 5.3.1 ist, sind $\mathfrak{e}_1$ und $\mathfrak{e}_2$ die beiden Ecken von $\mathcal{H}_k$. Es gilt also die Äquivalenz:

p hat Nullstellen von ungerader Vielfachheit in $[0,1]$ genau dann,
wenn der Stützpunkt $y(\mathfrak{l}, \mathcal{M}_{1,k})$ eine der Ecken $\mathfrak{e}_1$ oder $\mathfrak{e}_2$ von $\mathcal{H}_k$ ist.

Über die algebraische Beschreibung von $\mathcal{H}_k$ kann man Rückschlüsse ziehen, wann ein Polynom p in $[0,1]$ Nullstellen hat. Man kann zumindest numerisch auf der Sphäre

S_{k-1} Bereiche (=Nullstellengebiete) bestimmen, deren Vektoren Polynome festlegen, die keine Nullstellen ungerader Vielfachheit in $[0,1]$ haben.

Nach meinen bisherigen (noch sehr beschränkten) Kenntnissen von $\mathcal{H}_k$ für $k = 1, 2, 3$ vermute ich, dass es auf der Sphäre zwei zusammenhängende Bereiche gibt, welche keine Nullstellengebiete sind.

Je nach Aussehen dieser Gebiete könnte man dann vielleicht einen Referenzvektor $\mathfrak{r}$ und einer reellen Zahl s bestimmen, sodass alle $\mathfrak{l} \in S_{k-1}$ mit $\langle \mathfrak{l}, \mathfrak{r} \rangle \leq s$ in einem Nullstellengebiet liegen. Oder umgekehrt.

Damit könnte man dann ein effektives hinreichende Kriterium entwickeln um zu bestimmen, wann ein Polynom in $[0,1]$ keine Nullstellen von ungerader Vielfachheit hat.

Anhang A.

Inhalt der CDROM und Installation der Software

In diesem Kapitel des Anhangs folgt im ersten Abschnitt eine kurze Beschreibung des Inhaltes der beigefügten CDROM. Der zweite Abschnitt enthält einige allgemeine Bemerkung zur Compilierung und Installation der beigefügten Software.

A.1. Inhalt der CDROM

Im Folgenden soll "`/`" das Grundverzeichnis der CD bezeichnen. Unter Unix-Systemen ist dies ein Verzeichnis z.B. `/media/cdrom` und unter Windows ein Laufwerksbuchstabe z.B. `e:` . Jedes Verzeichnis enthält eine Datei `Liesmich.txt`, soweit dies nötig bzw. sinnvoll ist, welche Auskunft über den Inhalt des Verzeichnisses und evtl. über die Installation der Software gibt. Es folgt eine Liste der Verzeichnissen auf der CD mit der Beschreibung von deren Inhalt darunter.

`/Scilab`

Scilab[1] ist ein Softwarepaket zur numerischen Mathematik ähnlich wie *Matlab*. Es wird von INRIA und ENPC entwickelt und darf kostenlos benutzt und kopiert werden. Bitte beachten Sie auch die Lizenz, die in der Datei `license.txt` enthalten ist. Für das selbsterstellte numerische Programmpaket wurde eine Schnittstelle für *Scilab* beigefügt, die die Benutzung des Programmpaketes erheblich erleichtert. Mit dieser hier beigefügten Version von *Scilab* wurde die Schnittstelle getestet. Allerdings handelt es sich um eine schon für *Linux* compilierte Version, kann aber auch unter anderen Unix-Systemen laufen.

`/Literatur`

Dieses Verzeichnis enthält zwei Versionen des Skriptes von R.Baier und M. Gerdts "Intensive Course on Set-Valued Numerical Analysis and Optimal Control", das in dieser Arbeit häufig zitiert wurde. Sie wurden zur Unterstützung des Lesers beigefügt. Die Datei `num_analysis_opt_control.pdf` entspricht dem Literaturzitat

[1] Scilab ©INRIA-ENPC

[16] und die Datei `num_analysis_opt_control_details.pdf` dem Literaturzitat [17].

`/DA`

Dieses Verzeichnis enthält diese Diplomarbeit in verschiedenen Formen. Außerdem sind die LaTeX-Quellen dieser Arbeit hier abgelegt.

`/Software`

Dieses Verzeichnis enthält das selbsterstellte numerische Programmpaket dieser Arbeit. Bei der Installation dieser Software sollte unbedingt die Verzeichnisstruktur in diesem Verzeichnis beibehalten werden.

Es folgen nun wichtige Unterverzeichnisse dieses Verzeichnisses

`/Software/Quellcode`

Hier sind die C++ Quellen der Programmpaketes enthalten. Außerdem noch einige weitere Dateien, die für die Entwicklung wichtig sind. Desweiteren ist eine Steuerdatei für das Softwareentwicklungstool *make* enthalten und eine Anleitung zum Installieren des Programmpaketes.

`/Software/Quellcode/Dokumentation`

Dieses Verzeichnis enthält eine Dokumentation der Funktionen des Programmpaketes im html-Format.

`/Software/Scilab-Schnittstelle`

Hier sind *Scilab*-Funktionen, die das Arbeiten mit dem Programmpaket unter *Scilab* erleichtern, enthalten. Weiter ist ein *Scilab*-Skript enthalten um das in eine dynamische Bibliothek compilierte Programmpaket mit *Scilab* zu linken und die *Scilab*-Funktionen in Scilab zu laden. Außerdem befindet sich in einem Unterverzeichnis eine Dokumentation dieser Schnittstelle in Form von *Scilab*-Hilfedateien.

`/Software/Matlab-Schnittstelle`

Falls möglich wird hier später eine *Matlab*-Schnittstelle hinzugefügt. Nähere Informationen dazu sind in der Datei `Liesmich.txt` in diesem Verzeichnis enthalten.

A.2. Bemerkungen über die Installation der Software

Die Installation von *Scilab* unter *Linux* oder *Unix*-Systemen ist einfach. Eine Anleitung ist in der Datei `/Scilab/Liesmich.txt` enthalten. Nach dem Installieren beachten sie bitte auch die Lizenz in der Datei `license.txt` .

Das selbsterstellte Programmpaket wurde unter Linux programmiert und mit dem Gnu-C++ Compiler *GCC* compiliert. Auf der CD liegen aber nur die Quellen bereit, da die übersetzten Dateien zu sehr plattformabhängig sind. Zur Übersetzung wird noch die Bibliothek *uBlas* benötigt, die zu der Bibliothekensammlung *Boost* gehört. Diese Bibliothek ist auf der CD mit enthalten, mitsamt den anderen Bibliotheken der

Boost Familie. Die *Boost* Bibliotheken sind freie Software. Bitte beachten Sie die Lizenz in der Datei `/Software/Quellcode/boost/LICENSE_1_0.txt`. Ansonsten werden nur die Standardbibliotheken für C und C++ benötigt, die mit jedem Compiler mitgeliefert werden. Das Programmpaket ist also plattformunabhängig und kann unter vielen Betriebssystemen mit verschiedenen Compilern übersetzt werden. Da jedoch *uBlas* hohe Anforderungen an den Compiler stellt, sollte der verwendete Compiler nicht zu alt sein. Für die Tests wurde das Programmpaket mit der Version 4.1.0 des *GCC* compiliert.

Um den Übersetzungsprozess zu erleichtern ist noch eine Steuerdatei, ein so genanntes *Makefile*, für das Softwareentwicklungsprogramm Gnu-*make*, mit beigefügt. Damit kann die Übersetzungsprozess automatisiert werden. Allerdings muss das *Makefile* dafür angepasst werden.

Eine Anleitung zur Installation des Programmpaketes bzw. einen Verweis darauf findet man in der Datei `/Software/Quellcode/Liesmich.txt`.

Was die Schnittstellen zu *Scilab* und *Matlab* betrifft, so muss dafür zum einen das entsprechende Programm zur Verfügung stehen. Zum anderen muss das selbsterstellte Programmpaket in eine dynamische Bibliothek compiliert werden. Dafür gilt das gleich was vorher über die Compilierung gesagt wurde.

Eine Anleitung zur Installation der *Scilab*-Schnittstelle bzw. einen Verweis darauf findet man in den Dateien `/Software/Scilab-Schnittstelle/Liesmich.txt`.

Zum Zeitpunkt der Fertigstellung dieser Diplomarbeit steht noch keine *Matlab* Schnittstelle zur Verfügung. Es ist auch nicht klar ob und wie so eine Schnittstelle überhaupt erstellt werden kann. Falls möglich wird so eine Schnittstelle später nachgeliefert. In jedem Fall findet man aktuelle Informationen dazu in folgender Datei:
`/Software/Matlab-Schnittstelle/Liesmich.txt`

Anhang B.

Benutzung des selbsterstellten Programmpaketes

In diesem Kapitel soll ganz kurz und knapp anhand von einem Beispiel die Benutzung des numerischen Programmpaketes erläutert werden. Es wird vorausgesetzt, dass die Software korrekt compiliert und installiert ist. Wir werden das Beispiel 6.2.3 aus Kapitel 6 dazu verwenden. Das betreffende Kontrollproblem ist gegeben durch

$$\begin{aligned} x'(t) &= \mathfrak{A}x(t) + \mathfrak{B}\mathfrak{u}(t) \quad (\text{für } t \in I := [0, 2\pi]) \\ x(0) &= x_0 \\ \mathfrak{u}(t) &\in U \quad \forall t \in I \end{aligned}$$

mit

$$\mathfrak{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \ \mathfrak{B} = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

der Steuermenge $U = [-1, 1] \subset \mathbb{R}$ und der Anfangsbedingung $x_0 = (0, 0)^T$.

B.1. Scilab-Schnittstelle

Hier soll die Scilab-Schnittstelle vorgestellt werden. Für mehr Details und weitere Befehle steht ein Scilab-Hilfe zu dem Softwarepaket bereit. Nähere Informationen zur Scilab-Hilfe finden sich auf der CD in der Datei:
/Software/Scilab-Schnittstelle/Liesmich.txt

Es werden grundlegende Scilab-Kenntnisse vorausgesetzt.

Im Folgenden bezeichnet "`-->`" die Scilab-Eingabeaufforderung.

Zunächst definieren wir einige Variablen, die die Beispieldaten enthalten:

```
-->A=[0,1;-1,0];
-->B=[0;1];
-->x0=[0;0];
-->U=[-1,1];
```

```
-->I=[0,2*%pi];
```

Durch den Zeilenvektor `[-1,1]` wird das Intervall $[-1, 1]$ für den Steuerbereich U modelliert. Analog wird durch `[0,2*%pi]` das Zeitintervall $I = [0, 2\pi]$ dargestellt. Dabei ist `%pi` die eingebaute Scilab-Konstante für π.

Jetzt beschaffen wir uns 200 Richtungsvektoren, die durch die Parametrisierung aus 6.1.3.2 berechnet werden:

```
-->L=dim2Richt(200);
```

`L` ist nun eine 2×200-Matrix, deren Spalten die gewünschten Richtungsvektoren enthalten.

Jetzt berechnen wir zu den Richtungsvektoren in `L` die Stützpunkte für die erreichbare Menge mit dem Ferretti-Verfahren der Ordnung 8 für 60 Iterationen. Dabei benutzen wir die Variablen, die wir vorher definiert haben.

```
-->X=FerrettiVerfOrd8(60,L,A,B,x0,I,U);
```

Die Matrix `X` hat nun ebenfalls die Dimensionen 2×200. Die i-te Spalte von `X` enthält den Stützpunkt bezüglich der Richtung in der i-ten Spalte von `L`.

Jetzt berechnen wir dieselben Stützpunkte mit dem Ferretti-Verfahren der Ordnung 2 mit numerischer Integration für 10 Iterationen:

```
-->Y=FerrettiVerfOrd2_num(10,L,A,B,x0,I,U);
```

Anschließend berechnen wir den approximierten Hausdorff-Abstand für diese beiden Näherungen. Er wird nach (6.1) in 6.1.3.3 berechnet.

```
-->approxHA(X,Y,L)
```

Und Scilab liefert das Ergebnis. Mit dem folgenden Scilab-Befehl kann man das Ergebnis visualisieren.

```
-->plot(X(1,:),X(2,:)),plot(Y(1,:),Y(2,:));
```

B.2. C++ Quellcode

Hier soll ein kleines Beispielprogramm vorgestellt werden, das das Programmpaket in Form einer statischen Bibliothek, genannt *libFerretti.a*, nutzt. Für dieses Programmpaket wurde eine ausführliche HTML-Dokumentation geschrieben, in der die einzelnen Funktionen mit ihren Parametern erklährt werden. Nähere Informationen zur Scilab-Hilfe finden sich auf der CD in der Datei:
/Software/Quellcode/Dokumentation/Liesmich.txt

Zuerst soll hier das Beispielprogramm *test.cpp* als Ganzes dargestellt werden:

```
#include "FerrettiAlgorithmen.hpp"

typedef ferrettiAlgorithmen fA;
using namespace std;
```

```
int main(int argc, char* argv[])
{
linearesKontrollsystem kontr("input.txt");
erreichbareMenge menge1,menge2;


kontr.ausgebenInKonsole();

fA::FerrettiVerfahrenOrd8_UQ_2d( 60, 0.0, 2.0*pi, kontr, 200, menge1 );
fA::FerrettiVerfahrenOrd2_UQ_2d( 10, 0.0, 2.0*pi, kontr, 200, menge2 );

cout << menge1.hausdorffAbstandApprox( menge2 ) << endl;

menge1.ausgebenInDatei("output1.txt", 20, false, false);
menge2.ausgebenInDatei("output2.txt", 20, false, false);

return 0;
}
```

Und jetzt wollen wir dieses Programm Schritt für Schritt durchgehen.

```
#include "FerrettiAlgorithmen.hpp"
typedef ferrettiAlgorithmen fA;
```

In der Headerdatei "`FerrettiAlgorithmen.hpp`" sind alle Deklarationen für das ganze Softwarepaket enthalten. Die typedef-Anweisung dient hier nur zur Abkürzung. Die Klasse `ferrettiAlgorithmen` enthält alle numerischen Verfahren von diesem Programmpaket. Sie enthält nur statische Methoden und Variablen, und ist deswegen eigentlich nur ein besserer Namespace. Im Gegensatz zu einem Namespace besitzt sie auch private-Elemente.

```
linearesKontrollsystem kontr("input.txt");
erreichbareMenge menge1,menge2;

kontr.ausgebenInKonsole();
```

Die beiden Klassen `linearesKontrollsystem` und `erreichbareMenge` komplettieren dieses Softwarepaket. Es handelt sich dabei eigentlich nur um Hilfsklassen. Sie dienen zur Ein- und Ausgabe und als Parameter für die Algorithmen der Klasse `ferrettiAlgorithmen`.

Die Klasse `linearesKontrollsystem` modelliert, wie der Name schon sagt, ein lineares Konstrollsystem. Sie bietet Methoden zur Ein- und Ausgabe der Problemdaten an und prüft die Daten auf ihre Integrität. Sie dient vor allem als Eingabeparameter für die Funktionen der Klasse `ferrettiAlgorithmen`. Hier liest das Objekt `kontr` die Problemdaten aus der Datei "input.txt" ein. Über die Methode `ausgebenInKonsole()` werden diese Daten zur Kontrolle ausgegeben.

Die Klasse `erreichbareMenge` modelliert eine erreichbare Menge eines linearen Kontrollproblems (eigentlich eine beliebige konvexe Menge). Intern wird die Menge in Form von Stützpunkten und zugehörigen Richtungsvektoren abgespeichert. Sie bietet Methoden zur Ein- und Ausgabe ihrer Daten an. Sie dient als Rückgabeparameter für die Funktionen der Klasse `ferrettiAlgorithmen`. Hier werde zwei Objekte `menge1` und `menge2` erzeugt, die zunächst leer sind bzw. die leere Menge darstellen.

```
fA::FerrettiVerfahrenOrd8_UQ_2d( 60, 0.0, 2.0*pi, kontr, 200, menge1 );
fA::FerrettiVerfahrenOrd2_UQ_2d( 10, 0.0, 2.0*pi, kontr, 200, menge2 );
```

Mit diesen beiden Funktionen werden Stützpunkte an die erreichbare Menge in `200` Richtungen berechnet, und zwar mit den Ferretti-Verfahren der Ordnung 2 und 8 mit analytischer Integration. Die double-Werte `0.0` und `2.0*pi` (die Konstante pi wird in der Header-Datei definiert) bestimmen das Zeitintervall $[0, 2\pi]$ des Kontrollproblems, denn dies ist nicht in der Klasse `linearesKontrollsystem` enthalten (da es sich um autonome Systeme handelt). Die int-Werte `60` und `10` bestimmen die Iterationszahlen der Verfahren und damit zusammen mit dem Zeitintervall die Schrittweite. Der int-Wert `200`, legt die Anzahl der Richtungen fest, die mit Polarkoordinaten erzeugt werden, wie das in 6.1.3.2 auf Seite 132 gezeigt wurde. Das Objekt `kontr` liefert die Problemdaten und die Objekte `menge1` und `menge2` nehmen das Ergebnis auf.

```
cout << menge1.hausdorffAbstandApprox( menge2 ) << endl;

menge1.ausgebenInDatei("output1.txt", 20, false, false);
menge2.ausgebenInDatei("output2.txt", 20, false, false);
```

Die Funktion `hausdorffAbstandApprox` gehört zur Klasse erreichbareMenge und berechnet eine Näherung für den Hausdorff-Abstand zwischen `menge1` und `menge2`. Dies geschieht mit Hilfe von Stützfunktionen durch die Formel (6.1) auf Seite 134. Mit der Funktion `ausgabeInDatei` wird die errechneten Approximationen in eine Datei ausgegeben. Die Ausgabe erfolgt in zwei großen Bandmatrizen, die in ihren Spalten die Richtungsvektoren bzw. die Stützpunkt enthalten.

Dieses Beispielprogramm muss jetzt noch compiliert werden und mit der Bibliothek `libFerretti.a` gelinkt werden. Fertig.

Dies war nur eine kurzes Beispiel zur Orientierung. Für mehr Informationen konsultieren Sie bitte die mitgelieferte, ausführliche Dokumentation.

Literaturverzeichnis

[1] R. A. Adams. *Sobolev Spaces*. Academic Press, Inc., New York, 1975.

[2] J.-P. Aubin and H. Frankowska. *Set-Valued Analysis*, volume 2 of Systems & Control: Foundations and Applications. Birkhäuser, Boston-Basel-Berlin, 1990.

[3] R. J. Aumann. *Integrals of Set-Valued Functions*. J. Math. Anal. Appl., 12(1):1-12, 1965.

[4] R. Baier. *Mengenwertige Integration und die diskrete Approximation erreichbarer Mengen*. Bayreuther Mathematische Schriften, Band 50. Mathematisches Institut der Universität Bayreuth, Bayreuth, 1995.

[5] R. Baier. Extrapolation methods for the computation of set-valued integrals and reachable sets of linear differential inclusions. *Zeitschrift für Angewandte Mathematik und Mechanik*, 74(6): T 555- T 557, 1994.

[6] R. Baier and F. Lempio. Approximating reachable sets by extrapolation methods. In P.-J. Laurent, A. Le Méhauté, and L. L. Schumaker, editors, *Curves and Surfaces in Geometric Design,* pages 9-18, Wellesley, Massachusetts, 1994. A K Peters.

[7] H. Bauer. *Maß- und Integrationstheorie, 2. , überarbeitete Auflage*. de Gruyter, Berlin-New York, 1992.

[8] S. Bosch. *Algebra, 3. Auflage*. Springer-Verlag, Berlin-Heidelberg-New York, 1999.

[9] V. I. Burenkov. *Sobolev spaces on domains*. Teubner-Verlag, Stuttgart-Leipzig, 1998.

[10] J. C. Butcher. *Numerical Methods for Ordinary Differential Equations*. John Wiley & Sons, Ltd, 2003.

[11] S. B. Chae. *Lebesgue Integration*. Marcel Dekker, Inc. , New York and Basel, 1980.

[12] L. Collatz. *Funktionalanalysis und Numerische Mathematik*. Unveränderter Nachdruck der ersten Auflage von 1964. Die Grundlehren der Mathematischen Wissenschaften, Band 120. Springer-Verlag, Berlin-Heidelberg-New York, 1968.

[13] R. Ferretti. High-Order Approximations of Linear Control Systems via Runge-Kutta Schemes. *Computing*, 58:351-364, 1997

[14] V. I. Blagodatskikh and A. F. Filippov. Differential inclusions and optimal control. In *Topology, Ordinary Differential Equations, Dynamical Systems*, volume 1986, issue 4 of *Proc. Steklov Inst. Math.*, pages 199-259. AMS, Providence, Rhode Island, 1987.

[15] C. Geiger and C. Kanzow. *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer-Verlag, Berlin-Heidelberg-New York, 2002.

[16] M. Gerdts and R. Baier. *Intensive Course on Set-Valued Numerical Analysis and Optimal Control*. Preliminary Version: 29th September 2005.
Dieses Buch ist auf der beigefügten CD enthalten. Siehe AbschnittA.1 im Anhang.

[17] M. Gerdts and R. Baier. *Intensive Course on Set-Valued Numerical Analysis and Optimal Control*. Preliminary Version: 1st September 2005.
Dieses Buch ist auf der beigefügten CD enthalten. Siehe AbschnittA.1 im Anhang.

[18] A. Ghizzetti. *Ricerche sui momenti di una funzione limitata compresa tra limiti assegnati*. Atti della Reale Accademia d'Italia 13, p. 1165-1199, 1942.

[19] F. Hausdorff. *Momentenprobleme für ein endliches Intervall*. Math. Z. 16, p. 220-248, 1923.

[20] K. Königsberger. *Analysis 1, 4., neu bearbeitete und erweiterte Auflage* . Springer-Verlag, Berlin-Heidelberg-New York, 1999.

[21] K. Königsberger. *Analysis 2, 3., überarbeitete Auflage* . Springer-Verlag, Berlin-Heidelberg-New York, 2000.

[22] F. Lempio. *Numerische Mathematik I: Methoden der linearen Algebra*. Bayreuther Mathematische Schriften, Band 51. Mathematisches Institut der Universität Bayreuth, Bayreuth, 1997.

[23] F. Lempio. *Numerische Mathematik II: Methoden der Analysis*. Bayreuther Mathematische Schriften, Band 55. Mathematisches Institut der Universität Bayreuth, Bayreuth, 1998.

[24] W. Müller. *Algebra*. Bayreuther Mathematische Schriften, Band 57. Mathematisches Institut der Universität Bayreuth, Bayreuth, 1999.

[25] I. P. Natanson. *Theorie der Funktionen einer reellen Veränderlichen, 4. Auflage.* Verlag Harri Deutsch, Zürich-Frankfurt am Main-Thun, 1975.

[26] A. W. Roberts and D. E. Varberg. *Convex Functions*. Academic Press, New York and London, 1973.

[27] E. D. Sontag. *Mathematical Control Theory, 2. Auflage*. Springer-Verlag, Berlin-Heidelberg-New York, 1998.

[28] J. Stoer and R. Bulirsch. *Numerische Mathematik 2, 3. verbesserte Auflage*. Springer-Verlag, Berlin-Heidelberg-New York, 1990.

[29] W. Walter. *Gewöhnliche Differentialgleichungen, 7. Auflage*. Springer-Verlag, Berlin-Heidelberg-New York, 2000.

[30] E. Zeidler. *Teubner-Taschenbuch der Mathematik*. Teubner-Verlag, Stuttgart-Leipzig, 1996.

[31] R. Wheeden and A. Zygmund. *Measure and Integral*. Marcel Dekker, Inc. , New York and Basel, 1977.

Erklährung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.

Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen.

Bayreuth, 1. Oktober 2006

_____

Hannes Buchholzer