Das Newton-Verfahren zur Lösung von Optimierungsproblemen mit parabolischen Differentialgleichungen

Diplomarbeit

vorgelegt von
Tina Anne Schütz
aus
Kassel

angefertigt im
Institut für Numerische und Angewandte Mathematik
der Georg-August-Universität Göttingen
2007

Inhaltsverzeichnis

Kapitel 1

Einleitung

Wir werden in dieser Arbeit das Newton-Verfahren zur Lösung von Problemen der optimalen Steuerung parabolischer Differentialgleichungen untersuchen.

Bezeichnen wir mit J das zu optimierende Steuerungsfunktional, so betrachten wir Probleme der Form:

$$\text{Minimiere } J(q,u) \text{ unter der Bedingung}$$

$$\left.\begin{aligned} \partial_t u(t;x) + L(x)u(t;x) &= f(q)(t;x) \\ u(0;x) &= u_0(q)(x). \end{aligned}\right\} \tag{1.1}$$

Hierbei bezeichnen q die *Steuerungsvariable*, u die *Zustandsvariable* und L einen elliptischen Operator bei vorgegebenen Daten $f(q)$ und $u_0(q)$. Die parabolische Differentialgleichung (1.1) lässt sich physikalisch beispielsweise als Wärmeleitprozess interpretieren, den man unter gewissen Gesichtspunkten (beispielsweise dem Erreichen einer bestimmten Temperatur zu einem bestimmten Zeitpunkt) steuern möchte. Da wir also einen Zustand beeinflussen wollen, bezeichnen wir Gleichung (1.1) im Folgenden auch als *Zustandsgleichung*.
Wir werden uns in dieser Arbeit auf linear-quadratische Optimierungsprobleme beschränken und sowohl verteilte Steuerungen als auch Steuerungen der Anfangsbedingung bei homogenen Dirichlet-Randbedingungen betrachten.

Ziel dieser Arbeit ist es, das Newton-Verfahren als eine Lösungsmöglichkeit solcher Probleme vorzustellen und näher zu untersuchen. Dafür werden wir ein reduziertes unrestringiertes Steuerungsfunktional j einführen. Um dessen Minimum zu bestimmen, betrachten wir die Optimalitätsbedingung 1. Ordnung

$$j'(q)(\tau q) = 0, \quad \forall\, \tau q \in Q,$$

wobei Q den Raum der Steuerungen bezeichnet. Diese Gleichung werden wir der Idee in [BMV07], [BR01] und [MV07] folgend mittels des Newton-Verfahrens lösen.

In diesem Zusammenhang ist die Lösung mehrerer parabolischer Differentialgleichungungen nötig, die jedoch nur approximativ mittels eines numerischen Verfahrens möglich ist. Wir werden hierbei eine diskontinuierliche Galerkin Methode in der Zeit mit der Methode der Finiten Elemente im Raum verknüpfen. Durch diese Approximation entsteht ein Diskretisierungsfehler, der sich von den Differentialgleichungen auf das Steuerungsfunktional überträgt. Deshalb werden wir weiterhin einen a posteriori Fehlerschätzer für das Steuerungsfunktional einführen und unter dessen Verwendung ein Adaptivitätskonzept für die zeitliche und räumliche Diskretisierung vorstellen.

In den Kapiteln 2 und 3 werden wir nun die Zustandsgleichung und das Optimierungsproblem exakt formulieren und Aussagen zu deren Lösbarkeit zusammenfassen, bevor wir in Kapitel 4 das Newton-Verfahren einführen. Dieses werden wir hierbei zunächst allgemein darstellen, bevor wir es auf das konkrete Problem übertragen.
Kapitel 5 behandelt zunächst die numerische Lösung der Zustandsgleichung, woraus dann ein diskretisiertes Optimierungsproblem hergeleitet wird. Daraufhin werden wir in Kapitel 6 einen a posteriori Fehlerschätzer für die Lösung dieses diskretisierten Optimierungsproblems herleiten und einen adaptiven Lösungsalgorithmus vorstellen. Abschließend werden wir in Kapitel 7 an zwei Beispielen die praktische Umsetzung der in dieser Arbeit hergeleiteten Algorithmen und deren Ergebnisse dokumentieren.

Kapitel 2

Die parabolische Zustandsgleichung

In diesem Kapitel wollen wir die parabolische Zustandsgleichung, die den Untersuchungen in dieser Arbeit zugrunde liegt, näher betrachten. Dazu werden wir zunächst einige grundlegende Definitionen und Aussagen zusammenfassen, mit deren Hilfe wir dann die schwache Formulierung des Problems aufstellen können. Zuletzt werden wir sehen, unter welchen Annahmen eine eindeutige Lösung der Zustandsgleichung existiert.

2.1 Grundlagen

Die Lösung der Zustandsgleichung ist eine Funktion, die von einer räumlichen Variablen $x \in \mathbb{R}^n$ und einer Zeitvariablen $t \in \mathbb{R}$ abhängt. Zur Behandlung dieser Funktionenklasse wählen wir den Zugang über abstrakte Funktionen. Für diese benötigen wir zunächst das Bochner-Integral, bevor wir verallgemeinerte Ableitungen definieren können. Diese werden wir für die schwache Formulierung der Zustandsgleichung benötigen.

Abstrakte Funktionen

Definition 2.1 Sei V ein Banachraum und $T > 0$. Dann bezeichnen wir eine Funktion $u : [0, T] \to V$ als *abstrakte Funktion*.

Den linearen Raum aller auf $[0, T]$ stetigen, abstrakten Funktionen bezeichnen wir mit $C([0, T]; V)$.

Definition 2.2 Eine abstrakte Funktion $u : [0,T] \to V$ heißt *differenzierbar in* $t \in [0,T]$, wenn es ein $u'(t) \in V$ gibt, so dass

$$\lim_{h \to 0^+} \left\| u'(t) - \frac{u(t+h) - u(t)}{h} \right\|_V = 0.$$

Wir bezeichnen $u'(t)$ als *klassische Ableitung* von u in t.

Induktiv definiert man höhere Ableitungen $u^{(m)}(t)$, $m \in \mathbb{N}$.

Den Raum aller abstrakten, m-mal stetig differenzierbaren Funktionen bezeichnen wir mit $\mathcal{C}^m([0,T];V)$.

Satz 2.3 *Versehen mit der Norm*

$$\|u\|_{\mathcal{C}^m([0,T];V)} := \max_{t \in [0,T]} \sum_{j=0}^{m} \|u^{(j)}(t)\|_V,$$

ist $\mathcal{C}^m([0,T];V)$ *ein Banachraum.*

Beweis: Ein Beweis hierzu findet sich in [GGZ74], Kapitel IV, Satz 1.1. □

Das Bochner-Integral

Nachdem wir bereits einen ersten Ableitungsbegriff für abstrakte Funktionen eingeführt haben, wollen wir nun das Bochner-Integral vorstellen.

Ähnlich wie beim Lebesgue-Integral definieren wir das Bochner-Integral zunächst für einfache Funktionen:

Definition 2.4 Eine abstrakte Funktion $u : [0,T] \to V$ heißt *einfach* oder *endlich*, wenn es höchstens endlich viele, paarweise disjunkte, Lebesgue-messbare Mengen $E_i \subset [0,T]$ für $i = 1, \ldots, m$ mit $m \in \mathbb{N} \setminus \{0\}$ von endlichem Lebesgue-Maß $\mu(E_i)$ gibt, so dass u auf jeder dieser Mengen einen konstanten Wert $u_i \in V$, $u_i \neq 0$ annimmt und sonst verschwindet.

Das *Bochner-Integral* einer einfachen Funktion definieren wir als

$$\int_0^T u(t)\, dt := \sum_{i=1}^{m} u_i \mu(E_i) \quad \in V.$$

Nun übertragen wir diesen Begriff mittels Grenzwertbildung auf allgemeine abstrakte Funktionen:

Definition 2.5 Eine abstrakte Funktion $u : [0,T] \to V$ heißt *Bochner-messbar*, falls es eine Folge $\{u_n\}$ von einfachen Funktionen gibt, so dass

$$u_n(t) \to u(t), \qquad \text{für fast alle } t \in [0,T],$$

im Sinne der Normkonvergenz.

Definition 2.6 Die abstrakte Funktion $u : [0,T] \to V$ sei Bochner-messbar und $\{u_n\}$ sei eine Folge einfacher Funktionen, die fast überall gegen u konvergiert. Dann heißt u *Bochner-integrierbar*, falls es zu jedem $\varepsilon > 0$ ein $n_0 \in \mathbb{N}$ gibt, so dass gilt

$$\int_0^T \|u_n(t) - u_m(t)\|_V \, dt < \varepsilon, \qquad \forall\, n, m \geq n_0.$$

Das *Bochner-Integral* über einer Lebesgue-messbaren Menge $B \subset [0,T]$ ist definiert als

$$\int_B u(t) \, dt := \lim_{n\to\infty} \int_0^T u_n(t)\chi_B(t) \, dt,$$

wobei χ_B die charakteristische Funktion der Menge B bezeichnet.

Wir beobachten an dieser Stelle, dass das Bochner-Integral selbst wieder Element des Banachraumes V ist.

Wir nutzen im Folgenden die Schreibweise $\int_a^b u(t) \, dt = \int_B u(t) \, dt$, falls gilt: $B = (a,b)$.

Lemma 2.7 *Eine Bochner-messbare Funktion $u : [0,T] \to V$ ist genau dann Bochner-integrierbar, wenn $t \mapsto \|u(t)\|_V$ Lebesgue- integrierbar ist.*

Beweis: Ein Beweis hierzu findet sich in [Emm04], Satz 7.1.15. □

Satz 2.8 *Ist $u \in \mathcal{C}([0,T];V)$, so ist u Bochner-integrierbar.*

Beweis: Ein Beweis hierzu findet sich in [Emm04], Satz 7.1.16. □

Wir fassen nun Funktionen, die fast überall gleich sind, zu einer Äquivalenzklasse zusammen und können so Räume Bochner-integrierbarer Funktionen definieren:

Definition 2.9 Wir fassen die Äquivalenzklassen Bochner-integrierbarer Funktionen $u : [0,t] \to V$ mit

$$\int_0^T \|u(t)\|_V^p \, dt < \infty, \qquad p \in [1,\infty),$$

zum Raum $L^p(0,T;V)$ zusammen.

Satz 2.10 *Versehen mit der Norm*

$$\|u\|_{L^p(0,T;V)} := \left(\int_0^T \|u(t)\|_V^p \, dt \right)^{1/p},$$

ist $L^p(0,T;V)$ für $1 \leq p < \infty$ ein Banachraum.

Beweis: Ein Beweis hierzu findet sich in [Lub03], Satz 8.5. □

Für die Räume $L^p(0,T;V)$ gelten die folgenden Einbettungseigenschaften:

Bemerkung 2.11 Es gilt

1. Für $1 \leq p < \infty$ ist $\mathcal{C}([0,T];V)$ dicht und stetig eingebettet in $L^p(0,T;V)$.
2. Für $1 \leq q \leq p < \infty$ ist $L^p(0,T;V)$ stetig in $L^q(0,T;V)$ eingebettet.

Wir wollen als nächstes den Dualraum zu $L^p(0,T;V)$ charakterisieren:

Satz 2.12 *Sei V ein reflexiver und separabler Banachraum, V^* sein Dualraum und $1 < p,q < \infty$ mit $\frac{1}{p} + \frac{1}{q} = 1$.*

Dann ist $L^q(0,T;V^*) = (L^p(0,T;V))^*$ der Dualraum zu $L^p(0,T;V)$. Das Dualitätsprodukt ist gegeben durch

$$\langle v,u \rangle_{L^q(0,T;V^*)\times L^p(0,T;V)} = \int_0^T \langle v(t),u(t) \rangle_{V^*\times V} \, dt.$$

Beweis: Ein Beweis hierzu findet sich in [GGZ74], Kapitel IV, Satz 1.14. □

Verallgemeinerte Ableitungen

Jetzt sind wir in der Lage, einen zweiten, allgemeineren Ableitungsbegriff einzuführen:

Definition 2.13 Seien V und W Banachräume, $u \in L^1(0,T;V), v \in L^1(0,T;W)$ und gelte für alle $\phi \in \mathcal{C}_0^\infty(0,T)$

$$\int_0^T \phi^{(n)}(t)u(t)\,dt = (-1)^n \int_0^T \phi(t)v(t)\,dt, \qquad n \in \mathbb{N},$$

dann heißt v *n-te verallgemeinerte Ableitung* von u auf $[0,T]$.

Wir schreiben $u^{(n)}(t) := v(t)$.

Wir erinnern an folgende

Definition 2.14 Sei V ein reeller, separabler und reflexiver Banachraum, V^* sein Dualraum und H ein reeller und separabler Hilbert-Raum. Ist $V \subseteq H$ eine stetige und dichte Einbettung, so nennen wir $V \subseteq H \subseteq V^*$ ein *Evolutionstripel*.

Bezüglich der Existenz und Eindeutigkeit der verallgemeinerten Ableitung wollen wir den folgenden Satz festhalten:

Satz 2.15 *Seien $V \subseteq H \subseteq V^*$ ein Evolutionstripel und $1 \leq p,q < \infty$ mit $\frac{1}{p} + \frac{1}{q} = 1$. Dann gilt*

1. *Für $u \in L^p(0,T;V)$ ist $u^{(n)} \in L^q(0,T;V^*)$ eindeutig bestimmt.*
2. *Für $u \in L^p(0,T;V)$ existiert $u^{(n)} \in L^q(0,T;V^*)$ genau dann, wenn eine Funktion $w \in L^q(0,T;V^*)$ existiert mit*

$$\int_0^T (u(t),v)_H\, \phi^{(n)}(t)\,dt = (-1)^n \int_0^T \langle w(t),v\rangle_{V^*\times V}\, \phi(t)\,dt$$

für alle $v \in V, \phi \in \mathcal{C}_0^\infty(0,T)$. Dann ist $u^{(n)} = w$ und

$$\frac{d^n}{dt^n}(u(t),v)_H = \langle u^{(n)}(t),v\rangle_{V^*\times V}$$

für alle $v \in V$ und für fast alle $t \in (0,T)$.

Beweis: Ein Beweis hierzu findet sich in [Lub03], Satz 8.20. □

Nun benötigen wir noch einen Raum, in dem wir Bochner-integrierbare Funktionen, die eine verallgemeinerte Ableitung besitzen, zusammenfassen:

Definition 2.16 Sei $V \subseteq H \subseteq V^*$ ein Evolutionstripel. Dann definieren wir:

$$W(0,T) := \{v \in L^2(0,T;V) : v' \in L^2(0,T;V^*)\}.$$

Satz 2.17 *Versehen mit der Norm*

$$\|u\|_{W(0,T)} := \left(\|u\|^2_{L^2(0,T;V)} + \|u'\|^2_{L^2(0,T;V^*)}\right)^{1/2}$$

ist $W(0,T)$ ein Banachraum.
Für beliebige $u, v \in W(0,T)$, $0 \leq s \leq t \leq T$ gilt die Regel der partiellen Integration*:*

$$\int_s^t \left(\langle u'(\tau), v(\tau)\rangle_{V^*\times V} + \langle u(\tau), v'(\tau)\rangle_{V^*\times V}\right)\, d\tau = (u(t), v(t))_H - (u(s), v(s))_H.$$

Außerdem ist $W(0,T)$ stetig in $\mathcal{C}([0,T];H)$ eingebettet und $\mathcal{C}^\infty([0,T];V)$ liegt dicht in $W(0,T)$.

Beweis: Ein Beweis hierzu findet sich in [Emm04], Satz 8.1.9. □

2.2 Starke und schwache Formulierung der Zustandsgleichung

Die bereits in der Einleitung vorgestellte Zustandsgleichung (1.1) wollen wir nun exakt beschreiben und die zugehörige schwache Formulierung aufstellen.

Hierbei unterdrücken wir der besseren Lesbarkeit zuliebe im Folgenden überall dort die Abhängigkeit von x und t, wo sie offensichtlich ist.

Wir betrachten das Problem:

Gesucht ist $u \in \mathcal{C}^1([0,T];\mathcal{C}^2(\Omega)) \cap \mathcal{C}([0,T];\mathcal{C}(\overline{\Omega}))$,
so dass zu gegebener Steuerung $q \in Q$ gilt:

$$\left.\begin{aligned} \partial_t u + Lu &= f(q) && \text{in } \Omega \times (0,T) \\ u &= 0 && \text{auf } \partial\Omega \times (0,T) \\ u(0) &= u_0(q) && \text{in } \Omega. \end{aligned}\right\} \tag{2.1}$$

Hierbei setzen wir voraus, dass $\Omega \subset \mathbb{R}^n$ ein Gebiet ist, $\partial\Omega$ dessen Rand und $I := (0, T)$ ein Zeitintervall.

Zunächst sei die rechte Seite $f(q) \in \mathcal{C}([0,T]; \mathcal{C}(\Omega))$ und der Anfangswert $u_0(q) \in \mathcal{C}(\Omega)$ vorgegeben.
Auf die konkrete Abhängigkeit von $q \in Q$ und den Raum Q werden wir weiter unten eingehen.

Wie schon in der Einleitung gefordert, sei außerdem L ein linearer, elliptischer Operator, d.h.:

$$(Lu)(x) = -div(A(x)\nabla u(t;x)) + b(x)\nabla u(t;x) + c(x)u(t;x)$$

mit $A(x) = (a_{ij}(x))_{i,j=1}^n$ symmetrisch und positiv definit für fast alle $x \in \Omega$ und $a_{ij} \in \mathcal{C}^1(\overline{\Omega})$ für $i, j = 1, \ldots, n$. Außerdem gelte für $b(x) = (b_i(x))_{i=1}^n$, dass die einzelnen Funktionen b_i ebenso wie c stetig auf $\overline{\Omega}$ seien, also $b_i, c \in \mathcal{C}(\overline{\Omega})$ für $i = 1, \ldots, n$.
Die Operatoren *div* und ∇ beziehen sich nur auf die räumlichen Koordinaten und wir nutzen für $u \in \mathcal{C}([0,T]; \mathcal{C}(\Omega))$ die Schreibweise $u(t)(x) = u(t;x)$.

Die Voraussetzungen, die wir in dieser klassischen Formulierung beispielsweise an die rechte Seite oder die Anfangsbedingung gestellt haben, sind sehr restriktiv. Unter der Verwendung von verallgemeinerten Ableitungen (sowohl räumlich, als auch zeitlich) können wir die schwache Formulierung des Problems (2.1) aufstellen und so eine Lösung unter abgeschwächten Annahmen bestimmen.

Zunächst benötigen wir jedoch die folgenden

Bezeichnungen 2.18 *Wir verwenden die Räume:*

- $V := \{v \in H^1(\Omega) : v|_{\partial\Omega} = 0\} = H_0^1(\Omega)$
- $H := L^2(\Omega)$
- $X := W(0,T)$ *sei der Hilbertraum der Ansatz- und Testfunktionen mit dem Skalarprodukt*

$$(u, v) := (u, v)_{L^2(I,H)} = \int_I (u(t), v(t))_H \, dt. \tag{2.2}$$

- Q *sei ein Hilbert-Raum und* $(\cdot,\cdot)_Q$ *ein Skalarprodukt auf* Q.

Mit der obigen Wahl von V und H ist die Definition von X sinnvoll, da $V \subseteq H \subseteq V^*$ nach [Hoh05], Satz 6.8, ein Evolutionstripel ist und somit die Voraussetzungen von Definition 2.16 erfüllt sind.

Indem wir (2.1) nun mit einer Testfunktion $\phi \in X$ multiplizieren, über Ω und $(0, T)$ integrieren, räumlich außerdem partielle Integration anwenden und X als den Raum für die Ansatz- und Testfunktionen wählen, erhalten wir die schwache Formulierung von (2.1):

Gesucht ist $u \in X$, so dass zu gegebener Steuerung $q \in Q$ gilt:

$$\left.\begin{aligned} (\partial_t u, \phi) + a(u, \phi) = (f(q), \phi), \quad \forall\, \phi \in X, \\ u(0) = u_0(q). \end{aligned}\right\} \tag{2.3}$$

Hierbei ist

$$a(u, \phi) = \int_I \int_\Omega (A \nabla u \cdot \nabla \phi + \nabla u \cdot b\phi + cu\phi)\, dx\, dt \tag{2.4}$$

mit $a_{ij}, b_i, c \in L^\infty(\Omega)$ für $i, j = 1, \ldots, n$, bilinear und stetig (vgl. z.B. [Lub06], Lemma 6.8).

Außerdem seien u_0 und f affin linear von q abhängig, dass heißt, es gelte

$$u_0(q)(x) = u_0^{(1)}(q)(x) + u_0^{(2)}(x) \tag{2.5}$$

und

$$f(q)(t; x) = f^{(1)}(q)(t; x) + f^{(2)}(t; x). \tag{2.6}$$

Hierbei seien $u_0^{(1)} : Q \to H$ und $f^{(1)} : Q \to L^2(I, V^*)$ lineare Operatoren und $u_0^{(2)} \in H$ sowie $f^{(2)} \in L^2(I, V^*)$.
Diese Wahl von $f(q)$ und $u_0(q)$ bietet uns die Freiheit, die rechte Seite oder die Anfangsbedingung auch unabhängig von q zu wählen.

Bemerkung 2.19 Wir beobachten, dass offensichtlich jede Lösung des ursprünglichen Anfangs-Randwertproblems (2.1) auch die schwache Formulierung (2.3) erfüllt und für glatte Daten und Lösungen die beiden Formulierungen äquivalent sind.

2.3 Lösbarkeit der Zustandsgleichung

Wir werden nun zeigen, dass das Variationsproblem (2.3) eine eindeutige Lösung besitzt. Dazu führen wir die Notation

$$\int_I \widetilde{a}(u(t), v(t))\, dt := a(u, v) \tag{2.7}$$

mit $\widetilde{a} : V \times V \to \mathbb{R}$ ein und erhalten

Satz 2.20 *Unter den Voraussetzungen:*

- $0 < T < \infty$
- $\widetilde{a} : V \times V \to \mathbb{R}$ *ist eine stetige und strikt positive Bilinearform, d.h.*
  - *es existiert eine Konstante $C > 0$, so dass für alle $u, v \in V$ gilt:*

$$|\widetilde{a}(u, v)| \leq C\|u\|_V \|v\|_V$$

  - *es existiert eine Konstante $\gamma > 0$, so dass für alle $v \in V$ gilt:*

$$\widetilde{a}(v, v) \geq \gamma \|v\|_V^2$$

- *f und u_0 sind wie in* (2.5) *und* (2.6) *gegeben,*

existiert genau eine Lösung $u \in X$ der schwachen Formulierung (2.3)*:*

$$\begin{aligned} (\partial_t u, \phi) + a(u, \phi) &= (f(q), \phi), \quad \forall\, \phi \in X, \\ u(0) &= u_0(q). \end{aligned}$$

Beweis: Nach Theorem 9.5 und Satz 10.1 in [Lub03] existiert unter den obigen Voraussetzungen genau eine Lösung $u \in X$ von

$$\begin{aligned} \langle u'(t), v\rangle_{V^* \times V} + \widetilde{a}(u(t), v) &= \langle f(q)(t), v\rangle_{V^* \times V}, \quad \forall\, v \in V, \quad \forall\, t \in (0, T), \\ u(0) &= u_0(q). \end{aligned}$$

Setzen wir hier $\phi(t) := v$ mit $\phi \in X$ und integrieren zusätzlich über $(0, T)$, so erhalten wir die schwache Formulierung (2.3) und die eindeutige Lösbarkeit überträgt sich. □

Theorem 9.5 in [Lub03] liefert uns außerdem die stetige Abhängigkeit der Lösung u von den gegebenen Daten $f(q)$ und $u_0(q)$. Deshalb können wir schreiben

$$u = G\left(f(q), u_0(q)\right),$$

wobei $G : L^2(I, V^*) \times H \to X$ ein stetiger, linearer Operator ist.
Setzen wir für $f(q)$ und $u_0(q)$ die Darstellungen (2.6) bzw. (2.5) ein, so ergibt sich aus der Linearität

$$\begin{aligned} u &= G\left(f(q), u_0(q)\right) \\ &= G\left(f^{(1)}(q) + f^{(2)}, u_0^{(1)}(q) + u_0^{(2)}\right) \\ &= G\left(f^{(1)}(q), u_0^{(1)}(q)\right) + G\left(f^{(2)}, u_0^{(2)}\right). \end{aligned}$$

Nutzen wir für feste $f^{(1)}, f^{(2)}, u_0^{(1)}$ und $u_0^{(2)}$ die Bezeichnungen

$$S^{(1)}(q) := G\left(f^{(1)}(q), u_0^{(1)}(q)\right) \qquad \text{und} \qquad S^{(2)} := G\left(f^{(2)}, u_0^{(2)}\right),$$

mit $S^{(1)} : Q \to X$ und $S^{(2)} \in X$, so können wir folgende Definition festhalten, die wir im weiteren Verlauf der Arbeit noch benötigen werden:

Definition 2.21 Mit $S : Q \to X$, $S(q) := S^{(1)}(q) + S^{(2)}$ bezeichnen wir den Operator, der zu $q \in Q$ die Lösung $u = S(q) \in X$ der schwachen Formulierung (2.3) liefert.

Dieser Lösungsoperator ist unter den Voraussetzungen von Satz 2.20 wohldefiniert und wegen der Linearität von $S^{(1)}$ affin linear.

Bemerkung 2.22 Um in den folgenden Betrachtungen die eindeutige Lösbarkeit der Zustandsgleichung annehmen zu können, werden wir ab jetzt davon ausgehen, dass die Voraussetzungen von Satz 2.20 erfüllt sind.

Kapitel 3

Das Optimierungsproblem

In diesem Kapitel wollen wir uns dem Optimierungsproblem widmen. Dafür werden wir zunächst einige Grundlagen zusammenfassen, bevor wir das eigentliche Problem formulieren. Zuletzt stellen wir Aussagen zur Lösbarkeit und Optimalitätsbedingungen solcher Probleme zusammen.

3.1 Grundlagen

Wir wollen in diesem Abschnitt zunächst auf die Differenzierbarkeit in Banachräumen eingehen, bevor wir Definitionen und Aussagen zu Konvexität und schwacher Konvergenz wiederholen.

3.1.1 Differenzierbarkeit in Banachräumen

Da das Steuerungsfunktional, das wir im Folgenden betrachten werden, nicht unbedingt über dem $\mathbb{R}^n$ definiert ist, sondern allgemein über einem Banachraum, müssen wir den bekannten Ableitungsbegriff verallgemeinern. Diesen benötigen wir zur Herleitung und Formulierung notwendiger Optimalitätsbedingungen.

In diesem Abschnitt seien V und W reelle Banachräume sowie $F : V \to W$ eine Abbildung von V nach W.

Definition 3.1 Existiert zu gegebenen $v, h \in V$ der Grenzwert

$$\delta F(v,h) := \lim_{t \to 0^+} \frac{1}{t} \left(F(v+th) - F(v) \right)$$

in W, so heißt dieser *Richtungsableitung* von F an der Stelle v in Richtung h. Existiert der Grenzwert für alle $h \in V$, dann heißt die Abbildung $h \mapsto \delta F(v,h)$ *erste Variation* von F an der Stelle v.

Definition 3.2 Existieren die erste Variation $\delta F(v,h)$ an der Stelle v und ein linearer, stetiger Operator $A : V \to W$, so dass

$$\delta F(v,h) = Ah$$

für alle h aus V gilt, dann heißt A *Gâteaux-Ableitung* von F an der Stelle v. Wir schreiben $A := F'_G(v)$.

Bemerkung 3.3

- Aus der Definition folgt, dass man Gâteaux-Ableitungen als Richtungsableitungen berechnen kann.
- Ist $f : V \to \mathbb{R}$ ein Gâteaux-differenzierbares Funktional, dann ist seine Ableitung ein Element des zu V dualen Raumes V^*.

Ein weiterer Differentiationsbegriff, der eine Spezialisierung der Gâteaux-Ableitungen ist, ermöglicht es uns, einige zusätzliche Aussagen treffen zu können:

Definition 3.4 Eine Abbildung $F : V \to W$ heißt an der Stelle v *Fréchet-differenzierbar*, wenn ein Operator $A \in \mathcal{L}(V,W)$ und eine Abbildung $r : V \times V \to W$ mit den folgenden Eigenschaften existieren:
Für alle $h \in V$ gilt

$$F(v+h) = F(v) + Ah + r(v,h)$$

und das *Restglied* r genügt der Beziehung

$$\frac{\|r(v,h)\|_W}{\|h\|_V} \to 0 \qquad \text{für } \|h\|_V \to 0.$$

A heißt *Fréchet-Ableitung* von F an der Stelle v. Wir schreiben $A := F'(v)$.

Bemerkung 3.5 Eine alternative Bedingung für Fréchet-Differenzierbarkeit ist

$$\frac{\|F(v+h) - F(v) - Ah\|_W}{\|h\|_V} \to 0 \qquad \text{für } \|h\|_V \to 0.$$

Diese Bedingung ist offensichtlich äquivalent zu jener aus Definition 3.4, da $F(v+h) - F(v) - Ah = r(v,h)$ und damit $\frac{\|r(v,h)\|_W}{\|h\|_V} \to 0$ für $\|h\|_V \to 0$.

Für lineare, stetige Operatoren wissen wir, wie die Fréchet-Ableitung aussieht:

Lemma 3.6 *Jeder lineare stetige Operator A ist Fréchet-differenzierbar und es gilt*

$$A'(v)h = Ah.$$

Beweis: Wegen der Linearität gilt

$$A(v+h) = A(v) + A(h) + 0.$$

Für das Restglied r gilt also $r(v,h) = 0$. □

Korollar 3.7 *Die Bilinearform a*

$$a(u,\phi) = \int_I \int_\Omega (A\nabla u \cdot \nabla\phi + \nabla u \cdot b\phi + cu\phi)\, dx\, dt$$

ist in beiden Argumenten Fréchet-differenzierbar.

Beweis: Sei ϕ fest gewählt. Dann ist $a(\cdot,\phi)$ ein linearer, stetiger Operator und damit nach Lemma 3.6 Fréchet-differenzierbar. Analog ist $a(u,\cdot)$ Fréchet-differenzierbar. □

Lemma 3.8 *Jede Fréchet-differenzierbare Abbildung F ist auch Gâteaux-differenzierbar und es gilt*

$$F'_G(v) = F'(v).$$

Beweis: Ein Beweis hierzu findet sich in [Lue69], Kapitel 7.2, Proposition 2. □

Außerdem gilt für Fréchet-Ableitungen die Kettenregel:

Lemma 3.9 *Sind U, V und W Banachräume sowie $F : U \to V$ und $G : V \to W$ an den Stellen $u \in U$ bzw. $F(u) \in V$ Fréchet-differenzierbare Abbildungen, dann ist auch $H : U \to W$ definiert durch*

$$H(u) := G\left(F(u)\right)$$

Fréchet-differenzierbar an der Stelle u und es gilt

$$H'(u) = G'\left(F(u)\right) \cdot F'(u)$$

Beweis: Ein Beweis hierzu findet sich in [Lue69], Kapitel 7.3, Proposition 1. □

3.1.2 Konvexität und schwache Konvergenz

Nun wollen wir zwei Definitionen wiederholen, die wir zur Formulierung der Optimalitätsbedingung und der Existenzaussagen benötigen. Abschließend werden wir einige Begriffe und Aussagen zusammenfassen, die im Beweis des Existenzsatzes 3.22 verwendet werden.

Definition 3.10 Sei M eine Teilmenge eines reellen Vektorraumes. Wenn zu zwei beliebigen Punkten $a, b \in M$ auch deren Verbindungsstrecke ganz in M liegt, d.h. falls für alle $0 \leq \lambda \leq 1$ gilt

$$\lambda a + (1 - \lambda b) \in M,$$

so heißt M *konvex*.

Definition 3.11 Sei M eine konvexe Teilmenge eines reellen Vektorraumes und $f : M \to \mathbb{R}$ eine reellwertige Funktion. Wir nennen f *konvex*, falls für alle $x, y \in M$ und $0 \leq \lambda \leq 1$ gilt:

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

Gilt unter denselben Voraussetzungen sogar

$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y),$$

so bezeichnen wir f als *streng konvex*.

Nun kommen wir zum Begriff der schwachen Konvergenz:

Definition 3.12 Sei V ein reeller Banachraum und $(v_n)_{n\in\mathbb{N}} \subset V$ eine Folge in V. Existiert ein $v \in V$, so dass für alle Funktionale $f \in V^*$ gilt

$$f(v_n) \to f(v), \qquad \text{für } n \to \infty,$$

so heißt (v_n) *schwach konvergent*.
Wir schreiben $v_n \rightharpoonup v$ für $n \to \infty$.

Lemma 3.13 *Das Grenzelement $v \in V$ einer schwach konvergenten Folge ist eindeutig bestimmt.*

Beweis: Ein Beweis hierzu findet sich in [LS68], S. 148. □

Definition 3.14 Sei V ein reeller Banachraum und M eine Teilmenge von V. Gilt für jede schwach konvergente Folge $(v_n)_{n\in\mathbb{N}} \subset M$ mit $v_n \rightharpoonup v$ für $n \to \infty$, dass auch das Grenzelement v in M liegt, so heißt M *schwach folgenabgeschlossen*. Besitzt jede Folge $(v_n)_{n\in\mathbb{N}} \subset M$ eine in V schwach konvergente Teilfolge, so nennen wir M *relativ schwach folgenkompakt*.

Eine Aussage darüber, welche Mengen schwach folgenabgeschlossen oder relativ schwach folgenkompakt sind, treffen die beiden folgenden Sätze:

Satz 3.15 *Jede konvexe und abgeschlossene Teilmenge eines Banachraumes ist schwach folgenabgeschlossen.*

Beweis: Ein Beweis hierzu findet sich in [Alt92], Satz 5.10. □

Satz 3.16 *Jede beschränkte Teilmenge eines reflexiven Banachraumes ist relativ schwach folgenkompakt.*

Beweis: Wir wissen aus Satz 60.6 in [Heu86], dass jede beschränkte Folge in einem reflexiven Banachraum eine schwach konvergente Teilfolge besitzt.

Ist nun M eine beschränkte Teilmenge eines reflexiven Banachraumes V und $(v_n)_{n\in\mathbb{N}}$ eine Folge in M, so ist $(v_n)_{n\in\mathbb{N}}$ eine beschränkte Folge in V und der oben zitierte Satz ist anwendbar. Es existiert also eine schwach konvergente Teilfolge $(v_{n_k})_{k\in\mathbb{N}} \subset M$ von $(v_n)_{n\in\mathbb{N}}$ und damit ist M relativ schwach folgenabgeschlossen. □

Zuletzt wollen wir einen Zusammenhang von schwach konvergenten Folgen und Funktionalen festhalten:

Definition 3.17 Sei V ein Banachraum, $(v_n)_{n\in\mathbb{N}} \in V$ eine beliebige schwach konvergente Teilfolge und $f \in V^*$ ein Funktional über V. Folgt aus $v_n \rightharpoonup v$ für $n \to \infty$, dass

$$\lim_{n\to\infty} \inf f(v_n) \geq f(v)$$

gilt, so heißt f *schwach unterhalbstetig*.

Satz 3.18 *Jedes in einem Banachraum V konvexe und stetige Funktional f ist schwach unterhalbstetig.*

Beweis: Ein Beweis hierzu findet sich in [Zei90], Proposition 25.20. □

3.2 Das Optimierungsproblem und seine Lösbarkeit

In diesem Abschnitt wollen wir zunächst das Steuerungsfunktional definieren, bevor wir dass zentrale Optimierungsproblem formulieren. Dieses werden wir dann in ein unrestringiertes Optimierungsproblem überführen, für das wir festhalten werden, unter welchen Bedingungen es eindeutig lösbar ist.

Definition 3.19 Das *Steuerungsfunktional* $J : Q \times X \longrightarrow \mathbb{R}$ definieren wir als:

$$J(q,u) := \int_I I_1(u(t))\,dt + I_2(u(T)) + \frac{\alpha}{2}\|q-\overline{q}\|_Q^2 \tag{3.1}$$

mit

- dem Regularisierungsparameter $\alpha \geq 0$,
- der Referenzsteuerung $\overline{q} \in Q$ und
- $I_1 : V \longrightarrow \mathbb{R}$, $I_2 : H \longrightarrow \mathbb{R}$ definiert durch

$$I_1(u(t)) := \frac{1}{2}b_1(u(t),u(t)) + L_1(u(t)) + C_1$$

und

$$I_2(u(T)) := \frac{1}{2}b_2(u(T),u(T)) + L_2(u(T)) + C_2,$$

wobei b_1, b_2 stetig Fréchet-differenzierbare, symmetrische Bilinearformen, L_1, L_2 stetig Fréchet-differenzierbare Linearformen und $C_1, C_2 \in \mathbb{R}$ seien.

Diese Form ermöglicht es uns, beispielsweise Funktionale der Form

$$J(q,u) := \int_I \|u(t)-\overline{u}(t)\|^2\,dt + \frac{\alpha}{2}\|q-\overline{q}\|_Q^2$$

mit einer Referenzlösung $\overline{u} \in X$ zu betrachten.

Mit diesen Bezeichnungen erhalten wir nun das

Optimierungsproblem:

$$\text{Minimiere } J(q,u) \text{ unter der Bedingung (2.3) mit } (q,u) \in Q \times X. \tag{3.2}$$

Definition 3.20 Sei $S : Q \longrightarrow X$ der Lösungsoperator aus Definition 2.21. Dann definiert

$$\begin{aligned} j : Q &\longrightarrow \mathbb{R} \\ q &\longmapsto j(q) = J(q, S(q)) \end{aligned}$$

das sogenannte *reduzierte Steuerungsfunktional*.

Bemerkung 3.21 Das reduzierte Steuerungsfunktional j ist ein stetiges Funktional, da der Lösungsoperator S, die Operatoren I_1 und I_2 stetig sind und auch die Norm $\|\cdot\|_Q$ eine stetige Abbildung ist.

Mit diesem reduzierten Steuerungsfunktional betrachten wir das Minimierungsproblem (3.2) als

unrestringiertes Optimierungsproblem:

$$\text{Minimiere } j(q) \text{ mit } q \in Q. \tag{3.3}$$

Zur Lösbarkeit dieses Optimierungsproblems halten wir zunächst folgenden Satz fest:

Satz 3.22 *Seien Q ein Hilbertraum, $Q_{ad} \subset Q$ eine nichtleere, beschränkte, abgeschlossene und konvexe Teilmenge und j konvex. Dann besitzt das restringierte Optimierungsproblem*

$$\textit{Minimiere } j(q) \textit{ mit } q \in Q_{ad} \tag{3.4}$$

eine optimale Lösung $\widehat{q} \in Q_{ad}$.

Ist j streng konvex, so ist diese Lösung auch eindeutig bestimmt.

Beweis: Da das Funktional j stetig ist, wird die beschränkte Menge Q_{ad} auf eine wiederum beschränkte Menge $j(Q_{ad})$ abgebildet. Es existiert somit das Infimum von j über Q_{ad}, das wir mit $\widetilde{j}$ bezeichnen werden, also

$$\widetilde{j} := \inf_{q \in Q_{ad}} j(q).$$

Außerdem sei $(q_n)_{n\in\mathbb{N}}$ eine Folge aus Q_{ad} mit

$$j(q_n) \to \widetilde{j} \qquad \text{für } n \to \infty.$$

Nach Voraussetzung ist Q_{ad} beschränkt und damit nach Satz 3.16 relativ schwach folgenkompakt, d.h. es existiert eine in Q schwach konvergente Teilfolge $(q_{n_k})_{k\in\mathbb{N}}$, so dass

$$q_{n_k} \rightharpoonup \widehat{q} \qquad \text{für } k \to \infty \tag{3.5}$$

mit $\widehat{q} \in Q$.

Da Q_{ad} außerdem als abgeschlossen und konvex vorausgesetzt wird, wissen wir aus Satz 3.15, dass Q_{ad} auch schwach folgenabgeschlossen ist. Das Grenzelement $\widehat{q}$ aus (3.5) liegt also in Q_{ad}.

Aus Satz 3.18 wissen wir wegen der Konvexität und Stetigkeit von j, dass j schwach unterhalbstetig ist, dass also gilt:

$$j(\widehat{q}) \leq \liminf_{k\to\infty} j(q_{n_k}) = \widetilde{j}.$$

Da $\widehat{q}$ in Q_{ad} liegt und $j(\widehat{q})$ nicht kleiner als das Infimum $\widetilde{j}$ aller Funktionswerte über Q_{ad} sein kann, muss

$$j(\widehat{q}) = \widetilde{j}$$

gelten. Es existiert also eine optimale Steuerung $\widehat{q}$ von j.

Ist j sogar streng konvex, so ist diese optimale Steuerung eindeutig. Denn nehmen wir an, dass es zwei verschiedene optimale Steuerungen $\widehat{q}, \widehat{p} \in Q_{ad}$ mit $j(\widehat{q}) = j(\widehat{p}) = \widetilde{j}$ gibt, so folgt aus der strengen Konvexität von j mit $\lambda = \frac{1}{2}$:

$$\begin{aligned} j\left(\frac{1}{2}\widehat{q} + \frac{1}{2}\widehat{p}\right) &< \frac{1}{2}j(\widehat{q}) + \frac{1}{2}j(\widehat{p}) \\ &= \widetilde{j}. \end{aligned}$$

Dies ist aber ein Widerspruch dazu, dass $\widetilde{j}$ das Infimum von j über Q_{ad} ist. Also kann es nicht zwei optimale Steuerungen geben. □

Diese Aussage wollen wir nun auf unrestringierte Optimierungsprobleme verallgemeinern, indem wir zusätzliche Annahmen machen.

Satz 3.23 *Ist der Regularisierungsparameter $\alpha > 0$ und gilt für alle $q \in Q_{ad}$*

$$\int_I I_1(S(q)(t))\,dt + I_2(S(q)(T)) \geq 0,$$

so existiert auch dann eine Lösung $\widehat{q}$ des Problems (3.4)*, wenn wir Q_{ad} nur als konvex und abgeschlossen voraussetzen.*

Wie in Satz 3.22 ist diese Lösung eindeutig, falls j streng konvex ist.

Beweis: Wir bezeichnen erneut mit $\widetilde{j}$ das Infimum des Funktionals j über der Menge Q_{ad}, also

$$\widetilde{j} := \inf_{q \in Q_{ad}} j(q).$$

Für alle $q \in Q_{ad}$ mit $\|q - \overline{q}\|_Q^2 \geq \frac{2}{\alpha}(\widetilde{j} + 1)$ gilt:

$$\begin{aligned} j(q) &= \int_I I_1(S(q)(t))\, dt + I_2(S(q)(T)) + \frac{\alpha}{2}\|q - \overline{q}\|_Q^2 \\ &\geq \frac{\alpha}{2}\|q - \overline{q}\|_Q^2 \\ &\geq \widetilde{j} + 1 \\ &> \inf_{q \in Q_{ad}} j(q). \end{aligned}$$

Diese Steuerungen können also nicht Lösungen des Problems sein, so dass wir uns bei der Suche nach dem Optimum auf die Menge

$$\{q \in Q_{ad} : \ \|q - \overline{q}\|_Q^2 < \frac{2}{\alpha}\}$$

beschränken können. Diese Menge ist abgeschlossen, konvex und beschränkt, so dass die Aussage nun aus Satz 3.22 folgt. □

Korollar 3.24 *Unter den Voraussetzungen von Satz 3.23 an das reduzierte Steuerungsfunktional j, besitzt das unrestringierte Optimierungsproblem* (3.3) *eine eindeutige Lösung.*

Beweis: Setzt man in Satz 3.23 $Q_{ad} = Q$, so folgt die Aussage sofort. □

3.3 Optimalitätsbedingungen

Im folgenden Abschnitt wollen wir notwendige und hinreichende Bedingungen für die Optimalität einer Lösung herleiten. Die notwendige Optimalitätsbedingung 1. Ordnung wird dann im weiteren Verlauf dieser Arbeit eine grundlegende Rolle spielen.

Satz 3.25 *Sei X ein Banachraum und $f : X \to \mathbb{R}$ Gâteaux-differenzierbar auf X. Ferner habe f in $x_0 \in X$ ein* lokales Minimum, *d.h es gibt eine Umgebung $U \subset X$ von x_0, so dass $f(x_0) \leq f(x)$ für alle $x \in U$. Dann gilt für alle $h \in X$*

$$f'(x_0)(h) = 0.$$

Beweis: Sei x_0 ein lokales Minimum von f auf X.

Definieren wir für ein beliebiges, aber festes $h \in X$ die reelle Funktion $g_h : \mathbb{R} \to \mathbb{R}$ durch $g_h(\alpha) := f(x_0 + \alpha h)$, so hat g_h an $\alpha = 0$ ein lokales Minimum und wir wissen aus der reellen Differentialrechnung, dass

$$\frac{d}{d\alpha} g_h(\alpha)\Big|_{\alpha=0} = 0$$

gilt. Dies können wir nun auf f übertragen und erhalten:

$$\begin{aligned} 0 &= \frac{d}{d\alpha} f(x_0 + \alpha h)\Big|_{\alpha=0} \\ &= \lim_{t \to 0} \frac{1}{t}(f(x_0 + (\alpha + t)h) - f(x_0 + \alpha))|_{\alpha=0} \\ &= \lim_{t \to 0} \frac{1}{t}(f(x_0 + th) - f(x_0)) \\ &= f'(x_0)(h). \end{aligned}$$

□

Ist das Funktional f, für das wir ein Minimum bestimmen wollen, sogar konvex, so ist die obige Bedingung auch hinreichend:

Satz 3.26 *Das in Satz 3.25 definierte Funktional sei zusätzlich konvex. Erfüllt $x_0 \in X$ für alle $h \in X$ die Gleichung*

$$f'(x_0)(h) = 0,$$

so ist x_0 ein lokales Minimum von f.

Beweis: Da f konvex ist, gilt für alle $x, h \in X$

$$f(x + h) \geq f(x) + f'(x)(h).$$

Dies ist insbesondere auch für $x = x_0$ gültig, so dass wir wegen $f'(x_0)(h) = 0$

$$f(x_0 + h) \geq f(x_0)$$

erhalten. Da dies für alle $h \in X$ gilt, ist x_0 ein lokales Minimum. □

Unter der Voraussetzung, dass f konvex ist, wissen wir, dass ein lokales Minimum einer konvexen Menge sogar globales Minimum ist:

Satz 3.27 *Seien C eine konvexe Teilmenge eines normierten Raumes, $f : C \to \mathbb{R}$ ein konvexes Funktional $\mu := \inf_{x \in C} f(x)$ und $x_0 \in C$ ein lokales Minimum von f. Dann ist x_0 sogar ein globales Minimum von f auf C mit $f(x_0) = \mu$.*

Beweis: Da x_0 ein lokales Minimum von f ist, wissen wir, dass eine Umgebung U von x_0 existiert, so dass

$$f(x_0) \leq f(x), \qquad \forall\ x \in U,$$

gilt.
Wir zeigen, dass $f(x_0) \leq f(x)$ sogar für alle $x \in C$ gilt.

Bezeichne mit

$$\begin{aligned} g : \mathbb{R} &\to C \\ t &\mapsto tx + (1-t)x_0 \end{aligned}$$

die Konvexkombination von x_0 und einem beliebigen x aus C. Offensichtlich ist g stetig und wir erhalten:

$$\lim_{t \to 0} g(t) = x_0.$$

Ist t klein genug, so liegt $g(t)$ in der Umgebung U von x_0, und dies liefert uns:

$$\begin{aligned} f(x_0) \leq f(g(t)) &= f(tx + (1-t)x_0), && \text{für } t \text{ klein genug,} \\ &\leq tf(x) + (1-t)f(x_0), && \text{wegen der Konvexität von } f. \end{aligned}$$

Formen wir diese Ungleichung um und erinnern uns daran, dass $x \in C$ beliebig gewählt war, so erhalten wir

$$f(x_0) \leq f(x), \qquad \forall\ x \in C,$$

also globale Konvergenz. □

Zusammenfassend erhalten wir also für das reduzierte Steuerungsfunktional:

Korollar 3.28 *Ist j aus Definition 3.20 zusätzlich konvex, so ist $\widehat{q} \in Q$ genau dann ein globales Minimum, wenn die* Optimalitätsbedingung erster Ordnung

$$j'(\widehat{q})(\tau q) = 0, \qquad \forall\ \tau q \in Q,$$

erfüllt ist.

Beweis: Aus den Sätzen 3.25 und 3.26 folgt, dass $\widehat{q} \in Q$ genau dann ein lokales Minimum ist, wenn

$$j'(\widehat{q})(\tau q) = 0, \qquad \forall\ \tau q \in Q,$$

erfüllt ist. Satz 3.27 liefert uns außerdem, da Q konvex ist, dass dieses lokale Minimum auch ein globales Minimum ist. □

Kapitel 4

Das Newton-Verfahren

Dieses Kapitel beschäftigt sich mit der Lösung des Optimierungsproblems (3.3)

$$\text{Minimiere } j(q) \text{ mit } q \in Q.$$

Dazu werden wir zunächst das Newton-Verfahren vorstellen. Für dessen konkrete Umsetzung wollen wir dann die nötigen Voraussetzungen schaffen, bevor wir in Abschnitt 4.5 entsprechende Algorithmen präsentieren. Die Idee folgt hierbei [BMV07]. Zuletzt werden wir auf die Besonderheiten der in dieser Arbeit behandelten linear-quadratischen Aufgabenstellung eingehen.
Schon an dieser Stelle wollen wir darauf hinweisen, dass einige Untersuchungen in diesem Kapitel für solche linear-quadratischen Probleme nicht unbedingt nötig wären. Man erhält auf diese Weise jedoch einen Lösungsansatz, der sich leicht auf nichtlineare Probleme erweitern lässt.

4.1 Das Newton-Verfahren

Um das unrestringierte Optimierungsproblem (3.3)

$$\text{Minimiere } j(q) \text{ mit } q \in Q$$

zu lösen, wollen wir das Newton-Verfahren anwenden. Die Idee hierbei ist, ein $\widehat{q} \in Q$ zu finden, so dass die notwendige und hinreichende Optimalitätsbedingung erster Ordnung aus Korollar 3.28

$$j'(\widehat{q})(\tau q) = 0, \quad \forall\, \tau q \in Q, \tag{4.1}$$

erfüllt ist.

Der Algorithmus sieht in seiner Grobstruktur folgendermaßen aus:

Algorithmus 4.1 : Allgemeiner Newton-Algorithmus

Input : $n = 0$, $q^0 \in Q$ (Startwert)

1 **while** *Abbruchbedingung ist nicht erfüllt* **do**
2 Berechne δq als Lösung des linearen Problems
3 $j''(q^n)(\delta q, \tau q) = -j'(q^n)(\tau q), \quad \forall\, \tau q \in Q;$
4 $q^{n+1} = q^n + \delta q;$
5 $n = n + 1;$

Das Resultat dieses Algorithmus ist eine Näherung q^N an die Lösung $\widehat{q}$ der Gleichung (4.1) und damit an die optimale Lösung des Optimierungsproblems (3.3). Unter gewissen Voraussetzungen entspricht das Ergebnis des Algorithmus sogar der exakten Lösung (vergleiche Lemma 4.14).

Um Algorithmus 4.1 durchführen zu können, benötigt man in Zeile 3 die ersten beiden Ableitungen von j. Diese werden wir für das reduzierte Steuerungsfunktional im Folgenden mit Hilfe von Lagrange-Funktionalen herleiten.

Zunächst wollen wir aber auf das Konvergenzverhalten des Newton-Verfahrens eingehen.

Konvergenz des Newton-Verfahrens

Wir wollen nun das Newton-Verfahren für Funktionen auf einem Banachraum untersuchen. Dazu seien Y, Z Banachräume und $F : Y \to Z$ eine Fréchet-differenzierbare Funktion.

Gesucht ist die Lösung $\widetilde{y}$ der (nichtlinearen) Gleichung

$$F(y) = 0.$$

Die Idee ist, ausgehend von einem Startwert y_0 mittels der Iteration

$$y_{n+1} := y_n - \left(F'(y_n)\right)^{-1} \left(F(y_n)\right) \tag{4.2}$$

eine Approximation y_N an die exakte Lösung $\widetilde{y}$ zu bestimmen.

Liegt der Startwert y_0 nah genug an der gesuchten Lösung $\widetilde{y}$, so konvergiert das Newton-Verfahren. Diese Tatsache wird im folgenden Satz mathematisch exakt formuliert:

Satz 4.1 *Seien Y, Z Banachräume, $U \subset Y$ eine Teilmenge von Y und $F : U \to Z$ Fréchet-differenzierbar auf U. Außerdem seien die folgenden Voraussetzungen erfüllt:*

1. *Für alle $x, y \in U$ gelte mit einem $\varepsilon \in \mathbb{R}$*

$$\|F(x) - F(y) - F'(y)(x-y)\|_Z \leq \varepsilon \|x-y\|_Y.$$

2. *Die Fréchet-Ableitung $F'(y)$ sei für alle $y \in U$ invertierbar und es existieren $M, K \in \mathbb{R}$ mit $\varepsilon M =: c < 1$, so dass*

$$\|F'(y)\|_{L(Y,Z)} \leq K \qquad \textit{und} \qquad \| \left(F'(y)\right)^{-1} \|_{L(Y,X)} \leq M.$$

3. *Es existiere ein Startwert $y_0 \in U$ so, dass*

$$y_1 := y_0 - \left(F'(y_0)\right)^{-1} \left(F(y_0)\right)$$

in U liegt und auch der Ball

$$B_r(y_1) := \{x \in Y : \|x - y_1\|_Y < r\}$$

mit $r := \frac{c}{1-c}\|y_0 - y_1\|_Y$ ganz in U enthalten ist.

Dann existiert in $\overline{B}_r(y_1)$ eine Lösung $\widetilde{y}$ von $F(y) = 0$ und das Newton-Verfahren (4.2) *konvergiert bei der Wahl von y_1 als Startwert gegen die exakte Lösung $\widetilde{y}$.*

Gilt zusätzlich in einer Umgebung W von $\widetilde{y}$ die Abschätzung

$$\|F(x) - F(y) - F'(y)(x-y)\|_Z \leq L\|x-y\|_Y^2 \leq \varepsilon \|x-y\|_Y$$

für $x, y \in W$ und ein festes $L > 0$, so konvergiert das Newton-Verfahren quadratisch, d.h.

$$\lim_{n\to\infty} \frac{\|y_{n+1} - \widetilde{y}\|_Y}{\|y_n - \widetilde{y}\|_Y^2} \leq C < \infty$$

für ein $C > 0$.

Beweis: Ein Beweis hierzu findet sich in [SW92], Satz 8.6.12 und Korollar 8.6.20. □

Die Optimalitätsbedingung 1. Ordnung (4.1)

$$j'(q)(\tau q) = 0, \quad \forall\, \tau q \in Q,$$

die bisher den Ausgangspunkt des Newton-Verfahrens darstellt, hat im Moment eine andere Form als die Ausgangsgleichung

$$F(y) = 0,$$

die wir in diesem Abschnitt untersucht haben. Verwendet man aber den folgenden Zusammenhang von j' und dem Gradienten von j:

$$(\nabla j(q), \tau q)_Q = j'(q)(\tau q), \quad \forall\, \tau q \in Q,$$

so ist (4.1) äquivalent zu

$$\nabla j(\widehat{q}) = 0. \tag{4.3}$$

Setzt man also $F(y) := \nabla j(\widehat{q})$ und beachtet, dass die Iterationsvorschrift (4.2)

$$y_{n+1} := y_n - \left(F'(y_n)\right)^{-1} (F(y_n))$$

äquivalent ist zu

$$F'(y_n)(y_{n+1} - y_n) = -F(y_n),$$

so können wir Satz 4.1 anwenden und erhalten die Voraussetzungen, unter welchen das in Algorithmus 4.1 vorgestellte Newton-Verfahren konvergiert. Bei der Umsetzung des Newton-Verfahrens zur Lösung von $\nabla j(q) = 0$ müssen wir jetzt natürlich statt j'' die Hesse-Matrix $\nabla^2 j(q)$ verwenden. Dafür benutzen wir folgenden Zusammenhang:

$$(\nabla^2 j(q)\delta q, \tau q)_Q = j''(q)(\delta q, \tau q), \quad \forall\, \delta q, \tau q \in Q.$$

Damit ist Zeile 4 aus Algorithmus 4.1

$$j''(q^n)(\delta q, \tau q) = -j'(q^n)(\tau q), \quad \forall\, \tau q \in Q, \tag{4.4}$$

äquivalent zu

$$\nabla^2 j(q)\delta q = -\nabla j(q). \tag{4.5}$$

4.2 Das Lagrange-Funktional

Nun wollen wir das Lagrange-Funktional für das in Abschnitt 3.2 vorgestellte Optimierungsproblem einführen, das die Grundlage der weiteren Untersuchungen in diesem Kapitel darstellt.

Definition 4.2 Abhängig vom Steuerungsfunktional J und der Zustandsgleichung (2.3) definieren wir das Lagrange-Funktional

$$\begin{aligned}\mathcal{L} : Q \times X \times X &\longrightarrow \mathbb{R} \\ (q,u,z) &\longmapsto J(q,u) + (f(q) - \partial_t u, z) - a(u,z) \\ &\quad - (u(0) - u_0(q), z(0))_H. \end{aligned} \tag{4.6}$$

Im Folgenden werden wir vor allem die ersten und zweiten partiellen Ableitungen (im Sinne der Fréchet-Differenzierbarkeit nach Abschnitt 3.1.1) des Lagrange-Funktionals benötigen. Deshalb führen wir sie im folgenden Lemma einmal zusammenfassend auf. Dafür erinnern wir an die Definition des Steuerungsfunktionals J aus Definition 3.19:

$$\begin{aligned} J(q,u) = &\int_I \left\{ \frac{1}{2} b_1(u(t),u(t)) + L_1(u(t)) + C_1 \right\} dt \\ &+ \frac{1}{2} b_2(u(T),u(T)) + L_2(u(T)) + C_2 \\ &+ \frac{\alpha}{2} \|q - \overline{q}\|_Q^2. \end{aligned}$$

Lemma 4.3 *Die ersten und zweiten partiellen Ableitungen von $\mathcal{L}$ haben folgende Form:*

$$\mathcal{L}'_q(q,u,z)(\delta q) = \alpha(q - \overline{q}, \delta q)_Q + (f^{(1)}(\delta q), z) + (u_0^{(1)}(\delta q), z(0))_H \tag{4.7}$$

$$\begin{aligned} \mathcal{L}'_u(q,u,z)(\delta u) = &\int_I \{ b_1(\delta u(t), u(t)) + L_1(\delta u(t)) \}\, dt \\ &+ b_2(\delta u(T), u(T)) + L_2(\delta u(T)) \\ &- (\partial_t \delta u, z) - a(\delta u, z) - (\delta u(0), z(0))_H \end{aligned} \tag{4.8}$$

$$\begin{aligned} \mathcal{L}'_z(q,u,z)(\delta z) = &(f(q), \delta z) - (\partial_t u, \delta z) - a(u, \delta z) \\ &- (u(0) - u_0(q), \delta z(0))_H \end{aligned} \tag{4.9}$$

und

$$\begin{aligned} \mathcal{L}''_{qq}(q,u,z)(\delta q, \tau q) &= \alpha(\tau q, \delta q)_Q \\ \mathcal{L}''_{qu}(q,u,z)(\delta q, \tau u) &= 0 \\ \mathcal{L}''_{qz}(q,u,z)(\delta q, \tau z) &= (f^{(1)}(\delta q), \tau z) + (u_0^{(1)}(\delta q), \tau z(0))_H \\ \mathcal{L}''_{uq}(q,u,z)(\delta u, \tau q) &= 0 \\ \mathcal{L}''_{uu}(q,u,z)(\delta u, \tau u) &= \int_I b_1(\delta u(t), \tau u(t))\, dt + b_2(\delta u(T), \tau u(T)) \end{aligned}$$

$$\begin{aligned}
\mathcal{L}''_{uz}(q,u,z)(\delta u,\tau z) &= (-\partial_t \delta u,\tau z) - a(\delta u,\tau z) - (\delta u(0),\tau z(0))_H \\
\mathcal{L}''_{zq}(q,u,z)(\delta z,\tau q) &= (f^{(1)}(\tau q),\delta z) + (u_0^{(1)}(\tau q),\delta z(0))_H \\
\mathcal{L}''_{zu}(q,u,z)(\delta z,\tau u) &= (-\partial_t \tau u,\delta z) - a(\tau u,\delta z) - (\tau u(0),\delta z(0))_H \\
\mathcal{L}''_{zz}(q,u,z)(\delta z,\tau z) &= 0
\end{aligned}$$

mit $\delta q, \tau q \in Q$ und $\delta u, \delta z, \tau u, \tau z \in X$.

Die dritten partiellen Ableitungen sind alle identisch Null.

Beweis: Dies folgt durch elementare Rechnungen. □

4.3 Hilfsgleichungen

Um später Ausdrücke für die Ableitungen von j aufstellen zu können, benötigen wir einige Hilfsgleichungen.

Jede dieser Gleichungen wird auf zwei unterschiedliche Arten formuliert: Einerseits werden die Ableitungen des Lagrange-Funktionals benutzt, andererseits werden diese Ableitungen durch die expliziten Terme des Optimierungsproblems dargestellt.

Duales Problem: Zu gegebenem $q \in Q$ und $u = S(q) \in X$ finde $z \in X$, so dass gilt:

$$\mathcal{L}'_u(q,u,z)(\phi) = 0, \quad \forall\, \phi \in X, \tag{4.10a}$$

$$\Leftrightarrow$$

$$(-\phi, \partial_t z) + a(\phi, z) = \int_I \{b_1(\phi(t), u(t)) + L_1(\phi(t))\}\, dt, \quad \forall\, \phi \in X, \tag{4.10b}$$

$$z(T) = \frac{1}{2} b_2'(u(T), u(T)) + L_2'(u(T)) \tag{4.10c}$$

mit

$$\frac{1}{2} b_2'(u(T), u(T))(\phi(T)) + L_2'(u(T))(\phi(T)) = b_2(\phi(T), u(T)) + L_2(\phi(T))$$

für alle $\phi \in X$.

Tangentengleichung: Zu gegebenem $\delta q \in Q$ finde $\delta u \in X$, so dass gilt:

$$\mathcal{L}''_{qz}(q,u,z)(\delta q,\phi) + \mathcal{L}''_{uz}(q,u,z)(\delta u,\phi) = 0, \quad \forall\, \phi \in X, \tag{4.11a}$$

$$\Leftrightarrow$$

$$(\partial_t \delta u,\phi) + a(\delta u,\phi) = (f^{(1)}(\delta q),\phi), \quad \forall\, \phi \in X, \tag{4.11b}$$

$$\delta u(0) = u_0^{(1)}(\delta q). \tag{4.11c}$$

Duale Hesse-Gleichung: Zu gegebenem $\delta q \in Q$ sei $\delta u \in X$ die Lösung der Tangentengleichung (4.11). Gesucht ist $\delta z \in X$, so dass gilt:

$$\mathcal{L}''_{uu}(q,u,z)(\delta u,\phi) + \mathcal{L}''_{zu}(q,u,z)(\delta z,\phi) = 0, \quad \forall\, \phi \in X, \tag{4.12a}$$

$$\Leftrightarrow$$

$$(-\phi,\partial_t \delta z) + a(\phi,\delta z) = \int_I b_1(\delta u(t),\phi(t))\,dt, \quad \forall\, \phi \in X, \tag{4.12b}$$

$$\delta z(T) = b_2'(\delta u(T),u(T)) \tag{4.12c}$$

mit $b_2'(\delta u(T),u(T))(\phi(T)) = b_2(\delta u(T),\phi(T))$ für alle $\phi \in X$.

Die Äquivalenz der beiden Formulierungen erhält man, indem man die Ableitungen von $\mathcal{L}$ ausrechnet, einsetzt, partiell integriert und die Randterme separiert.
Für das duale Problem wollen wir nun stellvertretend für alle drei Gleichungen diese Äquivalenz der beiden Formulierungen zeigen. Die Beweisidee folgt hierbei [Trö05], S. 97.

Lemma 4.4 *Gleichung* (4.10a) *ist äquivalent zu den Gleichungen* (4.10b) *und* (4.10c)

Beweis: Wir beginnen mit der partiellen Ableitung des Lagrange-Funktionals:

$$\mathcal{L}'_u(q,u,z)(\phi) = 0, \quad \forall\, \phi \in X.$$

Hier setzen wir die Darstellung (4.8) ein:

$$\begin{aligned}(\partial_t \phi,z) + a(\phi,z) + (\phi(0),z(0))_H = \int_I \{b_1(\phi(t),u(t)) + L_1(\phi(t))\}\,dt \\ + b_2(\phi(T),u(T)) + L_2(\phi(T)).\end{aligned}$$

Mittels partieller Integration erhält man:

$$(\partial_t \phi,z) = (-\phi,\partial_t z) + (\phi(T),z(T))_H - (\phi(0),z(0))_H.$$

Dies liefert uns:

$$(-\phi, \partial_t z) + a(\phi, z) + (\phi(T), z(T))_H = \int_I \{b_1(\phi(t), u(t)) + L_1(\phi(t))\}\, dt \\ + b_2(\phi(T), u(T)) + L_2(\phi(T))$$

für alle $\phi \in X$.

Wählt man zuerst ϕ beliebig aus $C_0^\infty([0,T];\Omega)$, so gilt $\phi(T) = 0$. Deshalb sind die Terme $(\phi(T), z(T))_H$, $b_2(\phi(T), u(T))$ und $L_2(\phi(T))$ identisch Null und es bleibt:

$$(-\phi, \partial_t z) + a(\phi, z) = \int_I \{b_1(\phi(t), u(t)) + L_1(\phi(t))\}\, dt.$$

Verzichtet man nun auf diese Forderung, erhält man außerdem die Gleichung

$$(\phi(T), z(T))_H = \frac{1}{2} b_2(\phi(T), u(T)) + L_2(\phi(T)),$$

was nach dem Darstellungssatz von Fréchet-Riesz (siehe [Heu86], Satz 26.1) äquivalent zu

$$z(T) = b_2'(u(T), u(T)) + L_2'(u(T))$$

ist. Insgesamt erhalten wir also, da $C_0^\infty([0,T];\Omega)$ dicht in $L^2([0,T];\Omega)$ liegt:

$$(-\phi, \partial_t z) + a(\phi, z) = \int_I \{b_1(\phi(t), u(t)) + L_1(\phi(t))\}\, dt, \quad \forall\, \phi \in X,$$
$$z(T) = \frac{1}{2} b_2'(u(T), u(T)) + L_2'(u(T)). \qquad \square$$

An der expliziten Darstellung der Tangentengleichung (4.11b) und der dualen Hesse-Gleichung (4.12b) sieht man bereits, dass diese beiden Gleichungen unabhängig von der Lösung der Zustandsgleichung, des dualen Problems und der Steuerung q sind. Sie können entsprechend separat gelöst werden. Diese Tatsache erhält man auch aus der folgenden Darstellung mittels der Lösungsoperatoren:

Bemerkung 4.5 Analog zum affin linearen Lösungsoperator S der Zustandsgleichung (2.3) führen wir jetzt einen Lösungsoperator R des dualen Problems (4.10) ein:

$$R : X \longrightarrow X$$
$$u \longmapsto z.$$

Da z affin linear von u abhängig ist, können wir den Operator R wie folgt schreiben:

$$z = R(u) = R^{(1)}(u) + R^{(2)},$$

wobei $R^{(1)} : X \to X$ ein linearer Operator ist und $R^{(2)} \in X$ gilt.

- Wir beobachten, dass bei gegebenen $q, \delta q \in Q$ aus

  $$u = S(q) = S^{(1)}(q) + S^{(2)}$$

  mit der Bezeichnung

  $$\begin{aligned} \delta u &:= S'(q)(\delta q) \\ &= S^{(1)}(\delta q) \end{aligned}$$

  folgt, dass δu Lösung der Tangentengleichung (4.11) ist.

- Ähnliches gilt für das duale Problem und die duale Hesse-Gleichung: Mit

  $$z = R(u) = R\left(S(q)\right) = R^{(1)}\left(S(q)\right) + R^{(2)}$$

  Lösung des dualen Problems (4.10) gilt, dass

  $$\begin{aligned} \delta z &:= \left(R\left(S(q)\right)\right)'(\delta q) \\ &= R^{(1)}\left(S^{(1)}(\delta q)\right) \end{aligned}$$

  die Lösung der dualen Hesse-Gleichung (4.12) ist.

4.4 Ableitungen des Steuerungsfunktionals

Nun haben wir alle nötigen Voraussetzungen erarbeitet, um die folgenden Sätze zur ersten und zweiten Ableitung von j formulieren und beweisen zu können.

Satz 4.6 *Sei $q \in Q$ gegeben und es gelte:*

(i) *$u = S(q) \in X$ sei Lösung der Zustandsgleichung (2.3),*

(ii) *$z \in X$ sei Lösung des dualen Problems* (4.10),

dann gilt für die erste Ableitung des reduzierten Steuerungsfunktionals:

$$\begin{aligned} j'(q)(\tau q) &= \mathcal{L}'_q(q,u,z)(\tau q) \\ &= \alpha(q-\overline{q},\tau q)_Q + (f^{(1)}(\tau q),z) + (u_0^{(1)}(\tau q),z(0))_H \end{aligned} \tag{4.13}$$

für alle $\tau q \in Q$.

Beweis: Da $u = S(q) \in X$ nach (i) Lösung der Zustandsgleichung (2.3) ist, gilt:

$$\begin{aligned} j(q) &= J(q,S(q)) \\ &= \mathcal{L}(q,u,z) - (f(q) - \partial_t u, z) + a(u,z) + (u(0) - u_0(q), z(0))_H \\ &= \mathcal{L}(q,u,z). \end{aligned}$$

Mit $\tau u = u'_q(\tau q) = S'(q)(\tau q)$ und $\tau z = z'_q(\tau q) = R(S(q))'(\tau q)$ gilt dann:

$$\begin{aligned} j'(q)(\tau q) &= \mathcal{L}'_q(q,u,z)(\tau q) + \mathcal{L}'_u(q,u,z)(\tau u) + \mathcal{L}'_z(q,u,z)(\tau z) \\ &= \mathcal{L}'_q(q,u,z)(\tau q) + \mathcal{L}'_z(q,u,z)(\tau z) && \text{wegen (ii)} \\ &= \mathcal{L}'_q(q,u,z)(\tau q) + (f(q),\tau z) - (\partial_t u,\tau z) - a(u,\tau z) \\ &\quad - (u(0) - u_0(q),\tau z(0))_H && \text{wegen (4.9)} \\ &= \mathcal{L}'_q(q,u,z)(\tau q) && \text{wegen (i)} \\ &= \alpha(q-\overline{q},\tau q)_Q + (f^{(1)}(\tau q),z) + (u_0^{(1)}(\tau q),z(0))_H && \text{wegen (4.7)} \end{aligned}$$

für alle $\tau q \in Q$. □

Bemerkung 4.7 Unabhängig von der Lösung durch das Newton-Verfahren können wir nun ein System von Gleichungen festhalten, das die Lösung $\widehat{q}$ des Optimierungsproblems (3.3) bestimmt:
Gilt

$$\left.\begin{aligned} \mathcal{L}'_z(\widehat{q},u,z)(\tau z) &= 0, \quad \forall\, \tau z \in X, && \text{(Zustandsgleichung)} \\ \mathcal{L}'_u(\widehat{q},u,z)(\tau u) &= 0, \quad \forall\, \tau u \in X, && \text{(duales Problem)} \\ \mathcal{L}'_{\widehat{q}}(\widehat{q},u,z)(\tau q) &= 0, \quad \forall\, \tau q \in Q, && \text{(Optimalitätsbedingung)} \end{aligned}\right\} \tag{4.14}$$

so löst $\widehat{q}$ die Gleichung

$$j'(q)(\tau q) = 0, \quad \forall\, \tau q \in Q,$$

und damit das Optimierungsproblem (3.3).

Der folgende Satz liefert uns eine Darstellung für die zweite Ableitung von j und damit für die Hesse-Matrix $\nabla^2 j$.

Satz 4.8 *Seien $\delta q \in Q$ und $\tau q \in Q$ gegeben und es gelte:*

(i) $\delta u \in X$ sei Lösung der Tangentengleichung (4.11) *zu δq,*

(ii) $\tau u \in X$ sei Lösung der Tangentengleichung (4.11) *zu τq,*

dann gilt für die zweite Ableitung des reduzierten Steuerungsfunktionals:

$$\begin{aligned} j''(q)(\delta q, \tau q) &= \mathcal{L}''_{qq}(q,u,z)(\delta q, \tau q) + \mathcal{L}''_{uu}(q,u,z)(\delta u, \tau u) \\ &= \alpha(\tau q, \delta q)_Q + \int_I b_1(\delta u(t), \tau u(t))\, dt + b_2(\delta u(T), \tau u(T)). \end{aligned} \tag{4.15}$$

Beweis: Wir wissen bereits aus Satz 4.6, dass für die erste Ableitung des reduzierten Steuerungsfunktionals gilt:

$$j'(q)(\delta q) = \mathcal{L}'_q(q,u,z)(\delta q) + \mathcal{L}'_u(q,u,z)(\delta u) + \mathcal{L}'_z(q,u,z)(\delta z).$$

Um die zweite Ableitung von j zu bestimmen, müssen wir nun die totale Ableitung dieses Ausdrucks berechnen. Dabei beachten wir, dass

$$\delta u = u'_q(\delta q) = S^{(1)}(\delta q)$$

und

$$\delta z = z'_q(\delta q) = R^{(1)}(S^{(1)}(\delta q))$$

unabhängig von q sind. Es gilt also:

$$\begin{aligned} j''(q)(\delta q, \tau q) = {} & \mathcal{L}''_{qq}()(\delta q, \tau q) + \mathcal{L}''_{qu}()(\delta q, \tau u) + \mathcal{L}''_{qz}()(\delta q, \tau z) \\ & + \mathcal{L}''_{uq}()(\delta u, \tau q) + \mathcal{L}''_{uu}()(\delta u, \tau u) + \mathcal{L}''_{uz}()(\delta u, \tau z) \\ & + \mathcal{L}''_{zq}()(\delta z, \tau q) + \mathcal{L}''_{zu}()(\delta z, \tau u) + \mathcal{L}''_{zz}()(\delta z, \tau z), \end{aligned}$$

wobei jeweils der Übersichtlichkeit zuliebe die Abhängigkeit von q, u und z unterdrückt wurde. Hierbei verschwinden $\mathcal{L}''_{qu}()(\delta q, \tau u)$, $\mathcal{L}''_{uq}()(\delta u, \tau q)$ und $\mathcal{L}''_{zz}()(\delta z, \tau z)$ nach Lemma 4.3 und wegen (i) und (ii) gelten

$$\mathcal{L}''_{qz}()(\delta q, \tau z) + \mathcal{L}''_{uz}()(\delta u, \tau z) = 0$$

und

$$\mathcal{L}''_{zq}()(\delta z, \tau q) + \mathcal{L}''_{zu}()(\delta z, \tau u) = 0.$$

Also bleibt

$$\begin{aligned} j''(q)(\delta q, \tau q) &= \mathcal{L}''_{qq}(q,u,z)(\delta q, \tau q) + \mathcal{L}''_{uu}(q,u,z)(\delta u, \tau u) \\ &= \alpha(\tau q, \delta q)_Q + \int_I b_1(\delta u(t), \tau u(t))\, dt + b_2(\delta u(T), \tau u(T)). \end{aligned}$$

□

Eine leicht abgeänderte Version von Satz 4.8 liefert nicht eine explizite Darstellung für j'', sondern nur die Wirkung von j'' auf ein $\delta q \in Q$. Gewissermaßen gibt der folgende Satz also an, wie das Matrix-Vektor-Produkt der Hesse-Matrix mit einem beliebigen Vektor $\delta q \in Q$ zu berechnen ist.

Satz 4.9 *Sei $\delta q \in Q$ gegeben und es gelte:*

(i) $\delta u \in X$ sei Lösung der Tangentengleichung (4.11)*,*

(ii) $\delta z \in X$ sei Lösung der dualen Hesse-Gleichung (4.12)*,*

dann gilt für die zweite Ableitung des reduzierten Steuerungsfunktionals:

$$\begin{aligned} j''(q)(\delta q, \tau q) &= \mathcal{L}''_{qq}(q,u,z)(\delta q, \tau q) + \mathcal{L}''_{zq}(q,u,z)(\delta z, \tau q) \\ &= \alpha(\tau q, \delta q)_Q + (f^{(1)}(\tau q), \delta z) + (u_0^{(1)}(\tau q), \delta z(0))_H \end{aligned} \tag{4.16}$$

für alle $\tau q \in Q$.

Beweis: Wie im Beweis zu Satz 4.8 gilt

$$\begin{aligned} j''(q)(\delta q, \tau q) = {} & \mathcal{L}''_{qq}()(\delta q, \tau q) + \mathcal{L}''_{qu}()(\delta q, \tau u) + \mathcal{L}''_{qz}()(\delta q, \tau z) \\ & + \mathcal{L}''_{uq}()(\delta u, \tau q) + \mathcal{L}''_{uu}()(\delta u, \tau u) + \mathcal{L}''_{uz}()(\delta u, \tau z) \\ & + \mathcal{L}''_{zq}()(\delta z, \tau q) + \mathcal{L}''_{zu}()(\delta z, \tau u) + \mathcal{L}''_{zz}()(\delta z, \tau z), \end{aligned}$$

mit $\mathcal{L}''_{qu}()(\delta q, \tau u) = \mathcal{L}''_{uq}()(\delta u, \tau q) = \mathcal{L}''_{zz}()(\delta z, \tau z) = 0$ nach Lemma 4.3. Außerdem folgen aus (i) und (ii)

$$\mathcal{L}''_{qz}()(\delta q, \tau z) + \mathcal{L}''_{uz}()(\delta u, \tau z) = 0$$

und

$$\mathcal{L}''_{uu}()(\delta u, \tau u) + \mathcal{L}''_{zu}()(\delta z, \tau u) = 0$$

Es bleibt also

$$\begin{aligned} j''(q)(\delta q, \tau q) &= \mathcal{L}''_{qq}(q,u,z)(\delta q, \tau q) + \mathcal{L}''_{zq}(q,u,z)(\delta z, \tau q) \\ &= \alpha(\tau q, \delta q)_Q + (f^{(1)}(\tau q), \delta z) + (u_0^{(1)}(\tau q), \delta z(0))_H \end{aligned}$$

für alle $\tau q \in Q$. □

Bemerkung 4.10 An dieser Stelle wollen wir auf eine Besonderheit hinweisen, die in dem in dieser Arbeit behandelten Fall auftritt, da wir uns auf quadratische Funktionale und lineare Zustandsgleichungen beschränken: Die Darstellungen für die Hesse-Matrix und auch die für die Wirkung dieser auf einen Vektor sind unabhängig von q, da die Tangentengleichung und die duale Hesse-Gleichung unabhängig von q sind.

Dies ist deshalb so erwähnenswert, da diese Eigenschaft den Aufwand unseres Algorithmus im folgenden Abschnitt erheblich gegenüber dem allgemeineren Fall verringern wird.

4.5 Der konkrete Algorithmus

In diesem Abschnitt formulieren wir nun, wie der Newton-Algorithmus auszusehen hat. Dabei gehen wir von der folgenden Voraussetzung aus:

Voraussetzungen 4.11 *Sei Q ein endlich dimensionaler Raum und*

$$\{\tau q_i \in Q : i = 1, \ldots, \dim(Q)\}$$

die zugehörige Standardbasis.

Zur Erinnerung: In jedem Schritt des Newton-Algorithmus ist die Lösung des Systems

$$\nabla^2 j(q)\delta q = -\nabla j(q) \tag{4.17}$$

zu bestimmen. Deshalb müssen wir zunächst $\nabla j(q)$ und $\nabla^2 j(q)$ aufstellen. Da die Hessematrix $\nabla^2 j(q)$ nach Bemerkung 4.10 unabhängig von q ist, können wir diese einmal zu Beginn des Newton-Algorithmus berechnen. Nur der Gradient $\nabla j(q)$ ist abhängig von q und muss in jedem Schritt neu berechnet werden.

Damit ergibt sich unter Anwendung der Sätze 4.6 und 4.8 folgender konkreter Newton-Algorithmus:

Algorithmus 4.2 : Newton-Algorithmus mit Aufstellen der Hesse-Matrix

input : $n = 0$, $q^0 \in Q$

1 Berechne $\{\tau u_i^n | i = 1, \ldots, \dim(Q)\} \subset X$ für die entsprechende Basis von Q, d.h. löse die Tangentengleichung (4.11) für jeden Basisvektor τq_i von Q;
2 Stelle mittels Satz 4.8 $\nabla^2 j(q^n)$ auf, d.h., um die ij-te Komponente $(\nabla^2 j(q^n))_{ij}$ zu berechnen, werte die rechte Seite von (4.15) für $\delta q = \tau q_j$, $\tau q = \tau q_i$, $\delta u = \tau u_j$ und $\tau u = \tau u_i$ aus;
3 **while** $\|\nabla j(q^n)\| \geq TOL$ **do**
4 Berechne $u^n \in X$, d.h. löse die Zustandsgleichung (2.3);
5 Berechne $z^n \in X$, d.h. löse das duale Problem (4.10) ;
6 Stelle mittels Satz 4.6 $\nabla j(q^n)$ auf, d.h., um die i-te Komponente $(\nabla j(q^n))_i$ zu bestimmen, werte die rechte Seite von (4.13) für $\tau q = \tau q_i$ aus;
7 Löse $\nabla^2 j(q^n)\delta q = -\nabla j(q^n)$ mittels eines beliebigen linearen Lösers ;
8 $q^{n+1} = q^n + \delta q$;
9 $n = n + 1$;

Bemerkung 4.12 In jedem Newton-Schritt werden bei diesem Algorithmus

- einmal die Zustandsgleichung (2.3)
- einmal das duale Problem (4.10)

gelöst. Vor Beginn der Newton-Iteration ist $dim(Q)$-mal die Tangentengleichung zu lösen.

In Schritt 7 von Algorithmus 4.2 konnten wir zur Lösung des linearen Gleichungssystems $\nabla^2 j(q^n)\delta q = -\nabla j(q^n)$ jeden beliebigen linearen Löser verwenden, da die komplette Hesse-Matrix bekannt war.
Einen alternativen Algorithmus erhalten wir, wenn wir uns auf Lösungsverfahren einschränken, die nicht die Matrix an sich benötigen, sondern nur die Ergebnisse von Matrix-Vektor-Multiplikationen. Das *Konjugierte Gradientenverfahren* (kurz: CG-Verfahren) ist ein solches Verfahren. Eine ausführliche Darstellung dieses Verfahrens findet sich beispielsweise in [Saa96], Abschnitt 6.7.

Wählen wir also das CG-Verfahren zur Lösung des Gleichungssystems, so können wir den Algorithmus mit Hilfe der Sätze 4.6 und 4.9 wie folgt neu formulieren:

Algorithmus 4.3 : Newton-Algorithmus ohne Aufstellen der Hesse-Matrix

input : $n = 0$, $q^0 \in Q$

1 **while** $\|\nabla j(q^n)\| \geq TOL$ **do**
2 Berechne $u^n \in X$, d.h. löse die Zustandsgleichung (2.3);
3 Berechne $z^n \in X$, d.h. löse das duale Problem (4.10) ;
4 Stelle mittels Satz 4.6 $\nabla j(q^n)$ auf, d.h., um die i-te Komponente $(\nabla j(q^n))_i$ zu bestimmen, werte die rechte Seite von (4.13) für $\tau q = \tau q_i$ aus;
5 Löse $\nabla^2 j(q^n)\delta q = -\nabla j(q^n)$ mittels des CG-Verfahrens. Um die nötigen Matrix-Vektor-Produkte zu berechnen, verwende Algorithmus 4.4;
6 $q^{n+1} = q^n + \delta q$;
7 $n = n + 1$;

Die Berechnung der Matrix-Vektor-Produkte in Schritt 5 übernimmt der folgende Algorithmus:

Algorithmus 4.4 : Berechnung des Matrix-Vektor-Produktes

input : $u^n, z^n \in X$, die für gegebenes $q^n \in Q$ bereits berechnet wurden

1 Berechne $\delta u^n \in X$, d.h. löse die Tangentengleichung (4.11);
2 Berechne $\delta z^n \in X$, d.h. löse die duale Hesse-Gleichung (4.12);
3 Stelle mittels Satz 4.9 $\nabla^2 j(q^n)\delta q$ auf, d.h., um die i-te Komponente $(\nabla^2 j(q^n)\delta q)_i$ zu bestimmen, werte die rechte Seite von (4.16) für $\tau q = \tau q_i$ aus;

Bemerkung 4.13

- Algorithmus 4.4 muss in jedem Schritt des CG-Verfahrens ausgeführt werden.
- Sei n_{CG} die Anzahl der Schritte, die bei einem Aufruf des CG-Verfahrens ausgeführt werden, so ist es für einen Newton-Schritt insgesamt nötig,
  - einmal die Zustandsgleichung (2.3)
  - einmal das duale Problem (4.10)
  - n_{CG} mal die Tangentengleichung (4.11)
  - n_{CG} mal die duale Hesse-Gleichung (4.12)

  zu lösen.

Um den Aufwand der Algorithmen 4.2 und 4.3 vergleichen zu können, bezeichnen wir mit n_{Newton} die Anzahl der Newton-Schritte, bis die vorgegebene Fehlerschranke unterschritten wurde, und mit n_{CG} die durchschnittliche Anzahl der CG-Iterationen pro Newton-Schritt.
Zur Durchführung von

- Algorithmus 4.2 müssen insgesamt

$$dim(Q) + 2n_{Newton},$$

- Algorithmus 4.3 insgesamt

$$n_{Newton}(2 + 2n_{CG}) = 2n_{Newton} + 2n_{Newton}n_{CG}$$

parabolische Differentialgleichungen gelöst werden. Da dieses Lösen der Differentialgleichungen den wesentlichen Aufwand der beiden Algorithmen darstellt, heißt dies, dass Algorithmus 4.2 bevorzugt werden sollte, falls

$$n_{Newton}n_{CG} > \frac{dim(Q)}{2}$$

gilt.

Wir wollen an dieser Stelle noch betonen, welchen Aufwand beispielsweise das Berechnen der 1. Ableitung $j'(q)$ verursacht. Aus Satz 4.6 wissen wir, dass zunächst die Zustandsgleichung (2.3) und das duale Problem (4.10) gelöst werden müssen.
Der wesentliche Unterschied dieser beiden Probleme besteht neben verschiedenen rechten Seiten und vertauschten Argumenten der Bilinearform a darin, dass die Zustandsgleichung (2.3) ein Vorwärtsproblem beschreibt, also ein Anfangswert

$$u(0) = u_0(q)$$

vorgegeben ist, während das duale Problem (4.10) ein Rückwärtsproblem ist, also ein "Endwert"

$$z(T) = \frac{1}{2}b_2'(u(T), u(T)) + L_2'(u(T))$$

gegeben ist.
Außerdem ist noch zu beachten, dass sowohl die rechte Seite, als auch der Endwert des dualen Problems (4.10) von der Lösung u der Zustandsgleichung 2.3 abhängen. Es muss also zunächst diese Lösung u berechnet und gespeichert werden, bevor das duale Problem gelöst werden kann.

4.6 Besonderheit der Linearität

Da wir uns bei den Problemen, die in dieser Arbeit untersucht werden, auf *lineare* parabolische Differentialgleichungen und *quadratische* Steuerungsfunktionale beschränken, ist die 1. Ableitung des reduzierten Steuerungsfunktionals linear. Das Newton-Verfahren benötigt also nur einen Schritt, um die Lösung der Gleichung

$$j'(q)(\tau q) = 0, \qquad \forall\, \tau q \in Q,$$

zu bestimmen. Diese Tatsache soll in diesem Abschnitt formal gezeigt werden.

Lemma 4.14 *Das Newton-Verfahren zur Lösung des Optimierungsproblems 3.2 liefert nach einem Schritt die exakte Lösung.*

Beweis: Im ersten Schritt des Newton-Verfahrens wird bei gegebenem $q^0 \in Q$ die Lösung $\delta q \in Q$ der Gleichung (4.4)

$$j''(q^0)(\tau q, \delta q) = -j'(q^0)(\tau q), \qquad \forall\, \tau q \in Q,$$

bestimmt. Diese liefert uns $q^1 := q^0 + \delta q$. Damit das Newton-Verfahren schon nach einem Schritt abbricht, muss gelten

$$j'(q^1)(\tau q) = 0, \qquad \forall\, \tau q \in Q. \tag{4.18}$$

Dies wollen wir nun zeigen.

Da wir zwei verschiedene Darstellungen für die zweite Ableitung j'' haben, müssen wir auch zwei verschiedene Fälle untersuchen.

Fall 1: Dieser Fall entspricht Algorithmus 4.2, also der Anwendung von Satz 4.8.

Wir erinnern daran, dass man den Ausdruck für j' mittels der Lösungsoperatoren S und R formulieren kann:

$$\begin{aligned} j'(q)(\tau q) = \alpha\,(q - \bar{q}, \tau q)_Q + \Big(f^{(1)}(\tau q), R(S(q))\Big) \\ + \Big(u_0^{(1)}(\tau q), R(S(q))(0)\Big)_H, \qquad \forall\, \tau q \in Q, \end{aligned} \tag{4.19}$$

und dass für j'' bei festen $\delta q, \tau q \in Q$ gilt:

$$j''(q)(\delta q, \tau q) = \alpha(\tau q, \delta q)_Q + \int_I b_1(\delta u(t), \tau u(t))\, dt + b_2(\delta u(T), \tau u(T)).$$

Aus Gleichung (4.4) ergibt sich damit

$$\begin{aligned}\alpha(\tau q, \delta q) = &- \int_I b_1(\delta u(t), \tau u(t))\, dt - b_2(\delta u(T), \tau u(T)) \\ &- \alpha\left(q^0 - \overline{q}, \tau q\right)_Q - \left(f^{(1)}(\tau q), R(S(q))\right) \\ &- \left(u_0^{(1)}(\tau q), R(S(q))(0)\right)_H . \end{aligned} \tag{4.20}$$

Nun wollen wir $j'(q^1)$ betrachten.
Zuvor halten wir jedoch fest, dass

$$\begin{aligned} R\left(S\left(q^1\right)\right) &= R^{(1)}\left(S\left(q^0 + \delta q\right)\right) + R^{(2)} \\ &= R^{(1)}\left(S\left(q^0\right)\right) + R^{(2)} + R^{(1)}\left(S^{(1)}\left(\delta q\right)\right) \\ &= R\left(S\left(q^0\right)\right) + R^{(1)}\left(S^{(1)}\left(\delta q\right)\right) \\ &=: z_0 + \delta z \end{aligned}$$

gilt.

Setzen wir Gleichung (4.20) in die Darstellung (4.19) für $j'(q^1)(\tau q)$ ein, so erhalten wir

$$\begin{aligned} j'(q^1)(\tau q) &= \alpha\left((q^0 + \delta q) - \overline{q}, \tau q\right)_Q + \left(f^{(1)}(\tau q), R\left(S\left(q^1\right)\right)\right) \\ &\quad + \left(u_0^{(1)}(\tau q), R\left(S\left(q^1\right)\right)(0)\right)_H \\ &= - \int_I b_1(\delta u(t), \tau u(t))\, dt - b_2(\delta u(T), \tau u(T)) \\ &\quad - \left(f^{(1)}(\tau q), z_0\right) - \left(u_0^{(1)}(\tau q), z_0(0)\right)_H \\ &\quad + \left(f^{(1)}(\tau q), z_0 + \delta z\right) + \left(u_0^{(1)}(\tau q), z_0(0) + \delta z(0)\right)_H \\ &= - \int_I b_1(\delta u(t), \tau u(t))\, dt - b_2(\delta u(T), \tau u(T)) \\ &\quad + \left(f^{(1)}(\tau q), \delta z\right) + \left(u_0^{(1)}(\tau q), \delta z(0)\right)_H \\ &= \mathcal{L}''_{zq}()(\delta z, \tau q) - \mathcal{L}''_{uu}()(\delta u, \tau u) \\ &= - \left(\mathcal{L}_{zu}()(\delta z, \tau u) + \mathcal{L}''_{uu}()(\delta u, \tau u)\right) \\ &= 0, \end{aligned}$$

da $\delta z = R^{(1)}\left(S^{(1)}\left(\delta q\right)\right)$ die duale Hesse-Gleichung löst. Die vorletzte Gleichheit folgte dabei aus (ii) in Satz 4.8.

Fall 2: Nun wollen wir Algorithmus 4.3, also die Anwendung von Satz 4.9 betrachten. Hier gilt Gleichung (4.19) und außerdem:

$$j''(q)(\delta q, \tau q) = \alpha(\tau q, \delta q)_Q + (f^{(1)}(\tau q), \delta z) + (u_0^{(1)}(\tau q), \delta z(0))_H, \quad \forall\, \tau q \in Q.$$

Aus der Gleichung

$$j''(q^0)(\tau q, \delta q) = -j'(q^0)(\tau q), \quad \forall\, \tau q \in Q,$$

ergibt sich damit im ersten Newton-Schritt:

$$\begin{aligned}\alpha(\tau q, \delta q)_Q = &-\alpha\left(q^0 - \overline{q}, \tau q\right)_Q - \left(f^{(1)}(\tau q), z_0 + \delta z\right) \\ &- \left(u_0^{(1)}(\tau q), z_0(0) + \delta z(0)\right)_H.\end{aligned}$$

Auch hier setzen wir diesen Ausdruck in $j'(q^1)(\tau q)$ ein und erhalten:

$$\begin{aligned}j'(q^1)(\tau q) &= \alpha\left((q^0 + \delta q) - \overline{q}, \tau q\right)_Q + \left(f^{(1)}(\tau q), R\left(S\left(q^1\right)\right)\right) \\ &\quad + \left(u_0^{(1)}(\tau q), R\left(S\left(q^1\right)\right)(0)\right)_H \\ &= \left(f^{(1)}(\tau q), z_0 + \delta z\right) + \left(u_0^{(1)}(\tau q), z_0(0) + \delta z(0)\right)_H \\ &\quad - \left(f^{(1)}(\tau q), z_0 + \delta z\right) - \left(u_0^{(1)}(\tau q), z_0(0) + \delta z(0)\right)_H \\ &= 0, \qquad \forall\, \tau q \in Q.\end{aligned}$$

In beiden Fällen gilt also, dass wir bereits nach einem Newton-Schritt die gesuchte Lösung q^1 gefunden haben. □

Man hätte natürlich auch sofort einen iterativen Löser auf die lineare Gleichung $\nabla j(q) = 0$ anwenden können. Doch haben wir so den Vorteil, dass der hier gewählte Ansatz ohne Weiteres auf den Fall nichtlinearer parabolischer Differentialgleichungen und nicht-quadratischer Steuerungsfunktionale erweitert werden kann.

Außerdem sind wir bisher davon ausgegangen, dass wir die Lösungen der partiellen Differentialgleichungen exakt bestimmen können. Dies ist jedoch in der Regel nicht der Fall. Wenn wir die Differentialgleichungen aber, wie in Kapitel 5 beschrieben, näherungsweise lösen, erhalten wir aus den Sätzen 4.6, 4.8 und 4.9 auch nur Approximationen an j' und j''. Deshalb kann es auch in dem hier behandelten Fall dazu kommen, dass mehrere Newton-Schritte benötigt werden.

Kapitel 5

Numerische Umsetzung

Die parabolischen Differentialgleichungen, deren Lösungen für die Durchführung des Newton-Verfahrens benötigt werden, können meist nicht exakt gelöst werden. Daher verwendet man numerische Verfahren, um eine Approximation zu berechnen. Eine Möglichkeit hierfür werden wir am Beispiel der Zustandsgleichung (2.3) ausführlich vorstellen und dann auf die in Abschnitt 4.3 vorgestellten Hilfsgleichungen übertragen.

Wir beschränken uns in dieser Arbeit von hieran darauf, den Fall eines linearen elliptischen Operators L der Form

$$(Lu)(x) = -div(A(x)\nabla u(t;x)),$$

beziehungsweise der Bilinearform

$$a(u,\phi) = \int_I \int_\Omega A\nabla u \cdot \nabla\phi \, dx \, dt$$

zu behandeln. Allgemeinere Bilinearformen a bedürfen einer angepassten, spezielleren Untersuchung.

5.1 Diskretisierung der Zustandsgleichung

Zur approximativen Lösung der Zustandsgleichung (2.3)

$$\begin{aligned}(\partial_t u, \phi) + a(u,\phi) &= (f(q),\phi), \quad \forall\, \phi \in X,\\ u(0) &= u_0(q)\end{aligned}$$

diskretisieren wir zunächst das Zeitintervall I. So erhalten wir eine Anzahl von elliptischen Differentialgleichungen, die wir dann mittels Ortsdiskretisierung und Anwendung der Methode der Finiten Elemente numerisch lösen. Wir

werden dabei sowohl für die Semidiskretisierung in der Zeit, als auch für das volldiskrete Problem festhalten, unter welchen Annahmen eine eindeutige Lösung existiert und wann diese gegen die exakte Lösung konvergiert.

5.1.1 Zeitdiskretisierung

Die Idee der Zeitdiskretisierung ist, das Zeitintervall $\bar{I} = [0, T]$ in kleinere Teilintervalle I_m zu zerlegen und für das gesuchte u eine Approximation so zu wählen, dass diese auf den einzelnen Intervallen I_m ein Polynom in der Zeit mit Werten in V ist. Hierbei gibt es zwei Möglichkeiten:

1. Die Polynome werden über die Intervallgrenzen hinweg stetig miteinander verknüpft oder
2. sie dürfen an den Intervallgrenzen Unstetigkeiten haben.

Wir werden im Folgenden die zweite Methode, die auch als *diskontinuierliche Galerkin (dG) Methode* bekannt ist, anwenden und uns letzten Endes auf stückweise konstante Ansatzfunktionen beschränken.

Zunächst benötigen wir folgende

Bezeichnungen 5.1 *Wir zerlegen das Intervall*

$$\bar{I} = [0, T] = \{0\} \cup I_1 \cup I_2 \cup \ldots \cup I_M$$

in M Teilintervalle $I_m = (t_{m-1}, t_m]$ *der Länge* $k_m = t_m - t_{m-1}$, $m = 1, \ldots, M$, *mit* $0 = t_0 < t_1 < \ldots < t_{M-1} < t_M = T$. *Außerdem definieren wir den Parameter k als stückweise konstante Funktion von* $[0, T]$ *nach* $\mathbb{R}$ *mit der Eigenschaft, dass* $k|_{I_m} = k_m$ *gilt, und* $\widehat{k}$ *als das Maximum aller* k_m, *also* $\widehat{k} = \max_{m=1,\ldots,M} k_m$.

Hiermit können wir eine diskrete Version des Ansatzraumes

$$X = \{v \in L^2(I, V) : \partial_t v \in L^2(I, V^*)\}$$

definieren:

Definition 5.2 Es sei

$$X_k^r := \{v_k \in L^2(I, V) : v_k|_{I_m} \in P_r(I_m, V),\ m = 1, \ldots, M,\ v_k(0) \in H\} \tag{5.1}$$

die *Diskretisierung des Ansatzraumes X*, wobei $P_r(I_m, V)$ den Raum der Polynome vom Grad $\leq r$ definiert auf I_m mit Werten in V bezeichnet.

Die Diskretisierung einer Gleichung über dem Raum X_k^r werden wir abkürzend auch als *dG(r)-Diskretisierung* bezeichnen.

Außerdem werden wir noch weitere Notationen verwenden:

Bezeichnungen 5.3

$$v_{k,m}^{+} = \lim_{t \to 0^+} v_k(t_m + t), \qquad v_{k,m}^{-} = \lim_{t \to 0^+} v_k(t_m - t), \qquad [v_k]_m = v_{k,m}^{+} - v_{k,m}^{-}$$

Des Weiteren sei $\widetilde{a}$ die Bilinearform, die wie in (2.7) *die Gleichung*

$$\int_0^T \widetilde{a}(u(t), \phi(t))\, dt = a(u, \phi)$$

für alle $u, \phi \in X$ erfüllt.

Damit lautet die $dG(r)$-Diskretisierung der Zustandsgleichung (2.3):

Gesucht ist $u_k \in X_k^r$, so dass zu gegebener Steuerung $q \in Q$ gilt:

$$\left.\begin{aligned} u_{k,0}^{-} &= u_0(q) \\ \sum_{m=1}^{M} \int_{I_m} \{(\partial_t u_k, \phi)_H + \widetilde{a}(u_k, \phi)\}\, dt + \sum_{m=1}^{M} ([u_k]_{m-1}, \phi_{m-1}^{+})_H & \\ = \sum_{m=1}^{M} \int_{I_m} (f(q), \phi)_H\, dt, \quad \forall\, \phi \in X_k^r. & \end{aligned}\right\} \tag{5.2}$$

Motivieren kann man diese Formulierung durch folgende Herleitung (vgl. [Tho97]):

Die kontinuierliche Formulierung unserer Zustandsgleichung lautet:

Gesucht ist $u \in X$, so dass zu gegebener Steuerung $q \in Q$ gilt:

$$\begin{aligned} (\partial_t u, \phi) + a(u, \phi) &= (f(q), \phi), \quad \forall\, \phi \in X, \\ u(0) &= u_0(q). \end{aligned}$$

Dies lässt sich offensichtlich wegen der Notationen (2.2) und (2.7) äquivalent formulieren als

$$\begin{aligned} \int_0^T \{(\partial_t u, \phi)_H + \widetilde{a}(u, \phi)\}\, dt &= \int_0^T (f(q), \phi)_H\, dt, \quad \forall\, \phi \in X, \\ u(0) &= u_0(q). \end{aligned}$$

Nach partieller Integration erhält man

$$\left.\begin{aligned} u(0) &= u_0(q) \\ \int_0^T \{(u, -\partial_t\phi)_H + \widetilde{a}(u,\phi)\}\, dt + (u(T), \phi(T))_H & \\ = \int_0^T (f(q),\phi)_H\, dt + (u(0),\phi(0))_H, \quad \forall\, \phi \in X. & \end{aligned}\right\} \tag{5.3}$$

Nun ersetzen wir in dieser Gleichung $u \in X$ durch $u_k \in X_k^r$. Betrachten wir zunächst den ersten Term separat und integrieren partiell auf jedem Teilintervall I_m, so erhalten wir:

$$\begin{aligned} \int_0^T (u_k, -\partial_t\phi)_H\, dt &= \sum_{m=1}^{M} \int_{I_m} (u_k, -\partial_t\phi)_H\, dt \\ &= \sum_{m=1}^{M} \int_{I_m} (\partial_t u_k, \phi)_H\, dt - \sum_{m=1}^{M} (u_k,\phi)_H|_{t_{m-1}}^{t_m} \\ &= \sum_{m=1}^{M} \int_{I_m} (\partial_t u_k, \phi)_H\, dt - \sum_{m=1}^{M-1} (u_{k,m}^-, \phi(t_m))_H \\ &\quad - (u_{k,M}^-, \phi(T))_H + (u_{k,0}^+, \phi(0))_H \\ &\quad + \sum_{m=1}^{M-1} (u_{k,m}^+, \phi(t_m))_H \\ &= \sum_{m=1}^{M} \int_{I_m} (\partial_t u_k, \phi)_H\, dt + \sum_{m=1}^{M-1} ([u_k]_m, \phi(t_m))_H \\ &\quad + (u_{k,0}^+, \phi(0))_H - (u_{k,M}^-, \phi(T))_H. \end{aligned}$$

Dies liefert uns bei Einsetzen in (5.3) und der zusätzlichen Wahl von ϕ aus dem Raum X_k^r der semidiskreten Funktionen :

$$\left.\begin{aligned} u_{k,0}^- &= u_0(q) \\ \sum_{m=1}^{M} \int_{I_m} \{(\partial_t u_k, \phi)_H + \widetilde{a}(u_k,\phi)\}\, dt + \sum_{m=1}^{M} ([u_k]_{m-1}, \phi_{m-1}^+)_H & \\ = \int_0^T (f(q),\phi)_H\, dt, \quad \forall\, \phi \in X_k^r. & \end{aligned}\right\} \tag{5.4}$$

Hierbei haben wir uns bei den Funktionsauswertungen von ϕ jeweils für den rechtsseitigen Limes ϕ_m^+ entschieden.

Bemerkung 5.4 Da sowohl die Ansatz- als auch die Testfunktionen aus dem semidiskreten Raum X_k^r gewählt werden, ist es möglich M Testfunktionen $\phi_1, \ldots, \phi_M$ zu wählen, so dass für alle $i \neq j \in \{1, \ldots, M\}$ gilt:

$$\phi_i|_{I_j} = 0.$$

Man kann dann die einzelnen Teilintervalle getrennt betrachten und erhält für die semidiskrete Zustandsgleichung:

Gesucht ist $u_k \in X_k^r$, so dass zu gegebener Steuerung $q \in Q$ gilt:

$$\left.\begin{aligned} u_{k,0}^- &= u_0(q) \\ \int_{I_m} \{(\partial_t u_k, \phi)_H + \widetilde{a}(u_k,\phi)\}\, dt + (u_{k,m-1}^+, \phi_{m-1}^+)_H & \\ = (u_{k,m-1}^-, \phi_{m-1}^+)_H + \int_{I_m} (f(q), \phi)_H\, dt, & \\ \forall\, \phi \in X_k^r,\ 1 \leq m \leq M. & \end{aligned}\right\} \quad (5.5)$$

dG(0)-Diskretisierung

Wir wollen uns nun auf den speziellen Fall der stückweise konstanten Ansatz- und Testfunktionen beschränken, also $u_k \in X_k^0$. Da also $u_k|_{I_m} = \textit{konstant}$ ist, gilt $\partial_t u_k|_{I_m} = 0$. Selbiges gilt für ϕ, welches wir im folgenden Abschnitt immer als Element aus X_k^0 betrachten wollen. Wir nutzen abkürzend die Bezeichnungen $u_m := u_k|_{I_m}$ und $\phi_m := \phi|_{I_m}$.

Man erhält für die Zustandsgleichung zunächst:

$$\begin{aligned} u_{k,0}^- &= u_0(q) \\ \int_{I_m} \widetilde{a}(u_m, \phi)\, dt + (u_m, \phi_m)_H &= (u_{m-1}, \phi_m)_H + \int_{I_m} (f(q), \phi)_H\, dt, \\ &\forall\, \phi \in X_k^0,\ 1 \leq m \leq M. \end{aligned}$$

Approximieren wir nun noch die Integrale durch die rechte Eckpunktregel, also $\int_{t_{m-1}}^{t_m} g(t)\, dt \approx (t_m - t_{m-1}) g(t_m)$, so erhalten wir folgende Iterationsvorschrift für die Lösung der

semidiskreten Zustandsgleichung:

- $m = 0$

$$u_{k,0}^{-} = u_0(q)$$

- $m = 1, \ldots, M$

$$k_m \widetilde{a}(u_m, \phi_m) + (u_m, \phi_m)_H = (u_{m-1}, \phi_m)_H + k_m (f(q)(t_m), \phi_m)_H, \quad \forall\, \phi_m \in V.$$

Dies entspricht M elliptischen Differentialgleichungen, die mittels Ortsdiskretisierung gelöst werden müssen.

Eindeutige Lösbarkeit der semidiskreten Zustandsgleichung

Wir wollen nun zeigen, dass die Gleichungen in (5.5), die auf den einzelnen Zeitintervallen zu lösen sind, eine eindeutige Lösung besitzen. Wir folgen dabei der Beweisidee in [EJT85], S. 618/619.

Satz 5.5 *Es existiert genau eine Lösung $u_k \in X_k^r$ von* (5.5)*:*

$$u_{k,0}^{-} = u_0(q)$$

$$\int_{I_m} \{(\partial_t u_k, \phi)_H + \widetilde{a}(u_k, \phi)\}\, dt + (u_{k,m-1}^{+}, \phi_{m-1}^{+})_H = (u_{k,m-1}^{-}, \phi_{m-1}^{+})_H + \int_{I_m} (f(q), \phi)_H\, dt, \quad \forall\, \phi \in X_k^r,\ 1 \le m \le M.$$

Beweis: Da wir auf jedem Zeitintervall eine eigene Gleichung lösen müssen, zeigen wir die Existenz und Eindeutigkeit von $u_k \in X_k^r$, indem wir beweisen, dass jede Gleichung aus (5.5) ein eindeutige Lösung $u_k^m := u_k|_{I_m} \in P_r(I_m, V)$ besitzt.

Ist $\{p_j\}_{j=1,\ldots,r}$ eine Orthonormalbasis von $P_r(I_m, \mathbb{R})$, also den Polynomen auf I_m mit Werten in $\mathbb{R}$ vom Grad $\le r$, so hat jedes $v \in P_r(I_m, V)$ eine Darstellung als

$$v = \sum_{j=1}^{r} v_j p_j$$

mit $v_j \in V$. Definieren wir $\widehat{v} := (v_1, v_2, \dots, v_r)^T \in V^r =: \widehat{V}$, so können wir v mit $\widehat{v}$ identifizieren.

Nun definieren wir außerdem $\widehat{M}^m := (m_{ij}^m)_{i,j=1,\dots,r}$ und $\widehat{f}^m := (f_i^m)_{i=1,\dots,r}$ mit

$$m_{ij}^m := \int_{I_m} \partial_t p_j p_i \, dt \cdot I + \int_{I_m} p_j p_i \; dt \cdot \widetilde{A} + p_j(t_{m-1}) p_i(t_{m-1}) \cdot I \quad \text{und}$$
$$f_i^m := p_i(t_{m-1}) u_{k,m-1}^- + \int_{I_m} f p_i \, ds,$$

wobei I die Identität auf V bezeichnet und $\widetilde{A}$ der zur Bilinearform $\widetilde{a}$ korrespondierende Operator ist, also $\widetilde{a}(u,v) = (\widetilde{A}u, v)_H$.

Damit ist die $m-$ te Gleichung aus (5.5) äquivalent zur Gleichung

$$[\widehat{M}^m \widehat{y}^m, \widehat{v}] = [\widehat{f}^m, \widehat{v}], \qquad \forall\, \widehat{v} \in \widehat{V},$$

welche wiederum äquivalent ist zu

$$\widehat{M}^m \widehat{y}^m = \widehat{f}^m.$$

Um also zu zeigen, dass (5.5) eine eindeutige Lösung besitzt, zeigen wir, dass $\widehat{M}^m$ für beliebiges $m = 1, \dots, M$ bijektiv ist. Dazu setzen wir an dieser Stelle o.B.d.A. $\widehat{M} := \widehat{M}^m$ für ein beliebiges $m \in \{1, \dots, M\}$.

Injektivität: Es gilt für alle $\widehat{v} \in \widehat{V}$:

$$\begin{aligned}
[\widehat{M}\widehat{v}, \widehat{v}] &= \int_{I_m} (\partial_t v, v)_H \, dt + \int_{I_m} (\widetilde{A}v, v)_H \, dt + (v_{m-1}^+, v_{m-1}^+)_H \\
&= \int_{I_m} \frac{1}{2} \frac{d}{dt} \|v\|_H^2 \, dt + \int_{I_m} \widetilde{a}(v,v) \, dt + \|v_{m-1}^+\|_H^2 \\
&= \frac{1}{2}\|v_m^-\|_H^2 + \frac{1}{2}\|v_{m-1}^+\|_H^2 + \int_{I_m} \widetilde{a}(v,v) \, dt \\
&\geq \int_{I_m} \widetilde{a}(v,v) \, dt \\
&\geq c \int_{I_m} \|v\|_H^2 \, dt \qquad \text{(da } \widetilde{a} \text{ strikt positiv ist)} \\
&= c \int_{I_m} \| \sum_{j=1}^r v_j p_j \|_H^2 \, dt \\
&= c \int_{I_m} \int_\Omega (\sum_{i=1}^r \sum_{j=1}^r v_i v_j p_i p_j) \, dx \, dt \qquad \text{(wegen } p_i p_j = \delta_{ij})
\end{aligned}$$

$$= c \int_{I_m} \int_{\Omega} \sum_{i=1}^{r} v_i^2 \, dx \, dt$$
$$= \widehat{c} \, \|[\widehat{v}]\|^2,$$

wobei $\|[\cdot]\|$ die Norm auf $\widehat{V}$ bezeichnet.

Ist nun $\widehat{v} \in Ker(\widehat{M})$, gilt also $\widehat{M}\widehat{v} = 0$, so ist

$$0 = [0, \widehat{v}] = [\widehat{M}\widehat{v}, \widehat{v}] \geq \widehat{c} \, \|[\widehat{v}]\|^2.$$

Es muss also $\|[\widehat{v}]\| = 0$ gelten und damit auch $\widehat{v} = 0$. Also ist $\widehat{M}$ injektiv.

Surjektivität: Da $\widehat{M}$ injektiv ist, ist auch der adjungierte Operator $\widehat{M}^*$ injektiv, da nach Definition von $\widehat{M}^*$

$$[\widehat{v}, \widehat{M}^*\widehat{v}] = [\widehat{M}\widehat{v}, \widehat{v}] \geq \widehat{c} \, \|[\widehat{v}]\|^2$$

gilt. Aus Lemma 5.11 in [RY00] folgt damit, dass das Bild von $\widehat{M}$ dicht liegt in $\widehat{V}$.

Wir beobachten weiterhin , dass $\widetilde{A}$ eine abgeschlossene Abbildung ist, und sich diese Eigenschaft nach Konstruktion von $\widetilde{A}$ auf $\widehat{M}$ überträgt. Betrachten wir nun die Inverse $\widehat{M}^{-1} : \widehat{V} \longrightarrow Im(\widehat{M})$ und setzen in der obigen Ungleichung $v = \widehat{M}^{-1}w$ mit $w \in Im(\widehat{M})$, so gilt mit der Cauchy-Schwarzschen Ungleichung

$$\|[\widehat{M}^{-1}w]\|^2 \leq [w, \widehat{M}^{-1}w] \leq c\|[w]\| \cdot \|[\widehat{M}^{-1}w]\|.$$

$\widehat{M}^{-1}$ ist also beschränkt und als linearer Operator damit auch stetig. Aus den Sätzen 39.1 und 39.3 in [Heu86] folgt damit, dass das Bild von $\widehat{M}$ abgeschlossen ist.

Wir haben also gezeigt, dass das Bild $Im(\widehat{M})$ eine dichte und abgeschlossene Teilmenge von $\widehat{V}$ ist. Es gilt also insgesamt $Im(\widehat{M}) = \widehat{V}$ und damit ist $\widehat{M}$ auch surjektiv. □

Konvergenz der semidiskreten Zustandsgleichung

In diesem Abschnitt wollen wir zeigen, dass die Lösung der semidiskreten Zustandsgleichung (5.5) für $k \to 0$ gegen die exakte Lösung der Zustandsgleichung (2.3) konvergiert.

Zunächst benötigen wir eine Definition und einige Hilfsaussagen, die wir an dieser Stelle allerdings nur zitieren, nicht aber beweisen werden.

Definition 5.6 Wir definieren auf V in Abhängigkeit des elliptischen Operators L für $l \in \mathbb{R}$

$$\|v\|_l := \left(\sum_{j=1}^{\infty} \lambda_j^{2l} (v, \varphi_j)^2 \right)^{1/2},$$

wobei $\{\lambda_j\}_{j=1,\dots,\infty}$ die positiven Eigenwerte von L sind und $\{\varphi_j\}_{j=1,\dots,\infty}$ die zugehörigen Eigenvektoren.

Da nach Bezeichnungen 2.18 $H = L^2(\Omega)$ gilt, können wir beobachten, dass $\|\cdot\|_0 = \|\cdot\|_H$ der Norm auf H entspricht.

Lemma 5.7 *Ist $u \in X$ die Lösung der Zustandsgleichung* (2.3) *und $j \in N$. Dann gilt für alle $t \in I$ folgende Ungleichung:*

$$\|u^{(j)}(t)\|_l + \left(\int_0^t \|u^{(j)}(s)\|_{l+1/2}^2 \, ds \right)^{1/2} \leq C \left(\|u^{(j)}(0)\|_l + \int_0^t \|f(q)^{(j)}(s)\|_l \, ds \right)$$

mit einer Konstanten $C > 0$.

Beweis: Da $f(q) \in L^2(I, V^*)$ nach Gleichung (2.6) gilt, folgt dies direkt aus [EJT85], Lemma 2. □

Lemma 5.8 *Sei $\widetilde{u} \in P_r(I_m, V)$ eine Interpolation der Lösung u der Zustandsgleichung* (2.3) *auf dem Intervall I_m definiert durch $\widetilde{u}|_{I_m} := u(t_m)$. Dann existiert eine Konstante C, abhängig nur vom Grad r der Polynome auf I_m, so dass für $\zeta := u - \widetilde{u}$ und $j = 0, \dots, r$ gilt:*

$$\sup_{s \in I_m} \|\zeta(s)\|_l \leq C k_m^j \int_{I_m} \|u^{(j+1)}(s)\|_l \, ds.$$

Beweis: Ein Beweis hierzu findet sich in [EJT85], Lemma 5. □

Lemma 5.9 *Sei $u \in X$ Lösung der Zustandsgleichung* (2.3), *$\widetilde{u}$ die Interpolation aus Lemma 5.8 und $u_k \in X_k^r$ Lösung der semidiskreten Zustandsgleichung* (5.5).
Dann gilt für $\zeta := u - \widetilde{u}$ und $\theta := \widetilde{u} - u_k$

$$\|\theta_m\|_H = \|\theta(t_m)\|_H \leq \left(\int_0^{t_m} \|\zeta(s)\|_{1/2}^2 \, ds \right)^{1/2}.$$

Beweis: Vergleiche hierzu [EJT85], Beweis zu Theorem 1. □

Nun können wir die Hauptaussage dieses Abschnitts festhalten. Hierbei interessieren wir uns vor allem für den Spezialfall $r = 0$, den wir auch beweisen werden.

Satz 5.10 *Ist $u \in X$ die Lösung der Zustandsgleichung* (2.3) *und $u_k \in X_k^r$ Lösung der entsprechenden semidiskreten Zustandsgleichung* (5.5)*, so konvergiert u_k für $k \to 0$ gegen u gemäß der a priori Abschätzung*

$$\|u_k(t) - u(t)\|_H \leq C\widehat{k}^{r+1} \left\{ \|u^{(r+1)}(0)\|_H + \|f(q)^{(r)}(0)\|_H + \int_0^{t_m} \|f(q)^{(r+1)}(s)\|_H \, ds \right\}$$

für $0 \leq t \leq t_m$ mit $\widehat{k} = max_{1,\ldots,m} k_m$ wie in Bezeichnungen 5.1.
Für $r = 0$ gilt insbesondere:

$$\|u_k(t) - u(t)\|_H \leq C\widehat{k} \left\{ \|u'(0)\|_H + \int_0^{t_m} \|f(q)'(s)\|_H \, ds \right\}.$$

Beweis: Wir wollen an dieser Stelle, wie bereits erwähnt, nur die Aussage für $r = 0$ beweisen, da wir uns auch im Folgenden auf diesen Fall beschränken werden. Die allgemeine Behauptung folgt mit $f(q) \in L^2(I, V^*)$ und $r = q - 1$ direkt aus [EJT85], Theorem 4.

Sei $\widetilde{u}$ die Interpolation aus Lemma 5.8. Die Differenz $u - u_k$ teilen wir auf in

$$u - u_k = (u - \widetilde{u}) + (\widetilde{u} - u_k) =: \zeta + \theta.$$

Aus Lemma 5.8 wissen wir bereits, dass wir $\|\zeta(t)\|_H$ für $t \in I_m$ abschätzen können durch

$$\begin{aligned} \|\zeta(t)\|_H &\leq \sup_{s \in I_m} \|\zeta(s)\|_H \\ &\leq C \int_{I_m} \|u'(s)\|_H \, ds \\ &\leq C\widehat{k} \sup_{s \in I_m} \|u'(s)\|_H. \end{aligned} \tag{5.6}$$

Um eine Aussage über $\|\theta(t)\|_H$ mit $t \in I_m$ herzuleiten, genügt es $\theta_m = \theta(t_m)$ zu betrachten, da $\theta \in X_k^0$ auf I_m konstant ist. Mit Lemma 5.8, Lemma 5.9 und

der Hölderschen Ungleichung folgt:

$$\begin{aligned}
\|\theta_m\|_H^2 &\leq \int_0^{t_m} \|\zeta(s)\|_{1/2}^2 \, ds = \sum_{I_n, n \leq m} \int_{I_n} \|\zeta(s)\|_{1/2}^2 \, ds \\
&\leq \sum_{I_n, n \leq m} k_n \sup_{s \in I_n} \|\zeta(s)\|_{1/2}^2 \\
&\leq \sum_{I_n, n \leq m} k_n \left(C \int_{I_n} \|u'(s)\|_{1/2} \, ds \right)^2 \\
&\leq C \sum_{I_n} k_n \cdot k_n \int_{I_m} \|u'(s)\|_{1/2}^2 \, ds \\
&\leq C \widehat{k}^2 \int_0^{t_n} \|u'(s)\|_{1/2}^2 \, ds.
\end{aligned}$$

Wir erhalten also für $t \in I_m$

$$\|\theta(t)\|_H = \|\theta(t_m)\|_H = \|\theta_m\|_H \leq C\widehat{k} \left(\int_0^{t_n} \|u'(s)\|_{1/2}^2 \, ds \right)^{1/2}. \tag{5.7}$$

Lemma 5.7 und die soeben hergeleiteten Ungleichungen (5.6) und (5.7) liefern uns nun zusammen mit der Dreiecksungleichung die gewünschte Aussage

$$\begin{aligned}
\|u(t) - u_k(t)\|_H &\leq \|\zeta(t)\|_H + \|\theta(t)\|_H \\
&\leq C\widehat{k} \sup_{s \in I_m} \|u'(s)\|_H + Ck \left(\int_0^{t_n} \|u'(s)\|_{1/2}^2 \, ds \right)^{1/2} \\
&\leq C\widehat{k} \left(\|u'(0)\|_H + \int_0^t \|f(q)'(s)\|_H \, ds \right). \qquad \square
\end{aligned}$$

5.1.2 Ortsdiskretisierung

Nachdem wir im vorigen Abschnitt gesehen haben, dass im Zuge der dG(0) Diskretisierung der Zustandsgleichungen (2.3) jeweils $M - 1$ elliptische Differentialgleichungen entstehen, wollen wir nun kurz auf die approximative Lösung dieser Gleichungen eingehen. Für eine ausführliche Behandlung verweisen wir an dieser Stelle beispielsweise auf [Lub06], Teil II.

Wir wollen die entstehenden elliptischen Differentialgleichungen mittels der Methode der Finiten Elemente lösen. Dafür muss zunächst das Gebiet $\Omega \subset \mathbb{R}^n$ diskretisiert werden. Wir bezeichnen hierfür mit $\mathcal{T}_h$ ein Gitter, welches aus den

Zellen D_i, $i = 1, \ldots, N$, besteht, so dass $\Omega = \cup_{i=1}^{N} D_i$ und $D_i \cap D_j = \emptyset$ für $i \neq j$. Dabei sei der Parameter $h : \Omega \longrightarrow \mathbb{R}$ definiert durch $h|_{D_i} = diam\ D_i$.

Auf diesem Gitter konstruieren wir nun den Finite-Elemente-Raum $V_h^s \subset V$:

$$V_h^s := \{v \in V : v|_D \in \widetilde{Q}_s(D) \text{ für } D \in \mathcal{T}_h\}.$$

Dabei ergibt sich $\widetilde{Q}_s(D)$ wie folgt:
Ist $\widehat{D} = (0,1)^n$ eine Referenzzelle, $Q_s(\widehat{D})$ der Raum der Funktionen vom Ansatzgrad s auf $\widehat{D}$ und $F : D \to \widehat{D}$ die affin lineare Transformation eines Elementes $D \in \mathcal{T}_h$ auf das Referenzelement $\widehat{D}$, so ist

$$\widetilde{Q}_s(D) := \{v : D \to \mathbb{R} : v(x) = u(F(x)),\ \forall\, x \in D, u \in Q_s(\widehat{D})\}.$$

Wir werden bei der Behandlung von Adaptivität in Kapitel 6 *dynamische Gitter* benötigen, d.h. wir werden auf jedem Zeitintervall I_m ein eigenes Gitter verwenden. Aus diesem Grund bezeichnen wir mit $\mathcal{T}_{h,m}$, $m = 1, \ldots, M$ das Gitter auf dem Intervall I_m. Entsprechend sei $V_{h,m}^s$ der Finite-Elemente-Raum auf I_m.

Hiermit können wir unseren volldiskreten Raum $X_{k,h}^{r,s}$ definieren durch:

$$X_{k,h}^{r,s} := \{v_{kh} \in L^2(I, V) : v_{kh}|_{I_m} \in P_r(I_m, V_{h,m}^s), m = 1, \ldots, M, v_{kh}(0) \in V_{h,0}^s\}.$$

Unter Verwendung der dG(0) Diskretisierung in der Zeit formulieren wir damit die
volldiskrete Zustandsgleichung

Gesucht ist $u_{kh} \in X_{k,h}^{0,s}$, so dass zu gegebener Steuerung $q \in Q$ gilt:

$$\left.\begin{aligned} u_{kh,0}^{-} &= u_0(q) \\ k_m \widetilde{a}(U_m, \Phi_m) + (U_m, \Phi_m)_H &= (U_{m-1}, \Phi_m)_H + k_m (f(q)(t_m), \Phi_m)_H, \\ \forall\, \Phi_m \in V_{h,m}^s, &\qquad m = 1, \ldots, M. \end{aligned}\right\} \quad (5.8)$$

formulieren. Dabei haben wir an dieser Stelle die Bezeichnung $U_m := u_{kh}|_{I_m}$ verwendet.

Bemerkung 5.11 In der Praxis taucht bei der Berechnung der rechten Seite von (5.8) ein Problem auf: Im Skalarprodukt $(U_{m-1}, \Phi_m)_H$, das einer Integralauswertung entspricht, ist u_{m-1} ein Element aus $V_{h,m-1}^s$, während Φ_m aus $V_{h,m}^s$ stammt. Deshalb ist es nicht möglich das Integral zellenweise zu berechnen. Eine Lösung für dieses Problem wird in [SV06], Abschnitt 2.3 vorgestellt.

Analyse der volldiskreten Zustandsgleichung

In diesem Abschnitt wollen wir zeigen, dass auch das volldiskrete Problem eindeutig lösbar ist und gegen die exakte Lösung der Zustandsgleichung konvergiert.

Zunächst benötigen wir jedoch noch eine Definition, um alle Voraussetzungen formulieren zu können.

Definition 5.12 Für ein Gebiet D sei ρ^D der Radius einer größten Kugel, die in D enthalten ist, und h^D der Radius einer kleinsten Kugel, die D enthält. Eine Familie $\{\mathcal{T}_h\}$ von Zerlegungen eines Gebiets Ω heißt *quasi-uniform*, falls es eine von h unabhängige Konstante $\kappa > 0$ gibt mit

$$\sup_{D \in \mathcal{T}_h} \frac{h^D}{\rho^D} \leq \kappa.$$

Wir wollen außerdem an die Definition der Seminorm $|\cdot|_{H^l(\Omega)}$ und der Standardnorm $\|\cdot\|_{H^l(\Omega)}$ auf dem Sobolevraum $H^l(\Omega)$ erinnern. Diese sind für $v \in H^l(\Omega)$ definiert als

$$|v|_{H^l(\Omega)} := \left(\sum_{|\alpha|=l} \|D^\alpha v\|^2_{L^2(\Omega)} \right)^{1/2}$$

und

$$\|v\|_{H^l(\Omega)} := \left(\sum_{|\alpha|\leq l} \|D^\alpha v\|^2_{L^2(\Omega)} \right)^{1/2}.$$

Weiterhin erinnern wir an dieser Stelle an das Lemma von *Lax-Milgram*:

Lemma 5.13 *Seien X ein Hilbertraum, $f : X \to \mathbb{R}$ eine stetige Linearform und $b : X \times X \to \mathbb{R}$ eine stetige, strikt-positive Bilinearform. Dann existiert genau eine Lösung $u \in X$ der Variationsgleichung*

$$b(u, \phi) = f(\phi), \qquad \forall\, \phi \in X.$$

Beweis: Ein Beweis hierzu findet sich in [Lub06], Satz 6.4. □

Jetzt können wir folgenden Satz festhalten:

Satz 5.14 *Unter der Voraussetzung, dass Ω ein polyedrisches, konvexes Gebiet ist und $\{\mathcal{T}_h\}$ eine Familie von quasi-uniformen Zerlegungen, gilt:*

1. *Es existiert genau eine Lösung $u_{kh} \in X_{k,h}^{0,s}$ des volldiskreten Problems* (5.8).
2. *Ist $u \in X$ die Lösung der Zustandsgleichung* (2.3), *$u_k \in X_k^r$ Lösung der semidiskreten Zustandsgleichung* (5.2) *mit $r = 0$ und $u_{kh} \in X_{k,h}^{0,s}$ Lösung des zugehörigen volldiskreten Problems* (5.8), *so konvergiert u_{kh} für $k \to 0$, $h \to 0$ gegen u gemäß der a priori Abschätzung*

$$\|u(t) - u_{kh}(t)\|_H \leq C \left(\widehat{k} \left\{ \|u'(0)\|_H + \int_0^{t_m} \|f(q)'(s)\|_H \, ds \right\} + h^2 \left\{ \sum_{i=1}^{m} \|f(q)(t_i)\|_H + \|u_0(q)\|_H \right\} \right)$$

für $0 \leq t \leq t_m$ mit $\widehat{k} = max_{1,\ldots,m} k_m$ wie in Bezeichnungen 5.1.

Beweis: Da wir aus Satz 5.10 bereits wissen, dass die Lösung u_k des semidiskreten Problems für $k \to 0$ gegen die exakte Lösung u konvergiert, reicht es zu zeigen, dass die Lösung u_{kh} des volldiskreten Problems für $h \to 0$ gegen u_k konvergiert und dass die einzelnen Gleichungen des volldiskreten Problems eine eindeutige Lösung besitzen.

Wir betrachten nun für ein festes aber beliebiges $n \in \{1, 2, \ldots, M\}$ eine einzelne elliptische Gleichung aus (5.2) und zeigen, dass die Lösung deren diskreter Version aus (5.8) existiert, eindeutig bestimmt ist und gegen die exakte Lösung dieser Gleichung konvergiert. Damit erhalten wir insgesamt die gewünschte Konvergenz $u_{kh} \to u_k$ für $h \to 0$.

Da $b(v, w) = k_n \widetilde{a}(v, w) + (v, w)_H$ eine stetige, strikt positive Bilinearform auf V ist und $V_h^s \subset V$ gilt, ist das Lemma 5.13 von Lax-Milgram anwendbar. Dieses liefert uns die eindeutige Lösbarkeit von

$$k_n \widetilde{a}(U_n, \Phi_n) + (U_n, \Phi_n)_H = (U_{n-1}, \Phi_n)_H + k_n (f(q)(t_n), \Phi_n)_H, \quad \forall\, \Phi \in V_h^s,$$

mit $U_n = u_{kh}|_{I_n} \in V_h^s$.
Bezeichnen wir außerdem mit $u_n := u_k|_{I_n} \in V$ die Lösung der kontinuierlichen Formulierung dieser Gleichung:

$$k_n \widetilde{a}(u_n, \phi_n) + (u_n, \phi_n)_H = (u_{n-1}, \phi_n)_H + k_n (f(q)(t_n), \phi_n)_H, \quad \forall\, \phi_n \in V,$$

so liefert uns Satz 10.14 in [Lub06] außerdem die Konvergenz von U_n gegen u_n gemäß

$$\|U_n - u_n\|_{L^2(\Omega)} \leq C h^2 |u_n|_{H^2(\Omega)}.$$

Die Voraussetzungen für die Anwendung dieses Satzes sind hierbei nach dem Regularitätssatz 7.2 in [Bra97] erfüllt. Nach Definition der Seminorm $|\cdot|_{H^2(\Omega)}$ gilt weiterhin die Abschätzung

$$|u_n|_{H^2(\Omega)} \leq \|u_n\|_{H^2(\Omega)}$$

und aus dem oben genannten Regularitätssatz folgt zusammen mit der Dreiecksungleichung, dass

$$\begin{aligned}\|u_n\|_{H^2(\Omega)} &\leq c\|k_n f(q)(t_n) + u_{n-1}\|_{L^2(\Omega)} \\ &\leq c\left(k_n\|f(q)(t_n)\|_{L^2(\Omega)} + \|u_{n-1}\|_{L^2(\Omega)}\right)\end{aligned}$$

gilt. Wegen $\|u_{n-1}\|_{L^2(\Omega)} \leq \|u_{n-1}\|_{H^2(\Omega)}$ erhält man nach derselben Argumentation

$$\|u_{n-1}\|_{L^2(\Omega)} \leq c\left(k_{n-1}\|f(q)(t_{n-1})\|_{L^2(\Omega)} + \|u_{n-2}\|_{L^2(\Omega)}\right).$$

Rekursiv ergibt sich damit:

$$\begin{aligned}\|U_n - u_n\|_{L^2(\Omega)} &\leq Ch^2\|u_n\|_{H^2(\Omega)} \\ &\leq Ch^2\left(\sum_{i=1}^{n} k_i\|f(q)(t_i)\|_{L^2(\Omega)} + \|u_0(q)\|_{L^2(\Omega)}\right).\end{aligned}$$

Da dies für alle $n \in \{1, 2, \ldots, M\}$ gilt, liefert uns nach Nulladdition die Dreiecksungleichung zusammen mit Satz 5.10 für $r = 0$ und $0 \leq t \leq t_m$:

$$\begin{aligned}\|u(t) - u_{kh}(t)\|_H &\leq \|u(t) - u_k(t)\|_H + \|u_k(t) - u_{kh}(t)\|_H \\ &\leq C\left(\widehat{k}\left\{\|u'(0)\|_H + \int_0^{t_m}\|f(q)'(s)\|_H\,ds\right\}\right. \\ &\quad \left. + h^2\left\{\sum_{i=1}^{m}\|f(q)(t_i)\|_H + \|u_0(q)\|_H\right\}\right).\end{aligned}$$

Insgesamt konvergiert u_{kh} für $k \to 0$ und $h \to 0$ also gegen die gesuchte Lösung u. □

5.2 Diskretisierung des Optimierungsproblems

Nachdem wir uns im vorigen Abschnitt der Diskretisierung der Zustandsgleichung (2.3) gewidmet haben, können wir nun das **semidiskrete Optimierungsproblem** aufstellen:

Minimiere $J(q_k, u_k)$ unter der Bedingung (5.4) mit $(q_k, u_k) \in Q \times X_k^r$. (5.9)

Um die Ergebnisse der früheren Kapitel anwenden zu können, müssen wir ein diskretisiertes reduziertes Steuerungsfunktional, ein diskretes Lagrange-Funktional und die entsprechenden diskretisierten Hilfsgleichungen einführen.

Analog zum kontinuierlichen Fall sei $S_k : Q \longrightarrow X_k^r,\ q \mapsto u_k$ der Lösungsoperator zu (5.4) und entsprechend

$$\begin{aligned} j_k : Q &\longrightarrow \mathbb{R} \\ q_k &\longmapsto j_k(q_k) = J(q_k, S_k(q)) \end{aligned}$$

das *reduzierte semidiskrete Steuerungsfunktional*.

Das *semidiskrete Lagrange-Funktional* hat entsprechend die Form:

$$\begin{aligned} \mathcal{L}_k : Q \times X_k^r \times X_k^r &\longrightarrow \mathbb{R} \\ (q_k, u_k, z_k) &\longmapsto J(q_k, u_k) + \sum_{m=1}^{M} \int_{I_m} \{(f(q_k) - \partial_t u_k, z_k)_H - \widetilde{a}(u_k, z_k)\}\, dt \\ &\quad - \sum_{m=1}^{M} ([u_k]_{m-1}, z_{k,m-1}^+)_H - (u_{k,0}^- - u_0(q_k), z_{k,0}^+)_H. \end{aligned}$$

Hiermit können wir nun die drei Hilfsgleichungen aus Abschnitt 4.3 in semidiskreter Form aufstellen:

semidiskretes duales Problem: Zu gegebenem $q_k \in Q$ und $u_k = S_k(q_k) \in X_k^r$ finde $z_k \in X_k^r$, so dass gilt:

$$\mathcal{L}'_{k,u_k}(q_k, u_k, z_k)(\phi) = 0, \quad \forall\, \phi \in X_k^r,$$

$$\Leftrightarrow$$

$$\sum_{m=1}^{M} \int_{I_m} \{(-\phi, \partial_t z_k)_H + \widetilde{a}(\phi, z_k)\}\, dt - \sum_{m=1}^{M-1} (\phi_m^-, [z_k]_m)_H + (\phi_M^-, z_{k,M}^-)_H =$$

$$\sum_{m=1}^{M} \int_{I_m} \{b_1(\phi(t), u_k(t)) + L_1(\phi(t))\}\, dt + b_2(\phi_M^-, u_{k,M}^-) + L_2(\phi_M^-), \quad \forall\, \phi \in X_k^r.$$

semidiskrete Tangentengleichung: Zu gegebenem $\delta q_k \in Q$ finde $\delta u_k \in X_k^r$, so dass gilt:

$$\begin{gathered}
\mathcal{L}''_{k,qz_k}(q_k,u_k,z_k)(\delta q_k,\phi)+\mathcal{L}''_{k,u_kz_k}(q_k,u_k,z_k)(\delta u_k,\phi)=0, \quad \forall\,\phi\in X_r,\\
\Leftrightarrow\\
\delta u_{k,0}^- = u_0^{(1)}(\delta q_k)\\
\sum_{m=1}^{M}\int_{I_m}\{(\partial_t\delta u_k,\phi)_H+\widetilde{a}(\delta u_k,\phi)\}\,dt+\sum_{m=1}^{M}([\delta u_k]_{m-1},\phi_{m-1}^+)_H=\\
\sum_{m=1}^{M}\int_{I_m}(f^{(1)}(\delta q_k),\phi)_H\,dt, \quad \forall\,\phi\in X_k^r.
\end{gathered} \tag{5.10}$$

semidiskrete Duale Hesse-Gleichung: Zu gegebenem $\delta q_k \in Q$ sei $\delta u_k \in X_k^r$ die Lösung der semidiskreten Tangentengleichung (5.10). Gesucht ist $\delta z_k \in X_k^r$, so dass gilt:

$$\begin{gathered}
\mathcal{L}''_{k,u_ku_k}(q_k,u_k,z_k)(\delta u_k,\phi)+\mathcal{L}''_{k,z_ku_k}(q_k,u_k,z_k)(\delta z_k,\phi)=0, \quad \forall\,\phi\in X_k^r,\\
\Leftrightarrow\\
\sum_{m=1}^{M}\int_{I_m}\{(-\phi,\partial_t\delta z_k)_H+\widetilde{a}(\phi,\delta z_k)\}\,dt-\sum_{m=1}^{M-1}(\phi_m^-,[\delta z_k]_m)_H+(\phi_M^-,\delta z_{k_M}^-)_H=\\
\sum_{m=1}^{M}\int_{I_m}b_1(\delta u_k(t),\phi(t))\,dt+b_2(\delta u_{k,M}^-,\phi_M^-), \quad \forall\,\phi\in X_k^r.
\end{gathered}$$

Analog zur Zustandsgleichung diskretisieren wir nun auch noch im Ort und betrachten den Spezialfall der dG(0)-Diskretisierung:

Das **volldiskrete Optimierungsproblem** lautet:

$$\text{Minimiere } J(q_{kh},u_{kh}) \text{ unter der Bedingung (5.8) mit } (q_{kh},u_{kh})\in Q\times X_{k,h}^{r,s}. \tag{5.11}$$

Die Formulierungen für das entsprechende reduzierte Steuerungsfunktional, Lagrangefunktional und die drei Hilfsgleichungen übertragen sich aus der Zeitdiskretisierung, indem wir jeweils den Raum X_k^r auf $X_{k,h}^{r,s}$ einschränken. Dabei müssen wir im Gegensatz zur Zeitdiskretisierung das Lagrange-Funktional nicht weiter anpassen, da durch die Ortsdiskretisierung keine weiteren Terme, wie die Sprungterme an den Intervallgrenzen bei der Zeitdiskretisierung, hinzukommen.

Insgesamt erhalten wir also mit den Bezeichnungen $U_m := u_{kh}|_{I_m}$, $Z_m := z_{kh}|_{I_m}$, $\delta U_m := \delta u_{kh}|_{I_m}$ und $\delta Z_m := \delta z_{kh}|_{I_m}$ folgende Iterationsvorschriften für die drei Hilfsgleichungen:
Es gelte für alle $\Phi_m \in V_h^s$, $m = 1, \ldots, M$:

volldiskretes duales Problem: Zu gegebenem $q_{kh} \in Q$ und $u_{kh} \in X_{k,h}^{0,s}$ finde $z_{kh} \in X_{k,h}^{0,s}$, so dass für

- $m = M$

$$k_m \widetilde{a}(Z_M, \Phi_M) + (\Phi_M, Z_M)_H = k_M(b_1(\Phi_M, U_M) + L_1(\Phi_M)) + b_2(\Phi_M, U_M) + L_2(\Phi_M)$$

- $m = M-1, \ldots, 1$

$$k_m \widetilde{a}(\Phi_m, Z_m) + (\Phi_m, Z_m)_H = (\Phi_m, Z_{m+1})_H + k_m(b_1(\Phi_m, U_m) + L_1(\Phi_m))$$

- $m = 0$

$$z_{kh,0}^- = Z_1$$

gilt.

volldiskrete Tangentengleichung: Zu gegebenem $\delta q_{kh} \in Q$ finde $\delta u_{kh} \in X_{k,h}^{0,s}$, so dass für

- $m = 0$

$$\delta u_{kh,0}^- = u_0^{(1)}(\delta q_{kh})$$

- $m = 1, \ldots, M$

$$k_m \widetilde{a}(\delta U_m, \Phi_m) + (\delta U_m, \Phi_m)_H = (\delta U_{m-1}, \Phi_m)_H + k_m(f^{(1)}(\delta q_{kh})(t_m), \Phi_m)_H$$

gilt.

volldiskrete Duale Hesse-Gleichung: Zu gegebenem $\delta q_{kh} \in Q$ sei $\delta u_{kh} \in X_{k,h}^{0,s}$ die Lösung der volldiskreten Tangentengleichung. Gesucht ist $\delta z_{kh} \in X_{k,h}^{0,s}$, so dass für

- $m = M$

$$k_m \widetilde{a}(\Phi_M, \delta Z_M) + (\Phi_M, \delta Z_M)_H = k_M b_1(\Phi_M, \delta U_M) + b_2(\Phi_M, \delta U_M)$$

- $m = M-1, \ldots, 1$

$$k_m \widetilde{a}(\Phi_m, \delta Z_m) + (\Phi_m, \delta Z_m)_H = (\Phi_m, \delta Z_{m+1})_H + k_m b_1(\Phi_m, \delta U_m)$$

- $m = 0$

$$\delta z^-_{kh,0} = \delta Z_1$$

gilt.

Kapitel 6

A posteriori Fehlerschätzer und Adaptivität

In diesem Kapitel verfolgen wir die Idee aus [MV07], um Aussagen über den Fehler

$$J(q,u) - J(q_{kh},u_{kh}) \tag{6.1}$$

zu treffen. Dabei ist J das Steuerungsfunktional aus Definition 3.19 und $u_{kh} \in X$ die Lösung der volldiskreten Zustangsgleichung (5.8) zu $q_{kh} \in Q$. Zunächst werden wir eine exakte Fehlerdarstellung herleiten. Um den Fehler für die dG(0)cG(1) Diskretisierung berechnen zu können, approximieren wir dann diese allgemeine Darstellung. Daraufhin werden wir die Fehlerschätzung verwenden, um die Diskretisierungen der partiellen Differentialgleichungen dem Problem anzupassen. Dabei ist zu beachten, dass nicht die Minimierung des Fehler $u - u_{kh}$ unser Ziel ist, sondern dass $J(q_{kh},u_{kh})$ möglichst gut $J(q,u)$ approximieren soll.

6.1 Allgemeine Fehlerdarstellung

Als erstes werden wir ein allgemeines Resultat zu Fehlerdarstellungen festhalten (siehe [BR01], Abschnitt 2.1), das wir dann auf Gleichung (6.1) übertragen werden.

Seien Y ein Funktionenraum, $L(\cdot)$ ein Fréchet-differenzierbares Funktional auf Y und $y \in Y$ ein stationärer Punkt von L auf Y, also

$$L'(y)(\widehat{y}) = 0, \qquad \forall\, \widehat{y} \in\ Y. \tag{6.2}$$

Sei $Y_0 \subset Y$ ein endlich dimensionaler Unterraum, in dem mittels einer Galerkin-Methode die Lösung $y \in Y$ von Gleichung (6.2) durch $y_0 \in Y_0$ approximiert wird, d.h. wir betrachten ergänzend das diskrete Problem

$$L'(y_0)(\widehat{y}_0) = 0, \qquad \forall\, \widehat{y}_0 \in\ Y_0. \tag{6.3}$$

Der folgende Satz liefert uns eine Darstellung für den Fehler $L(y) - L(y_0)$.

Satz 6.1 *Für beliebiges $\widehat{y}_0 \in Y_0$ gilt folgende* a posteriori Fehlerdarstellung*:*

$$L(y) - L(y_0) = \frac{1}{2} L'(y_0)(y - \widehat{y}_0) + R \tag{6.4}$$

mit dem Restterm

$$R := \frac{1}{2} \int_0^1 L'''(y_0 + se)(e,e,e) \cdot s \cdot (s-1)\, ds$$

und dem Fehler $e := y - y_0$.

Beweis: Im Folgenden gelte die Notation $L'(\overline{yy_0})(\widehat{y}) := \int_0^1 L'(y_0 + se)(\widehat{y}) ds$, also insbesondere mit der Kettenregel aus Lemma 3.9

$$\begin{aligned} L'(\overline{yy_0})(e) &= \int_0^1 L'(y_0 + se)(e) ds \\ &= \int_0^1 \frac{\partial}{\partial s}(L(y_0 + se)) ds \\ &= L(y_0 + e) - L(y_0) \\ &= L(y) - L(y_0). \end{aligned}$$

Da $e \in Y$, gilt nach (6.2) außerdem, dass $L'(y)(e) = 0$. Eine Nullergänzung liefert uns damit:

$$L(y) - L(y_0) = L'(\overline{yy_0})(e) + \frac{1}{2} L'(y_0)(e) - \frac{1}{2} L'(y_0)(e) - \frac{1}{2} L'(y)(e).$$

Weiterhin erhalten wir mit $\widehat{y}_0 \in Y_0$

$$\begin{aligned} L'(y_0)(e) &= L'(y_0)(y - y_0) \\ &= L'(y_0)(y - \widehat{y}_0 + \widehat{y}_0 - y_0) \\ &= L'(y_0)(y - \widehat{y}_0) + L'(y_0)(\widehat{y}_0 - y_0) \\ &= L'(y_0)(y - \widehat{y}_0). \end{aligned}$$

Dabei haben wir im letzten Schritt die Eigenschaft (6.3) ausgenutzt. Insgesamt gilt also für alle $\widehat{y}_0 \in Y_0$:

$$L(y) - L(y_0) = \frac{1}{2}L'(y_0)(y - \widehat{y}_0) + L'(\overline{yy_0})(e) - \frac{1}{2}L'(y_0)(e) - \frac{1}{2}L'(y)(e).$$

Die Trapezregel liefert uns für die letzten drei Terme:

$$\int_0^1 L'(y_0 + se)(e)ds = \frac{1}{2}\left(L'(y_0)(e) + L'(y)(e)\right) + R.$$

Für den Restterm R gilt nach Konstruktion der Trapezregel:

$$\begin{aligned} R &= \int_0^1 (L'(y_0 + se)(e) - L_1 L'(y_0 + se)(e))ds \\ &= \int_0^1 \frac{1}{2}L'''(y_0 + se)(e,e,e) \cdot s \cdot (s-1)ds, \end{aligned}$$

wobei L_1 den linearen Interpolationsoperator mit den Eckpunkten 0 und 1 bezeichnet und die zweite Gleichheit der üblichen Interpolationsfehlerdarstellung (siehe z.B. [Lub05], Satz 9.10) entspricht. □

6.2 A Posteriori Fehlerdarstellung für das Steuerungsfunktional

Wir wollen im Folgenden eine Darstellung für den Fehler $J(q,u) - J(q_{kh}, u_{kh})$ herleiten, wobei $u \in X$ Lösung der kontinuierlichen Zustandsgleichung (2.3) ist, $u_{kh} \in X_{k,h}^{r,s}$ Lösung der volldiskreten Zustandsgleichung (5.8) sowie $q \in Q$ und $q_{kh} \in Q$ die optimalen Steuerungen des kontinuierlichen (3.2) beziehungsweise volldiskreten (5.11) Optimierungsproblems sind. Dabei wollen wir den Einfluss der Zeit- und Ortsdiskretisierung getrennt betrachten und schreiben deshalb:

$$\begin{aligned} J(q,u) - J(q_{kh}, u_{kh}) &= J(q,u) - J(q_k, u_k) && (6.5) \\ &\quad + J(q_k, u_k) - J(q_{kh}, u_{kh}) && (6.6) \end{aligned}$$

mit der semidiskreten Lösung $u_k \in X_k^r$ von (5.4) und der optimalen Steuerung $q_k \in Q$ des semidiskreten Optimierungsproblems (5.9).

Die Anwendung von Satz 6.1 liefert uns

Satz 6.2 *Sind $(q,u) \in Q \times X$ die Lösung des kontinuierlichen Optimierungsproblems* (3.2)*, $(q_k,u_k) \in Q \times X_k^h$ die Lösung des semidiskreten Optimierungsproblems* (5.9)*, $(q_{kh},u_{kh}) \in Q \times X_{k,h}^{r,s}$ die Lösung des volldiskreten Optimierungsproblems* (5.11) *und $z \in X, z_k \in X_k^r, z_{kh} \in X_{k,h}^{r,s}$ die Lösungen der entsprechenden dualen Probleme, so gilt*

$$J(q,u) - J(q_k,u_k) = \frac{1}{2}\mathcal{L}_k'(q_k,u_k,z_k)(q-\widehat{q}_k, u-\widehat{u}_k, z-\widehat{z}_k) \quad \textit{und} \tag{6.7}$$

$$J(q_k,u_k) - J(q_{kh},u_{kh}) = \frac{1}{2}\mathcal{L}_k'(q_{kh},u_{kh},z_{kh})(q_k-\widehat{q}_{kh}, u_k-\widehat{u}_{kh}, z_k-\widehat{z}_{kh}) \tag{6.8}$$

für beliebige $\widehat{u}_k, \widehat{z}_k \in X_k^r$, $\widehat{u}_{kh}, \widehat{z}_{kh} \in X_{k,h}^{r,s}$ und $\widehat{q}_k, \widehat{q}_{kh} \in Q$, wobei mit $\mathcal{L}_k'(\cdot,\cdot,\cdot)(\cdot,\cdot,\cdot)$ die totale Ableitung des diskreten Lagrange-Funktionals gemeint ist.

Beweis: Ist $(q,u) \in Q \times X$ die Lösung des kontinuierlichen Optimierungsproblems (3.2), so ist insbesondere u Lösung der kontinuierlichen Zustandsgleichung (2.3) und damit gilt

$$\mathcal{L}(q,u,z) = J(q,u).$$

Die analoge Aussage gilt natürlich auch für die diskreten Fälle, also

$$\mathcal{L}_k(q_k,u_k,z_k) = J(q_k,u_k) \qquad \text{und} \qquad \mathcal{L}_k(q_{kh},u_{kh},z_{kh}) = J(q_{kh},u_{kh}).$$

Somit gilt für alle $z \in X$, $z_k \in X_k^r$ und $z_{kh} \in X_{k,h}^{r,s}$

$$\begin{aligned} J(q,u) - J(q_k,u_k) &= \mathcal{L}(q,u,z) - \mathcal{L}_k(q_k,u_k,z_k) \\ &= \mathcal{L}_k(q,u,z) - \mathcal{L}_k(q_k,u_k,z_k) \end{aligned}$$

und

$$J(q_k,u_k) - J(q_{kh},u_{kh}) = \mathcal{L}_k(q_k,u_k,z_k) - \mathcal{L}_k(q_{kh},u_{kh},z_{kh}).$$

Hierbei nutzen wir aus, dass $\mathcal{L}(q,u,z) = \mathcal{L}_k(q,u,z)$ für $(q,u,z) \in Q \times X \times X$ gilt, da in diesem Fall die in $\mathcal{L}_k$ auftretenden Sprungterme aus der Zeitdiskretisierung identisch Null sind.

Weiterhin folgt aus Bemerkung 4.7, dass $(q,u,z) \in Q \times X \times X$ das Optimalitätssystem

$$\left.\begin{aligned} \mathcal{L}_z'(\widehat{q},u,z)(\tau z) = 0, \quad \forall\, \tau z \in X, \qquad & \text{(Zustandsgleichung)} \\ \mathcal{L}_u'(\widehat{q},u,z)(\tau u) = 0, \quad \forall\, \tau u \in X, \qquad & \text{(duales Problem)} \\ \mathcal{L}_{\widehat{q}}'(\widehat{q},u,z)(\tau q) = 0, \quad \forall\, \tau q \in Q, \qquad & \text{(Optimalitätsbedingung)} \end{aligned}\right\}$$

löst, falls $(q,u) \in Q \times X$ Lösung des Optimierungsproblems ist. Es gilt also insbesondere für alle $(\widehat{q},\widehat{u},\widehat{z}) \in Q \times X \times X$

$$\begin{aligned}\mathcal{L}'(q,u,z)(\widehat{q},\widehat{u},\widehat{z}) &= \mathcal{L}'_q(q,u,z)(\widehat{q}) + \mathcal{L}'_u(q,u,z)(\widehat{u}) + \mathcal{L}'_z(q,u,z)(\widehat{z}) \\ &= 0. \end{aligned} \tag{6.9}$$

Damit ist $(q,u,z) \in Q \times X \times X$ stationärer Punkt von $\mathcal{L}'_k$.
Analog sind $(q_k,u_k,z_k) \in Q \times X_k^r \times X_k^r$ und $(q_{kh},u_{kh},z_{kh}) \in Q \times X_{k,h}^{r,s} \times X_{k,h}^{r,s}$ stationäre Punkte von $\mathcal{L}'_k$, d.h. es gilt

$$\begin{aligned}\mathcal{L}'_k(q_k,u_k,z_k)(\widehat{q}_k,\widehat{u}_k,\widehat{z}_k) = 0, &\qquad \forall\, (\widehat{q}_k,\widehat{u}_k,\widehat{z}_k) \in Q \times X_k^r \times X_x^r, \\ \mathcal{L}'_k(q_{kh},u_{kh},z_{kh})(\widehat{q}_{kh},\widehat{u}_{kh},\widehat{z}_{kh}) = 0, &\qquad \forall\, (\widehat{q}_{kh},\widehat{u}_{kh},\widehat{z}_{kh}) \in Q \times X_{k,h}^{r,s} \times X_{k,h}^{r,s}.\end{aligned}$$

Setzen wir nun $Y = Q \times (X \cup X_k^r) \times (X \cup X_k^r)$ und $Y_0 = Q \times X_k^r \times X_k^r$, so sind die Voraussetzungen von Satz 6.1 erfüllt und wir erhalten:

$$\begin{aligned}J(q,u) - J(q_k,u_k) &= \mathcal{L}_k(q,u,z) - \mathcal{L}_k(q_k,u_k,z_k) \\ &= \frac{1}{2}\mathcal{L}'_k(q_k,u_k,z_k)(q-\widehat{q}_k, u-\widehat{u}_k, z-\widehat{z}_k).\end{aligned}$$

Der Restterm R verschwindet hierbei, da die dritten Ableitungen von $\mathcal{L}_k$ nach Lemma 4.3 identisch Null sind. Da X_k^r keine Teilmenge von X ist, haben wir außerdem in Y die Vereinigung $X \cup X_k^r$ verwendet, um $Y_0 \subset Y$ sicherzustellen. Gleichung (6.9) ist dabei weiterhin erfüllt, da $X \subset X \cup X_k^r$ eine dichte Einbettung ist.

Analog erhalten wir mit $Y = Q \times X_k^r \times X_x^r$ und $Y_0 = Q \times X_{k,h}^{r,s} \times X_{k,h}^{r,s}$ aus Satz 6.1 die Identität

$$J(q_k,u_k) - J(q_{kh},u_{kh}) = \frac{1}{2}\mathcal{L}'_k(q_{kh},u_{kh},z_{kh})(q_k-\widehat{q}_{kh}, u_k-\widehat{u}_{kh}, z_k-\widehat{z}_{kh}). \qquad \square$$

Bemerkung 6.3 Die Dichtheit $X \subset X \cup X_k^r$ kann mit Hilfe von Mollifiern und Faltungseigenschaften ähnlich wie in Abschnitt 4.2 bei [Hoh05] gezeigt werden.

Führen wir nun folgende Residuen ein

$$\widetilde{\rho}^u(q,u)(\phi) = \mathcal{L}'_{k,z}(q,u,z)(\phi), \qquad \phi \in X, \tag{6.10}$$

$$\widetilde{\rho}^z(u,z)(\phi) = \mathcal{L}'_{k,u}(q,u,z)(\phi), \qquad \phi \in X, \tag{6.11}$$

$$\widetilde{\rho}^q(q,z)(\phi) = \mathcal{L}'_{k,q}(q,u,z)(\phi), \qquad \phi \in Q, \tag{6.12}$$

so können die Aussagen aus Satz 6.2 geschrieben werden als

$$J(q,u) - J(q_k,u_k) = \frac{1}{2}(\widetilde{\rho}^u(q_k,u_k)(z-\widehat{z}_k) + \widetilde{\rho}^z(u_k,z_k)(u-\widehat{u}_k)) \tag{6.13}$$

und

$$J(q_k,u_k) - J(q,u_{kh}) = \frac{1}{2}(\widetilde{\rho}^u(q_{kh},u_{kh})(z_k-\widehat{z}_{kh}) + \widetilde{\rho}^z(u_{kh},z_{kh})(u_k-\widehat{u}_{kh})). \tag{6.14}$$

Die Terme $\widetilde{\rho}^q(q_k,z_k)(q-\widehat{q})$ und $\widetilde{\rho}^q(q_{kh},z_{kh})(q-\widehat{q})$ verschwinden hierbei, da $\widehat{q}$ jeweils beliebig aus Q gewählt werden darf, also insbesondere auch $\widehat{q}=q$.

6.3 Fehlerschätzer für die dG(0)cG(1) Diskretisierung

In die Fehlerdarstellungen des vorigen Abschnitts sind jeweils auch Interpolationsfehler (z.B. $z-\widehat{z}_k$) der Zeit- und Ortsdiskretisierung eingegangen. Diese wollen wir nun durch Interpolation in höher dimensionale Finite-Elemente-Räume approximieren. Wir beziehen uns hierbei wieder auf die dG(0) Diskretisierung in der Zeit und schränken uns bei der räumlichen Diskretisierung auf bilineare Ansatz- und Testfunktionen ein, d.h. wir setzen $s=1$. Da wir dabei stetige Übergänge an den Zellgrenzen haben, sprechen wir auch von einer cG(1) Diskretisierung im Ort. Außerdem betrachten wir in dieser Arbeit nur den Fall $\Omega \subset \mathbb{R}^2$.

6.3.1 Berechnung des Fehlerschätzers

Zum Zweck der Interpolation führen wir die Operatoren Π_k und Π_h mit den Eigenschaften

$$\begin{aligned} z-\widehat{z}_k &\approx \Pi_k z_k, & u-\widehat{u}_k &\approx \Pi_k u_k,\\ z_k-\widehat{z}_{kh} &\approx \Pi_h z_{kh}, & u_k-\widehat{u}_{kh} &\approx \Pi_h u_{kh} \end{aligned}$$

ein.
Hierbei haben Π_k und Π_h die Form

$$\Pi_k := I_k^{(1)} - id \qquad mit\ I_k^{(1)} : X_k^0 \longrightarrow X_k^1, \tag{6.15}$$

$$\Pi_h := I_{2h}^{(2)} - id \qquad mit\ I_{2h}^{(2)} : X_{k,h}^{0,1} \longrightarrow X_{k,2h}^{0,2}. \tag{6.16}$$

Dabei ist der lineare Interpolationsoperator $I_k^{(1)}$ für $v \in X_k^0$ definiert als:

$$I_k^{(1)} v(t) := \frac{t_m - t}{k_m} v(t_{m-1}) + \frac{t - t_{m-1}}{k_m} v(t_m), \qquad \text{für } t \in (t_{m-1}, t_m].$$

Die Wirkung von $I_k^{(1)}$ ist in Abbildung 6.1 dargestellt.

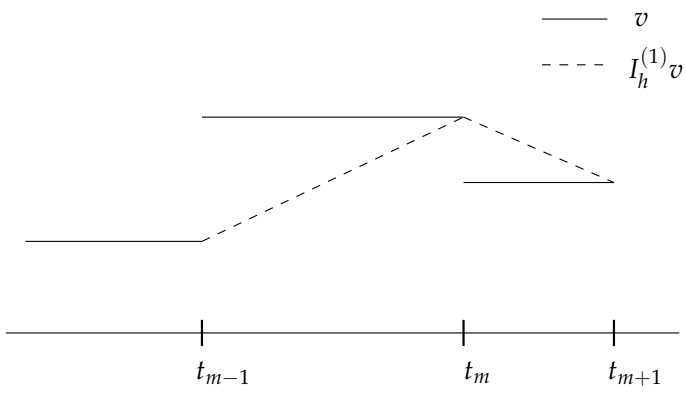


Abbildung 6.1: Wirkung des Interpolationsoperators $I_k^{(1)}$

Den räumliche Interpolationsoperator $I_{2h}^{(2)}$ definieren wir folgendermaßen: Das Gebiet $\Omega \subset \mathbb{R}^2$ diskretisieren wir mit Rechteckselementen mit einer Patch-Struktur (siehe Abbildung 6.2), d.h. jeweils 4 Zellen der feineren Diskretisierung bilden zusammen eine Makro-Zelle der gröberen Diskretisierung. Der Operator $I_{2h}^{(2)}$ beschreibt nun die biquadratische Interpolation einer linearen Funktion u, die auf dem feinen Gitter definiert ist, in den Raum der biquadratischen Funktionen auf dem groben Gitter.

Diese Art der Approximation der Interpolationsfehler ist sinnvoll, da bisher $\widehat{z}_k, \widehat{u}_k \in X_k^r$ und $\widehat{z}_{kh}, \widehat{u}_{kh} \in X_{k,h}^{r,s}$ beliebig gewählt werden konnten, also insbesondere auch $\widehat{z}_k = z_k, \widehat{u}_k = u_k, \widehat{z}_{kh} = z_{kh}$ und $\widehat{u}_{kh} = u_{kh}$. Des Weiteren stellt z.B. $I_k^{(1)} z_k$ eine Approximation an z dar, ebenso wie $I_{2h}^{(2)} z_{kh}$ eine Approximation an z_k ist. Damit haben wir also $z - \widehat{z}_k \approx I_k^{(1)} z_k - z_k$ und $z_k - \widehat{z}_{kh} \approx I_{2h}^{(2)} z_{kh} - z_{kh}$. Analog natürlich auch $u - \widehat{u}_k \approx I_k^{(1)} u_k - u_k$ und $u_k - \widehat{u}_{kh} \approx I_{2h}^{(2)} u_{kh} - u_{kh}$.

Um nun die bisher hergeleitete Fehlerdarstellungen (6.13) und (6.14) berechenbar zu machen, muss man die Lösungen der semidiskreten Probleme durch die Lösungen der volldiskreten Probleme ersetzen.

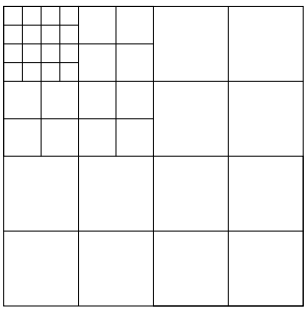


Abbildung 6.2: Beispiel für ein Gitter mit Patch-Struktur

Insgesamt erhalten wir also

$$J(q,u) - J(q_{kh},u_{kh}) \approx \eta_k + \eta_h$$

mit

$$\eta_k := \frac{1}{2}(\widetilde{\rho}^u(q_{kh},u_{kh})(\Pi_k z_{kh}) + \widetilde{\rho}^z(u_{kh},z_{kh})(\Pi_k u_{kh}) \qquad \textit{und}$$
$$\eta_h := \frac{1}{2}(\widetilde{\rho}^u(q_{kh},u_{kh})(\Pi_h z_{kh}) + \widetilde{\rho}^z(u_{kh},z_{kh})(\Pi_h u_{kh}).$$

Für die konkrete Wahl $u_{kh}, z_{kh} \in X_{k,h}^{0,1}$ gilt dann:

Satz 6.4 *Ist $f(q)$ nicht nur in $L^2(I,V^*)$, sondern sogar stückweise stetig bezüglich t, so gilt mit der Notation $z_{kh}|_{I_m} = z_{kh}(t_m) = Z_m$ und $u_{kh}|_{I_m} = u_{kh}(t_m) = U_m$:*

$$\begin{aligned}\widetilde{\rho}^u(q_{kh},u_{kh})(\Pi_k z_{kh}) \approx \sum_{m=1}^{M} \{&(U_m - U_{m-1}, Z_m - Z_{m-1})_H \\ &+ \frac{k_m}{2}\widetilde{a}(U_m, Z_m - Z_{m-1}) \\ &+ \frac{k_m}{2}((f(q_{kh})(t_{m-1}), Z_{m-1})_H \\ &\quad - (f(q_{kh})(t_m), Z_m)_H)\},\end{aligned} \tag{6.17a}$$

$$\begin{aligned}\widetilde{\rho}^z(u_{kh},z_{kh})(\Pi_k u_{kh}) \approx \sum_{m=1}^{M} \frac{k_m}{2}\{&\widetilde{a}(U_m - U_{m-1}, Z_m) \\ &+ b_1(U_{m-1} - U_m, U_m) + L_1(U_{m-1} - U_m)\},\end{aligned} \tag{6.17b}$$

$$\begin{aligned}
\widetilde{\rho}^u(q_{kh},u_{kh})(\Pi_h z_{kh}) \approx \sum_{m=1}^{M} \{&-(U_m - U_{m-1}, I_{2h}^{(2)} Z_m - Z_m)_H \\
&- k_m \widetilde{a}(U_m, I_{2h}^{(2)} Z_m - Z_m) \\
&+ k_m (f(q_{kh})(t_m), I_{2h}^{(2)} Z_m - Z_m)_H\}, \qquad (6.17c)\\
\widetilde{\rho}^z(u_{kh},z_{kh})(\Pi_h u_{kh}) \approx \sum_{m=1}^{M} k_m \{&-\widetilde{a}(I_{2h}^{(2)} U_m - U_m, Z_m) \\
&+ b_1(I_{2h}^{(2)} U_m - U_m, U_m) + L_1(I_{2h}^{(2)} U_m - U_m)\} \\
&+ \sum_{m=1}^{M-1} (I_{2h}^{(2)} U_m - U_m, Z_{m+1} - Z_m)_H \\
&+ b_2(I_{2h}^{(2)} U_M - U_M, U_M) + L_2(I_{2h}^{(2)} U_M - U_M) \\
&- (I_{2h}^{(2)} U_M - U_M, Z_M)_H. \qquad (6.17d)
\end{aligned}$$

Beweis: Nach Definition gelten

$$\begin{aligned}
\widetilde{\rho}^u(q_{kh},u_{kh})(\widehat{z}_{kh}) &= \mathcal{L}'_{k,z}(q_{kh},u_{kh},z_{kh})(\widehat{z}_{kh}) \\
&= \sum_{m=1}^{M} \int_{I_m} \{(f(q_{kh}) - \partial_t u_{kh}, \widehat{z}_{kh})_H - \widetilde{a}(u_{kh},\widehat{z}_{kh})\}\, dt \\
&\quad - \sum_{m=1}^{M} ([u_{kh}]_{m-1}, (\widehat{z}_{kh})^+(t_{m-1}))_H
\end{aligned} \qquad (6.18)$$

und

$$\begin{aligned}
\widetilde{\rho}^z(u_{kh},z_{kh})(\widehat{u}_{kh}) &= \mathcal{L}'_{k,u}(q_{kh},u_{kh},z_{kh})(\widehat{u}_{kh}) \\
&= \sum_{m=1}^{M} \int_{I_m} \{b_1(\widehat{u}_{kh}(t), u_{kh}(t)) + L_1(\widehat{u}_{kh}(t))\}\, dt \\
&\quad + b_2(\widehat{u}_{kh}(T), u_{kh}(T)) + L_2(\widehat{u}_{kh}(T)) \\
&\quad - \sum_{m=1}^{M} \int_{I_m} \{(-\widehat{u}_{kh}, \partial_t z_{kh})_H + \widetilde{a}(\widehat{u}_{kh}, z_{kh})\}\, dt \\
&\quad + \sum_{m=1}^{M-1} ((\widehat{u}_{kh})^-(t_m), [z_{kh}]_m)_H \\
&\quad - ((\widehat{u}_{kh})^-(T), z_{kh}^-(T))_H.
\end{aligned} \qquad (6.19)$$

Obwohl $u_{kh}, z_{kh} \in X_{k,h}^{0,1}$ die Lösungen der volldiskreten Zustandsgleichung beziehungsweise des volldiskreten dualen Problems sind, verschwinden die

Ausdrücke $\mathcal{L}'_{k,z}(q_{kh}, u_{kh}, z_{kh})(\widehat{z}_{kh})$ und $\mathcal{L}'_{k,u}(q_{kh}, u_{kh}, z_{kh})(\widehat{u}_{kh})$ nicht, da die Testfunktionen $\widehat{z}_{kh}$ und $\widehat{u}_{kh}$ aus den höher dimensionalen Räumen $X^{1,1}_{k,h}$ und $X^{0,2}_{k,h}$ gewählt wurden. Allerdings gilt wegen $u^-_{kh}(0) = u_0(q)$

$$(u^-_{kh}(0) - u_0(q_{kh}), \widehat{z}^+_{kh}(0))_H = 0,$$

so dass sich $\mathcal{L}'_{k,z}(q_{kh}, u_{kh}, z_{kh})(\widehat{z}_{kh})$ auf den in (6.18) angegeben Ausdruck reduziert.

Um nun Gleichung (6.17a) zu zeigen, wählen wir (6.18) mit

$$\widehat{z}_{kh} = \Pi_k z_{kh} = I^{(1)}_k z_{kh} - z_{kh}$$

als Ausgangspunkt.

Da u_{kh} auf den einzelnen Teilintervallen I_m konstant ist, gilt $\partial_t u_{kh}|_{I_m} = 0$.

Nun approximieren wir alle Integrale, die den Ausdruck $I^{(1)}_k z_{kh}$ enthalten, durch die Trapezregel (also $\int_{I_m} g(t)\,dt \approx \frac{k_m}{2}(g(t_{m-1}) + g(t_m))$) und alle restlichen mit der oberen Rechteckregel. Dies liefert uns unter der Berücksichtigung der Definition von $I^{(1)}_k$:

$$\begin{aligned}
\int_{I_m} (f(q_{kh}), \Pi_k z_{kh})_H\,dt &= \int_{I_m} (f(q_{kh}), I^{(1)}_k z_{kh})_H\,dt - \int_{I_m} (f(q_{kh}), z_{kh})_H\,dt \\
&\approx \frac{k_m}{2}((f(q_{kh})(t_{m-1}), I^{(1)}_k z_{kh}(t_{m-1}))_H \\
&\qquad + (f(q_{kh})(t_m), I^{(1)}_k z_{kh}(t_m))_H) \\
&\quad - k_m (f(q_{kh})(t_m), I^{(1)}_k z_{kh}(t_m))_H \\
&= \frac{k_m}{2}((f(q_{kh})(t_{m-1}), Z_{m-1})_H - (f(q_{kh})(t_m), Z_m)).
\end{aligned}$$

Da $\widetilde{a}$ unabhängig von t ist und $\Pi_k z_{kh}$ linear von t abhängt, ist die folgende Auswertung sogar exakt:

$$\begin{aligned}
-\int_{I_m} \widetilde{a}(u_{kh}, \Pi_k z_{kh}) &= \int_{I_m} \widetilde{a}(u_{kh}, z_{kh})\,dt - \int_{I_m} \widetilde{a}(u_{kh}, I^{(1)}_k z_{kh})\,dt \\
&= k_m \widetilde{a}(U_m, Z_m) - \frac{k_m}{2}(\widetilde{a}(U_m, Z_{m-1}) + \widetilde{a}(U_m, Z_m)) \\
&= \frac{k_m}{2}\widetilde{a}(U_m, Z_m - Z_{m-1}).
\end{aligned}$$

Einfaches Einsetzen liefert weiterhin

$$\begin{aligned}-([u_{kh}]_{m-1}, (\Pi_k z_{kh})^+(t_{m-1}))_H &= (U_m - U_{m-1}, Z_m - (I_k^{(1)} z_{kh})^+(t_{m-1}))_H \\ &= (U_m - U_{m-1}, Z_m - Z_{m-1})_H.\end{aligned}$$

Unter Berücksichtigung dieser Ergebnisse ergibt sich aus (6.18) die gewünschte Gleichung (6.17a).

Analog berechnen wir Gleichung (6.17b) aus (6.19), indem wir

$$\widehat{u}_{kh} = \Pi_k u_{kh} = I_k^{(1)} u_{kh} - u_{kh}$$

wählen.

Hierbei beachten wir, dass $\partial_t z_{kh} = \Pi_k u_{kh}(T) = 0$ nach den gleichen Argumenten wie oben erfüllt ist und dass außerdem

$$(\Pi_k u_{kh})^-(t_m) = (I_k^{(1)} u_{kh})^-(t_m) - u_{kh}^-(t_m) = U_m - U_m = 0$$

gilt.

Indem man in (6.19) nun ebenfalls die Integrale, die den Ausdruck $I_k^{(1)} u_{kh}$ enthalten, durch die Trapezregel und die Restlichen durch die obere Rechteckregel approximiert, erhalten wir Gleichung (6.17b).

Durch konsequentes Anwenden der oberen Rechteckregel erhalten wir analog Gleichung (6.17c) aus (6.18) mit $\widehat{z}_{kh} = \Pi_h z_{kh}$ und (6.17d) aus (6.19) durch Wahl von $\widehat{u}_{kh} = \Pi_h u_{kh}$. □

6.3.2 Lokalisierung des Fehlerschätzers

Um lokale Adaptivität zu ermöglichen, wollen wir die Fehlerschätzer des vorigen Abschnitts nun auf den einzelnen Zeitschritten und Zellen des räumlichen Gitters betrachten. Dafür setzen wir voraus, dass wir ein zeitabhängiges Gitter $\mathcal{T}_{h,m}$ haben, dass also die Zerlegung von Ω von Zeitschritt zu Zeitschritt variieren kann.

Wir beschränken uns im Folgenden auf die Wärmeleitungsgleichung, also

$$\widetilde{a}(u, v) = \int_\Omega \nabla u \cdot \nabla v \, dx.$$

Außerdem gehen wir davon aus, dass wir die Bilinearform b_1 und die Linearform L_1 räumlich lokal darstellen können als

$$b_1(u(t), v(t)) = \sum_{D \in \mathcal{T}_h} b_1|_D(u(t), v(t))$$

und

$$L_1(u(t)) = \sum_{D \in \mathcal{T}_h} L_1|_D(u(t)).$$

Für den zeitbezogenen Fehler gilt

Bemerkung 6.5 Mit

$$\begin{aligned}\eta_m^{time} := (U_m - U_{m-1}, Z_m - Z_{m-1})_H \\ + \frac{k_m}{2}\{\widetilde{a}(U_m, Z_m - Z_{m-1}) + \widetilde{a}(U_m - U_{m-1}, Z_m) \\ + (f(q_{kh})(t_{m-1}), Z_{m-1})_H - (f(q_{kh})(t_m), Z_m)_H \\ + b_1(U_{m-1} - U_m, U_m) + L_1(U_{m-1} - U_m)\}\end{aligned}$$

und Satz 6.4 gilt offensichtlich

$$\eta_k \approx \frac{1}{2} \sum_{m=1}^{M} \eta_m^{time}.$$

Für den räumlichen Fehlerschätzer erhalten wir:

Lemma 6.6 *Es gilt*

$$\eta_h \leq \frac{1}{2} \left(\sum_{m=1}^{M} k_m \sum_{D \in \mathcal{T}_h} \eta_{m,D}^{space} + \xi_M \right)$$

mit

$$\eta_{m,D}^{space} := \xi_1^{m,D}\xi_2^{m,D} + \xi_3^{m,D}\xi_4^{m,D} + \xi_5^{m,D} + \xi_6^{m,D},$$

$$\xi_1^{m,D} := \|f(q_{kh})(t_m) + \Delta U_m - \frac{U_m - U_{m-1}}{k_m}\|_D, \quad \xi_2^{m,D} := \|I_{2h}^{(2)} Z_m - Z_m\|_D,$$

$$\xi_3^{m,D} := \|\Delta Z_m + \frac{Z_{m+1} - Z_m}{k_m}\|_D, \qquad \xi_4^{m,D} := \|I_{2h}^{(2)} U_m - U_m\|_D,$$

$$\xi_5^{m,D} := b_1|_D(I_{2h}^{(2)}U_m - U_m, U_m), \qquad \xi_6^{m,D} := L_1|_D(I_{2h}^{(2)}U_m - U_m)$$

für $m = 1, \ldots, M$ *und*

$$\xi_M := b_2(I_{2h}^{(2)}U_M - U_M, U_M) + L_2(I_{2h}^{(2)}U_M - U_M) - (I_{2h}^{(2)}U_M - U_M, Z_{M+1})_H$$

für $Z_{M+1} \in X_{k,h}^{0,1}$ *beliebig. Mit* $\|\cdot\|_D$ *wird hierbei die Norm auf einer einzelnen Zelle* D *bezeichnet, d.h.* $\|g\|_D^2 = \int_D g(x)^2\,dx$.

Beweis: Es gilt durch Anwenden von partieller Integration und der Cauchy-Schwarzschen Ungleichung für $m = 1, \ldots, M$:

$$\begin{aligned}
&-(U_m - U_{m-1}, I_{2h}^{(2)}Z_m - Z_m)_H - k_m\widetilde{a}(U_m, I_{2h}^{(2)}Z_m - Z_m) \\
&\qquad\qquad + k_m(f(q_{kh})(t_m), I_{2h}^{(2)}Z_m - Z_m)_H \\
&= -\int_\Omega (U_m - U_{m-1})(I_{2h}^{(2)}Z_m - Z_m)\,dx \\
&\quad - k_m\int_\Omega \nabla U_m \cdot \nabla(I_{2h}^{(2)}Z_m - Z_m)\,dx \\
&\quad + k_m\int_\Omega f(q_{kh})(t_m)(I_{2h}^{(2)}Z_m - Z_m)\,dx \\
&= \sum_{D\in\mathcal{T}_h}\{\int_D (k_m f(q_{kh})(t_m) - U_m + U_{m-1})(I_{2h}^{(2)}Z_m - Z_m)\,dx \\
&\quad + k_m\int_D \Delta U_m(I_{2h}^{(2)}Z_m - Z_m)\,dx\} \\
&= k_m\sum_{D\in\mathcal{T}_h}\int_D (f(q_{kh})(t_m) + \Delta U_m - \frac{U_m - U_{m-1}}{k_m})(I_{2h}^{(2)}Z_m - Z_m)\,dx \\
&\le k_m\sum_{D\in\mathcal{T}_h}\|f(q_{kh})(t_m) + \Delta U_m - \frac{U_m - U_{m-1}}{k_m}\|_D\|I_{2h}^{(2)}Z_m - Z_m\|_D \\
&= k_m\sum_{D\in\mathcal{T}_h}\{\xi_1^{m,D}\cdot\xi_2^{m,D}\}.
\end{aligned}$$

Für beliebige $Z_{M+1} \in X_{k,h}^{0,1}$, also z.B. $Z_{M+1} = Z_M$, gilt

$$\begin{aligned}
&\sum_{m=1}^{M-1}(I_{2h}^{(2)}U_m - U_m, Z_{m+1} - Z_m)_H - (I_{2h}^{(2)}U_M - U_M, Z_M)_H \\
&\quad = \sum_{m=1}^{M}(I_{2h}^{(2)}U_m - U_m, Z_{m+1} - Z_m)_H - (I_{2h}^{(2)}U_M - U_M, Z_{M+1} - Z_M)_H \\
&\quad - (I_{2h}^{(2)}U_M - U_M, Z_M)_H
\end{aligned}$$

$$= \sum_{m=1}^{M} (I_{2h}^{(2)} U_m - U_m, Z_{m+1} - Z_m)_H - (I_{2h}^{(2)} U_M - U_M, Z_{M+1})_H.$$

Setzt man dies nun in (6.17d) ein und beachtet, dass

$$\begin{aligned}
k_m\{-\widetilde{a}(I_{2h}^{(2)} U_m - U_m, Z_m) + b_1(I_{2h}^{(2)} U_m - U_m, U_m) + L_1(I_{2h}^{(2)} U_m - U_m)\} \\
+ (I_{2h}^{(2)} U_m - U_m, Z_{m+1} - Z_m)_H \\
= k_m \int_\Omega -\nabla(I_{2h}^{(2)} U_{m-1} - U_{m-1}) \cdot \nabla Z_m \, dx \\
+ \int_\Omega (I_{2h}^{(2)} U_m - U_m)(Z_{m+1} - Z_m) \, dx \\
+ k_m\{b_1(I_{2h}^{(2)} U_m - U_m, U_m) + L_1(I_{2h}^{(2)} U_m - U_m)\} \\
= k_m \sum_{D \in \mathcal{T}_h} \{ \int_D (\Delta Z_m + \frac{Z_{m+1} - Z_m}{k_m})(I_{2h}^{(2)} U_m - U_m, U_m) \, dx \\
+ b_1|_D(I_{2h}^{(2)} U_m - U_m, U_m) + L_1|_D(I_{2h}^{(2)} U_m - U_m)\} \\
\leq k_m \sum_{D \in \mathcal{T}_h} \{ \|\Delta Z_m + \frac{Z_{m+1} - Z_m}{k_m}\|_D \| I_{2h}^{(2)} U_m - U_m\|_D \\
+ b_1|_D(I_{2h}^{(2)} U_m - U_m, U_m) + L_1|_D(I_{2h}^{(2)} U_m - U_m)\} \\
= k_m \sum_{D \in \mathcal{T}_h} \{\xi_3^{m,D} \cdot \xi_4^{m,D} + \xi_5^{m,D} + \xi_6^{m,D}\}
\end{aligned}$$

gilt, so ergibt sich mit (6.17c) und (6.17d) die gewünschte Fehlerabschätzung. □

6.4 Adaptiver Algorithmus

Nun wollen wir die im vorigen Abschnitt hergeleiteten Fehlerabschätzungen verwenden, um einen adaptiven Algorithmus zur Lösung des Optimierungsproblems vorzustellen. Dessen Anforderungen folgen dabei der Idee in [MV07].

Ziel ist es, den Gesamt-Fehler $J(q,u) - J(q_{kh}, u_{kh})$ zu minimieren, indem wir die Zeitschrittweite und das räumliche Gitter dem Problem anpassen. Außerdem sollen der Fehler, der durch die zeitliche Diskretisierung entsteht, und der, der auf der räumlichen Diskretisierung basiert, etwa gleich groß sein. Insgesamt soll also bei gegebener Toleranz TOL gelten:

$$J(q,u) - J(q_{kh}, u_{kh}) = \eta_k + \eta_h < TOL$$

und

$$|\eta_k| \approx |\eta_h|.$$

Das grundsätzliche Vorgehen sieht dabei so aus, dass man zunächst eine approximative Lösung des Optimierungsproblems bestimmt. Dann werden die Fehlerschätzer ausgewertet und, falls der Gesamt-Fehler zu groß ist, wird in Abhängigkeit von η_k und η_h zeitlich und/oder räumlich verfeinert.

Wir werden folgende Bezeichnungen verwenden:

- K^n steht stellvertretend für die zeitliche Diskretisierung im n-ten Schritt, also für die Zerlegung $\overline{I} = \{0\} \cup I_1^n \cup I_2^n \cup \ldots \cup I_{M_n}^n$ des Ursprungsintervalls in M_n Teilintervalle.
- $\mathcal{T}^n$ bezeichnet das zeitabhängige räumliche Gitter im n-ten Schritt. Zu ihm gehören die Gitter $\mathcal{T}_{h,1}^n$, $\mathcal{T}_{h,2}^n$, …, $\mathcal{T}_{h,M_n}^n$ in den Zeitschritten $1, 2, \ldots, M_n$.

Für die konkrete Umsetzung der Verfeinerung gibt es nun mehrere Möglichkeiten, die wir am Beispiel der Zeit-Diskretisierung darstellen wollen. Hierfür betrachten wir die lokalen Fehlerschätzer η_m^{time} aus Bemerkung 6.5 auf den einzelnen Zeitintervallen.

1. Die Zeitintervalle, die zu den z.B. 30% der Zeitintervalle mit dem größten Fehler η_m^{time} gehören, werden halbiert und die, die zu den z.B. 3% mit dem kleinsten Fehler gehören, werden mit einem Nachbarintervall zu einem gröberen Zeitintervall zusammengefasst.
2. Alle Zeitintervalle, deren Fehler η_m^{time} über einer vorgegebenen Verfeinerungs-Toleranz TOL_{fein} liegen, werden halbiert; alle Intervalle für die bei gegebener Vergröberungs-Toleranz TOL_{grob} gilt $\eta_m^{time} < TOL_{grob}$ werden mit einem Nachbarintervall zu einem gröberen Zeitintervall zusammengefasst.

Für die Verfeinerung und Vergröberung des räumlichen Gitters $\mathcal{T}^n$ überträgt sich das Vorgehen, indem wir die Fehlerschätzer $\eta_{m,D}^{space}$ betrachten. Dabei adaptieren wir die Gitter nicht getrennt voneinander, sondern suchen die Zellen mit dem größten/kleinsten Fehleranteil aus der Menge aller Gitter $\mathcal{T}_{h,1}^n$, $\mathcal{T}_{h,2}^n$, …, $\mathcal{T}_{h,M_n}^n$ und passen diese entsprechend an.

Insgesamt erhalten wir folgenden Algorithmus:

Algorithmus 6.1 : Adaptiver Verfeinerungs Algorithmus

Input : $n = 0$, $q^0 \in Q$, TOL,Anfangsdiskretisierung K^0, $\mathcal{T}^0$

Berechne eine optimale Lösung $(q_{kh}^{(0)}, u_{kh}^{(0)})$ mittels Algorithmus 4.2 oder 4.3;
Werte die Fehlerschätzer η_h und η_k aus;
while $\eta_k + \eta_h > TOL$ **do**
    if $|\frac{\eta_h}{\eta_k}| > 1$ **then** /* $|\eta_k| < |\eta_h|$ */
        Verfeinere das räumliche Gitter $\mathcal{T}^n \longrightarrow \mathcal{T}^{n+1}$ nach einer der oben genannten Strategien;
        $K^{n+1} = K^n$;
    else if $|\frac{\eta_k}{\eta_h}| > 1$ **then** /* $|\eta_k| > |\eta_h|$ */
        Verfeinere das zeitliche Gitter $K^n \longrightarrow K^{n+1}$ nach einer der oben genannten Strategien;
        $\mathcal{T}^{n+1} = \mathcal{T}^n$;
    else /* $|\eta_k| \approx |\eta_h|$ */
        Verfeinere räumliches und zeitliches Gitter $\mathcal{T}^n \longrightarrow \mathcal{T}^{n+1}$ und $K^n \longrightarrow K^{n+1}$ nach einer der oben genannten Strategien;
    Berechne eine neue optimale Lösung $(q_{kh}^{(n+1)}, u_{kh}^{(n+1)})$ mittels Algorithmus 4.2 oder 4.3;
    Werte die Fehlerschätzer η_k und η_h aus;

In der praktischen Umsetzung dieses Algorithmus wird man die Bedingungen $|\frac{\eta_h}{\eta_k}| > 1$ und $|\frac{\eta_k}{\eta_h}| > 1$ mit einer Ausbalancierungs-Toleranz tol_{equi} versehen und durch $|\frac{\eta_h}{\eta_k}| > 1 + tol_{equi}$ beziehungsweise $|\frac{\eta_k}{\eta_h}| > 1 + tol_{equi}$ ersetzen, da der *else*-Fall sonst so gut wie nie eintritt.

Kapitel 7

Numerische Resultate

In diesem Kapitel wollen wir die Anwendung der in dieser Arbeit vorgestellten Verfahren und deren Ergebnisse an zwei Beispielen darstellen. Zu Vergleichs- und Verifikationszwecken werden wir zunächst alternativ zu den Sätzen 4.6 und 4.8 die Ableitungen des reduzierten Steuerungsfunktionals unter der Verwendung von Differenzenquotienten darstellen. Dann wollen wir zwei Beispiele vorstellen, für die wir die in Kapitel 4 hergeleiteten Algorithmen durchführen. Dabei werden wir vor allem die Laufzeiten vergleichen und mit Hilfe der Differenzenquotienten-Darstellung die Korrektheit der Berechnung der Ableitungen dokumentieren.

7.1 Ableitungen mittels Differenzenquotient

Alternativ zum Vorgehen in Abschnitt 4.4 wollen wir im Folgenden die 1. und 2. Ableitung von j mittels Differenzenquotienten aufstellen. Wir verwenden hierbei jeweils zentrale Differenzen. Diese sind eine Approximation 2. Ordnung (vgl. z.B. [Lub03], S.75) und liefern daher für quadratische Funktionale sogar die exakte Ableitung.
Es gilt also bei gegebenem $\varepsilon > 0$:

$$(\nabla j(q), \tau q)_Q = j'(q)(\tau q) = \frac{j(q + \varepsilon\, \tau q) - j(q - \varepsilon\, \tau q)}{2\, \varepsilon} \tag{7.1}$$

und

$$\begin{aligned}(\nabla^2 j(q)\delta q, \tau q)_Q &= j''(q)(\delta q, \tau q) \\ &= \frac{1}{2\,\varepsilon}\left(\frac{j(q+\varepsilon\,\tau q+\varepsilon\,\delta q)-j(q+\varepsilon\,\tau q-\varepsilon\,\delta q)}{2\,\varepsilon}\right. \\ &\qquad \left.-\frac{j(q-\varepsilon\,\tau q+\varepsilon\,\delta q)-j(q-\varepsilon\,\tau q-\varepsilon\,\delta q)}{2\,\varepsilon}\right) \\ &= \frac{1}{4\,\varepsilon^2}(j(q+\varepsilon\,\tau q+\varepsilon\,\delta q)-j(q+\varepsilon\,\tau q-\varepsilon\,\delta q) \\ &\qquad -j(q-\varepsilon\,\tau q+\varepsilon\,\delta q)+j(q-\varepsilon\,\tau q-\varepsilon\,\delta q)). \qquad (7.2)\end{aligned}$$

Der wesentliche Grund, der gegen das Berechnen der Ableitungen mittels Differenzenquotienten spricht, ist der hohe Rechenaufwand:
Zunächst einmal wissen wir a-priori nicht, dass die Hesse-Matrix unabhängig von q ist. In der Darstellung (7.2) zumindest sieht es so aus, als ob eine Abhängigkeit von q existiert, da alle Funktionsauswertungen von j von q abhängig sind. Auch wenn wir aus Kapitel 4 bereits wissen, dass diese Abhängigkeit nicht existiert, müssen wir also eigentlich in jedem Newton-Schritt die Hesse-Matrix neu berechnen.

Des weiteren sind im Gegensatz zum Vorgehen in Kapitel 4 zum Aufstellen des Gradienten je Newton-Schritt nicht nur 2 parabolische Differentialgleichungen sondern $2 \cdot dim(Q)$ Differentialgleichungen zu lösen. Für die Berechnung der Hesse-Matrix nach (7.2) ist sogar die Lösung von $4 \cdot dim(Q)^2$ Zustandsgleichungen nötig, während bei der Verwendung von Satz 4.8 nur $dim(Q)$-mal die Tangentengleichung gelöst werden muss. Insgesamt müssen also

$$2n_{Newton}(dim(Q) + 2dim(Q)^2)$$

Differentialgleichungen gelöst werden.

Vorteil des Aufstellens der Ableitungen mittels Differenzenquotienten ist hingegen, dass nur noch voneinander unabhängige Vorwärtsgleichungen gelöst werden müssen. Dies ist algorithmisch natürlich weniger aufwändig und bietet bessere Möglichkeiten der Parallelisierung.

7.2 Ergebnisse

Wir wollen nun die beiden Beispiele vorstellen und die Ergebnisse unserer Rechnungen auswerten. Die Beispiele unterscheiden sich vor allem in der Dimension des Steuerungsraumes Q und in der Form der Steuerung (einmal steuern wir die rechte Seite, einmal die Anfangsbedingung). Zur Ausführung der Rechnungen wurde basierend auf der Software-Bibliothek deal.ii (http://www.dealii.org) ein Programm geschrieben, welches die dG(0)cG(1)-Diskretisierung der parabolischen Differentialgleichungen und die in Kapitel 4 vorgestellten Versionen des Newton-Algorithmus umsetzt. Gerechnet wurde auf einem Mehrbenutzer-System mit einem Pentium 2,8 GHz Prozessor und 2 GB Arbeitsspeicher.

7.2.1 Beispiele

Beispiel 7.1 Gegeben sei eine Referenzfunktion

$$\overline{u}(t; x_1, x_2) = e^{-25((x_1-t)^2+(x_2-t)^2)},$$

welche wir rekonstruieren wollen.

Hierbei stören wir die rechte Seite mit $q \in Q := \mathbb{R}$, wie folgt:

$$\begin{aligned} \partial_t u - \Delta u &= \frac{1}{2} \cdot q \cdot (\partial_t \overline{u} - \Delta \overline{u}) && \text{in } \Omega \times (0,1) \\ u &= 0 && \text{auf } \partial\Omega \times (0,1) \\ u(0) &= \overline{u}(0) && \text{in } \Omega. \end{aligned}$$

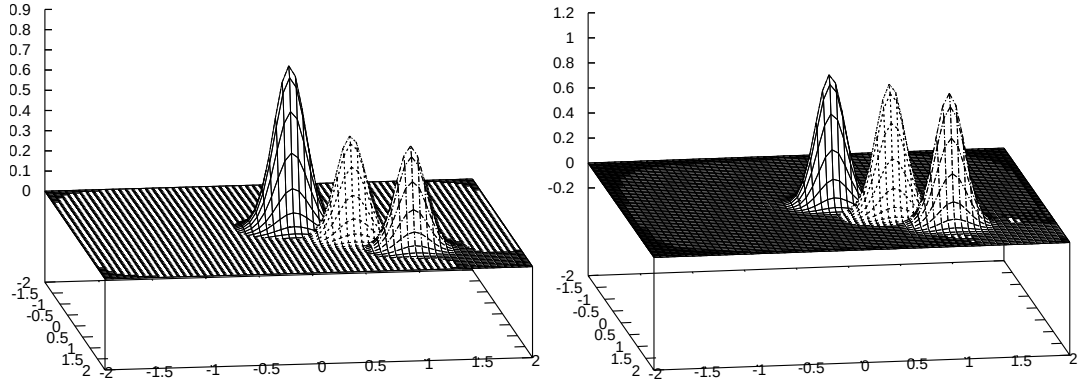
Wir rechnen dabei auf dem Gebiet $\Omega = [0,2] \times [0,2]$, um homogene Randbedingungen voraussetzen zu können.

Als Steuerungsfunktional wählen wir

$$J(q,u) = \int_0^1 \int_\Omega (u - \overline{u})^2 \, dx \, dt.$$

Das optimale q sollte also möglichst nah bei 2 liegen, um die Störung der rechten Seite zu beheben.

Die Lösung der Zustandsgleichung bei ungesteuertem und gesteuertem Zustand ist in Abbildung 7.1 dargestellt.



(a) (b)

Abbildung 7.1: Beispiel 7.1: Lösung der Zustandsgleichung zu den Zeitpunkten 0.01, 0.5 und 1; (a) ungesteuert, (b) gesteuert

Beispiel 7.2 Erneut geben wir uns eine Referenz-Funktion $\overline{u}(t;x)$ vor. In diesem Beispiel wollen wir jedoch die Anfangsbedingung mit $q \in Q := \mathbb{R}^5$ steuern. Insgesamt betrachten wir die Zustandsgleichung:

$$\begin{aligned} \partial_t u - \Delta u &= f && \text{in } \Omega \times (0,1) \\ u &= 0 && \text{auf } \partial\Omega \times (0,1) \\ u(0) &= \sum_{i=1}^{5} q_i \cdot g_i && \text{in } \Omega, \end{aligned}$$

wobei

- $\Omega = \{x \in \mathbb{R}^2 : \|x\| \leq 1\}$,
- $f = (\partial_t \overline{u} - \Delta \overline{u})$ und
- $g_i(x) = (1 - 0.5 \cdot \|x - \overline{x_i}\|)^{20}$ mit $\overline{x_1} = (0.5, 0.5)^T$, $\overline{x_2} = (-0.5, 0.5)^T$, $\overline{x_3} = (0.5, -0.5)^T$, $\overline{x_4} = (-0.5, -0.5)^T$ und $\overline{x_5} = (0.0, 0.0)^T$.

Dabei wählen wir als Referenzlösung $\overline{u}(t;x) = \frac{1}{\sqrt{4\pi t}} e^{\frac{-\|x\|^2}{4t}} e^{\frac{-1}{1-\|x\|^2}}$.

Das zu minimierende Steuerungsfunktional hat die Form:

$$J(q,u) = \int_0^1 \int_\Omega (u(T) - \overline{u}(T))^2 \, dx \, dt + \frac{\alpha}{2} \cdot \|q\|_Q,$$

wobei wir mit $\alpha = 10^{-4}$ rechnen werden.

Abbildung 7.2 zeigt die Lösung der Zustandsgleichung zu verschiedenen Zeitpunkten vor der Optimierung und nach Beendigung des Newton-Verfahrens.

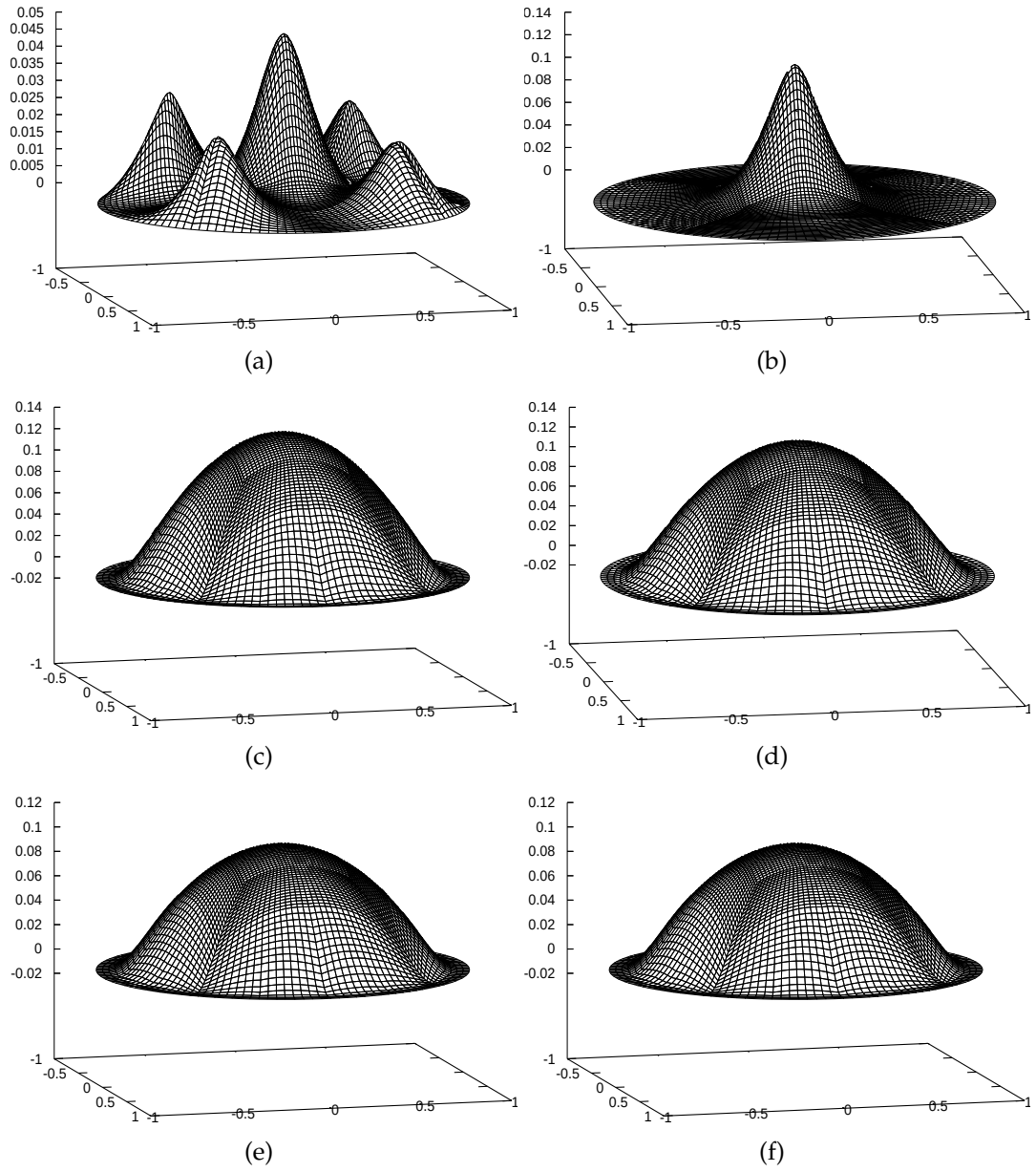


Abbildung 7.2: Beispiel 7.2: Lösung der Zustandsgleichung zu den Zeitpunkten 0.01, 0.5 und 1; (a),(c),(e) ungesteuert, (b),(d)(f) gesteuert

7.2.2 Auswertungen

Nun wollen wir anhand der beiden Beispiele die drei Möglichkeiten zur Durchführung des Newton-Verfahrens bezogen auf die Rechenzeit vergleichen. Zu diesem Zweck führen wir den Newton-Algorithmus ohne Adaptivität für die beiden Beispiele einmal

- in Form von **Algorithmus 4.2** (Die Hesse-Matrix wird zu Beginn einmal berechnet. Die Berechnung von Hesse-Matrix und Gradient erfolgt wie in Kapitel 4)
- in Form von **Algorithmus 4.3** (Es wird nicht die Hesse-Matrix berechnet, sondern nur Matrix-Vektor-Produkte, die für die Durchführung des CG-Verfahrens benötigt werden. Auch diese Berechnungen erfolgen wie in Kapitel 4.)
- mit Hilfe von **Differenzenquotienten** (Gradient und Hesse-Matrix werden in jedem Schritt mittels Differenzenquotienten neu berechnet.)

durch. Dabei vergleichen wir die Rechenzeiten für verschiedene zeitliche und räumliche Diskretisierungen.
Wir bezeichnen mit M die Anzahl der Zeitschritte und mit N die Zahl der räumlichen Freiheitsgrade. Als Abbruchkriterium für das Newton-Verfahren wurde die Bedingung $\|\nabla j(q^n)\| < TOL$ mit $TOL = 10^{-6}$ gewählt und die Angaben für die Laufzeit sind in Sekunden. n_{Newton} gibt die Anzahl der benötigten Newton-Schritte an.

Für Beispiel 7.1 ($dim(Q) = 1$) mit $q^0 = 1$ ergaben sich folgende Rechenzeiten:

		Algorithmus 4.2		Algorithmus 4.3		Diff.quotient	
M	N	Zeit	n_{Newton}	Zeit	n_{Newton}	Zeit	n_{Newton}
50	289	3.1	2	8.5	3	20.1	3
100	289	5.7	2	13.5	2	24.2	2
500	289	16.0	1	34.1	1	55.9	1
50	4225	53.5	2	145.6	3	234.6	3
100	4225	97.7	2	228.4	2	266.9	2
500	4225	263.2	1	563.3	1	575.8	1
50	66049	1235.7	2	3310.2	3	5276.2	3
100	66049	2093.5	2	4769.6	2	5748.1	2
500	66049	4852.0	1	10149.7	1	10634.8	1

Beispiel 7.2 ($dim(Q) = 5$) benötigte mehr Zeit ($q^0 = (1, 1, 1, 1)^T$):

		Algorithmus 4.2		Algorithmus 4.3		Diff.quotient	
M	N	Zeit	n_{Newton}	Zeit	n_{Newton}	Zeit	n_{Newton}
50	337	8.8	3	14.4	2	313.8	2
100	337	12.7	2	29.9	3	611.47	2
500	337	56.8	2	128.2	3	2686.67	2
50	5185	162.5	3	328.6	3	5282.83	2
100	5185	229.0	2	528.9	3	9782.35	2
500	5185	970.3	2	1770.1	2	44334.2	2
50	20609	785.1	3	1268.9	2	24743.6	2
100	20609	1094.1	2	1942.8	2	44143.4	2
500	20609	4363.5	2	7866.9	2	194780.8	2

Bei beiden Beispielen wurden höchstens 3 Newton-Schritte benötigt, bis die vorgegebene Toleranz unterschritten wurde. Und während für Beispiel 7.1 das Lösen des Gleichungssystems

$$\nabla^2 j(q)\delta q = -\nabla j(q)$$

dem Lösen einer linearen Gleichung entspricht (es gilt also immer $n_{CG} = 1$), waren bei Beispiel 7.2 in der Regel 2-3 CG-Schritte nötig.

Zu beachten ist noch, dass zur Auswertung des Abbruchkriteriums

$$\|\nabla j(q^n)\| < TOL$$

zunächst der Gradient $\|\nabla j(q^n)\|$ für die errechnete neue Steuerung q^n berechnet werden muss. Im Programm wird dies realisiert, indem ein weiterer Newton-Schritt gerechnet wird.

An beiden Tabellen ist zu erkennen, dass Algorithmus 4.2 eine kürzere Laufzeit als Algorithmus 4.3 hat und dieser wiederum schneller ist als die Berechnung mittels Differenzenquotienten. Diesen Zusammenhang wollen wir nun genauer an den Daten zu Beispiel 7.2 erläutern.
Wir haben in Kapitel 4 bereits festgestellt, dass der wesentliche Aufwand des Newton-Verfahrens im Lösen der parabolischen Differentialgleichungen liegt.

Deshalb wollen wir noch einmal wiederholen, wieviele Differentialgleichungen je Algorithmus zu lösen waren:

Algorithmus 4.2:	$dim(Q) + 2n^{(1)}_{Newton}$
Algorithmus 4.3:	$2n^{(2)}_{Newton}(1 + n_{CG})$
Differenzenquotienten:	$2n^{(3)}_{Newton}(dim(Q) + 2dim(Q)^2)$.

Dabei bezeichnen wir mit $n^{(1)}_{Newton}$ die Anzahl der benötigten Newton-Schritte bei Algorithmus 4.2, mit $n^{(2)}_{Newton}$ die Zahl der Newton-Schritte bei Algorithmus 4.3 und mit $n^{(3)}_{Newton}$ die Anzahl der nötigen Newton-Iterationen bei der Rechnung mittels Differenzenquotienten.

Wollen wir nun die Laufzeiten der Algorithmen 4.2 und 4.3 vergleichen, so betrachten wir den Quotienten

$$\frac{2n^{(2)}_{Newton}(1 + n_{CG})}{dim(Q) + 2n^{(1)}_{Newton}}. \tag{7.3}$$

Analog gilt für das Verhältnis der Laufzeit von Algorithmus 4.3 zur Berechnung mittels Differenzenquotienten:

$$\frac{2n^{(3)}_{Newton}(dim(Q) + 2dim(Q)^2}{2n^{(2)}_{Newton}(1 + n_{CG})}. \tag{7.4}$$

Da je nach Diskretisierung die Anzahl der Newton- und CG-Schritte bei allen drei Verfahren schwankt, kann man aus diesen Quotienten jedoch keine feste Zahl berechnen, die für alle Diskretisierungen das genaue Verhältnis angibt. Wir können aber feststellen, dass sich die Laufzeiten von Algorithmus 4.2 und Algorithmus 4.3 Gleichung (7.3) entsprechend etwa um den Faktor 1.7-3 unterscheiden und das Verhältnis (7.4) sich im Bereich 16-20 bewegt.

Wie bereits zu Beginn des Kapitels angekündigt, haben wir die Differenzenquotienten nicht nur zu Vergleichs- sondern auch zu Verifikationszwecken eingeführt. Sie liefern bei den Beispielen 7.1 und 7.2 unabhängig von der Wahl von ε den exakten Gradienten $\nabla_{diff} j(q)$ und die exakte Hesse-Matrix $\nabla^2_{diff} j(q)$. Einzig durch die numerische Lösung der Differentialgleichungen entsteht ein Fehler. Dieselbe Aussage trifft für die Darstellungen des Gradienten $\nabla j(q)$ und der Hesse-Matrix $\nabla^2 j(q)$ aus den Sätzen 4.6 und 4.9 zu. Abgesehen von diesen Diskretisierungsfehlern stimmen Gradient und Hesse-Matrix bei beiden Verfahren überein. Unter Verwendung der Bezeichnungen

$$e_1 := \|\nabla j(q) - \nabla_{diff} j(q)\|, \qquad e_2 := \|\nabla^2 j(q) - \nabla^2_{diff} j(q)\|.$$

ist in der folgenden Tabelle jeweils die Norm der Differenz von Gradient beziehungsweise Hesse-Matrix für verschiedene Diskretisierungen aufgelistet: Der größere Fehler für e_1 im Gegensatz zu e_2 erklärt sich durch die Art der

M	N	e_1	e_2
50	337	5.74224e-06	1.26547e-09
100	337	1.89186e-06	1.20703e-11
500	337	7.55209e-08	2.54497e-12
50	5185	5.7728e-06	3.21205e-09
100	5185	1.93685e-06	7.54323e-10
500	5185	1.16444e-07	3.80733e-12

Differentialgleichungen, die jeweils gelöst werden mussten. Während sowohl für $\nabla^2 j(q)$, $\nabla^2_{diff} j(q)$ als auch für $\nabla_{diff} j(q)$ nur unabhängige Vorwärtsgleichungen zu lösen waren, war für das Aufstellen von $\nabla j(q)$ nach dem Lösen der Zustandsgleichung, das Lösen des dualen Problems nötig. Dieses Rückwärtsproblem ist von der Lösung der Zustandsgleichung abhängig. Da diese aber bereits einen Diskretisierungsfehler enthält, ist der gesamte Fehler größer.

Zuletzt wollen wir noch darstellen, dass sich der Diskretisierungsfehler der Differentialgleichungen tatsächlich darin wiederspiegelt, dass der Gradient $\nabla j(q)$ von Newton-Schritt zu Newton-Schritt "kleiner" wird und also auch mehrere Newton-Iterationen nötig sind. Dafür betrachten wir erneut Beispiel 7.2. Bei 100 Zeitschritten und 5185 Freiheitsgraden im Ort ergibt sich:

Newton-Schritt	$\|\nabla j(q)\|$
0	4.09732e-04
1	1.57894e-05
2	7.83148e-07
3	3.91285e-08
4	1.55566e-09
5	7.85772e-11

Kapitel 8

Zusammenfassung und Ausblick

Wir haben in dieser Arbeit einen Algorithmus entwickelt, der das adaptive Lösen von linear-quadratischen Optimierungsproblemen mit parabolischen Differentialgleichungen ermöglicht.

Dabei haben wir gesehen, dass hierbei unabhängig von der Dimension des Steuerungsraumes Q zur Berechnung der benötigten Ableitungen ein Vorgehen nach Kapitel 4 gegenüber der Verwendung von Differenzenquotienten zu bevorzugen ist, da in diesem Fall der Rechenaufwand wesentlich geringer ist. Aus demselben Grund ist es leicht einzusehen, dass in Abhängigkeit von $dim(Q)$ unterschiedliche Ausführungen des Newton-Verfahrens, das den Kern des adaptiven Algorithmus darstellt, verwendet werden sollten.

Im Fall der linear-quadratischen Probleme, die in dieser Arbeit behandelt wurden, liefert das Newton-Verfahren bei exakten Daten und Rechnungen nach einem Schritt das exakte Ergebnis. Somit ist es also eigentlich ein überdimensioniertes Lösungsverfahren. Jedoch bietet dieser Ansatz viele Möglichkeiten für Erweiterungen der Problemstellung und wir konnten sehen, dass numerische Fehler dazu führen, dass mehrere Newton-Schritte nötig sind.

Durch die Herleitung und anschließende Verwendung von Fehlerschätzern, die die aus der Zeit- und Ortsdiskretisierung entstehenden Fehler trennen, war es schließlich möglich, die zeitliche und räumliche Diskretisierung adaptiv dem Optimierungsproblem anzupassen.

Wir haben also insgesamt gesehen, dass es lohnenswert ist, bei der Umsetzung des Newton-Verfahrens die Dimension des Steuerungsraumes Q zu berücksichtigen und die Fehlerschätzer aus Kapitel 6 zur Bestimmung einer optimalen Diskretisierung zu benutzen.

Abschließend wollen wir einige Punkte erwähnen, die in dieser Arbeit nicht untersucht wurden, deren Behandlung aber eine wichtige Erweiterung der vorgestellten Ideen darstellen würden.

Zunächst wäre, wie bereits erwähnt, eine Ausweitung der Problemstellung denkbar. So könnten beispielweise statt der homogenen Dirichlet-Bedingungen auch inhomogene Randbedingungen unterschiedlicher Art betrachtet werden. Weiterhin wäre es denkbar die vorgestellten Algorithmen so zu modifizieren, dass auch Randsteuerungsprobleme gelöst werden können. Auch nicht-lineare Zustandsgleichungen oder nicht-quadratische Steuerungsfunktionale könnten in Betracht gezogen werden.

Außerdem könnten der Fall $dim(Q) = \infty$ und eine damit einhergehende Diskretisierung des Steuerungsraumes Q sowie die Behandlung von Restriktionen an den Steuerungsraum Q untersucht werden.

Unabhängig von der Problemwahl könnten etwa die Auswirkungen eines anderen numerischen Lösungsverfahrens für die parabolischen Differentialgleichungen auf den Optimierungsalgorithmus und die Fehlerschätzer betrachtet werden. Für den Fall einer konvektionsdominanten Zustandsgleichung sollten außerdem Stabilisierungsverfahren berücksichtigt werden.

Neben diesen theoretischen Erweiterungsmöglichkeiten wäre außerdem eine Weiterentwicklung des im Rahmen dieser Arbeit erstellten Programmes wünschenswert, welches bereits die Auswertung der in Kapitel 6 vorgestellten Fehlerschätzer und die Verwendung von Adaptivität vorsieht. Dabei könnte etwa untersucht werden, ob der Zeitfehlerschätzer unabhängig von der Ortsdiskretisierung ist und entsprechend der Ortsfehlerschätzer unabhängig von der Zeitdiskretisierung, wie es die Theorie vermuten lässt. Auch wäre denkbar, bekannte Fehlerschätzer für parabolische Differentialgleichungen zur adaptiven Anpassung der Gitter zu verwenden und deren Auswirkungen auf das Steuerungsfunktional im Vergleich zum Vorgehen nach Kapitel 6 zu betrachten.

Literaturverzeichnis

[Alt92] ALT, H. W.: *Lineare Funktionalanalysis*. Berlin, Heidelberg, New York : Springer-Verlag, 1992

[BMV07] BECKER, R. ; MEIDNER, D. ; VEXLER, B.: Efficient Numerical Solution of Parabolic Optimization Problems by Finite Element Methods. In: *Optimization Methods and Software* (akzeptiert, 2007)

[BR01] BECKER, R. ; RANNACHER, R.: An optimal control approach to a posteriori error estimation in finite element methods. In: *Acta Numerica* (2001), S. 1–102

[Bra97] BRAESS, D.: *Finite Elemente*. Berlin, Heidelberg, New York : Springer-Verlag, 1997

[EJT85] ERIKSSON, K. ; JOHNSON, C. ; THOMÉE, V.: Time discretization of parabolic problems by the discontinuous Galerkin method. In: *RAIRO Modél. Math. Anal. Numér.* 19 (1985), S. 611–643

[Emm04] EMMRICH, E.: *Gewöhnliche und Operator-Differentialgleichungen*. Wiesbaden : Vieweg Verlag, 2004

[GGZ74] GAJEWSKI, H. ; GRÖGER, K. ; ZACHARIAS, K.: *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen*. Berlin : Akademie-Verlag, 1974

[Heu86] HEUSER, H.: *Funktionalanalysis*. Stuttgart : B. G. Teubner, 1986

[Hoh05] HOHAGE, T.: *Partielle Differentialgleichungen, Vorlesungsskript*. Vorlesungsskript Wintersemester 2004/2005, Göttingen, 2005

[LS68] LJUSTERNIK, L. ; SOBOLEW, W.: *Elemente der Funktionalanalysis*. Berlin : Akademie-Verlag, 1968

[Lub03] LUBE, G.: *Numerik instationärer partieller Differentialgleichungen*. Vorlesungsskript Wintersemester 2002/2003, Göttingen, 2003

[Lub05] LUBE, G.: *Numerische Mathematik I*. Vorlesungsskript Wintersemester 2004/2005, Göttingen, 2005

[Lub06] LUBE, G.: *Theorie und Numerik elliptischer Differentialgleichungen*. Vorlesungsskript Sommersemester 2006, Göttingen, 2006

[Lue69] LUENBERGER, D.: *Optimization by vector space methods*. New York, London, Sydney, Toronto : John Wiley & Sons, Inc., 1969

[MV07] MEIDNER, D. ; VEXLER, B.: Adaptive Space-Time Finite Element Methods for Parabolic Optimization Problems. In: *SIAM J. Control Optim.* 46 (2007), S. 116–142

[RY00] RYNNE, B. ; YOUNGSON, M.: *Linear Functional Analysis*. London : Springer-Verlag, 2000

[Saa96] SAAD, Y.: *Iterative methods for sparse linear systems*. Boston : PWS Publishing Company, 1996

[SV06] SCHMICH, M. ; VEXLER, B.: Adaptivity with dynamic meshes for space-time finite element discretizations of parabolic equations. (submitted, 2006)

[SW92] SCHABACK, R. ; WERNER, H.: *Numerische Mathematik*. Berlin, Heidelberg, New York : Springer-Verlag, 1992

[Tho97] THOMÉE, V.: *Galerkin Finite Element Methods for Parabolic Problems*. Berlin, Heidelberg : Springer-Verlag, 1997

[Trö05] TRÖLTZSCH, F.: *Optimale Steuerung partieller Differentialgleichungen*. Wiesbaden : Vieweg Verlag, 2005

[Zei90] ZEIDLER, E.: *Nonlinear functional analysis and its applications*. Bd. II/B: *Nonlinear monotone operators*. New York : Springer-Verlag, 1990

Danksagung

An dieser Stelle möchte ich all den Menschen herzlichst danken, die zum Entstehen dieser Diplomarbeit beigetragen haben.

Mein besonderer Dank gilt Herrn Professor Dr. Gert Lube für die Bereitstellung dieses interessanten Themas, die Zeit, die er sich immer für mich genommen hat, und die damit einhergehende hervoragende Betreuung während der Entstehung dieser Arbeit.

Weiterhin möchte ich mich bei Stephan Kramer bedanken, der mir aufgrund seiner Erfahrungen mit deal.ii in der Phase der Implementierung eine große Hilfe war und viele wertvolle Hinweise gegeben hat.

Ich danke meinen Eltern, die mir mein Studium erst ermöglicht haben, mir immer helfend mit Rat und Tat zur Seite standen und einfach für mich da sind.

Außerdem danke ich Christian Schuft für seine fachliche und seelisch-moralische Unterstützung.

Abschließend möchte ich mich bei meiner Schwester und meinen Freunden für das Verständnis und die Geduld, die sie gerade im letzten Abschnitt meiner Diplomarbeit für mich aufgebracht haben, bedanken.