Anwendung von stetigen Runge-Kutta-Verfahren auf Probleme der optimalen Steuerung

Dissertation
zur Erlangung des akademischen Grades

doctor rerum naturalium (Dr.rer.nat.)

vorgelegt dem Rat der Fakultät für Mathematik und Informatik

der Friedrich-Schiller-Universität Jena

von Dipl.-Math. oec. Dirk Hemmelmann

geboren am 13. Juli 1978 in Erfurt

Gutachter:
1. Prof. Dr. Walter Alt (Friedrich-Schiller-Universität Jena)
2. Prof. em. Dr. Frank Lempio (Universität Bayreuth)

Tag der öffentlichen Verteidigung: 03.02.2010

Inhaltsverzeichnis

Abbildungsverzeichnis

Notationen

$\mathbb{R}^n$	n-dimensionaler Euklidischer Raum
$0_{\mathcal{Y}},\ 0_n$	Nullvektor im Vektorraum $\mathcal{Y}$ bzw. im $\mathbb{R}^n$
f_x	Partielle Ableitung von f nach x
$\dot{x}$	Ableitung von $x(t)$ nach t
$L^\alpha(J;\mathbb{R}^n)$	L^α-Raum von Funktionen die das Intervall J in $\mathbb{R}^n$ abbilden
$W^{k,\alpha}(J;\mathbb{R}^m)$	Sobolev-Raum von vektorwertigen Funktionen deren j-te Ableitung fur alle $0 \leq j \leq k$ im Raum L^α liegt
$\langle\cdot,\cdot\rangle_{L^2}$	Skalarprodukt im Hilbertraum $L^2(J;\mathbb{R}^n)$
$\mathcal{O}(N)$	Landau-Symbol für wächst nicht wesentlich schneller als N
bzw.	Abkürzung für beziehungsweise
SRKV	Abkürzung für stetige Runge-Kutta-Verfahren bzw. Runge-Kutta-Verfahren mit stetigen Erweiterungen
GDNN	Abkürzung für Generalisierte Dynamische Neuronale Netzwerke
u.Nb.	Abkürzung für „unter der/den Nebenbedingung/en“

1 Einleitung

Häufig wird bei numerischen Verfahren zur Lösung von Problemen der optimalen Steuerung zwischen direkten und indirekten Methoden unterschieden (siehe zum Beispiel [Bet01]). Dabei bezeichnet man in der Regel mit direkten Methoden diejenigen, die vesuchen direkt die Zielfunktion zu minimieren. Mit indirekten Methoden bezeichnet man Algorithmen, bei denen man versucht, eine Lösung der notwendigen Optimalitätsbedingungen zu berechnen. Möchte man dabei eine numerische Lösung des stetigen Steuerungsproblems mit Hilfe einer direkten Methode berechnen, wird in der Regel das gesamte Problem diskretisiert und ein Verfahren der nichtlinearen Optimierung auf das entstehende endlich-dimensionale Problem angewandt. Dadurch erhält man im Allgemeinen eine diskrete Näherungslösung auf einem vorgegebenen Gitter. Untersucht man den Diskretisierungsfehler der dabei gemacht wird, bleibt es natürlich zunächst offen, ob man das entstehende diskrete Steuerungsproblem mit Hilfe eines direkten oder eines indirekten Verfahrens numerisch löst. Weitere Herangehensweisen, bei denen zu einem späterem Zeitpunkt im Algorithmus diskretisiert wird, ignorieren in der Regel den dabei entstehenden Diskretisierungsfehler.

Diese Arbeit beschäftigt sich damit, zu untersuchen, inwiefern stetige Runge-Kutta-Verfahren, im Folgenden mit SRKV abgekürzt, Vorteile bei der Diskretisierung von Steuerungsproblemen bieten. Es stellt sich dabei heraus, dass eine Diskretisierung des stetigen Steuerungsproblems mit Hilfe von SRKV weder für eine theoretische Betrachtung des Diskretisierungsfehlers geeignet ist, noch intuitiv bei der Entwicklung von Optimierungsverfahren erscheint. Somit untersuchen wir formal eine indirekte Methode um mit Hilfe von SRKV stetige Näherungslösungen von stetigen Problemen der optimalen Steuerung berechnen zu können. Wir werden zeigen, dass unter relativ allgemeinen Voraussetzungen Algorithmen möglich sind, bei denen man ohne weitere Interpolationen Näherungslösungen an jedem Punkt berechnen kann. Dies ist in Hinblick auf eine einfache und intuitive Anwendung solcher Algorithmen sehr interessant, wenn man mit großer Genauigkeit an beliebigen Punkten optimale Steuerungen berechnen möchte. Es wird insbesondere gezeigt, dass man dabei eine sehr hohe Konvergenzordnung der stetigen Näherungslösungen bezüglich der Schrittweite der Diskretisierung erhalten kann.

1.1 Runge-Kutta-Verfahren zur Diskretisierung

Runge-Kutta-Verfahren sind bei der numerischen Lösung von gewöhnlichen Differentialgleichungen weit verbreitet. Daher bieten sich diese Verfahren an, um numerische Lösungen von Problemen der optimalen Steuerung zu berechnen. Wichtige theoretische Arbeiten zu diesem Thema stammen von W. W. Hager und A. L. Dont-

chev [DHV98, Hag00, DH01]. Dabei wurden fast gleichzeitig die sich ergänzenden Arbeiten über Runge-Kutta-Verfahren mit Konvergenzordnung 2 in [DHV98] und die Arbeit zu Runge-Kutta-Verfahren mit höherer Konvergenzordnung in [Hag00] veröffentlicht. Die Beweise und die Herangehensweise, um die entsprechenden Konvergenzordnungen zu erhalten, unterscheiden sich dabei deutlich.

In [DMR06] werden Verfahren der numerischen Integration mit stetigen Runge-Kutta-Verfahren (SRKV) zur Berechnung von Näherungslösungen von stetigen Problemen der optimalen Steuerung gekoppelt. Es wird gezeigt, dass die Zielfunktion des diskretisierten Problems mit einer hohen Geschwindigkeit gegen die Zielfunktion des stetigen Problems konvergiert. Allerdings wird dabei keine theoretische Aussage darüber gemacht, ob sich die Konvergenzgeschwindigkeit auf die Näherungslösungen der optimalen Steuerung überträgt und ob es zu einer eindeutig bestimmten optimalen Steuerung, für hinreichend kleine Schrittweiten der Diskretisierung, eine eindeutig bestimmte Näherungslösung gibt. Die theoretischen Ergebnisse in Kapitel 3 und die numerischen Resultate im Kapitel 5, insbesondere das Beispiel 5.3, werden zeigen, dass die eigentlich interessanten Aussagen hierüber von dem konkreten Steuerungsproblem und dessen Lösung abhängen.

In [Hag00] werden zusätzliche Bedingungen an „diskrete“ Runge-Kutta-Verfahren gestellt, damit sie geeignet sind, um sie für stetige Steuerungsprobleme zu benutzen und um mit ihnen eine hohe Konvergenzordnung der Näherungslösungen zu erreichen. Es wird gezeigt, dass diese Einschränkungen für explizite Runge-Kutta-Verfahren mit Konvergenzordnung kleiner oder gleich 4 weitestgehend unbedeutend sind. Das klassische Runge-Kutta-Verfahren erfüllt zum Beispiel alle Bedingungen, um mit Hilfe der in [Hag00] vorgeschlagenen Diskretisierung bestimmte stetige Steuerungsprobleme näherungsweise mit Konvergenzordnung 4 zu lösen. Dabei werden allerdings keine Verfahren mit einer Konvergenzordnung größer als 4 vorgeschlagen, die alle aufgestellten Bedingungen erfüllen. Außerdem werden in [Hag00] keine Aussagen über implizite Runge-Kutta-Verfahren mit hoher Konvergenzordnung hergeleitet.

Ziel dieser Arbeit war zunächst zu prüfen, inwieweit man die von W.W. Hager (siehe [Hag00]) gezeigten Ergebnisse auf SRKV übertragen kann und dadurch stetige Näherungslösungen von stetigen Problemen der optimalen Steuerung erhält. Dabei stellt sich sehr schnell heraus, dass die zusätzliche Bedingungen, die an die Runge-Kutta-Verfahren gestellt werden, praktisch in keinem SRKV mit Konvergenzordnung größer als zwei erfüllt sind. Setzt man in die Rechnungen in [Hag00] an jede Stelle die stetigen Versionen der Konsistenzgleichungen ein, so sieht man, dass sich einige Vereinfachungen für die stetigen Konsistenzgleichungen nicht ergeben. Für stetige Verfahren mit mindestens quadratischer Interpolation zwischen den Gitterpunkten widersprechen sich dabei sogar die stetigen Varianten der Konsistenzbedingungen, die man nach [Hag00] aufstellen müsste.

In den Arbeiten [DHV98, Hag00] wird im Wesentlichen ein diskretes Steuerungsproblem aufgestellt und gezeigt, dass die Lösung für hinreichend kleine Schrittweiten der Diskretisierung mit hoher Geschwindigkeit gegen die Lösung des stetigen Steuerungsproblems an den Diskretisierungspunkten konvergiert. Es werden keine Aussagen darüber gemacht, wie man Lösungen außerhalb der Gitterpunkte berechnen kann. Dabei konvergiert nach [Hag00] zunächst nur der berechnete

optimale Zustand und der optimale adjungierte Zustand des diskreten Steuerungsproblems gegen die Lösung des stetigen Steuerungsproblems mit der Konvergenzordnung 3 oder 4 des expliziten Runge-Kutta-Verfahrens. Die optimale Steuerung des diskreten Problems kann dabei mit einer um ein oder zwei Ordnungen geringeren Geschwindigkeit konvergieren. An den Gitterpunkten der Diskretisierung kann man sich a posteriori zu dem optimalen Zustand und adjungierten Zustand eine Steuerung berechnen, die mit der selben Geschwindigkeit, in Abhängigkeit von der Schrittweite, gegen die Optimalsteuerung des stetigen Problems konvergiert. In [DHV98] gibt es diese Einschränkung für Verfahren der Ordnung 2 nicht.

Diese Probleme beim Übertragen der Konvergenzordnung der Runge-Kutta-Verfahren für Anfangswertprobleme auf Algorithmen zur Berechnung von Näherungslösungen von stetigen Steuerungsproblemen entstehen dabei dadurch, dass man aus dem Maximumprinzip nach Pontryagin (siehe [PBGM64]) für stetige Steuerungsprobleme zu einer gegebenen Steuerung ein Anfangswertproblem für die Zustände und ein Endwertproblem für die adjungierten Zustände lösen muss. Dabei kann man entweder in einem iterativen Verfahren daraus eine neue Steuerung berechnen, oder man kann in linear-quadratischen-Problemen versuchen, die Steuerung durch das Maximumprinzip als Funktion von Zustand und adjungierten Zustand auszudrücken, und löst das Randwertproblem aus Zustandsgleichung und der Gleichung für die adjungierten Zustände.

Diskretisiert man die Zustandsgleichungen und die Gleichungen für die adjungierten Zustände mit einem Runge-Kutta-Verfahren, werden im Allgemeinen nicht alle Diskretisierungspunkte aufeinander fallen. In diesem Fall benötigt man den Zustand an Punkten, an denen man die Zustandsgleichungen nicht diskretisiert hat. Bei SRKV kann man diese Punkte sehr leicht mit einer relativ hohen Konvergenzordnung berechnen, allerdings geht dies für diskrete Runge-Kutta-Verfahren nicht. Dieses Problem besteht nicht, wenn man das stetige Problem der optimalen Steuerung diskretisiert und daraus diskrete adjungierte Zustandsgleichungen berechnet. Im nichtlinearen Fall werden sich im Allgemeinen diese diskreten adjungierten Gleichungen von der Diskretisierung der adjungierten Gleichungen des stetigen Steuerungsproblems unterscheiden. Daher kann man einerseits die hohe Konvergenzordnung der Runge-Kutta-Verfahren nicht einfach auf ein diskretisiertes Steuerungsproblem übertragen, aber andererseits auch nicht die diskretisierten Zustandsgleichungen und diskretisierten adjungierten Gleichungen eines stetigen Problems einem diskretisiertem Steuerungsproblem zuordnen.

Wir zeigen in dieser Arbeit, dass die Verwendung von SRKV die Möglichkeit eröffnet, die Konvergenzordnung, die die stetigen Erweiterungen zwischen den Gitterpunkten für Anfangswertprobleme bieten, auf die Diskretisierung der Zustandsgleichungen und der Gleichungen für die adjungierten Zustände zu übertragen. Mit Hilfe der diskretisierten Zustandsgleichungen und adjungierten Gleichungen wird ein Gleichungssystem der diskretisierten notwendigen Optimalitätsbedingungen 1. Ordnung aufgestellt. Es wird gezeigt, dass man an SRKV keine weiteren Bedingungen stellen muss, um sie bei der Berechnung von Näherungslösungen stetiger Steuerungsprobleme einsetzen zu können und dabei eine hohe Konvergenzordnung zu erhalten.

Im Kapitel 2 werden SRKV vorgestellt und die grundlegenden Aussagen gezeigt,

die als Motivation dafür dienen, sie bei der Berechnung von Näherungslösungen stetiger Steuerungsprobleme einzusetzen. Danach wird im Kapitel 3 gezeigt, wie sich die Konvergenzordnung der stetigen Erweiterung des Runge-Kutta-Verfahrens auf die Lösung des Gleichungssystems aus den diskretisierten Optimalitätsbedingungen 1. Ordnung überträgt. Schließlich werden wir im Kapitel 4 die wesentlichen Schritte erarbeiten, um zu zeigen, dass man mit dem Newton-Verfahren eine Nullstelle der diskretisierten notwendigen Optimalitätsbedingungen 1. Ordnung berechnen kann. Im Kapitel 5 demonstrieren wir anhand von numerischen Resultaten, dass sich bei einfachen Problemen die Konvergenzordnung des stetigen Runge-Kutta-Verfahrens vollständig auf die Näherungslösung des Steuerungsproblems übertragen lässt. Im Kapitel 4 beschreiben wir die Algorithmen, mit denen die numerischen Resultate aus Kapitel 5 berechnet wurden. Dabei schlagen wir vor, gegebenenfalls direkte und indirekte Methoden gemeinsam zu benutzen, um ihre jeweiligen Vorteile zu kombinieren.

1.2 Dynamische Neuronale Netzwerke

Für eine große Zahl unterschiedlicher Probleme wurden in den vergangenen Jahren Lösungsansätze mit Hilfe von künstlichen Neuronalen Netzwerken erprobt (siehe zum Beispiel [Roj96, Zel94]). Dabei wurden für unterschiedliche Aufgaben verschiedene künstliche Neuronale Netzwerke entwickelt. In der Signalverarbeitung und der Analyse von Messwerten werden dabei unter anderem Dynamische Neuronale Netzwerke eingesetzt. Erweitert man diese zu Generalisierten Dynamischen Neuronalen Netzwerken (im Folgenden als GDNN bezeichnet, siehe [GLZW04]), dann kann man die Aufgabe des überwachten Lernens vorgegebener Muster zur Mustererkennung durch ein stetiges Steuerungsproblem beschreiben. Allgemein beschreibt man ein solches künstliches Dynamisches Neuronales Netzwerk zu jedem Zeitpunkt $t \in [0, T]$ durch den Zustand $x(t)$. Dieser Zustand hat die Dimension n für n Knoten aus denen das Netzwerk besteht. Damit gibt es n^2 mögliche gerichtete Verbindungen zwischen den Knoten. Modelliert man dazu Verbindungsgewichte durch einen zeitabhängigen $m = n^2$-dimensionalen Vektor $u(t)$, dann kann man die Dynamik des künstlichen Netzwerks durch eine Differentialgleichung

$$\dot{x}(t) = f(x(t), u(t), I^r(t)) \tag{1.2.1}$$

beschreiben, wobei $I^r(t)$ im einfachsten Fall das zu lernende Muster ist. Für dieses Muster möchte man die Verbindungsgewichte $u(t)$ so wählen, dass die Lösung der Differentialgleichung (1.2.1) einen möglichst geringen Abstand zu einem vorgegebenen Zustand $x^r(t)$ hat. Entsprechend wählt man eine Zielfunktion, die diesen Abstand beschreibt und welche minimiert werden soll. Eine formale Definition einer Zielfunktion für ein einzelnes Muster ist damit

$$\min \int_0^T \|x(t) - x^r(t)\|_2 \, dt.$$

Eine sinnvolle Variante, die Gleichung (1.2.1) zu spezifizieren, ist zum Beispiel

$$\dot{x}(t) = -x(t) + \arctan\big(W(t)\,x(t)\big) + I^r(t),$$

wobei in $W(t)$ die Steuerungen $u(t)$ zu einer Matrix angeordnet werden. Dabei werden neben dem Arkustangens verschiedene Aktivierungsfunktionen benutzt, die meist einen ähnlichen, S-förmigen Verlauf haben.

Erweitert man dieses Problem auf $\mathcal{K}$ Muster, die man gleichzeitig lernen möchte, so erhält man ein GDNN, welches zum Beispiel in [ZLM+04] erfolgreich zur Mustererkennung eingesetzt wird. Praktisch führt man dabei lediglich entsprechende Summen über $r = 1, \ldots, \mathcal{K}$ in den Gleichungen ein. Im Allgemeinen wird man dabei die $\mathcal{K}$ Trainingsmuster in wenige Klassen unterteilen, die jeweils den gleichen vorgegebenen Zustand $x^r(t)$ haben. Formal hat man dabei immer noch das gleiche stetige Steuerungsproblem, indem man die Zustände und das Eingangssignal $I(t)$ auf die Dimension $n\mathcal{K}$ erweitert und die Funktion f aus der Gleichung (1.2.1) sowie die Zielfunktion entsprechend anpasst. Dabei wird natürlich nicht die Dimension m der Steuerung verändert, da man eine Steuerung erhalten möchte, mit deren Hilfe man anschließend unbekannte Muster $x^r(t)$ mit $r > \mathcal{K}$ den verschiedenen Klassen zuordnen kann. Betrachtet man den Trainingsalgorithmus, also die Lösung eines stetigen Steuerungsproblem, sind dabei allerdings für das globale Minimum nicht die hinreichenden Optimalitätsbedingungen erfüllt, die meist für theoretische Resultate über die Konvergenz von Verfahren vorausgesetzt werden müssen. Dieses Problem kann man in der Regel durch die Einführung von Regularisierungstermen in der Zielfunktion lösen. Dies ist aber je nach Anwendung mitunter nicht erwünscht oder relativ unpraktisch.

Wir werden in Abschnitt 5.3 ein einfaches Problem beschreiben, welches mit einem Abstiegsverfahren trainiert werden kann. Dabei konvergiert auch für sehr gute Startwerte der Steuerung ein Newton-Verfahren zur Lösung der diskretisierten Optimatiätsbedingungen nicht. Außerdem erweist es sich praktisch als äußerst schwierig, diese Situation durch die Addition von Regularisierungstermen an die Zielfunktion zu verbessern, ohne dabei das globale Minimum zu verändern. Dieses Beispiel zeigt deutlich den Vorteil, der dadurch entstehen kann, dass die vorgeschlagene Diskretisierung der notwendigen Optimalitätsbedingungen 1. Ordnung unter anderem die Möglichkeit bietet, sehr genaue Näherungen für Gradienten des ursprünglichen Problems zu berechnen und damit einfache Abstiegsverfahren zu konstruieren.

2 Stetige Einschrittverfahren

In diesem Kapitel wollen wir zunächst Einschrittverfahren formal auf stetige Einschrittverfahren erweitern. Die folgenden Definitionen und Sätze sind in dieser oder ähnlicher Form in vielen Lehrbüchern über numerische Verfahren zur Lösung von gewöhnlichen Differentialgleichungen für diskrete Einschrittverfahren vorhanden (siehe unter anderem [But03, SB80, HNW93]). Prinzipiell bietet dabei eine Erweiterung auf stetige Einschrittverfahren keine wesentlichen Unterschiede. Für das Verständnis dieser Arbeit sind dabei der Beweis von Satz 2.1.5 und die Folgerung 2.1.6 sehr wichtig. Weiterhin entstehen Fragen, die sich so bei diskreten Einschrittverfahren nicht stellen, zum Beispiel kann man bei SRKV Ableitungen der lipschitzstetigen Näherungslösungen nach der unabhängigen Variable berechnen. Wir werden in diesem Kapitel stetige Einschrittverfahren und insbesondere stetige Erweiterungen von Runge-Kutta-Verfahren soweit vorstellen, dass man in Kapitel 3, bei dem Beweis der Existenz einer stetigen Näherungslösung des speziellen Steuerungsproblems, darauf zurückgreifen kann.

Wir betrachten im Folgenden Einschrittverfahren, welche eine stetige Näherungslösung x_h des Anfangswertproblems

$$\text{(AWP)} \qquad \dot{x}(t) = g(x(t), t),\ \forall t \in [0, T],\ x(0) = a$$

berechnen. Dabei gehen wir immer davon aus, dass wir das Intervall $[0, T]$ in $N \in \mathbb{N}$ gleich große Intervalle unterteilen, und bestimmen daraus die Schrittweite $h = \frac{T}{N}$ des Einschrittverfahrens. Die Näherungslösung x_h, ausgehend von den Anfangswerten $t_0 = 0$ und $x_h(t_0) = x_h(0) = a$, berechnet man für $k = 0, 1, \ldots, N-1$ und $\theta = \frac{t - t_j}{h} \in (0, 1]$ wie folgt:

$$\begin{aligned} t_{k+1} &= t_k + h \\ x_h(t_k + \theta h) &= x_h(t_k) + h\,\Phi(t_k, x_h(t_k), h, \theta, g). \end{aligned}$$

Hierbei ist Φ die Verfahrensfunktion, welche das konkrete Einschrittverfahren definiert. Für eine Verfahrensfunktion sollte dabei Φ stetig von $\theta \in (0, 1]$ abhängen und außerdem $\lim_{\theta \to 0} \Phi(t_k, x_h(t_k), h, \theta, g) = 0$ gelten, damit die berechnete Näherungslösung x_h stetig ist.

2.1 Konvergenz von stetigen Einschrittverfahren

Wir wollen zunächst die wichtigsten Begriffe für Einschrittverfahren für den stetigen Fall definieren, um damit ein Resultat über die Konvergenz von Einschrittverfahren aufzustellen. Dabei beschränken wir uns im Folgenden auf Funktionen g auf der rechten Seite des Problems (AWP), die hinreichend glatt sind, und definieren mit einer Konstanten $\kappa \geq 1$ den Begriff der Konsistenz.

Definition 2.1.1 *Gilt für alle $t \in [0,T]$, $z \in \mathbb{R}^n$ und $g(z,\cdot) \in C^\kappa([0,T],\mathbb{R}^n)$*

$$\lim_{h\to 0} \Phi(t,z,h,1,g) = g(z,t),$$

dann heißt das Einschrittverfahren konsistent.

Um einen Fehler des Näherungsverfahrens betrachten zu können, definieren wir zu der exakten Lösung $\tilde{x}$ des Anfangswertproblems (AWP) mit

$$\triangle(t,\tilde{x},\theta h,g) := \begin{cases} \frac{\tilde{x}(t+\theta h)-\tilde{x}(t)}{\theta h} & \text{für } h>0 \\ g(\tilde{x}(t),t) & \text{für } h=0 \end{cases} \tag{2.1.1}$$

einen Differenzenquotienten (vergleiche [SB80]) und definieren damit die Begriffe des Konsistenzfehlers und der Konvergenz des Einschrittverfahrens.

Definition 2.1.2 (Konsistenzfehler) *Die Differenz*

$$r_\theta(t_k,\tilde{x},h,g) = \theta\triangle(t_k,\tilde{x},\theta h,g) - \Phi(t_k,\tilde{x}(t_k),h,\theta,g),\ \theta\in(0,1]$$

heißt Konsistenzfehler *oder lokaler Verfahrensfehler des Einschrittverfahrens an der Stelle $t_k+\theta h$, mit $t_k = kh$, $k=0,\dots,N-1$. Gilt mit positiven, von h unabhängigen Konstanten η_p und η_q, für zwei positive reelle Zahlen p und q*

$$\sup_{k=0,\dots,N-1} \|r_1(t_k,\tilde{x},h,g)\|_2 \le \eta_p h^p,$$
$$\sup_{k=0,\dots,N-1}\ \sup_{0<\theta<1} \|r_\theta(t_k,\tilde{x},h,g)\|_2 \le \eta_q h^q,$$

so spricht man von einem stetigen Einschrittverfahren mit der Konsistenzordnung *(p,q). Dabei bezeichnet p die Konsistenzordnung auf den Gitterpunkten $t_k = kh$, $k=1,\dots,N$, und q die Interpolationsgüte im Intervall $[0,T]$ außerhalb der Gitterpunkte.*

Definition 2.1.3 (Konvergenz) *Es sei für alle $z\in\mathbb{R}^n$ die Funktion $g(z,\cdot)\in C^\kappa([0,T];\mathbb{R}^n)$, dann sagen wir ein Einschrittverfahren ist* konvergent*, wenn für den globalen Diskretisierungsfehler $e_h(t) := x_h(t)-\tilde{x}(t)$ für alle $t\in[0,T]$*

$$\lim_{h\to 0} e_h(t) = \lim_{h\to 0}(x_h(t)-\tilde{x}(t)) = 0 \tag{2.1.2}$$

gilt.

Um ein Resultat über die Konvergenz zu zeigen, benötigen wir folgendes Lemma:

Lemma 2.1.4. *Wenn für die Folge ξ_i, $i=0,1,2,\dots$ Ungleichungen der Form*

$$|\xi_{i+1}| \le (1+\delta)|\xi_i| + B,\ \delta>0,\ B\ge 0$$

gelten, dann gilt für das n-te Folgenglied

$$|\xi_n| \le e^{n\delta}|\xi_0| + \frac{e^{n\delta}-1}{\delta}B.$$

Den Beweis zu Lemma 2.1.4 findet man zum Beispiel in [SB80], Lemma (7.2.2.2). Mit diesen Grundlagen kann man den folgenden Satz über die Konvergenz von stetigen Einschrittverfahren formulieren.

Satz 2.1.5. *Es sei das Anfangswertproblem* (AWP) *mit der Lösung* $\tilde{x}(t)$ *gegeben. Die Verfahrensfunktion* $\Phi(\cdot, g)$ *sei stetig auf*

$$G := \{(t, z, h, \theta) : \ 0 \le t \le T, \ \|z - \tilde{x}(t)\|_2 \le \gamma, \ 0 \le h \le h_0, \ 0 < \theta \le 1\}$$

für ein $h_0 > 0$ *und* $\gamma > 0$. *Außerdem gelte die Lipschitzbedingung*

$$\|\Phi(t, z_1, h, \theta, g) - \Phi(t, z_2, h, \theta, g)\|_2 \le M\|z_1 - z_2\|_2$$

für alle $(t, z_i, h, \theta) \in G, \ i = 1, 2$ *und*

$$\sup_{k=0,\ldots,N-1} \|r_1(t_k, \tilde{x}, h, g)\|_2 \le \eta_p h^p,$$
$$\sup_{k=0,\ldots,N-1} \sup_{0<\theta<1} \|r_\theta(t_k, \tilde{x}, h, g)\|_2 \le \eta_q h^q,$$

dann existiert ein $\overline{h}$ *mit* $0 < \overline{h} \le h_0$, *und es gilt für den globalen Diskretisierungsfehler* $e_h(t) := x_h(t) - \tilde{x}(t)$

$$\|e_h(t_k + \theta h)\|_2 \le (1 + hM)\frac{e^{t_k M} - 1}{M} \ \eta_p h^p + \eta_q h^{q+1}$$

für alle $h \le \bar{h}, \ k = 0, 1, \ldots, N-1$ *und* $\theta \in (0, 1]$.

Beweis: Wir definieren auf

$$\hat{G} := \{(t, z, h, \theta) : \ 0 \le t \le T, \ z \in \mathbb{R}^n, \ 0 \le h \le h_0, \ 0 < \theta \le 1\}$$

die stetige Funktion

$$\hat{\Phi}(t, z, h, \theta, g) := \Phi(t, \tilde{x}(t) + \tau(z - \tilde{x}(t)), h, \theta, g) \tag{2.1.3}$$

für alle $0 \le t \le T$, $h \le h_0$ und $0 < \theta \le 1$ mit $\tau := \min\{1, \frac{\gamma}{\|z - \tilde{x}(t)\|_2}\}$ für $z \ne \tilde{x}(t)$. Durch die Konstruktion von τ und die Verwendung der Euklidischen-Norm für die Vektoren bleibt dabei die Lipschitzbedingung

$$\|\hat{\Phi}(t, z_1, h, \theta, g) - \hat{\Phi}(t, z_2, h, \theta, g)\|_2 \le M\|z_1 - z_2\|_2 \tag{2.1.4}$$

für alle $(t, z_i, h, \theta) \in \hat{G}, \ i = 1, 2$ erhalten. Weiterhin bleibt

$$\|\triangle(t_k, \tilde{x}, h, g) - \hat{\Phi}(t_k, \tilde{x}(t_k), h, 1, g)\|_2 \le \eta_p h^p, \tag{2.1.5}$$

$$\|\theta\triangle(t_k, \tilde{x}, \theta h, g) - \hat{\Phi}(t_k, \tilde{x}(t_k), h, \theta, g)\|_2 \le \eta_q h^q, \ \theta \in (0, 1), \tag{2.1.6}$$

erfüllt, da $\hat{\Phi}(t_k, \tilde{x}(t_k), h, \theta, g) = \Phi(t_k, \tilde{x}(t_k), h, \theta, g)$ ist. Das durch $\hat{\Phi}$ definierte Einschrittverfahren liefere die Näherungslösungen $\hat{x}_h(t)$. Dann gilt für die Näherungslösungen

$$\hat{x}_h(t_k + \theta h) = \hat{x}_h(t_k) + h\hat{\Phi}(t_k, \hat{x}_h(t_k), h, \theta, g)$$

und aus dem Differenzenquotienten folgt

$$\tilde{x}(t_k + \theta h) = \tilde{x}(t_k) + \theta h \triangle(t_k, \tilde{x}, \theta h, g).$$

Daraus ergibt sich für den globalen Diskretisierungsfehler $\hat{e}_h(t_k + \theta h) := \hat{x}_h(t_k + \theta h) - \tilde{x}(t_k + \theta h)$ an der Stelle $t_k + \theta h$

$$\begin{aligned}\hat{e}_h(t_k + \theta h) &= \hat{e}_h(t_k) + h\left[\hat{\Phi}(t_k, \hat{x}_h(t_k), h, \theta, g) - \theta\triangle(t_k, \tilde{x}, \theta h, g)\right] \\ &= \hat{e}_h(t_k) + h\left[\hat{\Phi}(t_k, \hat{x}_h(t_k), h, \theta, g) - \hat{\Phi}(t_k, \tilde{x}(t_k), h, \theta, g)\right] \\ &\quad + h\left[\hat{\Phi}(t_k, \tilde{x}(t_k), h, \theta, g) - \theta\triangle(t_k, \tilde{x}, \theta h, g)\right]\end{aligned}$$

Aus der Lipschitzbedingung (2.1.4), Gleichung (2.1.6) und der Dreiecksungleichung folgt für $0 < \theta < 1$

$$\begin{aligned}\|\hat{e}_h(t_k + \theta h)\|_2 &\leq \|\hat{e}_h(t_k)\|_2 + hM\ \|\hat{x}_h(t_k) - \tilde{x}(t_k)\|_2 + h\eta_q h^q \\ &= (1 + hM)\|\hat{e}_h(t_k)\|_2 + \eta_q h^{q+1}\end{aligned}$$

und analog mit Gleichung (2.1.5)

$$\|\hat{e}_h(t_{k+1})\|_2 = \|\hat{e}_h(t_k + h)\|_2 \leq (1 + hM)\|\hat{e}_h(t_k)\|_2 + \eta_p h^{p+1}$$

für $\theta = 1$. Insgesamt erhält man somit eine rekursive Formel für den globalen Diskretisierungsfehler. Mit Hilfe von Lemma 2.1.4 erhält man

$$\|\hat{e}_h(t_k)\|_2 \leq e^{khM}\|\hat{e}_h(0)\|_2 + \frac{e^{khM} - 1}{hM}\ \eta_p h^{p+1} \tag{2.1.7}$$

$$= e^{t_k M}\|\hat{e}_h(0)\|_2 + \frac{e^{t_k M} - 1}{M}\ \eta_p h^p. \tag{2.1.8}$$

Mit $\hat{e}_h(0) = \hat{x}_h(0) - \tilde{x}(0) = 0$ folgt für den globalen Diskretisierungsfehler an der Stelle $t_k + \theta h$

$$\|\hat{e}_h(t_k + \theta h)\|_2 \leq (1 + hM)\frac{e^{t_k M} - 1}{M}\ \eta_p h^p + \eta_q h^{q+1}$$

für alle $t_k = kh,\ k = 0, 1 \ldots, N-1$, $\theta \in (0,\ 1]$ und $h \leq h_0$. Da $\gamma > 0$ ist, können wir ein $\bar{h}$ mit $0 < \bar{h} \leq h_0$ so wählen, dass $\|\hat{e}_h(t)\|_2 \leq \gamma$ für alle $h \leq \bar{h}$, $t \in [0, T]$ gilt. Für diesen Fall gilt $e_h = \hat{e}_h$, $x_h = \hat{x}_h$ und $\Phi(t_k, x_h(t_k), h, \theta, g) = \hat{\Phi}(t_k, \hat{x}_h(t_k), h, \theta, g)$. Daraus folgt für alle Punkte $t_k + \theta h \in [0, T]$ und alle $h \leq \bar{h}$ die Behauptung

$$\|e_h(t_k + \theta h)\|_2 \leq (1 + hM)\frac{e^{t_k M} - 1}{M}\ \eta_p h^p + \eta_q h^{q+1}.$$

$\Diamond$

Folgerung 2.1.6. *Berechnet man eine Näherungslösung für das Anfangswertproblem* (AWP) *mit den Startwerten* $t_0 = 0$ *und* $x_h(0) = a + \varepsilon_h$ *durch*

$$\begin{aligned}t_{k+1} &= t_k + h \\ x_h(t_k + \theta h) &= x_h(t_k) + h\Phi(t_k, x_h(t_k), h, \theta, g), \\ &\qquad \textit{für } k = 0, \ldots, N-1 \textit{ und } \theta \in (0, 1]\end{aligned}$$

mit einem stetigen Einschrittverfahren der Konsistenzordnung $(p\,,q)$ *und gilt*

$$\lim_{h\to 0}\|\varepsilon_h\|_2 = 0,$$

dann gibt es ein $\tilde{h}$ *mit* $0 < \tilde{h} \leq \bar{h}$ *so dass*

$$\|e_h(t_k+\theta h)\|_2 \leq (1+hM)\left[e^{t_k M}\|\varepsilon_h\|_2 + \frac{e^{t_k M}-1}{M}\,\eta_p h^p\right] + \eta_q h^{q+1}$$

für alle $h \leq \tilde{h}$, $k = 0,1,\ldots,N-1$ *und* $\theta \in (0,1]$ *gilt.*

Beweis: Der Beweis ergibt sich in dem man in die rechte Seite der Gleichung (2.1.8) $\hat{e}_h(0) = \hat{x}_h(0) - \tilde{x}(0) = \varepsilon_h$ einsetzt und analog zum Beweis von Satz 2.1.5 folgert. $\Diamond$

Damit hat man eine Grundlage, um auch bei einer Störung des Anfangswertes der adjungierten Zustandsgleichung eine Aussage über die Konvergenz eines Einschrittverfahrens zu haben. Um die Anwendung auf Probleme der optimalen Steuerung deutlich zu vereinfachen, werden wir uns im Folgenden nur noch auf SRKV als Vertreter eines solchen abstrakten stetigen Einschrittverfahrens beschränken.

2.2 Stetige-Runge-Kutta-Verfahren (SRKV)

Bei Runge-Kutta-Verfahren benutzt man im einfachsten Fall ebenfalls die Startwerte $t_0 = 0$ und $x_h(0) = a$ für eine Näherungslösung vom Problem (AWP) und berechnet iterativ

$$\begin{aligned} t_{j+1} &= t_j + h, \\ x_h(t_{j+1}) &= x_h(t_j) + h\sum_{k=1}^{s} b_k g(y_{j,k}, t_j + c_k h) \end{aligned}$$

für alle $j = 0,\ldots,N-1$. Dabei benötigt man Hilfspunkte $y_{j,k}$, die man aus dem Gleichungssystem

$$y_{j,k} = x_h(t_j) + h\sum_{l=1}^{s} a_{k,l}\, g(y_{j,l}, t_j + c_l h) \tag{2.2.1}$$

berechnet, und es sind für das Verfahren die Parameter $a_{k,l}$, $c_k = \sum_{l=1}^{s} a_{k,l}$ und b_k für alle $k,l = 1,\ldots,s$ gegeben. Ein solches Verfahren hat üblicherweise $s \geq 2$ Stufen, und die notwendigen Parametersätze für $a_{k,l}$ kann man in einer $s \times s$ Matrix anordnen. Ist dabei diese Matrix nur unterhalb der Hauptdiagonalen besetzt, so spricht man von einem expliziten Runge-Kutta-Verfahren, und man kann die Gleichung für die Hilfspunkte $y_{j,k}$ leicht nacheinander lösen (siehe zum Beispiel [But03, HNW93]).

Ersetzt man die Parameter b_k durch Polynome $b_k(\theta)$, $\theta \in (0,1]$ mit

$$\lim_{\theta\to 0} b_k(\theta) = 0 \text{ und } b_k(1) = b_k,$$

erhält man ein stetiges Einschrittverfahren (siehe [Zen86, VZ95a, HNW93]). Man berechnet mit $t_0 = 0$ und $x_h(0) = a$ zu jedem Zeitpunkt $t \in (0, T]$ eine Näherungslösung $x_h(t) = x_h(t_j + \theta h)$ durch

$$\begin{aligned} t_{j+1} &= t_j + h, \\ x_h(t_j + \theta h) &= x_h(t_j) + h \sum_{k=1}^{s} b_k(\theta) g(y_{j,k}, t_j + c_k h), \end{aligned}$$

wobei man weiterhin für jedes $j = 0, 1, \ldots, N-1$ die s Hilfspunkte $y_{j,k}$ aus dem Gleichungssystem 2.2.1 berechnet. Somit benutzt man

$$\begin{aligned} \Phi(t_k, z, h, \theta, g) &= \sum_{k=1}^{s} b_k(\theta) g(y_{j,k}, t_j + c_k h) \text{ mit} \\ y_{j,k} &= z + h \sum_{l=1}^{s} a_{k,l}\, g(y_{j,l}, t_j + c_l h) \end{aligned}$$

formal als Verfahrensfunktion eines stetigen Einschrittverfahrens.

2.2.1 Konsistenzfehler

Wir nehmen im Folgenden an, dass wir ein SRKV der Ordnung $(p\,, q)$ gegeben haben, welches die folgenden Konsistenzfehler

$$\begin{aligned} \sup_{j=0,\ldots,N-1} \|r_1(t_j, \tilde{x}, h, g)\|_2 &\leq \eta_p h^p \\ \sup_{j=0,\ldots,N-1} \sup_{0<\theta<1} \|r_\theta(t_j, \tilde{x}, h, g)\|_2 &\leq \eta_q h^q \end{aligned}$$

mit von der Schrittweite h unabhängigen Konstanten η_p und η_q besitzt. Dies bedeutet, es gilt

$$\begin{aligned} \sup_{j=0,\ldots,N-1} \left\| \frac{\tilde{x}(t_j + h) - \tilde{x}(t_j)}{h} - \sum_{k=1}^{s} b_k(1) g(y_{j,k}, t_j + c_k h) \right\|_2 &\leq \eta_p h^p, \\ \sup_{j=0,1,\ldots,N-1} \sup_{0<\theta<1} \left\| \frac{\tilde{x}(t_j + \theta h) - \tilde{x}(t_j)}{h} - \sum_{k=1}^{s} b_k(\theta) g(y_{j,k}, t_j + c_k h) \right\|_2 &\leq \eta_q h^q. \end{aligned} \tag{2.2.2}$$

Für einige Aussagen ist diese Formulierung des Konsistenzfehlers außerhalb der Gitterpunkte ungünstig. Sieht man sich die Herleitung der Ordnungsbedingungen genauer an, fällt auf, dass sich statt der zweiten Bedingung mit den selben Konsistenzgleichungen bzw. Ordnungsbedingungen ebenfalls

$$\sup_{j=0,1,\ldots,N-1} \sup_{0<\theta<1} \left\| \frac{\tilde{x}(t) - \tilde{x}(t_j)}{\theta h} - \sum_{k=1}^{s} \frac{b_k(\theta)}{\theta} g(y_{j,k}, t_j + c_k h) \right\|_2 \leq \eta_q h^q \tag{2.2.3}$$

zeigen lässt. Dabei stört es nicht, dass die Differenz in der Norm durch $\theta > 0$ geteilt wurde.

Im Folgenden skizzieren wir kurz einen einfachen Weg, um die Konsistenzgleichungen für ein Verfahren der Ordnung $(2,2)$ herzuleiten. Wir nehmen an, dass die Lösung $\tilde{x} \in C^\kappa([0,T],\mathbb{R}^n)$ des Anfangswertproblems (AWP) mit einem hinreichend großem κ gegeben ist, und können damit für $\tilde{x}(t) = \tilde{x}(t_j + \theta h)$ eine Taylor-Entwicklung

$$\begin{aligned}\tilde{x}(t_j+\theta h) = \tilde{x}(t_j) + \dot{\tilde{x}}(t_j)\theta h + \frac{\ddot{\tilde{x}}(t_j)}{2}(\theta h)^2 \\ + \frac{\tilde{x}^{(3)}(t_j)}{6}(\theta h)^3 + \frac{1}{6}\int_{t_j}^{t_j+\theta h} \tilde{x}^{(4)}(\tau)(\tau - t_j)^3 d\tau.\end{aligned}$$

aufstellen. Da $\tilde{x}$ Lösung von (AWP) ist, gilt

$$\dot{\tilde{x}}(t_j) = g(\tilde{x}(t_j), t_j)$$

und daraus ergibt sich für $\ddot{\tilde{x}}$ und $\tilde{x}^{(3)}$

$$\begin{aligned}\ddot{\tilde{x}}(t_j) &= g_x(\tilde{x}(t_j),t_j)(\dot{\tilde{x}}(t_j)) + g_t(\tilde{x}(t_j),t_j) \\ &= g_x(\tilde{x}(t_j),t_j)(g(\tilde{x}(t_j),t_j)) + g_t(\tilde{x}(t_j),t_j) \\ \tilde{x}^{(3)}(t_j) &= g_{xx}(\tilde{x}(t_j),t_j)(g(\tilde{x}(t_j),t_j), g(\tilde{x}(t_j),t_j)) + 2g_{xt}(\tilde{x}(t_j),t_j)(g(\tilde{x}(t_j),t_j)) \\ &\quad + g_{tt}(\tilde{x}(t_j),t_j) + g_x(\tilde{x}(t_j),t_j)(g_x(\tilde{x}(t_j),t_j)(g(\tilde{x}(t_j),t_j))) \\ &\quad + g_x(\tilde{x}(t_j),t_j)(g_t(\tilde{x}(t_j),t_j)) + g_{tt}(\tilde{x}(t_j),t_j).\end{aligned}$$

Dabei sind Abbildungen der Form $g_{xx}(g,g)$ die entsprechend passend definierten, in beiden Argumenten linearen Ableitungen. Es berechnet sich zum Beispiel die i-te Komponente durch

$$(g_{xx}(g,g))_i := \sum_{l,k=1}^{n} \frac{\partial^2 g_i}{\partial x_k \partial x_l} g_k g_l,$$

wobei g_i, g_k und g_l hier jeweils die entsprechende Komponente der vektorwertigen Funktion $g(\tilde{x}(t_j),t_j)$ bezeichnet. Im zweiten Term müssen wir $g(y_{j,k},t_{j,k})$ abschätzen. Hierzu benutzen wir eine Taylor-Entwicklung von g bezüglich des Punktes $(\tilde{x}(t_j),t_j)$. Dabei ergibt sich unter Berücksichtigung von

$$\begin{aligned}y_{j,k} &= \tilde{x}(t_j) + h\sum_{l=1}^{s} a_{k,l}\, g(y_{j,l},t_{j,l}) \text{ und} \\ \gamma_{j,k} &:= \sum_{l=1}^{s} a_{k,l}\, g(y_{j,l},t_{j,l})\end{aligned}$$

als Taylor-Entwicklung für g zum Beispiel:

$$g(y_{j,k}, t_{j,k}) = g(\tilde{x}(t_j), t_j) + g_x(\tilde{x}(t_j), t_j)(h\gamma_{j,k}) + g_t(\tilde{x}(t_j), t_j)(c_k h) + \frac{1}{2}\int_0^1 g_{xx}(\tilde{x}(t_j), t_j)(\vartheta h\gamma_{j,k}, \vartheta h\gamma_{j,k}) + 2g_{xt}(\tilde{x}(t_j), t_j)(\vartheta h\gamma_{j,k}, \vartheta c_k h) + g_{tt}(\tilde{x}(t_j), t_j)(\vartheta c_k h, \vartheta c_k h)\, d\vartheta. \tag{2.2.4}$$

Diese Taylorreihe und die Taylor-Entwicklung für $\tilde{x}(t_j + \theta h)$ kann man in Gleichung (2.2.3) einsetzen und Bedingungen an die Parameter $a_{k,l}$ und die Polynome $b_k(\theta)$ so stellen, dass Terme, in denen die Schrittweite h in einer relativ geringen Potenz vorkommt, verschwinden. Fassen wir dabei die Terme, in denen die Schrittweite h nicht vorkommt, zusammen, erhalten wir:

$$g(\tilde{x}(t_j), t_j) - \frac{1}{\theta}\sum_{k=1}^{s} b_k(\theta) g(\tilde{x}(t_j), t_j).$$

Da der Funktionswert $g(\tilde{x}(t_j), t_j)$ unabhängig von k ist, ist es offensichtlich egal, wenn man die ursprüngliche Abschätzung des Konsistenzfehlers zwischen den Gitterpunkten durch θ teilt, und dieser Term verschwindet, wenn

$$\sum_{k=1}^{s} b_k(\theta) = \theta \tag{OB1}$$

erfüllt ist. Diese Bedingung wird auch allgemein als Konsistenz des Verfahrens bezeichnet, und wir werden sie später mehrfach benutzen.

Die Summe aller Terme, in denen h nur in der ersten Potenz steht, ist

$$\frac{\theta}{2}[g_x(\tilde{x}(t_j), t_j)(g(\tilde{x}(t_j), t_j)) + g_t(\tilde{x}(t_j), t_j)] - \frac{1}{\theta}\sum_{k=1}^{s} b_k(\theta)[g_x(\tilde{x}(t_j), t_j)(\gamma_{j,k}) + g_t(\tilde{x}(t_j), t_j)(c_k)].$$

Hier stört jedoch für einen Vergleich der beiden Teile der Ausdruck $\gamma_{j,k}$. Um diese Terme vergleichen zu können, kann man den „bootstrapping“Ansatz verfolgen (siehe [Her04]) und in die Definition von $\gamma_{j,k}$ wieder die Taylor-Entwicklung (2.2.4) für $g(y_{j,l}, t_{j,l})$ einsetzen. Dabei entstehen wieder Terme, in denen die Schrittweite h vorkommt, die man also dann bei den Termen mit höherer Potenz von h berücksichtigen muss. Da der einzige Term in Gleichung (2.2.4), in dem die Schrittweite h nicht vorkommt, $g(\tilde{x}(t_j), t_j))$ ist, erhält man als nächste Ordnungsbedingung

$$\sum_{k=1}^{s} c_k b_k(\theta) = \frac{\theta^2}{2}. \tag{OB2}$$

Dabei wurde die Definition $c_k = \sum_{l=1}^{s} a_{k,l}$ benutzt, um die Summe aus der Definition von $\gamma_{j,k}$ zu vereinfachen.

Da das Aufstellen von Konsistenzgleichungen für Verfahren mit höherer Ordnung immer komplexer wird, verweisen wir für eine systematische Herangehensweise auf die, in vielen Lehrbüchern beschriebene, Herleitung mit Hilfe von Wurzel-Bäumen (siehe zum Beispiel [But03, HNW93]). Diese Systematik wurde ursprünglich für diskrete Runge-Kutta-Verfahren entwickelt und kann sehr leicht für SRKV übernommen werden (siehe [OZ91, OZ92]). Man muss lediglich bei der Entwicklung konkreter SRKV aufpassen, dass sich bestimmte Vereinfachungen der Konsistenzgleichungen nicht analog zum diskreten Fall ergeben. Die stetige Version der Konsistenzgleichungen lautet nach [OZ91, OZ92]

$$\sum_{k=1}^{s} b_k(\theta)\phi_j(\mathcal{B}) = \frac{\theta^{\rho(\mathcal{B})}}{\gamma(\mathcal{B})} \text{ für alle Bäume } \mathcal{B} \text{ mit } \rho(\mathcal{B}) \leq q, \tag{2.2.5}$$

dabei ist $\phi_j(\mathcal{B})$ das j-te elementare Gewicht des Baums $\mathcal{B}$, $\rho(\mathcal{B})$ die Ordnung von $\mathcal{B}$ und $\gamma(\mathcal{B})$ ein Koeffizient, der von dem Baum $\mathcal{B}$ abhängt, siehe [But03, HNW93, OZ91, OZ92]. Die Konsistenzgleichungen (2.2.5) liefern damit Verfahren, die (2.2.3), mit passenden Konstanten q und η_q, erfüllen.

Weiterhin gilt mit derselben Herleitung für die Einschrittfehler auch

$$\begin{aligned}
\sup_{j=0,\ldots,N-1} \left\| \tilde{x}(t_j + h) - \tilde{x}(t_j) - h \sum_{k=1}^{s} b_k(1) g(y_{j,k}, t_j + c_k h) \right\|_2 &\leq \eta_p h^{p+1} \\
\sup_{j=0,1,\ldots,N-1} \sup_{0<\theta<1} \left\| \tilde{x}(t_j + \theta h) - \tilde{x}(t_j) - h \sum_{k=1}^{s} b_k(\theta) g(y_{j,k}, t_j + c_k h) \right\|_2 &\leq \eta_q h^{q+1}.
\end{aligned}$$

Gehen wir in dieser Darstellung davon aus, dass ein SRKV gegeben ist, welches so konstruiert ist, dass sich die ersten Terme einer Taylor-Entwicklung aufheben, dann kann man den Ausdruck

$$\tilde{x}(t) - \tilde{x}(t_j) - h \sum_{k=1}^{s} b_k \left(\frac{t - t_j}{h} \right) g(y_{j,k}, t_j + c_k h) \tag{2.2.6}$$

durch die Restterme beider Taylor-Entwicklungen in Integralform darstellen. Man erhält für ein Verfahren der Ordnung q mit einer Funktion $\mathcal{R}_q$ eine Darstellung der Art

$$\tilde{x}(t) - \tilde{x}(t_j) - h \sum_{k=1}^{s} b_k \left(\frac{t - t_j}{h} \right) g(y_{j,k}, t_j + c_k h) = \int_{t_j}^{t} \mathcal{R}_q(\tau)(\tau - t_j)^q \, d\tau.$$

Dabei muss die Funktion $\mathcal{R}_q$ wesentlich beschränkt sein, damit man die Abschätzungen (2.2.2) und (2.2.3) für beliebig kleine Schrittweiten h mit unabhängigen Konstanten η_p und η_q aufstellen kann. Damit ist offensichtlich, dass man eine zu der Gleichung (2.2.3) äquivalente Abschätzung erhält, wenn man nicht den Differenzenquotienten durch die Division der Gleichung (2.2.6) durch θh bildet, sondern

die Gleichung (2.2.6) im Intervall (t_j, t_{j+1}) nach t ableitet. Dabei verliert man in beiden Fällen jeweils eine Ordnung für die Konvergenz der entsprechenden Terme, und es gilt

$$\sup_{j=0,1,\ldots,N-1} \operatorname*{ess\,sup}_{0<\theta<1} \left\| \dot{\tilde{x}}(t_j + \theta h) - \sum_{k=1}^{s} b_k{}'(\theta) g(y_{j,k}, t_j + c_k h) \right\|_2 \leq \tilde{\eta}_q h^q, \tag{2.2.7}$$

wobei b_k' die Ableitung der Polynome $b_k'(\theta)$ nach θ bezeichnet. Diese Betrachtungsweise hat natürlich nur für SRKV Sinn, und der entsprechende Restterm $\mathcal{R}_q$ kann für Verfahren mit hoher Konvergenzordnung sehr kompliziert sein. Daher lohnt es sich nicht, ihn ohne eine systematische Herangehensweise mit Hilfe von Bäumen darstellen zu wollen.

2.2.2 Parametersätze

Satz 2.1.5 zeigt, dass jedes Runge-Kutta-Verfahren mit mindestens quadratischer Ordnung auch ohne die explizite Existenz einer stetigen Erweiterung durch einfache lineare Interpolation formal zu einem SRKV der Ordnung zwei wird. Es ergeben sich die Polynome $b_k(\theta)$ aus den Parametern b_k aus dem Runge-Kutta-Tableau (auch Butcher-Tableau genannt) durch $b_k(\theta) = b_k\theta$ mit $0 < \theta \leq 1$. Man beachte, dass dabei die Bedingung (OB2) im Allgemeinen nicht erfüllt ist, sondern nur die diskrete Variante $\sum_{k=1}^{s} c_k b_k = \frac{1}{2}$ und erst der Satz 2.1.5 ein stetiges Verfahren der Ordnung zwei zusichert.

Mit Hilfe der Bedingungen (OB1) und (OB2) lassen sich sehr leicht alle expliziten SRKV mit 2 Stufen und Ordnung $(2, 2)$ in Abhängigkeit von dem Parameter $0 < c_2 \leq 1$ beschreiben. Damit das 2-stufige Verfahren ein explizites Verfahren wird, muss offensichtlich $a_{1,1} = a_{1,2} = a_{2,2} = 0$, $c_1 = 0$ und $a_{2,1} = c_2$ gelten. Aus den Ordnungsbedingungen ergeben sich die Gleichungen

$$b_1(\theta) + b_2(\theta) = \theta \quad \text{und} \quad c_2 b_2(\theta) = \frac{\theta^2}{2}.$$

Die zweite Gleichung kann man offensichtlich für $c_2 \neq 0$ nach $b_2(\theta)$ auflösen und damit die Polynome $b_1(\theta) = \theta - \frac{\theta^2}{2c_2}$ und $b_2(\theta) = \frac{\theta^2}{2c_2}$ bestimmen.

Parametersätze für stetige Verfahren höherer Ordnung zu erzeugen, ist im Allgemeinen sehr kompliziert. Dabei werden entweder komplett neue Verfahren konstruiert, die die Konsistenzgleichungen (2.2.5) erfüllen, oder es werden zu existierenden Runge-Kutta-Verfahren stetige Erweiterungen, also Polynome $b_k(\theta)$, gesucht, die eine möglichst gute Interpolation zwischen den Gitterpunkten garantieren. Dazu werden oft bestehende Runge-Kutta-Verfahren um einige Stufen erweitert, um eine bessere Interpolation konstruieren zu können. Nach Satz 2.1.5 ist dabei $q = p - 1$ sinnvoll, damit man insgesamt ein stetiges Verfahren der Ordnung p bekommt.

Neben den grundlegende Arbeiten zu SRKV [Sha85, EJNT86, Zen86, VZ95b] findet man konkrete Parametersätze zum Beispiel in [OZ91, OZ92, Ver93, CT96] und auf der Internetseite von Jim Verner [Ver].

3 Stetige Näherungslösungen von Steuerungsproblemen

In diesem Kapitel zeigen wir, dass die SRKV sich sehr gut eignen, um damit bestimmte Probleme der optimalen Steuerung zu diskretisieren. Dazu betrachten wir folgendes Steuerungsproblem:

$$\begin{array}{lll} \text{(OS)} & \min & \Psi(x(T)) \\ & u.Nb. & \dot{x}(t) = f(x(t), u(t)) \quad \forall' t \in [0,T], \\ & & x(0) = a. \end{array}$$

Es sei $u(t) \in \mathbb{R}^m$ die Steuerung und $x(t) \in \mathbb{R}^n$ der Zustand zur Zeit t. Weiter seien $\Psi : \mathbb{R}^n \longrightarrow \mathbb{R}$, $f : \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n$ und $a \in \mathbb{R}^n$ gegeben. Mit $\dot{x}$ bezeichnen wir die schwache Ableitung $\frac{d}{dt}x$ und mit $L^\alpha(J;\mathbb{R}^m)$ bezeichnen wir den L_α-Raum von Funktionen $u : J \longrightarrow \mathbb{R}^m$ für die $\|u(\cdot)\|_2^\alpha$ Lebesgue-integrierbar ist. Für diesen Raum benutzen wir die Standardnorm:

$$\|u\|_{L^\alpha} = \left(\int_J \|u(t)\|_2^\alpha \, d\mu(t) \right)^{\frac{1}{\alpha}} \tag{3.0.1}$$

wobei alle Funktionen einer Äquivalenzklasse zugeordnet werden, die fast überall gleich sind. Dabei bezeichnet $\|\cdot\|_2$ die Euklidische Norm. Analog erhält man den Raum L^∞ der wesentlich beschränkten Funktionen. Weiterhin bezeichnen wir mit $W^{k,\alpha}(J;\mathbb{R}^m)$ den Sobolev-Raum von vektorwertigen Funktionen deren j-te Ableitung für alle $0 \leq j \leq k$ im Raum L^α liegt. Die dazugehörige Norm sei

$$\|u\|_{W^{k,\alpha}} = \sum_{j=0}^{k} \|u^{(j)}\|_{L^\alpha} \tag{3.0.2}$$

und $u^{(j)}$ bezeichnet die j-te schwache Ableitung der Funktion $u : J \longrightarrow \mathbb{R}^m$.

Natürlich kann man Zielfunktionen der Art $\int_0^T g(t, x(t), u(t))\, dt$ mit Hilfe von weiteren Zuständen auf die Form von Problem (OS) umschreiben. Durch die hier verwendete Schreibweise des Steuerungsproblems muss man ein solches Integral nicht explizit numerisch lösen, sondern hierzu wird ebenfalls das SRKV benutzt. Wir betrachten ausschließlich Steuerungsprobleme mit Startzeit 0 und fester Endzeit T, die außer der Zustandsgleichung an sich keine weiteren Zustands- oder Steuerungsbeschränkungen haben.

3.1 Hinreichende Voraussetzungen für die Existenz einer Näherungslösung

Wir stellen zunächst die Voraussetzungen zusammen, die notwendig sind, um in den nächsten Abschnitten die Existenz einer numerischen Lösung auf der Basis von SRKV zeigen zu können. Dabei setzen wir zunächst voraus, dass das Steuerungsproblem (OS) eine Lösung hat und diese folgende Eigenschaften erfüllt:
***Glattheit*:** *Für eine natürliche Zahl $\kappa \geq 3$ hat das Problem* (OS) *eine lokale Lösung (x^*, u^*) in $W^{\kappa,\infty}([0,T];\mathbb{R}^n) \times W^{\kappa-1,\infty}([0,T];\mathbb{R}^m)$. Es existiere eine offene Menge $\Omega \subset \mathbb{R}^n \times \mathbb{R}^m$, so dass es ein $\rho > 0$ gibt und die Kugeln $B_\rho(x^*(t), u^*(t))$ für alle $t \in [0,T]$ Teilmenge von Ω sind. Weiterhin seien die k-ten Ableitungen von f für $k = 0, 1, \ldots, \kappa$ in Ω lipschitzstetig bezüglich $x(t)$ und $u(t)$ mit Lipschitzkonstante M_f^k und die k-te Ableitung von Ψ für $k = 0, 1, \ldots, \kappa$ sei in der Kugel $B_\rho(x^*(T))$ lipschitzstetig mit Lipschitzkonstante M_Ψ^k.*
Mit diesen Voraussetzungen gibt es einen zu (x^*, u^*) gehörenden adjungierten Zustand $\lambda^* \in W^{\kappa,\infty}([0,T];\mathbb{R}^n)$, und $\lambda^*(t)$ ist damit auf dem Intervall $[0,T]$ wesentlich beschränkt (siehe zum Beispiel [PBGM64]). Als Abkürzung fassen wir im Folgenden den Zustand, adjungierten Zustand und die Steuerung häufig als neue Funktion $w(t) := (x(t), \lambda(t), u(t))$ zusammen und definieren damit die Hamilton-Funktion

$$H(w(t)) := \lambda(t)^\top f(x(t), u(t)) = f(x(t), u(t))^\top \lambda(t).$$

Weiterhin sei $w^* = (x^*, \lambda^*, u^*)$, und wir definieren die folgenden Abkürzungen

$$\begin{aligned} A(t) &= f_x(x^*(t), u^*(t)) & B(t) &= f_u(x^*(t), u^*(t)) & V &= \Psi_{xx}(x^*(T)) \\ Q(t) &= H_{xx}(w^*(t)) & R(t) &= H_{uu}(w^*(t)) & S(t) &= H_{ux}(w^*(t)). \end{aligned}$$

Mit f_x bezeichnen wir dabei die Ableitung $\frac{\partial f}{\partial x}$. Damit ist $f_x(x^*(t), u^*(t))$ eine lineare Abbildung von $\mathbb{R}^n$ in $\mathbb{R}^n$ und $B(t) : \mathbb{R}^m \to \mathbb{R}^n$ linear, das heißt wir können diese Funktionen zu jedem Zeitpunkt als Matrix $A(t) \in \mathbb{R}^{n\times n}$ und $B(t) \in \mathbb{R}^{n\times m}$ auffassen. Die Ableitungen $H_x(w(t))$ und $H_u(w(t))$ der Funktion $H(w(t))$ sind damit zu jedem Zeitpunkt Vektoren aus $\mathbb{R}^n$ bzw. $\mathbb{R}^m$. Die linearen Abbildungen V, $Q(t)$, $R(t)$, $S(t)$ wollen wir ebenfalls als Matrizen darstellen, wobei zum Beispiel für den Eintrag an der Stelle (i,j) der Matrix $S(t)$ gilt $(S(t))_{i,j} = \frac{\partial H}{\partial x_i \partial u_j}(w^*(t))$, und somit sind V, $Q(t) \in \mathbb{R}^{n\times n}$, $R(t) \in \mathbb{R}^{m\times m}$ und $S(t) \in \mathbb{R}^{n\times m}$. Entsprechend fassen wir allgemein die zweiten Ableitungen von $H(w(t))$ als Matrizen auf. Mit Hilfe dieser Matrizen definieren wir

$$\begin{aligned} \mathcal{B}(x,u) := \frac{1}{2} x(T)^\top V x(T) \\ + \frac{1}{2} \int_0^T x(t)^\top Q(t) x(t) + u(t)^\top R(t) u(t) + 2x(t)^\top S(t) u(t)\, dt \end{aligned}$$

Entsprechend den Resultaten in [Hag00, DHV98] benötigen wir ebenfalls eine starke hinreichende Optimalitätsbedingung 2. Ordnung.

***Koerzivität*:** *Mit der Menge*

$$\begin{aligned}\mathcal{C} := \{(x,u) : x \in W^{1,2}([0,T],\mathbb{R}^n),\, u \in L^2([0,T],\mathbb{R}^m),\\ \dot{x}(t) = A(t)x(t) + B(t)u(t),\, x(0) = 0,\, t \in [0,T]\}\end{aligned}$$

gelte

$$\mathcal{B}(x,u) \geq \alpha \left(\|x\|_{H^1}^2 + \|u\|_{L^2}^2\right) \quad \forall (x,u) \in \mathcal{C}$$

mit einer Konstanten $\alpha > 0$.
Hierbei benutzen wir für den Hilbertraum $W^{1,2}([0,T],\mathbb{R}^n)$ zur Verdeutlichung die Schreibweise $H^1([0,T],\mathbb{R}^n)$. Da natürlich jede Funktion aus $W^{1,\infty}([0,T],\mathbb{R}^n)$ oder $L^\infty([0,T],\mathbb{R}^m)$ auch in den Räumen $H^1([0,T],\mathbb{R}^n)$ bzw. $L^2([0,T],\mathbb{R}^m)$ ist, kann man hier die Normen $\|\cdot\|_{H^1}$ und $\|\cdot\|_{L^2}$ benutzen. Dies ist für den Beweis des Satzes 3.3.3 notwendig. Die Bedingung *Koerzivität* ist eine starke hinreichende Optimalitätsbedingung 2. Ordnung und identisch zu der Bedingung „Coercivity“ aus [Hag00], da man für $(x,u) \in \mathcal{C}$ den Term $\|x\|_{H^1}^2$ mit Hilfe von $\|u\|_{L^2}^2$ nach oben abschätzen kann (siehe Abschnitt 3.5).

3.2 Diskretisierung der notwendigen Bedingungen 1. Ordnung

In den Arbeiten [DH01, DHV98, Hag00] wird das Problem (OS) diskretisiert und anschließend werden zu dem diskretisierten Problem die Optimalitätsbedingungen 1. Ordnung aufgestellt. Mit Hilfe des diskretisierten Problems kann man damit sowohl direkte als auch indirekte Verfahren konstruieren. Dieses Vorgehen ist für den Beweis einer hohen Konvergenzordnung bei Verwendung von SRKV ungünstig, da sich die Diskretisierung der adjungierten Zustände des stetigen Problems von den diskreten adjungierten Zuständen aus dem diskretisierten Steuerungsproblems unterscheiden. Daher diskretisieren wir das System der notwendigen Bedingungen und zeigen, dass unter den Bedingungen *Glattheit* und *Koerzivität* aus Abschnitt 3.1 dieses diskretisierte System eine Lösung besitzt. Somit untersuchen wir die Grundlage für indirekte Verfahren, bei denen die gewöhnlichen Differentialgleichungen der notwendigen Bedingungen mit Hilfe von SRKV diskretisiert werden. Als notwendige Optimalitätsbedingungen 1. Ordnung für das Problem (OS) ergeben sich die Gleichungen (siehe [PBGM64, LW07])

$$\text{(FC)} \qquad \begin{aligned} \dot{x}(t) &= f(x(t),u(t)) && \text{mit} && x(0) = a,\\ \dot{\lambda}(t) &= -f_x(x(t),u(t))^\top \lambda(t) && \text{mit} && \lambda(T) = \Psi_x(x(T))^\top,\\ 0 &= \lambda(t)^\top f_u(x(t),u(t)), && \forall t \in [0,T]. \end{aligned}$$

Dabei bezeichnet man die erste Gleichung als Zustandsgleichung, die zweite als adjungierte Gleichung und die dritte mitunter nach [PBGM64] als Minimumprinzip.

Wir wollen diese System von Gleichungen mit einem SRKV diskretisieren. Dabei gehen wir davon aus, dass ein fester Endzeitpunkt T gegeben ist, und zerlegen das Intervall $[0,T]$ in N gleich große Intervalle mit Hilfe von insgesamt $N+1$ äquidistanten Gitterpunkten. Die Schrittweite dieser Diskretisierung $h > 0$ berechnet sich

dann aus $h = \frac{T}{N}$. Der erste Gitterpunkt soll auf $t_0 = 0$ und der letzte auf $t_N = T$ fallen, so dass die Gitterpunkte $t_j = jh$, für $j = 0, \ldots, N$, sind. Somit betrachten wir zur Vereinfachung nur Schrittweiten $h \in H$ mit

$$H := \left\{ h : h = \frac{T}{k},\, k \in \mathbb{N} \right\}.$$

Jedem gegebenen Zeitpunkt $t \in (0, T]$ ordnen wir einen Gitterpunkt $t_j := t_j(t) := jh$ durch $j = \lceil \frac{t}{h} - 1 \rceil := \min\{k \in \mathbb{Z} : \frac{t}{h} - 1 \leq k\}$ so zu, dass $jh < t \leq (j+1)h$ gilt. Daraus erhält man für jedes $t \in (0, T]$, $t \neq t_j$ die Darstellung $t = t_j + \theta h$ mit $\theta = \frac{t - t_j}{h} > 0$.

Zur Vereinheitlichung der Diskretisierung und der folgenden Abschätzungen wollen wir in den adjungierten Gleichungen die Substitution $\bar{\lambda}(t) = \lambda(T - t)$ und $\dot{\bar{\lambda}}(t) = -\dot{\lambda}(T - t)$ durchführen und zur Vereinfachung $\bar{\lambda}$ wieder mit λ bezeichnen. Anschließend stellen wir durch

$$\begin{aligned}
\text{(DFC)} \quad x(t_j + \theta h) &= x(t_j) + h \sum_{k=1}^{s} b_k(\theta) f(y_{j,k}, u(t_j + c_k h)) \\
&\qquad \forall \theta \in (0, 1],\, \forall j = 0, \ldots, N-1 \text{ mit } x(t_0) = x(0) = a \\
\lambda(t_j + \theta h) &= \lambda(t_j) + h \sum_{k=1}^{s} b_k(\theta) f_x(x(T - t_j - c_k h), u(T - t_j - c_k h))^\top v_{j,k} \\
&\qquad \forall \theta \in (0, 1],\, \forall j = 0, \ldots, N-1 \text{ mit } \lambda(0) = \Psi_x(x(T))^\top \\
0 &= \lambda(T - t)^\top f_u(x(t), u(t)).
\end{aligned}$$

das diskretisierte System der notwendigen Optimalitätsbedingungen auf. Es berechnen sich die Zwischenstellen $y_{j,k}$ der Zustände x und $v_{j,k}$ der adjungierten Zustände λ wie folgt:

$$y_{j,k} = x(t_j) + h \sum_{l=1}^{s} a_{k,l} f(y_{j,l}, u(t_{j,l})) \tag{3.2.1}$$

$$v_{j,k} = \lambda(t_j) + h \sum_{l=1}^{s} a_{k,l} f_x(x(\bar{t}_{j,l}), u(\bar{t}_{j,l}))^\top v_{j,l}, \tag{3.2.2}$$

wobei $t_{j,k} := t_j + c_k h$ und $\bar{t}_{j,k} := T - t_j - c_k h$ sind.

Da man für die Zustandsgleichung einen Anfangswert und für die Gleichung der adjungierten Zustände einen Endwert gegeben hat, muss man die Zustände bei der Berechnung der adjungierten Zustände zu Zeitpunkten $T - t_j - c_k h$ auswerten, die im Allgemeinen bei Runge-Kutta-Verfahren mit hoher Konvergenzordnung nicht genau auf Zeitpunkte $t_{\tilde{j}} + c_{\tilde{k}} h$ fallen. Die Berechnung von Zuständen an beliebigen Zeitpunkten mit einer hohen Konvergenzordnung wird letztlich erst durch die stetigen Erweiterungen der Runge-Kutta-Verfahren ermöglicht.

Man kann die Zustände und adjungierten Zustände zu jeder gegebener Schrittweite h aus der Steuerung an endlich vielen Zeitpunkten berechnen. Insgesamt reicht es für alle folgenden Rechnungen, die Steuerung, Zustände und adjungierten Zustände jeweils an endlich vielen Zeitpunkten auszuwerten. Dabei sind dann stetige

Funktionen eindeutig durch die stetigen Erweiterungen und das Minimumprinzip festgelegt. Wir werden, vor allem um eine größer Übersichtlichkeit der entstehenden Gleichungen zu erhalten, dennoch stetige Funktionen suchen, die das Gleichungssystem (DFC) erfüllen. Es stellt sich die Frage, ob zu einer vorgegeben Anzahl von Diskretisierungspunkten Lösungen x_h, u_h und λ_h existieren und vor allem welchen Abstand diese Näherungslösungen von Lösungen des Gleichungssystems (FC) haben.

3.3 Anwendung des Fixpunktsatzes

In den folgenden Abschnitten wird die Existenz einer Lösung des Gleichungssystems (DFC) gezeigt. Im Gegensatz zur Diskretisierung des Optimierungsproblems in [Hag00] mit diskreten Runge-Kutta-Verfahren wird man dabei feststellen, dass man keine weiteren Forderungen an das verwendete SRKV stellen muss. Um dies zu zeigen, benutzen wir für beliebige fest gewählte hinreichend kleine Schrittweiten $h \in H$ folgendes Lemma, welches eine einfache Anwendung des Fixpunktsatzes von Banach ist. Die Formulierung orientiert sich dabei an Lemma 2.1 in [DH98]:

Lemma 3.3.1. *Es sei X eine abgeschlossene Teilmenge eines vollständigen normierten Vektorraums $\mathcal{X}$ und $\mathcal{Y}$ ein normierten Vektorraum. Weiterhin sei ein $w^* \in X$ und $r > 0$ gegeben, und wir definieren die Menge $X_r := B_r(w^*) \cap X$, wobei wir mit $B_r(w^*)$ die abgeschlossene Kugel mit Radius r um den Punkt w^* bezeichnen. Außerdem fordern wir dass die Abbildungen $P : X_r \to \mathcal{Y}$ und $F : X \to \mathcal{Y}$ die folgenden Eigenschaften erfüllen:*

(B1) Die Abbildung F^{-1} existiert, für jedes $\pi \in \mathcal{Y}$ gilt $F^{-1}(\pi) \in X$. F^{-1} ist lipschitzstetig mit Lipschitzkonstante c_0.

(B2) Es gibt ein $\beta > 0$, so dass $\|P(w^*)\|_{\mathcal{Y}} \leq \beta$ ist.

(B3) Es gibt ein $\varepsilon > 0$, so dass $\|(F-P)(w_1) - (F-P)(w_2)\|_{\mathcal{Y}} \leq \varepsilon \|w_1 - w_2\|_{\mathcal{X}}$ für alle $w_1, w_2 \in X_r$ ist.

Wenn $\varepsilon c_0 < 1$ und $\frac{c_0\beta}{1-\varepsilon c_0} \leq r$ sind, dann gibt es ein eindeutig bestimmtes $\tilde{w} \in X_r$, so dass $P(\tilde{w}) = 0_{\mathcal{Y}}$ ist. Weiterhin gilt:

$$\|w^* - \tilde{w}\|_{\mathcal{X}} \leq \frac{c_0}{1-\varepsilon c_0}\|P(w^*)\|_{\mathcal{Y}} \leq \frac{c_0\beta}{1-\varepsilon c_0}. \tag{3.3.1}$$

Beweis: Wir definieren für alle $w \in X_r$ die Abbildung

$$\Phi(w) := F^{-1}((F-P)(w)).$$

Damit ist $\Phi(w)$ nach Bedingung (B1) zunächst eine Abbildung von X_r in X. Aufgrund von (B1) und (B3) gilt für beliebige $w_1, w_2 \in X_r$:

$$\begin{aligned} \|\Phi(w_1) - \Phi(w_2)\|_{\mathcal{X}} &= \|F^{-1}((F-P)(w_1)) - F^{-1}((F-P)(w_2))\|_{\mathcal{X}} \\ &\leq c_0\|(F-P)(w_1) - (F-P)(w_2)\|_{\mathcal{Y}} \\ &\leq c_0\varepsilon\|w_1 - w_2\|_{\mathcal{X}} \ . \end{aligned} \tag{3.3.2}$$

Aus der Bedingung (B1) folgt $w^* = F^{-1}(F(w^*))$. Es folgt aus (B1) und der Dreiecksungleichung für beliebiges $w \in X_r$:

$$\begin{aligned}\|w^* - \Phi(w)\|_{\mathcal{X}} &= \|F^{-1}(F(w^*)) - F^{-1}((F-P)(w))\|_{\mathcal{X}} \\ &\leq c_0 \|F(w^*) - (F-P)(w)\|_{\mathcal{Y}} \\ &\leq c_0(\|F(w^*) - (F-P)(w^*)\|_{\mathcal{Y}} \\ &\quad + \|(F-P)(w^*) - (F-P)(w)\|_{\mathcal{Y}}).\end{aligned}$$

Damit folgt

$$\|w^* - \Phi(w)\|_{\mathcal{X}} \leq c_0(\beta + \varepsilon r) \leq r$$

aus $r \geq \frac{c_0\beta}{1-\varepsilon c_0}$, $\varepsilon c_0 < 1$, (B2), (B3) und $\|w^* - w\|_{\mathcal{X}} \leq r$. Somit ist $\Phi(w) \in X_r$ für alle $w \in X_r$. Somit ist Φ eine Kontraktion in X_r, da $\varepsilon c_0 < 1$ ist und Gleichung (3.3.2) gilt. Daraus folgt mit dem Fixpunktsatz von Banach, dass es ein eindeutiges $\tilde{w} \in X_r$ mit $\tilde{w} = \Phi(\tilde{w}) = F^{-1}((F-P)(\tilde{w}))$ gibt. Da andererseits mit $\tilde{w} \in X_r$ ebenfalls $\tilde{w} = F^{-1}(F(\tilde{w}))$ gilt, folgt daraus, dass $F(\tilde{w}) = (F-P)(\tilde{w})$ und somit $P(\tilde{w}) = 0_{\mathcal{Y}}$ sein muss.

Weiterhin gilt für w^* und $\tilde{w}$:

$$\begin{aligned}\|w^* - \tilde{w}\|_{\mathcal{X}} &= \|F^{-1}(F(w^*)) - F^{-1}((F-P)(\tilde{w}))\|_{\mathcal{X}} \\ &\leq c_0 \|F(w^*) - (F-P)(\tilde{w})\|_{\mathcal{Y}} \\ &= c_0 \|P(w^*) + (F-P)(w^*) - (F-P)(\tilde{w})\|_{\mathcal{Y}} \\ &\leq c_0 \|P(w^*)\|_{\mathcal{Y}} + c_0 \|(F-P)(w^*) - (F-P)(\tilde{w})\|_{\mathcal{Y}} \\ &\leq c_0 \|P(w^*)\|_{\mathcal{Y}} + c_0 \varepsilon \|w^* - \tilde{w}\|_{\mathcal{X}}\end{aligned}$$

und daraus folgt die Ungleichung (3.3.1). ◊

Dieses Lemma kann man natürlich äquivalent formulieren, indem man eine metrische Menge bzw. einen vollständigen metrischen Raum X und die abgeschlossene Kugel B_r betrachtet, aber die Menge X dabei nicht als Teilmenge eines vollständigen normierten Vektorraums auffasst. Da wir später Aussagen benutzen wollen, bei denen wir einen reflexiven Banachraum ausnutzen (wobei $\mathcal{X}$ und damit auch X Teilmenge dieses reflexiven Banachraums sein werden), ist es an dieser Stelle günstig, die Menge X gleich als Teilmenge eines Banachraums aufzufassen.

3.3.1 Konstruktion der Vektorräume

Um mit Hilfe des Lemma 3.3.1 die Existenz von Lösungen $\tilde{x}$, $\tilde{\lambda}$ und $\tilde{u}$ der diskretisierten Gleichungen (DFC) für hinreichend kleines h zu zeigen, müssen wir zunächst geeignete Vektorräume $\mathcal{X}$, $\mathcal{Y}$, die Teilmenge $X = X_h$ sowie Abbildungen $P = P_h$ und F definieren. Hierbei wollen wir im Prinzip diese Existenz für jedes hinreichend kleine, fest gewählte $h \in H$ einzeln zeigen. Dabei kann man theoretisch die Teilmenge X_h in Abhängigkeit von der Schrittweite h wählen. Man muss lediglich darauf achten, dass X_h abgeschlossen und der Raum $\mathcal{X}$ vollständig sind. Den Raum $\mathcal{X}$ wählen wir als

$$\begin{aligned}\mathcal{X} = \{&w = (x, \lambda, u) : \\ &x \in W^{1,\infty}([0,T], \mathbb{R}^n),\ \lambda \in W^{1,\infty}([0,T], \mathbb{R}^n),\ u \in W^{1,\infty}([0,T], \mathbb{R}^m)\}\end{aligned}$$

und benutzen die Norm

$$\|w\|_{\mathcal{X}} := \|x\|_{W^{1,\infty}} + \|\lambda\|_{W^{1,\infty}} + \|u\|_{W^{1,\infty}}.$$

Weiterhin definieren wir die Teilmenge

$$X_h = \{w \in \mathcal{X} : x(0) = a\},$$

wobei der Index h hier nur zur besseren Unterscheidung von dem Raum $\mathcal{X}$ dient. Die Menge X_r aus Lemma 3.3.1 identifizieren wir mit der Menge

$$X_{h,r} := B_r(w^*) \cap X_h.$$

Außerdem werden wir Differenzen $w_1 - w_2$ von Elementen aus $X_{h,r}$ benötigen. Es gilt offensichtlich $\|w_1 - w_2\|_{\mathcal{X}} \leq 2r$ für $w_1, w_2 \in X_{h,r}$, und damit liegen solche Differenzen dann in der Menge

$$\tilde{X}_{h,2r} = \{w \in \mathcal{X} : x(0) = 0,\ \|w\|_{\mathcal{X}} \leq 2r\}.$$

Diese Menge unterscheidet sich durch $x(0) = 0$ statt $x(0) = a$ von einer Menge $X_{h,(2r)}$, welche wir im folgenden nicht benötigen.

Das Lemma 3.3.1 soll im Folgenden sicherstellen, dass es für eine Abbildung $P = P_h$ ein eindeutig bestimmtes $\tilde{w}$ mit $P(\tilde{w}) = 0_{\mathcal{Y}}$ gibt. Daher wählen wir die Abbildung P_h so, dass dies gleichbedeutend damit ist, dass die Gleichungen (DFC) erfüllt sind. Damit erklären wir die Abbildung P_h punktweise für $w \in X_{h,r}$, $t \in [0,T]$ mit

$$\begin{aligned} j := j(t) :&= \max\left\{0,\ \left\lceil \frac{tN}{T} - 1 \right\rceil\right\} \\ &= \max\left\{0,\ \min\{k \in \mathbb{Z} : \frac{t}{h} - 1 \leq k\}\right\}, \end{aligned} \tag{3.3.3}$$

$t_j = jh$, $\theta = \frac{t-t_j}{h}$ und für alle $h \in H$ als

$$P_h(w)(t) = \big(P_h^1(w)(t),\ P_h^2(w)(t),\ P_h^3(w)(t),\ P_h^4(w)(t)\big)^\top,$$

wobei die einzelnen Komponenten jeweils

$$\begin{aligned} P_h^1(w)(t) &= x(t) - a - \sum_{i=0}^{j-1} h \sum_{k=1}^{s} b_k f(y_{i,k}, u(t_{i,k})) - h \sum_{k=1}^{s} b_k(\theta) f(y_{j,k}, u(t_{j,k})), \\ P_h^2(w)(t) &= \lambda(t) - \lambda(0) - \sum_{i=0}^{j-1} h \sum_{k=1}^{s} b_k \big(f_x(x(\bar{t}_{i,k}), u(\bar{t}_{i,k}))\big)^\top v_{i,k} \\ &\quad - h \sum_{k=1}^{s} b_k(\theta) \big(f_x(x(\bar{t}_{j,k}), u(\bar{t}_{j,k}))\big)^\top v_{j,k}, \\ P_h^3(w)(t) &= f_u(x(t), u(t))^\top \lambda(\bar{t}), \\ P_h^4(w)(t) &= \lambda(0) - \big(\Psi_x(x(T))\big)^\top \end{aligned}$$

sind. Dabei haben wir für eine übersichtlichere Darstellung die Abkürzungen $t_{j,k} := t_j + c_k h$, $\bar{t} = T - t$, $\bar{t}_{j,k} := T - t_j - c_k h$ und $b_k = b_k(1)$ benutzt. Die Zwischenstellen $y_{j,k}$ und $v_{j,k}$ berechnen sich aus den Gleichungssystemen

$$y_{j,k} = x(t_j) + h \sum_{l=1}^{s} a_{k,l} f(y_{j,l}, u(t_{j,l})), \tag{3.3.4}$$

$$v_{j,k} = \lambda(t_j) + h \sum_{l=1}^{s} a_{k,l} f_x(x(\bar{t}_{j,l}), u(\bar{t}_{j,l}))^\top v_{j,l}. \tag{3.3.5}$$

Hierbei summieren wir in den ersten beiden Komponenten von P_h lediglich die Einschrittfehler auf, da wir für die nächsten Hilfspunkte $y_{j,k}$ und $v_{j,k}$ und damit für den Fehler im nächsten Schritt wieder die Punkte $x(t_j)$ und $\lambda(t_j)$ zugrunde legen. Würde man zu den ursprünglichen Differentialgleichungen die Näherungslösungen zu der aktuellen Steuerung u berechnen, dann müsste man bei der Berechnung der Hilfspunkte $y_{j,k}$ und $v_{j,k}$ die bis dahin berechnete Näherungslösung benutzen. Damit ist zum Beispiel $P_h^1(w)(t)$ nicht gleich der Differenz aus $x(t)$ und der Näherungslösung der Zustandsgleichung zur Steuerung $u(t)$. Trotzdem sind durch diese Konstruktion $P_h^1(w)$ und $P_h^2(w)$ lipschitzstetig. Allerdings ist selbst bei Verwendung eines C^1-Verfahrens (siehe Abschnitt 3.7) nicht zu erwarten, dass $P_h^1(w)$ und $P_h^2(w)$ für alle $w \in X_{h,r}$ einmal stetig differenzierbar sind. Ist hingegen $P_h^1(\tilde{w})(t) = 0$ für alle $t \in [0,T]$ für ein $\tilde{w} \in X_{h,r}$, dann ist zum Beispiel

$$\tilde{x}(t_1) = a + h \sum_{k=1}^{s} b_k f(y_{0,k}, \tilde{u}(t_{0,k})),$$

und es macht keinen Unterschied, ob man für die Berechnung der Hilfspunkte $y_{1,k}$ im 2. Schritt $\tilde{x}(t_1)$ oder die Summe auf der rechten Seite benutzt. Zur Vereinfachung haben wir uns ebenfalls mit Hilfe der Menge X_h auf Funktionen x mit $x(0) = a$ beschränkt und diese Bedingung nicht in den Operator P_h aufgenommen.

Weiterhin benötigen wir einen Operator F, für den wir die Bedingungen (B1) bis (B3) von Lemma 3.3.1 zeigen können. Wir definieren F punktweise für $t \in [0,T]$ als

$$F(w)(t) = \begin{pmatrix} x(t) - a - \int_0^t A(\tau)x(\tau) + B(\tau)u(\tau)\,d\tau \\ \lambda(t) - \lambda(0) - \int_0^t Q(\bar{\tau})x(\bar{\tau}) + S(\bar{\tau})u(\bar{\tau}) + A(\bar{\tau})^\top \lambda(\tau)\,d\tau \\ R(t)^\top u(t) + S(t)^\top x(t) + B(t)^\top \lambda(T-t) \\ \lambda(0) - V^\top x(T) \end{pmatrix},$$

dabei sei $\bar{t} = T - t$ und $\bar{\tau} = T - \tau$. Dieser bezüglich $w \in \mathcal{X}$ lineare Operator wurde so gewählt, dass man die Bedingung (B1) von Lemma 3.3.1 mit Hilfe eines Problems der optimalen Steuerung mit quadratischer Zielfunktion und linearen Nebenbedingungen zeigen kann (siehe Abschnitt 3.5). Dieser Operator F ist eine Vereinfachung eines entsprechenden Operators in [DH98], da wir nur Probleme ohne zusätzliche Steuerungs- oder Zustandsbeschränkungen betrachten. Im Gegensatz dazu wird in [Hag00] ein diskretisiertes linear-quadratisches Problem benutzt.

Die Eigenschaften von $P_h(w)$ und $F(w)$ nutzen wir, um einen geeignenten Raum $\mathcal{Y}$ zu wählen. Hierzu ordnen wir die Funktionswerte des Operators $P_h(w)$, entsprechend der Definitionen oben, zu einer Funktion $\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$ an. Da x, λ und u jeweils in $W^{1,\infty}$-Räumen sind, alle auftretenden Funktionen bzw. Operatoren von x, λ, u lipschitzstetig sind und ansonsten nur Polynome und Integrale auftreten, ist offensichtlich

$$\begin{aligned}\mathcal{Y} = \{\pi = (\pi_1, \pi_2, \pi_3, \pi_4) : \pi_1 \in W^{1,\infty}([0,T], \mathbb{R}^n),\ \pi_2 \in W^{1,\infty}([0,T], \mathbb{R}^n),\\ \pi_3 \in W^{1,\infty}([0,T], \mathbb{R}^m),\ \pi_4 \in \mathbb{R}^m,\\ \text{mit } \pi_1(0) = 0\}\end{aligned}$$

eine geeignete Wahl für den Raum $\mathcal{Y}$. Als Norm im Raum $\mathcal{Y}$ benutzen wir

$$\|\pi\|_{\mathcal{Y}} := \|\pi_1\|_{W^{1,\infty}} + \|\pi_2\|_{W^{1,\infty}} + \|\pi_3\|_{W^{1,\infty}} + \|\pi_4\|_2.$$

Die Definition der Räume $\mathcal{X}$ und $\mathcal{Y}$ wollen wir kurz damit abschließen, die Vollständigkeit dieser Räume festzuhalten.

Lemma 3.3.2. *Die Räume $\mathcal{X}$ und $\mathcal{Y}$ sind vollständig, und die Menge $X_h \subset \mathcal{X}$ ist abgeschlossen.*

Beweis: Nach [Sch87] sind die Sobolevräume $W^{1,\infty}([0,T], \mathbb{R}^n)$ mit der Norm $\|\cdot\|_{W^{1,\infty}}$ vollständig, und damit sind auch die Räume $\mathcal{X}$ und $\mathcal{Y}$ vollständig. Sei x^i eine Cauchy-Folge und für alle x^i gilt $x^i(0) = a$. Aus der Konvergenz in $W^{1,\infty}([0,T], \mathbb{R}^n)$ folgt für die Funktionen x^i die punktweise Konvergenz, und somit gilt für den Grenzwert der Cauchy-Folge in $\mathcal{X}$ ebenfalls $x(0) = a$, und X_h ist somit abgeschlossen. $\Diamond$

Mit Hilfe von Lemma 3.3.1 zeigen wir in den nächsten Abschnitten folgenden Satz:

Satz 3.3.3. *Für das Problem* (OS) *mögen aus Abschnitt 3.1* Glattheit *mit einem $\kappa \geq 3$ und* Koerzivität *gelten. Weiterhin seien die Parameter $a_{k,l}$ für $k, l = 1, 2, \ldots, s$ mit $c_k = \sum_{l=1}^{s} a_{k,l}$ sowie die Polynome $b_k(\cdot)$ eines SRKVs mit Konsistenzordnung (q, q) mit $q \leq \kappa$ gegeben. Dann gibt es ein $h_{3.3.6} > 0$ und $r_{3.3.6} > 0$, so dass für alle $h \leq h_{3.3.6}$ das Gleichungssystem* (DFC) *eine eindeutig bestimmte Lösung $\tilde{w} = (\tilde{x}, \tilde{\lambda}, \tilde{u}) \in X_{h,r}$ besitzt, und es gibt eine von der Schrittweite h unabhängige Konstante $\zeta_{3.3.6}$, mit der gilt*

$$\|\tilde{w} - w^*\|_{\mathcal{X}} \leq \zeta_{3.3.6}\, h^q. \tag{3.3.6}$$

Zum Beweis dieses Satzes werden wir das Lemma 3.3.1 anwenden. Wir benutzen dabei die oben definierten Operatoren P_h, F, die Räume $\mathcal{X}$, $\mathcal{Y}$, sowie die Teilmengen X_h und $X_{h,r}$. Dabei zeigt sich, dass man für hinreichend kleine Schrittweiten h und kleine Radien r die Konstante $\zeta_{3.3.6}$ unabhängig von h und r wählen kann. Wir werden den Beweis zunächst mit einigen Bemerkungen über die Hilfspunkte $y_{j,k}$ und $v_{j,k}$ des Runge-Kutta-Verfahrens im Abschnitt 3.3.2 beginnen. Der gesamte Beweis ist so konstruiert, dass wir im Abschnitt 3.3.3 die Bedingung (B2) des Lemma 3.3.1 auf die Bemerkungen aus Abschnitt 2.2.1 zurückführen können. Dadurch

ist sichergestellt, dass man existierende SRKV mit hoher Konvergenzordnung verwenden kann und man keine speziellen Verfahren entwickeln muss. Ein großer Teil des Beweises besteht schließlich darin im Abschnitt 3.4 Bedingung (B3) zu zeigen. Die Invertierbarkeit des linearen Operators F und die Beschränktheit des inversen Operators, das heißt Bedingung (B1), wird in Abschnitt 3.5 gezeigt.

Bemerkung 3.3.4: Für den Beweis auf den folgenden Seiten besteht die wesentliche Arbeit darin, die geforderten Eigenschaften für die Ableitungen der Operatoren nach t zu zeigen. Die ersten und zweiten Komponenten des Operators P_h verdeutlichen den Zusammenhang zu dem Gleichungssystem (DFC).

3.3.2 Die Hilfspunkte des Runge-Kutta-Verfahrens

Im weiteren Verlauf des Beweis benötigt man an einigen Stellen Aussagen über die Existenz und Beschränktheit der Hilfspunkte $y_{j,k}$ (bzw. $v_{j,k}$), die sich als Lösung des Gleichungssystems (3.3.4) (bzw. (3.3.5)) ergeben. Weiterhin werden wir die Ableitung dieser Hilfspunkte nach dem zugrundeliegenden $w \in X_{h,r}$ benötigen und werden daher ebenfalls eine Aussage über die Existenz und die Beschränktheit dieser Ableitungen zeigen. Wir beweisen zunächst analog zu den bekannten Resultaten in [HNW93, But03], dass die Hilfspunkte $y_{j,k}$ beschränkt sind. Um dabei die Abhängigkeit der Hilfspunkte von $w \in X_{h,r}$ deutlich zu machen, bezeichnen wir die Lösung des Gleichungssystems (3.3.4) in diesem Abschnitt mit $y_{j,k}(w)$.

Lemma 3.3.5. *Für Matrizen $y_j \in \mathbb{R}^{n\times s}$ mit Spalten $y_{j,k} \in \mathbb{R}^n$, das heißt $y_j = (y_{j,1}, y_{j,2}, \ldots, y_{j,s})$, definieren wir die Norm*

$$\|y_j\|_{max,2} := \max_{k=1,\ldots,s} \|y_{j,k}\|_2.$$

Man erhält somit einen vollständigen normierten Vektorraum der Matrizen aus $\mathbb{R}^{n\times s}$. Es sei das Gleichungssystem

$$y_{j,k} = x(t_j) + h\sum_{l=1}^{s} a_{k,l} f(y_{j,l}, u(t_j + c_l h)) \tag{3.3.7}$$

gegeben, und es gilt Glattheit *aus Abschnitt 3.1. Dann gibt es ein $h_{3.3.8} > 0$, so dass zu jedem $w = (x, \lambda, u) \in X_{h,r}$ und für alle $k = 1, 2, \ldots, s$ eine eindeutig bestimmte Lösung $y_{j,k}(w)$ von (3.3.7) existiert. Weiterhin gilt dann für alle $h \in H$ mit $h \le h_{3.3.8}$*

$$\|y_{j,k}(w) - x(t_j)\|_2 \le h\zeta_{3.3.8}, \quad j = 0, 1, \ldots, N-1, \quad k = 1, 2, \ldots, s \tag{3.3.8}$$

mit einer von $w \in X_{h,r}$ unabhängigen Konstanten $\zeta_{3.3.8}$.

Beweis: Man kann leicht nachvollziehen, dass die Definition von $\|y_j\|_{max,2}$ eine Norm darstellt und der entstehende normierte Vektorraum vollständig ist, da er endlichdimensional ist. Wir wollen zeigen, dass das Gleichungssystem (3.3.7) Lösungen $y_{j,k}(w)$ besitzt, diese beschränkt sind und Gleichung (3.3.8) gilt. Hierzu benutzen wir den Fixpunktsatz von Banach und zeigen zunächst, dass das Gleichungssystem aus dem die $y_{j,k}(w)$ berechnet werden für hinreichend kleine Schrittweiten

eine Kontraktion in $\mathbb{R}^{n\times s}$ darstellt. Wir ordnen die rechten Seite der Gleichung (3.3.7) für gegebenes $w \in X_{h,r}$ und beliebiges $y_j \in \mathbb{R}^{n\times s}$, analog zu y_j, zu einer Matrix $G_w(y_j)$ an. Dabei wollen wir durch den Index w verdeutlichen, dass diese Abbildung von $\mathbb{R}^{n\times s}$ in $\mathbb{R}^{n\times s}$ von dem gegebenen $w \in X_{h,r}$ abhängt. Es gilt damit für beliebige $y_j^1, y_j^2 \in \mathbb{R}^{n\times s}$

$$\begin{aligned}
&\|G_w(y_j^1) - G_w(y_j^2)\|_{max,2} \\
&\quad = \max_{k=1,\dots,s} \|h \sum_{l=1}^{s} a_{k,l}[f(y_{j,l}^1, u(t_j + c_l h)) - f(y_{j,l}^2, u(t_j + c_l h))]\|_2 \\
&\quad \le h \max_{k=1,\dots,s} \sum_{l=1}^{s} |a_{k,l}| \|f(y_{j,l}^1, u(t_j + c_l h)) - f(y_{j,l}^2, u(t_j + c_l h))\|_2 \\
&\quad \le h \max_{k=1,\dots,s} \sum_{l=1}^{s} |a_{k,l}| M_f^0 \|y_{j,l}^1 - y_{j,l}^2\|_2 \\
&\quad \le h \max_{k=1,\dots,s} \sum_{l=1}^{s} |a_{k,l}| M_f^0 (\max_{l=1,\dots,s} \|y_{j,l}^1 - y_{j,l}^2\|_2) \\
&\quad \le h M_f^0 \max_{k=1,\dots,s} \bar{c}_k \, \|y_j^1 - y_j^2\|_{max,2},
\end{aligned}$$

wobei wir $\bar{c}_k = \sum_{l=1}^{s} |a_{k,l}|$ definieren. Somit ist der Operator $G_w(y_j)$ für alle $w \in X_{h,r}$ eine Kontraktion auf $\mathbb{R}^{n\times s}$, wenn gilt

$$h M_f^0 \max_{k=1,\dots,s} \bar{c}_k < 1.$$

Daher wählen wir ein $h_{3.3.8} < (M_f^0 \max_{k=1,\dots,s} \bar{c}_k)^{-1}$ und wenden auf den Operator $G_w(y_j)$ den Fixpunktsatz von Banach an. Wir erhalten ein eindeutig bestimmtes $y_j(w)$ mit $y_j(w) = G_w(y_j(w))$, und es gilt für beliebiges $y_j \in \mathbb{R}^{n\times s}$

$$\|y_j - y_j(w)\|_{max,2} \le \frac{1}{1 - h_{3.3.8} M_f^0 \max_{k=1,\dots,s} \bar{c}_k} \|y_j - G_w(y_j)\|_{max,2}.$$

Für $y_j = (x(t_j), \dots, x(t_j))$ kann man $\|y_j - G_w(y_j)\|_{max,2}$ durch

$$\begin{aligned}
&\|y_j - G_w(y_j)\|_{max,2} = \max_{k=1,\dots,s} \|x(t_j) - x(t_j) - h \sum_{l=1}^{s} a_{k,l} f(x(t_j), u(t_j + c_l h))\|_2 \\
&\le h \max_{k=1,\dots,s} \sum_{l=1}^{s} |a_{k,l}| \|f(x(t_j), u(t_j + c_l h)) - f(x^*(t_j), u^*(t_j)) + f(x^*(t_j), u^*(t_j))\|_2 \\
&\quad \le h \max_{k=1,\dots,s} \sum_{l=1}^{s} |a_{k,l}| \big[M_f^0 \|x(t_j) - x^*(t_j)\|_2 + M_f^0 \|u(t_j + c_l h) - u^*(t_j)\|_2 \\
&\qquad\qquad + \|f(x^*(t_j), u^*(t_j))\|_2 \big]
\end{aligned}$$

abschätzen. Da $w \in X_{h,r}$ vorausgesetzt ist, gilt $\|x(t_j) - x^*(t_j)\|_2 \le r$ und

$$\begin{aligned}
\|u(t_{j,l}) - u^*(t_j)\|_2 &\le \|u(t_{j,l}) - u^*(t_{j,l})\|_2 + \|u^*(t_{j,l}) - u^*(t_j)\|_2 \\
&\le \qquad r \qquad + m^{\frac{1}{2}} \|\dot{u}^*\|_{L^\infty} h \max_{k=1,\dots,s} c_k.
\end{aligned}$$

Für $y_j = (x(t_j), \dots, x(t_j))$ folgt damit die Behauptung aus

$$\|y_j - G_w(y_j)\|_{max,2} \leq h\big[2M_f^0 r + h\, m^{\frac{1}{2}} M_f^0 \|\dot{u}^*\|_{L^\infty} + \|f(x^*(t_j), u^*(t_j))\|_2\big] \max_{k=1,\dots,s} \bar{c}_k,$$

da $\max_{k=1,\dots,s} c_k \leq 1$ ist und das zweite h auf der rechten Seite natürlich durch $h_{3.3.8}$ nach oben abgeschätzt werden kann. ◊

Lemma 3.3.5 zeigt, dass für hinreichend kleine Schrittweiten h es zu jedem $w = (x, \lambda, u) \in X_{h,r}$ und zu jedem Zeitpunkt $j = 0, 1, \dots, N-1$ ein eindeutig bestimmtes $y_j(w) = (y_{j,1}(w), y_{j,2}(w), \dots, y_{j,s}(w))$ gibt, welches das Gleichungssystem (3.3.7) löst.

Im Folgenden interessieren wir uns dafür, ob diese Lösungen $y_j(w)$ differenzierbar von dem Punkt $w \in X_{h,r}$ abhängen. Wir betrachten die k-te Spalte der Matrix $y_j(w)$, die das Gleichungssystems (3.3.7) löst. Bildet man auf beiden Seiten des Gleichungssystems (3.3.7) zu jedem $\nu \in \tilde{X}_{h,2r}$ die klassischen Ableitungen von $y_{j,k}(w + \vartheta\nu)$ nach $\vartheta \in \mathbb{R}$ an der Stelle $\vartheta = 0$, erhält man das Gleichungssystem

$$\begin{aligned}(3.3.9)\quad {y_{j,k}}'(w)(\nu) = \nu_1(t_j) + h \sum_{l=1}^{s} a_{k,l}[&f_x(y_{j,l}(w), u(t_{j,l}))({y_{j,l}}'(w)(\nu)) \\ &+ f_u(y_{j,l}(w), u(t_{j,l}))(\nu_3(t_{j,l}))]\end{aligned}$$

zur Bestimmung der Fréchet-Ableitungen von $y_j(w)$ nach w angewandt auf ein $\nu \in \tilde{X}_{h,2r}$. Dabei setzt sich ν entsprechend $w = (x, \lambda, u)$ als $\nu = (\nu_1, \nu_2, \nu_3)$ zusammen. Hat also das Gleichungssystem (3.3.9) für jedes $\nu \in \tilde{X}_{h,2r}$ eine eindeutig bestimmte Lösung, dann sind diese Lösungen natürlich auch die Ableitungen von $y_{j,k}(w)$ in Richtung ν.

Damit können wir analog zu Lemma 3.3.5 eine Abschätzung für die Ableitungen der Hilfspunkte $y_{j,k}(w)$ angeben.

Lemma 3.3.6. *Es gibt eine positive Zahl $h_{3.3.9}$ mit $0 < h_{3.3.9} \leq h_{3.3.8}$, so dass für alle $h \in H$ mit $h \leq h_{3.3.9}$, $w = (x, \lambda, u) \in X_{h,r}$, $\nu \in \tilde{X}_{h,2r}$ und für alle Zeitpunkte $j = 0, 1, \dots, N-1$ das Gleichungssystem* (3.3.9) *eine eindeutig bestimmte Lösung $\tilde{y}'_j$ hat.*

Beweis: Wie im Beweis von Lemma 3.3.5 ordnen wir die $y'_{j,k}(w)(\nu)$ und die Auswertungen der rechte Seite von Gleichung (3.3.9) für $k = 1, 2, \dots, s$ zu Matrizen y'_j und $G_{w,\nu}(y'_j)$ an. Dabei betrachten wir $y_j(w) \in \mathbb{R}^{n\times s}$, $w \in X_{h,r}$ und $\nu \in \tilde{X}_{h,2r}$ als gegeben. Es gilt dann für zwei $y^1, y^2 \in \mathbb{R}^{n\times s}$

$$\begin{aligned}\|G_{w,\nu}(y^1) - G_{w,\nu}(y^2)\|_{max,2} &= \max_{k=1,2,\dots,s} \|h \sum_{l=1}^{s} a_{k,l}[\, f_x(y_{j,l}(w), u(t_{j,l}))(y_l^1) \\ &\qquad - f_x(y_{j,l}(w), u(t_{j,l}))(y_l^2)]\|_2 \\ &\leq h \max_{k=1,2,\dots,s} \sum_{l=1}^{s} |a_{k,l}| \|f_x(y_{j,l}(w), u(t_{j,l}))(y_l^1 - y_l^2)\|_2.\end{aligned}$$

Da nach Voraussetzung *Glattheit* f_x lipschitzstetig mit Lipschitzkonstante M_f^1 und natürlich eine lineare Abbildung ist, gilt

$$\|f_x(y_{j,l}(w), u(t_{j,l}))(y_l^1 - y_l^2)\|_2 \leq \sup_{\substack{z\in\mathbb{R}^n \\ z\neq 0}} \frac{\|f_x(y_{j,l}(w), u(t_{j,l}))z\|_2}{\|z\|_2} \|y_l^1 - y_l^2\|_2.$$

Dabei schätzen wir den ersten Term auf der rechten Seite mit Hilfe von

$$\sup_{\substack{z\in\mathbb{R}^n\\ z\neq 0}} \frac{\|[f_x(y_{j,l}(w),u(t_{j,l}))-f_x(x^*(t_j),u^*(t_j))+f_x(x^*(t_j),u^*(t_j))]z\|_2}{\|z\|_2}$$

$$\leq M_f^1\|y_{j,l}(w)-x^*(t_j)\|_2+M_f^1\|u(t_{j,l})-u^*(t_j)\|_2+\sup_{\substack{z\in\mathbb{R}^n\\ z\neq 0}}\frac{\|f_x(x^*(t_j),u^*(t_j))z\|_2}{\|z\|_2}$$

ab. Benutzt man die Abkürzung

$$\|A\|_{\mathcal{M}} := \sup_{t\in[0,T]}\sup_{\substack{z\in\mathbb{R}^n\\ z\neq 0}}\frac{\|f_x(x^*(t),u^*(t))z\|_2}{\|z\|_2},$$

dann gilt nach *Glattheit* $\|A\|_{\mathcal{M}}<\infty$, und man erhält

$$\|G_{w,\nu}(y^1)-G_{w,\nu}(y^2)\|_{max,2}\leq h\max_{k=1,2,\ldots,s}\sum_{l=1}^{s}|a_{k,l}|$$
$$\cdot\left(M_f^1\left[\|y_{j,l}(w)-x^*(t_j)\|_2+\|u(t_{j,l})-u^*(t_j)\|_2\right]+\|A\|_{\mathcal{M}}\right)\|y_l^1-y_l^2\|_2.$$

Wählt man $h\leq h_{3.3.8}$ für $h_{3.3.8}$ aus Lemma 3.3.5, dann gilt $\|y_{j,l}(w)-x(t_j)\|_2\leq h\zeta_{3.3.8}$. Außerdem folgt mit $\|y_l^1-y_l^2\|_2\leq\|y^1-y^2\|_{max,2}$ und $\|x(t_j)-x^*(t_j)\|_2\leq r$ die Abschätzung $\|y_{j,l}(w)-x^*(t_j)\|_2\leq h\zeta_{3.3.8}+r$ und schließlich

$$\|G_{w,\nu}(y^1)-G_{w,\nu}(y^2)\|_{max,2}\leq h\left(M_f^1(h\zeta_{3.3.8}+r+r+m^{\frac{1}{2}}\|\dot{u}^*\|_{L^\infty}h\max_{k=1,\ldots,s}c_k)+\|A\|_{\mathcal{M}}\right)$$
$$\cdot\|y^1-y^2\|_2\max_{k=1,2,\ldots,s}\bar{c}_k.$$

Man kann daher ein hinreichend kleines $h_{3.3.9}\leq h_{3.3.8}$ so wählen, dass

$$\zeta_{3.3.9}:=h_{3.3.9}$$
$$\cdot\left(M_f^1(h_{3.3.8}\zeta_{3.3.8}+2\,r+m^{\frac{1}{2}}\|\dot{u}^*\|_{L^\infty}h_{3.3.8}\max_{k=1,\ldots,s}c_k)+\|A\|_{\mathcal{M}}\right)\max_{k=1,2,\ldots,s}\bar{c}_k<1$$

ist. Damit ist $G_{w,\nu}(\cdot)$ für alle $h\leq h_{3.3.9}$ ebenfalls eine Kontraktion auf $\mathbb{R}^{n\times s}$ und es existiert zu jedem $\nu\in\tilde{X}_{h,2r}$ ein eindeutig bestimmtes $\tilde{y}_j'(w)(\nu)$, welches das Gleichungssystem (3.3.9) löst. Setzt man $y^0=(\nu_1(t_j),\ldots,\nu_1(t_j))$, dann erhält man

$$\|\tilde{y}_j'(w)(\nu)-y^0\|_{max,2}\leq\frac{1}{1-\zeta_{3.3.9}}\|y^0-G_{w,\nu}(y^0)\|_{max,2} \tag{3.3.10}$$

mit

$$\|y^0-G_{w,\nu}(y^0)\|_{max,2}=\max_{k=1,2,\ldots,s}\|h\sum_{l=1}^{s}a_{k,l}[f_x(y_{j,l}(w),u(t_{j,l}))(\nu_1(t_j))$$
$$+f_u(y_{j,l}(w),u(t_{j,l}))(\nu_3(t_{j,l}))]\|_2. \tag{3.3.11}$$

$\Diamond$

3.3.3 Ausnutzen der Ordnung des SRKVs

In diesem Abschnitt zeigen wir, dass durch die Konstruktion des Operators P_h die Bedingung (B2) je nach verwendeten SRKVs mit einem $\beta = \zeta_\beta h^q$ erfüllt ist, wobei die Konstante ζ_β unabhängig von der Schrittweite h ist. Dadurch wird die Ordnung, mit der die Näherungslösungen der Differentialgleichungen mit sinkender Schrittweite h gegen die exakte Lösung konvergieren auf das Optimierungsproblem bzw. auf die Lösungen der Gleichungssysteme (FC) und (DFC) übertragen. Damit die Bedingung (B2) für die ersten Komponenten des Operators P_h erfüllt ist, muss

$$\sup_{t\in[0,T]} \Big\| x^*(t) - a - \sum_{i=0}^{j-1} h \sum_{k=1}^{s} b_k f(y^*_{i,k}, u^*(t_{i,k})) - h\sum_{k=1}^{s} b_k(\theta) f(y^*_{j,k}, u^*(t_{j,k})) \Big\|_2 \leq \beta \tag{3.3.12}$$

gelten. Dabei berechnet sich j aus Gleichung (3.3.3), und es sind $t_j = jh$, $\theta = \frac{t-t_j}{h}$ und $t_{j,k} = t_j + c_k h$. Weiterhin sei $y^*_{j,k}$ eine kürzere Schreibweise für $y_{j,k}(w^*)$. Da $x^* \in W^{\kappa,\infty}([0,T],\mathbb{R}^n)$ nach der Voraussetzung *Glattheit* Lösung der Differentialgleichung ist, die Funktion f nach beiden Argumenten κ-mal differenzierbar ist und außerdem die optimale Steuerung u^* κ-mal differenzierbar ist, erfüllt $g(x(t),t) := f(x^*(t),u^*(t))$ die Voraussetzung, damit man den globalen Diskretisierungsfehler nach Satz 2.1.5 durch ζh^q nach oben abschätzen kann. Mit einer ähnlichen Vorgehensweise wie in Satz 2.1.5 summiert man in Gleichung (3.3.12) lediglich die Einschrittfehler auf. Da $x^*(t_0) = a$ ist, gilt die Gleichung

$$x^*(t) - a = x^*(t) - x^*(t_j) + \sum_{k=0}^{j-1} x^*(t_{k+1}) - x^*(t_k). \tag{3.3.13}$$

Weiterhin gilt mit der Voraussetzung *Glattheit* und den Aussagen aus Kapitel 2

$$\Big\| x^*(t) - x^*(t_j) - h\sum_{k=1}^{s} b_k(\theta) f(y^*_{j,k}, u^*(t_{j,k})) \Big\|_2 \leq \eta_q h^{q+1}.$$

Setzt man die Gleichung 3.3.13 in die Ungleichung 3.3.12 ein, dann sieht man, dass man die Norm in Ungleichung 3.3.12 durch die Norm von $j+1$ Summanden abschätzen kann. Nach Abschnitt 2.2.1 können die $j+1 \leq N = \frac{T}{h}$ einzelnen Summanden jeweils durch $\eta_q h^{q+1}$ nach oben abgeschätzt werden. Weiterhin folgt mit Gleichung (2.2.7) sofort eine entsprechende Abschätzung der L^∞-Norm der schwachen Ableitung von $P_h^1(w^*)$ nach t. Man beachte dabei, dass in $P_h^1(w^*)$ gemäß der Definition in Abschnitt 3.3.1 nur der Term $x^*(t)$ und $\theta = \frac{t-t_j}{h}$ von t abhängen.

Entsprechende Abschätzungen gelten ebenfalls für $P_h^2(w^*)$, da $\lambda^*(T-\cdot)$ Lösung der Differentialgleichung

$$\dot\lambda(t) = f_x(x^*(\bar t), u^*(\bar t))^\top \lambda(t) \quad \text{mit } \lambda(0) = \Psi_x(x^*(T))^\top$$

ist. Für die letzten Komponenten $P_h^4(w^*)$ muss nichts gezeigt werden, da durch die Voraussetzung *Glattheit* $\lambda^*(0) = \Psi_x(x^*(T))^\top$ gilt und damit $P_h^4(w^*)$ verschwindet. Weiterhin gilt ebenfalls nach *Glattheit* $\lambda^*(T-t)^\top f_u(x^*(t),u^*(t)) = 0$ für alle

$t \in [0, T]$, womit man $P_h^3(w^*)$ ebenfalls durch $\beta = \zeta_\beta h^q$ nach oben abschätzen kann. Somit ist gezeigt, dass die Konvergenzordnung erhalten bleibt und β aus der Bedingung (B2) von Lemma 3.3.1 unabhänging von ε und r beliebig klein gewählt werden kann.

3.4 Lipschitzstetigkeit des Operators $F - P_h$

Als nächstes zeigen wir, dass die Bedingung (B3) von Lemma 3.3.1 für die ersten Komponenten der Abbildung $F - P_h$ erfüllt ist. Dazu benutzen wir einen Mittelwertsatz der Funktionalanalysis, den man zum Beispiel im Buch von Dirk Werner ([Wer07], Satz III.5.4) findet. Danach gilt für alle $w_1, w_2 \in X_{h,r}$

$$\|(F - P_h)(w_1) - (F - P_h)(w_2)\|_{\mathcal{Y}} \leq \sup_{0 \leq \tau \leq 1} \|(F - P_h)'(w_\tau)(w_1 - w_2)\|_{\mathcal{Y}}, \tag{3.4.1}$$

wobei $(F - P_h)'(w_\tau)$ für jeden Punkt $w_\tau = w_1 + \tau(w_2 - w_1)$ mit $0 \leq \tau \leq 1$ eine lineare Abbildung von $\tilde{X}_{h,2r}$ in $\mathcal{Y}$ ist. Offensichtlich gilt aufgrund der Konvexität der Menge X_h und damit von $X_{h,r}$, dass mit $w_1,\ w_2 \in X_{h,r}$ auch $w_\tau \in X_{h,r}$ ist. Damit gilt

$$\sup_{0 \leq \tau \leq 1} \|(F - P_h)'(w_\tau)(w_1 - w_2)\|_{\mathcal{Y}} \leq \sup_{w \in X_{h,r}} \|(F - P_h)'(w)(w_1 - w_2)\|_{\mathcal{Y}},$$

und wir führen zur Vereinfachung die Abkürzung $\|\cdot\|_{\mathcal{Z}}$ als Operatornorm

$$\|(F - P_h)'(w_\tau)\|_{\mathcal{Z}} := \sup_{\substack{\nu \in \tilde{X}_{h,2r} \\ \nu \neq 0}} \frac{\|(F - P_h)'(w_\tau)(\nu)\|_{\mathcal{Y}}}{\|\nu\|_{\mathcal{X}}}$$

ein.

Man kann damit die linke Seite von Gleichung (3.4.1) großzügig durch

$$\|(F - P_h)(w_1) - (F - P_h)(w_2)\|_{\mathcal{Y}} \leq \sup_{w \in X_{h,r}} \|(F - P_h)'(w)\|_{\mathcal{Z}} \|w_1 - w_2\|_{\mathcal{X}}$$

abschätzen, da für die entsprechenden Differenzen $(w_1 - w_2) \in \tilde{X}_{h,2r}$ gilt. Damit kann man sofort die Bedingung (B3) folgern, wenn man zeigen kann, dass

$$\sup_{w \in X_{h,r}} \sup_{\substack{\nu \in \tilde{X}_{h,2r} \\ \nu \neq 0}} \frac{\|(F - P_h)'(w)(\nu)\|_{\mathcal{Y}}}{\|\nu\|_{\mathcal{X}}} \leq \varepsilon. \tag{3.4.2}$$

gilt. Im Folgenden ist dabei zu beachten, dass man für die Anwendung von Lemma 3.3.1 diese Bedingung mit hinreichend kleiner Schrittweite h und hinreichend kleinem r für beliebig kleine ε erfüllen können muss, damit man für die Anwendung von Lemma 3.3.1 $\varepsilon c_0 < 1$ sicherstellen kann. Dabei muss gleichzeitig die Ungleichung $\frac{c_0 \beta}{1 - \varepsilon c_0} \leq r$ gelten, was dadurch erreicht wird, dass man β nach dem Abschnitt 3.3.3 in Abhängigkeit von der Schrittweite h, unabhängig von ε und r, beliebig klein wählen kann.

Wir bestimmen daher für ein gegebenes $w = (x, \lambda, u) \in X_{h,r}$ die Ableitung $(F-P_h)'(w)$. Dabei kann man nach der Voraussetzung *Glattheit* h und r hinreichend klein wählen, so dass die Abbildung $F - P_h$ auf einer offenen Menge, welche $X_{h,r}$ enthält, Fréchet-differenzierbar ist. Die ersten Komponenten von $(F - P_h)(w)$ sind an der Stelle $t \in [0, T]$

$$\sum_{i=0}^{j-1} h \sum_{k=1}^{s} b_k f(y_{i,k}, u(t_{i,k})) + h \sum_{k=1}^{s} b_k(\theta) f(y_{j,k}, u(t_{j,k})) - \int_0^t A(\tau)x(\tau) + B(\tau)u(\tau)\, d\tau, \tag{3.4.3}$$

wobei sich die Hilfspunkte $y_{i,k}$ für $i = 1, \ldots, j$ aus dem Gleichungssystem

$$y_{i,k} = x(t_i) + h \sum_{l=1}^{s} a_{k,l} f(y_{i,l}, u(t_{i,l}))$$

bestimmen und j in der Gleichung (3.3.3) definiert wird. Zu diesem Term muss man die Lipschitzstetigkeit mit einer, in Abhängigkeit von h und r, beliebig kleinen Lipschitzkonstante zeigen. Dabei sind natürlich die entsprechenden $W^{1,\infty}$ Normen zu benutzen. Den Ausdruck (3.4.3) kann man dabei, wegen der Konsistenzbedingung (OB1) und $\int_{t_j}^{t} d\tau = t - t_j = h\theta$, in

$$\sum_{i=0}^{j-1} \int_{t_i}^{t_{i+1}} \sum_{k=1}^{s} b_k \left[f(y_{i,k}, u(t_{i,k})) - A(\tau)x(\tau) - B(\tau)u(\tau) \right] d\tau + \int_{t_j}^{t} \sum_{k=1}^{s} \frac{b_k(\theta)}{\theta} \left[f(y_{j,k}, u(t_{j,k})) - A(\tau)x(\tau) - B(\tau)u(\tau) \right] d\tau \tag{3.4.4}$$

umformen. Stellt man hingegen die Polynome $b_k(\theta)$ als $\sum_{l=1}^{\tilde{q}} b_{k,l}\theta^l$ dar, ergibt sich für die Ableitung des Terms (3.4.3) nach t in den Intervallen (t_j, t_{j+1})

$$\sum_{k=1}^{s} \sum_{l=1}^{\tilde{q}} l b_{k,l} \theta^{l-1} f(y_{j,k}, u(t_{j,k})) - A(t)x(t) - B(t)u(t).$$

Leitet man die Konsistenzordnung (OB1) auf beiden Seiten nach t ab sieht man, dass

$$\sum_{k=1}^{s} \sum_{l=1}^{\tilde{q}} l b_{k,l} \theta^{l-1} = 1$$

gilt und man damit die Ableitung von (3.4.3) nach t als

$$\sum_{k=1}^{s} \sum_{l=1}^{\tilde{q}} l b_{k,l} \theta^{l-1} \left[f(y_{j,k}, u(t_{j,k})) - A(t)x(t) - B(t)u(t) \right]$$

darstellen kann. Kann man damit die Lipschitzstetigkeit der Differenz in der letzten Gleichung mit einer hinreichend kleinen Lipschitzkonstante und für beliebige Zeitpunkte $t \in (t_j, t_{j+1}]$, zeigen, dann folgt daraus die gewünschte Lipschitzstetigkeit in der entsprechenden $W^{1,\infty}$-Norm. In der Gleichung (3.4.4) entsteht dabei durch jedes einzelne Integral der Faktor h, der die Summe über bis zu $N = T/h$ Summanden ausgleicht. Daher wollen wir Fréchet-Ableitung von

$$f(y_{j,k}, u(t_{j,k})) - A(t)x(t) - B(t)u(t) \tag{3.4.5}$$

bezüglich w an einer Stelle $w \in X_{h,r}$ bestimmen, um die entsprechenden $W^{1,\infty}$-Normen abzuschätzen.

Die Terme $A(t)x(t)$ und $B(t)u(t)$ sind lineare Abbildungen von x bzw. u und die Ableitung ist damit an jedem Punkt w der lineare Operator selbst. Um die Ableitung des ersten Terms von Gleichung (3.4.5) zu bestimmen, benötigen wir die Ableitung von $y_{j,k}$ bezüglich w aus dem Gleichungssystem (3.3.9). Aus diesem Gleichungssystem kann man nach Lemma 3.3.6 die Ableitungen ${y_{j,k}}'$ für $1 \leq k \leq s$ und hinreichend kleines $h > 0$ bestimmen. Für die Ableitung vom Ausdruck (3.4.5) bezüglich $w = (x, \lambda, u)$ an der Stelle $w \in X_{h,r}$, angewandt auf $\nu = (\nu_1, \nu_2, \nu_3) \in \tilde{X}_{h,2r}$, erhalten wir an jeder Stelle $t \in (0, T]$:

$$\begin{aligned} E_h(w)(\nu)(t) := {} & f_x(y_{j,k}, u(t_{j,k}))({y_{j,k}}'(w)(\nu)) \\ & + f_u(y_{j,k}, u(t_{j,k}))(\nu_3(t_{j,k})) - A(t)\nu_1(t) - B(t)\nu_3(t). \end{aligned} \tag{3.4.6}$$

Wir betrachten das Supremum von $E_h(w)(\nu)(t)$ bezüglich aller Punkte $\nu \in \tilde{X}_{h,2r}$ und wollen somit zeigen, dass

$$\sup_{w \in X_{h,r}} \|E_h(w)\|_{\mathcal{Z}} := \sup_{w \in X_{h,r}} \sup_{\substack{\nu \in \tilde{X}_{h,2r} \\ \nu \neq 0}} \sup_{t \in (0,T]} \|E_h(w)(\nu)(t)\|_2 / \|\nu\|_{\mathcal{X}} \leq \varepsilon_{3.4.7} \tag{3.4.7}$$

ist, und definieren hierfür

$$\begin{aligned} \triangle A(w,j,k)(t) &:= f_x(y_{j,k}, u(t_{j,k})) - A(t) \\ \triangle \nu_1(w,j,k)(t) &:= {y_{j,k}}'(w)(\nu) - \nu_1(t) \\ \triangle B(w,j,k)(t) &:= f_u(y_{j,k}, u(t_{j,k})) - B(t) \\ \triangle \nu_3(w,j,k)(t) &:= \nu_3(t_{j,k}) - \nu_3(t). \end{aligned}$$

Die Abbildungen A, $\triangle A$, B und $\triangle B$ sind zu jedem Zeitpunkt $t \in (0, T]$ Matrizen und somit lineare Abbildungen. Wir können die rechte Seite von Gleichung (3.4.6) daher als mehrere Differenzen vom Typ

$$\mathcal{A}z - \tilde{\mathcal{A}}\tilde{z} = (\mathcal{A} - \tilde{\mathcal{A}})(z - \tilde{z}) + (\mathcal{A} - \tilde{\mathcal{A}})(\tilde{z}) + (\tilde{\mathcal{A}})(z - \tilde{z}), \tag{3.4.8}$$

mit Matrizen $\mathcal{A}$ und passend definierten Vektoren z schreiben. Insgesamt müssen wir zeigen, dass die entsprechenden Matrizen $\tilde{\mathcal{A}}$ und Vektoren $\tilde{z}$ beschränkt sind und die Differenzen $\mathcal{A} - \tilde{\mathcal{A}}$ und $z - \tilde{z}$ mit hinreichend kleiner Schrittweite h und Radius r beliebig klein werden.

Nach der Voraussetzung *Glattheit* sind die Matrizen $A(t)$ und $B(t)$ beschränkt, und wir definieren daher die Abkürzungen

$$\|A\|_{\mathcal{M}} := \sup_{\substack{z\in\mathbb{R}^n\\ z\neq 0}} \sup_{t\in(0,T]} \|A(t)z\|_2/\|z\|_2$$
$$\|B\|_{\mathcal{M}} := \sup_{\substack{z\in\mathbb{R}^n\\ z\neq 0}} \sup_{t\in(0,T]} \|B(t)z\|_2/\|z\|_2.$$

Weiterhin sind natürlich mit $\nu \in \tilde{X}_{h,2r}$ die Vektoren $\nu_1(t)$ und $\nu_3(t)$ für alle $t \in (0,T]$ beschränkt, und man kann diese Vektoren durch $\|\nu\|_{\mathcal{X}}$ nach oben abschätzen. Daher braucht man nach Gleichung (3.4.8) lediglich zeigen, dass

$$\sup_{w\in X_{h,r}} \sup_{t\in(0,T]} \sup_{\substack{z\in\mathbb{R}^n\\ z\neq 0}} \|\triangle A(w,j,k)(t)z\|_2/\|z\|_2 \leq \varepsilon_3^x(h,r)$$
$$\sup_{w\in X_{h,r}} \sup_{\substack{\nu\in\tilde{X}_{h,2r}\\ \nu\neq 0}} \sup_{t\in(0,T]} \|\triangle \nu_1(w,j,k)(t)\|_2/\|\nu\|_{\mathcal{X}} \leq \varepsilon_3^x(h,r)$$
$$\sup_{w\in X_{h,r}} \sup_{t\in(0,T]} \sup_{\substack{z\in\mathbb{R}^n\\ z\neq 0}} \|\triangle B(w,j,k)(t)z\|_2/\|z\|_2 \leq \varepsilon_3^x(h,r)$$
$$\sup_{w\in X_{h,r}} \sup_{\substack{\nu\in\tilde{X}_{h,2r}\\ \nu\neq 0}} \sup_{t\in(0,T]} \|\triangle \nu_3(w,j,k)(t)\|_2/\|\nu\|_{\mathcal{X}} \leq \varepsilon_3^x(h,r)$$

sind, wobei $\varepsilon_3^x(h,r)$ mit $h,r \to 0$ beliebig klein wird und ansonsten unabhängig von $w \in X_{h,r}$, $\nu \in \tilde{X}_{h,2r}$, t, k und j ist.

Wir betrachten zunächst $\triangle A(w,j,k)(t)$, mit $A(t) = f_x(x^*(t),u^*(t))$ gilt

$$\triangle A(w,j,k)(t) := f_x(y_{j,k},u(t_{j,k})) - f_x(x^*(t),u^*(t)).$$

Nach der Voraussetzung *Glattheit* sind die ersten κ Ableitungen von f lipschitzstetig mit einer Lipschitzkonstanten M_f^k, $k = 0,\ldots,\kappa$, und natürlich lineare Abbildungen, damit gilt

$$\sup_{w\in X_{h,r}} \sup_{t\in(0,T]} \sup_{\substack{z\in\mathbb{R}^n\\ z\neq 0}} \|\triangle A(w,j,k)(t)z\|_2/\|z\|_2 \leq \sup_{w\in X_{h,r}} \sup_{t\in(0,T]} M_f^1\big(\|y_{j,k}-x^*(t)\|_2 + \|u(t_{j,k})-u^*(t)\|_2\big).$$

Für die Terme $\|y_{j,k}-x^*(t)\|_2$ und $\|u(t_{j,k})-u^*(t)\|_2$ gilt für jedes $w=(x,\lambda,u)\in X_{h,r}$ nach Lemma 3.3.5

$$\begin{aligned}\|y_{j,k}-x^*(t)\|_2 &= \|y_{j,k}-x(t_j)+x(t_j)-x^*(t_j)+x^*(t_j)-x^*(t)\|_2\\ &\leq \|y_{j,k}-x(t_j)\|_2 + \|x(t_j)-x^*(t_j)\|_2 + \|x^*(t_j)-x^*(t)\|_2\\ &\leq h\zeta_{3.3.8} + r + h\|\dot{x}^*\|_{L^\infty} \qquad (3.4.9)\end{aligned}$$

und

$$\begin{aligned}\|u(t_{j,k})-u^*(t)\|_2 &= \|u(t_{j,k})-u^*(t_{j,k})+u^*(t_{j,k})-u^*(t)\|_2\\ &\leq \|u(t_{j,k})-u^*(t_{j,k})\|_2 + \|u^*(t_{j,k})-u^*(t)\|_2\\ &\leq r + h\|\dot{u}^*\|_{L^\infty}\ . \qquad (3.4.10)\end{aligned}$$

Daher können wir die Operatornorm zu $\triangle A(w,j,k)(t)$ durch $M_f^1(h\zeta_{3.3.8} + 2r + h\|\dot{x}^*\|_{L^\infty} + h\|\dot{u}^*\|_{L^\infty})$ abschätzen.

Weiterhin gilt für den nächsten Term

$$\sup_{w\in X_{h,r}} \sup_{\substack{\nu\in\tilde{X}_{h,2r}\\ \nu\neq 0}} \sup_{t\in(0,T]} \|\triangle\nu_1(w,j,k)(t)\|_2/\|\nu\|_{\mathcal{X}} = \sup_{w\in X_{h,r}} \sup_{\substack{\nu\in\tilde{X}_{h,2r}\\ \nu\neq 0}} \sup_{t\in(0,T]} \|{y_{j,k}}'(w)(\nu) - \nu_1(t)\|_2/\|\nu\|_{\mathcal{X}}.$$

Nach Gleichung (3.3.11) aus dem Beweis von Lemma 3.3.6 gilt

$$\|{y_{j,k}}'(w)(\nu) - \nu_1(t_j)\|_2 \leq \frac{1}{1-\zeta_{3.3.9}} \max_{k=1,2,\dots,s} \|h\sum_{l=1}^{s} a_{k,l}[f_x(y_{j,l},u(t_{j,l}))(\nu_1(t_j)) + f_u(y_{j,l},u(t_{j,l}))(\nu_3(t_{j,l}))]\|_2,$$

Damit folgt

$$\begin{aligned}
&\sup_{w\in X_{h,r}} \sup_{\substack{\nu\in\tilde{X}_{h,2r}\\ \nu\neq 0}} \sup_{t\in(0,T]} \|{y_{j,k}}'(w)(\nu) - \nu_1(t)\|_2/\|\nu\|_{\mathcal{X}} \\
&\leq \sup_{w\in X_{h,r}} \sup_{\substack{\nu\in\tilde{X}_{h,2r}\\ \nu\neq 0}} \sup_{t\in(0,T]} \left(\|{y_{j,k}}'(w)(\nu) - \nu_1(t_j)\|_2 + \|\nu_1(t_j) - \nu_1(t)\|_2\right)/\|\nu\|_{\mathcal{X}} \\
&\leq \frac{h}{1-\zeta_{3.3.9}} \sup_{w\in X_{h,r}} \sup_{\substack{\nu\in\tilde{X}_{h,2r}\\ \nu\neq 0}} \sup_{t\in(0,T]} \max_{k=1,2,\dots,s} \Big\| \sum_{l=1}^{s} a_{k,l}[f_x(y_{j,l},u(t_{j,l}))(\nu_1(t_j)) \\
&\qquad + f_u(y_{j,l},u(t_{j,l}))(\nu_3(t_{j,l}))]\Big\|_2/\|\nu\|_{\mathcal{X}} + h,
\end{aligned}$$

da für die zweite Differenz $\|\nu_1(t_j) - \nu_1(t)\|_2 \leq h\|\dot{\nu}\|_{L^\infty} \leq h\|\nu\|_{\mathcal{X}}$ gilt. In dem ersten Term schätzen wir jeden einzelnen Summanden ab, da die Parameter $a_{k,l}$ beschränkt sind. Weiterhin haben wir den Faktor h vor dem Term stehen und müssen lediglich zeigen, dass die Terme $f_x(\tilde{y}_{j,l},u(t_{j,l}))$ und $f_u(\tilde{y}_{j,l},u(t_{j,l}))$ beschränkt sind. Dazu benutzen wir

$$\begin{aligned}
f_x(y_{j,l},u(t_{j,l})) &= f_x(y_{j,l},u(t_{j,l})) - f_x(x^*(t_{j,l}),u^*(t_{j,l})) + f_x(x^*(t_{j,l}),u^*(t_{j,l})) \\
f_u(y_{j,l},u(t_{j,l})) &= f_u(y_{j,l},u(t_{j,l})) - f_u(x^*(t_{j,l}),u^*(t_{j,l})) + f_u(x^*(t_{j,l}),u^*(t_{j,l}))
\end{aligned}$$

und erhalten mit der Lipschitzstetigkeit von f_x und f_u und mit

$$\begin{aligned}
\|f_x(x^*(t_{j,l}),u^*(t_{j,l}))(\nu_1(t_j))\|_2 &\leq \|A\|_{\mathcal{M}}\|\nu\|_{\mathcal{X}}, \\
\|f_u(x^*(t_{j,l}),u^*(t_{j,l}))(\nu_3(t_{j,l}))\|_2 &\leq \|B\|_{\mathcal{M}}\|\nu\|_{\mathcal{X}}
\end{aligned}$$

die Abschätzung

$$\begin{aligned}
&\sup_{w\in X_{h,r}} \sup_{\substack{\nu\in\tilde{X}_{h,2r}\\ \nu\neq 0}} \sup_{t\in(0,T]} \|f_x(y_{j,l},u(t_{j,l}))(\nu_1(t_j)) + f_u(y_{j,l},u(t_{j,l}))(\nu_3(t_{j,l}))\|_2/\|\nu\|_{\mathcal{X}} \\
&\leq \|A\|_{\mathcal{M}} + \|B\|_{\mathcal{M}} + 2M_f^1 \sup_{w\in X_{h,r}} \sup_{t\in(0,T]} \left[\|y_{j,l} - x^*(t_{j,l})\|_2 + \|u(t_{j,l}) - u^*(t_{j,l})\|_2\right].
\end{aligned}$$

Analog zu den Abschätzungen in den Gleichungen (3.4.9) und (3.4.10) gilt

$$\begin{aligned}\|y_{j,l} - x^*(t_{j,l})\|_2 &\le \|y_{j,l} - x(t_j) + x(t_j) - x^*(t_{j,l})\|_2\\ &\le h\zeta_{3.3.8} + r + h\|\dot{x}^*\|_{L^\infty},\\ \|u(t_{j,l}) - u^*(t_{j,l})\|_2 &\le r\end{aligned}$$

und mit $\bar{c}_k := \sum_{l=1}^{s} |a_{k,l}|$ insgesamt

$$\begin{aligned}&\sup_{w\in X_{h,r}} \sup_{\substack{\nu\in\tilde{X}_{h,2r}\\ \nu\neq 0}} \sup_{t\in(0,T]} \|\triangle\nu_1(w,j,k)(t)\|_2/\|\nu\|_{\mathcal{X}}\\ &\quad\le h\Big[1 + \frac{\max_{k=1,2,\dots,s}\bar{c}_k}{1-\zeta_{3.3.9}}\big(\|A\|_{\mathcal{M}} + \|B\|_{\mathcal{M}} + 2M_f^1(h\zeta_{3.3.8} + 2r + h\|\dot{x}^*\|_{L^\infty})\big)\Big].\end{aligned}$$

Den Term $\triangle B(w,j,k)(t)$ schätzen wir mit $B(t) = f_u(x^*(t),u^*(t))$ und der Lipschitzstetigkeit von f_u wie folgt ab

$$\begin{aligned}&\sup_{w\in X_{h,r}} \sup_{t\in(0,T]} \sup_{\substack{z\in\mathbb{R}^n\\ z\neq 0}} \|\triangle B(w,j,k)(t)z\|_2/\|z\|_2\\ &\quad= \sup_{w\in X_{h,r}} \sup_{t\in(0,T]} \sup_{\substack{z\in\mathbb{R}^n\\ z\neq 0}} \|[f_u(y_{j,k},u(t_{j,k})) - f_u(x^*(t),u^*(t))]z\|_2/\|z\|_2\\ &\qquad\le \sup_{w\in X_{h,r}} \sup_{t\in(0,T]} M_f^1\left(\|y_{j,k} - x^*(t)\|_2 + \|u(t_{j,k}) - u^*(t)\|_2\right)\\ &\qquad\qquad\le M_f^1(h\zeta_{3.3.8} + 2r + h\|\dot{x}^*\|_{L^\infty} + h\|\dot{u}^*\|_{L^\infty}),\end{aligned}$$

wobei die letzte Ungleichung mit der gleichen Abschätzung wie in (3.4.9) und (3.4.10) folgt. Schließlich gilt für $\triangle\nu_3(w,j,k)(t)$ ebenfalls

$$\begin{aligned}&\sup_{w\in X_{h,r}} \sup_{\substack{\nu\in\tilde{X}_{h,2r}\\ \nu\neq 0}} \sup_{t\in(0,T]} \|\triangle\nu_3(w,j,k)(t)\|_2/\|\nu\|_{\mathcal{X}}\\ &\qquad= \sup_{w\in X_{h,r}} \sup_{\substack{\nu\in\tilde{X}_{h,2r}\\ \nu\neq 0}} \sup_{t\in(0,T]} \|\nu_3(t_{j,k}) - \nu_3(t)\|_2/\|\nu\|_{\mathcal{X}} \le h,\end{aligned}$$

weil aus $0 \le c_k \le 1$ folgt $|t_{j,k} - t| \le h$, und damit folgt $\|\nu_3(t_{j,k}) - \nu_3(t)\|_2 \le h\|\nu\|_{\mathcal{X}}$. Damit hat man insgesamt gezeigt, dass es in Ungleichung (3.4.7) ein $\varepsilon_{3.4.7}$ gibt mit $\varepsilon_{3.4.7} \to 0$ für $h,\ r \to 0$. Man hat ebenfalls gezeigt, dass man die Ungleichung (3.4.2) und damit Bedingung (B3) mit hinreichend kleinem h und r für jedes $\varepsilon > 0$ für die ersten Komponenten des Operators $F - P_h$ erfüllen kann.

Die zweiten Komponenten des Operators $(F - P_h)$ schätzen wir analog ab. Bildet man die Ableitungen bezüglich der Zeit t, um die entsprechenden $W^{1,\infty}$-Normen abzuschätzen, kann man genau wie bei den ersten Komponenten vorgehen. Möchte man eine entsprechende Abschätzung für (3.4.2) zeigen, kann man wieder die entsprechende Ableitung $(F - P_h)'(w)$ von

$$f_x(x(\overline{t}_{j,k}),u(\overline{t}_{j,k}))^\top v_{j,k} - Q(\overline{t})x(\overline{t}) - S(\overline{t})u(\overline{t}) - A(\overline{t})^\top\lambda(t) \tag{3.4.11}$$

bestimmen. Dabei gilt für den Spaltenvektor

$$f_x(x(\overline{t}_{j,k}), u(\overline{t}_{j,k}))^\top v_{j,k} = H_x(x(\overline{t}_{j,k}), v_{j,k}(w), u(\overline{t}_{j,k})),$$

und für die Ableitung dieses Vektors bzgl. w, angewandt auf ein $\nu \in \tilde{X}_{h,2r}$, erhalten wir

$$\begin{aligned} H_{xx}(x(\overline{t}_{j,k}), v_{j,k}(w), u(\overline{t}_{j,k}))\nu_1(\overline{t}_{j,k}) + H_{ux}(x(\overline{t}_{j,k}), v_{j,k}(w), u(\overline{t}_{j,k}))\nu_3(\overline{t}_{j,k}) \\ + f_x(x(\bar{t}_{j,k}), u(\bar{t}_{j,k}))^\top {v_{j,k}}'(w)(\nu). \end{aligned}$$

Dabei berechnen sich die Zwischenpunkte $v_{j,k}$ nach Gleichung (3.3.5), und daraus bestimmen wir die Ableitung analog zu Gleichungssystem (3.3.9) aus dem Gleichungssystem

$$\begin{aligned} {v_{j,k}}'(w)(\nu) = \nu_2(t_j) + h \sum_{l=1}^{s} a_{k,l} \Big(& H_{xx}(x(\overline{t}_{j,l}), v_{j,l}(w), u(\overline{t}_{j,l}))\nu_1(\overline{t}_{j,l}) \\ & + H_{ux}(x(\overline{t}_{j,l}), v_{j,l}(w), u(\overline{t}_{j,l}))\nu_3(\overline{t}_{j,l}) + f_x(x(\bar{t}_{j,l}), u(\bar{t}_{j,l}))^\top {v_{j,l}}'(w)(\nu) \Big). \end{aligned}$$

Weiterhin ist der Term

$$Q(\bar{t})x(\bar{t}) + S(\bar{t})u(\bar{t}) + A(\bar{t})^\top \lambda(t)$$

linear von $w = (x, \lambda, u)$ abhängig, und es ergibt sich für die Ableitung bezüglich w, angewandt auf ein $\nu \in \tilde{X}_{h,2r}$,

$$Q(\bar{t})\nu_1(\bar{t}) + S(\bar{t})\nu_3(\bar{t}) + A(\bar{t})^\top \nu_2(t).$$

Wir erhalten insgesamt als Ableitung von Gleichung (3.4.11) bezüglich $w \in X_{h,r}$, angewandt auf ein $\nu \in \tilde{X}_{h,2r}$, für jeden Zeitpunkt t

$$\begin{aligned} E_h^2(w)(\nu)(t) := {} & H_{xx}(x(\overline{t}_{j,k}), v_{j,k}(w), u(\overline{t}_{j,k}))\nu_1(\overline{t}_{j,k}) \\ & - Q(\bar{t})\nu_1(\bar{t}) + H_{ux}(x(\overline{t}_{j,k}), v_{j,k}(w), u(\overline{t}_{j,k}))\nu_3(\overline{t}_{j,k}) - S(\bar{t})\nu_3(\bar{t}) \\ & + f_x(x(\bar{t}_{j,k}), u(\bar{t}_{j,k}))^\top {v_{j,k}}'(w)(\nu) - A(\bar{t})^\top \nu_2(t), \end{aligned} \tag{3.4.12}$$

wobei

$$\begin{aligned} A(t) &= f_x(x^*(t), u^*(t)), \\ Q(t) &= H_{xx}(w^*(t)) = H_{xx}((x^*(t), \lambda^*(t), u^*(t))), \\ S(t) &= H_{ux}(w^*(t)) = H_{ux}((x^*(t), \lambda^*(t), u^*(t))) \end{aligned}$$

sind. Damit muss man zeigen, dass

$$\sup_{w \in X_{h,r}} \sup_{\substack{\nu \in \tilde{X}_{h,2r} \\ \nu \neq 0}} \sup_{t \in (0,T]} \|E_h^2(w)(\nu)(t)\|_2 / \|\nu\|_{\mathcal{X}_h} \leq \varepsilon_{3.4.13} \tag{3.4.13}$$

für beliebig kleines $\varepsilon_{3.4.13} > 0$ gilt, wenn man h und r hinreichend klein wählt. Wir wenden wiederum die Argumentation von Gleichung (3.4.8) an, und wollen

jeden der s Summanden, die sich aus dem s-stufigen Runge-Kutta-Verfahren bei der Berechnung von $v_{j,k}'(w)(\nu)$ ergeben, und dort jede der drei zusammenpassenden Differenzen der Art „$H_{xx}\nu_1 - Q\nu_1$“ in Gleichung (3.4.12) nach oben abschätzen. Dazu halten wir fest, dass aufgrund von *Glattheit*

$$\begin{aligned}
\|Q\|_{\mathcal{M}} &:= \sup_{\substack{\nu \in \tilde{X}_{h,2r} \\ \nu \neq 0}} \sup_{t\in(0,T]} \|Q(t)\nu_1(t)\|_2/\|\nu\|_{\mathcal{X}} < \infty \\
\|S\|_{\mathcal{M}} &:= \sup_{\substack{\nu \in \tilde{X}_{h,2r} \\ \nu \neq 0}} \sup_{t\in(0,T]} \|S(t)\nu_3(t)\|_2/\|\nu\|_{\mathcal{X}} < \infty \\
\|A^\top\|_{\mathcal{M}} &:= \sup_{\substack{\nu \in \tilde{X}_{h,2r} \\ \nu \neq 0}} \sup_{t\in(0,T]} \|A(t)^\top\nu_2(t)\|_2/\|\nu\|_{\mathcal{X}} < \infty
\end{aligned}$$

gilt und $\nu \in \tilde{X}_{h,2r}$ beschränkt ist. Somit genügt wiederum zu zeigen, dass jeweils die Norm der entsprechenden Differenzen von Matrizen und Vektoren in Gleichung (3.4.12) für $h, r \to 0$ gegen Null geht. Betrachtet man

$$H_{xx}(x(\overline{t}_{j,k}), v_{j,k}(w), u(\overline{t}_{j,k}))\nu_1(\overline{t}_{j,k}) - Q(\bar{t})\nu_1(\bar{t}),$$

dann gilt mit $\nu \in \tilde{X}_{h,2r}$, dass ν_1 lipschitzstetig mit Lipschitzkonstante kleiner oder gleich $\|\nu\|_{\mathcal{X}}$ ist. Mit $\bar{t} = T - (t_j + \theta h)$ und $\bar{t}_{j,k} = T - (t_j + c_k h)$ folgt

$$\|\nu_1(\overline{t}_{j,k}) - \nu_1(\bar{t})\|_2 \leq h\|\nu\|_{\mathcal{X}}.$$

Die Abbildung $H_{xx}((x(t), \lambda(t), u(t)))$ kann man als partielle Ableitung von

$$H_x((x(t), \lambda(t), u(t))) = f_x(x(t), u(t))^\top \lambda(t)$$

analog zu $Q(t)$ als Matrix anordnen, wobei an der (i, j)-ten Stelle der $n \times n$ Matrix der Eintrag

$$\frac{\partial}{\partial x_i \partial x_j} f(x(t), u(t))^\top \lambda(t) = \sum_{l=1}^{n} \frac{\partial}{\partial x_i \partial x_j} f^l(x(t), u(t))\lambda^l(t)$$

steht. Hierbei bezeichnen f^l und λ^l die l-ten Komponenten der vektorwertigen Funktionen f und λ. Definiert man mit f^l_{xx} eine entsprechende Matrix der zweiten Ableitungen der l-ten Komponente von f, kann man die zweiten Ableitungen von H auch als

$$\sum_{l=1}^{n} f^l_{xx}(x(t), u(t))\lambda^l(t)$$

darstellen. Mit der Voraussetzung, dass für jede Komponente von f die zweiten Ableitungen lipschitzstetig von der Stelle abhängen, an der die Ableitung ausgewertet wird, der Äquivalenz der Normen auf einem endlichdimensionalen Vektorraum

($\sum_{l=1}^{n} |\lambda^l(t)| \leq \zeta_{3.4.14} \|\lambda(t)\|_2$) und wiederum mit einer Argumentation entsprechend Gleichung (3.4.8) folgert man

$$
\begin{aligned}
(3.4.14) \quad & \sup_{\substack{z\in\mathbb{R}^n \\ z\neq 0}} \|\big(H_{xx}(x(\overline{t}_{j,k}), v_{j,k}(w), u(\overline{t}_{j,k})) - Q(\bar{t})\big)z\|_2 / \|z\|_2 \\
&= \sup_{\substack{z\in\mathbb{R}^n \\ z\neq 0}} \left\| \sum_{l=1}^{n} \big(f^l_{xx}(x(\overline{t}_{j,k}), u(\overline{t}_{j,k}))v^l_{j,k}(w) - f^l_{xx}(x^*(\bar{t}), u^*(\bar{t}))\lambda^{*l}(t)\big)z \right\|_2 / \|z\|_2 \\
&\leq M_f^2 \left(\|x(\overline{t}_{j,k}) - x^*(\bar{t})\|_2 + \|u(\overline{t}_{j,k}) - u^*(\bar{t})\|_2\right) \zeta_{3.4.14} \|v_{j,k}(w) - \lambda^*(t)\|_2 \\
&\quad + \zeta_{3.4.14} \|v_{j,k}(w) - \lambda^*(t)\|_2 \sup_{\substack{z\in\mathbb{R}^n \\ z\neq 0}} \sum_{l=1}^{n} \|f^l_{xx}(x^*(\bar{t}), u^*(\bar{t}))z\|_2 / \|z\|_2 \\
&\quad + M_f^2 \left(\|x(\overline{t}_{j,k}) - x^*(\bar{t})\|_2 + \|u(\overline{t}_{j,k}) - u^*(\bar{t})\|_2\right) \zeta_{3.4.14} \|\lambda^*(t)\|_2,
\end{aligned}
$$

wobei $\bar{t} := T - t$ ist. Aus *Glattheit* aus Abschnitt 3.1 folgt, dass mit einem $\zeta_{3.4.15}$

$$
(3.4.15) \qquad \sup_{t\in(0,T]} \sup_{\substack{z\in\mathbb{R}^n \\ z\neq 0}} \sum_{l=1}^{n} \left\| f^l_{xx}(x^*(\bar{t}), u^*(\bar{t}))z \right\|_2 / \|z\|_2 \leq \zeta_{3.4.15}
$$

gilt, da x^*, u^* beschränkt sind und die zweiten Ableitungen von f lipschitzstetig sind. Weiterhin gelten folgende Abschätzungen

$$
\begin{aligned}
\|x(\overline{t}_{j,k}) - x^*(\bar{t})\|_2 &= \|x(\overline{t}_{j,k}) - x^*(\overline{t}_{j,k}) + x^*(\overline{t}_{j,k}) - x^*(\bar{t})\|_2 \\
&\leq r + h\|\dot{x}^*\|_{L^\infty}, \\
\|u(\overline{t}_{j,k}) - u^*(\bar{t})\|_2 &= \|u(\overline{t}_{j,k}) - u^*(\overline{t}_{j,k}) + u^*(\overline{t}_{j,k}) - u^*(\bar{t})\|_2 \\
&\leq r + h\|\dot{u}^*\|_{L^\infty}.
\end{aligned}
$$

Um die Zwischenpunkte $v_{j,k}(w)$ abschätzen zu können, wenden wir Lemma 3.3.5 und 3.3.6 an und erhalten mit Gleichung (3.3.11) und einer Konstanten $\zeta_{3.4.16}$ für alle $j = 0, 1, \ldots, N-1$ und $k = 1, 2, \ldots, s$

$$
(3.4.16) \qquad \|v_{j,k}(w) - \lambda(t_j)\|_2 \leq h\zeta_{3.4.16},
$$

und damit gilt

$$
\begin{aligned}
\|v_{j,k}(w) - \lambda^*(t)\|_2 &= \|v_{j,k}(w) - \lambda(t_j) + \lambda(t_j) - \lambda^*(t_j) + \lambda^*(t_j) - \lambda^*(t)\|_2 \\
&\leq h\zeta_{3.4.16} + r + h\|\dot{\lambda}^*\|_{L^\infty}.
\end{aligned}
$$

Somit hat man insgesamt gezeigt, dass die erste Differenz in (3.4.12) beliebig klein wird, wenn man h und r hinreichend klein wählt. Völlig analog kann man dies für die zweite Differenz in (3.4.12) zeigen. Für die dritte Differenz in (3.4.12) benutzt man

$$
\begin{aligned}
(3.4.17) \quad & \|v_{j,k}{}'(w)(\nu) - \nu_2(t_j)\|_2 \\
&\leq \frac{h}{1 - \zeta_{3.4.17}} \max_{k=1,2,\ldots,s} \Big\| \sum_{l=1}^{s} a_{k,l} \Big(H_{xx}(x(\overline{t}_{j,l}), v_{j,l}(w), u(\overline{t}_{j,l}))\nu_1(\overline{t}_{j,l}) \\
&\qquad + H_{ux}(x(\overline{t}_{j,l}), v_{j,l}(w), u(\overline{t}_{j,l}))\nu_3(\overline{t}_{j,l}) + f_x(x(\bar{t}_{j,l}), u(\bar{t}_{j,l}))^\top \nu_2(t_j) \Big) \Big\|_2
\end{aligned}
$$

aus dem Beweis von Lemma 3.3.6. Mit den obigen Abschätzungen kann man zeigen, dass es ein $h_{3.4.18} > 0$ und $r_{3.4.18} > 0$ gibt, so dass die Summe auf der rechten Seite von (3.4.17) für alle $0 < h \leq h_{3.4.18}$ und $0 < r \leq r_{3.4.18}$ durch eine Konstante beschränkt ist, die wiederum nur von den entsprechenden Lipschitzkonstanten von f, w^*, $h_{3.4.18}$ und $r_{3.4.18}$ abhängt und für $r \to 0$, $h \to 0$ gegen 0 geht. Damit hat man ein entsprechendes Resultat ebenfalls für die dritte Differenz aus (3.4.12) und es gilt insgesamt

$$\begin{aligned}(3.4.18)\qquad \forall \varepsilon_{3.4.13} > 0 \quad \exists\, h_{3.4.18}, r_{3.4.18} > 0 : \forall\, 0 < h \leq h_{3.4.18} \wedge 0 < r \leq r_{3.4.18} : \\ \sup_{w \in X_{h,r}} \sup_{\substack{\nu \in \tilde{X}_{h,2r} \\ \nu \neq 0}} \sup_{t \in (0,T]} \|E_h^2(w)(\nu)(t)\|_2 / \|\nu\|_{\mathcal{X}} \leq \varepsilon_{3.4.13}.\end{aligned}$$

Da die letzten Komponenten von $F - P_h$ eine Bedingung an den Anfangswert für die zweiten Komponenten darstellen, wollen wir diese Komponenten jetzt betrachten, es gilt

$$\begin{aligned}\lambda(0) - Vx(T) - \lambda(0) + \Psi_x(x(T))^\top &= \Psi_x(x(T))^\top - Vx(T) \\ &= \Psi_x(x(T))^\top - \Psi_{xx}(x^*(T))(x(T)).\end{aligned}$$

Für die Ableitung bezüglich w an der Stelle $w \in X_{h,r}$, angewandt auf $\nu \in \tilde{X}_{h,2r}$, ergibt sich

$$\Psi_{xx}(x(T))(\nu_1(T)) - \Psi_{xx}(x^*(T))(\nu_1(T)) = [\Psi_{xx}(x(T)) - \Psi_{xx}(x^*(T))](\nu_1(T)),$$

und mit der Voraussetzung, dass Ψ_{xx} lipschitzstetig ist, konvergiert die Differenz mit $r \to 0$ gegen 0.

Für die dritten Komponenten von $(F - P_h)(w)(t)$ ist die, in Bedingung (B3) von Lemma 3.3.1 geforderte, Lipschitzstetigkeit mit einer hinreichend kleinen Lipschitzkonstanten leicht zu sehen. Man benutzt dazu lediglich die Lipschitzstetigkeit der Ableitungen von f. Die Abschätzungen der entsprechenden Ableitungen führen wir daher hier nicht detailliert auf. Die dritten Komponenten von $(F - P_h)(w)(t)$ sind

$$(3.4.19)\qquad R(t)^\top u(t) + S(t)^\top x(t) + B(t)^\top \lambda(T-t) - f_u(x(t), u(t))^\top \lambda(T-t),$$

wobei

$$\begin{aligned}R(t) &= \sum_{l=1}^{n} f_{uu}^l(x^*(t), u^*(t))\lambda^{*l}(t), \\ S(t) &= \sum_{l=1}^{n} f_{ux}^l(x^*(t), u^*(t))\lambda^{*l}(t), \\ B(t) &= f_u(x^*(t), u^*(t))\end{aligned}$$

sind. Für diesen Ausdruck muss man ebenfalls die Lipschitzstetigkeit bezüglich w in den entsprechenden $W^{1,\infty}$-Normen zeigen. Bestimmt man von dem Term (3.4.19)

die Ableitung bezüglich w an einer Stelle $w \in X_{h,r}$, angewandt auf ein $\nu \in \tilde{X}_{h,2r}$ erhält man

$$\begin{aligned}(3.4.20)\quad & R(t)^\top \nu_3(t) + S(t)^\top \nu_1(t) + B(t)^\top \nu_2(T-t) \\ & - \Big(\sum_{l=1}^{n} f_{uu}^l(x(t),u(t))\lambda^l(t)\Big)^\top \nu_3(t) - \Big(\sum_{l=1}^{n} f_{ux}^l(x(t),u(t))\lambda^l(t)\Big)^\top \nu_1(t) \\ & \qquad - f_u(x(t),u(t))^\top \nu_2(T-t),\end{aligned}$$

und es ist offensichtlich, dass aufgrund der Lipschitzstetigkeit der zweiten Ableitungen man die L^∞-Norm von diesem Term mit der Wahl eines hinreichend kleinen $r_{3.4.20}$ für alle $0 < r < r_{3.4.20}$ unabhängig von der Wahl von h unter jede fest vorgegebene Schranke bekommt. Berechnet man zu (3.4.20) die Ableitung nach der Zeit t, bzw. berechnet die Ableitung nach t vom Ausdruck (3.4.19), wendet den Mittelwertsatz an und benutzt die Lipschitzstetigkeit der dritten Ableitungen von f, so kann man die entstehenden Differenzen leicht mit Hilfe der Terme $\|x - x^*\|_{W^{1,\infty}}$, $\|\lambda - \lambda^*\|_{W^{1,\infty}}$ und $\|u - u^*\|_{W^{1,\infty}}$ abschätzen. Dabei sind mit $\|w - w^*\|_{\mathcal{X}} \leq r$ diese Terme für hinreichend kleines r beliebig klein. Insgesamt kann man für jedes $\varepsilon > 0$ mit der Wahl von hinreichend kleinen $\bar{r}$ und $\bar{h}$ für alle $0 < r < \bar{r}$ und $0 < h < \bar{h}$ die Bedingung (B3) von Lemma 3.3.1 erfüllen.

3.5 Invertierbarkeit des linearisierten Problems

Damit wir die Bedingung (B1) von Lemma 3.3.1 zeigen können, skizzieren wir zunächst den üblichen Weg, die Norm der Lösung eines linearen Differentialgleichungssystems durch die Ungleichung von Gronwall abzuschätzen (siehe zum Beispiel [Ama95]). In dem Operator F kommen lineare Differentialgleichungssysteme vom Typ

$$(3.5.1)\qquad \dot{x}(t) = A(t)x(t) + \chi(t)$$

mit $x(0) = a$ und $t \in [0,T]$ bzw. eine entsprechende Integralgleichung

$$x(t) = a + \int_0^t A(\tau)x(\tau) + \chi(\tau)\, d\tau$$

vor. Um die Lösung der Integralgleichung durch die Norm der Funktion χ abschätzen zu können bildet man auf beiden Seiten die euklidische Norm. Mit

$$\|A(\tau)x(\tau)\|_2 \leq \|A\|_{\mathcal{M}} \|x(\tau)\|_2$$

gewinnt man die Ungleichung

$$(3.5.2)\qquad \|x(t)\|_2 \leq \|a\|_2 + \left\| \int_0^t \chi(\tau)\, d\tau \right\|_2 + \int_0^t \|A\|_{\mathcal{M}} \|x(\tau)\|_2 \, d\tau.$$

Somit kann man für $\|x(t)\|_2$ in (3.5.2) das Lemma von Gronwall anwenden und erhält

$$\|x(t)\|_2 \leq \left(\|a\|_2 + \left\| \int_0^t \chi(\tau)\, d\tau \right\|_2 \right) + \int_0^t \left(\|a\|_2 + \left\| \int_0^\tau \chi(\vartheta)\, d\vartheta \right\|_2 \right) \|A\|_{\mathcal{M}}\, e^{T\|A\|_{\mathcal{M}}}\, d\tau,$$

da $\int_{t_1}^{t_2} \|A\|_{\mathcal{M}}\, dt \leq T\, \|A\|_{\mathcal{M}}$ für alle $0 \leq t_1 \leq t_2 \leq T$ gilt. Mit der Abschätzung

$$\left\| \int_0^\tau \chi(\vartheta)\, d\vartheta \right\|_2 \leq \int_0^\tau \|\chi(\vartheta)\|_2\, d\vartheta \leq \int_0^T \|\chi(\vartheta)\|_2\, d\vartheta = \|\chi\|_{L^1}$$

erhält man für $a = 0_n$

$$\|x(t)\|_2 \leq \|\chi\|_{L^1} \left(1 + T\, \|A\|_{\mathcal{M}}\, e^{T\|A\|_{\mathcal{M}}}\right) \tag{3.5.3}$$

und damit die Beschränktheit von x. Offensichtlich gilt dann auch

$$\|x\|_{L^\infty} \leq \sup_{t\in[0,T]} \|x(t)\|_2 \leq \|\chi\|_{L^1} \left(1 + T\, \|A\|_{\mathcal{M}}\, e^{T\|A\|_{\mathcal{M}}}\right).$$

Ebenfalls kann man mit einer Konstanten ζ die L^1-Norm von x abschätzen und erhält eine Ungleichung $\|x\|_{L^1} \leq \zeta\, \|\chi\|_{L^1}$, die wir jedoch im Folgenden in dieser Form nicht benötigen. Da wir in unserem Fall nur Funktionen auf einem Intervall $[0, T]$ betrachten, gilt nach der Hölder-Ungleichung $\|\chi\|_{L^q} \leq \zeta \|\chi\|_{L^p}$ und $L^p \subset L^q$ für $p > q \geq 1$ mit einer von χ unabhängigen Konstanten ζ, sofern $\chi \in L^p$ gilt. Damit erhält man sofort auch Abschätzungen der Form

$$\|x\|_{L^p} \leq \zeta_{3.5.4}\, \|\chi\|_{L^p} \tag{3.5.4}$$

mit einer von x und χ unabhängigen Konstanten $\zeta_{3.5.4}$. Weiterhin gilt natürlich aufgrund der Gleichung (3.5.1) für $1 \leq p \leq \infty$ für $a = 0_n$ insgesamt

$$\|\dot{x}\|_{L^p} \leq \zeta_{3.5.5} \|\chi\|_{L^p} \tag{3.5.5}$$

mit einer von χ und x unabhängigen Konstanten $\zeta_{3.5.5}$, sofern $A(\cdot) \in L^p$ gilt.

3.5.1 Invertierbarkeit von F

Mit Hilfe dieser Vorbemerkungen zeigen wir als nächstes, dass die Bedingung (B1) von Lemma 3.3.1 erfüllt ist. Dazu müssen wir zeigen, dass die inverse Abbildung von F existiert, lipschitzstetig mit Lipschitzkonstante c_0 ist und für jedes $\pi \in \mathcal{Y}$ gilt $F^{-1}(\pi) \in X_h$. Wir wollen daher zeigen, dass

$$\forall \pi \in \mathcal{Y}\ \exists^1 w \in X_h\ : F(w) - \pi = 0_{\mathcal{Y}} \tag{3.5.6}$$

gilt. Dazu betrachten wir für ein gegebenes π folgendes Optimierungsproblem

$$\begin{aligned}(\text{QLP}) \qquad \min \quad & \mathcal{B}(x,u) + \pi_4^\top x(T) - \int_0^T \dot{\pi}_2(t)^\top x(t)dt - \int_0^T \pi_3(t)^\top u(t)dt \\ u.Nb. \quad & \dot{x}(t) - A(t)x(t) - B(t)u(t) - \dot{\pi}_1(t) = 0, \quad \forall t \in (0,T], \\ & x(0) = a\end{aligned}$$

und bestimmen dazu die notwendigen Optimalitätsbedingungen 1. Ordnung:

$$\begin{aligned}&\dot{x}(t) - A(t)x(t) - B(t)u(t) - \dot{\pi}_1(t) = 0,\\ &\dot{\lambda}(t) + Q(t)x(t) + S(t)u(t) + A(t)\lambda(t) - \dot{\pi}_2(t) = 0,\\ &R(t)^\top u(t) + S(t)^\top x(t) + B(t)^\top \lambda(t) - \pi_3(t) = 0,\end{aligned}$$

wobei $t \in [0,T]$ ist und man die Anfangswerte $x(0) = a$ und $\lambda(T) = Vx(T) + \pi_4$ für die Differentialgleichungen gegeben hat. Ersetzt man die zweite Gleichung mit Hilfe der Substitutionen $\lambda(t) = \bar{\lambda}(T-t)$, $\dot{\pi}_2(t) = -\dot{\bar{\pi}}_2(T-t)$ und $\bar{t} = T - t$ durch

$$\dot{\bar{\lambda}}(t) - Q(\bar{t})x(\bar{t}) - S(\bar{t})u(\bar{t}) - A(\bar{t})\bar{\lambda}(t) - \dot{\bar{\pi}}_2(t) = 0$$

mit dem Anfangswert $\bar{\lambda}(0) = Vx(T) + \pi_4$, dann hat das entstehende Gleichungssystem genau dann die Lösung $w = (x, \lambda, u)$, wenn $F(w) - \pi = 0_{\mathcal{Y}}$ gilt. Wenn wir zeigen können, dass das Optimierungsproblem (QLP) zu jedem $a \in \mathbb{R}^n$ und dem entsprechenden $\pi \in \mathcal{Y}$ genau ein lokales Minimum $\bar{w} \in X_h$ hat, dann folgt daraus, dass auch die Gleichung $F(w) - \pi = 0_{\mathcal{Y}}$ für jedes $\pi \in \mathcal{Y}$ die eindeutig bestimmte Lösung $\bar{w} \in X_h$ hat und damit der Operator F invertierbar ist.

Dafür zeigen wir zunächst, dass es zu jedem $\pi \in \mathcal{Y}$ ein $\bar{w} \in \mathcal{Q} \supset X_h$ gibt, welches eindeutig bestimmtes Minimum des Problems (QLP) ist, und zeigen dann, dass dieses $\bar{w}$ auch in X_h liegt. Wir betrachten zu der Zielfunktion aus (QLP) und einem beliebigen fest gewählten $\pi \in \mathcal{Y}$ die zulässige Menge

$$\begin{aligned}U := \{(x,u) \in W^{1,2}([0,T],\mathbb{R}^n) \times L^2([0,T],\mathbb{R}^m) :\\ \dot{x}(t) - A(t)x(t) - B(t)u(t) - \dot{\pi}_1(t) = 0,\ \forall t \in [0,T],\ x(0) = a\},\end{aligned}$$

die nichtleer und konvex ist. Da $\pi \in \mathcal{Y}$ ist, folgt dass $\dot{\pi}_1 \in L^\infty$ und damit auch in L^2 ist. Mit den Eigenschaften der Funktionen $A(t)$ und $B(t)$, welche aus der Voraussetzung *Glattheit* folgen, ist U nichtleer. Wir zeigen zunächst die Konvexität von U. Es sei $(x^1,u^1),(x^2,u^2) \in U$ mit $(x^1,u^1) \neq (x^2,u^2)$, dann gilt für alle $0 < \vartheta < 1$ offensichtlich $\vartheta x^1 + (1-\vartheta)x^2 \in W^{1,2}([0,T],\mathbb{R}^n)$ und $\vartheta u^1 + (1-\vartheta)u^2 \in L^2([0,T],\mathbb{R}^m)$. Aufgrund der Linearität der Differentialgleichung ist diese ebenfalls für $(\vartheta x^1 + (1-\vartheta)x^2,\ \vartheta u^1 + (1-\vartheta)u^2)$ erfüllt, und es gilt natürlich $\vartheta x^1(0) + (1-\vartheta)x^2(0) = a$. Damit ist auch $(\vartheta x^1 + (1-\vartheta)x^2,\ \vartheta u^1 + (1-\vartheta)u^2) \in U$ und somit U konvex. Weiterhin ist für $(x^1,u^1),(x^2,u^2) \in U$ aufgrund der Linearität der Differentialgleichung $(x^1 - x^2, u^1 - u^2) \in \mathcal{C}$ aus der Bedingung *Koerzivität*. Definieren wir damit einen Raum

$$\mathcal{Q} = \{(x,u) : x \in W^{1,2}([0,T],\mathbb{R}^n),\ u \in L^2([0,T],\mathbb{R}^m)\}$$

mit der üblichen Norm (siehe Gleichung (3.0.2)), dann ist der normierte Raum $\mathcal{Q}$ ein reflexiver Banachraum (siehe z.B. [Wer07], Definition III.3.3). Es gilt $U \subset \mathcal{Q}$, im Raum $\mathcal{Q}$ ist U abgeschlossen, nichtleer und konvex.

Wir wollen zeigen, dass die Zielfunktion

$$Z(x,u) := \mathcal{B}(x,u) + \pi_4^\top x(T) - \int_0^T \dot{\pi}_2(t)^\top x(t)dt - \int_0^T \pi_3(t)^\top u(t)dt$$

auf U strikt konvex ist, das heißt für alle $0 < \vartheta < 1$ gilt

$$\vartheta Z(x^1,u^1) + (1-\vartheta)Z(x^2,u^2) - Z(\vartheta x^1 + (1-\vartheta)x^2,\, \vartheta u^1 + (1-\vartheta)u^2) > 0.$$

Dabei kürzen sich in der Differenz natürlich die in x und u linearen Terme heraus, und es genügt zu zeigen, dass

$$\vartheta \mathcal{B}(x^1,u^1) + (1-\vartheta)\mathcal{B}(x^2,u^2) - \mathcal{B}(\vartheta x^1 + (1-\vartheta)x^2,\, \vartheta u^1 + (1-\vartheta)u^2) > 0$$

ist. Dazu betrachten wir exemplarisch den Term $x(T)^\top V x(T)$ und wenden dieselbe Argumentation auf die restlichen Terme von $\mathcal{B}(x,u)$ an. Es gilt

$$\begin{aligned}\vartheta(x^1(T)^\top V x^1(T)) + (1-\vartheta)(x^2(T)^\top V x^2(T)) \\ - (\vartheta x^1(T) + (1-\vartheta)x^2(T))^\top V(\vartheta x^1(T) + (1-\vartheta)x^2(T)) \\ = \vartheta(1-\vartheta)[(x^1(T)-x^2(T))^\top V(x^1(T)-x^2(T))],\end{aligned}$$

und entsprechend erhält man

$$\begin{aligned}\vartheta \mathcal{B}(x^1,u^1) + (1-\vartheta)\mathcal{B}(x^2,u^2) - \mathcal{B}(\vartheta x^1 + (1-\vartheta)x^2,\, \vartheta u^1 + (1-\vartheta)u^2) \\ = \vartheta(1-\vartheta)\mathcal{B}(x^1-x^2,\, u^1-u^2).\end{aligned}$$

Da aus $(x^1,u^1),(x^2,u^2) \in U$ folgt dass $(x^1-x^2,u^1-u^2) \in \mathcal{C}$ ist, gilt die Bedingung *Koerzivität* für $\mathcal{B}(x^1-x^2,\, u^1-u^2)$. Wegen $0 < \vartheta < 1$ gilt außerdem $\vartheta(1-\vartheta) > 0$. Daher ist für $(x^1,u^1),(x^2,u^2) \in U$ mit $(x^1,u^1) \neq (x^2,u^2)$

$$\begin{aligned}\vartheta Z(x^1,u^1) + (1-\vartheta)Z(x^2,u^2) - Z(\vartheta x^1 + (1-\vartheta)x^2,\, \vartheta u^1 + (1-\vartheta)u^2) \\ \geq \vartheta(1-\vartheta)\alpha(\|x^1-x^2\|_{H^1}^2 + \|u^1-u^2\|_{L^2}^2) > 0\,,\end{aligned}$$

und die Zielfunktion ist strikt konvex auf der konvexen zulässigen Menge U.

Aufgrund der Linearität der Differentialgleichung kann man mit Hilfe eines beliebigen fest gewählten Elements $(x^1,u^1) \in U$ jedes Element aus U als Summe (x^1+x, u^1+u) mit einem $(x,u) \in \mathcal{C}$ darstellen. Da die einzigen nichtlinearen Terme in der Zielfunktion quadratische Terme in $\mathcal{B}$ sind, kann man die Zielfunktion $Z(x^1+x,u^1+u)$ in eine Summe aus $Z(x^1,u^1) + Z(x,u)$ und einem Rest zerlegen. Dabei stehen im Rest die gemischten Terme, die sich in $\mathcal{B}$ ergeben und welche bei fest gewählten $(x^1,u^1) \in U$ linear in $(x,u) \in \mathcal{C}$ sind. Benutzt man die Tatsache,

dass für reelle Zahlen a und b stets $a + b \geq a - |b|$ gilt, die Bedingung *Koerzivität* und wendet die Cauchy-Schwarzsche-Ungleichung für

$$\left| \int_0^T \dot{\pi}_2(t)^\top x(t) dt \right| \leq \|\dot{\pi}_2\|_{L^2} \|x\|_{L^2}, \quad \left| \int_0^T \pi_3(t)^\top u(t) dt \right| \leq \|\pi_3\|_{L^2} \|u\|_{L^2}$$

an, erhält man wegen $(x, u) \in \mathcal{C}$

$$\begin{aligned} (3.5.7) \quad Z(x,u) &= \mathcal{B}(x,u) + \pi_4^\top x(T) - \int_0^T \dot{\pi}_2(t)^\top x(t) dt - \int_0^T \pi_3(t)^\top u(t) dt \\ &\geq \alpha(\|x\|_{H^1}^2 + \|u\|_{L^2}^2) - \|\pi_4\|_2 \|x(T)\|_2 - \|\dot{\pi}_2\|_{L^2} \|x\|_{L^2} - \|\pi_3\|_{L^2} \|u\|_{L^2}. \end{aligned}$$

Schätzt man noch die gemischten Terme von $\mathcal{B}(x^1 + x, u^1 + u)$ nach unten mit Hilfe von $b \geq -|b|$ ab und berücksichtigt man dabei $\|\dot{\pi}_2\|_{L^2}$, $\|\pi_3\|_{L^2} < \infty$, $\|\pi_4\|_2 < \infty$ und dass alle Normen, in denen $(x, u) \in \mathcal{C}$ nicht vorkommt, endlich sind, ist offensichtlich dass die Zielfunktion nach unten beschränkt ist. Hierbei gilt

$$\|x(T)\|_2 = \left\| x(0) + \int_0^T \dot{x}(t) dt \right\|_2 \leq \|x(0)\|_2 + \int_0^T \|\dot{x}(t)\|_2 \, dt \leq \|x(0)\|_2 + \|\dot{x}\|_{L^1}$$

und mit einer Konstanten ζ, wie oben beschrieben, $\|\dot{x}\|_{L^1} \leq \zeta \|\dot{x}\|_{L^2} \leq \zeta \|x\|_{H^1}$. Wegen $x^1(0) = a$ und $x(0) = 0$ kann man die Terme $x^1(T)$ und $x(T)$ entsprechend abschätzen.

Aus Gleichung (3.5.7) folgt, dass die Zielfunktion koerzitiv ist (siehe z.B. [Wer07], Definition III.5.7), dass heißt, es gelten für Folgen x^j und u^j die Schlussfolgerungen

$$\begin{aligned} &\|x^j\|_{H^1} + \|u^j\|_{L^2} \to \infty \Rightarrow Z(x^j, u^j) \to \infty \text{ bzw.} \\ &\forall K \; \exists r_0 : \; \|x\|_{H^1} + \|u\|_{L^2} \geq r_0 \Rightarrow Z(x,u) \geq K. \end{aligned}$$

Wählt man $K > \inf_{(x,u) \in U} Z(x,u) > -\infty$ und eine Folge $(x^j, u^j)_{j \in \mathbb{N}} \in U$ mit

$$\lim_{j \to \infty} Z(x^j, u^j) = \inf_{(x,u) \in U} Z(x,u)$$

dann gibt es ein j_0 und r_0 mit $\|x^j\|_{H^1} + \|u^j\|_{L^2} \leq r_0 \; \forall j \geq j_0$. Daraus folgt, dass die Folge $(x^j, u^j)_{j \geq j_0} \in U$ beschränkt ist und es somit eine schwach konvergente Teilfolge $(x^i, u^i)_i \in U$ gibt, die gegen ein $(\bar{x}, \bar{u})$ konvergiert (siehe z.B. [Wer07], Definition III.3.6 und Theorem III.3.7). Da die zulässige Menge U eine abgeschlossene und konvexe Teilmenge eines reflexiven Banachraums ist, gilt $(\bar{x}, \bar{u}) \in U$ (siehe z.B. [Wer07], Satz III.3.8). Die Zielfunktion hängt stetig von x und u ab, und somit folgt $\lim_{i \to \infty} Z(x^i, u^i) = Z(\bar{x}, \bar{u})$. Gleichzeitig konvergiert die Folge $(Z(x^j, u^j))_j$ gegen $\inf_{(x,u) \in U} Z(x,u)$. Damit konvergiert die Folge $(Z(x^i, u^i))_i$ als Teilfolge der Folge $(Z(x^j, u^j))_j$ ebenfalls gegen $\inf_{(x,u) \in U} Z(x,u)$, und somit gilt

$$Z(\bar{x}, \bar{u}) = \inf_{(x,u) \in U} Z(x,u).$$

Bleibt noch zu zeigen, dass der Punkt $(\bar{x}, \bar{u}) \in U$ das einzige lokale und damit das eindeutig bestimmte globale Minimum des Problems (QLP) ist, dann ist auch die zugehörige Adjungierte $\bar{\lambda}$ eindeutig bestimmt. Wegen der Konvexität des Problems (QLP) ist auch jede Extremale (x, λ, u), die die notwendigen Bedingungen 1. Ordnung erfüllt, Optimallösung von (QLP). Dies folgert man aus der Konvexität von U und der strikten Konvexität von Z. Nimmt man an, es gibt einen zweiten Punkt $(\hat{x}, \hat{u}) \in U$, der lokales Minimum der Funktion Z ist, dann müsste es eine Umgebung $U_\varepsilon \subset U$ um diesen Punkt geben mit $Z(x, u) \geq Z(\hat{x}, \hat{u})$, $\forall (x, u) \in U_\varepsilon$. Aufgrund der Konvexität von U sind aber alle Punkte $(\vartheta\bar{x} + (1-\vartheta)\hat{x}, \vartheta\bar{u} + (1-\vartheta)\hat{u})$ für $0 < \vartheta < 1$ in U und in U_ε für hinreichend kleines ϑ. Außerdem ist $Z(\bar{x}, \bar{u}) \leq Z(\hat{x}, \hat{u})$, und damit gilt aufgrund der strikten Konvexität der Funktion Z

$$Z(\vartheta\bar{x} + (1-\vartheta)\hat{x}, \vartheta\bar{u} + (1-\vartheta)\hat{u}) < \vartheta Z(\bar{x}, \bar{u}) + (1-\vartheta) Z(\hat{x}, \hat{u}) \leq Z(\hat{x}, \hat{u}).$$

Dies ist aber ein Widerspruch zu der Annahme, dass $(\hat{x}, \hat{u})$ lokales Minimum ist. Damit gibt es zu jedem $\pi \in \mathcal{Y}$ genau ein $\bar{w} \in \mathcal{X}$, welches die notwendigen Optimalitätsbedingung des Problems (QLP) erfüllt. Daraus folgt, dass die Gleichung $F(w) - \pi = 0_{\mathcal{Y}}$ eine eindeutig bestimmte Lösung $\bar{w} \in \mathcal{Q}$ hat.

Damit die Bedingung (B1) erfüllt ist, müssen wir für die Lösung $\bar{w}$ des Gleichungssystems $F(w) - \pi = 0_{\mathcal{Y}}$ zeigen, dass $\bar{w}$ auch in X_h liegt. Dabei gilt mit $\bar{w} \in U$ die Zustandsgleichung

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + \dot{\pi}_1(t),$$

und man erhält mit $\dot{\pi}_1(t) \in L^\infty$ aus der Ungleichung von Gronwall, dass $x \in L^\infty$ gilt. Ebenso liegt der zugehörige adjungierte Zustand aus den notwendigen Optimalitätsbedingungen zum Problem (QLP) in L^∞. Da die notwendigen Optimalitätsbedingungen des Problems (QLP) für $\bar{w}$ gelten, erhält man folgende Beziehung

$$u(t) = R(t)^{-1} \left[-S(t)^\top x(t) - B(t)^\top \lambda(t) + \pi_3(t) \right]. \tag{3.5.8}$$

Dabei existiert nach der Bedingung *Koerzivität* zu der symmetrischen Matrix $R(t)$ punktweise eine Inverse $R(t)^{-1}$ (siehe [DH98]). Außerdem gilt für alle $\nu \in \mathbb{R}^m$ nach *Koerzivität* $\nu^\top R(t)\nu \geq \alpha \nu^\top \nu$. Setzt man $\nu = R(t)^{-1}\tilde{\nu}$ ein, erhält man

$$\|R(t)^{-1}\tilde{\nu}\|_2 \|\tilde{\nu}\|_2 \geq \tilde{\nu}^\top R(t)^{-1}\tilde{\nu} \geq \alpha \|R(t)^{-1}\tilde{\nu}\|_2^2$$

bzw. durch Umstellen

$$\|R(t)^{-1}\tilde{\nu}\|_2 \leq \frac{1}{\alpha}\|\tilde{\nu}\|_2$$

und damit, dass $R(t)^{-1}$ beschränkt ist, da $\tilde{\nu} \in \mathbb{R}^m$ beliebig ist. Außerdem liefert die Gleichung $R(t_1)R(t_1)^{-1} - R(t_2)R(t_2)^{-1} = 0$ nach Umformung

$$R(t_1)\left[R(t_1)^{-1} - R(t_2)^{-1}\right] - \left[R(t_2) - R(t_1)\right] R(t_2)^{-1} = 0$$

und schließlich die Gleichung

$$R(t_1)^{-1} - R(t_2)^{-1} = R(t_1)^{-1}\left[R(t_2) - R(t_1)\right] R(t_2)^{-1}.$$

Da $R(t)^{-1}$ beschränkt ist, folgt damit die Lipschitzstetigkeit von $R(t)^{-1}$ aus der Lipschitzstetigkeit von $R(t)$. Damit sind die matrixwertigen Funktionen A, B, Q, S und R^{-1} alle lipschitzstetig bezüglich t. Aus der Gleichung (3.5.8) folgt mit Hilfe von $\bar{x} \in L^\infty$, $\bar{\lambda} \in L^\infty$ und $\pi_3 \in L^\infty$ sofort, dass $\bar{u} \in L^\infty$ ist. Damit kann man aus der Gleichung für den Zustand und die adjungierten Zustände folgern, dass $x, \lambda \in W^{1,\infty}$ sind. Aus der Gleichung (3.5.8) erhält man, dass u ebenfalls in $W^{1,\infty}$ ist und schließlich $\bar{w} \in X_h$.

3.5.2 Lipschitzstetigkeit von F^{-1}

Damit ist gezeigt, dass der Operator F invertierbar ist, und wir müssen noch die, in Bedingung (B1) geforderte, Lipschitzstetigkeit zeigen. Dies ist dann natürlich gleichbedeutend damit, dass der lineare Operator F^{-1} stetig bzw. beschränkt ist. Dazu definieren wir einen Operator Λ, der jeder Funktion $\chi \in L^\infty([0,T], \mathbb{R}^n)$ die Lösung des Differentialgleichungssystems

$$\dot{x}(t) = A(t)x(t) + \chi(t),\ t \in [0,T],\ x(0) = 0$$

zuordnet. Weiterhin sei φ die Lösung des Differentialgleichungssystems

$$\dot{x}(t) = A(t)x(t),\ t \in [0,T],\ x(0) = a\,,$$

und somit ist aufgrund der Linearität $\Lambda(\chi) + \varphi$ eine Lösung von Gleichung (3.5.1). Zu gegebenen u ist dann $\Lambda(Bu) + \varphi$ die Lösung der Zustandsgleichungen, und zu jedem u mit $w = (x, \lambda, u) \in X_h$ ist das Paar aus u und $\Lambda(Bu + \dot{\pi}_1) + \varphi$ in U. Der so definierte Operator Λ ist linear und nach den Ausführungen zum Abschätzen der Lösungen einer linearen Differentialgleichung mit Hilfe der Gronwall-Ungleichung beschränkt.

Betrachten wir die Zielfunktion des Problems (QLP) und sehen $x = \Lambda(Bu + \dot{\pi}_1) + \varphi$ im Folgenden immer als Funktion von u an. Damit ist auch die Zielfunktion

$$Z(u) := Z(\Lambda(Bu + \dot{\pi}_1) + \varphi, u)$$

ein Funktional von u. Wir haben gezeigt, dass das Problem (QLP) eine eindeutig bestimmte Lösung $\bar{w} \in X_h$ besitzt. Aufgrund der Konvexität von U muss daher $Z(\bar{u}) < Z(\bar{u} + \vartheta(u - \bar{u}))$ für alle $u \neq \bar{u}$ und $0 < \vartheta \leq 1$ gelten. Damit muss ebenfalls

$$\frac{Z(\bar{u} + \vartheta(u - \bar{u})) - Z(\bar{u})}{\vartheta} > 0 \text{ und } \lim_{\vartheta \to 0} \frac{Z(\bar{u} + \vartheta(u - \bar{u})) - Z(\bar{u})}{\vartheta} \geq 0$$

gelten, falls dieser Grenzwert existiert. Diesen Grenzwert kann man allerdings direkt aus den entsprechenden Definitionen berechnen. Berücksichtigt man die Symmetrie der Matrizen V, Q, R und benutzt wie üblich das Skalarprodukt $\langle \cdot, \cdot \rangle$ im L^2 als Abkürzung für

$$\langle \bar{x}, Q(x - \bar{x}) \rangle_{L^2} := \int_0^T \bar{x}(t)^\top Q(t)(x(t) - \bar{x}(t))\, dt\,,$$

erhält man die Variationsungleichung

$$\begin{aligned}(\text{VI}) \qquad 0 \le{}& \bar{x}(T)^\top V(x(T)-\bar{x}(T)) + \langle \bar{x}, Q(x-\bar{x})\rangle_{L^2} + \langle \bar{u}, R(u-\bar{u})\rangle_{L^2} \\ &+ \langle \bar{x}, S(u-\bar{u})\rangle_{L^2} + \langle x-\bar{x}, S\bar{u}\rangle_{L^2} \\ &+ \pi_4^\top(x(T)-\bar{x}(T)) - \langle \dot{\pi}_2, x-\bar{x}\rangle_{L^2} - \langle \pi_3, u-\bar{u}\rangle_{L^2},\end{aligned}$$

wobei weiterhin $x = \Lambda(Bu+\dot{\pi}_1)+\varphi$ gilt und man formal natürlich zu den matrixwertigen Funktionen $Q(\cdot)$, $R(\cdot)$, $S(\cdot)$ lineare Operatoren Q, R, S definieren kann, die den entsprechenden vektorwertigen Funktionen wieder vektorwertige Funktionen zuordnen, so dass zum Beispiel $(S\bar{u})(t) = S(t)\bar{u}(t)$ gilt und man damit das übliche Skalarprodukt im L^2 verwendet. Jetzt setzen wir zunächst $x = \Lambda(Bu+\dot{\pi}_1)+\varphi = \Lambda Bu+\Lambda\dot{\pi}_1+\varphi$ und den entsprechenden Ausdruck für $\bar{u}$ in die Variationsungleichung ein und erhalten

$$\begin{aligned}(\Lambda B\bar{u}+\Lambda\dot{\pi}_1+\varphi)(T)^\top V((\Lambda B(u-\bar{u}))(T)) + \pi_4^\top((\Lambda B(u-\bar{u}))(T)) \\ +\langle \Lambda B\bar{u}+\Lambda\dot{\pi}_1+\varphi, Q\Lambda B(u-\bar{u})\rangle_{L^2} + \langle \bar{u}, R(u-\bar{u})\rangle_{L^2} \\ +\langle \Lambda B\bar{u}+\Lambda\dot{\pi}_1+\varphi, S(u-\bar{u})\rangle_{L^2} + \langle \Lambda B(u-\bar{u}), S\bar{u}\rangle_{L^2} \\ -\langle \dot{\pi}_2, \Lambda B(u-\bar{u})\rangle_{L^2} - \langle \pi_3, u-\bar{u}\rangle_{L^2} \ge 0.\end{aligned}$$

Die Variationsungleichung (VI) liefert die gewünschte lipschitzstetige Abhängigkeit der eindeutig bestimmten Lösung $\bar{w} \in X_h$ bzw. der entsprechenden Steuerungen $\bar{u}$ des Problems (QLP) von $\pi \in \mathcal{Y}$. Betrachten wir dazu die eindeutig bestimmte Lösung $\bar{u}^i$ zu gegebenen $\pi^i = (\pi_1^i, \pi_2^i, \pi_3^i, \pi_4^i)$ für $i = 1, 2$ und die Variationsungleichung sowohl für $\bar{u} = \bar{u}^1$, $u = \bar{u}^2$, $\pi = \pi^1$ als auch für $\bar{u} = \bar{u}^2$, $u = \bar{u}^1$, $\pi = \pi^2$. Dann erhält man als Summe beider Variationsungleichungen

$$\begin{aligned}(\Lambda B(\bar{u}^2-\bar{u}^1)+\Lambda(\dot{\pi}_1^2-\dot{\pi}_1^1))(T)^\top V((\Lambda B(\bar{u}^2-\bar{u}^1))(T)) \\ +(\pi_4^2-\pi_4^1)^\top((\Lambda B(\bar{u}^2-\bar{u}^1))(T)) \\ +\langle \Lambda B(\bar{u}^2-\bar{u}^1)+\Lambda(\dot{\pi}_1^2-\dot{\pi}_1^1), Q\Lambda B(\bar{u}^2-\bar{u}^1)\rangle_{L^2} + \langle \bar{u}^2-\bar{u}^1, R(\bar{u}^2-\bar{u}^1)\rangle_{L^2} \\ +\langle \Lambda(\dot{\pi}_1^2-\dot{\pi}_1^1), S(\bar{u}^2-\bar{u}^1)\rangle_{L^2} + 2\langle \Lambda B(\bar{u}^2-\bar{u}^1), S(\bar{u}^2-\bar{u}^1\rangle_{L^2} \\ -\langle \dot{\pi}_2^2-\dot{\pi}_2^1, \Lambda B(\bar{u}^2-\bar{u}^1)\rangle_{L^2} - \langle \pi_3^2-\pi_3^1, \bar{u}^2-\bar{u}^1\rangle_{L^2} \le 0,\end{aligned}$$

wenn man dabei

$$\begin{aligned}&\langle \bar{u}^1, R(\bar{u}^2-\bar{u}^1)\rangle_{L^2} + \langle \bar{u}^2, R(\bar{u}^1-\bar{u}^2)\rangle_{L^2} \\ &\quad = \langle \bar{u}^1, R(\bar{u}^2-\bar{u}^1)\rangle_{L^2} - \langle \bar{u}^2, R(\bar{u}^2-\bar{u}^1)\rangle_{L^2} = \langle \bar{u}^1-\bar{u}^2, R(\bar{u}^2-\bar{u}^1)\rangle_{L^2} \\ &\quad = -\langle \bar{u}^2-\bar{u}^1, R(\bar{u}^2-\bar{u}^1)\rangle_{L^2}\end{aligned}$$

berücksichtigt.

Betrachtet man zu gegebenen $\bar{u}^i$ die Funktion $\Lambda B\bar{u}^i+\varphi$, dann erfüllt diese die ungestörten Zustandsgleichungen, es gilt $(\Lambda B(\bar{u}^2-\bar{u}^1), \bar{u}^2-\bar{u}^1) \in \mathcal{C}$ und damit $\mathcal{B}(\Lambda B(\bar{u}^2-\bar{u}^1), \bar{u}^2-\bar{u}^1) \ge \alpha\|\bar{u}^2-\bar{u}^1\|_{L^2}^2$. Deshalb schreiben wir die Ungleichung in

$$\begin{aligned}&2\mathcal{B}(\Lambda B(\bar{u}^2-\bar{u}^1), \bar{u}^2-\bar{u}^1) \le \\ &\quad -(\Lambda(\dot{\pi}_1^2-\dot{\pi}_1^1))(T)^\top V((\Lambda B(\bar{u}^2-\bar{u}^1))(T)) - (\pi_4^2-\pi_4^1)^\top((\Lambda B(\bar{u}^2-\bar{u}^1))(T)) \\ &\qquad -\langle \Lambda(\dot{\pi}_1^2-\dot{\pi}_1^1), Q\Lambda B(\bar{u}^2-\bar{u}^1)\rangle_{L^2} - \langle \Lambda(\dot{\pi}_1^2-\dot{\pi}_1^1), S(\bar{u}^2-\bar{u}^1)\rangle_{L^2} \\ &\qquad\qquad +\langle \dot{\pi}_2^2-\dot{\pi}_2^1, \Lambda B(\bar{u}^2-\bar{u}^1)\rangle_{L^2} + \langle \pi_3^2-\pi_3^1, \bar{u}^2-\bar{u}^1\rangle_{L^2}\end{aligned}$$

um. Jetzt können wir direkt die rechte Seite mit Hilfe der Hölder- bzw. Cauchy-Schwarzschen-Ungleichung nach oben abschätzen. Es folgt mit der Definition von Λ und Ungleichung (3.5.3) mit geeigneter Konstante ζ

$$\begin{aligned}
&2\mathcal{B}(\Lambda B(\bar{u}^2-\bar{u}^1),\bar{u}^2-\bar{u}^1) \leq \\
&\quad \zeta\|\dot{\pi}_1^2-\dot{\pi}_1^1\|_{L^1}\|V\|_{\mathcal{M}}\zeta\|B\|_{\mathcal{M}}\|\bar{u}^2-\bar{u}^1\|_{L^1}+\|\pi_4^2-\pi_4^1\|_2\zeta\|B\|_{\mathcal{M}}\|\bar{u}^2-\bar{u}^1\|_{L^1} \\
&+\zeta\|\dot{\pi}_1^2-\dot{\pi}_1^1\|_{L^1}\|Q\|_{\mathcal{M}}\zeta\|B\|_{\mathcal{M}}\|\bar{u}^2-\bar{u}^1\|_{L^1}+\zeta\|\dot{\pi}_1^2-\dot{\pi}_1^1\|_{L^1}\|S\|_{\mathcal{M}}\|\bar{u}^2-\bar{u}^1\|_{L^2} \\
&\quad +\|\dot{\pi}_2^2-\dot{\pi}_2^1\|_{L^2}\zeta\|B\|_{\mathcal{M}}\|\bar{u}^2-\bar{u}^1\|_{L^1}+\|\pi_3^2-\pi_3^1\|_{L^2}\|\bar{u}^2-\bar{u}^1\|_{L^2},
\end{aligned}$$

Da wir nur Funktionen auf dem Intervall $[0,T]$ betrachten, kann man wiederum mit einer anderen Konstanten $\|\bar{u}^2-\bar{u}^1\|_{L^1}\leq\zeta\|\bar{u}^2-\bar{u}^1\|_{L^2}$ abschätzen. Danach schätzt man die linke Seite mit Hilfe der Bedingung *Koerzivität* nach unten ab und teilt die gewonnene Ungleichung durch $\|\bar{u}^2-\bar{u}^1\|_{L^2}$. Man erhält insgesamt

$$\begin{aligned}
&2\alpha\|\bar{u}^2-\bar{u}^1\|_{L^2}\leq \\
&\qquad \zeta_{3.5.9}\Big[\|\dot{\pi}_1^2-\dot{\pi}_1^1\|_{L^1}+\|\dot{\pi}_2^2-\dot{\pi}_2^1\|_{L^2}+\|\pi_3^2-\pi_3^1\|_{L^2}+\|\pi_4^2-\pi_4^1\|_2\Big]
\end{aligned}\tag{3.5.9}$$

mit einer Konstanten $\zeta_{3.5.9}$. Da weiterhin mit $\pi\in\mathcal{Y}$ folgt, dass $\dot{\pi}_1\in L^\infty$, $\dot{\pi}_2\in L^\infty$ und $\pi_3\in L^\infty$ sind, kann man wiederum die L^2-Norm durch die L^∞-Norm und diese durch die $W^{1,\infty}$-Norm abschätzen. Man erhält damit, mit einer Konstanten $\zeta_{3.5.10}$, die Ungleichung

$$\begin{aligned}
&\|\bar{u}^2-\bar{u}^1\|_{L^2}\leq \\
&\qquad \zeta_{3.5.10}\Big[\|\pi_1^2-\pi_1^1\|_{W^{1,\infty}}+\|\pi_2^2-\pi_2^1\|_{W^{1,\infty}}+\|\pi_3^2-\pi_3^1\|_{W^{1,\infty}}+\|\pi_4^2-\pi_4^1\|_2\Big].
\end{aligned}\tag{3.5.10}$$

Aus den Zustandsgleichungen und den Gleichungen für die adjungierten Zustände erhält man mit Gleichung (3.5.3) und entsprechenden Konstanten Abschätzungen bezüglich der Supremumsnorm für $\bar{x}^2-\bar{x}^1$ und $\bar{\lambda}^2-\bar{\lambda}^1$

$$\begin{aligned}
\|\bar{x}^2-\bar{x}^1\|_{L^\infty} &\leq \zeta\|\bar{u}^2-\bar{u}^1\|_{L^1}\leq\tilde{\zeta}\|\bar{u}^2-\bar{u}^1\|_{L^2} \\
&\leq\hat{\zeta}\|\pi^2-\pi^1\|_{\mathcal{Y}} \\
\|\bar{\lambda}^2-\bar{\lambda}^1\|_{L^\infty} &\leq \zeta\left[\|\bar{u}^2-\bar{u}^1\|_{L^1}+\|\bar{x}^2-\bar{x}^1\|_{L^1}\right]\leq\tilde{\tilde{\zeta}}\left[\|\bar{u}^2-\bar{u}^1\|_{L^2}+\|\bar{x}^2-\bar{x}^1\|_{L^\infty}\right] \\
&\leq\hat{\tilde{\zeta}}\|\pi^2-\pi^1\|_{\mathcal{Y}}.
\end{aligned}$$

Da das Minimumprinzip

$$R(t)^\top\bar{u}^i(t)+S(t)^\top\bar{x}^i(t)+B(t)^\top\bar{\lambda}^i(t)-\pi_3^i(t)=0$$

für das Problem (QLP) punktweise gilt, kann man daraus Abschätzungen bezüglich der Supremumsnorm für $\bar{u}^2-\bar{u}^1$ gewinnen. Stellt man diese Gleichung nach $\bar{u}^i$ um und berücksichtigt, dass $R(t)^{-1}$ beschränkt ist, dann folgt aus der Beschränktheit und Linearität der auftretenden Matrizen

$$\begin{aligned}
\bar{u}^2(t)-\bar{u}^1(t) &\leq \zeta\left[\|\bar{x}^2(t)-\bar{x}^1(t)\|_2+\|\bar{\lambda}^2(t)-\bar{\lambda}^1(t)\|_2+\|\pi_3^2(t)-\pi_3^1(t)\|_2\right] \\
&\leq\tilde{\zeta}\left[\|\bar{x}^2-\bar{x}^1\|_{L^\infty}+\|\bar{\lambda}^2-\bar{\lambda}^1\|_{L^\infty}+\|\pi_3^2-\pi_3^1\|_{L^\infty}\right]
\end{aligned}$$

mit neuen Konstanten ζ und $\tilde{\zeta}$. Daraus folgt mit einer Konstanten $\zeta_{3.5.11}$

$$\|\bar{u}^2 - \bar{u}^1\|_{L^\infty} \leq \zeta_{3.5.11} \|\pi^2 - \pi^1\|_{\mathcal{Y}}, \tag{3.5.11}$$

Abschließend kann man, wiederum mit neuen Konstanten, aus der Zustandsgleichung und der Differentialgleichung für die adjungierten Zustände des Problems (QLP) $\dot{\bar{x}}^2 - \dot{\bar{x}}^1$ und $\dot{\bar{\lambda}}^2 - \dot{\bar{\lambda}}^1$ nach oben abschätzen, es gilt

$$\begin{aligned}
\|\dot{\bar{x}}^2 - \dot{\bar{x}}^1\|_{L^\infty} &\leq \zeta \left[\|\bar{x}^2 - \bar{x}^1\|_{L^\infty} + \|\bar{u}^2 - \bar{u}^1\|_{L^\infty}\right] \\
&\leq \tilde{\zeta} \|\pi^2 - \pi^1\|_{\mathcal{Y}}, \\
\|\dot{\bar{\lambda}}^2 - \dot{\bar{\lambda}}^1\|_{L^\infty} &\leq \hat{\zeta} \left[\|\bar{x}^2 - \bar{x}^1\|_{L^\infty} + \|\bar{u}^2 - \bar{u}^1\|_{L^\infty} + \|\bar{\lambda}^2 - \bar{\lambda}^1\|_{L^\infty}\right] \\
&\leq \tilde{\tilde{\zeta}} \|\pi^2 - \pi^1\|_{\mathcal{Y}}.
\end{aligned}$$

Aus dem Minimumprinzip erhält man für $\|\dot{\bar{u}}^2 - \dot{\bar{u}}^1\|_{L^\infty}$ eine entsprechende Abschätzung, zunächst gilt punktweise

$$\begin{aligned}
\bar{u}^2(t) - \bar{u}^1(t) = R(t)^{-1^\top} \Big[& S(t)^\top \left[\bar{x}^2(t) - \bar{x}^1(t)\right] \\
& + B(t)^\top \left[\bar{\lambda}^2(t) - \bar{\lambda}^1(t)\right] - \left[\pi_3^2(t) - \pi_3^1(t)\right] \Big].
\end{aligned} \tag{3.5.12}$$

Am Ende von Abschnitt 3.5.1 haben wir gezeigt, dass aus der *Koerzivität* aus Abschnitt 3.1 folgt, dass $R(t)^{-1}$ beschränkt und lipschitzstetig ist. Damit ist die Ableitung von $R(t)^{-1}$ wiederum beschränkt. Nach *Glattheit* aus Abschnitt 3.1 sind $S(t)$ und $B(t)$ lipschitzstetig, und somit sind die entsprechenden Ableitungen nach t beschränkt. Dabei hängen die Schranken jeweils nur von dem Problem (OS), das heißt von der Funktion f und dem gegebenen Optimum w^* ab. Somit kann man die Gleichung (3.5.12) nach t ableiten und die rechte Seite nach oben abschätzen. Man erhält mit den bisher gezeigten Ungleichungen eine Abschätzung für $\|\dot{\bar{u}}^2 - \dot{\bar{u}}^1\|_{L^\infty}$ und hat insgesamt gezeigt, dass eine Konstante ζ_M existiert mit der

$$\|\bar{w}^2 - \bar{w}^1\|_{\mathcal{X}} \leq \zeta_M \|\pi^2 - \pi^1\|_{\mathcal{Y}}$$

gilt. Somit hängt die eindeutig bestimmte Lösung von (QLP) lipschitzstetig von dem Parameter $\pi \in \mathcal{Y}$ ab.

3.6 Verfahren der Ordnung $(p, p-1)$

Mit dem Ende des vorigen Abschnitts ist der Satz 3.3.3 gezeigt. Im Beweis benötigt man für die Interpolation zwischen den Gitterpunkten die gleiche Ordnung wie für einen kompletten Schritt von einem Gitterpunkt zum nächsten. Löst man eine Differentialgleichung mit einem SRKV reicht allerdings bei der Interpolation zwischen den Gitterpunkten, wie man im Satz 2.1.5 sieht, die Ordnung $p-1$ aus, um insgesamt die Konvergenzordnung p zu erhalten. Dies überträgt sich bei den praktischen Beispielen im Kapitel 5 auch auf die Verwendung bei Problemen der optimalen Steuerung. Beim Beweis von Satz 2.1.5 wird ausgenutzt, dass man

für einen einzelnen Schritt von einem Gitterpunkt zum nächsten Gitterpunkt im Prinzip eine um eins höhere Konvergenzordnung hat. Summiert man die Fehler der einzelnen Schritte, verliert man diese Ordnung prinzipiell wieder. Allerdings kann man sich für den Fehler zwischen den Gitterpunkten immer wieder auf den vorherigen Gitterpunkt stützen. Man verliert dadurch zwischen den Gitterpunkten keine Ordnung und erhält somit ein stetiges Verfahren der Ordnung p.

Wir nutzen in unserem Beweis diese um eins höhere Konvergenzordnung für einen einzelnen Schritt aus, um die niedrigere Konvergenzordnung der Ableitungen zu kompensieren. Dabei verlieren wir keine weitere Ordnung durch eine Summation und erhalten für ein stetiges Verfahren der Konvergenzordnung (p, p) insgesamt die Ordnung p. Allerdings verlieren wir bei Verfahren der Ordnung $(p, p-1)$, die für einen einzelnen Schritt die Ordnung $(p+1, p)$ haben, genauso eine Ordnung für die Interpolation zwischen den Gitterpunkten bei der Berechnung der Ableitung und erhalten auch nur eine Konvergenzordnung von $p-1$. Wir können hier nicht den Unterschied zwischen den Ordnungen ausgleichen in dem man sich bei der Summation der Fehler auf die um eins höhere Ordnung auf den Gitterpunkten stützt, da im Beweis, der auf Lemma 3.3.1 beruht, die Ordnung nicht durch eine Summation, sondern durch die Verwendung der entsprechenden $W^{1,\infty}$-Normen verringert wird.

Man erhält zwar durch den Beweis von Satz 3.3.3 nicht nur die Konvergenz von $\tilde{w}$ gegen w^*, sondern auch eine entsprechende Konvergenz für die Ableitungen $\|\dot{\tilde{x}} - \dot{x}^*\|_{L^\infty} \leq \zeta_{3.3.3}\, h^{p-1}$, $\|\dot{\tilde{\lambda}} - \dot{\lambda}^*\|_{L^\infty} \leq \zeta_{3.3.3}\, h^{p-1}$ und $\|\dot{\tilde{u}} - \dot{u}^*\|_{L^\infty} \leq \zeta_{3.3.3}\, h^{p-1}$. Es ist aber ein technisches Problem des Beweises, dass man sich die eventuell höhere Ordnung des SRKVs auf den Gitterpunkten nicht ausnutzen kann. Könnte man zeigen, dass man auf den Gitterpunkten die Ordnung p erhält, ist sofort die Ordnung p auch für Punkte zwischen dem Gitter gezeigt, da

$$\begin{aligned}
\|\tilde{x}(t) - x^*(t)\|_2 &= \|\tilde{x}(t_j) - x^*(t_j) + \int_{t_j}^{t} \dot{\tilde{x}}(\tau) - \dot{x}^*(\tau)\, d\tau\|_2 \\
&\leq \|\tilde{x}(t_j) - x^*(t_j)\|_2 + \int_{t_j}^{t} \|\dot{\tilde{x}}(\tau) - \dot{x}^*(\tau)\|_2\, d\tau \\
&\leq \|\tilde{x}(t_j) - x^*(t_j)\|_2 + \zeta_{3.3.3}\, h^{p-1} \int_{t_j}^{t} d\tau \\
&\leq \|\tilde{x}(t_j) - x^*(t_j)\|_2 + \zeta_{3.3.3}\, h^{p-1} h
\end{aligned}$$

für alle $t \in (t_j, t_{j+1})$ gilt. Analog schließt man natürlich sofort von einer Konvergenz auf dem Gitter der Ordnung p auf die Konvergenz der stetigen Funktionen $\tilde{\lambda}$ und $\tilde{u}$ mit Ordnung p.

Der kritische Punkt ist zu zeigen, dass die Ordnung p auf dem Gitter erhalten bleibt. Dies kann man in den Beweis zu Satz 3.3.3 nicht einfach integrieren. Um eine Chance zu haben, diese höhere Ordnung auf dem Gitter und den Diskretisierungspunkten der Steuerung zu zeigen, müsste man den Operator F ebenfalls dahingehend verändern, dass er die notwendigen Optimalitätsbedingungen eines diskreten

linear-quadratischen Steuerungsproblems widerspiegelt, welches eine eindeutig bestimmte Lösung hat. Dieser Operator F muss dann immer noch hinreichend nah an einem abgeänderten Operator P_h sein (im Sinn von Bedingung (B2) von Lemma 3.3.1). Allerdings kann man die diskretisierten Optimalitätsbedingungen (DFC) nicht einfach einem diskreten Steuerungsproblem zuordnen. Daher ist es ebenfalls nicht offensichtlich, wie man ein diskretes linear-quadratisches Steuerungsproblem wählen kann, so dass der Operator F die notwendigen Optimalitätsbedingungen enthält und man die Bedingungen aus dem Lemma 3.3.1 erfüllen kann.

3.7 C^1-Runge-Kutta-Verfahren

Von praktischem Interesse ist natürlich, inwiefern die stetigen Näherungslösungen auch an den Gitterpunkten stetig differenzierbar sind, da man in dem Beweis zu Satz 3.3.3 fordert, dass die optimale Steuerung des Problems (OS) hinreichend glatt ist. Berechnet man durch

$$\begin{aligned} x(t_j+\theta h) &= x(t_j) + h\sum_{k=1}^{s} b_k(\theta) f(y_{j,k}, u(t_j+c_k h)) \\ &\qquad \forall \theta \in (0,1],\, \forall j = 0,\dots,N-1 \text{ mit } x(t_0) = x(0) = a\,, \\ y_{j,k} &= x(t_j) + h\sum_{l=1}^{s} a_{k,l} f(y_{j,l}, u(t_{j,l})) \end{aligned}$$

eine stetige Approximation der Lösung des Anfangswertproblems

$$\dot{x}(t) = f(x(t), u(t)),\ x(0) = a$$

zu einer gegebenen lipschitzstetigen Steuerung $u(t)$, dann interessiert man sich dafür, ob

$$\lim_{\vartheta \searrow 0} \frac{x(t_j) - x(t_j - \vartheta)}{t_j - (t_j - \vartheta)} = \lim_{\vartheta \searrow 0} \frac{x(t_j + \vartheta) - x(t_j)}{(t_j + \vartheta) - t_j}$$

gilt. Setzt man

$$\begin{aligned} x(t_j) &= x(t_{j-1}) + h\sum_{k=1}^{s} b_k(1) f(y_{j-1,k}, u(t_{j-1}+c_k h)), \\ x(t_j - \vartheta) &= x(t_{j-1}) + h\sum_{k=1}^{s} b_k\left(\frac{h-\vartheta}{h}\right) f(y_{j-1,k}, u(t_{j-1}+c_k h)), \\ x(t_j + \vartheta) &= x(t_j) + h\sum_{k=1}^{s} b_k\left(\frac{\vartheta}{h}\right) f(y_{j,k}, u(t_j+c_k h)), \end{aligned}$$

ein, stellt sich die Frage, ob

$$\begin{aligned} &\lim_{\vartheta \searrow 0} \frac{h}{\vartheta} \sum_{k=1}^{s} \left[b_k(1) - b_k\left(\frac{h-\vartheta}{h}\right)\right] f(y_{j-1,k}, u(t_{j-1}+c_k h)) \\ =&\lim_{\vartheta \searrow 0} \frac{h}{\vartheta} \sum_{k=1}^{s} b_k\left(\frac{\vartheta}{h}\right) f(y_{j,k}, u(t_j+c_k h)). \end{aligned}$$

Weiterhin gehen wir davon aus, dass sich $b_k(\theta)$ als Polynom in der Form

$$b_k(\theta) = \sum_{l=1}^{q} b_{k,l}\theta^l \tag{3.7.1}$$

darstellen lässt. Damit gilt bei der Berechnung des zweiten Grenzwert

$$\frac{h}{\vartheta} b_k\left(\frac{\vartheta}{h}\right) = \sum_{l=1}^{q} b_{k,l}\left(\frac{\vartheta}{h}\right)^{l-1}, \quad \text{und mit} \quad \lim_{\vartheta \searrow 0} \frac{h}{\vartheta}\ b_k\left(\frac{\vartheta}{h}\right) = b_{k,1}$$

folgt schließlich

$$\lim_{\vartheta \searrow 0} \frac{h}{\vartheta} \sum_{k=1}^{s} b_k\left(\frac{\vartheta}{h}\right) f(y_{j,k}, u(t_j + c_k h)) = \sum_{k=1}^{s} b_{k,1} f(y_{j,k}, u(t_j + c_k h)).$$

Für den ersten Grenzwert gilt

$$\lim_{\vartheta \searrow 0} \frac{h}{\vartheta}\left[b_k(1) - b_k\left(\frac{h-\vartheta}{h}\right)\right] = \lim_{\vartheta \searrow 0} \frac{\left[b_k(1) - b_k\left(1 - \frac{\vartheta}{h}\right)\right]}{\frac{\vartheta}{h}} = b_k'(1),$$

und mit Gleichung (3.7.1) folgt

$$b_k'(1) = \sum_{l=1}^{q} l b_{k,l}.$$

Für das SRKV muss daher insgesamt

$$\sum_{k=1}^{s}\left(\sum_{l=1}^{q} l b_{k,l}\right) f(y_{j-1,k}, u(t_{j-1} + c_k h)) = \sum_{k=1}^{s} b_{k,1} f(y_{j,k}, u(t_j + c_k h)) \tag{3.7.2}$$

gelten, damit es eine stetig differenzierbare Näherungslösung liefert. Man kann die Parameter des SRKVs daher so wählen, dass beide Seiten $f(x(t_j), u(t_j))$ entsprechen, um die Bedingung (3.7.2) für möglichst allgemeine Funktionen f und u zu erfüllen. Eine offensichtliche Möglichkeit, dies für die rechte Seite von (3.7.2) zu erreichen, ist, ein $\tilde{k} \in \{1, \ldots, s\}$ mit $b_{\tilde{k},1} = 1$ und $a_{\tilde{k},l} = 0$, $l = 1, \ldots, s$ zu wählen und dementsprechend alle anderen $b_{k,1} = 0$, $k \in \{1, \ldots, s\} \setminus \tilde{k}$ zu setzen. Daraus ergibt sich für die linke Seite von (3.7.2) die Bedingung, dass es ein $\bar{k} \in \{1, \ldots, s\}$ mit $\bar{k} \neq \tilde{k}$ geben muss mit $\sum_{l=1}^{q} l b_{\bar{k},l} = 1$, $c_{\bar{k}} = \sum_{l=1}^{s} a_{\bar{k},l} = 1$ und $a_{\bar{k},l} = b_l(1)$. Für die restlichen $k \in \{1, \ldots, s\} \setminus \bar{k}$ muss gelten $\sum_{l=1}^{q} l b_{k,l} = 0$. Sind diese Bedingungen erfüllt gilt

$$\begin{aligned} y_{j-1,\bar{k}} &= x(t_{j-1}) + h \sum_{l=1}^{s} a_{\bar{k},l} f(y_{j-1,l}, u(t_{j-1,l})) \\ &= x(t_{j-1}) + h \sum_{l=1}^{s} b_l(1) f(y_{j-1,l}, u(t_{j-1,l})) \ = x(t_j) \end{aligned}$$

und $u(t_{j-1,k}) = u(t_{j-1} + c_k h) = u(t_{j-1} + h) = u(t_j)$. Damit ergibt die linke Seite von (3.7.2) ebenfalls $f(x(t_j), u(t_j))$, und die beiden Seiten sind gleich. Diese zusätzlichen Bedingung werden meist so erfüllt, dass $\tilde{k} = 1$ und $\bar{k} = s$ sind. Dabei ergibt sich, dass bei der letzten Stufe des Schritts $j - 1$ dieselbe Funktionsauswertung wie bei der ersten Stufe des Schritts j zu berechnen ist, dies wird FSAL („first same as last“) oder „stage reuse“ genannt.

Viele verbreitete SRKV liefern C^1-Approximationen der Lösung. In den Verfahren nach Sarafyan, die im Kapitel 5 benutzt werden, sind zum Beispiel exakt die obigen Bedingungen erfüllt. Für weitere Details und Aussagen zu Verfahren mit höherer Glattheit sei auf die Arbeiten [Hig91, Ver93, PT97] verwiesen.

4 Iterationsverfahren zur Bestimmung einer Nullstelle

In diesem Kapitel stellen wir Verfahren vor, um die Nullstelle des Operators P_h zu bestimmen. Dabei werden wir zunächst im Abschnitt 4.1 die wesentlichen Schritte für den Beweis zeigen, dass ein auf den Banachräumen $\mathcal{X}$ und $\mathcal{Y}$ operierendes Newton-Verfahren konvergiert. Diese Konvergenz sichert die Konvergenz eines Newton-Verfahrens, welches nur auf den Diskretisierungspunkten $u(t_{j,k})$ und $u(\bar{t}_{j,k})$ der Steuerung beruht. Dies folgt aus der Tatsache, dass diese Diskretisierungspunkte eindeutig den Zustand x und damit ebenfalls den adjungierten Zustand λ bestimmen. Die Optimalsteuerung an nicht Diskretisierungspunkten kann man dann aus dem Minimumprinzip berechnen. Wir werden die Konvergenz des Newton-Verfahrens auf den endlichdimensionalen, diskretisierten Steuerungen nicht explizit zeigen, da dies nicht in den Rahmen dieser Arbeit passt. Aus praktischer Sicht verweisen wir insbesondere auf die numerischen Resultate in Kapitel 5.

Weiterhin schlagen wir eine Kombination aus einem Abstiegsverfahren und dem Newton-Verfahren zur numerischen Bestimmung der Nullstelle vom Gleichungssystem (DFC) vor. Wir werden auf die Möglichkeit verweisen, das Newton-Verfahren durch ein so genanntes Forward-Backward-Sweep-Verfahren zu ersetzen, ohne für ein solches Verfahren die Konvergenz theoretisch zu untersuchen. Anschließend werden wir kurz die Algorithmen vorstellen, die wir in Kapitel 5 benutzen, um die Konvergenzordnung in Abhängigkeit von der Schrittweite der Diskretisierung an numerischen Beispielen zu demonstrieren.

4.1 Anwendung des Newton-Verfahren

In diesem Abschnitt werden wir die Konvergenz des Newton-Verfahrens zeigen. Wir betrachten für den Operator $P_h : X_{h,r} \to \mathcal{Y}$ die Newton-Iteration

$$w_{n+1} = w_n - P_h'(w_n)^{-1} P_h(w_n), \tag{4.1.1}$$

die sich als Lösung des Gleichungssystems

$$P_h(w_n) + P_h'(w_n)(w - w_n) = 0$$

ergibt. Dabei ist $P_h'(w_n)$ die Fréchet-Ableitung des Operators P_h an der Stelle w_n und die Inverse dazu $P_h'(w_n)^{-1}$. In einigen Resultaten über die Konvergenz des Newton-Verfahrens in Banachräumen wird die Existenz der Nullstelle des Operators P_h ebenfalls gezeigt. Da wir die Existenz einer Nullstelle $\tilde{w}_h$ mit $P_h(\tilde{w}_h) = 0$

bereits gezeigt haben, können wir hier ein einfacheres Resultat benutzen. Wir verwenden den folgenden Satz 4.1.1, welcher der Proposition 5.1 in [Zei95] über die lokale Konvergenz des Newton-Verfahrens entspricht. Dort ist dieser Satz für einen Operator der von einem Banachraum $\mathcal{X}$ in $\mathcal{X}$ abbildet ausgeführt. Da sich an dem Beweis nichts wesentliches ändert, wenn der Operator von dem Raum X in einen Banachraum $\mathcal{Y}$ abbildet, verzichten wir auf den Beweis.

Satz 4.1.1. *Es seien die Banachräume $\mathcal{X}$, $\mathcal{Y}$ und die abgeschlossenen Teilmengen $X \subseteq \mathcal{X}$, $Y \subseteq \mathcal{Y}$ gegeben. Weiterhin seien $P : X \to Y$ und ein $\tilde{w} \in X$ mit $P(\tilde{w}) = 0$ gegeben. Die Fréchet-Ableitung $P'(w)$ existiere für alle w in einer offenen Umgebung U von $\tilde{w}$ in X und sei dort lipschitzstetig, das heißt es gilt*

$$\|P'(w_1) - P'(w_2)\| \leq M_{P'}\|w_1 - w_2\|_{\mathcal{X}}$$

für alle $w_1, w_2 \in U$ mit einer Konstanten $M_{P'}$. Außerdem existiere $P'(\tilde{w})^{-1}$ als stetiger lineare Operator von $\mathcal{Y}$ nach $\mathcal{X}$.

Dann gibt es ein $\delta > 0$, so dass für jeden Startpunkt w_0 mit $\|w_0 - \tilde{w}\|_{\mathcal{X}} \leq \delta$ die Folge w_n aus Gleichung (4.1.1) gegen die Lösung $\tilde{w}$ konvergiert. Weiterhin gilt mit einer Konstanten M, die von der Umgebung U und der Lipschitzkonstante $M_{P'}$ abhängt

$$\|w_n - \tilde{w}\|_{\mathcal{X}} \leq M\|w_{n+1} - \tilde{w}\|_{\mathcal{X}}^2, \quad \|w_n - \tilde{w}\|_{\mathcal{X}} \leq \frac{(M\delta)^{2n}}{M}.$$

Im Satz 4.1.1 verzichten wir auf eine besondere Kennzeichnung der entsprechenden Operatorennorm und bezeichnen sie mit $\|\cdot\|$. Um diesen Satz auf die Operatoren P_h anwenden zu können, muss man im wesentlichen Eigenschaften des Operators P_h zeigen, die wir durch die Bedingungen (B1) bis (B3) des Lemma 3.3.1 sichern können. Dabei hat man lediglich die Lipschitzstetigkeit der Fréchet-Ableitung noch nicht gezeigt. Daher werden wir für den Operator P_h zeigen dass für alle $w_1, w_2 \in X_{h,r}$ und $\nu \in \mathcal{X}$

$$\sup_{\substack{\nu \in \mathcal{X} \\ \nu \neq 0}} \frac{\|P_h{}'(w_1)(\nu) - P_h{}'(w_2)(\nu)\|_{\mathcal{Y}}}{\|\nu\|_{\mathcal{X}}} \leq \zeta_{4.1.2}\|w_1 - w_2\|_{\mathcal{X}} \tag{4.1.2}$$

gilt, bevor wir den Satz 4.1.1 anwenden.

4.1.1 Lipschitzstetigkeit von P_h'

Im Abschnitt 3.4 haben wir bereits die Ableitung des Operators $P_h(w)$ für $w \in X_{h,r}$ berechnet. Die ersten Komponenten von $P_h(w)$ sind für $t \in (0, T]$ mit j aus Gleichung (3.3.3)

$$x(t) - a - \sum_{i=0}^{j-1} h \sum_{k=1}^{s} b_k f(y_{i,k}, u(t_{i,k})) - h \sum_{k=1}^{s} b_k(\theta) f(y_{j,k}, u(t_{j,k})).$$

Bestimmt man die Ableitung nach w an der Stelle $w = (x, \lambda, u) \in X_{h,r}$, angewandt auf ein $\nu \in \mathcal{X}$, erhält man

$$\nu_1(t) - \sum_{i=0}^{j-1} h \sum_{k=1}^{s} b_k \mathcal{R}_{i,k}(w, \nu) - h \sum_{k=1}^{s} b_k(\theta) \mathcal{R}_{j,k}(w, \nu)\,,$$

wobei wir $\mathcal{R}_{i,k}(w, \nu)$ für $i = 0, \ldots, N-1$ und $k = 1, \ldots, s$ durch

$$\mathcal{R}_{i,k}(w, \nu) = f_x(y_{i,k}(w), u(t_{i,k}))(y_{i,k}{}'(w)(\nu)) + f_u(y_{i,k}(w), u(t_{i,k}))(\nu_3(t_{i,k}))$$

definieren. Berechnet man die Differenz auf der linken Seite von Gleichung (4.1.2) erhält man

$$\sum_{i=0}^{j-1} h \sum_{k=1}^{s} b_k \left(\mathcal{R}_{i,k}(w_2, \nu) - \mathcal{R}_{i,k}(w_1, \nu)\right) + h \sum_{k=1}^{s} b_k(\theta) \left(\mathcal{R}_{j,k}(w_2, \nu) - \mathcal{R}_{j,k}(w_1, \nu)\right)$$

und wir betrachten die Terme

$$\begin{aligned}(4.1.3)\quad & \mathcal{R}_{i,k}(w_1, \nu) - \mathcal{R}_{i,k}(w_2, \nu) = \\ & \Big(\big[f_u(y_{j,k}(w_1), u_1(t_{j,k})) - f_u(y_{j,k}(w_2), u_2(t_{j,k}))\big](\nu_3(t_{j,k})) \\ & + f_x(y_{j,k}(w_1), u_1(t_{j,k}))(y_{j,k}{}'(w_1)(\nu)) - f_x(y_{j,k}(w_2), u_2(t_{j,k}))(y_{j,k}{}'(w_2)(\nu))\Big).\end{aligned}$$

Damit man Ungleichung (4.1.2) zeigen kann, muss man also lediglich für $y_{j,k}(w)$ und $y_{j,k}{}'(w)(\nu)$ die lipschitzstetige Abhängigkeit von $w \in X_{h,r}$ zeigen. Die Abschätzung der entsprechenden $W^{1,\infty}$-Normen bei der Norm auf $\mathcal{Y}$ in Gleichung (4.1.2) folgt wieder analog zu der Lipschitzstetigkeit von $F - P_h$ in Abschnitt 3.4, da hier die auftretenden Differenzen von $\mathcal{R}_{i,k}(w_{1/2}, \nu)$ auf den Intervallen (t_j, t_{j+1}) ebenfalls nicht von t abhängen. Die Lipschitzstetigkeit der dritten und vierten Komponenten in P_h folgt durch die Lipschitzstetigkeit der Funktionen f und Ψ und ihrer Ableitungen. Somit muss man zeigen, dass

$$(4.1.4)\qquad \|y_{j,k}(w_1) - y_{j,k}(w_2)\|_2 \le \zeta_{4.1.4} \|w_1 - w_2\|_{\mathcal{X}},$$

$$(4.1.5)\qquad \sup_{\substack{\nu \in \mathcal{X} \\ \nu \neq 0}} \frac{\|y_{j,k}{}'(w_1)(\nu) - y_{j,k}{}'(w_2)(\nu)\|_2}{\|\nu\|_{\mathcal{X}}} \le \zeta_{4.1.5} \|w_1 - w_2\|_{\mathcal{X}}$$

für alle $j = 0, 1, 2, \ldots, N-1$ und alle $k = 1, 2 \ldots, s$ gilt.

Als erstes zeigen wir, dass die Hilfspunkte $y_{j,k}(w)$ lipschitzstetig von $w \in X_{h,r}$ abhängen. Benutzt man die Bezeichnungen aus dem Beweis von Lemma 3.3.5 und ordnet die einzelnen Spaltenvektoren $y_{j,k}(w)$ zu Matrizen $y_j(w)$ an, dann ist $y_j(w_i)$ Fixpunkt von $G_{w_i}(\cdot)$, $i = 1, 2$. Da im Beweis von Satz 3.3.5 gezeigt wird, dass $G_{w_i}(\cdot)$ für hinreichend kleine Schrittweiten h eine Kontraktion ist, folgt mit $\zeta_{4.1.6} := h_{3.3.8}\, M_f^0\, \max_{k=1,\ldots,s} \bar{c}_k < 1$

$$\begin{aligned}(4.1.6)\quad & \|y_j(w_1) - y_j(w_2)\|_{max,2} = \|G_{w_1}(y_j(w_1)) - G_{w_2}(y_j(w_2))\|_{max,2} \\ & \le \|G_{w_1}(y_j(w_1)) - G_{w_1}(y_j(w_2))\|_{max,2} + \|G_{w_1}(y_j(w_2)) - G_{w_2}(y_j(w_2))\|_{max,2} \\ & \le \zeta_{4.1.6} \|y_j(w_1) - y_j(w_2)\|_{max,2} + \|G_{w_1}(y_j(w_2)) - G_{w_2}(y_j(w_2))\|_{max,2}\end{aligned}$$

und daraus

$$\|y_j(w_1) - y_j(w_2)\|_{max,2} \leq \frac{1}{1-\zeta_{4.1.6}} \|G_{w_1}(y_j(w_2)) - G_{w_2}(y_j(w_2))\|_{max,2}.$$

Die Differenz auf der rechten Seite besteht dabei nach Gleichung (3.3.7) aus den Spaltenvektoren

$$x_1(t_j) + h\sum_{l=1}^{s} a_{k,l} f(y_{j,l}(w_2), u_1(t_j + c_l h)) \\ - x_2(t_j) + h\sum_{l=1}^{s} a_{k,l} f(y_{j,l}(w_2), u_2(t_j + c_l h)).$$

Mit der Lipschitzstetigkeit von f folgt daraus sofort, dass

$$\|G_{w_1}(y_j(w_2)) - G_{w_2}(y_j(w_2))\|_{max,2} \leq \zeta_{4.1.7} \|w_1 - w_2\|_{\mathcal{X}} \tag{4.1.7}$$

mit einer geeigneten Konstanten $\zeta_{4.1.7}$ gilt, und es folgt

$$\|y_j(w_1) - y_j(w_2)\|_{max,2} \leq \frac{\zeta_{4.1.7}}{1-\zeta_{4.1.6}} \|w_1 - w_2\|_{\mathcal{X}}.$$

Natürlich kann man die Lipschitzstetigkeit der Hilfspunkte $y_{j,k}(w)$ auch sofort aus der Existenz und Beschränktheit der Ableitungen $y_{j,k}{}'$ folgern. Allerdings möchten wir jetzt diesen Weg für die Hilfspunkte $y_{j,k}{}'$ wiederholen, um die Lipschitzstetigkeit gemäß Ungleichung (4.1.5) zu zeigen. Wir benutzen hierzu im Folgenden die Bezeichnungen aus dem Beweis von Lemma 3.3.6 und ordnen die Spaltenvektoren für jedes vorgegebene $\nu \in \mathcal{X}$ wieder zu Matrizen $y_j{}'(w)(\nu)$ und $G_{w,\nu}(y_j{}'(w)(\nu))$ an. Dabei sind $w_i \in X_{h,r}$, $i = 1, 2$, und es gilt

$$\begin{aligned}\|y_j{}'(w_1)(\nu) - y_j{}'(w_2)(\nu)\|_{max,2} \\ &\leq \|G_{w_1,\nu}(y_j{}'(w_1)(\nu)) - G_{w_1,\nu}(y_j{}'(w_2)(\nu))\|_{max,2} \\ &+ \|G_{w_1,\nu}(y_j{}'(w_2)(\nu)) - G_{w_2,\nu}(y_j{}'(w_2)(\nu))\|_{max,2} \\ &\leq \zeta_{3.3.9} \|y_j{}'(w_1)(\nu) - y_j{}'(w_2)(\nu)\|_{max,2} \\ &\quad + \|G_{w_1,\nu}(y_j{}'(w_2)(\nu)) - G_{w_2,\nu}(y_j{}'(w_2)(\nu))\|_{max,2},\end{aligned}$$

da $y_j{}'(w_i)(\nu)$ für $i = 1, 2$ jeweils Fixpunkte von $G_{w_i,\nu}(\cdot)$ sind. Damit folgt

$$\begin{aligned}\|y_j{}'(w_1)(\nu) - y_j{}'(w_2)(\nu)\|_{max,2} \\ &\leq \frac{1}{1-\zeta_{3.3.9}} \|G_{w_1,\nu}(y_j{}'(w_2)(\nu)) - G_{w_2,\nu}(y_j{}'(w_2)(\nu))\|_{max,2}\,,\end{aligned}$$

und in der Differenz auf der rechten Seite ist für $i = 1, 2$ die k-te Spalte der Matrix $G_{w_i,\nu}(y_j{}'(w_2)(\nu))$

$$\nu_1(t_j) + h\sum_{l=1}^{s} a_{k,l} \big[f_x(\tilde{y}_{j,l}(w_i), u_i(t_{j,l}))(y_{j,l}{}'(w_2)(\nu)) + f_u(\tilde{y}_{j,l}(w_i), u_i(t_{j,l}))(\nu_3(t_{j,l})) \big].$$

Da sowohl die Parameter $a_{k,l}$ beschränkt sind als auch die Hilfspunkte $\tilde{y}_{j,l}(w_i)$ lipschitzstetig von w_i abhängen, erhält man die gewünschte Lipschitzstetigkeit aus der Lipschitzstetigkeit der ersten Ableitungen von f und der Beschränktheit der Vektoren $\nu_3(t_{j,l})$. Es gilt $\|\nu\|_{\mathcal{X}} < \infty$, weil man die Differenzen im Zähler der linken Seite von Gleichung (4.1.5) nur für fest gewähltes $\nu \in \mathcal{X}$ betrachtet. Da nach Lemma 3.3.5 die Hilfspunkte $y_{j,k}(w)$ beschränkt sind, $\|w_2 - w^*\|_{\mathcal{X}} \leq r$ ist und die ersten Ableitungen der Funktion f lipschitzstetig sind, gibt es nach Gleichung (3.3.11) eine positive Konstante $\zeta_{4.1.8}$, so dass

$$\|{y_{j,l}}'(w_2)(\nu)\|_2 \leq (1 + \zeta_{4.1.8})\|\nu\|_{\mathcal{X}} \tag{4.1.8}$$

gilt. Somit ist gezeigt, dass es eine Konstante $\zeta_{4.1.9}$ mit

$$\|G_{w_1,\nu}({y_j}'(w_2)(\nu)) - G_{w_2,\nu}({y_j}'(w_2)(\nu))\|_{max,2} \leq \zeta_{4.1.9}\|w_1 - w_2\|_{\mathcal{X}}\|\nu\|_{\mathcal{X}} \tag{4.1.9}$$

gibt, woraus sofort die Behauptung folgt.

Damit hat man für alle auftretenden Terme in Gleichung (4.1.3) eine entsprechende Beschränktheit oder Lipschitzstetigkeit gezeigt und somit für die ersten Komponenten die Ungleichung (4.1.2). Berücksichtigt man, dass in den zweiten Komponenten von P_h schon die ersten Ableitungen von f vorkommen und man dementsprechend die Lipschitzstetigkeit höherer Ableitungen von f benötigt, kann man für diese Komponenten eine entsprechende Abschätzung völlig analog zeigen. Die letzten Komponenten, also $\lambda(0) - \big(\Psi_x(x(T))\big)^\top$, bereiten ebenfalls keine Problem. Für diese Komponenten steht in ${P_h}'(w_1)(\nu) - {P_h}'(w_2)(\nu)$ der Term

$$[\Psi_{xx}(x_2(T)) - \Psi_{xx}(x_1(T))]\,\nu_1(T)$$

und man führt die Abschätzung in (4.1.2) direkt auf die Lipschitzstetigkeit der zweiten Ableitung von Ψ zurück. Bildet man für die Komponenten $P_h^3(w)(t) = f_u(x(t), u(t))^\top \lambda(T - t)$ die erste Ableitung nach w, so folgt eine entsprechende Abschätzung (4.1.2) aus der Lipschitzstetigkeit der zweiten Ableitungen von f und der Beschränktheit der $w \in X_{h,r}$. Damit hat man insgesamt (4.1.2) gezeigt und kann die lokale Konvergenz des Newton-Verfahrens folgern. Dies fassen wir kurz im nächsten Abschnitt zusammen.

4.1.2 Konvergenz des Newton-Verfahrens

Wir wollen Satz 4.1.1 anwenden, um die Konvergenz des Newton-Verfahrens zu zeigen. Daher werden wir kurz zusammenfassen, dass die Voraussetzungen hierfür erfüllt sind. Die Lipschitzstetigkeit der ersten Ableitung des Operators P_h haben wir im Abschnitt 4.1.1 gezeigt. Außerdem haben wir für hinreichend kleine Schrittweiten h und Umgebungen um w^* die Existenz von einem $\tilde{w} \in X_{h,r}$ mit $P_h(\tilde{w}) = 0_{\mathcal{Y}}$ in Satz 3.3.3 gezeigt.

Damit bleibt zu zeigen, dass $[{P_h}'(\tilde{w})]^{-1}$ existiert. In Abschnitt 3.5 wurde gezeigt, dass F^{-1} existiert und jedem $\pi \in \mathcal{Y}$ ein $F^{-1}(\pi) \in X_h$ zuordnet. Da der Operator F linear ist, ist die Ableitung an jeder Stelle der Operator selbst. Es gilt somit $F'(\tilde{w}) = F$, und damit existiert $[F'(\tilde{w})]^{-1} = F^{-1}$, und F^{-1} ist lipschitzstetig mit

Lipschitzkonstante c_0. Weiterhin haben wir in Abschnitt 3.4 gezeigt, dass $\|F'(\tilde{w}) - P_h'(\tilde{w})\| \leq \varepsilon$ ist, da $\tilde{w} \in X_{h,r}$ gilt. Außerdem kann man $h_{3.3.6}$ und $r_{3.3.6}$ so wählen, dass $\varepsilon c_o < 1$ für alle $h \leq h_{3.3.6}$ und $r \leq r_{3.3.6}$ ist. Es folgt

$$\|I - [F'(\tilde{w})]^{-1} P_h'(\tilde{w})\| \leq \varepsilon c_0 < 1\,,$$

und man kann mit Hilfe der Neumann-Reihe schließen, dass die Inverse von

$$I - I + [F'(\tilde{w})]^{-1} P_h'(\tilde{w}) = F^{-1} P_h'(\tilde{w})$$

existiert. Weiterhin gilt $\|\left[F^{-1}P_h'(\tilde{w})\right]^{-1}\| \leq (1-\delta)^{-1}$ mit beliebigen $\varepsilon c_0 < \delta < 1$. Damit existiert auch der lineare Operator $\left[F^{-1}P_h'(\tilde{w})\right]^{-1} F^{-1} = \left[P_h'(\tilde{w})\right]^{-1}$, und die Voraussetzungen, um Satz 4.1.1 anzuwenden, sind erfüllt.

4.2 Numerische Bestimmung von Näherungslösungen

In diesem Abschnitt stellen wir konkrete Verfahren vor, um Näherungslösungen $\tilde{w}$ für w^* zu bestimmen. Mit diesen Verfahren wurden die numerischen Resultate im Kapitel 5 berechnet, welche die hohe Konvergenzordnung der Näherungslösungen bezüglich der Schrittweite h verdeutlichen. Dabei ist im Allgemeinen eine Schwierigkeit bei der Berechnung von Näherungslösungen $\tilde{w}$ die Wahl geeigneter Startwerte der Steuerungen an den Diskretisierungspunkten $u(t_{j,k})$ und $u(\bar{t}_{j,k})$. Weiterhin hat das Newton-Verfahren Nachteile, wodurch die Verwendung eines anderen Verfahrens sinnvoll sein kann. Ein einfacher Ansatz eines alternativen Verfahrens ist das sogenannte Forward-Backward-Sweep-Verfahren aus [LW07]. Bei diesem Verfahren kann die Wahl geeigneter Startwerte ebenfalls kritisch sein.

Löst man das Gleichungssystem (FC) mit Hilfe von SRKV numerisch, so hat man im Allgemeinen keine Information über die Art des kritischen Punktes vom Problem (OS), den man näherungsweise bestimmt. Es ergibt sich insbesondere kein diskretisiertes Optimierungsproblem, welches man benutzen kann, um Abstiegsverfahren zu konstruieren. Daher entsteht wiederum die Notwendigkeit, Startwerte zu bestimmen, die hinreichend nah an einem Minimum w^* liegen.

4.2.1 Kombinationen mit einem Abstiegsverfahren

Eine relativ einfache Methode, gute Startwerte für das Newton-Verfahren oder ein Forward-Backward-Sweep-Verfahren zu berechnen, ist einen Näherungswert für den Gradienten des ursprünglichen Problems zu berechnen und mit diesem ein „Abstiegsverfahren“ zu konstruieren. Gibt man eine Steuerung u vor, so kann man Zustände x_h näherungsweise mit Hilfe des SRKVs bestimmen und danach die adjungierten Zustände λ_h Näherungsweise berechnen. Dabei gilt, dass die Zustände x_h und adjungierten Zustände λ_h entsprechend nah bei den Funktionen x und λ liegen, die sich ergeben würden, könnte man die entsprechenden Anfangswertprobleme zu der gegebenen Steuerung u exakt lösen.

Für das ursprüngliche Problem (OS) kann man einen Gradienten zu jedem Zeitpunkt $t \in (0,T)$ durch

$$\lambda(T-t)^\top f_u(x(t),u(t))$$

berechnen (siehe [PBGM64] oder [LW07]). Setzt man hingegen die approximierten Zustände x_h und λ_h ein, approximiert man damit den Gradienten. Somit hat man für Gradienten mit einer verhältnismäßig großen Norm durch $\lambda_h(T-t)^\top f_u(x_h(t),u(t))$ eine Abstiegsrichtung gegeben. Diese Richtung kann jedoch offensichtlich keine gradientenbezogene Suchrichtung (siehe [Alt02]) für das Problem (OS) darstellen und wird hinreichend nah am Minimum keine Abstiegsrichtung sein. Weiterhin ist nicht klar, ob ein so konstruiertes Verfahren gegen den Punkt $\tilde{w}$ aus Satz 3.3.3 konvergiert.

Insgesamt bietet sich somit folgender, relativ einfach implementierbarer Algorithmus an, um das Gleichungssystem (DFC) zu lösen.

Algorithmus 4.2.1.

1. *Wähle Startwerte für die Steuerung an den Diskretisierungspunkten $u^0(t_{j,k})$ und $u^0(\bar{t}_{j,k})$. Setze den Iterationszähler $i=0$.*

2. *Berechne mit der ersten Gleichungen von* (DFC) *und* (3.3.4) *die Zustände $x^i(t_j)$ an den Gitterpunkten und die Vektoren $f(y^i_{j,k},u^i(t_{j,k}))$.*

3. *Berechne mit den zweiten Gleichungen von* (DFC) *und* (3.3.5) *die adjungierten Zustände $\lambda^i(t_j)$ und die Vektoren $f_x(x^i(\bar{t}_{j,k}),u^i(\bar{t}_{j,k}))^\top v^i_{j,k}$.*

4. *Berechne zu jedem Zeitpunkt $t_{j,k}$ und $\bar{t}_{j,k}$ die Vektoren*

$$d_{j,k} := f_u(x^i(t_{j,k}),u^i(t_{j,k}))^\top \lambda^i(T-t_{j,k})$$
$$\bar{d}_{j,k} := f_u(x^i(\bar{t}_{j,k}),u^i(\bar{t}_{j,k}))^\top \lambda^i(t_{j,k}).$$

5. *Berechne eine geeignete Schrittweite σ, aktualisiere die Steuerungen durch*

$$u^{i+1}(t_{j,k}) = u^i(t_{j,k}) + \sigma d_{j,k}$$
$$u^{i+1}(\bar{t}_{j,k}) = u^i(\bar{t}_{j,k}) + \sigma \bar{d}_{j,k}$$

   und erhöhe den Iterationszähler i um eins.

6. *Wenn die Änderungen $\sigma d_{j,k}$, $\sigma \bar{d}_{j,k}$ hinreichend groß und die Iterationszahl klein genug ist gehe zu Schritt 2, ansonsten weiter mit Schritt 7.*

7. *Setze einen neuen Zähler $l=0$. Berechne die Zustände und adjungierten Zustände zur aktuellen Steuerung anhand der Schritte 2 und 3.*

8. *Berechne die Ableitungen $J(t_{j,k})$ und $J(\bar{t}_{j,k})$ der Terme $d_{j,k}$ und $\bar{d}_{j,k}$ aus Schritt 4 nach den Steuerungen zu den Diskretisierungszeitpunkten.*

9. *Berechne den Vektor $\tilde{d}$ aus dem linearen Gleichungssystem $J\tilde{d} = d$ (siehe Bemerkung 4.2.3).*

10. *Berechne die neuen Steuerungen nach dem Prinzip $u^{i+1} = u^i - \hat{d}$, wobei $\hat{d}$ zu jedem Diskretisierungszeitpunkt aus den entsprechenden Komponenten von $\tilde{d}$ besteht. Erhöhe den Iterationszähler l um eins und berechne die Zustände und adjungierten Zustände zur aktuellen Steuerung entsprechend den Schritten 2 und 3 neu.*

11. *Berechne die Vektoren $d_{j,k}$ und $\bar{d}_{j,k}$ entsprechend Schritt 4 neu und wiederhole die Schritte 8 bis 10 solange bis die Norm der Vektoren $d_{j,k}$ und $\bar{d}_{j,k}$ hinreichend klein ist, oder eine maximale Anzahl an Iterationen erreicht ist. Beende danach das Programm mit den notwendigen Ausgaben.*

Bemerkung 4.2.2: In Schritt 3 und 4 benötigt man neben den Steuerungen $u(\bar{t}_{j,k})$ die Zustände $x(\bar{t}_{j,k}) = x(T - t_j - c_k h)$. Diese berechnet man natürlich ebenfalls mit Hilfe der ersten Gleichung in (DFC). Dabei benötigt man keine weiteren Auswertungen der Funktion f, wenn man jeweils die aktuellen Vektoren $f(y_{j,k}, u(t_{j,k}))$ in Schritt 2 speichert. Ebenso benötigt man in Schritt 4 die adjungierten Zustände an Zwischenpunkten, die sich wiederum durch die Verwendung eines SRKVs aus den Gleichungen in (DFC) berechnen lassen.

Bemerkung 4.2.3: In Schritt 8 ist es natürlich notwendig, um ein Newton- oder Sekantenverfahren zu realisieren, die Auswirkungen einer Änderung der Steuerung zu allen Diskretisierungszeitpunkten auf die Zustände und adjungierten Zustände an den jeweiligen Zeitpunkten $t_{j,k}$ und $\bar{t}_{j,k}$ zu berücksichtigen. Daher kann man diese Ableitungen zu einer quadratischen Matrix J der Dimension $2mNs$ zusammenfassen. Außerdem kann man die Terme $d_{j,k}$ und $\bar{d}_{j,k}$ als einen Vektor d der Dimension $2mNs$ auffassen.

4.2.2 Forward-Backward-Sweep-Verfahren

In Bemerkung 4.2.3 wird deutlich, dass man bei der Implementierung des Newton-Verfahrens die $(2mNs)^2$ Einträge der Matrix J benötigt. Da hier der Speicherplatz quadratisch von der Anzahl der Diskritisierungszeitpunkte $t_{j,k}$ und $\bar{t}_{j,k}$ abhängt, wächst der benutzte Speicherplatz sehr schnell mit kleiner werdender Schrittweite an. Außerdem benötigt man für die Lösung des linearen Gleichungssystems $J\tilde{d} = d$ in der Regel $\mathcal{O}((2mNs)^3)$ und bei Verwendung von iterativen Verfahren immer noch $\mathcal{O}((2mNs)^2)$ Rechenoperationen.

Neben modernen Verfahren für die Newton-Iterationen, die den Rechenaufwand und den Speicherplatzbedarf für große Probleme verringern sollen, kann es daher sinnvoll sein, andere Verfahren zu benutzen, um eine Nullstelle der dritten Gleichung in (DFC) zu berechnen. Eine sehr einfach zu implementierende Variante ist das sogenannte Forward-Backward-Sweep-Verfahren aus [LW07]. Eine ähnliche Herangehensweise findet sich auch in dem Algorithmus von Y. Sakawa in [Sak81]. Bei dem Verfahren aus [LW07] berechnet man zu einer gegebenen Steuerung die Zustände und adjungierten Zustände an den jeweiligen Diskretisierungszeitpunkten der Steuerung und sucht Nullstellen $u_{j,k}$ der Funktionen $d_{j,k}(u) := f_u(x^i(t_{j,k}), u)^\top \lambda^i(\bar{t}_{j,k})$ und völlig analog entsprechende Nullstellen an den Zeitpunkten $\bar{t}_{j,k}$. Im einfachsten Fall benutzt man in der nächsten Iteration die

Steuerungen $u^{i+1}(t_{j,k}) = u_{j,k}$ und berechnet damit die Zustände und adjungierten Zustände neu.

Bei dieser Vorgehensweise tauscht man den Aufwand, das große lineare Gleichungssystem zu lösen, dagegen ein, dass man in jeder Iteration $2Ns$-mal eine Nullstelle der Funktion $f_u(x^i(t_{j,k}), u)^\top \lambda^i(\bar{t}_{j,k})$ berechnen muss. Selbst wenn man dabei wiederum Ableitungen dieser Terme nach u berechnet, hängt der Speicherplatzbedarf zwar quadratisch von der Dimension m ab, aber nur linear von $2Ns$. Der benötigte Rechenaufwand ist hier kaum abschätzbar, da man in der Regel die Nullstellen von $d_{j,k}(u)$ über ein iteratives Verfahren suchen muss. Allerdings steigt der Rechenaufwand ebenfalls nur linear mit $2Ns$ an, das heißt der Rechenaufwand steigt mit kleiner werdender Schrittweite nur linear. Weiterhin kann man die Bestimmung der $2Ns$ Nullstellen sehr leicht parallel durchführen, da man die Zustände und adjungierten Zustände nur einmal in jeder Iteration neu berechnen muss.

Demgegenüber stehen schlechtere Konvergenzeigenschaften im Vergleich mit dem Newton-Verfahren. Ist der Faktor $2Ns$ relativ klein, so ist das Newton-Verfahren deutlich schneller. In der praktischen Anwendung hat sich gezeigt (siehe [LW07]), dass man die Konvergenzeigenschaften teilweise verbessern kann, wenn man nicht direkt die gefundenen Nullstellen als neue Steuerungen im nächsten Iterationsschritt benutzt, sondern Konvexkombinationen der Art

$$u^{i+1}(t_{j,k}) = \iota u^i(t_{j,k}) + (1-\iota)u_{j,k} \text{ mit } 0 < \iota < 1.$$

Diese Konvexkombination kann man auch in Abhähngigkeit von der aktuellen Iterationszahl zum Beispiel durch $u^{i+1}(t_{j,k}) = \iota^i u^i(t_{j,k}) + (1-\iota^i)u_{j,k}$ wählen und dadurch versuchen, die Konvergenzgeschwindigkeit wieder zu verbessern.

Da dieses Verfahren im Allgemeinen nicht für alle Startwerte konvergiert, bietet sich wiederum eine Kombination analog zum Algorithmus 4.2.1 an, indem man nur das Newton-Verfahren durch eine entsprechende Forward-Backward-Sweep-Iteration ersetzt.

5 Numerische Resultate

In diesem Kapitel demonstrieren wir anhand von einfachen Beispielen die hohe Konvergenzordnung, die man mit Hilfe von SRKV bei Steuerungsproblemen erreichen kann. Außerdem werden wir die mit Hilfe der SRKV berechneten Steuerungen oder Zustände darstellen. Die berechneten Zustände und adjungierten Zustände haben dabei natürlich immer die entsprechnenden Eigenschaften aus den SRKV und sind zum Beispiel C^1-Funktionen, wenn man C^1-Verfahren benutzt. Sind dabei die Voraussetzungen von Satz 3.3.3 erfüllt und gilt die dritte Gleichung aus (DFC), dann kann man die Steuerung aus dem Minimumprinzip berechnen und diese hat ebenfalls die entsprechenden Eigenschaften.

Wir betrachten als erstes die beiden linear-quadratischen Probleme (P1) und (P2) aus [Hag76]. Diese Probleme eignen sich hierfür sehr gut, da man die eindeutig bestimmte optimale Steuerung analytisch berechnen kann und sie ebenfalls in [Hag00] benutzt werden. Danach werden wir durch ein nichtlineares Beispiel zeigen, dass die hohe Konvergenzordnung nicht auf linear-quadratische Probleme beschränkt ist. Abschließend werden wir anhand eines sehr einfachen Beispiels mit Hilfe der GDNN aus Abschnitt 1.2 zeigen, dass man mit Hilfe der SRKV immer noch ein Abstiegsverfahren konstruieren kann, auch wenn die theoretischen Voraussetzungen für die Existenz einer Lösung des Gleichungssystems (DFC) und damit für die Konvergenz des Newton-Verfahrens nicht erfüllt sind.

5.1 Linear-Quadratische Testprobleme

Mit Hilfe der in Kapitel 4 vorgeschlagenen Algorithmen berechnen wir Näherungslösungen für feste Schrittweiten $h = T/N$ der folgenden Steuerungsprobleme:

$$\text{(P1)} \quad \min \quad \frac{1}{2}\int_0^1 u(t)^2 + 2x(t)^2\,dt$$
$$u.Nb. \quad \dot{x}(t) = 0.5x(t) + u(t) \quad \forall t \in [0,1], \quad x(0) = 1$$

und

$$\text{(P2)} \quad \min \quad \frac{1}{2}\int_0^1 u(t)^2 + \frac{5}{4}x(t)^2 + x(t)u(t)\,dt$$
$$u.Nb. \quad \dot{x}(t) = 0.5x(t) + u(t) \quad \forall t \in [0,1], \quad x(0) = 1.$$

Der wesentliche Unterschied zwischen beiden Problemen besteht in dem gemischten Term $x(t)u(t)$ in der Zielfunktion von (P2). Beide Probleme entsprechen zwar

zunächst nicht der Form (OS), können aber mit Hilfe eines weiteren Zustands in diese Form umgeschrieben werden. Alle auftretenden Funktionen sind offensichtlich hinreichend glatt.

Die analytisch berechneten Lösungen zu den Problemen (P1) und (P2) (siehe [Hag00]) sind:

$$\text{(S1)} \qquad x^*(t) = \frac{2e^{3t} + e^3}{e^{\frac{3t}{2}}(2 + e^3)}, \quad u^*(t) = \frac{2\,(e^{3t} - e^3)}{e^{\frac{3t}{2}}(2 + e^3)}$$

beziehungsweise

$$\text{(S2)} \qquad x^*(t) = \frac{\cosh(1-t)}{\cosh(1)}, \quad u^*(t) = -\frac{(\tanh(1-t) + 0.5)\cosh(1-t)}{\cosh(1)}.$$

Zu diesen Problemen haben wir numerische Lösungen mit Schrittweiten $h = 1/N$ für Werte von $N = 1$ bis $N = 544$ berechnet. Dabei erhält man direkt, aus dem Gleichungssystem (DFC), stetige Näherungslösungen der Zustände und adjungierten Zustände. Speichert man die entsprechenden Funktionswerte an den Zwischenpunkten, dann kann man den Zustand und den adjungierten Zustand zu jedem beliebigen Punkt $t \in [0, T]$ ohne weitere Auswertungen der Funktion f angeben.

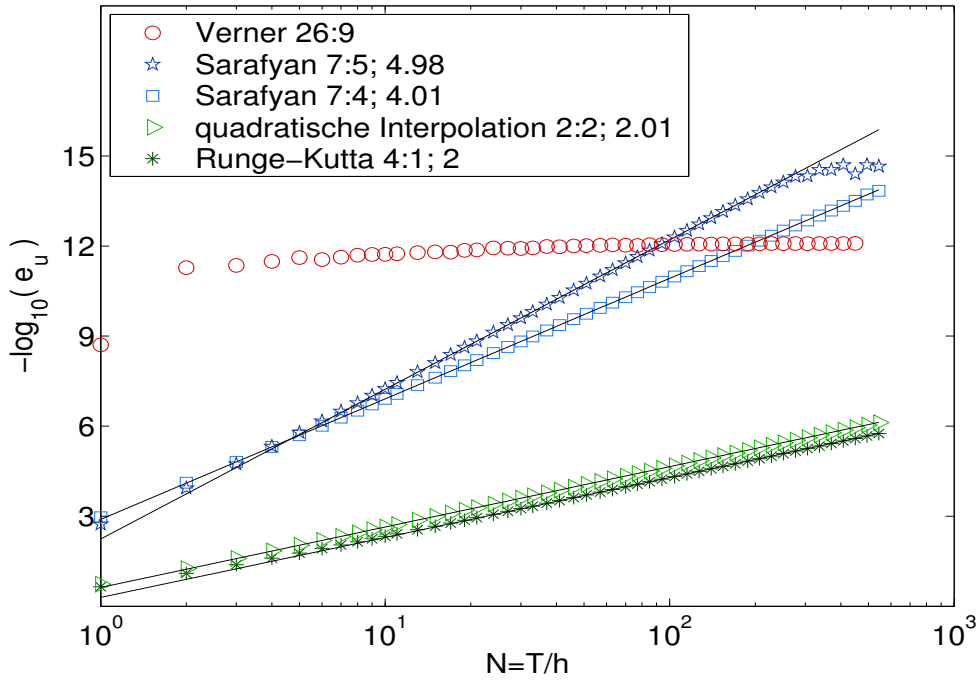
Die jeweiligen Steuerungen benötigt man während der Berechnung nur an den Diskretisierungspunkten $u(t_{j,k})$ und $u(\bar{t}_{j,k})$. An diesen Punkten liefern die Algorithmen direkt eine Näherungslösung für die optimale Steuerung, und an allen anderen Punkten kann man die Steuerung über das Minimumprinzip berechnen. Wie in Kapitel 4 beschrieben, muss dies nur einmal am Ende der Iteration für diejenigen Punkte durchgeführt werden, an denen man eine Näherungslösung der Steuerung benötigt. Dabei ist die letzte Gleichung von (DFC) natürlich an den Diskretisierungspunkten $u(t_{j,k})$ und $u(\bar{t}_{j,k})$ erfüllt. Damit kann man sehr leicht gute Startwerte für ein iteratives Verfahren gewinnen, um eine entsprechende Nullstelle der dritten Gleichung von (DFC) zu berechnen.

Einen Diskretisierungsfehler der Steuerungen, im Sinn von $\|\tilde{u} - u^*\|_{L^\infty}$, kann man somit natürlich nur schätzen. Dazu haben wir, unabhängig von der Schrittweite h der Diskretisierung, ein äquidistantes Gitter mit 100 000 Gitterpunkten benutzt. Fassen wir diese Gitterpunkte in der Menge $\mathcal{G}$ zusammen, benutzen wir

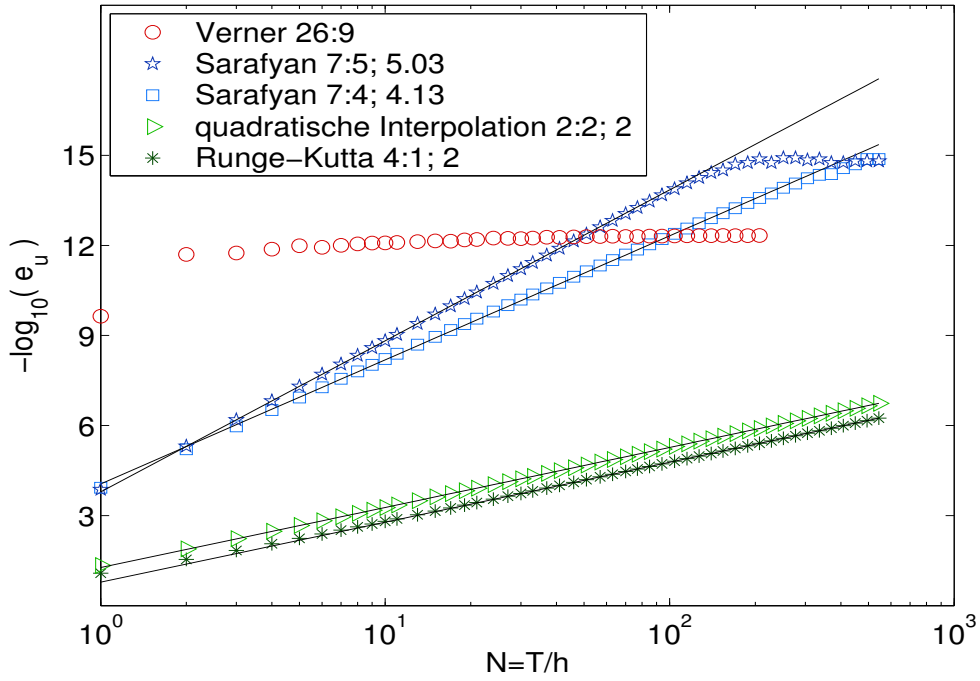
$$e_u := \max_{t \in \mathcal{G}} \|\tilde{u}(t) - u^*(t)\|_2$$

zum Schätzen des Diskretisierungsfehlers, der beim Lösen des Gleichungssystems (DFC) statt (FC) entsteht. Für beide Probleme konvergierten die Iterationsverfahren aus Kapitel 4 sehr gut, und man kann zum Beispiel $u(t) \equiv 0$ als Startwert für die Steuerung benutzen.

In Abbildung 5.1 ist dieser Diskretisierungsfehler für verschiedene SRKV in Abhängigkeit von der Anzahl N der Schritte des Einschrittverfahrens abgebildet. Zur besseren Darstellung wird N auf einer logarithmischen Achse dargestellt, und auf der Ordinate ist $-\log_{10}(e_u)$ aufgetragen. Somit entsprechen höhere Werte einem kleineren Fehler, und die Steigung der entstehenden Kurven spiegelt die Konvergenzordnung des Verfahrens wider. Da neben dem Diskretisierungsfehler weitere



(a) FBS-Verfahren für Problem (P1)



(b) Newton-Verfahren für Problem (P2)

Abbildung 5.1: Geschätzter Diskretisierungsfehler e_u der Steuerung in Abhängigkeit von der Schrittweite für Verfahren der Konvergenzordnung 9 (für gewöhnliche Differentialgleichungen) nach Verner (siehe [Ver]), der Ordnung 4 nach Sarafyan (siehe [CT96]), für Verfahren mit quadratischer Interpolation aus Abschnitt 2.2.2 mit dem Parameter $c_2 = 1/2$ und für das klassische Runge-Kutta-Verfahren mit linearer Interpolation.

Verfahren	Schritte N	Fehler e_u	CPU-Zeit
Verner 26:9	1	$2.29 * 10^{-10}$	0.34
Sarafyan 7:5	13	$4.03 * 10^{-10}$	0.6
Sarafyan 7:4	21	$2.76 * 10^{-10}$	1.49
quadratische Interpolation 2:2	494	$2.20 * 10^{-07}$	41.53
Runge-Kutta 4:1	494	$6.89 * 10^{-07}$	226.34

Tabelle 5.1: Ausgewählte real benötigte Rechenzeit in Sekunden zum Problem (P2) mit der jeweiligen Anzahl der Schritte der Runge-Kutta-Verfahren. Die Schrittweite wurde so gewählt, dass ein Fehler erreicht wird, der ungefähr in der selben Größenordnung liegt wie der Fehler des Verfahrens Verner 26:9 mit einem Schritt.

Rundungsfehler bei der Berechnung gemacht werden, kann man die Genauigkeit der Näherungslösungen nicht beliebig steigern.

Für beide Probleme liefern das Forward-Backward-Sweep- und das Newton-Verfahren jeweils identische Ergebnisse. Daher stellen wir jeweils nur die Fehler für ein Verfahren dar. Dabei konnte in beiden Fällen auf die Bestimmung eines guten Startwerts mit Hilfe des näherungsweisen Abstiegsverfahrens verzichtet werden.

In den Legenden der Abbildungen 5.1(a) und 5.1(b) kann man neben dem Namen der Verfahren bzw. dem Namen der Entwickler der Verfahren als erstes die Anzahl der Stufen des SRKVs, danach den Grad der Polynome $b_k(\theta)$ und als letztes den Anstieg der zu den jeweiligen abgebildeten Verfahren gehörenden Geraden ablesen. Allen Verfahren liegen explizite Runge-Kutta-Verfahren zugrunde. Das klassische Runge-Kutta-Verfahren mit linearer Interpolation hat nach der Definition 2.1.2 die Konvergenzordnung $(4, 1)$, womit Satz 3.3.3 lineare Konvergenz in Abhängigkeit von der Schrittweite sichert. Das 2-stufige explizite Verfahren mit quadratischer Interpolation hat theoretisch und praktisch quadratische Konvergenz. In Abbildung 5.1 sieht man deutlich, dass das klassische Runge-Kutta-Verfahren mit linearer Interpolation für kleinere Schrittweiten bei diesen Beispielen ebenfalls quadratische Konvergenz zeigt.

Das Verfahren Sarafyan 7:5 hat nach [CT96] die Konvergenzordnung $(5, 4)$ und das Verfahren Sarafyan 7:4 die Ordnung $(4, 4)$. Beide Verfahren sind 7-stufige C^1-Verfahren und erfüllen die Bedingung FSAL aus Abschnitt 3.7. Damit sind bei einer entsprechenden Implementierung jeweils 6 Funktionsauswertungen pro Schritt notwendig. In Abbildung 5.1(a) sieht man für das Verfahren Sarafyan 7:5 ebenfalls, dass man die höhere Konvergenzordnung auf dem Gitter erst für kleinere Schrittweiten ausnutzen kann. Im Gegensatz dazu sind die Beispiele wiederum einfach genug, um von Anfang an (ab $N = 1$ Schritte) die Konvergenzordnung 4 für das Verfahren Sarafyan 7:4 demonstrieren zu können.

Dem Verfahren Verner 26:9 liegt ein 16-stufiges, explizites, eingebettetes Runge-Kutta-Verfahren der Ordnung 9(8) zugrunde. Dieses Verfahren wurde nach [Ver] um jeweils 5 Stufen erweitert, um Interpolationen der Ordnung 8 und mit einem 26-stufigen Verfahren Interpolationen der Ordnung 9 angeben zu können. Die beiden Beispiele sind offensichtlich zu einfach, um damit die hohe Konvergenzordnung des Verfahrens Verner 26:9 demonstrieren zu können. Man sieht dabei den Ein-

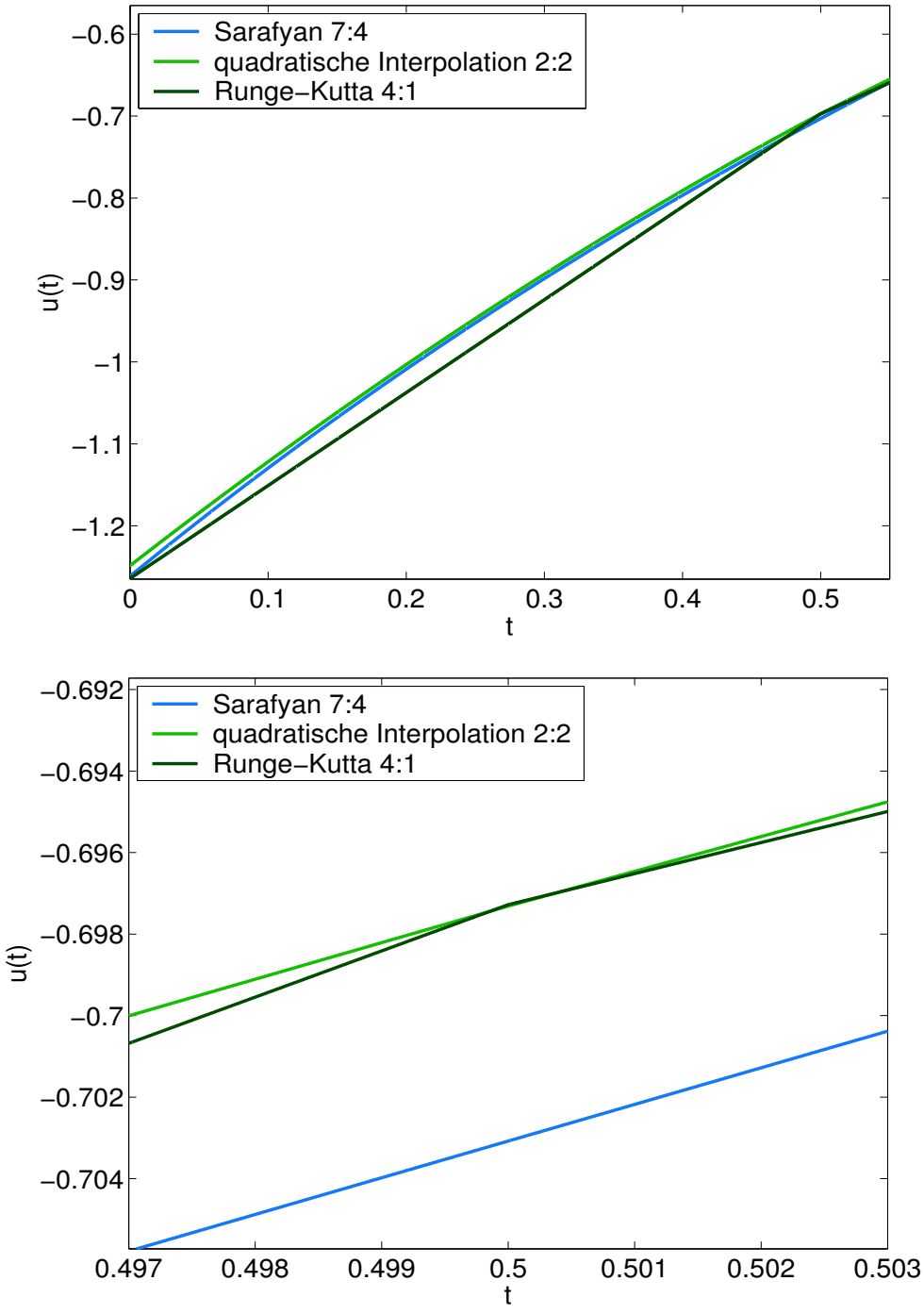


Abbildung 5.2: Ausschnitte aus den berechneten Näherungslösungen für die optimalen Steuerungen von Problem (P2), es wurden jeweils $N = 2$ Schritte benutzt, womit die Grenze zwischen beiden Schritten bei $t = 0.5$ ist.

fluss der Rundungsfehler sehr deutlich. Die größtmögliche Genauigkeit kann bei diesen Beispielen offensichtlich durch Verfahren mit kleinerer Konvergenzordnung und deutlich mehr Schritten erziehlt werden.

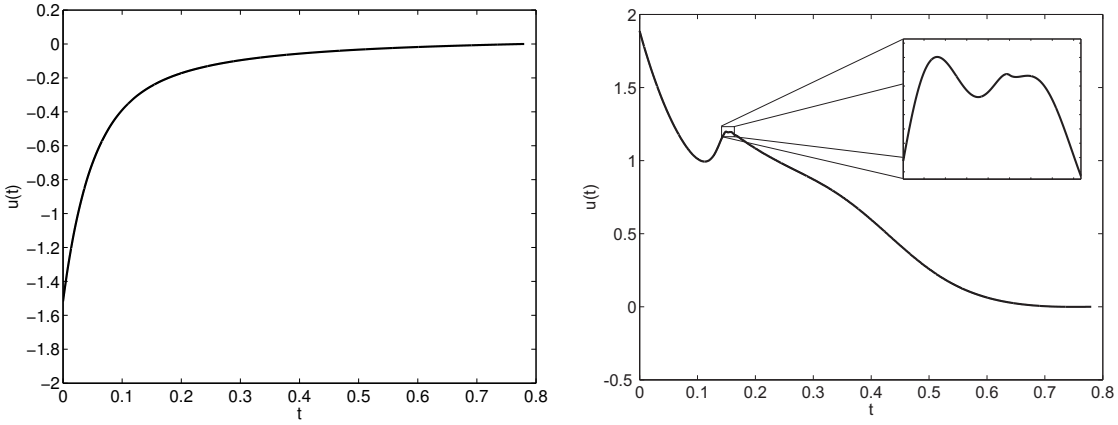
Da unsere Implementation der Algorithmen und Verfahren eher auf Flexibilität ausgelegt ist und damit relativ ineffizient bezüglich der Rechenzeit gestaltet ist, lohnt es sich nicht, den unterschiedlichen Rechenaufwand der verschiedenen Verfahren detailliert gegenüber zu stellen. Um eine grobe Einordnung vornehmen zu können, haben wir in Tabelle 5.1 zu jedem der fünf SRKV eine real benötigte CPU-Zeit dem jeweiligen damit erreichten Fehler gegenübergestellt. Dabei wurden die Werte aus den Ergebnissen zum Problem (P2) so ausgewählt, dass jeweils 3 Iterationen des Newton-Verfahrens durchgeführt wurden, bis das Abbruchkriterium erfüllt war und man dem Fehler des Verfahrens von J. Verner möglichst nahe kommt. Man sieht sehr deutlich, dass für das Newton-Verfahren dabei die Verfahren mit relativ niedriger Konvergenzordnung deutlich länger zur Berechnung benötigen, wenn man entsprechend mehr Schritte macht, um einen annähernd gleich guten Diskretisierungsfehler zu erhalten. Ein ähnliches Verhalten kann man ebenfalls für das Forward-Backward-Sweep-Verfahren bei den Problemen (P1) und (P2) beobachten, mit demselben Algorithmus benötigt man tendenziell weniger Rechenzeit, um denselben Fehler zu erreichen, wenn man Verfahren mit einer höheren Konvergenzordnung benutzt.

In der Abbildung 5.2 sind jeweils Ausschnitte aus den berechneten Näherungslösungen abgebildet. Dabei sieht man, dass das klassische Runge-Kutta-Verfahren mit linearer Interpolation natürlich entsprechend stückweise lineare Näherungslösungen liefert. Dagegen sieht man bei der berechneten Näherungslösung des C^1-Verfahrens Sarafyan 7:4 die Grenze zwischen den Schritten bei dem Problem (P2) schon bei $N = 2$ Schritten nicht mehr. Dabei ist diese Näherungslösung so genau, dass bei der Darstellung von Abbildung 5.2 die analytisch berechnete Lösung von dieser Näherungslösung überdeckt werden würde.

5.2 Nichtlineares Testproblem

Als nächstes betrachten wir ein einfaches nichtlineares Testproblem aus [LVV03] (siehe auch [FPA+99, AST97]). Dieses Testproblem stammt aus einem Modell zu einer irreversiblen chemischen Reaktion in einem gekühlten kontinuierlich betriebenen Rührkesselreaktor. Die chemische Reaktion wird dabei durch die Zustandsgleichungen

$$\dot{x}_1 = -(2+u)(x_1+0.25) + (x_2+0.5)\exp\left(\frac{25x_1}{x_1+2}\right)$$
$$\dot{x}_2 = 0.5 - x_2 - (x_2+0.5)\exp\left(\frac{25x_1}{x_1+2}\right)$$



(a) globale Lösung **(b)** weitere lokale Lösung

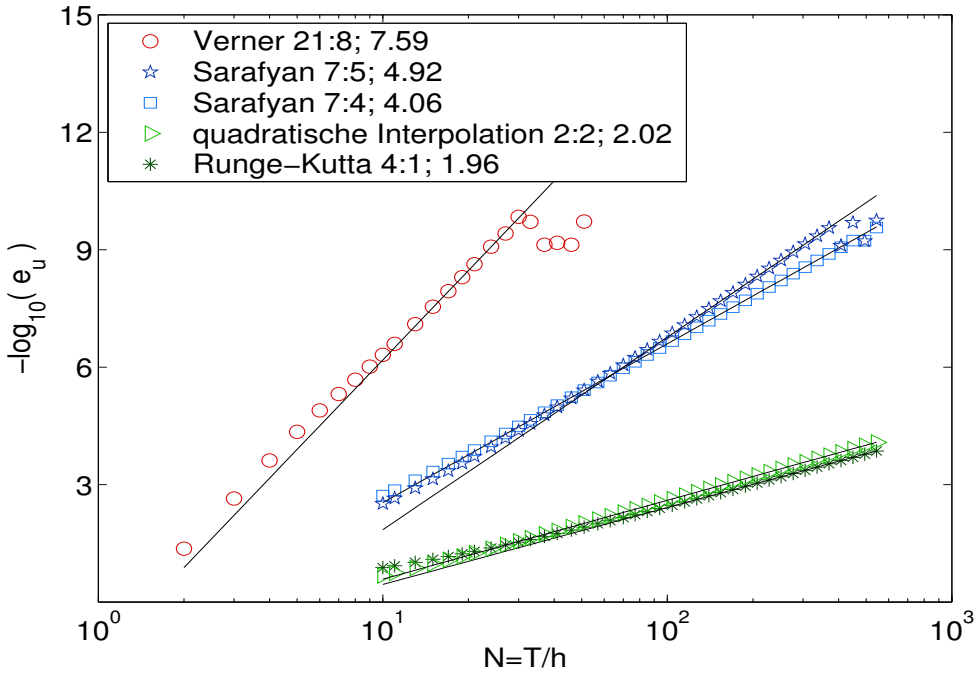
Abbildung 5.3: Näherungslösungen der optimalen Steuerungen für nichtlineares Testproblem, berechnet mit dem Verfahren Verner 21:8 mit $N = 70$ Gitterpunkten.

modelliert. Dabei kann über die Steuerung u die Kühlung beeinflusst werden. Das Problem besteht darin, die optimale Steuerung zu der Zielfunktion

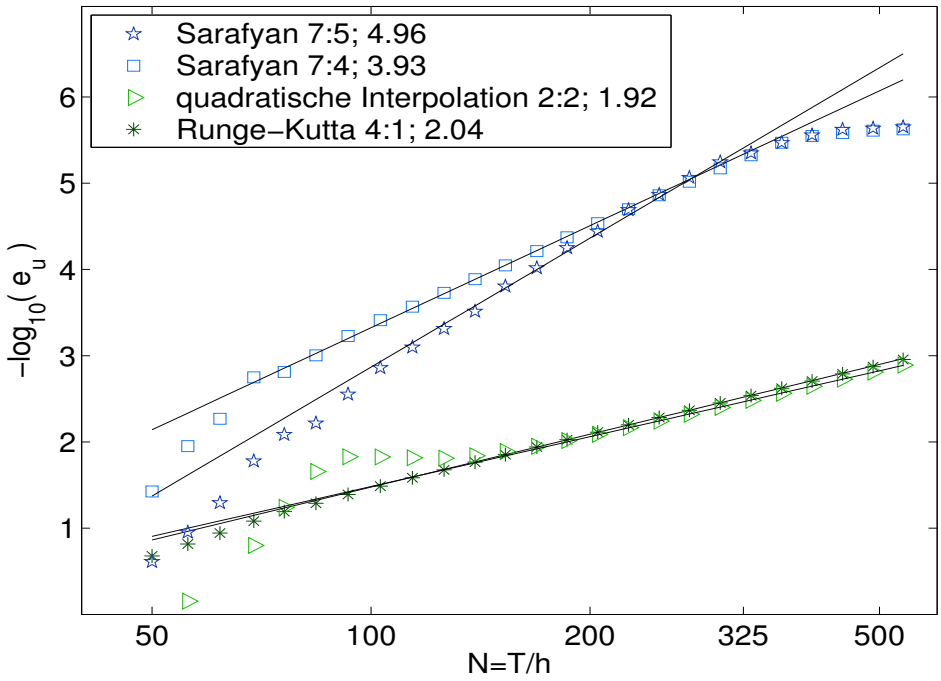
$$\min \quad \int_0^T (x_1(t)^2 + x_2(t)^2 + 0.1u(t)^2)\, dt$$

zu finden. Dabei sind $T = 0.78$ und $x_1(0) = x_2(0) = 0.09$. Dieses Problem besitzt nach [LVV03] ein globales und ein weiteres lokales Minimum. Für die beiden Minima sind keine analytisch berechneten Lösungen bekannt, und in Abbildung 5.3 sind Näherungslösungen dargestellt. Dabei hat insbesondere das lokale Minimum einen sehr interessanten Verlauf. Entsprechend ist auch die Näherungslösung zu dem lokalen Minimum, welche mit dem Verfahren Verner 21:8 und $N = 70$ Schritten berechnet wurde, etwas ungenauer als die Näherungslösung zu dem globalen Minimum. Dies sieht man deutlich in der Abbildung 5.4. Für eine Schätzung des Diskretisierungsfehlers wurden die Näherungslösungen aus Abbildung 5.3 als Referenz benutzt und der L^∞-Abstand zu weiteren Näherungslösungen mit geringerer Anzahl an Schritten geschätzt. Dabei wurde wiederum der maximale euklidische Abstand zwischen den beiden Näherungslösungen für die Steuerung auf einem äquidistanten Gitter mit 100 000 Gitterpunkten bestimmt. Der negative dekadische Logarithmus dieses Abstands zwischen der Referenzlösung und den Näherungslösungen ist in Abbildung 5.4 dargestellt.

Bei diesem Problem konvergieren die Iterationsverfahren aus Kapitel 4 nicht für jede natürliche Zahl von Schritten N. Dieses Verhalten hängt zudem von dem Startpunkt der Steuerung ab. Außerdem konvergieren die lokal konvergenten Verfahren, wenn man eine entsprechend große Anzahl an Schritten N benutzt, in Abhängigkeit von der initialen Steuerung zu einem der beiden Minima. Daher haben wir das näherungsweise Abstiegsverfahren und im Anschluss das Newton-Verfahren benutzt, um die Referenzlösungen zu berechnen. Wählt man dabei für das Abstiegsverfahren die Steuerung konstant $u(\cdot) = 2$, dann konvergiert das Verfahren zu dem lokalen



(a) globale Lösung



(b) weitere lokale Lösung

Abbildung 5.4: Geschätzter Diskretisierungsfehler für das nichtlineare Testproblem in Abhängigkeit von der Anzahl der Schritte N.

Minimum und für eine initiale Steuerung von konstant $u(\cdot) = 0$ zum globalen Minimum. Für alle weiteren Berechnungen wurden dann aus den Referenzlösungen die initialen Steuerungen für das Newton-Verfahren bestimmt.

Dabei sieht man in Abbildung 5.4(a), dass das Verfahren Verner 21:8 hier für einige Schrittweiten praktisch die Konvergenzordnung 7.6 aufweist. Bei diesem Problem ist allerdings die Grenze, an der man keine Verbesserung mehr erhält, nicht nur durch die Rundungsfehler, sondern vor allem auch durch den Abstand der Referenzlösung zu dem unbekannten globalen Minimum gegeben. Es ist somit anzunehmen, dass der L^∞-Abstand unserer Referenzlösung zum globalen Minimum ungefähr 10^{-9} beträgt.

Für das weitere lokale Minimum wurde die Referenzlösung trotz gleicher Anzahl an Einzelschritten der SRKV offensichtlich weniger genau bestimmt, und alle Verfahren zeigen keine systematische Verbesserung unter einen Abstand von ungefähr $2 \cdot 10^{-6}$ zur Referenzlösung hinaus. In Abbildung 5.4(b) sieht man außerdem deutlich, dass das Verfahren Sarafyan 7:4 trotz geringerer Konvergenzordnung für die gleiche Anzahl an Schritten einen besseren Fehler als das Verfahren Sarafyan 7:5 liefert. Man kann allerdings auch erahnen, dass dies sich für kleinere Schrittweiten wieder umdrehen würde, hätte man eine genauerer Referenzlösung zur Verfügung. Da für die Konvergenz des Newton-Verfahrens die Schrittweite kleiner gewählt werden musste als bei der Konvergenz zu dem globalen Minimum, stehen hier nicht mehr genügend Daten zur Verfügung, um eine entsprechende Auswertung für das Verfahren Verner 21:8 darzustellen. Insgesamt sieht man in beiden Fällen sehr deutlich, dass man eine Konvergenz an sich und insbesondere eine hohe Konvergenzordnung praktisch nur für hinreichend kleine Schrittweiten erwarten kann.

5.3 Lernen stetiger Trajektorien

Als letztes Beispiel betrachten wir ein Problem, welches die Voraussetzungen für die Anwendung von Satz 3.3.3 nicht erfüllt. Dafür formulieren wir ein einfaches Problem mit Hilfe der künstlichen Neuronalen Netzwerke aus Abschnitt 1.2. Wir betrachten dabei ein GDNN aus zwei Knoten, welche mit Hilfe der Zustandsgleichung

$$\begin{aligned} \dot{x}_1(t) &= -x_1(t) + u_5(t)\, \arctan\big(u_1(t)x_1(t) + u_2(t)x_2(t)\big) \\ \dot{x}_2(t) &= -x_2(t) + u_5(t)\, \arctan\big(u_3(t)x_1(t) + u_4(t)x_2(t)\big) \end{aligned}$$

beschrieben werden. Wir möchten dieses Netzwerk darauf trainieren, dass die beiden Zustände für $t \in [0, 6.28]$ einen Kreis in der x_1-x_2-Ebene beschreiben. Dementsprechend wählen wir

$$\min \int_0^{6.28} \big((x_1(t) - \cos(t))^2 + (x_2(t) - \sin(t))^2\big)^{\frac{1}{2}}\ dt$$

als Zielfunktion. Offensichtlich muss man bei diesem Problem ohne einen externen „Input“ $I(t)$ mindestens einen Zustand zum Zeitpunkt 0 ungleich 0 wählen, daher wählen wir den Startpunkt auf dem zu beschreibenden Kreis durch $x_1(0) = 1$ und

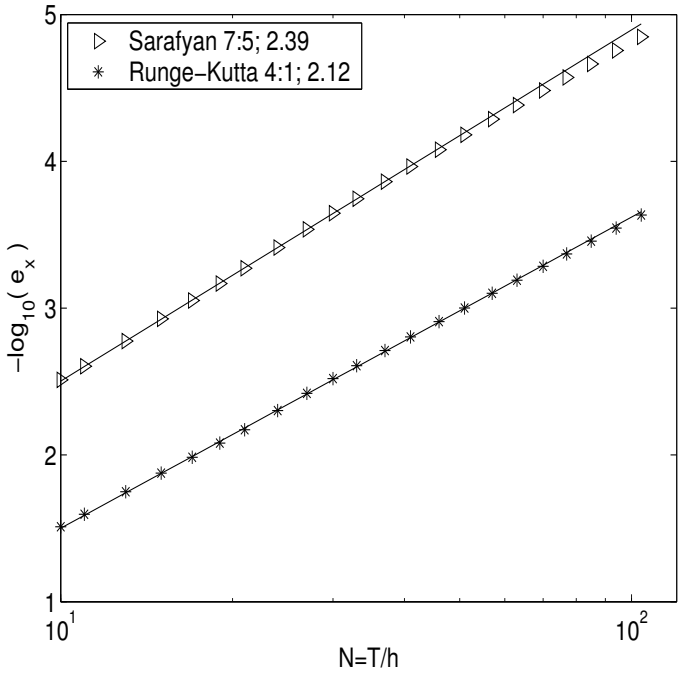
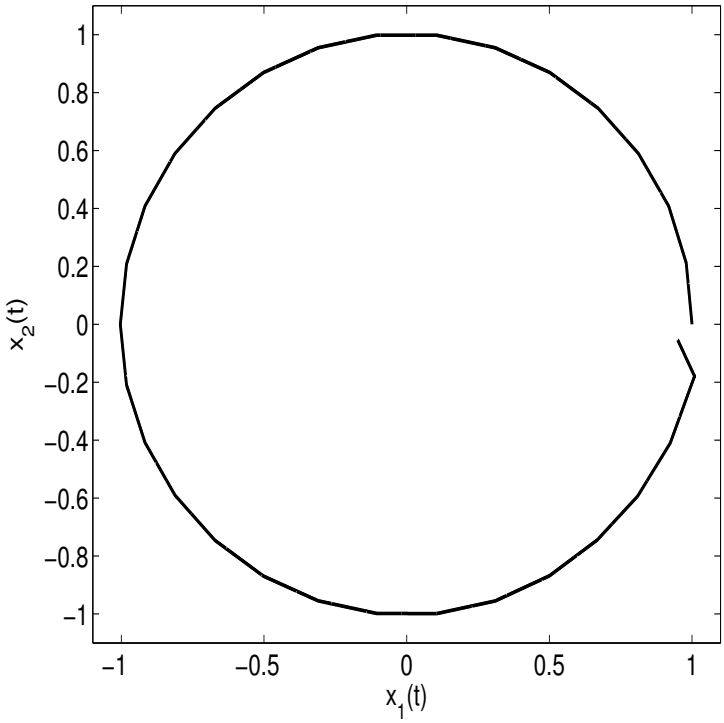


Abbildung 5.5: Wert der Zielfunktion für das Problem des Lernens stetiger Trajektorien in Abhängigkeit von der Schrittweite.

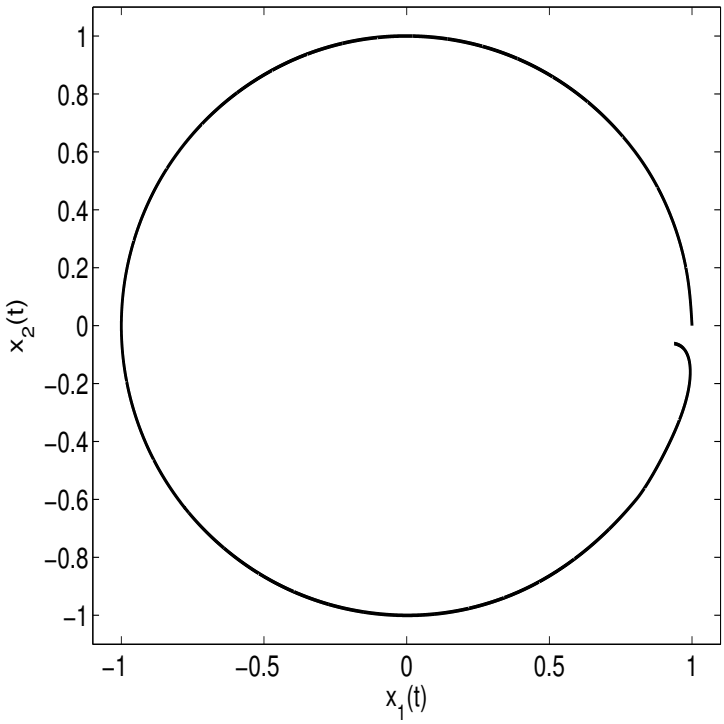
$x_2(0) = 0$. Dieses Problem stellt damit eine sehr einfache Variante einer Reihe von Standardproblemen dar, mit deren Hilfe man Trainingsalgorithmen und Eigenschaften Dynamischer Neuronaler Netzwerke untersucht (siehe unter anderem [Pea89, GLW99]).

Zu diesem Problem ist keine analytische Lösung bekannt. Allerdings kann man mit einem einfachen GDNN aus 2 Neuronen dem zu beschreibenden Kreis beliebig nah kommen (siehe z.B. [GLW99]). Bei den Berechnungen zu diesem Problem ist es mit keiner Schrittweite h gelungen, dass das Newton- oder das FBS-Verfahren zuverlässig für alle kleineren Schrittweiten zu einem globalen Minimum konvergiert. Selbst für sehr gute Anfangswerte für die Steuerung, mit denen der Kreis relativ gut beschrieben wird, konvergieren die beiden indirekten Verfahren nicht gegen ein lokales bzw. globales Minimum.

Allerdings konvergiert das näherungsweise Abstiegsverfahren aus Kapitel 4 dennoch für hinreichend kleine Schrittweiten h und endet nach endlich vielen Schritten. In Abbildung 5.5 kann man erkennen, dass man dabei trotzdem nicht die entsprechend hohe Konvergenzordnung der SRKV erhählt. Man sieht praktisch nur noch einen geringen Vorteil des Verfahrens Sarafyan 7:5 gegenüber dem klassischen Runge-Kutta-Verfahren mit linearer Interpolation bezüglich der Konvergenzordnung. Der Rechenaufwand, in Bezug auf die benötigte Rechenzeit, war dabei für ähnliche Werte der Zielfunktion immer in einer ähnlichen Größenordnung. In Abbildung 5.6 sieht man eine entsprechende Lösung zu dem klasssichen Runge-Kutta-Verfahren mit linearer Interpolation mit $N = 30$ Schritten und zu dem Verfahren Sarafyan 7:5 mit $N = 10$ Schritten. Die Zielfunktionswerte, die dabei erreicht wurden, sind ähnlich groß. Man kann erkennen, dass die Eigenschaften der Näherungslösungen

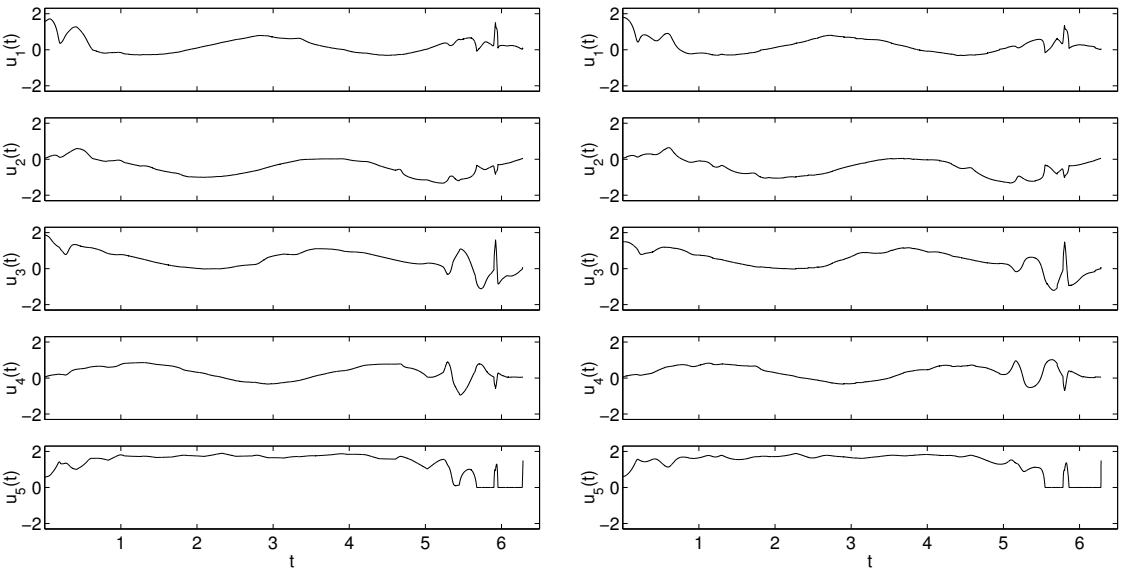


(a) klassisches Runge-Kutta-Verfahren mit linearer Interpolation und $N = 30$ Schritten



(b) Verfahren Sarafyan 7:5 mit $N = 10$ Schritten

Abbildung 5.6: Zustände zweier Näherungslösungen in der x_1-x_2-Ebene für das Problem des Lernens stetiger Trajektorien.



(a) klassisches Runge-Kutta-Verfahren mit linearer Interpolation und $N = 30$ Schritten

(b) Verfahren Sarafyan 7:5 mit $N = 10$ Schritten

Abbildung 5.7: Steuerungen zu den Näherungslösungen aus Abbildung 5.6.

für den Zustand, die sich aus den SRKV ergeben, natürlich erhalten bleiben und dementsprechend das klassische Runge-Kutta-Verfahren mit linearer Interpolation nur Zustände aus stückweise affine-linearen Funktionen liefert, wohingegen das Verfahren Sarafyan 7:5 mit weniger Schritten sichtbar glattere Zustände liefert.

Da man dabei allerdings keine Lösung des Gleichungssystems (DFC) gefunden hat, gelten diese Eigenschaften nicht mehr für Steuerungen, die man aus dem Minimumprinzip berechnet (siehe Abbildung 5.8). Die zugehörigen berechneten Steuerungen aus dem Minimumprinzip sieht man in Abbildung 5.7. Bei näherer Betrachtung fällt auf, dass diese Steuerungen deutliche Sprungstellen aufweisen. Somit können die aus dem Minimumprinzip berechneten Steuerungen von den zuvor mit Hilfe des Abstiegsverfahrens berechneten Steuerungen an den Diskretisierungspunkten abweichen. Dies ist, wenn man Rundungsfehler vernachlässigt, nicht möglich, wenn man eine Lösung des Gleichungssystems (DFC) berechnet hat, da hier für die Steuerungen an den Diskretisierungspunkten das Minimumprinzip bereits erfüllt ist.

Bei diesem Beispiel sieht man eindrucksvoll, dass man mit Hilfe der SRKV sehr zuverlässig stetige und sogar stetig-differenzierbare Näherungslösungen berechnen kann. Sind dabei die theoretischen Voraussetzungen von Satz 3.3.3 nicht erfüllt, ist im Allgemeinen nicht klar, ob es sinnvoll ist, Steuerungen an Punkten, welche keine Diskretisierungspunkte sind, über das Minimumprinzip zu berechnen.

Bei der Anwendung der GDNN in [GLW99, GLZW04, ZLM$^+$04] wird ein sehr ähnlicher Algorithmus verwendet, um die künstlichen Neuronalen Netzwerke zu trainieren. Dabei wird eine stetige Steuerung mit Hilfe einer fest vorgegebenen Anzahl von Basisfunktionen definiert. In einem iterativen Verfahren werden dann die

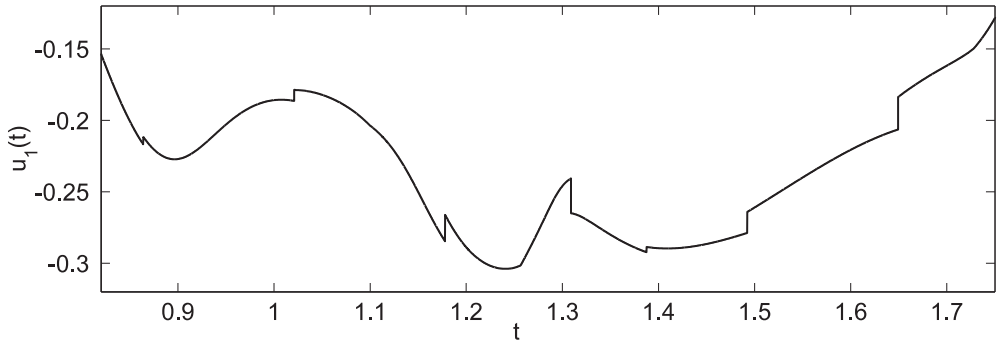


Abbildung 5.8: Ausschnitt aus der berechneten Steuerung u_1 zu dem Verfahren Sarafyan 7:5 mit $N = 10$ Schritten.

Koeffizienten dieser Steuerung geändert. Um die neuen Steuerungen zu berechnen, werden näherungsweise Gradienten dieser Koeffizienten berechnet. Dabei wird in der praktischen Anwendung ebenfalls zunächst die Zustandsgleichung vorwärts mit einem eingebetteten Runge-Kutta-Verfahren mit Schrittweitensteuerung gelöst. Danach wird die Gleichung für die adjungierten Zustände rückwärts mit dem gleichen Verfahren gelöst. Dabei benötigt man natürlich ebenfalls für die Berechnung der adjungierten Zustände die Zustände an Zwischenpunkten. Diese Zustände werden dann entweder als stückweise konstant angenommen oder linear interpoliert. In dem Fall dass man sie linear interpoliert, wird im Prinzip die einfachste Variante SRKV verwendet.

Da bei dem Algorithmus aus [GLW99] die Steuerungen über Koeffizenten und vorgegebene Basisfunktionen bestimmt werden, entsteht dabei in der Anwendung meist nicht die Notwendigkeit, die Steuerungen mit Hilfe anderer Basisfunktionen auszudrücken. Die Anzahl und Art der Basisfunktionen wird dabei zu einer wichtigen Eigenschaft des künstlichen Neuronalen Netzwerks, mit deren Hilfe man unter anderem auch Einfluss auf die Generalisierungsfähigkeit des GDNNs hat (siehe [GLZW04]).

Ein dazu ähnliches Vorgehen bietet sich mit Hilfe der SRKV an. Dabei gibt man sich statt der Basisfunktionen lediglich eine Anzahl an Gitterpunkten vor und speichert die Steuerung an den Diskretisierungspunkten, die zu einem SRKV gehören. Da es bei diesem Problem sinnvoll ist die Diskretisierung der Steuerung während einer Iteration nicht ständig zu verändern, kann man dabei keine eingebetteten Runge-Kutta-Verfahren mit Schrittweitensteuerung benutzen, obwohl solche ebenfalls mit SRKV möglich sind. Während des Trainings eines GDNNs kann man dabei natürlich sehr leicht Verfahren konstruieren, die im Laufe der Iterationen die Anzahl und die Lage der Gitterpunkte anpassen.

Ist das GDNN trainiert, könnte man das SRKV und die dazugehörigen Diskretisierungspunkte ebenfalls als Eigenschaft des künstlichen Neuronalen Netzwerks auffassen. Insgesamt sind damit sehr komplexe Algorithmen denkbar, bei denen man ebenfalls mit wenigen Parametern mit einer klaren Bedeutung Einfluss darauf nehmen kann, wie exakt das GDNN die Trainingsmuster lernen kann. Dies ist im Hinblick auf Überanpassung bzw. die Generalisierungsfähigkeit sehr wichtig.

Im Unterschied zu der oben beschriebenen Variante kann man das SRKV jedoch so wählen, dass man, wie in Abbildung 5.6 zu sehen ist, nur stetig differenzierbare Zustände und adjungierte Zustände erhält. Dadurch sind im Allgemeinen Unterschiede in der Generalisierungsfähigkeit der GDNNs zu erwarten.

6 Zusammenfassung und Ausblick

Im Kapitel 3 wurde gezeigt, dass unter den Voraussetzungen *Glattheit* und *Koerzivität* es zu hinreichend glatten optimalen Steuerungen eindeutig bestimmte stetige Näherungslösungen gibt, wenn man die auftretenden Differentialgleichungen für den Zustand und für die adjungierten Zustände mit SRKV numerisch löst. Es konnte ebenfalls gezeigt werden, dass man durch den Einsatz von SRKV sehr hohe Konvergenzordnungen bezüglich der Schrittweite der Diskretisierung erhalten kann. Dabei haben wir gezeigt, dass man für SRKV der Konvergenzordnung (p, p) ebenfalls die Konvergenzordnung p für die numerischen Lösungen des Steuerungsproblems erhält. Diese theoretischen Erkenntnisse wurden im Kapitel 5 von numerischen Ergebnissen zu Testproblemen bestätigt.

Als Grundlage der Ergebnisse diskretisierten wir die notwendigen Optimalitätsbedingungen erster Ordnung. Bisherige theoretische Erkenntnisse (siehe [Hag00, DHV98]) diskretisieren dagegen das Steuerungsproblem und erhalten mit einer ähnlichen Vorgehensweise, unter ähnlichen Voraussetzungen, Resultate für diskrete Runge-Kutta-Verfahren. Dabei kann man allerdings in [Hag00] nicht alle Runge-Kutta-Verfahren verwenden, die für entsprechende Anfangswertprobleme eine hohe Konvergenzordnung aufweisen. Für unsere Ergebnisse kann man dagegen alle existierenden SRKV benutzen. Dabei benötigen SRKV im Allgemeinen mehr Stufen für eine bestimmte Konvergenzordnung als entsprechende diskrete Runge-Kutta-Verfahren.

Die numerischen Resultate im Kapitel 5 zeigen, dass man bei den einfachsten Testproblemen mit hinreichend kleiner Schrittweite h auch eine Konvergenz der Ordnung p mit Hilfe SRKV der Ordnung $(p, p-1)$ erhalten kann. Dabei sieht man, dass dafür eventuell die Schrittweite h weiter verkleinert werden muss. Im Vergleich dazu wird im Beweis der Konvergenzordnung in [Hag00, DHV98] anstatt der Bedingung *Koerzivität* eine diskretisierte Koerzivitätsbedingung ausgenutzt, welche für hinreichend kleine Schrittweiten h gilt. Es ist daher nicht auszuschließen, dass man die Konvergenzordnung p für SRKV der Ordnung $(p, p-1)$ mit einem anderen Beweis zeigen kann, welcher ebenfalls eine vergleichbare Bedingung enthält.

In der Tabelle 5.1 sieht man unter anderem, dass die Verfahren aus Kapitel 4 bisher nur relativ ineffizient implementiert sind. Auf der Internetseite [Hag] gibt es das frei verfügbare Softwarepaket Optcon, welches ein Vorgehen nach dem Vorbild der Arbeit [Hag00] implementiert und in der Programmiersprache C geschrieben ist. Im Gegensatz dazu sind die vorgeschlagenen Algorithmen in erster Linie zur Demonstration der hohen Konvergenzordnung und der errechneten Steuerungen entstanden und in objektorientiertem C++ implementiert, ohne Rücksicht darauf zu nehmen, benötigten Speicher möglichst geschickt zu reservieren. Die derzeitige Implementierung ist lediglich darauf ausgelegt, relativ wenig Funktionsauswertungen der Zustandsgleichungen und adjungierten Gleichungen durchzuführen. Da in

den Beispielen in Kapitel 5 lediglich extrem einfache Zustandsgleichungen vorkommen, fällt schnell der hohe Speicherbedarf und der Aufwand der Additionen und Multiplikationen zur Auswertung des Runge-Kutta-Schemas ins Gewicht.

Sehr interessant wäre daher der Laufzeitvergleich einer effizienten Implementierung von weiterentwickelten Algorithmen auf der Basis des Satzes 3.3.3 mit dem Softwarepaket Optcon. Dabei sind sicherlich, neben komplett anderen Algorithmen, welche eine Nullstelle des Gleichungssystems (DFC) berechnen, auch bei den vorgeschlagenen Algorithmen aus Kapitel 4 an vielen Stellen noch deutliche Verbesserungen möglich. In dem Softwarepaket Optcon wird zum Beispiel statt eines einfachen Gradientenverfahrens ein Verfahren der konjugierten Gradienten benutzt. Durch den Trend zu immer leistungsfähigeren Prozessoren, die sich aus vielen einzelnen Recheneinheiten zusammensetzen und welche dadurch sehr viele Berechnungen parallel durchführen können, sind gerade Algorithmen, die dem Forward-Backward-Sweep-Verfahren aus [LW07] ähneln, für die Zukunft interessant, da diese sehr gut parallelisiert werden können. Weiterhin ist dabei interessant, inwieweit die verwendeten Runge-Kutta-Verfahren parallelisiert werden können. Einen ersten Überblick hierzu bietet die Arbeit [JN90]. In [BP94] beschäftigt man sich direkt mit der Entwicklung von parallelen expliziten SRKV.

Neben diesen Fragen entstehen Möglichkeiten, die man bei Verfahren auf der Basis der Arbeit [Hag00] nicht hat. Man kann eventuell, wie es beim numerischen Lösen von gewöhnlichen Differentialgleichungen üblich ist, sowohl für die Differentialgleichung der Zustände als auch für die adjungierten Zustände unabhängig voneinander Schrittweitensteuerungen benutzen. Um dies effizient implementieren zu können, müsste man allerdings das Minimumprinzip geschickt ausnutzen, um gegebenenfalls die Diskretisierungspunkte der Steuerung wechseln zu können.

Weiterhin sind Verfahren denkbar, die das Steuerungsproblem zunächst mit einer relativ groben Schrittweite lösen und diese Lösung als Start für ein Verfahren mit einer kleineren Schrittweite benutzen. Bei Verwendung von SRKV kann man dabei sicherstellen, dass das Minimumprinzip an den Diskretisierungspunkten, bis auf Rundungsfehler, erfüllt ist, und damit eventuell die Berechnung neuer initialer Steuerung an Zwischenpunkten erheblich verbessern. Es kann damit nicht abschließend geklärt werden, ob bei hinreichend glatten Steuerungsproblemen eine Implementation auf Basis von SRKV sogar Laufzeitenvorteile gegenüber einer Implementation auf Basis der transformierten adjungierten Gleichungen aus [Hag00] bieten kann. Dies liegt nicht zuletzt daran, dass die Effizienz eines Algorithmus stark von dem jeweiligen zu lösenden Steuerungsproblems und ebenfalls von der Rechnerarchitektur der verwendeten Computer abhängt. Bei den Algorithmen aus Kapitel 4 kann man zum Beispiel sehr einfach einen geringeren Speicherbedarf dadurch erreichen, dass man die Funktionen aus den Zustandsgleichungen gegebenenfalls mehrfach auswertet.

Die SRKV bieten allerdings schon jetzt einen deutlichen Vorteil beim Lösen von Steuerungsproblemen, wenn man die Näherungslösungen möglichst genau und an sehr vielen oder sogar an beliebigen Punkten benötigt. Man kann sich die berechneten Zustände und adjungierten Zustände so speichern, dass man sehr schnell an jedem Punkt eine gute Näherungslösung berechnen kann. Sind die Voraussetzungen dafür erfüllt, kann man daraus sehr leicht eine numerische Lösung für die Steue-

rung aus dem Minimumprinzip berechnen. Dabei muss man keine einzelnen Diskretisierungspunkte interpolieren, und daher braucht man sich auch keine Gedanken darüber zu machen, wie man dies geschickt realisiert. Wie man im Abschnitt 3.7 sieht, kann man dabei sogar sicherstellen, dass die Näherungslösungen stetig differenzierbar sind.

Aus theoretischer und praktischer Sicht wirft diese Arbeit ebenfalls die Frage auf, wie sich zusätzliche Zustands- oder Steuerungsbeschränkungen auf die vorgeschlagene Herangehensweise auswirken. In den Arbeiten [Hag00, DHV98] werden zumindest einfache Steuerungsbeschränkungen betrachtet. Eine zusätzliche theoretische oder praktische Betrachtung in dieser Arbeit hätte jedoch den Rahmen deutlich gesprengt.

Aus theoretischer Sicht ist außerdem interessant, ob für das Newton-Verfahren eine „Mesh-Independence“ Eigenschaft (siehe unter anderem [ABPR86, DHV00]) erfüllt ist. In den Ergebnissen aus Kapitel 5 deutet sich dies bei den einfachen linearquadratischen Testproblemen an. Hier benötigte das Newton-Verfahren jeweils nur 3 oder 4 Iterationen bei fast allen Berechnungen. Lediglich bei zwei Ausnahmen wurden 5 Iterationen benötigt, bis das Abbruchkriterium erfüllt war. Allerdings kann eine solche Systematik bei dem nichtlinearen Beispiel aus Abschnitt 5.2 numerisch nicht bestätigt werden.

In Abschnitt 5.3 wurde gezeigt, dass es interessant wäre, einen Trainingsalgorithmus für GDNNs aus Abschnitt 1.2 auf Basis SRKV zu implementieren und deren Generalisierungsfähigkeit zu untersuchen. Inwiefern man dabei deutlich besser generalisierende künstliche Neuronale Netzwerke erhalten kann, ist völlig offen und wird wiederum von dem konkreten Problem abhängen.

Literaturverzeichnis

[ABPR86] Allgower, E. L. ; Böhmer, K. ; Potra, F. A. ; Rheinboldt, W. C.: A mesh-independence principle for operator equations and their discretizations. In: *SIAM J. Numer. Anal.* 23 (1986), Nr. 1, S. 160–169

[Alt02] Alt, W.: *Nichtlineare Optimierung: eine Einführung in Theorie, Verfahren und Anwendungen*. Braunschweig : Vieweg, 2002 (Vieweg Studium: Aufbaukurs Mathematik). – ISBN 3–528–03193–X

[Ama95] Amann, H.: *Gewöhnliche Differentialgleichungen*. 2. Berlin : Walter de Gruyter, 1995 (De Gruyter Lehrbuch). – 499 S. – ISBN 3110145839

[AST97] Ali, M. M. ; Storey, C. ; Törn, A.: Application of stochastic global optimization algorithms to practical problems. In: *Journal of Optimization Theory and Applications* 95 (1997), Nr. 3, S. 545–563. – ISSN 0022–3239

[Bet01] Betts, J. T.: *Advances in Design and Control*. Bd. 3: *Practical methods for optimal control using nonlinear programming*. Philadelphia, PA : Society for Industrial and Applied Mathematics (SIAM), 2001. – 190 S. – ISBN 0–89871–488–5

[BP94] Baker, C. T. H. ; Paul, C. A. H.: A Global Convergence Theorem for a Class of Parallel Continuous Explicit Runge-Kutta Methods and Vanishing Lag Delay Differential Equations. In: *SIAM Journal on Numerical Analysis* 33 (1994), Nr. 4, S. 1559–1576

[But03] Butcher, J. C.: *Numerical methods for ordinary differential equations*. Wiley, 2003

[CT96] Corwin, S. P. ; Thompson, S.: Error estimation and step size control for delay differential equation solvers based on continuously embedded Runge-Kutta-Sarafyan methods. In: *Computers & Mathematics with Applications* 31 (1996), Nr. 6, S. 1 – 11. – ISSN 0898–1221

[DH98] Dontchev, A. L. ; Hager, W. W.: Lipschitzian Stability for State Constrained Nonlinear Optimal Control. In: *SIAM J. Control Optim.* 36 (1998), Nr. 2, S. 698–718. – ISSN 0363–0129

[DH01] Dontchev, A. L. ; Hager, W. W.: The Euler approximation in state constrained optimal control. In: *Math. Comput.* 70 (2001), Nr. 233, S. 173–203. – ISSN 0025–5718

[DHV98] Dontchev, A. L. ; Hager, W. W. ; Veliov, V. M.: Second-order Runge-Kutta approximations in constrained optimal control. In: *SIAM Journal of Numerical Analysis* (1998), Nr. 38, S. 202–226

[DHV00] Dontchev, A. L. ; Hager, W. W. ; Veliov, V. M.: Uniform Convergence and Mesh Independence of Newton's Method for Discretized Variational Problems. In: *SIAM J. Control Optim.* 39 (2000), Nr. 3, S. 961–980. – ISSN 0363–0129

[DMR06] Diele, F. ; Marangi, C. ; Ragni, S.: Coupling quadrature and continuous Runge-Kutta methods for optimal control problems. In: *Optimization Methods and Software* 21 (2006), S. 961–975(15)

[EJNT86] Enright, W. H. ; Jackson, K. R. ; Nørsett, S. P. ; Thomsen, P. G.: Interpolants for Runge-Kutta formulas. In: *ACM Trans. Math. Softw.* 12 (1986), Nr. 3, S. 193–218. – ISSN 0098–3500

[FPA+99] Floudas, C. A. ; Pardalos, P. M. ; Adjiman, C. ; Esposito, W. R. ; Gümüs, Z. H. ; Harding, S. T. ; Klepeis, J. L. ; Meyer, C. A. ; Schweiger, C. A.: *Handbook of Test Problems in Local and Global Optimization*. Kluwer Academic Publishers, 1999 (Nonconvex Optimization and Its Applications). – ISBN 978–0–7923–5801–5

[GLW99] Galicki, M. ; Leistritz, L. ; Witte, H.: Learning Continuous Trajectories in Recurrent Neural Networks with Time-Dependent Weights. In: *IEEE Transactions on Neural Networks* 10 (1999), Juli, Nr. 4, S. 741–755

[GLZW04] Galicki, M. ; Leistritz, L. ; Zwick, E. B. ; Witte, H.: Improving generalization capabilities of dynamic neural networks. In: *Neural Computation* 16 (2004), Nr. 6, S. 1253–1282

[Hag] Hager, W. W.: *William W. Hager*. `http://www.math.ufl.edu/~hager/`, Abruf: 08. Juli. 2009

[Hag76] Hager, W. W.: Rates of convergence for discrete approximations to unconstrained control problems. In: *SIAM J. Numer. Anal* (1976), Nr. 13, S. 449–472

[Hag00] Hager, W. W.: Runge-Kutta methods in optimal control and the transformed adjoint system. In: *Numerische Mathematik* 87 (2000), Nr. 2, S. 247–282

[Her04] Hermann, M.: *Numerik gewöhnlicher Differentialgleichungen*. München : Oldenbourg Wissenschaftsverlag, 2004. – ISBN 978–3–486–27606–0

[Hig91] Higham, D. J.: Highly continuous Runge-Kutta interpolants. In: *ACM Trans. Math. Softw.* 17 (1991), Nr. 3, S. 368–386. – ISSN 0098–3500

[HNW93] Hairer, E. ; Nørsett, S. P. ; Wanner, G.: *Solving Ordinary Differential Equations I. Nonstiff Problems*. Berlin : Springer-Verlag, 1993

[JN90] Jackson, K. R. ; Nørsett, S. P.: The Potential for Parallelism in Runge-Kutta Methods. Part 1: RK Formulas in Standard Form. In: *SIAM J. Numer. Anal* 32 (1990), S. 49–82

[LVV03] Lopez Cruz, I. L. ; Van Willigenburg, L. G. ; Van Straten, G.: Efficient Differential Evolution algorithms for multimodal optimal control problems. In: *Applied Soft Computing* 3 (2003), Nr. 2, S. 97 – 122. – ISSN 1568–4946

[LW07] Lenhart, S. ; Workman, J. T.: *Optimal Control Applied to Biological Models*. London : CRC Press, Taylor and Francis Group, 2007 (Chapman & Hall/CRC Mathematical and Computational Biology Series). – ISBN 1–58488–640–4; 978–1–58488–640–2

[OZ91] Owren, B. ; Zennaro, M.: Order Barriers for Continuous Explicit Runge-Kutta Methods. In: *Math. Comp* (1991), Nr. 56, S. 645–661

[OZ92] Owren, B. ; Zennaro, M.: Derivation of Efficient Continuous Explicit Runge-Kutta Methods. In: *SIAM J. Sci. Stat. Comput* (1992), Nr. 13, S. 1488–1501

[PBGM64] Pontryagin, L. C. ; Boltyanski, V. G. ; Gamkrelidze, R. V. ; Mischenko, E. F.: *The mathematical theory of optimal processes*. New York : MacMillan, 1964

[Pea89] Pearlmutter, B. A.: Learning State Space Trajectories in Recurrent Neural Networks. In: *Neural Computation* 1 (1989), S. 263–269

[PT97] Papakostas, S. N. ; Tsitouras, Ch.: Highly Continuous Interpolants for One-Step Ode Solvers and their Application to Runge-Kutta Methods. In: *SIAM Journal on Numerical Analysis* 34 (1997), Nr. 1, S. 22–47. – ISSN 00361429

[Roj96] Rojas, R.: *Neural networks : a systematic introduction*. Berlin : Springer-Verlag, 1996

[Sak81] Sakawa, Y.: On Local Convergence of an Algorithm for Optimal Control. In: *Numerical Functional Analysis and Optimization* 3 (1981), Nr. 3, S. 301–319. – ISSN 0163–0563

[SB80] Stoer, J. ; Bulirsch, R.: *Introduction to Numerical Analysis*. Berlin : Springer-Verlag, 1980. – ISBN 3–540–90420–4

[Sch87] Schmeisser, H.-J.: *Vector-valued Sobolev and Besov spaces*. Seminar Analysis of the Karl-Weierstraß-Institute, Berlin/GDR 1985/86, Teubner-Texte Math. 96, 4-44 (1987), 1987

[Sha85] Shampine, L. F.: Interpolation for Runge-Kutta methods. In: *SIAM J. Numer. Anal* (1985), Nr. 22, S. 1014–1027

[Ver] VERNER, J. H.: *Jim Verner's Refuge for Runge-Kutta Pairs*. `http://www.math.sfu.ca/~jverner/`, Abruf: 07. Juli. 2009

[Ver93] VERNER, J. H.: Differentiable Interpolants for High-Order Runge-Kutta Methods. In: *SIAM Journal on Numerical Analysis* 30 (1993), Nr. 5, S. 1446–1466. – ISSN 00361429

[VZ95a] VERNER, J. H. ; ZENNARO, M.: Continuous Explicit Runge-Kutta Methods of Order 5. In: *Mathematics of Computation* 64 (1995), Nr. 211, S. 1123–1146. – ISSN 00255718

[VZ95b] VERNER, J. H. ; ZENNARO, M.: The orders of embedded continuous explicit Runge-Kutta methods. In: *BIT Numerical Mathematics* 35 (1995), Nr. 3, S. 406–416. – ISSN 00361429

[Wer07] WERNER, D.: *Funktionalanalysis*. 6., korr. Aufl. Berlin : Springer, 2007 (Springer-Lehrbuch). – 531 S. – ISBN 978–3–540–72533–6; 3–540–72533–4

[Zei95] ZEIDLER, E.: *Applied Mathematical Sciences*. Bd. Applications to Mathematical Physics: *Applied Functional Analysis*. New York : Springer-Verlag, 1995. – ISBN 0–387–94442–7

[Zel94] ZELL, A.: *Simulation Neuronaler Netze*. Bonn : Addison-Wesley, 1994

[Zen86] ZENNARO, M.: Natural Continuous Extensions of Runge-Kutta Methods. In: *Mathematics of Computation* 46 (1986), Nr. 173, S. 119–133. – ISSN 00255718

[ZLM+04] ZWICK, E. B. ; LEISTRITZ, L. ; MILLEIT, B. ; SARAPH, V. ; ZWICK, G. ; GALICKI, M. ; WITTE, H. ; STEINWENDER, G.: Classification of equinus in ambulatory children with cerebral palsy - discrimination between dynamic tightness and fixed contracture. In: *Gait & Posture* 20 (2004), Nr. 3, S. 273 – 279. – ISSN 0966–6362

Ehrenwörtliche Erklärung

Hiermit erkläre ich,

- dass mir die Promotionsordnung der Fakultät bekannt ist,
- dass ich die Dissertation selbst angefertigt habe, keine Textabschnitte oder Ergebnisse eines Dritten oder eigenen Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönliche Mitteilungen und Quellen in meiner Arbeit angegeben habe,
- dass ich die Hilfe eines Promotionsberaters nicht in Anspruch genommen habe und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen,
- dass ich die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder wissenschaftliche Prüfung eingereicht habe.

Bei der Auswahl und Auswertung des Materials, sowie bei der Herstellung des Manuskripts haben mich folgende Personen unterstützt:

Prof. Dr. Walter Alt

Ich habe die gleiche, eine in wesentlichen Teilen ähnliche bzw. eine andere Abhandlung bereits bei einer anderen Hochschule als Dissertation eingereicht: Nein

Jena, den .....

Dirk Hemmelmann