Optimale Steuerung partieller Differentialgleichungen

Skript zur Vorlesung
im
Frühjahrsemester 2018

Helmut Harbrecht

Stand: 1. Juni 2018

Vorwort

Diese Mitschrift kann und soll nicht ganz den Wortlaut der Vorlesung wiedergeben. Sie soll das Nacharbeiten des Inhalts der Vorlesung erleichtern. Kapitel, die mit einem Stern markiert sind, beinhalten ergänzendes Material. Die Vorlesung orientiert sich hauptsächlich an dem unten genannten Buch von Fredi Tröltzsch.

Hilfreich, aber nicht notwendig, zum Verstehen der Vorlesung sind Kenntnisse aus der Numerischen Mathematik, wie man sie beispielsweise in folgenden Büchern findet:

- M. Hanke-Bourgeois: *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*, Teubner-Verlag
- R. Schaback und H. Wendland: *Numerische Mathematik*, Springer-Verlag
- J. Stoer und R. Bulirsch: *Numerische Mathematik I+II*, Springer-Verlag

Literatur zur Vorlesung:

- D. Braess: *Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*, Springer-Verlag
- F. Tröltzsch: *Optimale Steuerung partieller Differentialgleichungen*, Vieweg-Verlag

Inhaltsverzeichnis

1. Einführung

Gegeben sei die stationäre Wärmeleitungsgleichung in einem Gebiet $\Omega \in \mathbb{R}^d$:

$$-\Delta y = u \text{ in } \Omega, \quad y = 0 \text{ auf } \Gamma := \partial\Omega. \tag{1.1}$$

Dabei entspricht y der Temperaturverteilung und u der Wärmeleistungsdichte im Gebiet. Ist die Wärmeleistungsdichte u bekannt, so kann die Temperaturverteilung y im Gebiet berechnet werden. Häufig kann allerdings die Wärmeleistungsdichte innerhalb gewisser Grenzen gewählt werden, um die Temperatur im Gebiet zu steuern. Die Wärmeleitungsdichte heißt dann *Steuerung* des Problems. In der Praxis kann sie etwa durch Mikrowellenstrahlung oder elektromagnetische Induktion verwirklicht werden. Demgegenüber ist die Temperaturverteilung y die von der Steuerung u abhängige Größe, auf die man keinen direkten Einfluss hat. Sie ergibt sich als Lošung von (1.1) und wird *Zustand* genannt. Die zugehörige partielle Differentialgleichung (1.1) nennt man entsprechend *Zustandsgleichung*.

Um festzulegen, wann eine Wahl von u "besser" ist als eine andere, müssen wir eine Bewertung der Steuerung u und der durch sie erzeugten Temperaturverteilung y vornehmen. Diese Bewertung übernimmt ein *Zielfunktional* J, das minimiert werden soll. Ein wichtiges Beispiel für ein solches Zielfunktional ist der Abstand zu einer gewünschten Temperaturverteilung y_d bezüglich der $L^2(\Omega)$-Norm:

$$J(y,u) = \frac{1}{2}\int_\Omega |y - y_d|^2 \,\mathrm{d}\mathbf{x}. \tag{1.2}$$

Im allgemeinen führt ein Zielfunktional der Form (1.2) ohne weitere Bedingungen nicht auf eine wohlgestellte Aufgabe, da die Steuerung im optimalen Fall unbeschränkt sein kann. Daher muss sichergestellt werden, dass Steuerung beschränkt bleibt, etwa durch eine *punktweise Steuerbeschränkung*:

$$u_a \le u \le u_b \text{ in } \Omega. \tag{1.3}$$

In der Anwendung entspricht eine solche Beschänkung der Aufgabenstellung, denn Mikrowellenstrahler oder Induktionsquellen können keine beliebig große Leistungsdichten erzielen. Kann außerdem nur geheizt und nicht gekühlt werden, so folgt $u_a \equiv 0$.

Oft ist die Steuerung mit gewissen Kosten verbunden, beispielsweise durch den Verbrauch elektrischer Energie. Daher kann man das Zielfunktional um einen sogenannten *Kontrollkostenterm* zu ergänzen:

$$J(y,u) = \frac{1}{2}\int_\Omega |y - y_d|^2 \,\mathrm{d}\mathbf{x} + \frac{\lambda}{2}\int_\Omega |u|^2 \,\mathrm{d}\mathbf{x}. \tag{1.4}$$

Hierbei ist $\lambda > 0$ der *Kontrollkostenparameter*, der als relative Gewichtung der beiden Anteile im Zielfunktional verstanden werden kann. Wie wir noch sehen werden, führt die Minimierung dieses Funktional auch ohne die Steuerbeschränkung (1.3) für $\lambda > 0$ auf ein wohlgestelltes Problem.

Zusammenfassend werden wir uns mit folgender *Optimalsteuerungsaufgabe* befassen:

$$\left.\begin{array}{c} \text{minimiere } J(y,u) = \dfrac{1}{2}\displaystyle\int_\Omega |y - y_d|^2 \,\mathrm{d}\mathbf{x} + \dfrac{\lambda}{2}\int_\Omega |u|^2 \,\mathrm{d}\mathbf{x} \\ \text{unter den Nebenbedingungen } -\Delta y = u \text{ in } \Omega,\ y = 0 \text{ auf } \Gamma \\ \text{und } u_a \le u \le u_b \text{ in } \Omega. \end{array}\right\} \tag{1.5}$$

Da die Steuerung u auf dem Gebiet Ω wirkt, bezeichnet man (1.5) als eine Aufgabe mit *verteilter Steuerung*.

Zu beantwortende Fragestellungen für die Optimalsteuerungsaufgabe (1.5) sind:

- Existenz und Eindeutigkeit einer optimalen Lösung.
- Herleitung von notwendigen und hinreichenden Optimalitätsbedingungen, die die optimale Lösung charakterisieren.
- Konstruktion und Analyse von Optimierungsalgorithmen auf Basis der Optimalitätsbedingungen.
- Diskretisierung und Implementierung der Optimierungsalgorithmen.

2. Grundkonzepte

2.1 Endlichdimensionale Optimalsteuerungsaufgabe

Wir wollen zunächst die Lösung von Optimalsteuerungsaufgaben im endlich-dimensionalen Fall untersuchen. Dazu sei

- ein Zielfunktional $J : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$,
- eine invertierbare Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ und
- eine nichtleere Menge U_{ad}

gegeben. Gesucht ist eine Lösung $(\mathbf{y}, \mathbf{u}) \in \mathbb{R}^n \times \mathbb{R}^n$ der Optimalsteuerungsaufgabe

$$\left.\begin{array}{c} \text{minimiere } J(\mathbf{y}, \mathbf{u}) \\ \text{unter den Nebenbedingungen } \mathbf{Ay} = \mathbf{u} \text{ und } \mathbf{u} \in U_{ad} \end{array}\right\} \tag{2.1}$$

Definition 2.1 Das Paar $(\mathbf{y}^\star, \mathbf{u}^\star) \in \mathbb{R}^n \times \mathbb{R}^n$ heißt eine **lokal optimale Lösung** von (2.1), falls eine Umgebung $U(\mathbf{y}^\star, \mathbf{u}^\star)$ von $(\mathbf{y}^\star, \mathbf{u}^\star)$ existiert, so dass $J(\mathbf{y}^\star, \mathbf{u}^\star) \leq J(\mathbf{y}, \mathbf{u})$ gilt für alle $(\mathbf{y}, \mathbf{u}) \in \{(\mathbf{y}, \mathbf{u}) \in U(\mathbf{y}^\star, \mathbf{u}^\star) : \mathbf{Ay} = \mathbf{u} \text{ und } \mathbf{u} \in U_{ad}\}$. Es heißt **lokal optimal**, wenn die Ungleichung sogar für alle $(\mathbf{y}, \mathbf{u}) \in \{(\mathbf{y}, \mathbf{u}) \in \mathbb{R}^n \times \mathbb{R}^n : \mathbf{Ay} = \mathbf{u} \text{ und } \mathbf{u} \in U_{ad}\}$ erfüllt ist.

Wegen der Invertierbarkeit von $\mathbf{A}$, kann das lineare Gleichungssystem in (2.1) auch nach $\mathbf{y}$ aufgelöst werden:

$$\mathbf{y} = \mathbf{A}^{-1}\mathbf{u}. \tag{2.2}$$

Für jede Wahl von $\mathbf{u} \in \mathbb{R}^n$ gibt es also ein eindeutig bestimmtes $\mathbf{y} \in \mathbb{R}^n$, das die Gleichungsnebenbedingung in (2.1) erfüllt. Deshalb ist die Steuerung $\mathbf{u}$ die "eigentliche" Optimierungsvariable des Problems, da der Zustand $\mathbf{y}$ von $\mathbf{u}$ abhängt. Gleichung (2.2) führt auf den *Lösungsoperator* oder *Steuerungs-Zustands-Operator*

$$\mathbf{S} := \mathbf{A}^{-1} \in \mathbb{R}^{n\times n}.$$

Da für zulässige Paare $(\mathbf{y}, \mathbf{u})$ stets $\mathbf{y} = \mathbf{Su}$ gilt, können wir $\mathbf{y}$ eliminieren und die Optimierungsaufgabe nur anhängig von $\mathbf{u}$ formulieren. Wir definieren das *reduzierte Zielfunktional*

$$f(\mathbf{u}) := J(\mathbf{Su}, \mathbf{u})$$

und erhalten gleichwertig zu (2.1) die Optimierungsaufgabe

$$\text{minimiere } f(\mathbf{u}) \text{ unter der Nebenbedingung } \mathbf{u} \in U_{ad}, \tag{2.3}$$

bei der nur noch die Steuerung als Unbekannte auftritt.

2.2 Existenz optimaler Lösungen

Definition 2.2 Eine Steuerung $\mathbf{u}^\star \in U_{ad}$ heißt **lokal optimal** für (2.3), wenn es eine Umgebung $U(\mathbf{u}^\star)$ von $\mathbf{u}^\star$ gibt, so dass $f(\mathbf{u}^\star) \leq f(\mathbf{u})$ für alle $\mathbf{u} \in U(\mathbf{u}^\star) \cap U_{ad}$ gibt. Eine Steuerung $\mathbf{u}^\star$ heißt **global optimal**, wenn $f(\mathbf{u}^\star) \leq f(\mathbf{u})$ sogar für alle $\mathbf{u} \in U_{ad}$ gilt.

Satz 2.3 Es seien J stetig auf $\mathbb{R}^n \times U_{ad}$ und $\mathbf{A}$ invertierbar. Ferner sei U_{ad} nichtleer, beschränkt und abgeschlossen. Dann besitzt (2.3) mindestens eine global optimale Lösung $\mathbf{u}^\star$, das heißt, auch (2.1) besitzt mindestens ein global optimales Paar $(\mathbf{y}^\star, \mathbf{u}^\star)$.

Beweis. Es ist f stetig auf U_{ad}. Außerdem ist U_{ad} als beschränkte und abgeschlossene Menge eines endlichdimensionalen Raums kompakt. Deshalb nimmt f nach dem Satz von Weierstrass auf U_{ad} sein globales Minimum an. Es existiert also ein $\mathbf{u}^\star \in U_{ad}$ mit der Eigenschaft $f(\mathbf{u}^\star) = \min_{\mathbf{u} \in U_{ad}} f(\mathbf{u})$. □

Später werden wir nicht mehr so argumentieren können, weil in unendlichdimensionalen Räumen beschränkte und abgeschlossene Mengen nicht notwendig kompakt sind.

2.3 Notwendige Optimalitätsbedingungen

Ein zentrales Thema der Vorlesung ist die Frage, wie man optimale Lösungen finden kann. Zu diesem Zweck setzen wir nun voraus, dass J stetig partiell nach $\mathbf{y}$ und $\mathbf{u}$ differenzierbar ist. Aufgrund der Kettenregel ist dann auch $f(\mathbf{u}) = J(\mathbf{S}\mathbf{u}, \mathbf{u})$ stetig differenzierbar und wir erhalten die folgende *notwendige Bedingung* für die lokale Optimalität von $\mathbf{u}^\star$:

Satz 2.4 Ist U_{ad} konvex und ist $\mathbf{u}^\star$ eine lokal optimale Steuerung für die Optimierungsaufgabe (2.3), so gilt die Variationsungleichung

$$f'(\mathbf{u}^\star)(\mathbf{u} - \mathbf{u}^\star) \geq 0 \quad \text{für alle } \mathbf{u} \in U_{ad}. \tag{2.4}$$

Beweis. Da U_{ad} konvex ist, folgt aus $\mathbf{u} \in U_{ad}$ auch $(1-t)\mathbf{u}^\star + t\mathbf{u} \in U_{ad}$ für jedes $t \in [0,1]$. Wir können also f entlang der Verbindungstrecke von $\mathbf{u}$ und $\mathbf{u}^\star$ linearisieren. Da $\mathbf{u}^\star$ ein lokales Minimum ist, ergibt sich daher

$$0 \leq f\big((1-t)\mathbf{u}^\star + t\mathbf{u}\big) - f(\mathbf{u}^\star) = t f'(\mathbf{u}^\star)(\mathbf{u} - \mathbf{u}^\star) + t\, r(t)$$

mit $|r(t)| \to 0$ für $t \to 0$. Hieraus folgt

$$0 \leq f'(\mathbf{u}^\star)(\mathbf{u} - \mathbf{u}^\star) + r(t),$$

das ist für $t \to 0$ die Behauptung. □

Die Variationsungleichung drückt die Beobachtung aus, dass die Funktion f, ausgehend von einem Minimum $\mathbf{u}^\star$, in keiner Richtung fallen kann.

Beispiel 2.5 Im Fall des Funktionals

$$J(\mathbf{y}, \mathbf{u}) = \frac{1}{2}\|\mathbf{y} - \mathbf{y}_d\|^2 + \frac{\lambda}{2}\|\mathbf{u}\|^2. \tag{2.5}$$

folgt

$$f(\mathbf{u}) = \frac{1}{2}\|\mathbf{S}\mathbf{u} - \mathbf{y}_d\|^2 + \frac{\lambda}{2}\|\mathbf{u}\|^2.$$

Daher ergibt sich

$$\nabla f(\mathbf{u}) = \mathbf{S}^\top(\mathbf{S}\mathbf{u} - \mathbf{y}_d) + \lambda\mathbf{u},$$

dies bedeutet

$$f'(\mathbf{u})\mathbf{v} = \left[\mathbf{S}^\top(\mathbf{S}\mathbf{u} - \mathbf{y}_d) + \lambda\mathbf{u}\right]^\top\mathbf{v}.$$

△

In der Variationsungleichung (2.4) folgt mit Hilfe der Kettenregel, dass

$$\begin{aligned} f'(\mathbf{u}^\star)\mathbf{v} &= J_\mathbf{y}(\mathbf{S}\mathbf{u}^\star, \mathbf{u}^\star)\mathbf{S}\mathbf{v} + J_\mathbf{u}(\mathbf{S}\mathbf{u}^\star, \mathbf{u}^\star)\mathbf{v} \\ &= J_\mathbf{y}(\mathbf{S}\mathbf{u}^\star, \mathbf{u}^\star)\mathbf{A}^{-1}\mathbf{v} + J_\mathbf{u}(\mathbf{S}\mathbf{u}^\star, \mathbf{u}^\star)\mathbf{v} \\ &= \langle \mathbf{A}^{-\top}\nabla_\mathbf{y} J(\mathbf{S}\mathbf{u}^\star, \mathbf{u}^\star) + \nabla_\mathbf{u} J(\mathbf{S}\mathbf{u}^\star, \mathbf{u}^\star), \mathbf{v}\rangle. \end{aligned}$$

Folglich lässt sich die Variationsungleichung (2.4) in der expliziten Form

$$\langle \mathbf{A}^{-\top}\nabla_\mathbf{y} J(\mathbf{y}^\star, \mathbf{u}^\star) + \nabla_\mathbf{u} J(\mathbf{y}^\star, \mathbf{u}^\star), \mathbf{u} - \mathbf{u}^\star\rangle \geq 0 \quad \text{für alle } \mathbf{u} \in U_{ad} \tag{2.6}$$

schreiben.

2.4 Adjungierter Zustand

Dem Term

$$\mathbf{p}^\star := \mathbf{A}^{-\top}\nabla_\mathbf{y} J(\mathbf{y}^\star, \mathbf{u}^\star)$$

in der Variationsungleichung (2.6) kommt eine besondere Bedeutung zu. Er wird *adjungierte Variable* oder *adjungierter Zustand* genannt und lässt sich als Lösung der *adjungierten Gleichung*

$$\mathbf{A}^\top\mathbf{p}^\star := \nabla_\mathbf{y} J(\mathbf{y}^\star, \mathbf{u}^\star) \tag{2.7}$$

bestimmen, das heißt, als Lösung eines linearen Gleichungssystems mit $\mathbf{A}^\top$. Da $\mathbf{A} \in \mathbb{R}^{n\times n}$ als invertierbar vorausgesetzt wurde, ist auch $\mathbf{A}^\top$ invertierbar und die adjungierte Gleichung (2.7) ist eindeutig lösbar.

Beispiel 2.6 (Fortsetzung von Beispiel 2.5) Für das Funktional (2.5) erhalten wir die adjungierte Gleichung

$$\mathbf{A}^\top\mathbf{p}^\star = \mathbf{y}^\star - \mathbf{y}_d.$$

△

Mit dem adjungierten Zustand können wir die Variationsungleichung (2.6) vereinfacht schreiben gemäß

$$\langle \mathbf{p}^\star + \nabla_\mathbf{u} J(\mathbf{y}^\star, \mathbf{u}^\star), \mathbf{u} - \mathbf{u}^\star\rangle \geq 0 \quad \text{für alle } \mathbf{u} \in U_{ad}. \tag{2.8}$$

Zusammenfassend haben wir folgendes *Optimalitätssystem* für die drei unbekannten Vektoren $\mathbf{y}^\star$, $\mathbf{u}^\star$, $\mathbf{p}^\star$ hergeleitet, das zur Berechnung der optimalen Steuerung herangezogen werden kann:

$$\left.\begin{aligned} \mathbf{A}\mathbf{y} &= \mathbf{u}, \quad \mathbf{u} \in U_{ad}, \\ \mathbf{A}^\top \mathbf{p} &= \nabla_{\mathbf{y}} J(\mathbf{y}, \mathbf{u}), \\ \langle \mathbf{p} + \nabla_{\mathbf{u}} J(\mathbf{y}, \mathbf{u}), \mathbf{v} - \mathbf{u} \rangle &\geq 0 \quad \text{für alle } \mathbf{v} \in U_{ad}. \end{aligned}\right\} \tag{2.9}$$

Jede Lösung $(\mathbf{y}^\star, \mathbf{u}^\star)$ der Optimalsteuerungsaufgabe (2.1) muss diesem System genügen. Man beachte, dass sich im Spezialfall $U_{ad} = \mathbb{R}^n$ für die Variationsungleichung (2.8) die Gleichung

$$\mathbf{p}^\star + \nabla_{\mathbf{u}} J(\mathbf{y}^\star, \mathbf{u}^\star) = \mathbf{0}$$

ergibt, denn $\mathbf{u} - \mathbf{u}^\star$ kann jeden Wert $\mathbf{v} \in \mathbb{R}^n$ annehmen. Demnach ist das Optimalitätssystem (2.9) in diesem Fall ein reines im allgemeinen nichtlineares Gleichungssystem, das sich beispielsweise mit dem Newton-Verfahren lösen lässt.

Beispiel 2.7 (Fortsetzung von Beispiel 2.6) Für $U_{ad} = \mathbb{R}^n$ erhalten wir im Fall des Funktionals aus Beispiel 2.5 in Anbetracht von (2.8) die Gleichung

$$\mathbf{p}^\star = -\lambda \mathbf{u}^\star.$$

Die Lösung der Optimalsteuerungsaufgabe wird demnach beschrieben durch

$$\lambda \mathbf{A}\mathbf{y}^\star = -\mathbf{p}^\star, \quad \mathbf{A}^\top \mathbf{p}^\star = \mathbf{y}^\star - \mathbf{y}_d,$$

Dies bedeutet

$$\begin{bmatrix} \lambda \mathbf{A} & \mathbf{I} \\ \mathbf{I} & \mathbf{A}^\top \end{bmatrix} \begin{bmatrix} \mathbf{y}^\star \\ \mathbf{p}^\star \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{y}_d \end{bmatrix}$$

beziehungsweise

$$(\lambda \mathbf{A}^\top \mathbf{A} + \mathbf{I})\mathbf{y}^\star = \mathbf{y}_d.$$

△

2.5 Lagrange-Funktion

Der adjungierte Zustand lässt sich als Lagrange-Multiplikator interpretieren. Zu diesem Zweck benötigen wir die Lagrange-Funktion, welche sich aus der Zielfunktion durch "Ankoppeln" der Nebenbedingung $\mathbf{A}\mathbf{y} - \mathbf{u} = \mathbf{0}$ mit Hilfe des Lagrange-Multiplikators $\mathbf{p} \in \mathbb{R}^n$ ergibt:

Definition 2.8 Die Funktion $L : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ mit

$$L(\mathbf{y}, \mathbf{u}, \mathbf{p}) = J(\mathbf{y}, \mathbf{u}) - \langle \mathbf{A}\mathbf{y} - \mathbf{u}, \mathbf{p} \rangle$$

heißt **Lagrange-Funktion** des Problems (2.1).

Die Ungleichungsbeschränkungen $\mathbf{u} \in U_{ad}$ werden in dieser Lagrange-Funktion nicht berücksichtigt, sondern explizit beibehalten. Nachrechnen zeigt, dass das Optimalitätssystem (2.9) nun äquivalent ist zu

$$\left.\begin{aligned} \nabla_{\mathbf{p}} L(\mathbf{y}^\star, \mathbf{u}^\star, \mathbf{p}^\star) &= \mathbf{0}, \\ \nabla_{\mathbf{y}} L(\mathbf{y}^\star, \mathbf{u}^\star, \mathbf{p}^\star) &= \mathbf{0}, \\ \langle \nabla_{\mathbf{u}} L(\mathbf{y}^\star, \mathbf{u}^\star, \mathbf{p}^\star), \mathbf{v} - \mathbf{u}^\star \rangle \geq 0 \quad &\text{für alle } \mathbf{v} \in U_{ad}. \end{aligned}\right\} \tag{2.10}$$

Die adjungierte Gleichung ergibt sich also aus $\nabla_{\mathbf{y}} L(\mathbf{y}^\star, \mathbf{u}^\star, \mathbf{p}^\star) = \mathbf{0}$ und die Gleichungsnebenbedingung aus $\nabla_{\mathbf{p}} L(\mathbf{p}^\star, \mathbf{u}^\star, \mathbf{p}^\star) = \mathbf{0}$.

Bemerkung Die "Herleitung" des Optimalitätssystems (2.10) ist nicht rigoros, sondern ist nur ein formales Aufstellen der entsprechenden Gleichungen und Ungleichungen. Diese "formale" Lagrange-Technik ist jedoch nützlich, weil Optimalitätssysteme oftmals eine komplizierte Struktur haben. Man stellt mit ihrer Hilfe einen "Kandidaten" für ein Optimalitätssystem auf, dessen Notwendigkeit für lokale Optimalität dann noch gezeigt werden muss. △

2.6 Diskussion der Variationsungleichung

Die Menge U_{ad} ist oftmals durch obere und untere Schranken, sogenannte *Box-Beschränkungen*, gegeben. Deshalb wollen wir diese nun näher betrachten:

$$U_{ad} = \{\mathbf{u} \in \mathbb{R}^d : \mathbf{u}_a \leq \mathbf{u} \leq \mathbf{u}_b\}. \tag{2.11}$$

Hier sind $\mathbf{u}_a \leq \mathbf{u}_b$ fest vorgegebene Schranken und alle Ungleichungen sind komponentenweise zu verstehen.
Wir können nun die Variationsungleichung (2.8) schreiben als

$$\langle \mathbf{p}^\star + \nabla_{\mathbf{u}} J(\mathbf{y}^\star, \mathbf{u}^\star), \mathbf{u}^\star \rangle \leq \langle \mathbf{p}^\star + \nabla_{\mathbf{u}} J(\mathbf{y}^\star, \mathbf{u}^\star), \mathbf{u} \rangle \quad \text{für alle } \mathbf{u} \in U_{ad}.$$

Daher löst $\mathbf{u}^\star$ die lineare Optimierungsaufgabe

$$\langle \mathbf{p}^\star + \nabla_{\mathbf{u}} J(\mathbf{y}^\star, \mathbf{u}^\star), \mathbf{u}^\star \rangle = \min_{\mathbf{u} \in U_{ad}} \sum_{i=1}^{n} \left[\mathbf{p}^\star + \nabla_{\mathbf{u}} J(\mathbf{y}^\star, \mathbf{u}^\star)\right]_i u_i.$$

Aufgrund der speziellen Wahl von U_{ad} wird das Minimum genau dann angenommen, wenn in der Summe jede einzelne Komponente minimiert wird. Dies bedeutet, es gilt

$$\left[\mathbf{p}^\star + \nabla_{\mathbf{u}} J(\mathbf{y}^\star, \mathbf{u}^\star)\right]_i u_i^\star = \min_{u_{a,i} \leq u_i \leq u_{b,i}} \left[\mathbf{p}^\star + \nabla_{\mathbf{u}} J(\mathbf{y}^\star, \mathbf{u}^\star)\right]_i u_i$$

für alle $i = 1, \ldots, n$. Folglich ist

$$u_i^\star = \begin{cases} u_{b,i}, & \text{wo } \left[\mathbf{p}^\star + \nabla_{\mathbf{u}} J(\mathbf{y}^\star, \mathbf{u}^\star)\right]_i < 0, \\ u_{a,i}, & \text{wo } \left[\mathbf{p}^\star + \nabla_{\mathbf{u}} J(\mathbf{y}^\star, \mathbf{u}^\star)\right]_i > 0. \end{cases} \tag{2.12}$$

Für die Komponenten mit $\left[\mathbf{p}^\star + \nabla_{\mathbf{u}} J(\mathbf{y}^\star, \mathbf{u}^\star)\right]_i = 0$ bringt die Variationsungleichung keine Information. Allerdings liefert die Tatsache, dass die Komponente verschwindet, meist auch eine auswertbare Gleichung.

2.7 Formulierung als Karush-Kuhn-Tucker-System

Um die Box-Beschränkungen (2.11) behandeln zu können, führen wir die Größen

$$\begin{aligned} \boldsymbol{\mu}_a^\star &:= \left[\mathbf{p}^\star + \nabla_{\mathbf{u}} J(\mathbf{y}^\star, \mathbf{u}^\star)\right]^+ = \quad \max\left\{\mathbf{0}, \mathbf{p}^\star + \nabla_{\mathbf{u}} J(\mathbf{y}^\star, \mathbf{u}^\star)\right\}, \\ \boldsymbol{\mu}_b^\star &:= \left[\mathbf{p}^\star + \nabla_{\mathbf{u}} J(\mathbf{y}^\star, \mathbf{u}^\star)\right]^- = -\min\left\{\mathbf{0}, \mathbf{p}^\star + \nabla_{\mathbf{u}} J(\mathbf{y}^\star, \mathbf{u}^\star)\right\}, \end{aligned} \tag{2.13}$$

ein. Dabei ist das Maximum beziehungsweise Minimum jeweils komponentenweise zu verstehen.

Lemma 2.9 Es seien $\mathbf{u}^\star$ eine lokal optimale Lösung von (2.3) und U_{ad} von der Form (2.11). Weiter seien $\mathbf{y}^\star$ der optimale Zustand, $\mathbf{p}^\star$ der zugehörige adjungierte Zustand und $\boldsymbol{\mu}_a^\star$ und $\boldsymbol{\mu}_b^\star$ definiert wie oben. Dann gilt

$$\begin{aligned} &\boldsymbol{\mu}_a^\star \geq \mathbf{0}, \quad \mathbf{u}_a - \mathbf{u}^\star \leq \mathbf{0}, \quad \langle \mathbf{u}_a - \mathbf{u}^\star, \boldsymbol{\mu}_a^\star \rangle = 0, \\ &\boldsymbol{\mu}_b^\star \geq \mathbf{0}, \quad \mathbf{u}^\star - \mathbf{u}_b \leq \mathbf{0}, \quad \langle \mathbf{u}^\star - \mathbf{u}_b, \boldsymbol{\mu}_a^\star \rangle = 0. \end{aligned} \tag{2.14}$$

Beweis. Alle Ungleichungen in (2.14) ergeben sich direkt aus $\mathbf{u}^\star \in U_{ad}$ und der Definition von $\boldsymbol{\mu}_a^\star$ und $\boldsymbol{\mu}_b^\star$ in (2.13). Es verbleibt daher nur, die Gleichungen in (2.14) nachzuweisen. Dazu betrachten wir die erste Gleichung, die zweite beweist man analog. Sei $u_{a,i} < u_i^\star$. Dann folgt aus (2.12), dass $[\mathbf{p}^\star + \nabla_{\mathbf{u}} J(\mathbf{y}^\star, \mathbf{u}^\star)]_i \leq 0$ und daher ist $\mu_{a,i}^\star = 0$ nach Konstruktion von $\boldsymbol{\mu}_a^\star$. Dies impliziert $(u_{a,i} - u_i^\star)\mu_{a,i}^\star = 0$. Gilt umgekehrt $\mu_{a,i}^\star > 0$, dann ergibt sich aus (2.13), dass $[\mathbf{p}^\star + \nabla_{\mathbf{u}} J(\mathbf{y}^\star, \mathbf{u}^\star)]_i > 0$ und (2.12) liefert $u_i^\star = u_{a,i}$. Es ergibt sich also auch in diesem Fall $(u_{a,i} - u_i^\star)\mu_{a,i}^\star = 0$. □

Das System (2.14) ist das *Komplementaritätssystem*, während die Gleichungen in (2.14) *komplementäre Schlupfbedingungen* genannt werden.

Definition 2.10 Die Funktion $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ mit

$$\mathcal{L}(\mathbf{y}, \mathbf{u}, \mathbf{p}, \boldsymbol{\mu}_a, \boldsymbol{\mu}_b) = J(\mathbf{y}, \mathbf{u}) - \langle \mathbf{A}\mathbf{y} - \mathbf{u}, \mathbf{p} \rangle + \langle \mathbf{u}_a - \mathbf{u}, \boldsymbol{\mu}_a \rangle + \langle \mathbf{u} - \mathbf{u}_b, \boldsymbol{\mu}_b \rangle$$

heißt die **(erweiterte) Lagrange-Funktion** des Optimalsteuerungsaufgabe (2.1) mit zulässigen Steuerungen (2.11).

Im Gegensatz zur einfachen Lagrange-Funktion werden nun die Ungleichungsbeschränkungen durch zusätzliche Lagrange-Multiplikatoren $\boldsymbol{\mu}_a \in \mathbb{R}^n$ und $\boldsymbol{\mu}_b \in \mathbb{R}^n$ ebenfalls mit angekoppelt.

Korollar 2.11 Ist $\mathbf{u}^\star$ eine optimale Steuerung mit zugehörigem optimalen Zustand $\mathbf{y}^\star$ für (2.1) mit Box-Beschränkungen (2.11). Dann existieren ein adjungierter Zustand $\mathbf{p}^\star \in \mathbb{R}^n$ und Lagrange-Multiplikatoren $\boldsymbol{\mu}_a^\star, \boldsymbol{\mu}_b^\star \in \mathbb{R}^n$ derart, dass folgendes Optimalitätssystem

erfüllt ist:

$$\left.\begin{aligned} \nabla_{\mathbf{y}}\mathcal{L}(\mathbf{y}^\star,\mathbf{u}^\star,\mathbf{p}^\star,\boldsymbol{\mu}_a^\star,\boldsymbol{\mu}_b^\star) &= \mathbf{0},\\ \nabla_{\mathbf{u}}\mathcal{L}(\mathbf{y}^\star,\mathbf{u}^\star,\mathbf{p}^\star,\boldsymbol{\mu}_a^\star,\boldsymbol{\mu}_b^\star) &= \mathbf{0},\\ \nabla_{\mathbf{p}}\mathcal{L}(\mathbf{y}^\star,\mathbf{u}^\star,\mathbf{p}^\star,\boldsymbol{\mu}_a^\star,\boldsymbol{\mu}_b^\star) &= \mathbf{0},\\ \mathbf{u}^\star \geq \mathbf{u}_a, \quad \boldsymbol{\mu}_a^\star \geq \mathbf{0}, \quad \langle \mathbf{u}_a - \mathbf{u}^\star, \boldsymbol{\mu}_a^\star\rangle &= 0,\\ \mathbf{u}^\star \leq \mathbf{u}_b, \quad \boldsymbol{\mu}_b^\star \geq \mathbf{0}, \quad \langle \mathbf{u}^\star - \mathbf{u}_b, \boldsymbol{\mu}_a^\star\rangle &= 0. \end{aligned}\right\} \tag{2.15}$$

Beweis. Die letzten beiden Gleichungen sind das Komplimentaritätssystem (2.14). Ferner zeigt nachrechnen zeigt, dass die beiden Gleichungen $\nabla_{\mathbf{p}}\mathcal{L}(\mathbf{y}^\star,\mathbf{u}^\star,\mathbf{p}^\star,\boldsymbol{\mu}_a^\star,\boldsymbol{\mu}_b^\star) = \mathbf{0}$ und $\nabla_{\mathbf{y}}\mathcal{L}(\mathbf{y}^\star,\mathbf{u}^\star,\mathbf{p}^\star,\boldsymbol{\mu}_a^\star,\boldsymbol{\mu}_b^\star) = \mathbf{0}$ gerade der Zustandsgleichung und der adjungierten Gleichung entsprechen. Um schließlich die zweite Gleichung zu zeigen, beachten wir

$$\nabla_{\mathbf{u}}\mathcal{L}(\mathbf{y}^\star,\mathbf{u}^\star,\mathbf{p}^\star,\boldsymbol{\mu}_a^\star,\boldsymbol{\mu}_b^\star) = \nabla_{\mathbf{u}}J(\mathbf{y}^\star,\mathbf{u}^\star) + \mathbf{p}^\star - \boldsymbol{\mu}_a^\star + \boldsymbol{\mu}_b^\star.$$

Definieren wir nun $\boldsymbol{\mu}_a^\star$ und $\boldsymbol{\mu}_b^\star$ gemäß (2.13), dann gilt aber gerade der Gleichung

$$\nabla_{\mathbf{u}}J(\mathbf{y}^\star,\mathbf{u}^\star) + \mathbf{p}^\star - \boldsymbol{\mu}_a^\star + \boldsymbol{\mu}_b^\star = \mathbf{0}.$$

□

Das System (2.15) wird auch als die *Karush-Kuhn-Tucker-Bedingungen*, kurz *KKT-Bedingungen* zur Optimalsteuerungsaufgabe (2.1) bezeichnet.

2.8 Ausblick auf partielle Differentialgleichungen

Die Diskussion von Optimalsteuerungsproblemen mit partiellen Differentialgleichungen läuft analog ab, wobei eine partielle Differentialgleichung an die Stelle der Gleichungsnebenbedingung $\mathbf{Ay} = \mathbf{u}$ tritt. $\mathbf{A}$ stellt dann einen dann einen Differentialoperator dar. Aus der Matrix $\mathbf{S} = \mathbf{A}^{-1}$ wird der Lösungsoperator der Differentialgleichung. Außerdem kann man nicht mehr alles im $\mathbb{R}^n$ diskutieren, sondern muss in unendlich-dimensionalen Funktionenräumen argumentieren. Folgende Tabelle bietet einen Überblick der Gegenstücke:

endlichdimensional	kontinuierlich
Zustand $\mathbf{y} \in \mathbb{R}^n$	Zustand y in geeignetem Funktionenraum Y
Steuerung $\mathbf{u} \in \mathbb{R}^n$	Steuerung u in geeignetem Funktionenraum U
Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$	Differentialoperator $A : Y \to U$
Lösungsoperator $\mathbf{S} \in \mathbb{R}^{n\times n}$	Lösungsoperator $S : U \to Y$ der Differentialgleichung
Skalarprodukt $\langle\cdot,\cdot\rangle$ im $\mathbb{R}^n$	Skalarprodukt $(\cdot,\cdot)$ im Hilbert-Raum der Steuerungen

3. Existenz optimaler Steuerungen

3.1 Schwache Konvergenz

Zum Nachweis der Existenz optimaler Steuerungen benötigen wir den Begriff der schwachen Kompaktheit. Dieser wird uns als Ersatz für den Satz von Heine-Borel dienen, um den endlichdimensionalen Existenzsatz (2.3) ins Unendlichdimensionale zu übertragen.

Definition 3.1 Es sei U ein reeller Banach-Raum mit Dual U'. Eine Folge $\{u_n\}_{n=1}^{\infty}$ von Elementen aus U heisst **schwach konvergent**, falls ein $u \in U$ derart existiert, dass

$$f(u_n) \overset{n\to\infty}{\longrightarrow} f(u)$$

für alle Funktionale $f \in U'$. Wir schreiben kurz

$$u_n \rightharpoonup u.$$

Bemerkungen

1. In endlichdimensionalen Räumen stimmen starke und schwache Konvergenz überein.
2. Der schwache Grenzwert ist eindeutig:

$$u_n \rightharpoonup u \text{ und } u_n \rightharpoonup v \quad \Longrightarrow \quad u = v.$$

3. Jede (stark) gegen $u \in U$ konvergente Folge $\{u_n\}_{n=1}^{\infty}$ konvergiert auch schwach gegen u:

$$u_n \to u \quad \Longrightarrow \quad u_n \rightharpoonup u.$$

4. Im Hilbert-Raum H mit Skalarprodukt $(\cdot,\cdot)$ ist die schwache Konvergenz gleichbedeutend mit

$$(u_n, v) \to (u, v)$$

für alle $v \in H$. Gilt $u_n \to u$ und $v_n \rightharpoonup v$, so folgt auch $(u_n, v_n) \to (u, v)$. Das Skalarprodukt einer schwach konvergenten mit einer stark konvergenten Folge konvergiert demnach gegen das Skalarprodukt der Grenzelemente.

5. Schwach konvergente Folgen sind beschränkt:

$$u_n \rightharpoonup u \quad \Longrightarrow \quad \|u_n\| \le C.$$

6. In einem Hilbert-Raum implizieren schwache Konvergenz und Normkonvergenz die starke Konvergenz:

$$u_n \rightharpoonup u \text{ und } \|u_n\| \to \|u\| \quad \Longrightarrow \quad u_n \to u.$$

△

Beispiel 3.2 Im Hilbert-Raum $L^2([0,2\pi])$ betrachten wir die Funktionenfolge

$$u_n = \frac{1}{\sqrt{2\pi}} \sin(nx), \quad n \in \mathbb{N}. \tag{3.1}$$

Diese Folge stellt ein Orthonormalsystem dar, denn es ist

$$(u_n, u_m)_{L^2([0,2\pi])} = \frac{1}{2\pi} \int_0^{2\pi} \sin(nx)\sin(mx)\,\mathrm{d}x = \begin{cases} 1, & n = m, \\ 0, & n \neq m. \end{cases}$$

Da für solche Systeme die Besselsche Ungleichung

$$\sum_{n=1}^{\infty} |(f, u_n)_{L^2([0,2\pi])}|^2 \leq \|f\|^2_{L^2([0,2\pi])}$$

gilt, ist für ein beliebiges $f \in L^2([0,2\pi])$ demnach $\sum_{n=1}^{\infty} |(f, u_n)_{L^2([0,2\pi])}|^2 < \infty$ und es folgt

$$(f, u_n)_{L^2([0,2\pi])} \to 0 = (f, 0)_{L^2([0,2\pi])}.$$

Dies bedeutet $u_n \rightharpoonup 0$. Weil sich andererseits

$$\|u_n - u_m\|^2_{L^2([0,2\pi])} = \underbrace{\|u_n\|^2_{L^2([0,2\pi])}}_{=1} + 2\underbrace{(u_n, u_m)_{L^2([0,2\pi])}}_{=0} + \underbrace{\|u_m\|^2_{L^2([0,2\pi])}}_{=1} = 2$$

für $n \neq m$ ergibt, ist $\{u_n\}_{n=1}^{\infty}$ keine Cauchy-Folge und somit nicht stark konvergent. △

Das Beispiel illustriert, dass es schwach gegen 0 konvergente Folgen gibt, deren Elemente alle auf der Oberfläche der Einheitskugel liegen, also insbesondere nicht stark gegen 0 konvergieren.

Definition 3.3 Eine Abbildung $F : U \to V$ zwischen zwei reellen Banach-Räumen U und V heißt **schwach folgenstetig**, wenn aus der schwachen Konvergenz einer beliebigen Folge $\{u_n\}_{n=1}^{\infty}$ von Elementen aus U gegen $u \in U$ die schwache Konvergenz der Bildfolge $\{F(u_n)\}_{n=1}^{\infty}$ gegen $F(u)$ folgt.

Dass nachfolgende Beispiel zeigt, dass stetige Funktionen nicht unbedingt schwach folgenstetig sein müssen.

Beispiel 3.4 Das Funktional $f(u) = \|u\|_{L^2([0,2\pi])}$ ist im Hilbert-Raum $L^2([0,2\pi])$ nicht schwach folgenstetig. Denn betrachten wir wieder die Folge $\{u_n\}_{n=1}^{\infty}$ mit (3.1), so folgt $u_n \rightharpoonup 0$, aber es ist

$$1 = f(u_n) = \|u\|_{L^2([0,2\pi])} \to 1 \neq 0 = \|0\|_{L^2([0,2\pi])} = f(0).$$

Die Norm im $L^2([0,2\pi])$ ist demnach nicht schwach folgenstetig. Diese Aussage gilt sogar für jeden unendlichdimensionalen Hilbert-Raum. △

Definition 3.5 Eine Teilmenge M des reellen Banach-Raums U heißt **schwach folgenabgeschlossen**, wenn aus $\{u_n\}_{n=1}^{\infty} \subset M$ mit $u_n \rightharpoonup u$ folgt $u \in M$. Sie heißt **relativ schwach folgenkompakt**, wenn jede Folge $\{u_n\}_{n=1}^{\infty} \subset M$ eine schwach konvergente Teilfolge besitzt, und **schwach folgenkompakt**, wenn sie zusätzlich schwach folgenabgeschlossen ist.

Eine schwach folgenabgeschlossene Menge ist stets auch (stark) abgeschlossen. Da jede stark konvergente Folge auch schwach konvergiert, es also viel mehr schwach konvergente Folgen gibt, ist die Umkehrung im allgemeinen falsch. Die Oberfläche der Einheitskugel im Raum $L^2([0, 2\pi])$ ist abgeschlossen, aber gemäß Beispiel 3.2 nicht schwach folgenabgeschlossen.

3.2 Konvexität

Definition 3.6 Ein im reellen Banach-Raum U definiertes Funktional f heißt **konvex**, falls

$$f\big(tu + (1-t)v\big) \leq tf(u) + (1-t)f(v)$$

für alle $t \in (0, 1)$ und alle $u, v \in U$ ist. Es heißt **strikt konvex**, falls das strenge Ungleichheitszeichen gilt. Eine Teilmenge $C \subset U$ heißt konvex, wenn mit zwei beliebigen Elementen $u, v \in C$, auch die Konvexkombination $tu + (1-t)v$ für alle $t \in (0, 1)$ zu C gehört.

Wir benötigen folgende zwei Sätze, für deren Beweis wir auf die einschlägige Fachliteratur verweisen. Der erste folgt aus dem Satz von Hahn-Banach, der zweite aus dem Trennungssatz.

Satz 3.7 Jede beschränkte Menge eines reflexiven Banach-Raums ist relativ schwach folgenkompakt.

Satz 3.8 Jede konvexe und abgeschlossene Menge ist schwach folgenabgeschlossen.

Die Kombination dieser beiden Sätzen liefert sofort das unendlichdimensionale Analogon zum Satz von Heine-Borel: Jede beschränkte, konvexe und abgeschlossene Menge in einem reflexivem Banach-Raum ist schwach folgenkompakt. Eine weitere Konsequenz ist das unendlichdimensionale Analogon zum Satz von Bolzano-Weierstrass:

Korollar 3.9 Jede beschränkte Folge eines reflexiven Banach-Raums besitzt eine schwach konvergente Teilfolge.

Beweis. Ist $\{u_n\}_{n=1}^{\infty}$ eine beschränkte Folge des Banach-Raums U, dann existiert ein $C > 0$ derart, dass $\|u_n\| \leq C$ für alle $n \in \mathbb{N}$. Die Menge $M = \{u \in U : \|u\| \leq C\}$ ist beschränkt, konvex und abgeschlossen, und damit auch schwach folgenkompakt. Demnach besitzt $\{u_n\}_{n=1}^{\infty} \subset M$ eine schwach konvergente Teilfolge. □

Satz 3.10 Jedes in einem Banach-Raum U konvexe und stetige Funktional f ist schwach unterhalbstetig, das heißt, es gilt

$$u_n \rightharpoonup u \quad \Longrightarrow \quad \liminf_{n\to\infty} f(u_n) \geq f(u).$$

Beispiel 3.11 Die Norm ist schwach unterhalbstetig, denn sie ist stetig und aufgrund der Dreiecksungleichung auch konvex:

$$\|tu + (1-t)v\| \leq t\|u\| + (1-t)\|v\| \quad \text{für alle } u, v \in U \text{ und } t \in (0,1).$$

△

3.3 Existenzaussagen

Wir untersuchen nun, ob für das Beispielproblem (1.5) eine Lösung existiert. Diese Frage ist nicht nur von theoretischem Interesse, denn gegen was soll eine numerische Lösung bei feiner werdender Diskretisierung konvergieren, wenn die Aufgabe gar keine Lösung besitzt?

Definition 3.12 Seien U ein reeller Banach-Raum, $f : U \to \mathbb{R}$ und $U_{ad} \subset U$ gegeben. Eine Lösung $u^\star \in U_{ad}$ heißt **global optimal**, wenn für alle $u \in U_{ad}$ gilt

$$f(u^\star) \leq f(u).$$

Satz 3.13 Seien $U_{ad} \neq \emptyset$ eine konvexe, beschränkte und abgeschlossene Teilmenge des reflexiven Banach-Raums U und $f : U \to \mathbb{R}$ ein konvexes und stetiges Funktional. Existiert $c > -\infty$ derart, dass

$$f(u) \geq c \quad \text{für alle } u \in U_{ad},$$

dann besitzt die Optimierungsaufgabe

$$\min_{u\in U_{ad}} f(u) \tag{3.2}$$

eine global optimale Lösung $u^\star \in U_{ad}$. Ist f strikt konvex, dann ist die optimale Lösung eindeutig.

Beweis. Es ist $f(u) \geq c$ auf U_{ad}. Daher existiert

$$\underline{f} = \inf_{u \in U_{ad}} f(u).$$

Es gibt also eine Folge $\{u_n\}_{n=1}^{\infty} \subset U_{ad}$ mit $f(u_n) \to \underline{f}$. Weil U_{ad} beschränkt ist, besitzt $\{u_n\}_{n=1}^{\infty}$ gemäß Korollar 3.9 eine schwach konvergente Teilfolge. Wir nehmen ohne Einschränkung der Allgemeinheit an, dass es sich dabei um die ganze Folge handelt. Folglich gibt es ein $u^\star \in U$ mit $u_n \rightharpoonup u^\star$. Da die Menge U_{ad} konvex und abgeschlossen ist, ist sie nach Satz 3.8 auch schwach folgenabgeschlossen. Damit ist $u^\star \in U_{ad}$, also zulässig.

Wir zeigen nun, dass $u^\star$ auch globales Minimum ist. Da das Funktional f nach Voraussetzung stetig und konvex ist, ist es nach Satz 3.10 auch schwach unterhalbstetig. Aus $u_n \rightharpoonup u^\star$ folgt demnach

$$\underline{f} = \lim_{n \to \infty} f(u_n) \geq f(u^\star).$$

Da $u^\star \in U_{ad}$ und $\underline{f}$ das Infimum ist, muss $f(u^\star) = \underline{f}$ gelten.

Die Eindeutigkeit von $u^\star$ im Fall von strikter Konvexität von f sieht man vermittels Widerspruchsannahme: sei $v^\star \in U_{ad}$ eine weitere globale Lösung von (3.2) mit $u^\star \neq v^\star$. Aufgrund der strikten Konvexität von f folgt dann für beliebiges $t \in (0,1)$

$$f\big(tu^\star + (1-t)v^\star\big) < tf(u^\star) + (1-t)f(v^\star) = \underline{f}. \tag{3.3}$$

Setzen wir etwa $t = 1/2$, so ergäbe sich ein kleinerer Funktionswert als $\underline{f}$, was ein Widerspruch darstellt. □

Bemerkungen

1. Anhand von (3.3) mit "$\leq$" anstelle von "$<$" erkennt man, dass die Menge der globalen Optima im Fall eines konvexen Zielfunktionals konvex ist.
2. Es genügt in Satz 3.13 die Konvexität von f nur auf U_{ad} vorauszusetzen.

△

Lemma 3.14 Seien U und H Hilbert-Räume, $S : U \to H$ linear und stetig, $y_d \in H$ und $\lambda \geq 0$. Dann ist das Funktional

$$f(u) := \frac{1}{2}\|Su - y_d\|_H^2 + \frac{\lambda}{2}\|u\|_U^2$$

konvex. Ist zudem $\lambda > 0$, dann ist f strikt konvex.

Beweis. Für alle $t \in (0,1)$ gilt

$$\begin{aligned}
\|tu + (1-t)v\|_U^2 &= t^2\|u\|_U + 2t(1-t)(u,v)_U + (1-t)^2\|v\|_U^2 \\
&= t\|u\|_U^2 + (1-t)\|v\|_U^2 \\
&\quad - (t-t^2)\|u\|_U^2 + 2t(1-t)(u,v)_U - \big(\underbrace{(1-t) - (1-t)^2}_{=t-t^2}\big)\|v\|_V^2 \\
&= t\|u\|_U^2 + (1-t)\|v\|_U^2 - (t-t^2)\|u-v\|_U^2.
\end{aligned}$$

Wegen $t^2 < t$ ergibt sich daraus

$$\|tu + (1-t)v\|_U^2 \leq t\|u\|_U^2 + (1-t)\|v\|_U^2,$$

das heißt, die Funktion $\|\cdot\|_U^2$ ist konvex. Genauso ist auch die Funktion $\|\cdot\|_H^2$ konvex und deshalb gilt

$$\begin{aligned} f\big(tu+(1-t)v\big) &\le \frac{t}{2}\|Su-y_d\|_H^2 + \frac{\lambda t}{2}\|u\|_U^2 + \frac{1-t}{2}\|Sv-y_d\|_H^2 + \frac{\lambda(1-t)}{2}\|v\|_U^2 \\ &= tf(u)+(1-t)f(v). \end{aligned}$$

Die strikte Konvexität folgt aus der Tatsache, dass für $u \neq v$ sogar gilt

$$\|tu+(1-t)v\|_U^2 < t\|u\|_U^2 + (1-t)\|v\|_U^2.$$

□

Bemerkung Am Beweis von Lemma 3.14 fällt auf, dass f auch im Fall $\lambda = 0$ strikt konvex ist, sofern S injektiv ist. Denn im Falle $u \neq v$ folgt dann auch $Su \neq Sv$, woraus wie oben die strikte Konvexität von f folgt:

$$\begin{aligned} &\overbrace{\big\|S\big(tu+(1-t)v\big)-y_d\big\|_H^2}^{=2f(tu+(1-t)v)} \\ &\quad= t^2\|Su-y_d\|_H^2 + 2t(1-t)(Su-y_d, Sv-y_d)_H + (1-t)^2\|Sv-y_d\|_H^2 \\ &\quad= t\|Su-y_d\|_H^2 + (1-t)\|Sv-y_d\|_H^2 - (t-t^2)\|S(u-v)\|_H^2 \\ &\quad< t\underbrace{\|Su-y_d\|_H^2}_{=2f(u)} + (1-t)\underbrace{\|Sv-y_d\|_H^2}_{=2f(v)}. \end{aligned}$$

△

Wir wollen nun die Ergebnisse aus Satz 3.13 auf das Optimalsteuerungsproblem (1.5) anwenden. Dazu seien Ω ein beschränktes Gebiet mit Lipschitz-glattem Rand und $U := L^2(\Omega)$. Unter dieser Voraussetzung existiert zu jedem $u \in L^2(\Omega)$ eine eindeutige schwache Lösung $y = y(u) \in H_0^1(\Omega)$ der Poisson-Gleichung. Der Raum $Y = H_0^1(\Omega)$ wird dementsprechend *Zustandsraum* genannt.

Die Abbildung $G : L^2(\Omega) \to H_0^1(\Omega)$, welche jeder Steuerung u einen Zustand $y = y(u)$ zuordnet, ist linear und stetig. Da wir allerdings im Zielfunktional nur $y \in L^2(\Omega)$ benutzen, ist der *Steuerungs-Zustands-Operator* $S : L^2(\Omega) \to L^2(\Omega)$ gegeben durch

$$S = E \circ G,$$

wobei $E : H_0^1(\Omega) \to L^2(\Omega)$ den Einbettungsoperator von $H_0^1(\Omega)$ in $L^2(\Omega)$ bezeichnet, der jeder Funktion $y \in H_0^1(\Omega)$ dieselbe Funktion $y \in L^2(\Omega)$ zuordnet. Dieser ist wegen

$$\|Ey\|_{L^2(\Omega)} = \|y\|_{L^2(\Omega)} \le \|y\|_{H^1(\Omega)}$$

stetig, weshalb S ebenfalls stetig ist.

Definition 3.15 Eine Steuerung $u^\star \in U_{ad}$ heißt **(global) optimal**, und $y^\star = y^\star(u^\star)$ zugehöriger **optimaler Zustand**, wenn für alle $u \in U_{ad}$ und $y = y(u) \in Y$ gilt

$$J(u^\star, y^\star) \le J(u, y).$$

Um Satz 3.13 anwenden zu können, benötigen wir noch das folgende Resultat über die Menge der zulässigen Zustände.

Lemma 3.16 Es seien $u_a, u_b \in L^2(\Omega)$. Dann ist die Menge der zulässigen Steuerungen

$$U_{ad} = \{u \in L^2(\Omega) : u_a \leq u \leq u_b \text{ fast überall}\}$$

beschränkt, konvex und abgeschlossen.

Beweis. Für alle $u \in U_{ad}$ gilt $\|u\|_{L^2(\Omega)} \leq \|u_a\|_{L^2(\Omega)} + \|u_b\|_{L^2(\Omega)} < \infty$, das heißt, U_{ad} ist beschränkt.
Die Abgeschlossenheit zeigen wir indirekt. Angenommen, $\{u_n\}_{n=1}^{\infty} \subset U_{ad}$ ist eine konvergente Folge mit Grenzelement $u \notin U_{ad}$. Ohne Beschränkung der Allgemeinheit sei die untere Schranke $u_a \leq u$ verletzt. Dann existiert eine Menge $M \subset \Omega$ mit positivem Maß, so dass $u < u_a$ fast überall in M gilt. Es folgt

$$0 < (u_a - u, 1)_{L^2(M)} = \lim_{n\to\infty} \underbrace{(u_a - u_n, 1)_{L^2(M)}}_{\leq 0} \leq 0.$$

Dies ist ein Widerspruch und es folgt $u \in U_{ad}$.
Die Konvexität folgt schließlich trivialerweise aus der Struktur von U_{ad}. □

Satz 3.17 (Existenz und Eindeutigkeit) Das Optimalsteuerungsproblem (1.5) besitzt eine optimale Steuerung $u^\star \in L^2(\Omega)$, welche im Fall $\lambda > 0$ auch eindeutig ist.

Beweis. Wir wählen $U = L^2(\Omega)$ und definieren das reduzierte Zielfunktional

$$f(u) = \frac{1}{2}\|Su - y_d\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\|u\|^2_{L^2(\Omega)}.$$

Es ist nach unten durch Null beschränkt und stetig. Da es mit $H = L^2(\Omega)$ die in Lemma 3.14 geforderte Struktur besitzt, ist es auch konvex. Ferner ist U_{ad} gemäß Lemma 3.16 beschränkt, konvex und abgeschlossen. Damit sind alle Voraussetzungen von Satz 3.13 erfüllt, so dass mindestens eine global optimale Lösung $u^\star \in L^2(\Omega)$ existiert. Da f strikt konvex ist für $\lambda > 0$, ergibt sich in diesem Fall sogar die Eindeutigkeit. □

Bemerkung Wir haben bereits bemerkt, dass f auch dann strikt konvex ist, falls $S : L^2(\Omega) \to L^2(\Omega)$ injektiv ist. Dies ist bei der Poisson-Gleichung gerade der Fall. Wenn $y = 0$ ist, dann folgt aus der schwachen Formulierung

$$0 = \int_\Omega \langle \nabla y, \nabla v\rangle \,\mathrm{d}\mathbf{x} = \int_\Omega uv \,\mathrm{d}\mathbf{x} \quad \text{für alle } v \in H^1_0(\Omega).$$

Dies bedeutet, es ist auch $u = 0$ und somit S injektiv. △

Wir wollen uns nun mit dem Fall beschäftigen, dass die Zustände auch unbeschränkt sein dürfen, also etwa $U_{ad} = L^2(\Omega)$ gilt.

Satz 3.18 Ist $\lambda > 0$, so besitzt das Optimalsteuerungsproblem (1.5) auch dann genau eine optimale Steuerung, wenn $U_{ad} \neq \emptyset$ nur konvex und abgeschlossen ist.

Beweis. Es sei $\underline{c} = \inf_{u \in U_{ad}} f(u) \geq 0$ das Infimum von $f(u)$. Da $|f(u)| \to \infty$ für $\|u\|_{L^2(\Omega)} \to \infty$ gilt, existiert ein $r > 0$ derart, dass gilt

$$f(u) > \underline{c} + 1 \quad \text{für alle } \|u\|_{L^2(\Omega)} > r. \tag{3.4}$$

Wir zerlegen nun U_{ad} in die Menge $M = \overline{B_r(0)} \cap U_{ad}$ und $N = U_{ad} \setminus M$. Wegen (3.4) ist das Minimum auf der Menge M zu suchen, welche beschränkt, konvex, und angeschlossen. Gemäß Satz 3.17 existiert dort aber eine eindeutige optimale Lösung. □

4. Optimalitätsbedingungen erster Ordnung

4.1 Differenzierbarkeit in Banach-Räumen

Es seien U und V reelle Banach-Räume und $F : U \to V$ eine Abbildung.

Definition 4.1 Existiert zu $u, v \in U$ der Grenzwert

$$\delta F(u)[v] = \lim_{t \to 0+} \frac{f(u+tv) - f(u)}{t}$$

in V, so heißt dieser die **Richtungsableitung** von F an der Stelle u in Richtung v und F heißt **richtungsdifferenzierbar** an der Stelle u in Richtung v.

Die Abbildung $v \mapsto \delta F(u)[v]$ braucht nicht linear zu sein, etwa bei $F(u) = \max\{u, 0\}$ oder $F(u) = |u|$ auf $U = V = \mathbb{R}$. Sie ist aber immer positiv homogen, das heißt, es gilt

$$\delta F(u)[\lambda v] = \lambda \delta F(u)[v] \quad \text{für alle } \lambda \geq 0.$$

Definition 4.2 Existiert die Richtungsableitung in $u \in U$ für alle Richtungen $v \in U$ und ist

$$\delta F(u)[v] = Av$$

mit einem Operator $A \in \mathcal{L}(U, V)$, so heißt A **Gâteaux-Ableitung** von F an der Stelle u. Die Funktion F ist dann **Gâteaux-differenzierbar** an der Stelle u.

Eine Gateaux-differenzierbare Funktion ist demnach richtungsdifferenzierbar mit linearer und stetiger Ableitung. Die Gateaux-Ableitung kann man daher vermittels Richtungsableitungen berechnen. Ist speziell $F : U \to \mathbb{R}$ Gateaux-differenzierbar, so folgt $\delta F(u) \in U'$.

Beispiele 4.3

- Ist $F : \mathbb{R}^n \to \mathbb{R}$ stetig differentierbar, so stimmen Richtungs- beziehungsweise Gâteaux-Ableitung mit der bekannten Ableitung überein:

$$\delta F(\mathbf{u})[\mathbf{v}] = F'(\mathbf{u})\mathbf{v} = \nabla F(\mathbf{u})^\top \mathbf{v}.$$

- Ist $U = C([0,1])$ und $f(u) = \sin\big(u(1)\big) \in V = \mathbb{R}$. Berechne zunächst an irgendeiner Stelle die Richtungsableitung in Richtung $v \in C([0,1])$:

$$\begin{aligned}\lim_{t\to 0+} \frac{f(u+tv)-f(u)}{t} &= \lim_{t\to 0+} \frac{\sin\big(u(1)+tv(1)\big)-\sin\big(u(1)\big)}{t}\\ &= \frac{\mathrm{d}}{\mathrm{d}t}\sin\big(u(1)+tv(1)\big)\Big|_{t=0}\\ &= \cos\big(u(1)+tv(1)\big)v(1)\Big|_{t=0}\\ &= \cos\big(u(1)\big)v(1).\end{aligned}$$

  Die Richtungsableitung ist linear in v und wegen

$$|\cos\big(u(1)\big)v(1)| = \big|\cos\big(u(1)\big)\big||v(1)| \le \|v\|_{C([0,1])}$$

  auch stetig von $C([0,1])$ nach $\mathbb{R}$. Demnach ist f Gateaux-differenzierbar mit Ableitung

$$\delta f(u)[v] = \cos\big(u(1)\big)v(1).$$

△

Definition 4.4 Die Funktion $F : U \to V$ heißt an der Stelle $u \in U$ **Fréchet-differenzierbar**, wenn es ein $A \in \mathcal{L}(U,V)$ und eine Abbildung $r : U \to V$ gibt mit der Eigenschaft

$$F(u+v) = F(u) + Av + r(v), \tag{4.1}$$

wobei das Restglied der Abschätzung

$$\frac{\|r(v)\|_V}{\|v\|_U} \to 0 \quad \text{für } v \to 0 \tag{4.2}$$

genüge. Wir schreiben $F'(u) = A$ für die **Fréchet-Ableitung** von F an der Stelle u.

Lemma 4.5 Ist $F : U \to V$ an einer Stelle $u \in U$ Fréchet-differenzierbar, dann ist F dort auch Gâteaux-differenzierbar und Gâteaux- und Fréchet-Ableitung stimmen überein.

Beweis. Sei $v \in U$ eine beliebige, aber feste Richtung und $A \in \mathcal{L}(U,V)$ die Fréchet-Ableitung von F an der Stelle u. Mit (4.1) erhalten wir für $t > 0$

$$\left\|\frac{F(u+tv)-F(u)}{t} - Av\right\|_V = \frac{\|r(tv)\|_V}{t} = \frac{\|r(tv)\|_V}{\|tv\|_V}\|v\|_V.$$

Aus (4.2) folgt $\|r(tv)\|_V/\|tv\|_V \to 0$ für $t \to 0+$ und damit die Gâteaux-Differenzierbarkeit von F in u mit $\delta F(u)[v] = Av$. □

Jede Fréchet-differenzierbare Funktion ist demnach auch Gâteaux-differenzierbar. Zusätzlich haben wir allerdings die Restgliedabschätzung (4.2).

Beispiele 4.6

- Jeder Operator $A \in \mathcal{L}(U,V)$ ist überall Fréchet-differenzierbar mit Restglied Null, denn es ist
$$A(u+v) = Au + Av + 0.$$
Für gegebenes $z \in H^1(\Omega)$ ist etwa der Operator
$$A : H^1(\Omega) \to \mathbb{R}, \quad u \mapsto Au = \int_\Omega \langle \nabla z, \nabla u \rangle \, \mathrm{d}\mathbf{x}$$
linear, der wegen
$$|Au| \le \|z\|_{H^1(\Omega)} \|u\|_{H^1(\Omega)}$$
auch stetig ist. Die Fréchet-Ableitung lautet demnach
$$F'(u)v = Av = \int_\Omega \langle \nabla z, \nabla v \rangle \, \mathrm{d}\mathbf{x}.$$
- Die Gâteaux-Ableitung
$$\delta f(u)[v] = \cos\big(u(1)\big)v(1)$$
von $f(u) = \sin\big(u(1)\big)$ ist sogar die Fréchet-Ableitung. Es gibt nämlich gemäß der Taylor-Entwicklung ein ξ zwischen $u(1)$ und $u(1)+v(1)$, so dass gilt
$$\frac{\big|\sin\big(u(1)+v(1)\big)\big) - \sin\big(u(1)\big) - \cos\big(u(1)\big)v(1)\big|}{\|v\|_{C([0,1])}}$$
$$= \frac{1}{2}\frac{\big| -\sin(\xi)\big(v(1)\big)^2\big|}{\|v\|_{C([0,1])}} \le \frac{1}{2}\frac{\big|\big(v(1)\big)^2\big|}{\|v\|_{C([0,1])}} \to 0$$
für $\|v\|_{C([0,1])} \to 0$.
- In einem Hilbert-Raum ist $f(u) = \|u\|^2$ Fréchet-differenzierbar mit $f'(u)v = 2(u,v)$.

△

Für Fréchet-differenzierbare Abbildungen gilt die *Kettenregel*. Sind U, V, W Banach-Räume und $F : U \to V$ und $G : V \to W$ an der Stellen u beziehungsweise $F(u)$ Fréchet-differenzierbar, dann ist auch
$$E : U \to W, \quad u \mapsto E(u) = G\big(F(u)\big)$$
Fréchet-differenzierbar mit
$$E'(u) = G'\big(F(u)\big)F'(u).$$

Beispiel 4.7 Es seien U und H reelle Hilbert-Räume, $z \in H$ fest, $S \in \mathcal{L}(U,H)$ und
$$E(u) = \|Su - z\|_H^2.$$
Es ist $E(u) = G\big(F(u)\big)$ also die Verkettung von $G(v) = \|v\|_H^2$ und $F(u) = Su - z$. Es gilt
$$G'(v)w = 2(v,w)_H, \quad F'(u)w = Sw.$$
Die Kettenregel ergibt
$$E'(u)w = \big(2v, F'(u)w\big)_H = 2\big(Su - z, F'(u)w\big)_H = 2\big(S^\top(Su-z), w\big)_U.$$
Dabei ist $S^\top : H \to U$ der *adjungierte Operator* zu S, gegeben durch
$$(S^\top h, u)_U = (h, Su)_H \quad \text{für alle } u \in U \text{ und } h \in H.$$

△

4.2 Optimierungsprobleme im Hilbert-Raum

Es sei U ein reeller Banach-Raum, $U_{ad} \subset U$ eine konvexe Menge und $f : U_{ad} \to \mathbb{R}$ ein reellwertiges Funktional. Wir werden im folgenden für die Aufgabe

$$\text{minimiere } f(u) \text{ unter der Nebenbedingung } u \in U_{ad} \tag{4.3}$$

lokal optimale Lösungen charakterisieren.

Definition 4.8 Ein $u^\star \in U_{ad}$ ist eine **lokal optimale Lösung** für (4.3), wenn es ein $r > 0$ gibt, so dass

$$f(u^\star) \leq f(u) \quad \text{für alle } u \in U_{ad} \cap B_r(u^\star). \tag{4.4}$$

Satz 4.9 (notwendige Optimalitätsbedingung) Es sei U ein reeller Banach-Raum, $U_{ad} \subset U$ eine konvexe Menge und $u^\star$ eine lokal optimale Lösung von (4.3). Darüberhinaus sei $f : U_{ad} \to \mathbb{R}$ auf U_{ad} richtungsdifferenzierbar in $u^\star$ in alle Richtungen v, für die es ein $u \in U_{ad}$ gibt mit $v = u - u^\star$. Dann gilt die Variationsungleichung

$$\delta f(u^\star)[u - u^\star] \geq 0 \quad \text{für alle } u \in U_{ad}. \tag{4.5}$$

Beweis. Sei $u \in U_{ad}$ beliebig, aber fest. Dann ist die Konvexkombination

$$u(t) = u^\star + t(u - u^\star)$$

wieder in U_{ad} für alle $t \in [0, 1]$. Wird t hinreichend klein gewählt, so ist $u(t) \in B_r(u^\star)$, wobei der Radius r so gewählt sei, dass (4.4) gilt. Es folgt $f\big(u(t)\big) \geq f(u^\star)$, beziehungweise

$$\frac{1}{t}\Big[f\big(u^\star + t(u - u^\star)\big) - f(u^\star)\Big] \geq 0$$

Da f in $u^\star$ richtungsdifferenzierbar ist, erhalten wir für $t \to 0+$ das Behauptete. □

Bemerkung In Satz 4.9 wurde nicht gefordert, dass f konvex ist. Ausschlaggebend für die Variationsungleichung ist lediglich die Konvexität der zulässigen Menge U_{ad}. Die Aussage des Satzes ist daher für eine große Klasse von Optimalsteuerungsproblemen anwendbar. △

Ist zusätzlich f ebenfalls konvex ist, so ist die Variationsungleichung (4.5) auch hinreichend für die lokale Optimalität:

Satz 4.10 (hinreichende Optimalitätsbedingung) Sei U ein reeller Banach-Raum, $U_{ad} \subset U$ konvex und $f : U_{ad} \to \mathbb{R}$ auf U_{ad} konvex und richtungsdifferenzierbar in alle Richtungen. Dann ist jedes $u^\star \in U_{ad}$, das die Variationsungleichung (4.5) erfüllt, eine global optimale Lösung der Aufgabe (4.3).

Beweis. Erfüllt $u^\star \in U_{ad}$ die Variationsungleichung (4.5), so folgt für beliebiges $u \in U_{ad}$ aufgrund der Konvexität von f,

$$f\big(tu + (1-t)u^\star\big) \leq tf(u) + (1-t)f(u^\star),$$

dass

$$\frac{f\big(u^\star + t(u-u^\star)\big) - f(u^\star)}{t} \leq f(u) - f(u^\star).$$

Durch den Grenzübergang $t \to 0+$ ergibt sich hieraus

$$f(u) - f(u^\star) \geq \delta f(u^\star)[u - u^\star].$$

Nach Voraussetzung ist die rechte Seite ≥ 0, was schließlich die Behauptung liefert. □

Wir wollen nun die sehr allgemeine Aufgabe (4.3) konkretisieren, um die obigen Ergebnisse auf unser Optimalsteuerungsproblem (1.5) anwenden zu können. Dazu betrachten wir das Optimierungsproblem

$$\begin{gathered}\text{minimiere } f(u) = \frac{1}{2}\|Su - y_d\|_H + \frac{\lambda}{2}\|u\|_U^2 \\ \text{unter der Nebenbedingung } u \in U_{ad},\end{gathered} \tag{4.6}$$

wobei U und H im Hilbert-Räume, $S \in \mathcal{L}(U,H)$ und $y_d \in H$ seien. Das Funktional f heißt wieder das *reduzierte Zielfunktional*.

Satz 4.11 Es seien U und H Hilbert-Räume und $U_{ad} \subset U$ eine konvexe Menge. Ferner seien $S \in \mathcal{L}(U,H)$, $y_d \in H$ und $\lambda \geq 0$. Dann gilt ist $u^\star \in U_{ad}$ genau dann eine global optimale Lösung des Optimierungsproblems (4.6), wenn die Variationsungleichung

$$\big(S^\top(Su^\star - y_d) + \lambda u^\star, u - u^\star\big)_U \geq 0 \quad \text{für alle } u \in U_{ad} \tag{4.7}$$

erfüllt ist.

Beweis. Nach Beispielen 4.6 und 4.7 ist die Fréchet-Ableitung von f durch

$$f'(u^\star)v = \big(S^\top(Su^\star - y_d), v\big)_H + \lambda(u^\star, v)_H$$

gegeben. Die Aussage folgt daher direkt aus den Sätzen 4.9 und 4.10. □

Wenn $U_{ad} = U$ gilt, so ist die Variationsungleichung (4.7) äquivalent zu

$$S^\top(Su^\star - y_d) + \lambda u^\star = 0.$$

Bemerkung Satz (4.11) macht keine Aussage zur Existenz eines Minimums. Daher wird auch nicht die Abgeschlossenheit von U_{ad} benötigt. △

4.3 Optimale Steuerung der Poisson-Gleichung

Wir betrachten nun unser Optimalsteuerungsproblem (1.5). Hier ist S der Steuerungs-Zustands-Operator $S : L^2(\Omega) \to L^2(\Omega)$, der durch $S(u) = y$ gegeben ist. Damit hat (1.5) genau die Form (4.6), wobei $H = U = L^2(\Omega)$ ist. Damit ist Satz 4.11 anwendbar, und $u^\star \in U$ ist genau dann global optimale Lösung von (1.5), wenn Variationsungleichung (4.7) gilt. Da f strikt konvex ist, ist das globale Optimum eindeutig. Zu bestimmen bleibt aber noch der Operator $S^\top : L^2(\Omega) \to L^2(\Omega)$. Wir werden sehen, dass dieser Operator als Lösungsoperator der *adjungierten Gleichung* aufgefasst werden kann.

Lemma 4.12 Es seien $u, z \in L^2(\Omega)$ und $y, p \in H_0^1(\Omega)$ die schwachen Lösungen von

$$\begin{aligned} -\Delta y &= u & -\Delta p &= z & &\text{in } \Omega, \\ y &= 0 & p &= 0 & &\text{auf } \Gamma. \end{aligned}$$

Dann gilt

$$(z, y)_{L^2(\Omega)} = (p, u)_{L^2(\Omega)}.$$

Beweis. Wir schreiben die Variationsformulierungen für beide Randwertprobleme auf. Für $y \in H_0^1(\Omega)$ folgt mit Testfunktion $p \in H_0^1(\Omega)$

$$\int_\Omega \langle \nabla y, \nabla p \rangle \, \mathrm{d}\mathbf{x} = \int_\Omega up \, \mathrm{d}\mathbf{x}.$$

Umgekehrt gilt für $p \in H_0^1(\Omega)$ mit Testfunktion $y \in H_0^1(\Omega)$

$$\int_\Omega \langle \nabla p, \nabla y \rangle \, \mathrm{d}\mathbf{x} = \int_\Omega zy \, \mathrm{d}\mathbf{x}.$$

Da die linken Seiten gleich sind, sind es auch die rechten Seiten, was zu zeigen war. □

Lemma 4.13 Im Fall des Optimalsteuerungsproblems (1.5) ist der adjungierte Steuerungs-Zustands-Operator $S^\top : L^2(\Omega) \to L^2(\Omega)$ gegeben durch

$$S^\top z = p,$$

wobei $p \in H_0^1(\Omega)$ die schwache Lösung der Poisson-Gleichung

$$-\Delta p = z \text{ in } \Omega, \quad p = 0 \text{ auf } \Gamma$$

ist.

Beweis. Der Operator $S^\top$ ist festgelegt durch

$$(z, Su)_{L^2(\Omega)} = (S^\top z, u)_{L^2(\Omega)} \quad \text{für alle } u, z \in L^2(\Omega).$$

Wenden wir die Aussage aus Lemma 4.12 an, dann folgt

$$(z, Su)_{L^2(\Omega)} = (z, y)_{L^2(\Omega)} = (p, u)_{L^2(\Omega)} \quad \text{für alle } u, z \in L^2(\Omega).$$

Da die Zuordnung $z \mapsto p$ linear und stetig ist von $L^2(\Omega)$ nach $H_0^1(\Omega) \hookrightarrow L^2(\Omega)$, haben wir $S^\top z \mapsto p$ bewiesen. □

Die Konstruktion von $S^\top$ durch Lemma 4.13 ist nicht sehr intuitiv. Mit Hilfe der formalen Lagrange-Technik wird sich die Form der partiellen Differentialgleichung für $S^\top$ leicht ermitteln lassen.

Bemerkung Der Steuerungs-Zustands-Operator war definiert gemäß $S = E \circ G$, wobei $G : L^2(\Omega) \to H_0^1(\Omega)$ der Lösungsoperator der Poisson-Gleichung ist und $E : H_0^1(\Omega) \hookrightarrow L^2(\Omega)$ der Einbettungsoperator von $H_0^1(\Omega)$ in $L^2(\Omega)$ ist. Es ist also $S^\top = G^\top \circ E^\top$ mit $G^\top : H^{-1}(\Omega) \to L^2(\Omega)$ und $E^\top : L^2(\Omega) \hookrightarrow H^{-1}(\Omega)$. △

Definition 4.14 Die eindeutige schwache Lösung der **adjungierten Gleichung**

$$-\Delta p = y - y_d \text{ in } \Omega, \quad p = 0 \text{ auf } \Gamma \tag{4.8}$$

heißt der zu y gehörige **adjungierte Zustand**.

Die rechte Seite $y - y_d$ der adjungierten Gleichung liegt in $L^2(\Omega)$. Demnach existiert genau eine Lösung $p \in H_0^1(\Omega)$ der adjungierten Gleichung. Wegen Lemma 4.13 gilt

$$S^\top(Su - y_d) = S^\top(y - y_d) = p.$$

Daher lautet die Variationsungleichung (4.7)

$$(p^\star + \lambda u^\star, u - u^\star)_{L^2(\Omega)} \geq 0 \quad \text{für alle } u \in U_{ad}.$$

Zusammen mit Satz 4.11 ergibt sich hieraus das folgende Korollar:

Korollar 4.15 (Variationsungleichung) Ist $u^\star \in U_{ad}$ eine optimale Steuerung des Problems (1.5) und $y^\star$ der zugehörige optimale Zustand, dann existiert genau eine schwache Lösung $p^\star \in H_0^1(\Omega)$ der adjungierten Gleichung (4.8), so dass

$$\int_\Omega (p^\star + \lambda u^\star)(u - u^\star)\,\mathrm{d}\mathbf{x} \geq 0 \quad \text{für alle } u \in U_{ad} \tag{4.9}$$

gilt. Umgekehrt ist jedes $u^\star \in U_{ad}$ optimal, das mit dem zugehörigen Zustand $y^\star = y(u^\star)$ und der Lösung $p^\star$ von (4.8) der obigen Variationsungleichung genügt.

Eine Steuerung $u^\star \in U_{ad}$ ist demnach genau dann optimal für (1.5), wenn sie gemeinsam mit $y^\star$ und $p^\star$ das folgende *Optimalitätssystem* erfüllt:

$$\left.\begin{aligned} -\Delta y^\star = u^\star \qquad -\Delta p^\star = y^\star - y_d \qquad & \text{in } \Omega, \\ y^\star = 0 \qquad\qquad p^\star = 0 \qquad\qquad & \text{auf } \Gamma, \\ (p^\star + \lambda u^\star, u - u^\star)_{L^2(\Omega)} \geq 0 \quad & \text{für alle } u \in U_{ad}. \end{aligned}\right\} \tag{4.10}$$

Lemma 4.16 Die Variationsungleichung (4.9) gilt genau dann, wenn

$$u^\star(\mathbf{x}) \begin{cases} = u_a(\mathbf{x}), & \text{wo } p^\star(\mathbf{x}) + \lambda u^\star(\mathbf{x}) > 0, \\ \in [u_a(\mathbf{x}), u_b(\mathbf{x})], & \text{wo } p^\star(\mathbf{x}) + \lambda u^\star(\mathbf{x}) = 0, \\ = u_b(\mathbf{x}), & \text{wo } p^\star(\mathbf{x}) + \lambda u^\star(\mathbf{x}) < 0, \end{cases} \tag{4.11}$$

für fast alle $\mathbf{x} \in \Omega$ erfüllt ist.

Beweis. Wir setzen

$$z(\mathbf{x}) := p^\star(\mathbf{x}) + \lambda u^\star(\mathbf{x})$$

und argumentieren mittels Widerspruchsannahme. Dazu betrachten wir zunächst die untere Schranke. Ist das Behauptete falsch, so gibt es eine Menge $E \subset \Omega$ mit $|E| > 0$ derart, dass

$$z(\mathbf{x}) > 0 \quad \text{und} \quad u^\star(\mathbf{x}) > u_a(\mathbf{x}) \quad \text{fast überall auf } E.$$

Wir definieren

$$\widehat{u}(\mathbf{x}) := \begin{cases} u_a(\mathbf{x}), & \text{falls } \mathbf{x} \in E, \\ u^\star(\mathbf{x}), & \text{falls } \mathbf{x} \notin E, \end{cases}$$

und erhalten wegen $\widehat{u} \in U_{ad}$

$$\int_\Omega z(\widehat{u} - u^\star)\,\mathrm{d}\mathbf{x} = \int_E \underbrace{z(\mathbf{x})}_{>0} \big(\underbrace{u_a(\mathbf{x}) - u^\star(\mathbf{x})}_{<0} \big)\,\mathrm{d}\mathbf{x} < 0$$

im Widerspruch zu (4.9). Für die obere Schranke $u_b(\mathbf{x})$ argumentiert man analog, während der zweite Fall in (4.11) klar ist.

Es verbleibt zu zeigen, dass aus (4.11) die Variationsungleichung (4.9) folgt. Dazu seien $\mathbf{x} \in \Omega$ so gewählt, dass (4.11) gilt, und $v \in [u_a(\mathbf{x}), u_b(\mathbf{x})]$ beliebig. Ist $z(\mathbf{x}) > 0$, dann folgt aus (4.11), dass

$$z(\mathbf{x})\big(v - u^\star(\mathbf{x})\big) = \underbrace{z(\mathbf{x})}_{>0} \big(\underbrace{v - u_a(\mathbf{x})}_{\geq 0} \big) \geq 0$$

ist. Ist $z(\mathbf{x}) < 0$, dann folgt aus (4.11), dass

$$z(\mathbf{x})\big(v - u^\star(\mathbf{x})\big) = \underbrace{z(\mathbf{x})}_{<0} \big(\underbrace{v - u_b(\mathbf{x})}_{\leq 0} \big) \geq 0.$$

Weil sich $z(\mathbf{x})\big(v - u^\star(\mathbf{x})\big) = 0$ für $z(\mathbf{x}) = 0$ ergibt, erhalten wir demnach stets

$$z(\mathbf{x})\big(v - u^\star(\mathbf{x})\big) \geq 0.$$

Diese Ungleichung gilt fast überall, da (4.11) fast überall gilt. Da $v \in [u_a(\mathbf{x}), u_b(\mathbf{x})]$ beliebig war, erhalten wir

$$z(\mathbf{x})\big(u(\mathbf{x}) - u^\star(\mathbf{x})\big) \geq 0$$

und wegen der Monotonie des Integrals (4.9). □

Aus dem obigen Beweis folgt direkt, dass die Variationsungleichung (4.9) äquivalent zu der folgenden punktweisen Variationsungleichung in $\mathbb{R}$:

$$\langle p^\star(\mathbf{x}) + u^\star(\mathbf{x}), v - u^\star(\mathbf{x}) \rangle \geq 0 \quad \text{für alle } v \in [u_a(\mathbf{x}), u_b(\mathbf{x})].$$

Bemerkungen

- Im Fall $\lambda = 0$ vereinfacht sich (4.11) gemäß

$$u^\star(\mathbf{x}) = \begin{cases} u_a(\mathbf{x}), & \text{wo } p^\star(\mathbf{x}) > 0, \\ u_b(\mathbf{x}), & \text{wo } p^\star(\mathbf{x}) < 0. \end{cases}$$

  Wir erhalten aber keinerlei Aussage auf

$$\Omega_0 = \{\mathbf{x} \in \Omega : p(\mathbf{x}) = 0\}.$$

  Falls Ω_0 eine Nullmenge ist, so nimmt $u^\star(\mathbf{x})$ fast überall nur die Werte $\{u_a(\mathbf{x}), u_b(\mathbf{x})\}$ an. In diesem Fall spricht man von *Bang-Bang-Steuerung*.
- Im Fall $\lambda > 0$ gilt auf der Menge

$$\Omega_0 = \{\mathbf{x} \in \Omega : p^\star(\mathbf{x}) + \lambda u^\star(\mathbf{x}) = 0\}$$

  die Gleichung

$$u^\star(\mathbf{x}) = -\frac{1}{\lambda} p^\star(\mathbf{x}).$$

△

Im folgenden benötigen wir die Projektion von $\mathbb{R}$ auf ein kompaktes Intervall $[a, b]$, welche definiert ist durch

$$P_{[a,b]}(u) = \begin{cases} a, & \text{falls } u < a, \\ u, & \text{falls } a \le u \le b, \\ b, & \text{falls } u > b. \end{cases}$$

Satz 4.17 (Projektionsformel) Sei $\lambda > 0$. Dann ist $u^\star$ genau dann die optimale Steuerung von (1.5), falls der zugehörigen Zustand $y^\star$ und der adjungierte Zustand $p^\star$ der Projektionsformel

$$u^\star(\mathbf{x}) = P_{[u_a(\mathbf{x}), u_b(\mathbf{x})]}\left(-\frac{1}{\lambda} p^\star(\mathbf{x}) \right) \tag{4.12}$$

fast überall in Ω genügen.

Beweis. Die Projektionsformel kann äquivalent geschrieben werden als

$$u^\star(\mathbf{x}) = \begin{cases} u_a(\mathbf{x}), & \text{falls } -p^\star(\mathbf{x})/\lambda < u_a(\mathbf{x}), \\ -p^\star(\mathbf{x})/\lambda, & \text{falls } u_a(\mathbf{x}) \le -p^\star(\mathbf{x})/\lambda \le u_b(\mathbf{x}), \\ u_b(\mathbf{x}), & \text{falls } -p^\star(\mathbf{x})/\lambda > u_b(\mathbf{x}). \end{cases}$$

Dies entspricht aber genau (4.11), das heißt, nach Lemma 4.16 der notwendigen und hinreichenden Optimalitätsbedingung von (1.5). □

Bemerkung Ist $\lambda > 0$ und $U_{ad} = L^2(\Omega)$, dann folgt aus der Projektionsformel (4.12)

$$u^\star = -\frac{1}{\lambda} p^\star \text{ in } \Omega.$$

Man kann deshalb die Steuerung aus den notwendigen Bedingungen eliminieren und erhält das gekoppelte Randwertproblem

$$\begin{aligned} -\Delta y &= -\frac{1}{\lambda} p & -\Delta p &= y - y_d & &\text{in } \Omega, \\ y &= 0 & p &= 0 & &\text{auf } \Gamma. \end{aligned}$$

△

Anhand des Optimalitätssystems kann man Rückschlüsse auf die Regularität der optimalen Steuerung ziehen. Es gilt nämlich nicht nur $u^\star \in L^2(\Omega)$, sondern man hat sogar $u^\star \in H^1(\Omega)$.

Satz 4.18 (Regularität) Es sei $\lambda > 0$ und die Schranken u_a und u_b mögen zu $H^1(\Omega)$ gehören. Dann ist die optimale Steuerung $u^\star$ des Optimalsteuerungsproblem (1.5) eine Funktion aus $H^1(\Omega)$.

Beweis. Es seien

$$\Omega_a := \{\mathbf{x} \in \Omega : u^\star(\mathbf{x}) = u_a(\mathbf{x})\}, \quad \Omega_b := \{\mathbf{x} \in \Omega : u^\star(\mathbf{x}) = u_b(\mathbf{x})\}$$

und $\Omega_0 := \Omega \setminus (\Omega_a \cup \Omega_b)$. Da $p^\star \in H^1(\Omega)$ ist und daher keinen Sprung haben kann, folgt aus der Projektionsformel (4.12)

$$\|u^\star\|^2_{H^1(\Omega)} = \|u_a\|^2_{H^1(\Omega)} + \frac{1}{\lambda^2}\|p^\star\|^2_{H^1(\Omega_0)} + \|u_b\|^2_{H^1(\Omega)}.$$

Weiter ergibt sich aus (4.10), dass

$$\|p^\star\|_{H^1(\Omega)} \le c\|y^\star - y_d\|_{L^2(\Omega)} \le c\big\{\|y^\star\|_{H^1(\Omega)} + \|y_d\|_{L^2(\Omega)}\big\} \le c\big\{\|u^\star\|_{L^2(\Omega)} + \|y_d\|_{L^2(\Omega)}\big\},$$

das heißt, es ist tatsächlich $\|u^\star\|_{H^1(\Omega)} < \infty$. □

4.4 Formulierung als Karush-Kuhn-Tucker-System

Die Variationsungleichung (4.9) kann mit Hilfe von Lagrange-Multiplikatoren als weitere Gleichung formuliert werden.

Satz 4.19 Die Variationsungleichung (4.9) ist äquivalent zur Existenz von nichtnegativen Funktionen $\mu_a^\star, \mu_b^\star \in L^2(\Omega)$, so dass

$$p^\star + \lambda u^\star - \mu_a^\star + \mu_b^\star = 0 \text{ in } \Omega \tag{4.13}$$

und die komplementären Schlupfbedingungen

$$\mu_a^\star(u_a - u^\star) = \mu_b^\star(u^\star - u_b) = 0 \text{ in } \Omega \tag{4.14}$$

erfüllt sind.

Beweis. Wir zeigen zunächst, dass (4.13) und (4.14) aus der aus der Variationsungleichung (4.9) folgen. Dazu definieren wir

$$\left.\begin{aligned} \mu_a^\star(\mathbf{x}) &= \quad \max\{0, p^\star(\mathbf{x}) + \lambda u^\star(\mathbf{x})\} \\ \mu_b^\star(\mathbf{x}) &= -\min\{0, p^\star(\mathbf{x}) + \lambda u^\star(\mathbf{x})\} \end{aligned}\right\} \quad \mathbf{x} \in \Omega.$$

Es sind $\mu_a^\star, \mu_b^\star$ nichtnegative Funktionen aus $L^2(\Omega)$ und es gilt (4.13). Aus (4.11) ergeben sich

$$\begin{aligned} p^\star(\mathbf{x}) + \lambda u^\star(\mathbf{x}) > 0 &\implies u^\star(\mathbf{x}) = u_a(\mathbf{x}) \text{ und } \mu_b^\star(\mathbf{x}) = 0, \\ p^\star(\mathbf{x}) + \lambda u^\star(\mathbf{x}) < 0 &\implies u^\star(\mathbf{x}) = u_b(\mathbf{x}) \text{ und } \mu_a^\star(\mathbf{x}) = 0, \end{aligned}$$

und

$$u_a(\mathbf{x}) < u^\star(\mathbf{x}) < u_b(\mathbf{x}) \implies p^\star(\mathbf{x}) + \lambda u^\star(\mathbf{x}) = 0 \text{ und } \mu_a^\star(\mathbf{x}) = \mu_b^\star(\mathbf{x}) = 0$$

fast überall in Ω, dies bedeutet (4.14).
Es verbleibt zu zeigen, dass aus (4.13) und (4.14) die Variationsungleichung (4.9) folgt. Dazu sei $u \in U_{ad}$ beliebig, aber fest gewählt. Ferner setzen wir

$$\Omega_a := \{\mathbf{x} \in \Omega : u^\star(\mathbf{x}) = u_a(\mathbf{x})\}, \quad \Omega_b := \{\mathbf{x} \in \Omega : u^\star(\mathbf{x}) = u_b(\mathbf{x})\}$$

und $\Omega_0 := \Omega \setminus (\Omega_a \cup \Omega_b)$.
Auf Ω_0 folgt $u_a < u^\star < u_b$ und daher nach (4.13) und (4.14)

$$\int_{\Omega_0} (\underbrace{p^\star + \lambda u^\star}_{=0})(u - u^\star)\,\mathrm{d}\mathbf{x} = 0.$$

Auf Ω_a schließen wir aus (4.13) und (4.14), dass

$$\int_{\Omega_a} (\underbrace{p^\star + \lambda u^\star}_{\geq 0})(\underbrace{u - u_a}_{\geq 0})\,\mathrm{d}\mathbf{x} \geq 0,$$

und analog auf Ω_b, dass

$$\int_{\Omega_b} (\underbrace{p^\star + \lambda u^\star}_{\leq 0})(\underbrace{u - u_b}_{\leq 0})\,\mathrm{d}\mathbf{x} \geq 0.$$

Damit ist der Beweis vollständig erbracht. □

Mit Satz 4.19 können wir das Optimalitätssystem (4.10) auch als *Karush-Kuhn-Tucker-System* schreiben:

$$\left.\begin{aligned} -\Delta y^\star = u^\star \qquad -\Delta p^\star = y^\star - y_d \qquad & \text{in } \Omega, \\ y^\star = 0 \qquad p^\star = 0 \qquad & \text{auf } \Gamma, \\ p^\star + \lambda u^\star - \mu_a^\star + \mu_b^\star = 0 \quad & \text{auf } \Omega, \\ \mu_a^\star \geq 0, \qquad u_a - u^\star \leq 0, \qquad \mu_a^\star(u_a - u^\star) = 0 \qquad & \text{auf } \Omega, \\ \mu_b^\star \geq 0, \qquad u^\star - u_b \leq 0, \qquad \mu_a^\star(u^\star - u_b) = 0 \qquad & \text{auf } \Omega. \end{aligned}\right\} \tag{4.15}$$

Definition 4.20 Die in Satz 4.19 eingeführten Funktionen $\mu_a^\star, \mu_b^\star \in L^2(\Omega)$ heißen **Lagrange-Multiplikatoren** zu den Ungleichungsbedingungen $u_a \leq u$ beziehungsweise $u \leq u_b$.

4.5 Formales Lagrange-Prinzip

Bei der Herleitung des Optimalitätssystems (4.10) haben wir $y = S(u)$ benutzt und die reduzierte Aufgabe nur in u formuliert. Anschließend haben wir den Term $S^\top(Su - y_d)$ in der Variationsungleichung (4.7) durch ein geeignete Definition der adjungierten Variable p ersetzt. Die Form der zugehörigen adjungierten Gleichung haben wir wenig intuitiv in Lemma 4.13 hergeleitet. Das ist bei komplizierteren Aufgaben allerdings oftmals sehr schwer. Abhilfe schafft hier das formale Lagrange-Prinzip.

Wir betrachten die Optimalsteuerungsaufgabe

$$\text{minimiere } J(y,u) = \frac{1}{2}\int_\Omega |y - y_d|^2 \,\mathrm{d}\mathbf{x} + \frac{\lambda}{2}\int_\Omega |u|^2 \,\mathrm{d}\mathbf{x}$$
$$\text{unter den Nebenbedingungen } -\operatorname{div}\big(\mathbf{A}\nabla y\big) + \langle \mathbf{b}, \nabla y\rangle + cy = u \text{ in } \Omega,\ y = 0 \text{ auf } \Gamma$$
$$\text{und } u_a \leq u \leq u_b \text{ in } \Omega,$$

wobei $\mathbf{A} : \Omega \to \mathbb{R}^{d\times d}$ symmetrisch und uniform elliptisch, $\mathbf{b} : \Omega \to \mathbb{R}^d$ und $c : \Omega \to \mathbb{R}$ seien. Wie wir wissen, ist die zugehörige Bilinearform

$$a(y,z) = \int_\Omega \big\{\langle \mathbf{A}\nabla y, \nabla z\rangle + \langle \mathbf{b}, \nabla y\rangle z + cyz\big\}\,\mathrm{d}\mathbf{x}$$

elliptisch falls $c(\mathbf{x}) \geq \underline{c}$ hinreichend groß und $\|\mathbf{b}(\mathbf{x})\|_2 \leq \overline{b}$ hinreichend klein ist. Somit ist in diesem Fall die Zustandsgleichung eindeutig lösbar.

Um die Zustandsgleichung mit Hilfe des Lagrange-Multiplikators $p \in H_0^1(\Omega)$ an das Funktional $J(y,u)$ anzukoppeln, definieren wir das Lagrange-Funktional $L : H_0^1(\Omega) \times L^2(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ gemäß

$$L(y,u,p) = J(y,u) - \int_\Omega \big\{\langle \mathbf{A}\nabla y, \nabla p\rangle + \langle \mathbf{b}, \nabla y\rangle p + cyp\big\}\,\mathrm{d}\mathbf{x} + \int_\Omega up\,\mathrm{d}\mathbf{x}.$$

Weil im Optimum gilt

$$\delta_y L(y^\star, u^\star, p^\star)[z] = \int_\Omega (y^\star - y_d)z\,\mathrm{d}\mathbf{x} - \int_\Omega \big\{\langle \mathbf{A}\nabla z, \nabla p^\star\rangle + \langle \mathbf{b}, \nabla z\rangle p^\star + czp^\star\big\}\,\mathrm{d}\mathbf{x} \overset{!}{=} 0$$

für alle $z \in H_0^1(\Omega)$, genügt der adjungierte Zustand $p \in H_0^1(\Omega)$ der Gleichung

$$-\operatorname{div}\big(\mathbf{A}\nabla p\big) - \langle \mathbf{b}, \nabla p\rangle + (c - \operatorname{div}\mathbf{b})p = y - y_d \text{ in } \Omega, \quad p = 0 \text{ auf } \Gamma. \tag{4.16}$$

Dabei haben wir benutzt, dass

$$-\int_\Omega \langle \mathbf{b}, \nabla p\rangle z\,\mathrm{d}\mathbf{x} = \int_\Omega \big\{\langle \mathbf{b}, \nabla z\rangle p + \operatorname{div}\mathbf{b}zp\big\}\,\mathrm{d}\mathbf{x}.$$

Speziell sehen wir, dass die Gleichung (4.16) für den adjungierten Zustand auf diese Weise nur hingeschrieben werden kann, falls div $\mathbf{b}$ existiert. Ferner ist (4.16) nicht offensichtlich elliptisch. Zum Nachweis der Elliptizität kann man allerdungs die partielle Integration wieder rückgängig machen und landet bei der ursprünglichen, elliptischen Bilinearform für die Zustandsgleichung.

Wie in Lemma 4.13 muss nun noch nachweisen, dass es sich bei (4.16) tatsächlich um die adjungierte Gleichung handelt, deren Lösungsoperator also mit dem adjungierten Operator $S^\top$ des Steuerungs-Zustand-Operators übereinstimmt. Im vorliegenden Fall funktioniert das genauso wie beim Poisson-Problem, indem man die Zustandsgleichung und die adjungierte Gleichung mit der jeweils anderen Lösung testet und die entstehenden Gleichungen voneinander subtrahiert.

Auch die zugehörige Variationsungleichung im Optimalitätssystems ergibt sich aus der Lagrange-Funktion. Wie in (2.10) folgt aus $\delta_u L(y^\star, u^\star, p^\star)[u - u^\star] \geq 0$ nämlich

$$\delta_u L(y^\star, u^\star, p^\star)[u - u^\star] = \int_\Omega (p^\star + \lambda u^\star)(u - u^\star)\,\mathrm{d}\mathbf{x} \geq 0$$

für alle $u \in U_{ad}$.

5. Diskretisierung

5.1 Steuerungs-Zustands-Operator

Wir wollen nun die Diskretisierung des Optimalsteuerungsproblem (1.5) mit Hilfe von Finiten Elementen durchführen. Als Grundvorraussetzung in diesem Kapitel wollen wir annehmen, dass das Gebiet $\Omega \subset \mathbb{R}^d$ konvex und polygonal ($d = 2$) beziehungsweise polyhedral ($d = 3$) berandet ist. Ferner seien die Box-Beschränkungen u_a und u_b der Einfachheit halber konstante Funktionen und y_d eine Funktion aus $H^2(\Omega)$. Zu einer quasi-uniforme Familie $\{\mathcal{T}_h\}$ von Zerlegungen von Ω werden wir stückweise konstante Finite-Elemente-Räume

$$U_h = \left\{u_h \in L^2(\Omega) : u_h\big|_T \text{ ist konstant für alle } T \in \mathcal{T}_h\right\} \subset L^2(\Omega)$$

und stetige, stückweise lineare Finite-Elemente-Räume

$$V_h = \left\{v_h \in C(\overline{\Omega}) : v_h\big|_\Gamma = 0 \text{ und } v_h\big|_T \text{ ist linear für alle } T \in \mathcal{T}_h\right\} \subset H_0^1(\Omega)$$

betrachten.

Den Steuerungs-Zustands-Operator $S : L^2(\Omega) \to L^2(\Omega)$, der einem gegebenen Zustand $u \in L^2(\Omega)$ die zugehörige Steuerung $y \in L^2(\Omega)$ zuordnet, können wir mit Hilfe der Bilinearform

$$a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}, \quad a(y, z) = \int_\Omega \langle \nabla y, \nabla z \rangle \, \mathrm{d}\mathbf{x}$$

variationell schreiben gemäß

$$\text{suche } y \in H_0^1(\Omega), \text{ so dass } a(y, z) = (u, z)_{L^2(\Omega)} \text{ für alle } z \in H_0^1(\Omega).$$

Damit ist des Optimalsteuerungsproblem (1.5) äquivalent zu

$$\left.\begin{array}{c} \text{minimiere } J(y, u) = \dfrac{1}{2}\displaystyle\int_\Omega |y - y_d|^2 \, \mathrm{d}\mathbf{x} + \dfrac{\lambda}{2}\int_\Omega |u|^2 \, \mathrm{d}\mathbf{x} \\ \text{unter den Nebenbedingungen } a(y, z) = (u, z)_{L^2(\Omega)} \text{ für alle } z \in H_0^1(\Omega) \\ \text{und } u_a \le u \le u_b \text{ in } \Omega. \end{array}\right\} \tag{5.1}$$

Analog können wir den *diskreten Steuerungs-Zustands-Operator* $S_h : L^2(\Omega) \to L^2(\Omega)$ definieren, der jeder Steuerung $u \in L^2(\Omega)$ einen diskreten Zustand $y_h \in V_h$ zuordnet vermittels der Variationsformulierung

$$\text{suche } y_h \in V_h, \text{ so dass } a(y_h, z_h) = (u, z_h)_{L^2(\Omega)} \text{ für alle } z_h \in V_h.$$

Der Operator S_h ist offensichtlich linear und beschränkt und aufgrund der Annahmen über Gebiet und Triangulierung haben wir die Fehlerabschätzung

$$\|(S - S_h)u\|_{L^2(\Omega)} = \|y - y_h\|_{L^2(\Omega)} \leq ch^2 \|u\|_{L^2(\Omega)}$$

für alle $u \in L^2(\Omega)$. Sie impliziert sofort:

Lemma 5.1 Der diskrete Steuerungs-Zustands-Operator $S_h : L^2(\Omega) \to L^2(\Omega)$ erfüllt

$$\|S - S_h\| \leq ch^2$$

mit einer von h unabhängigen Konstante $c > 0$.

5.2 Variationelle Diskretisierung

Im folgenden werden wir zunächst anstelle von (5.1) semidiskrete Problem

$$\left.\begin{array}{c} \text{minimiere } J_h(y_h, u) = \dfrac{1}{2}\displaystyle\int_\Omega |y_h - I_h y_d|^2 \,\mathrm{d}\mathbf{x} + \dfrac{\lambda}{2}\int_\Omega |u|^2 \,\mathrm{d}\mathbf{x} \\ \text{unter den Nebenbedingungen } a(y_h, z_h) = (u, z_h)_{L^2(\Omega)} \text{ für alle } z_h \in V_h \\ \text{und } u_a \leq u \leq u_b \text{ in } \Omega \end{array}\right\} \tag{5.2}$$

für das Funktional $J_h(y_h, u) : V_h \times L^2(\Omega)$ studieren. Dabei haben wir y_d durch die stückweise lineare Interpolation $I_h y_d \in V_h$ und die Variationsformulierung in $H_0^1(\Omega)$ durch die diskrete Variationsformulierung in V_h ersetzt.
Wie im kontinuierlichen Fall, betrachten wir wieder das zu (5.2) äquivalente reduzierte Optimierungsproblem

$$\left.\begin{array}{c} \text{minimiere } f(u) = \dfrac{1}{2}\displaystyle\int_\Omega |S_h u - I_h y_d|^2 \,\mathrm{d}\mathbf{x} + \dfrac{\lambda}{2}\int_\Omega |u|^2 \,\mathrm{d}\mathbf{x} \\ \text{unter den Nebenbedingung } u_a \leq u \leq u_b \text{ in } \Omega. \end{array}\right\} \tag{5.3}$$

Satz 5.2 Das reduzierte Problem (5.3) besitzt eine eindeutige Lösung $\overline{u}_h^\star \in L^2(\Omega)$. Diese ist zusammen mit $\overline{y}_h^\star \in V_h$ und $\overline{p}_h^\star \in V_h$ eindeutig durch die Lösung des folgenden Optimalitätssystem gegeben:

$$\left.\begin{array}{l} a(\overline{y}_h^\star, z_h) = (\overline{u}_h^\star, z_h)_{L^2(\Omega)} \text{ für alle } z_h \in V_h, \\ a(z_h, \overline{p}_h^\star) = (\overline{y}_h^\star - I_h y_d, z_h)_{L^2(\Omega)} \text{ für alle } z_h \in V_h, \\ (\overline{p}_h^\star + \lambda \overline{u}_h^\star, u - \overline{u}_h^\star)_{L^2(\Omega)} \geq 0 \text{ für alle } u \in U_{ad}. \end{array}\right\} \tag{5.4}$$

Beweis. Die Existenz und Eindeutigkeit einer optimalen Lösung $\overline{u}_h^\star \in L^2(\Omega)$ zu (5.3) beweist man wie im Satz 3.17 unter Verwendung von Lemma 3.16 und Satz 3.13. Da S_h

wie S ein linearer und stetiger Operator ist, ist das Kostenfunktional in (5.3) differenzierbar und Satz 4.11 anwendbar. Wir erhalten daher als notwendige und hinreichende Optimalitätsbedingung:

$$\left(S_h^\top (S_h \overline{u}_h^\star - I_h y_d) + \lambda \overline{u}_h^\star, u - \overline{u}_h^\star\right)_{L^2(\Omega)} \geq 0 \quad \text{für alle } u \in U_{ad}.$$

Es verbleibt, die Form von $S_h^\top : L^2(\Omega) \to L^2(\Omega)$ nachzuweisen. Dazu beachten wir, dass $y_h = S_h f$ für $f \in L^2(\Omega)$ der Gleichung

$$(f, z_h)_{L^2(\Omega)} = a(y_h, z_h)_{L^2(\Omega)} \quad \text{für alle } z_h \in V_h$$

entspricht. Andererseits bedeutet $p_h = S_h^\top g$ bei gegebenem $g \in L^2(\Omega)$, dass

$$(g, z_h)_{L^2(\Omega)} = a(z_h, p_h)_{L^2(\Omega)} \quad \text{für alle } z_h \in V_h.$$

Setzen wir in der ersten Gleichung p_h und in der zweiten Gleichung y_h als Testfunktion ein, dann folgt

$$(f, S_h^\top g)_{L^2(\Omega)} = (f, p_h)_{L^2(\Omega)} = (y_h, g)_{L^2(\Omega)} = (S_h f, g)_{L^2(\Omega)}.$$

Da f und g beliebig waren, zeigt dies, dass $S_h^\top$ tatsächlich die Adjungierte von S_h ist. □

Lemma 5.3 Mit einer von h unabhängigen Konstante $c > 0$ gilt

$$\left\|S^\top - S_h^\top\right\| \leq ch^2.$$

Beweis. Aus

$$\begin{aligned}
\left\|(S^\top - S_h^\top)p\right\|_{L^2(\Omega)} &= \sup_{u \in L^2(\Omega)\setminus\{0\}} \frac{\left((S^\top - S_h^\top)p, u\right)_{L^2(\Omega)}}{\|u\|_{L^2(\Omega)}} \\
&= \sup_{u \in L^2(\Omega)\setminus\{0\}} \frac{\left(p, (S - S_h)u\right)_{L^2(\Omega)}}{\|u\|_{L^2(\Omega)}} \\
&\leq \|S - S_h\| \|p\|_{L^2(\Omega)}
\end{aligned}$$

folgt

$$\left\|S^\top - S_h^\top\right\|_{L^2(\Omega)} \leq \|S - S_h\|.$$

Deshalb ergibt sich das Behauptete nun mit Hilfe von Lemma 5.1. □

Mit Hilfe der Lemmata 5.1 und 5.3 sowie den Variationsungleichungen für (5.1) und (5.2) können wir den Approximationsfehler im semidiskreten Fall abschätzen:

Satz 5.4 Seien $0 < \lambda \leq 1$ und $u^\star$ und $\overline{u}_h^\star$ die Lösungen von (5.1) beziehungsweise (5.2) mit den zugehörigen Zuständen $y^\star$ und $\overline{y}_h^\star$. Dann existiert eine von h unabhängige Konstante $c > 0$ derart, dass

$$\|u^\star - \overline{u}_h^\star\|_{L^2(\Omega)} + \|y^\star - \overline{y}_h^\star\|_{L^2(\Omega)} \leq \frac{c}{\lambda} h^2.$$

Beweis. Wir betrachten die Variationsungleichungen für (5.1) und (5.2):

$$\left.\begin{aligned}\big(S^\top(Su^\star - y_d) + \lambda u^\star, u - u^\star\big)_{L^2(\Omega)} &\geq 0\\ \big(S_h^\top(S_h\overline{u}_h^\star - I_h y_d) + \lambda \overline{u}_h^\star, u - \overline{u}_h^\star\big)_{L^2(\Omega)} &\geq 0\end{aligned}\right\} \quad \text{für alle } u \in U_{ad}.$$

Setzen wir in die erste Variationsungleichung $\overline{u}_h^\star$ ein und in die zweite $u^\star$ und summieren wir anschließend die beiden Ungleichungen, dann folgt

$$\begin{aligned}
0 &\leq \big(S^\top(Su^\star - y_d) + \lambda u^\star, \overline{u}_h^\star - u^\star\big)_{L^2(\Omega)} + \big(S_h^\top(S_h\overline{u}_h^\star - I_h y_d) + \lambda\overline{u}_h^\star, u^\star - \overline{u}_h^\star\big)_{L^2(\Omega)}\\
&= \big(Su^\star - y_d, S(\overline{u}_h^\star - u^\star)\big)_{L^2(\Omega)} + \big(S_h\overline{u}_h^\star - I_h y_d, S_h(u^\star - \overline{u}_h^\star)\big)_{L^2(\Omega)}\\
&\quad + \big(Su^\star - y_d, S_h\big[\underbrace{u^\star - \overline{u}_h^\star - (u^\star - \overline{u}_h^\star)}_{=0}\big]\big)_{L^2(\Omega)} + \lambda\big(\overline{u}_h^\star - u^\star, u^\star - \overline{u}_h^\star\big)_{L^2(\Omega)}\\
&= \big(Su^\star - y_d, (S - S_h)(\overline{u}_h^\star - u^\star)\big)_{L^2(\Omega)} + \big(S_h\overline{u}_h^\star - Su^\star + y_d - I_h y_d, S_h(u^\star - \overline{u}_h^\star)\big)_{L^2(\Omega)}\\
&\quad + \big(S_h\overline{u}_h^\star - Su^\star + y_d - I_h y_d, S(\underbrace{u^\star - u^\star}_{=0})\big)_{L^2(\Omega)} - \lambda\|u^\star - \overline{u}_h^\star\|_{L^2(\Omega)}^2\\
&= \big((S - S_h)^\top(Su^\star - y_d), \overline{u}_h^\star - u^\star\big)_{L^2(\Omega)}\\
&\quad + \big(S_h\overline{u}_h^\star - Su^\star, (S_h - S)u^\star\big)_{L^2(\Omega)} + \big(S_h\overline{u}_h^\star - Su^\star, Su^\star - S_h\overline{u}_h^\star\big)\big)_{L^2(\Omega)}\\
&\quad + \big(y_d - I_h y_d, (S_h - S)u^\star\big)_{L^2(\Omega)} + \big(y_d - I_h y_d, Su^\star - S_h\overline{u}_h^\star\big)_{L^2(\Omega)}\\
&\quad - \lambda\|u^\star - \overline{u}_h^\star\|_{L^2(\Omega)}^2.
\end{aligned}$$

Demzufolge ergibt sich

$$\begin{aligned}
&\lambda\|u^\star - \overline{u}_h^\star\|_{L^2(\Omega)}^2 + \big\|Su^\star - S_h\overline{u}_h^\star\big\|_{L^2(\Omega)}^2\\
&\quad\leq \big(y_d - I_h y_d, (S_h - S)u^\star\big)_{L^2(\Omega)} + \big(S_h\overline{u}_h^\star - Su^\star, (S_h - S)u^\star\big)_{L^2(\Omega)}\\
&\qquad + \big(y_d - I_h y_d, Su^\star - S_h\overline{u}_h^\star\big)_{L^2(\Omega)} + \big((S - S_h)^\top(Su^\star - y_d), \overline{u}_h^\star - u^\star\big)_{L^2(\Omega)}.
\end{aligned}$$

Die letzten drei Terme schätzen wir jeweils mit der Cauchy-Schwarzschen Ungleichung und der Youngschen Ungleichung

$$ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$$

ab und erhalten

$$\begin{aligned}
&\lambda\|u^\star - \overline{u}_h^\star\|_{L^2(\Omega)}^2 + \big\|Su^\star - S_h\overline{u}_h^\star\big\|_{L^2(\Omega)}^2\\
&\quad\leq \big(y_d - I_h y_d, (S_h - S)u^\star\big)_{L^2(\Omega)} + \frac{1}{2}\big\|Su^\star - S_h\overline{u}_h^\star\big\|_{L^2(\Omega)}^2 + \frac{1}{2}\big\|(S - S_h)u^\star\big\|_{L^2(\Omega)}^2\\
&\qquad + \frac{1}{2}\|y_d - I_h y_d\|_{L^2(\Omega)}^2 + \frac{1}{2}\big\|Su^\star - S_h\overline{u}_h^\star\big\|_{L^2(\Omega)}^2\\
&\qquad + \frac{1}{2\lambda}\big\|(S - S_h)^\top(Su^\star - y_d)\big\|_{L^2(\Omega)}^2 + \frac{\lambda}{2}\|\overline{u}_h^\star - u^\star\|_{L^2(\Omega)}^2.
\end{aligned}$$

Wir fassen die Terme zusammen und schätzen weiter ab:

$$\begin{aligned}
\frac{\lambda}{2}\|u^\star - \overline{u}_h^\star\|_{L^2(\Omega)}^2 &\leq \big(y_d - I_h y_d, (S_h - S)u^\star\big)_{L^2(\Omega)} + \frac{1}{2}\|(S - S_h)u^\star\|_{L^2(\Omega)}^2\\
&\qquad + \frac{1}{2}\|y_d - I_h y_d\|_{L^2(\Omega)}^2 + \frac{1}{2\lambda}\big\|(S - S_h)^\top(Su^\star - y_d)\big\|_{L^2(\Omega)}^2\\
&\leq \|y_d - I_h y_d\|_{L^2(\Omega)}\|S - S_h\|\|u^\star\|_{L^2(\Omega)} + \frac{1}{2}\|S - S_h\|^2\|u^\star\|_{L^2(\Omega)}^2\\
&\qquad + \frac{1}{2}\|y_d - I_h y_d\|_{L^2(\Omega)}^2 + \frac{1}{2\lambda}\big\|S^\top - S_h^\top\big\|^2\|Su^\star - y_d\|_{L^2(\Omega)}^2.
\end{aligned}$$

Die Interpolationsfehlerabschätzung für lineare Finite Elemente liefert

$$\|y_d - I_h y_d\|_{L^2(\Omega)} \leq ch^2 \|y_d\|_{H^2(\Omega)},$$

da $y_d \in H^2(\Omega)$ nach Voraussetzung. Ferner können wir $\|S - S_h\|$ und $\left\|S^\top - S_h^\top\right\|$ vermittels Lemmata 5.1 und 5.3 abschätzen. Aufgrund der Existenz und Eindeutigkeit der Lösung $u \in L^2(\Omega)$ für $\lambda > 0$ gemäß den Sätzen 3.17 und 3.18 folgt $\|u^\star\|_{L^2(\Omega)} < \infty$ und folglich $\|Su^\star - y_d\|_{L^2(\Omega)} < \infty$, weil der Operator $S : L^2(\Omega) \to L^2(\Omega)$ beschränkt ist. Zusammengefasst erhalten wir damit

$$\frac{\lambda}{2}\|u^\star - \overline{u}_h^\star\|_{L^2(\Omega)}^2 \leq c\left(1 + \frac{1}{\lambda}\right)h^4,$$

was für $\lambda \leq 1$ äquivalent ist zu

$$\|u^\star - \overline{u}_h^\star\|_{L^2(\Omega)} \leq \frac{c}{\lambda}h^2.$$

Die noch fehlende Abschätzung für $y^\star - \overline{y}_h^\star$ ergibt sich schließlich aus

$$\begin{aligned}
\|y^\star - \overline{y}_h^\star\|_{L^2(\Omega)} &= \|Su^\star - S_h\overline{u}_h^\star\|_{L^2(\Omega)} \\
&\leq \|(S - S_h)u^\star\|_{L^2(\Omega)} + \|S_h(u^\star - \overline{u}_h^\star)\|_{L^2(\Omega)} \\
&\leq \|S - S_h\|\|u^\star\|_{L^2(\Omega)} + \|S_h\|\|u^\star - \overline{u}_h^\star\|_{L^2(\Omega)} \\
&\leq \frac{c}{\lambda}h^2.
\end{aligned}$$

□

Der Beweis zeigt, dass man im Fall von $\lambda > 1$ die Abschätzung

$$\|u^\star - \overline{u}_h^\star\|_{L^2(\Omega)} + \|y^\star - \overline{y}_h^\star\|_{L^2(\Omega)} \leq ch^2.$$

erhält. In der Praxis ist allerdings der Fall $\lambda \ll 1$ wichtiger, weshalb in Satz 5.4 die Abhängigkeit des Approximationsfehlers von λ explizit berücksichtigt wurde. Speziell sieht man, dass für kleines λ der Approximationsfehler durchaus groß sein kann.

5.3 Volldiskretisierung

Wir wollen nun den Fall betrachten, dass auch die Steuerung aus einem endlichdimensionalen Raum U_h gewählt wird. Anstelle von (5.2) betrachten wir also

$$\left.\begin{array}{c}
\text{minimiere } J_h(y_h, u_h) = \dfrac{1}{2}\displaystyle\int_\Omega |y_h - I_h y_d|^2 \,\mathrm{d}\mathbf{x} + \dfrac{\lambda}{2}\int_\Omega |u_h|^2 \,\mathrm{d}\mathbf{x} \\
\text{unter den Nebenbedingungen } a(y_h, z_h) = (u_h, z_h)_{L^2(\Omega)} \text{ für alle } z_h \in V_h \\
\text{und } u_a \leq u_h \leq u_b \text{ in } \Omega.
\end{array}\right\} \tag{5.5}$$

Da $u_h \in U_h$ ist, lässt sich die Box-Beschränkung sehr leicht umsetzen. Denn ist $\{\chi_T : T \in \mathcal{T}_h\}$ die Menge der stückweise konstanten Basisfunktionen, so erfüllt

$$u_h = \sum_{T \in \mathcal{T}_h} u_T \chi_T \in U_h$$

genau dann die Box-Beschränkung

$$u_a \leq u_h \leq u_b \text{ in } \Omega,$$

wenn $u_a \leq u_T \leq u_b$ gilt für alle $T \in \mathcal{T}_h$.
Um auch im Fall der Volldiskretisierung den Approximationsfehler zu untersuchen, beginnen wir wieder mit der Definition des reduzierten, volldiskreten Problems:

$$\left.\begin{array}{c} \text{minimiere } f(u_h) = \dfrac{1}{2}\displaystyle\int_\Omega |S_h u_h - I_h y_d|^2 \,\mathrm{d}\mathbf{x} + \dfrac{\lambda}{2}\int_\Omega |u_h|^2 \,\mathrm{d}\mathbf{x} \\ \text{unter den Nebenbedingung } u_a \leq u_h \leq u_b \text{ in } \Omega. \end{array}\right\} \tag{5.6}$$

Setzen wir

$$U_{h,ad} := \{u_h \in U_h : u_a \leq u_h \leq u_b \text{ in } \Omega\},$$

so haben wir das folgende Resultat:

Satz 5.5 Das reduzierte Optimalsteuerungsproblem (5.6) besitzt im Fall $\lambda > 0$ eine eindeutige Lösung $u_h^\star \in U_h$, welche der Variationsungleichung

$$\big(S_h^\top (S_h u_h^\star - I_h y_d) + \lambda u_h^\star, u_h - u_h^\star\big)_{L^2(\Omega)} \quad \text{für alle } u_h \in U_{h,ad} \tag{5.7}$$

genügt.

Beweis. Die zulässige Menge $U_{h,ad} \subset U_{ad}$ ist beschränkt, weil U_{ad} beschränkt ist. Ferner ist U_h ein endlichdimensionaler Unterraum von $L^2(\Omega)$. Damit ist $U_{h,ad} = U_h \cap U_{ad}$ als Durchschnitt zweier abgeschlossener und konvexer Mengen selbst abgeschlossen und konvex. Folglich sichert der allgemeine Satz 3.13 die Existenz einer Lösung zu (5.6), die sogar eindeutig ist aufgrund der strikten Konvexität des Funktionals f. Die Variationsungleichung (5.7) folgt schließlich aus Satz 4.11, da $U_{h,ad}$ konvex ist. □

Es bezeichne $\Pi_h : L^2(\Omega) \to U_h$ die $L^2(\Omega)$-Orthoprojektion auf U_h. Für $u \in L^2(\Omega)$ ist sie offensichtlich elementweise gegeben durch

$$(\Pi_h u)|_T = \frac{1}{|T|}\int_T u \,\mathrm{d}\mathbf{x} \quad \text{für alle } T \in \mathcal{T}_h.$$

Wir haben das folgende Resultat für die Approximationsgüte der $L^2(\Omega)$-Orthoprojektion. Dabei bezeichne $\widetilde{H}^{-1}(\Omega) := \big(H^1(\Omega)\big)'$ den Dualraum des $H^1(\Omega)$ bezüglich des Skalarprodukts in $L^2(\Omega)$, ausgestattet mit der Norm

$$\|u\|_{\widetilde{H}^{-1}(\Omega)} := \sup_{v \in H^1(\Omega)\setminus\{0\}} \frac{(u,v)_{L^2(\Omega)}}{\|v\|_{H^1(\Omega)}}.$$

Lemma 5.6 Ist $u \in H^1(\Omega)$, dann gelten die Fehlerabschätzungen

$$\|u - \Pi_h u\|_{L^2(\Omega)} \leq ch\|u\|_{H^1(\Omega)},$$
$$\|u - \Pi_h u\|_{\widetilde{H}^{-1}(\Omega)} \leq ch^2\|u\|_{H^1(\Omega)}.$$

Beweis. Die erste Aussage folgt direkt aus dem Bramble-Hilbert-Lemma. Die zweite Aussage

$$\begin{aligned}\|u - \Pi_h u\|_{\tilde{H}^{-1}(\Omega)} &= \sup_{v \in H^1(\Omega)\setminus\{0\}} \frac{(u - \Pi_h u, v)_{L^2(\Omega)}}{\|v\|_{H^1(\Omega)}} \\ &= \sup_{v \in H^1(\Omega)\setminus\{0\}} \frac{(u - \Pi_h u, v - \Pi_h v)_{L^2(\Omega)}}{\|v\|_{H^1(\Omega)}} \\ &\leq \sup_{v \in H^1(\Omega)\setminus\{0\}} \|u - \Pi_h u\|_{L^2(\Omega)} \frac{\|v - \Pi_h v\|_{L^2(\Omega)}}{\|v\|_{H^1(\Omega)}} \\ &\leq ch^2 \|u\|_{H^1(\Omega)}.\end{aligned}$$

□

Satz 5.7 Seien $0 < \lambda \leq 1$ und $u^\star$ und $u_h^\star$ die Lösungen von (5.1) beziehungsweise (5.5) mit den zugehörigen Zuständen $y^\star$ und $y_h^\star$. Dann existiert eine von h unabhängige Konstante $c > 0$ derart, dass

$$\|u^\star - u_h^\star\|_{L^2(\Omega)} + \|y^\star - y_h^\star\|_{L^2(\Omega)} \leq \frac{c}{\lambda} h.$$

Beweis. Der Beweis verläuft im wesentlichen wie der Beweis von Satz 5.4. Die Variationsungleichungen für (5.1) und (5.5) lauten

$$\left(S^\top(Su^\star - y_d) + \lambda u^\star, u - u^\star\right)_{L^2(\Omega)} \geq 0 \quad \text{für alle } u \in U_{ad}, \tag{5.8}$$

$$\left(S_h^\top(S_h u_h^\star - I_h y_d) + \lambda u_h^\star, u_h - u_h^\star\right)_{L^2(\Omega)} \geq 0 \quad \text{für alle } u_h \in U_{h,ad}. \tag{5.9}$$

Im Gegensatz zur Semidiskretisierung ist die kontinuierliche Lösung $u^\star$ jetzt nicht mehr zulässig für (5.9), denn es gilt im allgemeinen $u^\star \notin U_h$. Daher kann $u^\star$ nicht als Testfunktion in (5.9) eingesetzt werden. Wir wählen stattdessen $\Pi_h u^\star$ als Testfunktion. Diese Testfunktion ist zulässig, weil

$$(\Pi_h u^\star)|_T = \frac{1}{|T|} \int_T \underbrace{u^\star}_{\geq u_a} \mathrm{d}\mathbf{x} \geq \frac{1}{|T|} \int_T u_a \,\mathrm{d}\mathbf{x} = u_a$$

und analog auch $(\Pi_h u^\star)|_T \leq u_b$ ist für jedes $T \in \mathcal{T}_h$, dies bedeutet, es ist tatsächlich $\Pi_h u^\star \in U_{h,ad}$.
Wir setzen noch $u_h^\star$ als Testfunktion in (5.8) ein und addieren dann die beiden Variationsungleichungen:

$$\begin{aligned}0 &\leq \left(S^\top(Su^\star - y_d) + \lambda u^\star, u_h^\star - u^\star\right)_{L^2(\Omega)} + \left(S_h^\top(S_h u_h^\star - I_h y_d) + \lambda u_h^\star, \Pi_h u^\star - u_h^\star\right)_{L^2(\Omega)} \\ &= \left(S^\top(Su^\star - y_d) + \lambda u^\star, u_h^\star - u^\star\right)_{L^2(\Omega)} + \left(S_h^\top(S_h u_h^\star - I_h y_d) + \lambda u_h^\star, u^\star - u_h^\star\right)_{L^2(\Omega)} \\ &\quad + \left(S_h^\top(S_h u_h^\star - I_h y_d) + \lambda u_h^\star, \Pi_h u^\star - u^\star\right)_{L^2(\Omega)}.\end{aligned}$$

Die ersten beiden Terme auf der rechten Seite haben dieselbe Struktur wie bei der Abschätzung für die Semidiskretisierung im Beweis von Satz 5.4 und können deshalb genauso abgeschätzt werden. Dies führt auf

$$\frac{\lambda}{2}\|u^\star - u_h^\star\|_{L^2(\Omega)}^2 \leq c\left(1 + \frac{1}{\lambda}\right)h^4 + \left(S_h^\top(S_h u_h^\star - I_h y_d) + \lambda u_h^\star, \Pi_h u^\star - u^\star\right)_{L^2(\Omega)}. \tag{5.10}$$

Für die Abschätzung des letzten Terms setzen wir $p_h^\star := S_h^\top(S_h u_h^\star - I_h y_d)$ und erhalten

$$\begin{aligned}\big(S_h^\top(S_h u_h^\star - I_h y_d) + \lambda u_h^\star, \Pi_h u^\star - u^\star\big)_{L^2(\Omega)} \\ = \underbrace{\int_\Omega p_h^\star(\Pi_h u^\star - u^\star)\,\mathrm{d}\mathbf{x}}_{=:A} + \lambda \underbrace{\int_\Omega u_h^\star(\Pi_h u^\star - u^\star)\,\mathrm{d}\mathbf{x}}_{=:B}.\end{aligned}$$

Nach Konstruktion ist $p_h^\star \in V_h \subset H^1(\Omega)$ und es folgt

$$\begin{aligned}\|p_h^\star\|_{H^1(\Omega)} &\le c\,\|S_h u_h^\star - I_h y_d\|_{L^2(\Omega)} \\ &\le c\big\{\|S_h\|\|u^\star - u_h^\star\|_{L^2(\Omega)} + \|S_h\|\|u^\star\|_{L^2(\Omega)} + \|y_d - I_h y_d\|_{L^2(\Omega)} + \|y_d\|_{L^2(\Omega)}\big\} \\ &\le c\left\{\|u^\star - u_h^\star\|_{L^2(\Omega)} + \|u^\star\|_{L^2(\Omega)} + h^2\|y_d\|_{H^2(\Omega)} + \|y_d\|_{L^2(\Omega)}\right\} \\ &\le c\big\{\|u^\star - u_h^\star\|_{L^2(\Omega)} + 1\big\}.\end{aligned}$$

Da das $L^2(\Omega)$-Skalarprodukt ein lineares und stetiges Funktional auf $H^1(\Omega)$ darstellt, schließen wir in Anbetracht von Lemma 5.6

$$\begin{aligned}A &\le \|p_h^\star\|_{H^1(\Omega)}\|\Pi_h u^\star - u^\star\|_{\widetilde{H}^{-1}(\Omega)} \\ &\le ch^2\big\{\|u^\star - u_h^\star\|_{L^2(\Omega)} + 1\big\}\|u^\star\|_{H^1(\Omega)} \\ &\le \frac{\lambda}{8}\|u^\star - u_h^\star\|_{L^2(\Omega)}^2 + \frac{2c}{\lambda}h^4\|u^\star\|_{H^1(\Omega)}^2 + ch^2\|u^\star\|_{H^1(\Omega)}.\end{aligned}$$

Weiter erhalten wir mit Hilfe der Cauchy-Schwarzschen Ungleichung und der Youngschen Ungleichung

$$\begin{aligned}B &= \int_\Omega u^\star(\Pi_h u^\star - u^\star)\,\mathrm{d}\mathbf{x} + \int_\Omega (u_h^\star - u^\star)(\Pi_h u^\star - u^\star)\,\mathrm{d}\mathbf{x} \\ &\le \|u^\star\|_{H^1(\Omega)}\|u^\star - \Pi_h u^\star\|_{\widetilde{H}^{-1}(\Omega)} + \frac{1}{8}\|u^\star - u_h^\star\|_{L^2(\Omega)}^2 + 2\|u^\star - \Pi_h u^\star\|_{L^2(\Omega)}^2 \\ &\le ch^2\|u^\star\|_{H^1(\Omega)}^2 + \frac{1}{8}\|u^\star - u_h^\star\|_{L^2(\Omega)}^2.\end{aligned}$$

Da $\|u^\star\|_{H^1(\Omega)} < \infty$ gemäß Satz 4.18 ist, haben wir

$$A + \lambda B \le c\left(\underbrace{1+\lambda}_{\le 2} + \frac{h^2}{\lambda}\right)h^2 + \frac{\lambda}{4}\|u^\star - u_h^\star\|_{L^2(\Omega)}^2.$$

Dies eingesetzt in (5.10) liefert

$$\frac{\lambda}{4}\|u^\star - u_h^\star\|_{L^2(\Omega)}^2 \le c\left(1 + \frac{1}{\lambda}\right)h^4 + c\left(2 + \frac{h^2}{\lambda}\right)h^2,$$

woraus wir das Behauptete erhalten, wenn wir beachten, dass sich die noch fehlende Abschätzung für $y^\star - y_h^\star$ analog wie im Beweis von Satz 5.4 ergibt. □

Wir beobachten, dass die Konvergenzordnung bei der Volldiskretisierung schlechter ist als bei der Semidiskretisierung. Daran ändert auch eine Diskretisierung der Steuerung mit stückweise linearen Finiten Elementen nichts, da im allgemeinen nur $u^\star \in H^1(\Omega)$ gilt. Das zeigen auch numerische Rechenbeispiele. Das Ergebnis kann allerdings als optimal angesehen werden, da die Konvergenzordnung genauso gut ist wie der Fehler der Bestapproximation an $u^\star$ mit stückweise konstanten Ansatzfunktionen.

5.4 Umwandlung in ein quadratisches Programm

Wir setzen $N = \dim V_h$ und $M = \dim U_h$. Ferner bezeichne $\{\varphi_j\}_{i=1}^N$ die stückweise lineare nodale Basis in V_h und $\{\chi_i\}_{i=1}^M$ die stückweise konstanten Ansatzfunktionen in U_h. Um das volldiskrete Optimalsteuerungsproblem (5.5) als endlichdimensionales Optimierungsproblem in Matrix-Vektor-Form zu schreiben, benötigen wir die Steifigkeitsmatrix $\mathbf{A}_h \in \mathbb{R}^{N\times N}$ und die Massenmatrix $\mathbf{M}_h \in \mathbb{R}^{N\times N}$ der stückweise linearen Finite-Elemente-Methode:

$$\mathbf{A}_h = \left[\int_\Omega \langle \nabla\varphi_i, \nabla\varphi_j\rangle \,\mathrm{d}\mathbf{x}\right]_{i,j=1}^N, \quad \mathbf{M}_h = \left[\int_\Omega \varphi_i\varphi_j \,\mathrm{d}\mathbf{x}\right]_{i,j=1}^N.$$

Ferner seien

$$\mathbf{B}_h = \left[\int_\Omega \varphi_i\chi_j \,\mathrm{d}\mathbf{x}\right]_{i,j=1}^{N,M} \in \mathbb{R}^{N\times M}$$

diejenige Matrix, welche stückweise konstante und stückweise lineare Finite Elemente koppelt, und

$$\mathbf{N}_h = \left[\int_\Omega \chi_i\chi_j \,\mathrm{d}\mathbf{x}\right]_{i,j=1}^{M} \in \mathbb{R}^{M\times M}$$

die Massenmatrix für stückweise konstante Finite Elemente.

Für die Übersetzung von (5.5) benötigen wir außerdem die zu $I_h y_d$, u_a und u_b gehörigen Datenvektoren. Der Vektor $\mathbf{y}_d \in \mathbb{R}^N$ ergibt sich gemäß Definition als stückweise lineare Interpolante von y_d. Ist also $\mathbf{x}_i$ derjenige Knoten, der mit der Basisfunktion φ_i durch $\varphi_i(\mathbf{x}_i) = 1$ assoziiert ist, so gilt $[\mathbf{y}_d]_i = y_d(\mathbf{x}_i)$. Die diskreten Box-Beschränkungen $\mathbf{u}_a \in \mathbb{R}^M$ und $\mathbf{u}_b \in \mathbb{R}^M$ erhalten wir durch Lösen von

$$\mathbf{N}_h\mathbf{u}_a = \left[\int_\Omega u_a\chi_i \,\mathrm{d}\mathbf{x}\right]_{i=1}^M, \quad \mathbf{N}_h\mathbf{u}_b = \left[\int_\Omega u_b\chi_i \,\mathrm{d}\mathbf{x}\right]_{i=1}^M.$$

Dann führt der Ansatz

$$u_h = \sum_{i=1}^M u_i\chi_i \in U_h, \quad y_h = \sum_{i=1}^N y_i\varphi_i \in V_h$$

mit den Vektoren

$$\mathbf{u}_h = \left[u_i\right]_{i=1}^M \in \mathbb{R}^M, \quad \mathbf{y}_h = \left[y_i\right]_{i=1}^N \in \mathbb{R}^N$$

auf das folgende Optimierungsproblem in Matrix-Vektor-Form:

$$\left.\begin{aligned} &\text{minimiere } \frac{1}{2}(\mathbf{y}_h - \mathbf{y}_d)^\top \mathbf{M}_h(\mathbf{y}_h - \mathbf{y}_d) + \frac{\lambda}{2}\mathbf{u}_h^\top \mathbf{N}_h\mathbf{u}_h \\ &\text{unter den Nebenbedingungen } \mathbf{A}_h\mathbf{y}_h = \mathbf{B}_h\mathbf{u}_h \text{ und } \mathbf{u}_a \le \mathbf{u}_h \le \mathbf{u}_b. \end{aligned}\right\} \tag{5.11}$$

Definition 5.8 (quadratisches Programm) Es sei $\mathbf{Q} \in \mathbb{R}^{n\times n}$ eine symmetrische Matrix und $\mathbf{q} \in \mathbb{R}^n$ ein Vektor. Ferner seien $\mathbf{G} \in \mathbb{R}^{m\times n}$ und $\mathbf{H} \in \mathbb{R}^{p\times n}$ beliebige Matrizen mit vollem Rang und $\mathbf{g} \in \mathbb{R}^m$ und $\mathbf{h} \in \mathbb{R}^p$. Ein Optimierungsproblem für $\mathbf{z} \in \mathbb{R}^n$ der Form

$$\begin{aligned} &\text{minimiere } f(\mathbf{z}) = \frac{1}{2}\mathbf{z}^\top\mathbf{Q}\mathbf{z} + \mathbf{q}^\top\mathbf{z} \\ &\text{unter den Nebenbedingungen } \mathbf{G}\mathbf{z} = \mathbf{g} \text{ und } \mathbf{H}\mathbf{z} \le \mathbf{h} \end{aligned}$$

heißt **quadratisches Programm**.

Wir erhalten aus (5.11) ein quadratisches Programm, indem wir die Matrizen

$$\mathbf{Q} = \begin{bmatrix} \mathbf{M}_h & \mathbf{0} \\ \mathbf{0} & \lambda \mathbf{N}_h \end{bmatrix} \in \mathbb{R}^{(N+M)\times(N+M)}, \quad \mathbf{G} = \begin{bmatrix} \mathbf{A}_h & -\mathbf{B}_h \end{bmatrix} \in \mathbb{R}^{N\times(N+M)}$$

und

$$\mathbf{H} = \begin{bmatrix} \mathbf{0} & -\mathbf{I} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \in \mathbb{R}^{2M\times(N+M)}$$

sowie die Vektoren

$$\mathbf{z} = \begin{bmatrix} \mathbf{y}_h \\ \mathbf{u}_h \end{bmatrix} \in \mathbb{R}^{N+M}, \quad \mathbf{q} = \begin{bmatrix} -\mathbf{M}_h \mathbf{y}_d \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{N+M}, \quad \mathbf{h} = \begin{bmatrix} -\mathbf{u}_a \\ \mathbf{u}_b \end{bmatrix} \in \mathbb{R}^{2M}.$$

einführen. Damit erhalten wir das zu (5.11) äquivalente Problem

$$\left.\begin{aligned} &\text{minimiere } f(\mathbf{z}) = \frac{1}{2}\mathbf{z}^\top \mathbf{Q}\mathbf{z} + \mathbf{q}^\top \mathbf{z} \\ &\text{unter den Nebenbedingungen } \mathbf{G}\mathbf{z} = \mathbf{0} \text{ und } \mathbf{H}\mathbf{z} \leq \mathbf{h}. \end{aligned}\right\} \tag{5.12}$$

Hierbei haben wir im Zielfunktional den konstanten Anteil $\mathbf{y}_d^\top \mathbf{M}_h \mathbf{y}_d$ weggelassen, da er bei der Optimierung keine Rolle spielt.

Das quadratisches Programm (5.12) kann mit Hilfe von Standardsoftware gelöst werden, vorausgesetzt, die Dimension ist nicht zu groß. So gibt es in MATLAB etwa die Funktion `quadprog`. Um allerdings große Optimalsteuerungsprobleme lösen zu können, werden wir im nächsten Abschnitt gezielt auf das projizierte Gradientenverfahren und die Aktive-Mengen-Strategie schauen.

6. Nichtlineare Optimierung

6.1 Projiziertes Gradientenverfahren

Es seien $f : \mathbb{R}^n \to \mathbb{R}$ eine stetig differenzierbare Funktion und $K \subset \mathbb{R}^n$ eine abgeschlossene, konvexe Menge. Das projizierten Gradientenverfahren ist ein Verfahren zur numerischen Lösung des Optimierungsproblems

$$\text{minimiere } f(\mathbf{x}) \text{ unter der Nebenbedingung } \mathbf{x} \in K. \tag{6.1}$$

Wie wir in in Satz 4.9 gezeigt haben, lautet die notwendige Optimalitätsbedingung für ein Minimum $\mathbf{x}^\star \in K$ von (6.1)

$$\nabla f(\mathbf{x}^\star)^\top (\mathbf{x} - \mathbf{x}^\star) \geq 0 \text{ für alle } \mathbf{x} \in K. \tag{6.2}$$

Beispiel 6.1 Im Fall des volldiskreten Optimierungsproblem (5.5) ist f das reduzierte Funktional

$$f(\mathbf{u}_h) = \frac{1}{2}(\mathbf{y}_h - \mathbf{y}_d)^\top \mathbf{M}_h (\mathbf{y}_h - \mathbf{y}_d) + \frac{\lambda}{2} \mathbf{u}_h^\top \mathbf{N}_h \mathbf{u}_h \text{ mit } \mathbf{A}_h \mathbf{y}_h = \mathbf{B}_h \mathbf{u}_h$$

und $K := [\mathbf{u}_a, \mathbf{u}_b]$. Die notwendige Bedingung erster Ordnung für $\mathbf{u}_h^\star \in K$ lautet

$$\big(\nabla f(\mathbf{u}_h^\star)\big)^\top (\mathbf{u}_h - \mathbf{u}_h^\star) \geq 0 \text{ für alle } \mathbf{u}_h \in K,$$

wobei

$$\nabla f(\mathbf{u}_h) = \mathbf{B}_h^\top \mathbf{p}_h + \lambda \mathbf{N}_h \mathbf{u}_h \text{ mit } \mathbf{A}_h^\top \mathbf{p}_h = \mathbf{M}_h (\mathbf{y}_h - \mathbf{y}_d).$$

△

Grundlage des projizierten Gradientenverfahrens ist die orthogonale Projektion auf die zulässige Menge.

Definition 6.2 Es sei $K \subset \mathbb{R}^n$ eine abgeschlossene, konvexe Menge. Dann ist die **orthogonale Projektion** $\mathbf{P}_K : \mathbb{R}^n \to K$ definiert durch die Bedingung

$$\|\mathbf{P}_K(\mathbf{x}) - \mathbf{x}\|_2 = \min_{\mathbf{y} \in K} \|\mathbf{y} - \mathbf{x}\|_2.$$

Der Punkt $\mathbf{P}_K(\mathbf{x}) \in K$ besitzt also die Eigenschaft, den kürzesten Abstand zu einem gegebenen Punkt $\mathbf{x} \in \mathbb{R}^n$ zu besitzen.

Beispiel 6.3 Im Fall der Box-Beschränkungen $K = [\mathbf{u}_a, \mathbf{u}_b]$ ist $\mathbf{P}_K$ koordinatenweise gegeben durch

$$\big[\mathbf{P}_K(\mathbf{u})\big]_i = \begin{cases} [\mathbf{u}_a]_i, & \text{falls } [\mathbf{u}]_i < [\mathbf{u}_a]_i, \\ [\mathbf{x}]_i, & \text{falls } [\mathbf{u}_a]_i \le [\mathbf{u}]_i \le [\mathbf{u}_b]_i, \\ [\mathbf{u}_b]_i, & \text{falls } [\mathbf{u}_b]_i < [\mathbf{u}]_i, \end{cases} \qquad i = 1, \dots, M.$$

△

Die Grundversion des projizierten Gradientenverfahren ist im folgenden Algorithmus beschrieben:

Algorithmus 6.4 (projiziertes Gradientenverfahren)

input: Funktion $f : \mathbb{R}^n \to \mathbb{R}$, konvexe zulässige Menge $K \subset \mathbb{R}^n$ und Startnäherung $\mathbf{x}_0 \in K$

output: Folge von Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$

① Initialisierung: wähle $\sigma \in (0,1)$ und setze $k := 0$

② berechne den Antigradienten $\mathbf{d}_k := -\nabla f(\mathbf{x}_k)$ und setze $\alpha_k := 1$

③ solange

$$f\big(\mathbf{P}_K(\mathbf{x}_k + \alpha_k \mathbf{d}_k)\big) > f(\mathbf{x}_k) - \sigma \mathbf{d}_k^\top \big(\mathbf{P}_K(\mathbf{x}_k + \alpha_k \mathbf{d}_k) - \mathbf{x}_k\big) \tag{6.3}$$

setze $\alpha_k := \alpha_k/2$

④ setze $\mathbf{x}_{k+1} := \mathbf{P}_K(\mathbf{x}_k + \alpha_k \mathbf{d}_k)$

⑤ erhöhe $k := k + 1$ und gehe nach ②

Für den Fall der Minimierung ohne Nebenbedingungen, das heißt $K = \mathbb{R}^n$, stellt obiger Algorithmus das klassische Gradientverfahren dar. Insbesondere geht die Bedingung an die Reduktion des Funktionals über in die Armijo-Goldstein-Bedingung

$$f(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k) - \sigma \alpha_k \|\nabla f(\mathbf{x}_k)\|_2^2.$$

Man kann zeigen, dass für das projizierte Gradientenverfahren ein $\alpha_k > 0$ existiert, für das die Reduktionsbedingung erfüllt ist.

Lemma 6.5 Die orthogonale Projektion $\mathbf{P}_K$ besitzt die folgenden Eigenschaften:

(*i.*) Es gilt $\big(\mathbf{P}_K(\mathbf{x}) - \mathbf{x}\big)^\top \big(\mathbf{P}_K(\mathbf{x}) - \mathbf{y}\big) \le 0$ für alle $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in K$.

(*ii.*) Es gilt $\big(\mathbf{P}_K(\mathbf{y}) - \mathbf{P}_K(\mathbf{x})\big)^\top (\mathbf{y} - \mathbf{x}) \ge \|\mathbf{P}_K(\mathbf{y}) - \mathbf{P}_K(\mathbf{x})\|_2^2 \ge 0$ für alle $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, das heißt, $\mathbf{P}_K$ ist monoton.

(*iii.*) Es gilt $\|\mathbf{P}_K(\mathbf{y}) - \mathbf{P}_K(\mathbf{x})\|_2 \le \|\mathbf{y} - \mathbf{x}\|_2$ für alle $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, das heißt, $\mathbf{P}_K$ ist nicht expandierend.

Beweis. (*i.*) Wegen der Konvexität folgt aus $\mathbf{y} \in K$ auch $\widehat{\mathbf{y}} := (1-t)\mathbf{P}_K(\mathbf{x}) + t\mathbf{y} \in K$ für alle $t \in [0,1]$. Aus

$$\begin{aligned} \|\widehat{\mathbf{y}} - \mathbf{x}\|_2^2 &= \|\widehat{\mathbf{y}} - \mathbf{P}_K(\mathbf{x}) + \mathbf{P}_K(\mathbf{x}) - \mathbf{x}\|_2^2 \\ &= \|\widehat{\mathbf{y}} - \mathbf{P}_K(\mathbf{x})\|_2^2 + \|\mathbf{P}_K(\mathbf{x}) - \mathbf{x}\|_2^2 - 2\big(\mathbf{P}_K(\mathbf{x}) - \mathbf{x}\big)^\top \big(\mathbf{P}_K(\mathbf{x}) - \widehat{\mathbf{y}}\big) \end{aligned}$$

folgt aufgrund der Minimierungseigenschaft von $\mathbf{P}_K$, dass

$$\|\widehat{\mathbf{y}} - \mathbf{P}_K(\mathbf{x})\|_2^2 - 2\big(\mathbf{P}_K(\mathbf{x}) - \mathbf{x}\big)^\top \big(\mathbf{P}_K(\mathbf{x}) - \widehat{\mathbf{y}}\big) = \|\widehat{\mathbf{y}} - \mathbf{x}\|_2^2 - \|\mathbf{P}_K(\mathbf{x}) - \mathbf{x}\|_2^2 \geq 0.$$

Einsetzen von $\mathbf{P}_K(\mathbf{x}) - \widehat{\mathbf{y}} = t\big(\mathbf{P}_K(\mathbf{x}) - \mathbf{y}\big)$ führt auf

$$t^2\|\mathbf{y} - \mathbf{P}_K(\mathbf{x})\|_2^2 - 2t\big(\mathbf{P}_K(\mathbf{x}) - \mathbf{x}\big)^\top \big(\mathbf{P}_K(\mathbf{x}) - \mathbf{y}\big) \geq 0,$$

was für $t \to 0$ die gewünschte Aussage liefert.
(*ii.*) Die bereits bewiesene Aussage (*i.*) impliziert

$$\begin{aligned}\big(\mathbf{P}_K(\mathbf{x}) - \mathbf{x}\big)^\top \big(\mathbf{P}_K(\mathbf{x}) - \mathbf{P}_K(\mathbf{y})\big) &\leq 0,\\ \big(\mathbf{P}_K(\mathbf{y}) - \mathbf{y}\big)^\top \big(\mathbf{P}_K(\mathbf{y}) - \mathbf{P}_K(\mathbf{x})\big) &\leq 0.\end{aligned}$$

Zusammen führt dies auf

$$\big(\mathbf{P}_K(\mathbf{y}) - \mathbf{y} + \mathbf{x} - \mathbf{P}_K(\mathbf{x})\big)^\top \big(\mathbf{P}_K(\mathbf{y}) - \mathbf{P}_K(\mathbf{x})\big) \leq 0,$$

das ist Aussage (*ii.*).
(*iii.*) Diese Aussage folgt sofort aus Aussage (*ii.*) durch Anwenden der Cauchy-Schwarzschen Ungleichung. □

Bemerkung Aus der Monotonieeigenschaft (*ii.*) folgt wegen $\mathbf{x}_k = \mathbf{P}_K(\mathbf{x}_k)$ für die neue Iterierte $\mathbf{x}_{k+1} = \mathbf{P}_K\big(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)\big)$ des projizierten Gradientenverfahrens, dass

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 \leq (\mathbf{x}_{k+1} - \mathbf{x}_k)^\top (\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k) - \mathbf{x}_k).$$

Wir erhalten daher

$$-\nabla f(\mathbf{x}_k)^\top (\mathbf{x}_{k+1} - \mathbf{x}_k) \geq \frac{1}{\alpha_k}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2. \tag{6.4}$$

Dies bedeutet, dass durch die Abstiegsbedingung (6.3) des Algorithmus 6.4 tatsächlich $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$ erreicht wird. △

Lemma 6.6 Für beliebige $\mathbf{x} \in \mathbb{R}^n$ und $\mathbf{d} \in \mathbb{R}^n$ ist die Funktion

$$\varphi(\alpha) := \frac{\|\mathbf{P}_K(\mathbf{x} + \alpha\mathbf{d}) - \mathbf{x}\|_2}{\alpha}$$

für alle $\alpha > 0$ monoton fallend.

Beweis. (*i.*) Für $0 < \alpha < \beta$ setzen wir

$$\mathbf{u} := \mathbf{P}_K(\mathbf{x} + \alpha\mathbf{d}) - \mathbf{x}, \quad \mathbf{v} := \mathbf{P}_K(\mathbf{x} + \beta\mathbf{d}) - \mathbf{x}$$

und erhalten unter Verwendung von Lemma 6.5 (*i.*)

$$\begin{aligned}\mathbf{u}^\top(\mathbf{u} - \mathbf{v}) &= \{\mathbf{P}_K(\mathbf{x} + \alpha\mathbf{d}) - (\mathbf{x} + \alpha\mathbf{d}) + \alpha\mathbf{d}\}^\top \{\mathbf{P}_K(\mathbf{x} + \alpha\mathbf{d}) - \mathbf{P}_K(\mathbf{x} + \beta\mathbf{d})\}\\ &\leq \alpha\mathbf{d}^\top \{\mathbf{P}_K(\mathbf{x} + \alpha\mathbf{d}) - \mathbf{P}_K(\mathbf{x} + \beta\mathbf{d})\}\end{aligned}$$

und analog

$$\mathbf{v}^\top(\mathbf{v}-\mathbf{u}) \leq \beta\mathbf{d}^\top\{\mathbf{P}_K(\mathbf{x}+\beta\mathbf{d}) - \mathbf{P}_K(\mathbf{x}+\alpha\mathbf{d})\}.$$

Zusammen ergibt dies

$$\frac{\mathbf{u}^\top(\mathbf{u}-\mathbf{v})}{\alpha} \leq \frac{\mathbf{v}^\top(\mathbf{u}-\mathbf{v})}{\beta}. \tag{6.5}$$

(*ii.*) Weiter erhalten wir mit Lemma 6.5 (*ii.*)

$$\begin{aligned}
\mathbf{u}^\top(\mathbf{u}-\mathbf{v}) &\leq \alpha\mathbf{d}^\top\{\mathbf{P}_K(\mathbf{x}+\alpha\mathbf{d}) - \mathbf{P}_K(\mathbf{x}+\beta\mathbf{d})\} \\
&= -\frac{\alpha}{\beta-\alpha}(\alpha\mathbf{d}-\beta\mathbf{d})^\top\{\mathbf{P}_K(\mathbf{x}+\beta\mathbf{d}) - \mathbf{P}_K(\mathbf{x}+\alpha\mathbf{d})\} \\
&\leq -\frac{\alpha}{\beta-\alpha}\|\mathbf{P}_K(\mathbf{x}+\beta\mathbf{d}) - \mathbf{P}_K(\mathbf{x}+\alpha\mathbf{d})\|_2^2 \\
&\leq 0.
\end{aligned}$$

Aus der Cauchy-Schwarzschen Ungleichung ergibt sich

$$\mathbf{u}^\top\mathbf{v}(\|\mathbf{u}\|_2 + \|\mathbf{v}\|_2) \leq \|\mathbf{u}\|_2\|\mathbf{v}\|_2(\|\mathbf{u}\|_2 + \|\mathbf{v}\|_2),$$

woraus

$$\|\mathbf{u}\|_2\mathbf{v}^\top(\mathbf{u}-\mathbf{v}) = \|\mathbf{u}\|_2(\mathbf{u}^\top\mathbf{v} - \|\mathbf{v}\|_2^2) \leq \|\mathbf{v}\|_2(\|\mathbf{u}\|_2^2 - \mathbf{u}^\top\mathbf{v}) = \|\mathbf{v}\|_2\mathbf{u}^\top(\mathbf{u}-\mathbf{v}) \tag{6.6}$$

folgt.

(*iii.*) Wir unterscheiden nun zwei Fälle: Für $\mathbf{u}^\top(\mathbf{u}-\mathbf{v}) = 0$ gilt $\mathbf{P}_K(\mathbf{x}+\alpha\mathbf{d}) = \mathbf{P}_K(\mathbf{x}+\beta\mathbf{d})$ und somit $\mathbf{u} = \mathbf{v}$. Hieraus folgt unmittelbar auch

$$\varphi(\alpha) = \frac{\|\mathbf{u}\|_2}{\alpha} \geq \frac{\|\mathbf{v}\|_2}{\beta} = \varphi(\beta).$$

Für den Fall $\mathbf{u}^\top(\mathbf{u}-\mathbf{v}) < 0$ folgt aus (6.5)

$$\frac{\beta}{\alpha} \geq \frac{\mathbf{v}^\top(\mathbf{u}-\mathbf{v})}{\mathbf{u}^\top(\mathbf{u}-\mathbf{v})}$$

und aus (6.6)

$$\|\mathbf{v}\|_2 \leq \|\mathbf{u}\|_2\frac{\mathbf{v}^\top(\mathbf{u}-\mathbf{v})}{\mathbf{u}^\top(\mathbf{u}-\mathbf{v})}.$$

Kombiniert man diese zwei Ungleichungen, so erhält man wieder

$$\varphi(\alpha) = \frac{\|\mathbf{u}\|_2}{\alpha} \geq \frac{\|\mathbf{v}\|_2}{\beta} = \varphi(\beta).$$

□

Satz 6.7 Die Funktion $f : \mathbb{R}^n \to \mathbb{R}$ sei auf K stetig differenzierbar und nach unten beschränkt. Weiter sei ∇f auf K gleichmäßig stetig. Dann gilt für die Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ des projizierten Gradientenverfahrens

$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2}{\alpha_k} = 0.$$

Beweis. Wir führen einen Widerspruchsbeweis. Angenommen, es existiert zu jedem $\varepsilon > 0$ eine unendliche Teilfolge $\{k_\ell\}_{\ell\in\mathbb{N}}$, so dass

$$\frac{\|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2}{\alpha_{k_\ell}} \geq \varepsilon.$$

Dann gilt insbesondere auch

$$\frac{\|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2^2}{\alpha_{k_\ell}} \geq \varepsilon \max\{\varepsilon\alpha_{k_\ell}, \|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2\}. \tag{6.7}$$

Da die Folge $\{f(\mathbf{x}_{k_\ell})\}_{\ell\in\mathbb{N}}$ monoton fallend und nach unten beschränkt ist, folgt aus der Abstiegsbedingung (6.3) des projizierten Gradientenverfahrens

$$\lim_{\ell\to\infty} \nabla f(\mathbf{x}_{k_\ell})^\top(\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}) = 0,$$

was wiederum gemäß (6.4)

$$\lim_{\ell\to\infty} \frac{\|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2^2}{\alpha_{k_\ell}} = 0 \tag{6.8}$$

nach sich zieht. Aufgrund von (6.7) erhalten wir hieraus

$$\lim_{\ell\to\infty} \alpha_{k_\ell} = 0 \quad \text{und} \quad \lim_{\ell\to\infty} \|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2 = 0.$$

Für $\mathbf{y}_{k_\ell+1} := \mathbf{P}_K(\mathbf{x}_{k_\ell} + 2\alpha_{k_\ell}\mathbf{d}_{k_\ell})$ gilt aufgrund der algorithmischen Umsetzung des projizierten Gradientenverfahrens

$$f(\mathbf{y}_{k_\ell+1}) > f(\mathbf{x}_{k_\ell}) + \sigma\nabla f(\mathbf{x}_{k_\ell})^\top(\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}),$$

also auch

$$f(\mathbf{x}_{k_\ell}) - f(\mathbf{y}_{k_\ell+1}) < \sigma\nabla f(\mathbf{x}_{k_\ell})^\top(\mathbf{x}_{k_\ell} - \mathbf{y}_{k_\ell+1}). \tag{6.9}$$

Aus Lemma 6.6 folgt

$$\frac{\|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2^2}{\alpha_{k_\ell}} \geq \|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2 \frac{\|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2}{2\alpha_{k_\ell}} \geq \alpha_{k_\ell}\varepsilon \frac{\|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2}{2\alpha_{k_\ell}} = \frac{\varepsilon}{2}\|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2.$$

Weiter ergibt sich mit Lemma 6.5 (*ii.*)

$$(\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell})^\top \Big\{ \underbrace{\mathbf{x}_{k_\ell} - \alpha_{k_\ell}\nabla f(\mathbf{x}_{k_\ell}) - \mathbf{x}_{k_\ell}}_{=-\alpha_{k_\ell}\nabla f(\mathbf{x}_{k_\ell})} \Big\} \geq \|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2^2,$$

$$(\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell+1})^\top \Big\{ \underbrace{\mathbf{x}_{k_\ell} - 2\alpha_{k_\ell}\nabla f(\mathbf{x}_{k_\ell}) - \big(\mathbf{x}_{k_\ell} - \alpha_{k_\ell}\nabla f(\mathbf{x}_{k_\ell})\big)}_{=-\alpha_{k_\ell}\nabla f(\mathbf{x}_{k_\ell})} \Big\} \geq 0.$$

Zusammen führt dies auf

$$\begin{aligned}(\mathbf{x}_{k_\ell} - \mathbf{y}_{k_\ell+1})^\top\nabla f(\mathbf{x}_{k_\ell}) &\geq (\mathbf{x}_{k_\ell} - \mathbf{x}_{k_\ell+1})^\top\nabla f(\mathbf{x}_{k_\ell}) \\ &\geq \frac{\|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2^2}{\alpha_{k_\ell}} \\ &\geq \frac{\varepsilon}{2}\|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2.\end{aligned}$$

Speziell ergibt sich wegen (6.8) auch $\|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2 \to 0$ für $\ell \to \infty$. Die gleichmäßige Stetigkeit von ∇f impliziert daher

$$\begin{aligned}\left|1 - \frac{f(\mathbf{x}_{k_\ell}) - f(\mathbf{y}_{k_\ell+1})}{(\mathbf{x}_{k_\ell} - \mathbf{y}_{k_\ell+1})^\top \nabla f(\mathbf{x}_{k_\ell})}\right| &= \frac{o(\|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2)}{(\mathbf{x}_{k_\ell} - \mathbf{y}_{k_\ell+1})^\top \nabla f(\mathbf{x}_{k_\ell})} \\ &\leq \frac{2}{\varepsilon} \frac{o(\|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2)}{\|\mathbf{x}_{k_\ell} - \mathbf{y}_{k_\ell+1}\|_2} \\ &\overset{\ell\to\infty}{\longrightarrow} 0.\end{aligned}$$

Dies steht jedoch im Widerspruch zu der aus (6.9) folgenden Abschätzung

$$\frac{f(\mathbf{x}_{k_\ell}) - f(\mathbf{y}_{k_\ell+1})}{(\mathbf{x}_{k_\ell} - \mathbf{y}_{k_\ell+1})^\top \nabla f(\mathbf{x}_{k_\ell})} < \sigma < 1.$$

□

Bemerkung Da $\alpha_k \leq 1$ ist, folgt aus Satz 6.7 speziell $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 \to 0$ für $k \to \infty$. △

Definition 6.8 Eine Menge $C \subset \mathbb{R}^n$ heißt **Kegel**, wenn aus $\mathbf{x} \in C$ auch $\lambda\mathbf{x} \in C$ folgt für alle $\lambda \geq 0$. Der **Tangentialkegel** $T_D(\mathbf{x})$ von der Menge $D \subset \mathbb{R}^n$ an einen Punkt $\mathbf{x} \in D$ ist der kleinste abgeschlossener Kegel, der die Menge

$$M := \{\mathbf{d} = \mathbf{y} - \mathbf{x} : \mathbf{y} \in D\}$$

enthält.

Bemerkung Es sei $\mathbf{x} \in K$ und $\{\mathbf{y}_k\}_{k\in\mathbb{N}} \subset K \setminus \{\mathbf{x}\}$ eine Folge mit $\lim_{k\to\infty} \mathbf{y}_k = \mathbf{x}$. Dann ist

$$\mathbf{d} = \lim_{k\to\infty} \frac{\mathbf{y}_k - \mathbf{x}}{\|\mathbf{y}_k - \mathbf{x}\|_2}$$

offenbar im Tangentialkegel $T_K(\mathbf{x})$ enthalten. Die Richtung $\mathbf{d}$ wird *Grenzrichtung* der Folge genannt. Umgekehrt gibt es zu jedem $\mathbf{d} \in T_K(\mathbf{x})$ mit $\|\mathbf{d}\|_2 = 1$ eine Folge $\{\mathbf{y}_k\}_{k\in\mathbb{N}} \subset K$ derart, dass

$$\mathbf{d} = \lim_{k\to\infty} \frac{\mathbf{y}_k - \mathbf{x}}{\|\mathbf{y}_k - \mathbf{x}\|_2} \quad \text{und} \quad \lim_{k\to\infty} \mathbf{y}_k = \mathbf{x}. \tag{6.10}$$

Der Tangentialkegel enthält also gerade die Grenzrichtungen von allen Folgen $\{\mathbf{y}_k\}_{k\in\mathbb{N}} \subset K \setminus \{\mathbf{x}\}$ mit $\lim_{k\to\infty} \mathbf{y}_k = \mathbf{x}$. Insbesondere ist der Tangentialkegel konvex, weil K konvex ist. △

Lemma 6.9 Für jeden Punkt $\mathbf{x} \in K$ erfüllt die orthogonale Projektion $\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)$ der Richtung des steilsten Abstiegs auf den Tangentialkegel $T_K(\mathbf{x})$ die folgenden Eigenschaften:

(*i.*) Es gilt

$$\nabla f(\mathbf{x})^\top \mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) = -\big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2^2.$$

(*ii.*) Es ist

$$\min\{\nabla f(\mathbf{x})^\top \mathbf{d} : \mathbf{d} \in T_K(\mathbf{x}),\ \big\|\mathbf{d}\|_2 \leq 1\} = -\big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2.$$

(*iii.*) Der Punkt $\mathbf{x}$ ist genau dann ein stationärer Punkt des Minimierungsproblems mit Nebenbedingungen (6.1), wenn $\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) = \mathbf{0}$.

Beweis. (*i.*) Nach Definition der Orthogonalprojektion besitzt die Funktion

$$g(\lambda) := \frac{1}{2}\big\|\lambda \mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) + \nabla f(\mathbf{x})\big\|_2^2$$

ein Minimum bei $\lambda = 1$. Daher gilt

$$g'(1) := \big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2^2 + \nabla f(\mathbf{x})^\top \mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) = 0.$$

(*ii.*) Wegen Aussage (*i.*) gilt

$$\big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) + \nabla f(\mathbf{x})\big\|_2^2 = \|\nabla f(\mathbf{x})\|_2^2 - \big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2^2.$$

Für alle $\mathbf{d} \in T_K(\mathbf{x})$ mit $\|\mathbf{d}\|_2 \leq \big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2$ gilt nach Definition der orthogonalen Projektion

$$\begin{aligned}\big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) + \nabla f(\mathbf{x})\big\|_2^2 &\leq \|\mathbf{d} + \nabla f(\mathbf{x})\|_2^2 \\ &\leq \big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2^2 + 2\nabla f(\mathbf{x})^\top \mathbf{d} + \|\nabla f(\mathbf{x})\|_2^2.\end{aligned}$$

Zusammen ergibt dies

$$\nabla f(\mathbf{x})^\top \mathbf{d} \geq -\big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2^2.$$

Das Behauptete erhält man, indem man $\widehat{\mathbf{d}} = \mathbf{d}/\big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2$ setzt.

(*iii.*) Definitionsgemäß ist $\mathbf{x} \in K$ genau dann ein stationärer Punkt, wenn $\nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) \geq 0$ für alle $\mathbf{y} \in K$ ist. Dies ist gleichbedeutend damit, dass $\nabla f(\mathbf{x})^\top \mathbf{d} \geq 0$ für alle $\mathbf{d} \in T_K(\mathbf{x})$ ist. Aussage (*ii.*) impliziert, dass dies genau dann der Fall ist, wenn $\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) = \mathbf{0}$ erfüllt ist. □

Bemerkung Ist $\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) \neq \mathbf{0}$, so kann Aussage (*ii.*) des obigen Lemmas auch als

$$\min\{\nabla f(\mathbf{x})^\top \mathbf{d} : \mathbf{d} \in T_K(\mathbf{x}),\ \big\|\mathbf{d}\|_2 = 1\} = -\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2 \tag{6.11}$$

geschrieben werden, denn das Minimum wird für $\|\mathbf{d}\|_2 = 1$ angenommen. △

Satz 6.10 Die Funktion $f : \mathbb{R}^n \to \mathbb{R}$ sei auf K stetig differenzierbar und nach unten beschränkt. Weiter sei ∇f auf K gleichmäßig stetig. Dann gilt für die Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ des projizierten Gradientenverfahrens

$$\lim_{k\to\infty} \mathbf{P}_{T_K(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big) = \mathbf{0}.$$

Beweis. Zu beliebigen $\varepsilon > 0$ gibt es nach Lemma 6.9 (*ii.*) zu jeder Iterierten $\mathbf{x}_k$ ein $\mathbf{d}_k \in T_K(\mathbf{x}_k)$ mit $\|\mathbf{d}_k\|_2 = 1$, so dass

$$\nabla f(\mathbf{x}_k)^\top \mathbf{d}_k \leq -\big\|\mathbf{P}_{T_K(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big)\big\|_2 + \varepsilon \tag{6.12}$$

gilt. Da $\mathbf{d}_k$ Grenzrichtung einer zulässigen Folge ist, gibt es ein $\mathbf{y}_k \in K$ mit

$$\left\| \frac{\mathbf{y}_k - \mathbf{x}_k}{\|\mathbf{y}_k - \mathbf{x}_k\|_2} - \mathbf{d}_k \right\|_2 \leq \varepsilon.$$

Aus Lemma 6.5 (*i.*) folgt

$$\begin{aligned}
&\big\{\mathbf{x}_{k+1} - \big(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)\big)\big\}^\top (\mathbf{x}_{k+1} - \mathbf{y}_{k+1}) \\
&\quad = \big\{\mathbf{P}_K\big(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)\big) - \big(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)\big)\big\}^\top \big\{\mathbf{P}_K\big(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)\big) - \mathbf{y}_{k+1}\big\} \\
&\quad \leq 0,
\end{aligned}$$

was auf

$$\alpha_k \nabla f(\mathbf{x}_k)^\top (\mathbf{x}_{k+1} - \mathbf{y}_{k+1}) \leq \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 \|\mathbf{x}_{k+1} - \mathbf{y}_{k+1}\|_2,$$

beziehungsweise

$$-\frac{\nabla f(\mathbf{x}_k)^\top (\mathbf{y}_{k+1} - \mathbf{x}_{k+1})}{\|\mathbf{x}_{k+1} - \mathbf{y}_{k+1}\|_2} \leq \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2}{\alpha_k},$$

führt. Insgesamt erhalten wir deshalb

$$\begin{aligned}
-\nabla f(\mathbf{x}_k)^\top \mathbf{d}_{k+1} &\leq \|\nabla f(\mathbf{x}_k)\|_2 \left\| \frac{\mathbf{y}_{k+1} - \mathbf{x}_{k+1}}{\|\mathbf{y}_{k+1} - \mathbf{x}_{k+1}\|_2} - \mathbf{d}_{k+1} \right\|_2 - \frac{\nabla f(\mathbf{x}_k)^\top (\mathbf{y}_{k+1} - \mathbf{x}_{k+1})}{\|\mathbf{y}_{k+1} - \mathbf{x}_{k+1}\|_2} \\
&\leq \varepsilon \|\nabla f(\mathbf{x}_k)\|_2 + \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2}{\alpha_k}.
\end{aligned}$$

Die Kombination mit (6.12) ergibt

$$\begin{aligned}
&\big\|\mathbf{P}_{T_K(\mathbf{x}_{k+1})}\big(-\nabla f(\mathbf{x}_{k+1})\big)\big\|_2 \\
&\quad \leq -\nabla f(\mathbf{x}_{k+1})^\top \mathbf{d}_{k+1} + \varepsilon \\
&\quad \leq -\nabla f(\mathbf{x}_k)^\top \mathbf{d}_{k+1} + \|\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\|_2 \underbrace{\|\mathbf{d}_{k+1}\|_2}_{=1} + \varepsilon \\
&\quad \leq \varepsilon \|\nabla f(\mathbf{x}_k)\|_2 + \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2}{\alpha_k} + \|\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\|_2 + \varepsilon.
\end{aligned}$$

Weil $\varepsilon > 0$ beliebig war, folgt hieraus schließlich

$$\lim_{k\to\infty} \big\|\mathbf{P}_{T_K(\mathbf{x}_{k+1})}\big(-\nabla f(\mathbf{x}_{k+1})\big)\big\|_2 \leq \lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2}{\alpha_k} + \lim_{k\to\infty} \|\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\|_2 = 0$$

wobei Satz 6.7 und die gleichmäßige Stetigkeit von ∇f zur Anwendung kommt. □

In der Regel folgt aus der Stetigkeit von ∇f nicht, dass auch $\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)$ stetig ist. Um sicherzustellen, dass die Iterierten des projizierten Gradientenverfahrens tatsächlich gegen einen stationären Punkt konvergieren, benötigen wir daher das folgende Resultat.

Satz 6.11 Die Funktion $f : \mathbb{R}^n \to \mathbb{R}$ sei auf K stetig differenzierbar. Dann folgt für jede Folge $\{\mathbf{x}_k\}_{k\in\mathbb{N}} \subset K$ mit $\mathbf{x}_k \to \mathbf{x}^\star \in K$

$$\big\|\mathbf{P}_{T_K(\mathbf{x}^\star)}\big(-\nabla f(\mathbf{x}^\star)\big)\big\|_2 \le \liminf_{k\to\infty} \big\|\mathbf{P}_{T_K(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big)\big\|_2.$$

Beweis. Aus Lemma 6.9 (*ii.*) folgt für jedes $\mathbf{y} \in K$

$$-\nabla f(\mathbf{x}_k)^\top(\mathbf{y}-\mathbf{x}_k) \le \big\|\mathbf{P}_{T_K(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big)\big\|_2 \|\mathbf{y}-\mathbf{x}_k\|_2,$$

woraus sich für $k \to \infty$ die Ungleichung

$$-\nabla f(\mathbf{x}^\star)^\top(\mathbf{y}-\mathbf{x}^\star) \le \liminf_{k\to\infty} \big\|\mathbf{P}_{T_K(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big)\big\|_2 \|\mathbf{y}-\mathbf{x}^\star\|_2$$

ergibt. Zu jedem $\mathbf{d} \in T_K(\mathbf{x}^\star)$ mit $\|\mathbf{d}\|_2 = 1$ lässt sich eine Folge $\{\mathbf{y}_k\}_{k\in\mathbb{N}}$ aus K derart finden, dass

$$\mathbf{d} = \lim_{k\to\infty} \frac{\mathbf{y}_k - \mathbf{x}^\star}{\|\mathbf{y}_k - \mathbf{x}^\star\|_2} \quad \text{und} \quad \lim_{k\to\infty} \mathbf{y}_k = \mathbf{x}^\star.$$

Somit erhalten wir

$$-\nabla f(\mathbf{x}^\star)^\top \mathbf{d} \le \liminf_{k\to\infty} \big\|\mathbf{P}_{T_K(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big)\big\|_2$$

und daraus wegen (6.11) die Behauptung:

$$\begin{aligned}\big\|\mathbf{P}_{T_K(\mathbf{x}^\star)}\big(-\nabla f(\mathbf{x}^\star)\big)\big\|_2 &= \max\{-\nabla f(\mathbf{x}^\star)^\top\mathbf{d} : \mathbf{d} \in T_K(\mathbf{x}^\star),\ \|\mathbf{d}\|_2 = 1\} \\ &\le \liminf_{k\to\infty} \big\|\mathbf{P}_{T_K(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big)\big\|_2.\end{aligned}$$

□

Bemerkung Die Kombination der Sätze 6.7, 6.10 und 6.11 liefert die folgende Aussage: Ist die zulässige Menge $K \subset \mathbb{R}^n$ konvex und abgeschlossen und ist die Funktion $f : \mathbb{R}^n \to \mathbb{R}$ auf K stetig differenzierbar mit gleichmäßig stetigem Gradienten und nach unten beschränkt, dann gilt für jeden Häufungspunkt $\mathbf{x}^\star \in K$ der Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ des projizierten Gradientenverfahrens 6.4

$$\mathbf{P}_{T_K(\mathbf{x}^\star)}\big(-\nabla f(\mathbf{x}^\star)\big) = \mathbf{0}.$$

Gemäß Lemma 6.9 (*iii.*) bedeutet dies, dass $\mathbf{x}^\star$ ein stationärer Punkt ist. △

6.2 Aktive-Mengen-Strategie

Wir betrachten das Optimierungsproblem

$$\text{minimiere } f(\mathbf{x}) \text{ unter den Nebenbedingungen } g_j(\mathbf{x}) \le 0 \text{ für alle } j = 1,\dots,m, \tag{6.13}$$

wobei $f : \mathbb{R}^n \to \mathbb{R}$ und $g_j : \mathbb{R}^n \to \mathbb{R}$ gegebene Funktionen seien. Die Idee der *Aktive-Mengen-Strategie* ist es, zu unterscheiden, ob eine Nebenbedingung im Optimum $\mathbf{x}^\star \in \mathbb{R}^n$

aktiv ist, es also $g_j(\mathbf{x}^\star) = 0$ gilt, oder ob sie inaktiv ist, also weggelassen werden kann, da $g_j(\mathbf{x}^\star) < 0$ gilt.
Die Aktive-Mengen-Startegie funktioniert nun wie folgt. Es sei $A_k \subset \{1, \dots, m\}$ die Menge der aktiven Nebenbedingungen im k-ten Iterationsschritt. Wir lösen das Optimierungsproblem

$$\text{minimiere } f(\mathbf{x}) \text{ unter der Nebenbedingung } g_j(\mathbf{x}) = 0 \text{ für alle } j \in A_k. \tag{6.14}$$

Es können im k-ten Iterationsschritt zwei Fälle auftreten:

- Das Optimum $\mathbf{x}_k^\star$ verletzt eine inaktive Nebenbedingung. Es ist also $g_j(\mathbf{x}^\star) > 0$ für ein $j \notin A_k$. In diesem Fall wird der entsprechende Index j der neuen aktiven Menge A_{k+1} hinzugefügt. Dies ist der *Aktivierungsschritt*.
- Der zur Nebenbedingung $g_j(\mathbf{x}^\star) = 0$ gehörige Lagrange-Parameter λ_j besitzt das falsche Vorzeichen. In diesem Fall wird der entsprechende Index j aus der aktiven Menge entfernt. Dies ist der *Inaktivierungsschritt*.

Man iteriert nun solange, bis $A_{k+1} = A_k$ ist und man folglich das Optimum $\mathbf{x}^\star = \mathbf{x}_k^\star$ von (6.13) gefunden hat. Der Unterschied zwischen verschiedenen Aktive-Mengen-Verfahren sind die Strategien der Aktivierung und Inaktivierung von Nebenbedingungen.

Algorithmus 6.12 (Aktive-Mengen-Strategie)
input: Funktion $f : \mathbb{R}^n \to \mathbb{R}$ und Nebenbedingungen $g_j : \mathbb{R}^n \to \mathbb{R}$, $j = 1, \dots, m$
output: Optimum $\mathbf{x}^\star \in \mathbb{R}^n$

① Initialisierung: $A_1 := \emptyset$ und $k := 1$
② berechne das Optimum $\mathbf{x}^\star \in \mathbb{R}^n$ des Problems

$$\text{minimiere } f(\mathbf{x}) \text{ unter der Nebenbedingung } g_j(\mathbf{x}) = 0 \text{ für alle } j \in A_k$$

③ bestimme die neue aktive Menge A_{k+1}
④ ist $A_{k+1} \neq A_k$, dann erhöhe $k := k + 1$ und gehe nach ②

Für das volldiskrete Optimierungsproblem (5.5) betrachten wir die folgende Aktive-Mengen-Strategie. Wir definieren die Mengen

$$A_k^a = \{i \in \{1, \dots, M\} : [\mathbf{u}]_i = [\mathbf{u}_a]_i\},$$
$$A_k^b = \{i \in \{1, \dots, M\} : [\mathbf{u}]_i = [\mathbf{u}_b]_i\},$$

der aktiven Indizes, für die die untere Schranke $\mathbf{u}_a$ beziehungsweise die obere Schranke $\mathbf{u}_b$ erzwungen werden soll. Wir wollen also die Funktion

$$f(\mathbf{u}_h) = \frac{1}{2}(\mathbf{y}_h - \mathbf{y}_d)^\top \mathbf{M}_h(\mathbf{y}_h - \mathbf{y}_d) + \frac{\lambda}{2}\mathbf{u}_h^\top \mathbf{N}_h \mathbf{u}_h \text{ mit } \mathbf{A}_h\mathbf{y}_h = \mathbf{B}_h\mathbf{u}_h \tag{6.15}$$

unter den Nebenbedingungen

$$\mathbf{I}_{A_k^a}\mathbf{u}_h = \mathbf{I}_{A_k^a}\mathbf{u}_a \text{ und } \mathbf{I}_{A_k^b}\mathbf{u}_h = \mathbf{I}_{A_k^b}\mathbf{u}_b \tag{6.16}$$

minimieren. Hierbei mögen $\mathbf{I}_{A_k^a} \in \mathbb{R}^{|A_k^a| \times M}$ und $\mathbf{I}_{A_k^b} \in \mathbb{R}^{|A_k^b| \times M}$ diejenigen Matrizen bezeichnen, welche aus der Einheitsmatrix des $\mathbb{R}^M$ durch das Streichen derjeniger Zeilen hervorgeht, deren Index *nicht* in der jeweiligen aktiven Menge enthalten ist.

Führen wir die Lagrange-Parameter $\boldsymbol{\mu}_a \in \mathbb{R}^{|A_k^a|}$ und $\boldsymbol{\mu}_b \in \mathbb{R}^{|A_k^b|}$ ein, so erhalten wir in Anbetracht von

$$\nabla f(\mathbf{u}_h) = \mathbf{B}_h^\top \mathbf{p}_h + \lambda \mathbf{N}_h \mathbf{u}_h \text{ mit } \mathbf{A}_h^\top \mathbf{p}_h = \mathbf{M}_h(\mathbf{y}_h - \mathbf{y}_d)$$

das diskrete Sattelpunktproblem

$$\begin{bmatrix} \mathbf{M}_h & -\mathbf{A}_h^\top & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\mathbf{A}_h & \mathbf{0} & \mathbf{B}_h & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_h^\top & \lambda \mathbf{N}_h & -\mathbf{I}_{A_k^a}^\top & \mathbf{I}_{A_k^b}^\top \\ \mathbf{0} & \mathbf{0} & -\mathbf{I}_{A_k^a} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{A_k^b} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{y}_h \\ \mathbf{p}_h \\ \mathbf{u}_h \\ \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix} = \begin{bmatrix} \mathbf{M}_h \mathbf{y}_d \\ \mathbf{0} \\ \mathbf{0} \\ -\mathbf{I}_{A_k^a} \mathbf{u}_a \\ \mathbf{I}_{A_k^b} \mathbf{u}_b. \end{bmatrix} \tag{6.17}$$

Die Systemmatrix ist symmetrisch, aber indefinit. Sie kann mit Hilfe von iterativen Lösern für Sattelpunktprobleme effizient gelöst werden.

Hat man (6.17) gelöst, so muss man nur die diskrete Steuerung $\mathbf{u}_h$ dahingehend überprüfen, ob sie die Box-Beschränkung $\mathbf{u}_a \leq \mathbf{u}_h \leq \mathbf{u}_b$ erfüllt. Dadurch werden die neuen aktiven Mengen A_{k+1}^a und A_{k+1}^b bestimmt. Man setzt $A_{k+1}^a = A_k^a$ und fügt alle Indizes $i \in \{1, \ldots, M\} \setminus A_k^a$ hinzu, für die $[\mathbf{u}_a]_i > [\mathbf{u}_h]_i$ gilt. Umgekehrt entfernt man alle Indizes aus A_{k+1}^a, für die die Bedingung $[\boldsymbol{\mu}_a]_i \geq 0$ verletzt ist. Analog wird mit der aktiven Menge A_{k+1}^b verfahren. Das Verfahren endet, falls $A_{k+1}^a = A_k^a$ und $A_{k+1}^b = A_k^b$ gilt. Diese Aktive-Mengen-Strategie wird auch primal-duale Aktive-Mengen-Strategie genannt, weil sowohl die primale Variable $\mathbf{u}_h$ als auch die dualen Variablen $\boldsymbol{\mu}_a$ und $\boldsymbol{\mu}_b$ in die Bestimmung der neuen aktiven Mengen eingehen.

Es sei noch angemerkt, dass man die Lagrange-Parameter $\boldsymbol{\mu}_a$ und $\boldsymbol{\mu}_b$ auch aus (6.17) eliminieren kann. Denn sind $\mathbf{P}_{A_k^a}, \mathbf{P}_{A_k^b} \in \mathbb{R}^{M \times M}$ diejenigen Projektionen, die durch

$$[\mathbf{P}_{A_k^a}\mathbf{v}]_i = \begin{cases} [\mathbf{v}]_i, & \text{falls } i \in A_k^a, \\ 0, & \text{falls } i \notin A_k^a, \end{cases} \qquad [\mathbf{P}_{A_k^b}\mathbf{v}]_i = \begin{cases} [\mathbf{v}]_i, & \text{falls } i \in A_k^b, \\ 0, & \text{falls } i \notin A_k^b, \end{cases}$$

gegeben sind, so gilt offenbar

$$\mathbf{P}_{A_k^a}\mathbf{u}_h = \mathbf{P}_{A_k^a}\mathbf{u}_a, \quad \mathbf{P}_{A_k^b}\mathbf{u}_h = \mathbf{P}_{A_k^b}\mathbf{u}_b.$$

Wir können damit $\mathbf{u}_h$ also schreiben gemäß

$$\begin{aligned} \mathbf{u}_h &= (\mathbf{I} - \mathbf{P}_{A_k^a} - \mathbf{P}_{A_k^b})\mathbf{u}_h + \mathbf{P}_{A_k^a}\mathbf{u}_h + \mathbf{P}_{A_k^a}\mathbf{u}_h \\ &= -(\lambda \mathbf{N}_h)^{-1}(\mathbf{I} - \mathbf{P}_{A_k^a} - \mathbf{P}_{A_k^b})\mathbf{B}_h^\top \mathbf{p}_h + \mathbf{P}_{A_k^a}\mathbf{u}_a + \mathbf{P}_{A_k^a}\mathbf{u}_a \end{aligned}$$

Dies eingesetzt in (6.17) führt unmittelbar auf

$$\begin{bmatrix} \mathbf{M}_h & -\mathbf{A}_h^\top & \mathbf{0} \\ -\mathbf{A}_h & \mathbf{0} & \mathbf{B}_h \\ \mathbf{0} & (\lambda \mathbf{N}_h)^{-1}(\mathbf{I} - \mathbf{P}_{A_k^a} - \mathbf{P}_{A_k^b})\mathbf{B}_h^\top & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{y}_h \\ \mathbf{p}_h \\ \mathbf{u}_h \end{bmatrix} = \begin{bmatrix} \mathbf{M}_h \mathbf{y}_d \\ \mathbf{0} \\ \mathbf{P}_{A_k^a}\mathbf{u}_a + \mathbf{P}_{A_k^b}\mathbf{u}_b \end{bmatrix}.$$

Dieses System ist deutlich kleiner als (6.17), allerdings ist es nicht mehr symmetrisch. Um über die Aktivierung und Inaktivierung der Indizes in der aktiven Menge zu entscheiden, muss man den Lagrange-Parameter durch

$$\boldsymbol{\mu}_a = \mathbf{I}_{A_k^a}\big(\mathbf{B}_h^\top \mathbf{p}_h + \lambda \mathbf{N}_h \mathbf{u}_a\big), \qquad \boldsymbol{\mu}_b = -\mathbf{I}_{A_k^b}\big(\mathbf{B}_h^\top \mathbf{p}_h + \lambda \mathbf{N}_h \mathbf{u}_b\big)$$

extrahieren und kann dann wie zuvor die aktive Mengen aufdatieren.

6.3 Halbglattes Newton-Verfahren

Wir starten nun direkt mit dem Optimalitätssystem für das diskrete Optimalsteuerproblem. Für das Optimum $(\mathbf{y}_h^\star, \mathbf{p}_h^\star, \mathbf{u}_h^\star) \in \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^M$ müssen Lagrange-Parameter $\boldsymbol{\mu}_a{}^\star, \boldsymbol{\mu}_b{}^\star \in \mathbb{R}^M$ derart existieren, dass

$$\begin{aligned}
\mathbf{M}_h\mathbf{y}_h^\star - \mathbf{A}_h^\top\mathbf{p}_h^\star &= \mathbf{M}_h\mathbf{y}_d,\\
-\mathbf{A}_h\mathbf{y}_h^\star + \mathbf{B}_h\mathbf{u}_h^\star &= \mathbf{0},\\
\mathbf{B}_h^\top\mathbf{p}_h^\star + \lambda\mathbf{N}_h\mathbf{u}_h^\star - \boldsymbol{\mu}_a^\star + \boldsymbol{\mu}_b^\star &= \mathbf{0},\\
\boldsymbol{\mu}_a^\star \geq \mathbf{0}, \quad \mathbf{u}_a - \mathbf{u}_h^\star \leq \mathbf{0}, \quad (\mathbf{u}_h^\star - \mathbf{u}_a)^\top\boldsymbol{\mu}_a^\star &= 0,\\
\boldsymbol{\mu}_b^\star \geq \mathbf{0}, \quad \mathbf{u}_h^\star - \mathbf{u}_b \leq \mathbf{0}, \quad (\mathbf{u}_b - \mathbf{u}_h^\star)^\top\boldsymbol{\mu}_b^\star &= 0,
\end{aligned}$$

vergleiche (2.15) beziehungsweise (4.15). Hierin können wir die Komplementaritätsbedingungen auch umschreiben gemäß

$$\left.\begin{aligned}
\boldsymbol{\mu}_a^\star &= \max\{\mathbf{0}, \boldsymbol{\mu}_a^\star + c(\mathbf{u}_a^\star - \mathbf{u}_h)\}\\
\boldsymbol{\mu}_b^\star &= \max\{\mathbf{0}, \boldsymbol{\mu}_b^\star + c(\mathbf{u}_h - \mathbf{u}_b^\star)\}
\end{aligned}\right\} \quad c > 0.$$

Folglich ist das Optimalitätssystem äquivalent zum Auffinden einer Nullstelle:

$$\mathbf{F}(\mathbf{y}_h, \mathbf{p}_h, \mathbf{u}_h, \boldsymbol{\mu}_a, \boldsymbol{\mu}_b) := \begin{bmatrix}
\mathbf{M}_h\mathbf{y}_h - \mathbf{A}_h^\top\mathbf{p}_h - \mathbf{M}_h\mathbf{y}_d\\
-\mathbf{A}_h\mathbf{y}_h + \mathbf{B}_h\mathbf{u}_h\\
\mathbf{B}_h^\top\mathbf{p}_h + \lambda\mathbf{N}_h\mathbf{u}_h - \boldsymbol{\mu}_a + \boldsymbol{\mu}_b\\
\boldsymbol{\mu}_a - \max\{\mathbf{0}, \boldsymbol{\mu}_a + c(\mathbf{u}_a - \mathbf{u}_h)\}\\
\boldsymbol{\mu}_b - \max\{\mathbf{0}, \boldsymbol{\mu}_b + c(\mathbf{u}_h - \mathbf{u}_b)\}
\end{bmatrix} \stackrel{!}{=} \mathbf{0}. \tag{6.18}$$

Definition 6.13 (Newton-Differenzierbarkeit) Eine Funktion $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ heißt **Newton-differenzierbar** im Punkt $\mathbf{x} \in \mathbb{R}^n$, wenn eine offene Umgebung $U_\mathbf{x} \subset \mathbb{R}^n$ von $\mathbf{x}$ und eine Funktion $\mathbf{G}_\mathbf{x} : U_\mathbf{x} \to \mathbb{R}^{m\times n}$ derart existieren, dass

$$\frac{\|\mathbf{F}(\mathbf{x}+\mathbf{h}) - \mathbf{F}(\mathbf{x}) - \mathbf{G}_\mathbf{x}(\mathbf{x}+\mathbf{h})\mathbf{h}\|}{\|\mathbf{h}\|} \to 0 \quad \text{für } \|\mathbf{h}\| \to 0.$$

Die Funktion $\mathbf{G}_\mathbf{x}$ wird **Newton-Ableitung** von $\mathbf{F}$ in $\mathbf{x}$ genannt. Ist die Funktion in jedem Punkt $\mathbf{x} \in \mathbb{R}^n$ Newton-differenzierbar und hängt die Newton-Ableitung nicht von $\mathbf{x}$ ab, so schreiben wir einfach nur $\mathbf{G}$.

Beispiel 6.14 Die Funktion $F : \mathbb{R} \to \mathbb{R}$, definiert durch $F(x) = \max\{x, 0\}$, ist in allen Punkten $x \neq 0$ Fréchet-differenzierbar und daher auch Newton-differenzierbar. Im Punkt $x = 0$ ist sie allerdings nur Newton-differenzierbar. Es gilt

$$G(x) = \begin{cases} 1, & \text{falls } x > 0,\\ \delta, & \text{falls } x = 0,\\ 0, & \text{falls } x < 0, \end{cases}$$

wobei $\delta \in \mathbb{R}$ beliebig ist. Folglich ist die Funktion $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$ mit $\mathbf{F}(\mathbf{x}) = \max\{\mathbf{x}, \mathbf{0}\}$

ebenfalls Newton-differenzierbar und es gilt

$$[\mathbf{G}(\mathbf{x})]_i = \begin{cases} 1, & \text{falls } [\mathbf{x}]_i > 0, \\ \delta, & \text{falls } [\mathbf{x}]_i = 0, \\ 0, & \text{falls } [\mathbf{x}]_i < 0, \end{cases} \qquad i = 1, \dots, n.$$

Im folgenden wählen wir $\delta = 0$. △

Mit Hilfe der Newton-Ableitung kann man ein *verallgemeinertes Newton-Verfahren* zur Lösung von $\mathbf{F}(\mathbf{x}^\star) = \mathbf{0}$ formulieren:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \big(\mathbf{G}(\mathbf{x}_k)\big)^{-1}\mathbf{F}(\mathbf{x}_k) \quad \text{bzw.} \quad \mathbf{G}(\mathbf{x}_k)\mathbf{x}_{k+1} = \mathbf{G}(\mathbf{x}_k)\mathbf{x}_k - \mathbf{F}(\mathbf{x}_k) \tag{6.19}$$

Das folgende Resultat sichert uns lokal superlineare Konvergenz dieses Verfahrens zu.

Satz 6.15 Sei $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$ gegeben mit $\mathbf{F}(\mathbf{x}^\star) = \mathbf{0}$, so dass $\mathbf{F}$ Newton-differenzierbar in $\mathbf{x}^\star \subset \mathbb{R}^n$ ist. Es existiere eine Umgebung $U_{\mathbf{x}^\star}$ von $\mathbf{x}^\star$ und eine Newton-Ableitung $\mathbf{G}$ mit auf $U_{\mathbf{x}^\star}$ gleichmäßig beschränkter Inverser:

$$\big\|\mathbf{G}^{-1}(\mathbf{x})\big\| \le M \quad \text{für alle } \mathbf{x} \in U_{\mathbf{x}^\star}.$$

Dann gibt es eine Umgebung $U \subset U_{\mathbf{x}^\star}$ von $\mathbf{x}^\star$, so dass die verallgemeinerte Newton-Iteration (6.19) mit Startwert $\mathbf{x}_0 \in U$ superlinear gegen $\mathbf{x}^\star$ konvergiert.

Beweis. Sei $R > 0$ so gewählt, dass $B_R(\mathbf{x}^\star) \subset U_{\mathbf{x}^\star}$. Aufgrund der Newton-Differenzierbarkeit existiert ein $0 < r \le R$ mit der Eigenschaft

$$\|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}^\star) - \mathbf{G}(\mathbf{x})(\mathbf{x} - \mathbf{x}^\star)\| \le \frac{1}{2M}\|\mathbf{x} - \mathbf{x}^\star\| \quad \text{für alle } \mathbf{x} \in B_r(\mathbf{x}^\star).$$

Es folgt

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{x}^\star\| &= \Big\|\mathbf{x}_k - \mathbf{x}^\star - \big(\mathbf{G}(\mathbf{x}_k)\big)^{-1}\mathbf{F}(\mathbf{x}_k)\Big\| \\ &= \Big\|\big(\mathbf{G}(\mathbf{x}_k)\big)^{-1}\big(\mathbf{G}(\mathbf{x}_k)(\mathbf{x}_k - \mathbf{x}^\star) - \mathbf{F}(\mathbf{x}_k) + \mathbf{F}(\mathbf{x}^\star)\big)\Big\| \\ &\le \Big\|\big(\mathbf{G}(\mathbf{x}_k)\big)^{-1}\Big\|\|\mathbf{G}(\mathbf{x}_k)(\mathbf{x}_k - \mathbf{x}^\star) - \mathbf{F}(\mathbf{x}_k) + \mathbf{F}(\mathbf{x}^\star)\| \\ &\le \frac{M}{2M}\|\mathbf{x}_k - \mathbf{x}^\star\|, \end{aligned}$$

dies bedeutet, die Iteration mit Startwert $\mathbf{x}_k \in B_r(\mathbf{x}^\star)$ konvergiert gegen $\mathbf{x}^\star$. Dieselbe Abschätzung zeigt insbesondere auch

$$\|\mathbf{x}_{k+1} - \mathbf{x}^\star\| = o(\|\mathbf{x}_k - \mathbf{x}^\star\|),$$

was die superlineare Konvergenz beweist. □

Wir wollen uns nun mit der Umsetzung des Newton-Verfahrens zur Lösung von (6.18) befassen. Dazu betrachten wir zunächst folgendes Beispiel.

Beispiel 6.16 Das Optimalitätssystem des Problems

$$\text{minimiere } \mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{x}^\top \mathbf{b} \text{ unter der Nebenbedingung } \mathbf{x} \leq \mathbf{d}$$

lautet

$$\mathbf{F}(\mathbf{x}, \boldsymbol{\mu}) := \begin{bmatrix} \mathbf{A}\mathbf{x} - \mathbf{b} + \boldsymbol{\mu} \\ \boldsymbol{\mu} - \max\{\mathbf{0}, \boldsymbol{\mu} + c(\mathbf{x} - \mathbf{d})\end{bmatrix} \stackrel{!}{=} \mathbf{0}.$$

Definieren wir für $(\mathbf{x}, \boldsymbol{\mu})$ die Menge

$$A = \{i \in \{1, \ldots, n\} : [\boldsymbol{\mu} + c(\mathbf{x} - \mathbf{d})]_i > 0\},$$

so lässt sich die Newton-Ableitung von $\mathbf{F}(\mathbf{x}, \boldsymbol{\mu})$ schreiben als

$$\mathbf{G}(\mathbf{x}, \boldsymbol{\mu}) = \begin{bmatrix} \mathbf{A} & \mathbf{I} \\ -c\,\mathbf{P}_A & \mathbf{I} - \mathbf{P}_A \end{bmatrix}.$$

Das Newton-Verfahren (6.19) lautet daher

$$\begin{bmatrix} \mathbf{A} & \mathbf{I} \\ -c\,\mathbf{P}_{A_k} & \mathbf{I} - \mathbf{P}_{A_k} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{k+1} \\ \boldsymbol{\mu}_{k+1} \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{I} \\ -c\,\mathbf{P}_{A_k} & \mathbf{I} - \mathbf{P}_{A_k} \end{bmatrix} \begin{bmatrix} \mathbf{x}_k \\ \boldsymbol{\mu}_k \end{bmatrix} - \begin{bmatrix} \mathbf{A}\mathbf{x}_k - \mathbf{b} + \boldsymbol{\mu}_k \\ -c\,\mathbf{P}_{A_k}(\mathbf{x}_k - \mathbf{d}) + (\mathbf{I} - \mathbf{P}_{A_k})\boldsymbol{\mu}_k \end{bmatrix}.$$

Die erste Zeile dieses Systems lautet kurz

$$\mathbf{A}\mathbf{x}_{k+1} - \mathbf{b} + \boldsymbol{\mu}_{k+1} = \mathbf{0}.$$

Für einen Index $i \in A_k$ folgt außerdem aus der zweiten Zeile dieses Systems die Gleichung

$$-c\,[\mathbf{x}_{k+1}]_i = -c\,[\mathbf{x}_k]_i + c\,[\mathbf{x}_k - \mathbf{d}]_i \quad \Longleftrightarrow \quad [\mathbf{x}_{k+1}]_i = [\mathbf{d}]_i,$$

während sich für einen Index $i \notin A_k$ die Gleichung

$$[\boldsymbol{\mu}_{k+1}]_i = [\boldsymbol{\mu}_k]_i - [\boldsymbol{\mu}_k]_i = 0$$

ergibt. Zusammengefasst ist der Newton-Schritt also gegeben durch das System

$$\begin{aligned} \mathbf{A}\mathbf{x}_{k+1} - \mathbf{b} + \boldsymbol{\mu}_{k+1} &= \mathbf{0}, \\ \mathbf{x}_{k+1} &= \mathbf{d} \quad \text{auf } A_k, \\ \boldsymbol{\mu}_{k+1} &= \mathbf{0} \quad \text{auf } A_k^c. \end{aligned}$$

Dieses ist aber äquivalent zur primal-dualen Aktive-Mengen-Strategie aus dem vorhergehenden Kapitel, wenn man beachtet, dass stets $[\boldsymbol{\mu}_k]_i = 0$ oder $[\mathbf{x}_k]_i = [\mathbf{d}]_i$ gilt, sofern man das Newton-Verfahren mit $\boldsymbol{\mu}_0 = \mathbf{0}$ startet. Denn daher stimmt die Menge A_k des verallgemeinerten Newton-Verfahrens in der Tat mit der aktiven Menge aus Kapitel 6.2 überein. △

Genauso wie in diesem Beispiel lässt sich zeigen, dass das verallgemeinerte Newton-Verfahren zur Lösung von (6.18) auf das Gleichungssystem (6.17) führt, wenn man die Nullen in den Lagrange-Parametern $\boldsymbol{\mu}_a$ und $\boldsymbol{\mu}_b$ einfach ignoriert. Daher können wir schliessen, dass die primal-duale Aktive-Mengen-Strategie superlinear konvergiert.

7. Formoptimierung

7.1 Bernoullis freies Randproblem

Gegeben seien ein einfach zusammenhängendes Gebiet $T \subset \mathbb{R}^d$ mit einem *freien Rand* $\partial T = \Gamma$ und ein einfach zusammenhängendes Gebiet $S \subset T$ mit einem *festem Rand* $\partial S = \Sigma$, vergleiche Abbildung 7.1. Wir setzen $\Omega = T \setminus \overline{S}$ und suchen für eine gegebene Konstante $g > 0$ nun dasjenige Gebiet Ω und die dazugehörige Funktion u, welche das Randwertproblem

$$\begin{aligned} \Delta u &= 0 && \text{in } \Omega, \\ u &= 1 && \text{auf } \Sigma, \\ -\frac{\partial u}{\partial \mathbf{n}} &= g,\ u = 0 && \text{auf } \Gamma. \end{aligned} \tag{7.1}$$

erfüllen.

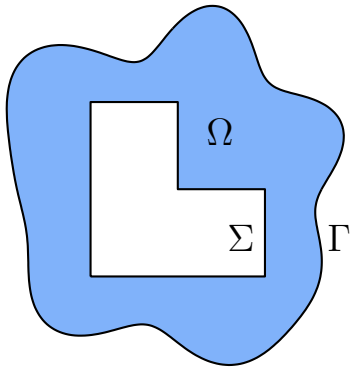

Abbildung 7.1: Das Gebiet Ω und seine Ränder Γ and Σ.

Das Problem (7.1) ist ein *freies Randproblem*, da der freie Rand Γ gesucht wird. Das nächste Lemma zeigt, dass es sinnvoll formuliert ist.

Lemma 7.1 Es sei $\Omega \subset \mathbb{R}^d$ ein Gebiet wie oben mit den Rändern Γ und Σ. Die Funktion $u \in C^1(\overline{\Omega})$ genüge der Laplace-Gleichung $\Delta u = 0$ auf Ω mit homogegenen Dirichlet-Randbedingungen auf Γ und positiven Dirichlet-Randbedingungen auf Σ. Dann gilt

$$0 \leq -\frac{\partial u}{\partial \mathbf{n}} = \|\nabla u\| \text{ auf } \Gamma.$$

Beweis. Wir führen den Beweis der Einfachheit halber in $d = 2$ Dimensionen durch und bemerken zunächst, dass die Dirichlet-Randbedingungen aufgrund des Maximumprinzips die Nichtnegativität von u implizieren: $u > 0$ in Ω. Deshalb gilt

$$0 \leq -\frac{\partial u}{\partial \mathbf{n}} \text{ auf } \Gamma.$$

Sei nun $\boldsymbol{\gamma} : [0,1] \to \Gamma$ eine Parametrisierung der Randkurve Γ. Die homogene Dirichlet-Ranbedingung bedeutet $u\big(\boldsymbol{\gamma}(s)\big) = 0$ für alle $t \in [0,1]$. Hieraus folgt

$$0 = \frac{\mathrm{d}}{\mathrm{d}s} u\big(\boldsymbol{\gamma}(s)\big) = \langle \nabla u\big(\boldsymbol{\gamma}(s)\big), \boldsymbol{\gamma}'(s)\rangle = \langle \nabla u\big(\boldsymbol{\gamma}(s)\big), \mathbf{t}(s)\rangle \|\boldsymbol{\gamma}'(s)\| = \frac{\partial u}{\partial \mathbf{t}}\big(\boldsymbol{\gamma}(s)\big)\|\boldsymbol{\gamma}'(s)\|,$$

dies bedeutet

$$0 = \frac{\partial u}{\partial \mathbf{t}} \text{ auf } \Gamma.$$

Aus der Identität

$$\nabla u = \langle \nabla u, \mathbf{t}\rangle \mathbf{t} + \langle \nabla u, \mathbf{n}\rangle \mathbf{n} = \frac{\partial u}{\partial \mathbf{t}}\mathbf{t} + \frac{\partial u}{\partial \mathbf{n}}\mathbf{n} = \frac{\partial u}{\partial \mathbf{n}}\mathbf{n} \text{ auf } \Gamma$$

ergibt sich das Behauptete:

$$\|\nabla u\| = \left|\frac{\partial u}{\partial \mathbf{n}}\right| \|\mathbf{n}\| = -\frac{\partial u}{\partial \mathbf{n}} \text{ auf } \Gamma.$$

□

Das freie Randproblem (7.1) ist nach Daniel Bernoulli (1700–1782) benannt, da entlang des freien Rands die Bedingung $\|\nabla u\| \equiv const.$ erfüllt ist. Wir werden sehen, dass sich die Lösung von *Bernoullis freiem Randproblem* (7.1) als Minimum des Dirichlet-Energiefunktionals

$$\begin{gathered} J(\Omega) = \int_\Omega \left\{\|\nabla u\|^2 + g^2\right\} \mathrm{d}\mathbf{x} \to \min \\ \text{unter den Nebenbedingung } \Delta u = 0 \text{ in } \Omega,\ u = 1 \text{ auf } \Sigma,\ u = 0 \text{ auf } \Gamma \end{gathered} \tag{7.2}$$

bestimmen lässt.

Das Gebiet $\Omega \subset \mathbb{R}^d$ spielt in (7.2) die Rolle der Steuerung und $u \in H_0^1(\Omega)$ ist der Zustand. Im Gegensatz zur Theorie in den vorangegangenen Kapiteln ist die Abhängigkeit des Zustandes von der Steuerung *nicht* mehr linear.

7.2 Formableitungen

Um das Funktional $J(\Omega)$ aus (7.2) zu minimieren, benötigen wir dessen *Formableitung*. Definieren wir für ein hinreichend glattes Verschiebungsfeld $\mathbf{V} : \overline{\Omega} \to \mathbb{R}^d$ das *gestörte Gebiet* Ω_ε gemäß

$$\Omega_\varepsilon := \{\mathbf{x} + \varepsilon \mathbf{V}(\mathbf{x}) : \mathbf{x} \in \Omega\}, \tag{7.3}$$

dann erhalten wir die Formableitung eines Funktionals $J(\Omega)$ wie üblich per Grenzwertbildung

$$\delta J(\Omega)[\mathbf{V}] := \lim_{\varepsilon \to 0} \frac{J(\Omega_\varepsilon) - J(\Omega)}{\varepsilon},$$

vergleiche Abbildung 7.3 für eine Illustration.

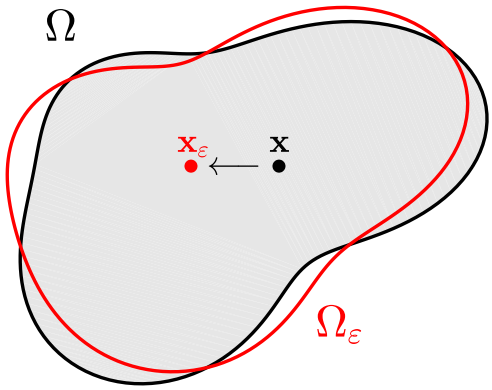

Abbildung 7.2: Das Gebiet Ω und das gestörte Gebiet Ω_ε.

Definition 7.2 (Formdifferenzierbarkeit) Es sei $\Omega \subset \mathbb{R}^d$ mit C^2-stetigem Rand und $\Omega_\varepsilon \subset \mathbb{R}^d$ sei wie in (7.3) definiert. Ein Funktional $J : \Omega \to \mathbb{R}$ heißt **formdifferenzierbar**, wenn der Grenzwert

$$\delta J(\Omega)[\mathbf{V}] := \lim_{\varepsilon \to 0} \frac{J(\Omega_\varepsilon) - J(\Omega)}{\varepsilon}$$

existiert für alle Richtungen $\mathbf{V} \in [C^2(\overline{\Omega})]^d$ und die Abbildung

$$\mathbf{V} \mapsto \delta J(\Omega)[\mathbf{V}]$$

linear und stetig ist. Wir nennen $\delta J(\Omega)[\mathbf{V}]$ die **Formableitung** des Funktionals $J(\Omega)$.

Für Volumenintegrale über das Gebiet Ω kann man die Formableitung leicht bestimmen.

Lemma 7.3 Die Formableitung des Funktionals

$$J(\Omega) = \int_\Omega f \, \mathrm{d}\mathbf{x}$$

lautet

$$\delta J(\Omega)[\mathbf{V}] = \int_{\partial\Omega} \langle \mathbf{V}, \mathbf{n} \rangle f \, \mathrm{d}\sigma,$$

vorausgesetzt es ist $f \in C^1(\mathbb{R}^d)$.

Beweis. Die Transformationsformel für Integrale und anschließende Taylor-Entwicklung liefern

$$\begin{aligned} \int_{\Omega_\varepsilon} f \, \mathrm{d}\mathbf{x} &= \int_\Omega f\big(\mathbf{x} + \varepsilon \mathbf{V}(\mathbf{x})\big) \det\big(\mathbf{I} + \varepsilon \mathbf{V}'(\mathbf{x})\big) \, \mathrm{d}\mathbf{x} \\ &= \int_\Omega \big\{ f(\mathbf{x}) + \varepsilon f'(\mathbf{x}) \mathbf{V}(\mathbf{x}) + o(\varepsilon) \big\} \big\{ 1 + \varepsilon \operatorname{tr}\big(\mathbf{V}'(\mathbf{x})\big) + \mathcal{O}(\varepsilon^2) \big\} \, \mathrm{d}\mathbf{x}. \end{aligned}$$

Hieraus folgt mit dem Gaußschen Integralsatz

$$\begin{aligned} J(\Omega_\varepsilon) - J(\Omega) &= \varepsilon \int_\Omega \big\{ f'(\mathbf{x})\mathbf{V}(\mathbf{x}) + f(\mathbf{x}) \operatorname{div}\big(\mathbf{V}(\mathbf{x})\big) \big\} \, \mathrm{d}\mathbf{x} + o(\varepsilon) \\ &= \varepsilon \int_\Omega \operatorname{div}(\mathbf{V} f) \, \mathrm{d}\mathbf{x} + o(\varepsilon) \\ &= \varepsilon \int_{\partial\Omega} \langle \mathbf{V}, \mathbf{n} \rangle f \, \mathrm{d}\sigma + o(\varepsilon), \end{aligned}$$

was dem Behaupteten entspricht. □

Die Lösung der Differentialgleichung hängt natürlich ebenfalls vom Gebiet ab. Daher müssen wir uns überlegen, wie man diese Ableitung bilden kann. Dazu betrachten wir die partielle Differentialgleichung

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ auf } \partial\Omega. \tag{7.4}$$

Für $\mathbf{V} \in [C^2(\overline{\Omega})]^d$ bezeichne Ω_ε das gestörtes Gebiet (7.3) und $u_\varepsilon \in H_0^1(\Omega_\varepsilon)$ die Lösung von

$$-\Delta u_\varepsilon = f \text{ in } \Omega_\varepsilon, \quad u_\varepsilon = 0 \text{ auf } \partial\Omega_\varepsilon. \tag{7.5}$$

Setzen wir

$$u^\varepsilon(\mathbf{x}) := u_\varepsilon\big(\mathbf{x} + \varepsilon \mathbf{V}(\mathbf{x})\big), \quad \mathbf{x} \in \Omega,$$

so ist $u^\varepsilon \in H_0^1(\Omega)$ und wir können u^ε mit der Lösung $u \in H_0^1(\Omega)$ von (7.4) vergleichen.

Definition 7.4 (Materialableitung) Der Grenzwert

$$\dot{u}(\mathbf{x}) = \lim_{\varepsilon \to 0} \frac{u^\varepsilon(\mathbf{x}) - u(\mathbf{x})}{\varepsilon}, \quad \mathbf{x} \in \Omega \tag{7.6}$$

heißt **Materialableitung** der Differentialgleichung (7.4).

Andererseits können wir aber auch für jedes $\mathbf{x} \in \Omega \cap \Omega_\varepsilon$ den Grenzwert

$$\delta u(\mathbf{x}) = \lim_{\varepsilon \to 0} \frac{u_\varepsilon(\mathbf{x}) - u(\mathbf{x})}{\varepsilon}$$

betrachten:

Definition 7.5 (lokale Formableitung) Der Grenzwert

$$\delta u(\mathbf{x}) = \lim_{\varepsilon \to 0} \frac{u_\varepsilon(\mathbf{x}) - u(\mathbf{x})}{\varepsilon}, \quad \mathbf{x} \in \Omega \cap \Omega_\varepsilon$$

heißt **lokale Formableitung** der Differentialgleichung (7.4).

Vorausgesetzt $\dot{u}$ und δu existieren beide, so folgt

$$\begin{aligned}\dot{u}(\mathbf{x}) &= \lim_{\varepsilon\to 0} \frac{u_\varepsilon\big(\mathbf{I}+\varepsilon\mathbf{V}(\mathbf{x})\big) - u(\mathbf{x})}{\varepsilon} \\ &= \lim_{\varepsilon\to 0} \frac{u_\varepsilon(\mathbf{x}) + \varepsilon\langle\nabla u_\varepsilon(\mathbf{x}), \mathbf{V}(\mathbf{x})\rangle - u(\mathbf{x})}{\varepsilon} \\ &= \lim_{\varepsilon\to 0} \frac{u_\varepsilon(\mathbf{x}) - u(\mathbf{x})}{\varepsilon} + \langle\nabla u(\mathbf{x}), \mathbf{V}(\mathbf{x})\rangle.\end{aligned}$$

Dies bedeutet:

$$\dot{u} = \delta u + \langle\nabla u, \mathbf{V}\rangle. \tag{7.7}$$

Lemma 7.6 Der Grenzwert in (7.6) konvergiert in der $H_0^1(\Omega)$-Norm gegen die Materialableitung $\dot{u} \in H_0^1(\Omega)$ zur Poisson-Gleichung (7.4). Diese genügt der Variationsformulierung

$$\int_\Omega \langle\nabla\dot{u}, \nabla v\rangle \,\mathrm{d}\mathbf{x} = \int_\Omega \operatorname{div}(f\,\mathbf{V}) v \,\mathrm{d}\mathbf{x} - \int_\Omega \big\langle \big[\operatorname{div}(\mathbf{V})\,\mathbf{I} - (\mathbf{V}')^\top - \mathbf{V}'\big]\nabla u, \nabla v\big\rangle \tag{7.8}$$

für alle $v \in H_0^1(\Omega)$ und damit der Differentialgleichung

$$-\Delta\dot{u} = \operatorname{div}(f\,\mathbf{V}) + \operatorname{div}\big(\big[\operatorname{div}(\mathbf{V})\,\mathbf{I} - (\mathbf{V}')^\top - \mathbf{V}'\big]\nabla u\big) \text{ in } \Omega, \quad \dot{u} = 0 \text{ auf } \partial\Omega.$$

Beweis. Wir vollziehen den Beweis in drei Schritten.
(*i.*) Es bezeichne u_ε wie zuvor die Lösung der Zustandsgleichung (7.5) auf dem gestörten Gebiet. Wir setzen

$$\mathbf{U}_\varepsilon(\mathbf{x}) := \mathbf{x} + \varepsilon\mathbf{V}(\mathbf{x})$$

und schließen aus der Beziehung $u^\varepsilon(\mathbf{x}) = u_\varepsilon\big(\mathbf{U}_\varepsilon(\mathbf{x})\big)$, dass

$$\nabla u_\varepsilon\big(\mathbf{U}_\varepsilon(\mathbf{x})\big) = \big(\mathbf{U}_\varepsilon'(\mathbf{x})\big)^{-\top}\nabla u^\varepsilon(\mathbf{x}), \quad \mathbf{x}\in\Omega.$$

Daher folgt die Variationsformulierung

$$\int_\Omega \langle\mathbf{A}_\varepsilon(\mathbf{x})\nabla u^\varepsilon(\mathbf{x}), \nabla v(\mathbf{x})\rangle \,\mathrm{d}\mathbf{x} = \int_\Omega f\big(\mathbf{U}_\varepsilon(\mathbf{x})\big)\det\big(\mathbf{U}_\varepsilon'(\mathbf{x})\big) v(\mathbf{x}) \,\mathrm{d}\mathbf{x} \quad \text{für alle } v\in H_0^1(\Omega),$$

wobei

$$\mathbf{A}_\varepsilon(\mathbf{x}) = \big(\mathbf{U}_\varepsilon'(\mathbf{x})^\top\mathbf{U}_\varepsilon'(\mathbf{x})\big)^{-1}\det\big(\mathbf{U}_\varepsilon'(\mathbf{x})\big).$$

Wir ziehen die Variationsformulierung von $u\in H_0^1(\Omega)$ ab, teilen durch ε und erhalten

$$\begin{aligned}&\int_\Omega \Big\langle \mathbf{A}_\varepsilon(\mathbf{x})\frac{\nabla\big(u^\varepsilon(\mathbf{x}) - u(\mathbf{x})\big)}{\varepsilon}, \nabla v(\mathbf{x})\Big\rangle \,\mathrm{d}\mathbf{x} \\ &\quad = \int_\Omega \frac{f\big(\mathbf{U}_\varepsilon(\mathbf{x})\big)\det\big(\mathbf{U}_\varepsilon'(\mathbf{x})\big) - f(\mathbf{x})}{\varepsilon} v(\mathbf{x}) \,\mathrm{d}\mathbf{x} - \int_\Omega \Big\langle \frac{\mathbf{A}_\varepsilon(\mathbf{x}) - \mathbf{I}}{\varepsilon}\nabla u(\mathbf{x}), \nabla v(\mathbf{x})\Big\rangle \,\mathrm{d}\mathbf{x}.\end{aligned} \tag{7.9}$$

Indem wir $v := u^\varepsilon - u$ als Testfunktion einsetzen, sehen wir, dass der Term $\nabla(u^\varepsilon - u)/\varepsilon$ in der $L^2(\Omega)$-Norm beschränkt ist, das heißt, $(u^\varepsilon - u)/\varepsilon$ ist in der $H_0^1(\Omega)$-Norm beschränkt. Folglich besitzt $(u^\varepsilon - u)/\varepsilon$ für $\varepsilon\to 0$ einen schwachen Grenzwert $\dot{u}\in H_0^1(\Omega)$.

(*ii.*) Mit Hilfe der Neumannschen Reihe ergibt sich

$$\begin{aligned}\mathbf{A}_\varepsilon(\mathbf{x}) &= \Big(\mathbf{I} - \varepsilon\Big[\big(\mathbf{V}'(\mathbf{x})\big)^\top + \mathbf{V}'(\mathbf{x})\Big] + \mathcal{O}(\varepsilon^2)\Big)\Big(1 + \varepsilon \operatorname{tr}\big(\mathbf{V}'(\mathbf{x})\big) + \mathcal{O}(\varepsilon^2)\Big)\\ &= \mathbf{I} + \varepsilon\Big[\operatorname{tr}\big(\mathbf{V}'(\mathbf{x})\big)\,\mathbf{I} - \big(\mathbf{V}'(\mathbf{x})\big)^\top - \mathbf{V}'(\mathbf{x})\Big] + \mathcal{O}(\varepsilon^2).\end{aligned}$$

Wir erinnern uns ferner, dass

$$f\big(\mathbf{U}_\varepsilon(\mathbf{x})\big)\det\big(\mathbf{U}_\varepsilon'(\mathbf{x})\big) = f(\mathbf{x}) + \varepsilon \operatorname{div}\big(f(\mathbf{x})\mathbf{V}(\mathbf{x})\big) + o(\varepsilon).$$

Zusammengefasst erhalten wir demnach aus (7.9) durch Grenzübergang $\varepsilon \to 0$ die gewünschte Variationsformulierung (7.8) für $\dot{u} \in H_0^1(\Omega)$.

(*iii.*) Wir müssen nun noch die starke Konvergenz von $(u^\varepsilon - u)/\varepsilon$ gegen $\dot{u}$ nachweisen. Dazu setzten wir $v_\varepsilon := (u^\varepsilon - u)/\varepsilon$ und erhalten

$$\begin{aligned}\int_\Omega \langle \mathbf{A}_\varepsilon \nabla v_\varepsilon, \nabla v_\varepsilon\rangle \,\mathrm{d}\mathbf{x} &= \int_\Omega \frac{(f\circ\mathbf{U}_\varepsilon)\det(\mathbf{U}_\varepsilon') - f}{\varepsilon} v_\varepsilon \,\mathrm{d}\mathbf{x} - \int_\Omega \Big\langle \frac{\mathbf{A}_\varepsilon - \mathbf{I}}{\varepsilon}\nabla u, \nabla v_\varepsilon\Big\rangle \,\mathrm{d}\mathbf{x}\\ &\xrightarrow{\varepsilon\to 0} \int_\Omega \operatorname{div}(f\,\mathbf{V})\dot{u}\,\mathrm{d}\mathbf{x} - \int_\Omega \big\langle \big[\operatorname{div}(\mathbf{V})\,\mathbf{I} - (\mathbf{V}')^\top - \mathbf{V}'\big]\nabla u, \nabla\dot{u}\big\rangle \,\mathrm{d}\mathbf{x}.\end{aligned}$$

Unter Beachtung von (7.8) bedeutet dies

$$\int_\Omega \langle \mathbf{A}_\varepsilon \nabla v_\varepsilon, \nabla v_\varepsilon\rangle \,\mathrm{d}\mathbf{x} \xrightarrow{\varepsilon\to 0} \int_\Omega \langle \nabla\dot{u}, \nabla\dot{u}\rangle \,\mathrm{d}\mathbf{x},$$

was die starke Konvergenz von $(u^\varepsilon - u)/\varepsilon$ gegen $\dot{u}$ in $H_0^1(\Omega)$ impliziert. □

Lemma 7.7 Die lokale Formableitung $\delta u \in H_0^1(\Omega)$ zur Poisson-Gleichung (7.4) genügt der Differentialgleichung

$$\Delta\delta u = 0 \text{ in } \Omega, \quad \delta u = -\langle \mathbf{V}, \mathbf{n}\rangle \frac{\partial u}{\partial \mathbf{n}} \text{ auf } \partial\Omega.$$

Beweis. Wir wollen die Gleichung bestimmen, die erfüllt ist für δu gemäß (7.7). Dazu sei $u \in H_0^1(\Omega)$ die Lösung der Poisson-Gleichung (7.4) und $v \in H_0^1(\Omega)$ beliebig. Setzen wir

$$\mathbf{b} := \langle \mathbf{V}, \nabla u\rangle\nabla v + \langle \mathbf{V}, \nabla v\rangle\nabla u - \langle \nabla u, \nabla v\rangle \mathbf{V},$$

so liefert Nachrechnen die Identität

$$-\big\langle \big[\operatorname{div}(\mathbf{V})\,\mathbf{I} - (\mathbf{V}')^\top - \mathbf{V}'\big]\nabla u, \nabla v\big\rangle = \operatorname{div}\mathbf{b} - \langle \mathbf{V}, \nabla u\rangle\Delta v - \langle \mathbf{V}, \nabla v\rangle\Delta u.$$

Weiter gilt aufgrund des Gaußschen Integralsatzes

$$\int_\Omega \operatorname{div}(f\,\mathbf{V})\, v \,\mathrm{d}\mathbf{x} = -\int_\Omega \langle \mathbf{V}, \nabla v\rangle f \,\mathrm{d}\mathbf{x} = \int_\Omega \langle \mathbf{V}, \nabla v\rangle \Delta u \,\mathrm{d}\mathbf{x}.$$

Für alle $v \in H_0^1(\Omega) \cap H^2(\Omega)$ erfüllt die Materialableitung daher

$$\begin{aligned}\int_\Omega \langle \nabla\dot{u}, \nabla v\rangle \,\mathrm{d}\mathbf{x} &= \int_\Omega \operatorname{div}(f\,\mathbf{V})\, v \,\mathrm{d}\mathbf{x} + \int_\Omega \big\{ \operatorname{div}\mathbf{b} - \langle \mathbf{V}, \nabla u\rangle\Delta v - \langle \mathbf{V}, \nabla v\rangle\Delta u\big\} \,\mathrm{d}\mathbf{x}\\ &= \int_{\partial\Omega} \langle \mathbf{b}, \mathbf{n}\rangle \,\mathrm{d}\sigma + \int_\Omega \big\langle \nabla\langle \mathbf{V}, \nabla u\rangle, \nabla v\big\rangle \,\mathrm{d}\mathbf{x} - \int_{\partial\Omega} \langle \mathbf{V}, \nabla u\rangle \frac{\partial v}{\partial \mathbf{n}} \,\mathrm{d}\sigma.\end{aligned}$$

Folglich gilt

$$\begin{aligned}\int_\Omega \langle \nabla(\dot{u} - \langle \mathbf{V}, \nabla u\rangle), \nabla v\rangle \,\mathrm{d}\mathbf{x} &= \int_{\partial\Omega} \langle \mathbf{b}, \mathbf{n}\rangle \,\mathrm{d}\sigma - \int_{\partial\Omega} \langle \mathbf{V}, \nabla u\rangle \frac{\partial v}{\partial \mathbf{n}} \,\mathrm{d}\sigma \\ &= \int_{\partial\Omega} \left\{ \langle \mathbf{V}, \nabla v\rangle \frac{\partial u}{\partial \mathbf{n}} - \langle \nabla u, \nabla v\rangle \langle \mathbf{V}, \mathbf{n}\rangle \right\} \mathrm{d}\sigma = 0,\end{aligned}$$

wobei wir im letzten Schritt

$$\begin{aligned}\langle \nabla u, \nabla v\rangle \langle \mathbf{V}, \mathbf{n}\rangle &= \langle \nabla u, \mathbf{n}\rangle \langle \mathbf{n}, \nabla v\rangle \langle \mathbf{V}, \mathbf{n}\rangle + \langle \nabla u, \mathbf{t}\rangle \underbrace{\langle \mathbf{t}, \nabla v\rangle}_{=0} \langle \mathbf{V}, \mathbf{n}\rangle \\ &= \frac{\partial u}{\partial \mathbf{n}} \langle \mathbf{n}, \nabla v\rangle \langle \mathbf{V}, \mathbf{n}\rangle + \frac{\partial u}{\partial \mathbf{n}} \underbrace{\langle \mathbf{t}, \nabla v\rangle}_{=0} \langle \mathbf{V}, \mathbf{t}\rangle \\ &= \frac{\partial u}{\partial \mathbf{n}} \langle \mathbf{V}, \nabla v\rangle\end{aligned}$$

benutzt haben. Dies bedeutet $\Delta \delta u = 0$ in Ω.
Die Randbedingungen von δu ergeben sich schließlich direkt aus der Einschränkung von (7.7) auf den Rand. Es gilt aufgrund der homogenen Randbedingung von u

$$0 = \delta u + \langle \nabla u, \mathbf{V}\rangle = \delta u + \langle \mathbf{V}, \mathbf{n}\rangle \frac{\partial u}{\partial \mathbf{n}} + \langle \mathbf{V}, \mathbf{t}\rangle \underbrace{\frac{\partial u}{\partial \mathbf{t}}}_{=0} \quad \text{auf } \partial\Omega.$$

□

Lemma 7.8 Vorgelegt sei die Poisson-Gleichung (7.4). Die Formableitung der Dirichlet-Energie

$$J(\Omega) = \int_\Omega \left\{ \|\nabla u\|^2 - 2fu \right\} \mathrm{d}\mathbf{x}$$

lautet dann

$$\delta J(\Omega)[\mathbf{V}] = -\int_{\partial\Omega} \langle \mathbf{V}, \mathbf{n}\rangle \left(\frac{\partial u}{\partial \mathbf{n}} \right)^2 \mathrm{d}\sigma.$$

Beweis. Wir führen den Beweis in drei Schritten.
(*i.*) Wir setzen

$$C_\varepsilon := \int_{\Omega_\varepsilon} \|\nabla u_\varepsilon\|^2 \,\mathrm{d}\mathbf{x} - \int_\Omega \|\nabla u\|^2 \,\mathrm{d}\mathbf{x}, \qquad D_\varepsilon := \int_{\Omega_\varepsilon} f u_\varepsilon \,\mathrm{d}\mathbf{x} - \int_\Omega fu \,\mathrm{d}\mathbf{x}$$

und zeigen

$$\frac{J(\Omega_\varepsilon) - J(\Omega)}{\varepsilon} = \frac{C_\varepsilon}{\varepsilon} - 2\frac{D_\varepsilon}{\varepsilon} \xrightarrow{\varepsilon \to 0} \int_{\partial\Omega} \langle \mathbf{V}, \mathbf{n}\rangle \left(\frac{\partial u}{\partial \mathbf{n}} \right)^2 \mathrm{d}\sigma - 2\int_{\partial\Omega} \langle \mathbf{V}, \mathbf{n}\rangle \left(\frac{\partial u}{\partial \mathbf{n}} \right)^2 \mathrm{d}\sigma,$$

was dem Behaupteten entspricht.

(*ii.*) Es gilt

$$\begin{aligned}\frac{C_\varepsilon}{\varepsilon} &= \frac{1}{\varepsilon}\Big\{ \int_\Omega \langle \mathbf{A}_\varepsilon \nabla u^\varepsilon, \nabla u^\varepsilon\rangle \,\mathrm{d}\mathbf{x} - \int_\Omega \langle \nabla u, \nabla u\rangle \,\mathrm{d}\mathbf{x}\Big\} \\ &= \int_\Omega \Big\langle \frac{\mathbf{A}_\varepsilon - \mathbf{I}}{\varepsilon}\nabla u^\varepsilon, \nabla u^\varepsilon\Big\rangle \,\mathrm{d}\mathbf{x} + \int_\Omega \Big\langle \nabla \frac{u^\varepsilon - u}{\varepsilon}, \nabla(u^\varepsilon + u)\Big\rangle \,\mathrm{d}\mathbf{x} \\ &\xrightarrow{\varepsilon\to 0} \int_\Omega \big\langle \big[\operatorname{div}(\mathbf{V})\,\mathbf{I} - (\mathbf{V}')^\top - \mathbf{V}'\big]\nabla u, \nabla u\big\rangle \,\mathrm{d}\mathbf{x} + 2\int_\Omega \langle \nabla \dot{u}, \nabla u\rangle \,\mathrm{d}\mathbf{x}.\end{aligned}$$

Wir benutzen (7.8) und erhalten

$$\frac{C_\varepsilon}{\varepsilon} \xrightarrow{\varepsilon\to 0} \int_\Omega \operatorname{div}(f\,\mathbf{V})\, u \,\mathrm{d}\mathbf{x} + \int_\Omega \langle \nabla \dot{u}, \nabla u\rangle \,\mathrm{d}\mathbf{x}.$$

Partielle Integration liefert die Identität

$$\begin{aligned}\int_\Omega \operatorname{div}(f\,\mathbf{V})\, u \,\mathrm{d}\mathbf{x} &= -\int_\Omega \langle \mathbf{V}, \nabla u\rangle f \,\mathrm{d}\mathbf{x} \\ &= \int_\Omega \langle \mathbf{V}, \nabla u\rangle \Delta u \,\mathrm{d}\mathbf{x} \\ &= \int_{\partial\Omega} \langle \mathbf{V}, \nabla u\rangle \frac{\partial u}{\partial \mathbf{n}} \,\mathrm{d}\sigma - \int_\Omega \big\langle \nabla\langle \mathbf{V}, \nabla u\rangle, \nabla u\big\rangle \,\mathrm{d}\mathbf{x}.\end{aligned}$$

Es folgt demnach

$$\frac{C_\varepsilon}{\varepsilon} \xrightarrow{\varepsilon\to 0} \int_{\partial\Omega} \underbrace{\langle \mathbf{V}, \nabla u\rangle}_{=\langle \mathbf{V},\mathbf{n}\rangle\langle \mathbf{n},\nabla u\rangle} \frac{\partial u}{\partial \mathbf{n}} \,\mathrm{d}\sigma + \underbrace{\int_\Omega \langle \nabla \delta u, \nabla u\rangle \,\mathrm{d}\mathbf{x}}_{=0} = \int_{\partial\Omega} \langle \mathbf{V}, \mathbf{n}\rangle \left(\frac{\partial u}{\partial \mathbf{n}}\right)^2 \mathrm{d}\sigma.$$

(*iii.*) Wir finden

$$\begin{aligned}\frac{D_\varepsilon}{\varepsilon} &= \frac{1}{\varepsilon}\int_\Omega \big\{ (f\circ \mathbf{U}_\varepsilon)\det(\mathbf{U}_\varepsilon')u^\varepsilon - fu\big\} \,\mathrm{d}\mathbf{x} \\ &= \int_\Omega \left\{ \frac{(f\circ \mathbf{U}_\varepsilon)\det(\mathbf{U}_\varepsilon') - f}{\varepsilon} u^\varepsilon + f\frac{u^\varepsilon - u}{\varepsilon}\right\} \mathrm{d}\mathbf{x} \\ &\xrightarrow{\varepsilon\to 0} \int_\Omega \big\{ \operatorname{div}(f\,\mathbf{V})\, u + f\dot{u}\big\} \,\mathrm{d}\mathbf{x}.\end{aligned}$$

Mit

$$\int_\Omega \operatorname{div}(f\,\mathbf{V})\, u \,\mathrm{d}\mathbf{x} = -\int_\Omega \langle \mathbf{V}, \nabla u\rangle f \,\mathrm{d}\mathbf{x}$$

folgt demnach

$$\begin{aligned}\frac{D_\varepsilon}{\varepsilon} &\xrightarrow{\varepsilon\to 0} \int_\Omega f(\dot{u} - \langle \mathbf{V}, \nabla u\rangle) \,\mathrm{d}\mathbf{x} \\ &= \int_\Omega \underbrace{f}_{=-\Delta u} \delta u \,\mathrm{d}\mathbf{x} \\ &= \underbrace{-\int_\Omega \Delta\delta u\, u \,\mathrm{d}\mathbf{x} + \int_{\partial\Omega} \frac{\partial \delta u}{\partial \mathbf{n}} u \,\mathrm{d}\sigma}_{=0} - \int_{\partial\Omega} \delta u \frac{\partial u}{\partial \mathbf{n}} \,\mathrm{d}\sigma \\ &= \int_{\partial\Omega} \langle \mathbf{V}, \mathbf{n}\rangle \left(\frac{\partial u}{\partial \mathbf{n}}\right)^2 \mathrm{d}\sigma.\end{aligned}$$

□

Die Kombination von Lemma 7.3 und Lemma 7.8 liefert schließlich den Formgradienten des Funktionals (7.2).

Satz 7.9 Es sei $\Omega \subset \mathbb{R}^d$ ein ringförmiges Gebiet mit innerem Rand Σ und äußerem Rand Γ. Die Formableitung des Funktionals

$$J(\Omega) = \int_\Omega \left\{ \|\nabla u\|^2 + g^2 \right\} \mathrm{d}\mathbf{x}$$
$$\text{unter der Nebenbedingung } \Delta u = 0 \text{ in } \Omega,\ u = 1 \text{ auf } \Sigma,\ u = 0 \text{ auf } \Gamma$$

lautet dann

$$\delta J(\Omega)[\mathbf{V}] = \int_\Gamma \langle \mathbf{V}, \mathbf{n} \rangle \left\{ g^2 - \left(\frac{\partial u}{\partial \mathbf{n}} \right)^2 \right\} \mathrm{d}\sigma \tag{7.10}$$

für alle Verschiebungsfelder $\mathbf{V} : \overline{\Omega} \to \mathbb{R}^d$, welche in einer Umgebung des inneren Randes Σ verschwinden.

Beweis. Es bezeichne $U \subset \mathbb{R}^d$ die Umgebung des inneren Randes Σ, auf der das Verschiebungsfeld jeweils verschwindet. Ferner sei $v \in H^1(\Omega)$ eine Funktion, welche der Wert 1 am inneren Rand Σ annimmt und außerhalb von U verschwindet. Demnach erfüllt v die Randbedingungen

$$v = 1 \text{ auf } \Sigma, \quad v = 0 \text{ auf } \Gamma.$$

Genügt $w \in H_0^1(\Omega)$ der Poisson-Gleichung

$$-\Delta w = \Delta v \text{ in } \Omega, \quad w = 0 \text{ auf } \Sigma, \quad w = 0 \text{ auf } \Gamma,$$

so gilt offensichtlich $u = v + w$ und es folgt

$$J(\Omega) = \int_\Omega \left\{ \|\nabla(v+w)\|^2 + g^2 \right\} \mathrm{d}\mathbf{x} = \int_\Omega \left\{ \|\nabla w\|^2 - 2\Delta v w + \|\nabla v\|^2 + g^2 \right\} \mathrm{d}\mathbf{x}.$$

Die Anwendung von Lemma 7.3 und Lemma 7.8 liefert dann wegen $\mathbf{V} = \mathbf{0}$ auf Σ und

$$\delta \left(\int_\Omega \|\nabla v\|^2 \, \mathrm{d}\mathbf{x} \right)[\mathbf{V}] = 0$$

die Formableitung

$$\delta J(\Omega)[\mathbf{V}] = \int_\Gamma \langle \mathbf{V}, \mathbf{n} \rangle \left\{ g^2 - \left(\frac{\partial w}{\partial \mathbf{n}} \right)^2 \right\} \mathrm{d}\sigma,$$

welche wegen $\partial v / \partial \mathbf{n} = 0$ auf Γ mit der gewünschten Formableitung (7.10) übereinstimmt. □

Es verbleibt, die Frage zu beantworten, warum die Lösung des Formoptimierungsproblems (7.2) mit der Lösung des freien Randproblems (7.1) übereinstimmt.

Satz 7.10 (notwendige Optimalitätsbedingung) In einem Minimum $\Omega^\star \subset \mathbb{R}^d$ von (7.2) gilt

$$g = -\frac{\partial u}{\partial \mathbf{n}} \text{ auf } \Gamma^\star.$$

Beweis. Aus der notwendigen Optimalitätsbedingung

$$\delta J(\Omega^\star)[\mathbf{V}] = \int_{\Gamma^\star} \langle \mathbf{V}, \mathbf{n} \rangle \left\{ g^2 - \left(\frac{\partial u}{\partial \mathbf{n}} \right)^2 \right\} \mathrm{d}\sigma = 0$$

für alle Verschiebungsfelder $\mathbf{V} \in [C^2(\overline{\Omega})]^d$ mit $\mathbf{V} = \mathbf{0}$ auf Σ folgt die Identität

$$g^2 = \left(\frac{\partial u}{\partial \mathbf{n}} \right)^2 \text{ auf } \Gamma^\star.$$

Da $\partial u / \partial \mathbf{n} \leq 0$ ist laut Lemma 7.1, entspricht dies dem Behaupteten. □

Bemerkung Die Lösung $\Omega^\star$ des freien Randproblems (7.1) entspricht wirklich einem lokalen Minimum des Funktionals $J(\Omega)$ aus (7.2). Der genaue Nachweis würde allerdings den hier gesetzten Rahmen sprengen, da man dazu den Formhessian untersuchen muss. △

7.3 Diskretisierung

Wir wollen nun die numerische Lösung des Formoptimierungsproblems (7.2) behandeln, wobei wir uns der Einfachheit halber auf die zweidimensionale Situation einschränken.

Zuerst überlegen wir uns, dass ein Gebiet Ω eindeutig durch Vorgabe seines freien Randes Γ beschrieben wird. Wir können also das Gebiet Ω mit einer Parametrisierung $\boldsymbol{\gamma} : [0,1] \to \Gamma$ identifizieren. Unter der Annahme, dass das gesuchte Gebiet Ω sternförmig bezüglich dem Ursprung ist, gibt es eine Funktion $r \in C^2_{\mathrm{per}}([0,1])$, so dass

$$\boldsymbol{\gamma} : [0,1] \to \Gamma, \quad \boldsymbol{\gamma}(s) = r(s) \begin{bmatrix} \cos(2\pi s) \\ \sin(2\pi s) \end{bmatrix}.$$

Man beachte, dass die radiale Funktion sogar eindeutig ist und somit das Gebiet Ω eindeutig mit einer Funktion $r \in C^2_{\mathrm{per}}([0,1])$ identifiziert werden kann.

Das Formfunktional hängt demzufolge nur noch von der Funktion r ab. Speziell sehen wir, dass die Formableitung (7.10) nur vom Verschiebungsfeld $\mathbf{V}$ auf dem Rand abhängt. Aufgrund dieser Tatsache, ist es ausreichend, das Verschiebungsfeld nur am Rand zu definieren $\mathbf{V} : \Gamma \to \mathbb{R}^2$. Daher müssen wir nur Variationen $q \in C^2_{\mathrm{per}}([0,1])$ bezüglich der Funktion r betrachten:

$$\mathbf{V}\big(\boldsymbol{\gamma}(s)\big) := q(s) \begin{bmatrix} \cos(2\pi s) \\ \sin(2\pi s) \end{bmatrix}.$$

Das gestörte Gebiet wäre damit eindeutig durch die Parametrisierung

$$\boldsymbol{\gamma}_\varepsilon : [0,1] \to \Gamma_\varepsilon, \quad \boldsymbol{\gamma}_\varepsilon(s) = \big(r(s) + \varepsilon q(s)\big) \begin{bmatrix} \cos(2\pi s) \\ \sin(2\pi s) \end{bmatrix}.$$

gegeben. Der Formgradient (7.10) kann in Polarkoordinaten ausgedrückt werden:

$$\delta J(r)[q] = 2\pi \int_0^1 q(s) \left\langle \begin{bmatrix} \cos(2\pi s) \\ \sin(2\pi s) \end{bmatrix}, \mathbf{n}\big(\boldsymbol{\gamma}(s)\big) \right\rangle \left\{ g^2 - \left(\frac{\partial u}{\partial \mathbf{n}}\big(\boldsymbol{\gamma}(s)\big) \right)^2 \right\} \sqrt{\big(r(s)\big)^2 + \big(r'(s)\big)^2}\, \mathrm{d}s.$$

Die Aufgabe ist es nun, diejenige radiale Funktion $r^\star \in C^2_{\text{per}}([0,1])$ zu finden, so dass $\delta J(r^\star)[q] = 0$ gilt für alle $q \in C^2_{\text{per}}([0,1])$. Diese Aufgabenstellung diskretisieren wir dadurch, dass wir einen endlichdimensionalen Ansatzraum

$$U_h = \text{span}\{\phi_1, \phi_2, \ldots, \phi_m\} \subset C^2_{\text{per}}([0,1])$$

einführen und die nichtlineare Gleichung $\delta J(r^\star)[q] = 0$ bezüglich diesem Raum lösen:

$$\text{suche } r_h^\star \in U_h, \text{ so dass } \delta J(r_h^\star)[q_h] = 0 \text{ für alle } q_h \in U_h. \tag{7.11}$$

Ein geeigneter Ansatzraum U_h besteht etwa aus $[0,1]$-periodischen kubischen Splines oder aus trigonometrischen Polynomen festen Grades.

Die nichtlineare Gleichung (7.11) kann iterativ mit dem Gradientenverfahren gelöst werden. Ein besserer Löser ist das Quasi-Newton-Verfahren aus Kapitel 8. Bei der numerischen Lösung muss man aber beachten, dass für jede Iterierte das Gebiet Ω ein anderes ist und man daher zur Berechnung des Zustands mit Hilfe der Finite-Elemente-Methode jeweils eine neue Triangulierung erzeugen muss. Letzteres kann effizient mit Hilfe einer Netzabbildungsmethode geschehen, zumindest solange die Netzdeformation nicht zu gross ist. Am einfachsten ist diese mit *Coons-Patches* umzusetzen.

Wird das Coons-Patch P berandet durch die vier Randkurven $\kappa_i : [0,1] \to \partial P$, so definiert die *Coons-Abbildung* $C(s,t) : [0,1]^2 \to P$ mit

$$C(s,t) = \kappa_1(s)\big(1-\phi(t)\big) + \kappa_2(t)\big(1-\phi(s)\big) + \kappa_3(s)\phi(t) + \kappa_4(t)\phi(s)$$

eine Abbildung vom Einheitsquadrat auf des Coons-Patch P. Hierbei ist $\phi : [0,1] \to [0,1]$ eine *Blending-Funktion*, das ist eine stetige und streng monoton steigende Funktion mit $\phi(0) = 0$ und $\phi(1) = 1$. Die einfachste Blending-Funktion ist linear, das heißt $\phi(s) = s$, aber auch Splinekurven oder eine Sinus-Funktion sind üblich.

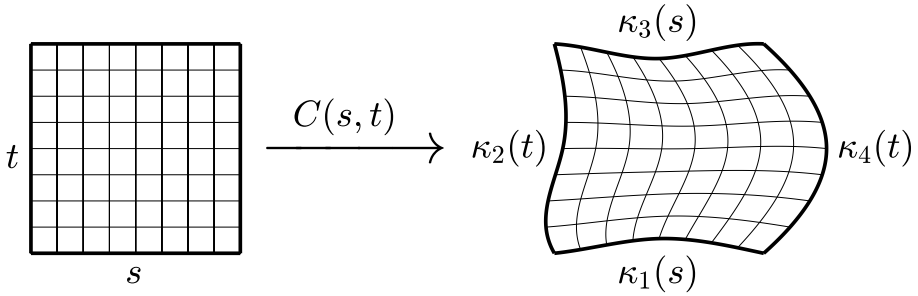

Abbildung 7.3: Illustration eines Coons-Patches.

Eine reguläre Triangulierung des Einheitsquadrats erzeugt nämlich eine reguläre Triangulierung des Coons-Patches, solange die Coons-Abbildung bijektiv bleibt. Um eine Triangulierung des gesamten Gebietes Ω zu erhalten, zerlegt man dieses regulär in Coons-Patches. Wenn zur Darstellung der Kanten zwischen den Coons-Patches jeweils dieselbe Parametrisierung benutzt wird, dann führt diese Konstruktion zum Schluss auf eine reguläre Triangulierung des Gebiets Ω, vergleiche Abbildung 7.4.

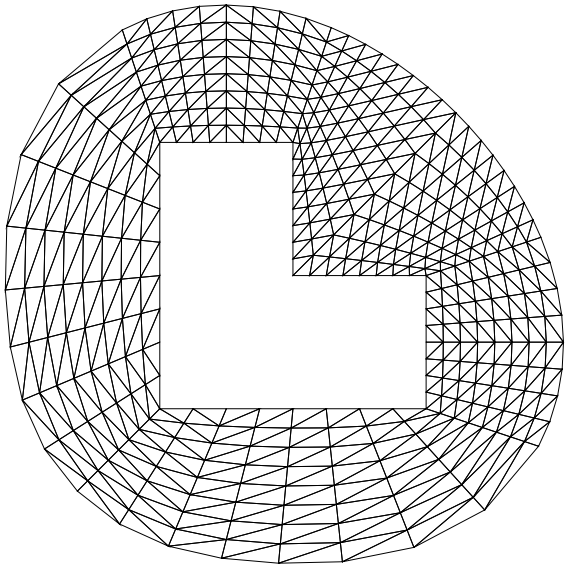

Abbildung 7.4: Reguläre Triangulierung für das freies Randproblem von Bernoulli im Fall eines L-Gebiets als innerer Rand. Die Triangulierung wird durch sechs Coons-Patches erzeugt.

8. Quasi-Newton-Verfahren*

Beim Newton-Verfahren ist das Update $\mathbf{d}_k$ durch die Newton-Gleichung $\nabla^2 f(\mathbf{x}_k)\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$ gegeben. Da das Berechnen der Hesse-Matrix und das Lösen dieses Gleichungssystems oftmals zu teuer ist, versucht man, $\big(\nabla^2 f(\mathbf{x}_k)\big)^{-1}$ durch einfach zu berechnende Matrizen $\mathbf{H}_k$ zu ersetzen und die Suchrichtung

$$\mathbf{d}_k := -\mathbf{H}_k \nabla f(\mathbf{x}_k)$$

zu benutzen. Man spricht von einem *Quasi-Newton-Verfahren*, wenn für alle $k \geq 0$ die Matrix $\mathbf{H}_{k+1}$ der *Quasi-Newton-Gleichung*

$$\mathbf{H}_{k+1}\big\{\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\big\} = \mathbf{x}_{k+1} - \mathbf{x}_k \tag{8.1}$$

genügt. Diese Bedingung stellt sicher, dass sich $\mathbf{H}_{k+1}$ in der Richtung $\mathbf{x}_{k+1} - \mathbf{x}_k$ ähnlich wie die Newton-Matrix $\big(\nabla^2 f(\mathbf{x}_k)\big)^{-1}$ verhält, für die gilt

$$\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k) = \nabla^2 f(\mathbf{x}_k)(\mathbf{x}_{k+1} - \mathbf{x}_k) + \mathcal{O}(\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2).$$

Für eine quadratische Funktion $q(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$ mit positiv definiter Matrix $\mathbf{A}$ gilt (8.1) wegen $\nabla q(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ sogar exakt. Ferner erscheint es sinnvoll, als $\mathbf{H}_k$ nur positiv definite Matrizen zu wählen. Dies garantiert, dass für $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$ die Richtung $\mathbf{d}_k = -\mathbf{H}_k \nabla f(\mathbf{x}_k)$ eine Abstiegsrichtung von f wird

$$\nabla f(\mathbf{x}_k)^\top \mathbf{d}_k = -\nabla f(\mathbf{x}_k)^\top \mathbf{H}_k \nabla f(\mathbf{x}_k) < 0.$$

Beide Forderungen lassen sich erfüllen: Mit den Abkürzungen

$$\mathbf{p}_k = \mathbf{x}_{k+1} - \mathbf{x}_k, \quad \mathbf{q}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$$

und frei wählbaren Parametern

$$\gamma_k > 0, \quad \nu_k \geq 0$$

ist $\mathbf{H}_{k+1}$ rekursiv gegeben durch

$$\begin{aligned} \mathbf{H}_{k+1} &:= \Phi(\mathbf{H}_k, \mathbf{p}_k, \mathbf{q}_k, \gamma_k, \nu_k), \\ \Phi(\mathbf{H}, \mathbf{p}, \mathbf{q}, \gamma, \nu) &:= \gamma \mathbf{H} + \left(1 + \gamma\nu \frac{\mathbf{q}^\top \mathbf{H}\mathbf{q}}{\mathbf{p}^\top \mathbf{q}}\right) \frac{\mathbf{p}\mathbf{p}^\top}{\mathbf{p}^\top \mathbf{q}} - \gamma \frac{1-\nu}{\mathbf{q}^\top \mathbf{H}\mathbf{q}} \mathbf{H}\mathbf{q}\mathbf{q}^\top \mathbf{H} \\ &\quad - \frac{\gamma\nu}{\mathbf{p}^\top \mathbf{q}} (\mathbf{p}\mathbf{q}^\top \mathbf{H} + \mathbf{H}\mathbf{q}\mathbf{p}^\top). \end{aligned} \tag{8.2}$$

Die Update-Funktion Φ ist nur für $\mathbf{p}^\top \mathbf{q} \neq 0$ und $\mathbf{q}^\top \mathbf{H}\mathbf{q} \neq 0$ erklärt. Man beachte, dass man $\mathbf{H}_{k+1}$ aus $\mathbf{H}_k$ dadurch erhält, dass man zur Matrix $\gamma_k \mathbf{H}_k$ eine Korrekturmatrix vom Rang ≤ 2 addiert:

$$\operatorname{rang}(\mathbf{H}_{k+1} - \gamma_k \mathbf{H}_k) \leq 2.$$

Man nennt dieses Verfahren daher auch *Rang-2-Verfahren*.

Folgende Spezialfälle sind in (8.2) enthalten:

1. $\gamma_k \equiv 1$, $\nu_k \equiv 0$: Verfahren von Davidon, Fletcher und Powell (*DFP-Verfahren*).
2. $\gamma_k \equiv 1$, $\nu_k \equiv 1$: Rang-2-Verfahren von Broydon, Fletcher, Goldfarb und Shanno (*BFGS-Verfahren*).
3. $\gamma_k \equiv 1$, $\nu_k = \mathbf{p}_k^\top \mathbf{q}_k / (\mathbf{p}_k^\top \mathbf{q}_k - \mathbf{p}_k^\top \mathbf{H}_k \mathbf{q}_k)$: *symmetrisches Rang-1-Verfahren von Broydon.*

Letzteres Verfahren ist nur für $\mathbf{p}_k^\top \mathbf{q}_k \neq \mathbf{p}_k^\top \mathbf{H}_k \mathbf{q}_k$ definiert; $\nu_k < 0$ ist möglich: in diesem Fall kann $\mathbf{H}_{k+1}$ auch indefinit werden, auch wenn $\mathbf{H}_k$ positiv definit ist (vergleiche Satz 8.2). Setzt man den gewählten Wert in (8.2) ein, erhält man für $\mathbf{H}_k$ eine Rekursionformel, die den Namen Rang-1-Verfahren erklärt:

$$\mathbf{H}_{k+1} := \mathbf{H}_k + \frac{\mathbf{z}_k \mathbf{z}_k^\top}{\alpha_k}, \quad \mathbf{z}_k := \mathbf{p}_k - \mathbf{H}_k \mathbf{q}_k, \quad \alpha_k := \mathbf{p}_k^\top \mathbf{q}_k - \mathbf{q}_k^\top \mathbf{H}_k \mathbf{q}_k.$$

Algorithmus 8.1 (Quasi-Newton-Verfahren)
input: Funktion $f : \mathbb{R}^n \to \mathbb{R}$ und Startnäherung $\mathbf{x}_0 \in \mathbb{R}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$

① Initialisierung: setze $\mathbf{H}_0 := \mathbf{I}$ und $k := 0$
② berechne die Quasi-Newton-Richtung $\mathbf{d}_k = -\mathbf{H}_k \nabla f(\mathbf{x}_k)$
③ löse
$$\alpha_k \approx \operatorname{argmin}_{\alpha \in \mathbb{R}} f(\mathbf{x}_k + \alpha \mathbf{d}_k)$$
④ setze $\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k \mathbf{d}_k$, $\mathbf{p}_k := \mathbf{x}_{k+1} - \mathbf{x}_k$ und $\mathbf{q}_k := \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$
⑤ wähle $\gamma_k > 0$, $\nu_k \geq 0$ und berechne $\mathbf{H}_{k+1} := \Phi(\mathbf{H}_k, \mathbf{p}_k, \mathbf{q}_k, \gamma_k, \nu_k)$ gemäß (8.2)
⑥ erhöhe $k := k + 1$ und gehe nach ②

Das Verfahren ist eindeutig durch die Wahl der Parameter γ_k, ν_k und die Minimierung in Schritt ③ fixiert. Die Minimierung $\mathbf{x}_k \mapsto \mathbf{x}_{k+1}$ und ihre Qualität kann man mit Hilfe eines Parameters σ_k beschreiben, der durch

$$\nabla f(\mathbf{x}_{k+1})^\top \mathbf{d}_k = \sigma_k \nabla f(\mathbf{x}_k)^\top \mathbf{d}_k = -\sigma_k \nabla f(\mathbf{x}_k)^\top \mathbf{H}_k \nabla f(\mathbf{x}_k)$$

definiert ist. Falls $\mathbf{d}_k$ eine Abstiegsrichtung ist, das heißt $\nabla f(\mathbf{x}_k)^\top \mathbf{d}_k < 0$, dann ist σ_k eindeutig bestimmt. Bei exakter Liniensuche ist $\sigma_k = 0$ wegen

$$\nabla f(\mathbf{x}_{k+1})^\top \mathbf{d}_k = \varphi_k'(\alpha_k) = 0, \quad \text{wobei} \quad \varphi_k(\alpha) := f(\mathbf{x}_k + \alpha \mathbf{d}_k).$$

Wir setzen für das folgende

$$\sigma_k < 1 \tag{8.3}$$

voraus. Falls $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$ und $\mathbf{H}_k$ positiv definit ist, folgt aus (8.3) $\alpha_k > 0$ und deshalb

$$\begin{aligned} \mathbf{q}_k^\top \mathbf{p}_k &= \alpha_k \{\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\}^\top \mathbf{d}_k \\ &= \alpha_k (\sigma_k - 1) \nabla f(\mathbf{x}_k)^\top \mathbf{d}_k \\ &= -\alpha_k (\sigma_k - 1) \nabla f(\mathbf{x}_k)^\top \mathbf{H}_k \nabla f(\mathbf{x}_k) \\ &> 0, \end{aligned}$$

also auch $\mathbf{q}_k \neq \mathbf{0}$ und $\mathbf{q}_k^\top \mathbf{H}_k \mathbf{q}_k > 0$. Die Matrix $\mathbf{H}_{k+1}$ ist damit durch (8.2) wohldefiniert.

Die Forderung (8.3) kann nur dann nicht erfüllt werden, wenn

$$\varphi_k'(\alpha) = \nabla f(\mathbf{x}_k + \alpha \mathbf{d}_k)^\top \mathbf{d}_k \leq \nabla f(\mathbf{x}_k)^\top \mathbf{d}_k = \varphi_k'(0) < 0$$

für alle $\alpha \geq 0$ gilt. Dann ist aber

$$f(\mathbf{x}_k + \alpha \mathbf{d}_k) - f(\mathbf{x}_k) = \int_0^\alpha \varphi_k'(t)\,\mathrm{d}t \leq \alpha \nabla f(\mathbf{x}_k)^\top \mathbf{d}_k < 0 \quad \text{für alle } \alpha \geq 0,$$

so dass $f(\mathbf{x}_k + \alpha \mathbf{d}_k)$ für $\alpha \to \infty$ nicht nach unten beschränkt ist. Die Forderung (8.3) bedeutet also keine wesentliche Einschränkung. Damit ist bereits der erste Teil des folgenden Satzes gezeigt, der besagt, dass das Quasi-Newton-Verfahren 8.1 unsere oben aufgestellten Forderungen erfüllt.

Satz 8.2 Falls im Quasi-Newton-Verfahren 8.1 die Matrix $\mathbf{H}_k$ für ein $k \geq 0$ positiv definit ist, $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$ und $\sigma_k < 1$ ist, dann ist für alle $\gamma_k > 0$, $\nu_k \geq 0$ die Matrix $\mathbf{H}_{k+1} := \Phi(\mathbf{H}_k, \mathbf{p}_k, \mathbf{q}_k, \gamma_k, \nu_k)$ wohldefiniert und wieder positiv definit. Insbesondere erfüllt sie die Quasi-Newton-Gleichung

$$\mathbf{H}_{k+1}\mathbf{q}_k = \mathbf{p}_k.$$

Beweis. Die Wohldefiniertheit von $\mathbf{H}_{k+1}$ haben wir bereits gezeigt, so dass wir nur noch die positive Definitheit nachweisen müssen. Sei $\mathbf{y} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ ein beliebiger Vektor und $\mathbf{H}_k = \mathbf{L}L^\top$ die Cholesky-Zerlegung von $\mathbf{H}_k$. Mit Hilfe der Vektoren

$$\mathbf{u} := \mathbf{L}^\top \mathbf{y}, \quad \mathbf{v} := \mathbf{L}^\top \mathbf{q}_k$$

lässt sich $\mathbf{y}^\top \mathbf{H}_{k+1}\mathbf{y}$ wegen (8.2) so schreiben:

$$\begin{aligned}
\mathbf{y}^\top \mathbf{H}_{k+1}\mathbf{y} &= \gamma_k \mathbf{u}^\top \mathbf{u} + \left(1 + \gamma_k \nu_k \frac{\mathbf{v}^\top \mathbf{v}}{\mathbf{p}_k^\top \mathbf{q}_k}\right) \frac{(\mathbf{p}_k^\top \mathbf{y})^2}{\mathbf{p}_k^\top \mathbf{q}_k} - \gamma_k \frac{1-\nu_k}{\mathbf{v}^\top \mathbf{v}} (\mathbf{u}^\top \mathbf{v})^2 - \frac{2\gamma_k \nu_k}{\mathbf{p}_k^\top \mathbf{q}_k} (\mathbf{p}_k^\top \mathbf{y})(\mathbf{u}^\top \mathbf{v}) \\
&= \gamma_k \left(\mathbf{u}^\top \mathbf{u} - \frac{(\mathbf{u}^\top \mathbf{v})^2}{\mathbf{v}^\top \mathbf{v}}\right) + \frac{(\mathbf{p}_k^\top \mathbf{y})^2}{\mathbf{p}_k^\top \mathbf{q}_k} + \gamma_k \nu_k \left(\sqrt{\mathbf{v}^\top \mathbf{v}} \frac{\mathbf{p}_k^\top \mathbf{y}}{\mathbf{p}_k^\top \mathbf{q}_k} - \frac{\mathbf{u}^\top \mathbf{v}}{\sqrt{\mathbf{v}^\top \mathbf{v}}}\right)^2 \\
&\geq \gamma_k \left(\mathbf{u}^\top \mathbf{u} - \frac{(\mathbf{u}^\top \mathbf{v})^2}{\mathbf{v}^\top \mathbf{v}}\right) + \frac{(\mathbf{p}_k^\top \mathbf{y})^2}{\mathbf{p}_k^\top \mathbf{q}_k}.
\end{aligned}$$

Die Cauchy-Schwarzsche Ungleichung ergibt

$$\mathbf{u}^\top \mathbf{u} - \frac{(\mathbf{u}^\top \mathbf{v})^2}{\mathbf{v}^\top \mathbf{v}} \geq 0,$$

mit Gleichheit genau dann, wenn $\mathbf{u} = \lambda \mathbf{v}$ für ein $\lambda \neq 0$ (wegen $\mathbf{y} \neq \mathbf{0}$). Für $\mathbf{u} \neq \lambda \mathbf{v}$ ist also $\mathbf{y}\mathbf{H}_{k+1}\mathbf{y} > 0$. Für $\mathbf{u} = \lambda \mathbf{v}$ folgt aus der Nichtsingularität von $\mathbf{H}_k$ und $\mathbf{L}$ auch $\mathbf{0} \neq \mathbf{y} = \lambda \mathbf{q}_k$, so dass

$$\mathbf{y}\mathbf{H}_{k+1}\mathbf{y} \geq \frac{(\mathbf{p}_k^\top \mathbf{y})^2}{\mathbf{p}_k^\top \mathbf{q}_k} = \lambda^2 \mathbf{p}_k^\top \mathbf{q}_k > 0.$$

Da $\mathbf{y} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ beliebig war, muss $\mathbf{H}_{k+1}$ positiv definit sein.

Die Quasi-Newton-Gleichung $\mathbf{H}_{k+1}\mathbf{q}_k = \mathbf{p}_k$ verifiziert man schließlich sofort mittels (8.2). □

Ein wesentliches Resultat ist, dass das Quasi-Newton-Verfahren im Fall einer quadratischen Funktion $f : \mathbb{R}^n \to \mathbb{R}$ das Minimum nach höchstens n Schritten liefert, sofern die Minimierung in ③ exakt sind. Da sich jede genügend oft differenzierbare Funktion f in der Nähe ihres Minimums beliebig genau durch eine quadratische Funktion approximieren lässt, lässt diese Eigenschaft vermuten, dass das Verfahren auch bei der Anwendung auf nichtquadratische Funktionen rasch konvergiert.

Satz 8.3 Sei

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$$

eine quadratische Funktion mit einer positiv definiten Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$. Wendet man das Quasi-Newton-Verfahren 8.1 zur Minimierung von f mit den Startwerten $\mathbf{x}_0$ und $\mathbf{H}_0$ an, wobei man die Minimierungen in ③ exakt durchführt, so liefert das Verfahren Folgen $\{\mathbf{x}_k\}_{k\geq 0}$, $\{\mathbf{H}_k\}_{k\geq 0}$, $\{\nabla f(\mathbf{x}_k)\}_{k\geq 0}$, $\{\mathbf{p}_k\}_{k\geq 0}$ und $\{\mathbf{q}_k\}_{k\geq 0}$ mit den Eigenschaften:

(*i.*) Es gibt ein kleinstes $m \leq n$ mit $\mathbf{x}_m = \mathbf{x}^\star = -\mathbf{A}^{-1}\mathbf{b}$, das heißt, $\mathbf{x}_m$ ist das eindeutige Minimum von f, insbesondere gilt also $\nabla f(\mathbf{x}_m) = \mathbf{0}$.

(*ii.*) Es ist $\mathbf{p}_k^\top \mathbf{q}_k > 0$ und $\mathbf{p}_k^\top \mathbf{q}_\ell = \mathbf{p}_k^\top \mathbf{A} p_\ell = 0$ für alle $0 \leq k \neq \ell < m$. Die Vektoren $\mathbf{p}_k$ sind demnach $\mathbf{A}$-*konjugiert*.

(*iii.*) Es gilt $\mathbf{p}_k^\top \nabla f(\mathbf{x}_\ell) = 0$ für alle $0 \leq k < \ell \leq m$.

(*iv.*) Es ist $\mathbf{H}_\ell \mathbf{q}_k = \lambda_{k,\ell}\mathbf{p}_k$ für alle $0 \leq k < \ell \leq m$ mit

$$\gamma_{k,\ell} := \begin{cases} \gamma_k\gamma_{k+1}\cdots\gamma_{\ell-1}, & \text{für } k < \ell - 1, \\ 1, & \text{für } k = \ell - 1. \end{cases}$$

(*v.*) Falls $m = n$, so gilt zusätzlich

$$\mathbf{H}_m = \mathbf{H}_n = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}\mathbf{A}^{-1},$$

wobei

$$\mathbf{D} = \operatorname{diag}(\gamma_{0,n}, \gamma_{1,n}, \ldots, \gamma_{n-1,n}), \quad \mathbf{P} = [\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_{n-1}].$$

Für $\gamma_k \equiv 1$ folgt $\mathbf{H}_n = \mathbf{A}^{-1}$.

Beweis. Wir zeigen zunächst induktiv, dass die Bedigungen (*ii.*)–(*iv.*) für ein beliebiges $m \geq 0$ gelten, falls $\mathbf{H}_j$ positiv definit und $\nabla f(\mathbf{x}_j) \neq \mathbf{0}$ für alle $j < m$ ist. Da die Aussagen für $m = 0$ trivialerweise erfüllt sind, können wir annehmen, dass sie für ein beliebiges $m \geq 0$ gelten. Der Induktionsschritt $m \mapsto m + 1$ ergibt sich nun wie folgt.
Da $\mathbf{H}_m$ positiv definit ist, folgt aus $\nabla f(\mathbf{x}_m) \neq \mathbf{0}$ sofort $\mathbf{d}_m = -\mathbf{H}_m \nabla f(\mathbf{x}_m) \neq \mathbf{0}$ und $\nabla f(\mathbf{x}_m)^\top \mathbf{H}_m \nabla f(\mathbf{x}_m) > 0$. Weil exakt minimiert wird, ist α_m die Nullstelle von

$$0 = \nabla f(\mathbf{x}_{m+1})^\top \mathbf{d}_m = \left\{\nabla f(\mathbf{x}_m) + \alpha_m \mathbf{A}\mathbf{d}_m\right\}^\top \mathbf{d}_m, \quad \alpha_m = \frac{\nabla f(\mathbf{x}_m)\mathbf{H}_m \nabla f(\mathbf{x}_m)}{\mathbf{d}^\top \mathbf{A}\mathbf{d}_m},$$

also $\mathbf{p}_m = \alpha_m \mathbf{d}_m$ und

$$\nabla f(\mathbf{x}_{m+1})^\top \mathbf{p}_m = \alpha_m \nabla f(\mathbf{x}_{m+1})^\top \mathbf{d}_m = 0. \tag{8.4}$$

Deshalb gilt

$$\begin{aligned}\mathbf{p}_m^\top \mathbf{q}_m &= \alpha_m \mathbf{d}_m^\top \big\{\nabla f(\mathbf{x}_{m+1}) - \nabla f(\mathbf{x}_m)\big\} \\ &= -\alpha_m \mathbf{d}_m^\top \nabla f(\mathbf{x}_m) \\ &= \alpha_m \nabla f(\mathbf{x}_m)^\top \mathbf{H}_m \nabla f(\mathbf{x}_m) \\ &> 0\end{aligned}$$

und folglich ist $\mathbf{H}_{m+1}$ nach Satz 8.2 positiv definit. Weiter ist für $k < m$ wegen $\mathbf{A}p_k = \mathbf{q}_k$

$$\mathbf{p}_k^\top \mathbf{q}_m = \mathbf{p}_k^\top \mathbf{A} p_m = \mathbf{q}_k^\top \mathbf{p}_m = -\alpha_m \mathbf{q}_k^\top \mathbf{H}_m \nabla f(\mathbf{x}_m) \overset{(iv.)}{=} -\alpha_m \gamma_{k,m} \mathbf{p}_k^\top \nabla f(\mathbf{x}_m) \overset{(iii.)}{=} 0. \quad (8.5)$$

Das ist der Induktionsschritt für Aussage (*ii.*).
Weiter gilt für $k < m$

$$\mathbf{p}_k^\top \nabla f(\mathbf{x}_{m+1}) = \mathbf{p}_k^\top \left(\nabla f(\mathbf{x}_{k+1}) + \sum_{j=k+1}^{m} \mathbf{q}_j \right) = 0$$

nach dem eben bewiesenen und Aussage (*iii.*). Zusammen mit (8.4) ergibt dies Aussage (*iii.*) für $m+1$.
Den Induktionsschritt für Aussage (*iv.*) sieht man wie folgt ein. Anhand von (8.2) verifiziert man sofort

$$\mathbf{H}_{m+1} \mathbf{q}_m = \mathbf{p}_m.$$

Wegen Aussage (*ii.*) für $m+1$ und der Induktionsvoraussetzung hat man ferner für $k < m$

$$\mathbf{p}_m^\top \mathbf{q}_k \overset{(ii.)}{=} \mathbf{0}, \quad \mathbf{q}_m^\top \mathbf{H}_m \mathbf{q}_k \overset{(iv.)}{=} \gamma_{k,m} \mathbf{q}_m^\top \mathbf{p}_m \overset{(ii.)}{=} \mathbf{0},$$

so dass für $k < m$ aus (8.2) folgt

$$\mathbf{H}_{m+1} \mathbf{q}_k = \gamma_m \mathbf{H}_m \mathbf{q}_k \overset{(iv.)}{=} \gamma_m \gamma_{k,m} \mathbf{p}_k = \gamma_{k,m+1} \mathbf{p}_k.$$

Der restliche Beweis ist nun einfach. Die Aussagen (*ii.*)–(*iv.*) können nur für $m \leq n$ richtig sein, da die Vektoren $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_{m-1}$ linear unabhängig sind. Aus $\mathbf{0} = \sum_{\ell=0}^{m-1} \lambda_\ell \mathbf{p}_\ell$ folgt nämlich durch Multiplikation mit $\mathbf{p}_k^\top \mathbf{A}$, $k = 0, 1, \ldots, m-1$, wegen Aussage (*ii.*) $\lambda_k \mathbf{p}_k^\top \mathbf{A} \mathbf{p}_k = 0$, das heißt, $\lambda_k = 0$.
Da wir bewiesen haben, dass die Aussagen (*ii.*)–(*iv.*) für beliebiges m gelten, solange $\nabla f(\mathbf{x}_m) \neq \mathbf{0}$ ist, muss es also einen ersten Index $m \leq n$ geben mit

$$\nabla f(\mathbf{x}_m) = \mathbf{0}, \quad \mathbf{x}_m = -\mathbf{A}^{-1}\mathbf{b},$$

dies bedeutet, es gilt Aussage (*i.*).
Für den Fall $m = n$ gilt wegen Aussage (*iv.*) zusätzlich $\mathbf{H}_n \mathbf{Q} = \mathbf{P}\mathbf{D}$ für die Matrizen

$$\mathbf{D} = \operatorname{diag}(\gamma_{0,n}, \gamma_{1,n}, \ldots, \gamma_{n-1,n}), \quad \mathbf{P} = [\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_{n-1}], \quad \mathbf{Q} = [\mathbf{q}_0, \mathbf{q}_1, \ldots, \mathbf{q}_{n-1}].$$

Wegen $\mathbf{A}\mathbf{P} = \mathbf{Q}$ ergibt sich schließlich wegen der Nichtsingularität der Matrix $\mathbf{P}$ die Beziehung

$$\mathbf{H}_n = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}\mathbf{A}^{-1},$$

Damit ist der Satz vollständig bewiesen. □

Es stellt sich nun die Frage, wie man die Parameter γ_k und ν_k wählen soll, um ein möglichst gutes Verfahren zu erhalten. Aussage $(v.)$ aus Satz 8.3 legt die Wahl $\gamma_k \equiv 1$ nahe, weil dies $\mathbf{D} = \mathbf{I}$ und folglich $\lim_m \mathbf{H}_m = \left(\nabla^2 f(\mathbf{x}^\star)\right)^{-1}$ vermuten lässt, weshalb das Verfahren voraussichtlich ähnlich schnell wie ein Newton-Verfahren konvergiert. Im allgemeinen ist diese Vermutung für nichtquadratische Funktionen aber nur unter zusätzlichen Voraussetzungen richtig. Nach praktischen Erfahrungen ist die Wahl

$$\gamma_k \equiv 1, \quad \nu_k \equiv 1 \quad \text{(BFGS-Verfahren)}$$

am besten.

Bemerkungen

1. Sowohl das DFP-Verfahren als auch das BFGS-Verfahren konvergieren superlinear in der Umgebung eines lokalen Minimus $\mathbf{x}^\star$, falls $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar ist und die Hesse-Matrix in der Umgebung von $\mathbf{x}^\star$ Lipschitz-stetig ist.
2. Eine andere Startmatrix $\mathbf{H}_0 \neq \mathbf{I}$ ist denkbar, solange sie symmetrisch und positiv definit ist.
3. In der Praxis macht man gelegentlich Restarts, setzt also $\mathbf{H}_k := \mathbf{H}_0$, falls $k \in m\mathbb{Z}$ mit festem $m \in \mathbb{N}$, beispielsweise $m = 100$.
4. Gerade bei großen Optimierungsproblemen stellt man die Matrix $\mathbf{H}_k$ nicht direkt auf, sondern berechnet sie rekursiv aus den Vektoren $\{(\gamma_k, \nu_k, \mathbf{p}_k, \mathbf{q}_k)\}_{k\geq 0}$. Damit auch bei vielen Schritten der Speicherplatz nicht überhand nimmt, speichert man nur die höchstens letzten m Vektoren. Man erlaubt also ein "Gedächtnis" von m Updates und ersetzt die unbekannte Matrix $\mathbf{H}_{k-m}$ durch $\mathbf{H}_0$. Man spricht von einem *Limited-Memory-Quasi-Newton-Verfahren*.

△

Index