JOHANNES KEPLER
UNIVERSITÄT LINZ | JKU

TNF

Technisch-Naturwissenschaftliche
Fakultät

Effiziente Löser für optimales Kontrollproblem für die instationären Stokes Gleichungen im zeitharmonischen Fall

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplomingenieur

im Diplomstudium

Industriemathematik

Eingereicht von:
Krendl Wolfgang

Angefertigt am:
Institut für Numerische Mathematik

Beurteilung:
A. Univ.-Prof. Dipl.-Ing. Dr. Walter Zulehner

Linz, Oktober, 2011

Zusammenfassung

In dieser vorliegenden Diplomarbeit untersuchen wir optimale Kontrollprobleme für die Stokes Gleichungen mit unbeschränkter Kontrolle im zeitabhängigen und zeitharmonischen Fall.

In beiden Fällen, wobei für die zeitliche Diskretisierung im zeitabhängigen Fall eine unstetige Galerkin Methode angewendet wurde, entspricht das diskretisierte Optimalitätssystem der Lösung von sehr großen Systemen linearer Gleichungen in Sattelpunktform. Diese linearen Systeme hängen empfindlich von dem Modellparameter α ab, welcher auch als Regularisierungs- oder Kostenparameter gesehen werden kann.

Für den zeitharmonischen Fall, studieren wir basierend auf bekannten und erst vor kurzem gezeigten Resultaten aus der Theorie der gemischten Variationsprobleme, die Konstruktion eines Präkonditioniers in Blockform, welcher robuste Konvergenzraten bezüglich der Modellparameter und der Feinheit der Diskretisierung für das MINRES (Minimal Residual)-Verfahren zu Folge hat.

Schließlich bestätigen wir die theoretischen Ergebnisse an Hand von numerischen Beispielen.

Abstract

In this master thesis we study optimal control problems for the Stokes equations with distributed control in the time-dependent and time-harmonic case.

In both cases, where for the time discretization in the time-dependent case we use a discontinuous Galerkin method, the discretized optimality system leads to a very large system of linear equations in saddlepoint form. This linear system depends in a sensitive way on a model parameter α, which can be viewed as a regularization or cost parameter.

For the time-harmonic case, we will study, based on recent results from the theory for mixed variational problems, the construction of a block preconditioner, which results in robust convergence rates for the Minimal Residual (MINRES) Method, with respect to the involved model parameters and the meshsize.

Finally some numerical tests will illustrate the theoretical results.

Danksagung

Ich möchte mich an dieser Stelle bei all jenen bedanken, die mir beim Schreiben dieser Arbeit und während meiner gesamten Studienzeit zur Seite gestanden sind.

Besonderer Dank gebührt dabei Professor Walter Zulehner, der sich ohne zögern als mein Betreuer zur Verfügung gestellt hat und sich immer Zeit für meine zahlreichen Fragen genommen hat. Mit großem Stolz erfüllt mich die Tatsache, mit dieser Diplomarbeit an einer brandneuen Erkenntnis im Gebiet der robusten Präkonditionierung mitgewirkt zu haben.

Ein großes Dankeschön gilt auch meinen Freunden, meiner Familie und meiner Freundin, die mir stets das vollste Vertrauen entgegen brachten und mich in jeder Hinsicht unterstützten.

Wolfgang Krendl
Linz, Oktober 2011

Inhaltsverzeichnis

Kapitel 1

Einführung

Die Steuerung von Strömungen spielt in der Mathematik und in den Ingenieurwissenschaften eine wichtige Rolle.

Das große Interesse lässt sich aus den zahlreichen Anwendungsbeispielen folgern: Durch optimale Formgebung der Herzklappen lassen sich zum Beispiel die Strömungsverhältnisse in einem künstlichen Herzen verbessern. Weiters kann durch Strömungskontrolle die optimale Form einer Schiffsturbine im Bezug auf den Abtrieb ermittelt werden.

Betrachtet man speziell die Steuerung von stark viskosen instationären Flüssigkeiten, so lässt sich dieses Problem mathematisch mit Hilfe eines optimalen Kontrollproblems für die instationären Stokes Gleichungen modelieren.

Die Berechnung einer Näherungslösung bedarf dabei der Lösung eines sehr großen linearen Systems, welches auf empfindlichem Wege von dem im Modell involvierten Parametern und der bei der Diskretisierung gewählten Feinheit abhängt.

Ziel ist nun ein Lösungsverfahren zu konstruieren dessen Konvergenzrate unabhängig von den Modellparametern und der Feinheit beschränkt ist. In diesem Zusammenhang spricht man auch von einem robusten Löser.

Im Rahmen dieser Arbeit werden wir zeigen, wie sich für den zeitharmonischen Fall ein solch robuster Löser konstruieren lässt.

1.1 Inhalt dieser Arbeit

In **Kapitel 2** wird die Problemstellung und die entsprechende mathematische Formulierung als optimales Kontrollproblem erläutert.

Eine Analyse des Problems bezüglich der Existenz, Eindeutigkeit und Charakterisierung einer Lösung wird dann in **Kapitel 3** vorgenommen.

Für die nummerische Berechnung einer Näherungslösung, wird das optimale Kontrollproblem in **Kapitel 4** zunächst diskretisiert. Die Ortsdiskretisierung erfolgt mit einer stetigen Galerkin-Methode. Motiviert bei einer Kurzeinführung verwenden wir anschließend für die zeitliche Diskretisierung ein Beispiel einer unstetigen Galerkin-Methode. Das entsprechende räumlich und zeitlich diskretisierte Optimalitätssystem lässt sich dann in Form eines linearen Gleichungssystems mit dünnbesetzter, regulärer, symmetrischer und indefiniter Systemmatrix in Sattelpunktsform schreiben.

In **Kapitel 5** widmen wir uns dann dem zeitharmonischen Fall: Unter Verwendung der für das zeitabhängige Problem gezeigten Resultate, führen wir auch für diesen Fall beginnend bei der Analyse bis zur Diskretisierung alle Schritte durch. Im Gegensatz zum zeitabhängigen Problem zerfällt hier das diskretisierte Optimalitätssystem für die verschiedenen Frequenzen in ein entkoppeltes System. Die Lösung der einzelnen Teilsysteme bedarf aber wiederum der Lösung eines Gleichungssystems mit großer, dünnbesetzter, regulärer, symmetrischer und indefiniter Systemmatrix in Sattelpunktsform.

Als mögliches Verfahren zur Lösung von symmetrisch und indefiniter Systeme, stellen wir in **Kapitel 6** das MINRES-Verfahren und seine präkonditinonierte Variante vor.

Das **Kernresultat** dieser Arbeit folgt dann in **Kapitel 7**: Wir werden zeigen, dass sich für den zeitharmonischen Fall ein bezüglich der im Modell auftretenden Parameter und der Schrittweite robuster Block-Präkonditionier konstruieren lässt. Unter anderem wird im Nachweis dieses Resultats, eine höchst aktuelle Aussage über die **Verbesserung der Abschätzungen** vom *Satz von Brezzi* gezeigt und angewendet. Als Folgerung der robusten Präkonditonierung erhält man darüber hinaus die Robustheit der benötigten Iterationszahlen für die MINRES-Methode.

In **Kapitel 8** verifizieren wir das gezeigte Resultat noch an Hand von nummerischen Beispielen für den zweidimensionalen Fall.

Abschließend geben wir in **Kapitel 9** eine grobe Zusammenfassung des Gezeigten wieder, verweisen auf, mit der Implementierung verbundene Probleme und geben weiterführende Ziele.

Kapitel 2

Problemstellung und Problemformulierung

Das folgende Kapitel basierend auf der Literatur *„Optimale Steuerung partieller Differentialgleichungen [7], Kapitel 1 - 3 “*, widmet sich der Analyse des Problems.

2.1 Problemstellung

Wir betrachten über den Zeitintervall $I = (0, T)$ eine inkompressible (Dichte ρ ist konstant) und hoch viskose (Reynoldszahl Re $\ll 1$) Flüssigkeit in einem gegebenen Gebiet $\Omega \subset \mathbb{R}^d$.

Solche Flüssigkeiten lassen sich durch die *instationären Stokes Gleichungen* beschreiben,

$$\begin{aligned} \rho\frac{\partial v}{\partial t} - \mu\Delta_x v + \nabla_x p = \rho f \quad &\text{in } \Omega \times (0,T), \\ \operatorname{div}_x v = 0 \quad &\text{in } \Omega \times (0,T), \end{aligned} \tag{2.1}$$

wobei für eine vollständige Formulierung, es der Angabe von Rand- und Anfangsbedingung

$$\begin{aligned} v = g_\Gamma \quad &\text{auf } \Sigma = \Gamma \times (0,T), \\ v(\cdot, 0) = v_0 \quad &\text{in } \Omega, \end{aligned}$$

bedarf, welche in unserem Fall der Einfachheit mit

$$g_\Gamma = 0 \;\text{ und }\; v_0 = 0.$$

gewählt werden.

In (2.1) bezeichnet $v(x) \in \mathbb{R}^d$ den Geschwindigkeitsvektor des im Ortspunkt $x \in \Omega$ befindlichen Partikels, p den Druck, f die Dichte der im Gebiet wirkenden Volumenkraft und die Konstante μ den Reibwert der Flüssigkeit.

Ziel ist es nun jene Zielströmung, charaktersiert durch das Tripel (v, f, p), zu ermitteln, sodass v ein gegebenes Geschwindigkeitsfeld v_d am besten approximiert.

2.2 Klassische Formulierung

Mathematisch gesehen, betrachten wir also ein optimales Kontrollproblem mit unbeschränkter Steuerung der Gestalt:

(OP) Optimales Kontrollproblem

Gegeben

- Ortsgebiet $\Omega \subset \mathbb{R}^d$ mit $d \in \{1, 2, 3\}$,
- Zeitintervall $I = (0, T)$,
- Geschwindigkeitsfeld $v_d \in C^1([0, T], (C^2(\Omega))^d)$,
- **Kontroll- oder Kostenparamteter** $\alpha > 0$.

Finde

- **Zustand** $v \in C^1([0, T], (C^2(\overline{\Omega}))^d)$ (Geschwindigkeitsfeld),
- **Zustand** $p \in C((0, T), C^1(\overline{\Omega}))$ (Druck),
- **Steuerung** $f \in C((0, T), (C(\overline{\Omega}))^d)$ (Volumenkraftdichte),

welche die **Zielfunktion**

$$J(v, f) := \frac{1}{2} \int_0^T \int_\Omega |v(x, t) - v_d(x, t)|^2 \, \mathrm{d}x \, \mathrm{d}t + \frac{\alpha}{2} \int_0^T \int_\Omega |f(x, t)|^2 \, \mathrm{d}x \, \mathrm{d}t, \tag{2.2}$$

minimiert, sodass die **Zustandsgleichungen**

$$\begin{aligned} \frac{\partial v}{\partial t} - \Delta_x v + \nabla_x p &= f \quad \text{in } Q = \Omega \times (0, T), \\ \operatorname{div}_x v &= 0 \quad \text{in } Q = \Omega \times (0, T), \\ v &= 0 \quad \text{auf } \Sigma = \Gamma \times (0, T), \\ v(\cdot, 0) &= 0 \quad \text{in } \Omega, \end{aligned} \tag{2.3}$$

erfüllt sind.

Bemerkung 2.1. *Für den weiteren Verlauf dieser Arbeit nehmen wir an, dass unser Ortsgebiet $\Omega \subset \mathbb{R}^d$ ein Lipschitz-Gebiet ist. Ein Gebiet bezeichnet eine offene, beschränkte und zusammenhängende Menge. Lipschitz-Gebiete besitzen zusätzlich die Eigenschaft, dass sich ihr Rand mit endlich vielen Graphen von Lipschitz stetigen Funktionen beschreiben lässt.*

Nachdem es für die mathematische Untersuchung keine Rolle spielt und um Schreibaufwand zu sparen, wurde für die Zustandsgleichungen die dimensionslose Formulierung der instationären Stokes Gleichungen verwendet.

Mittels des zweiten Terms der Zielfunktion lassen sich bei der Berechnung der optimalen Strömung (v, f, p) auch die Kosten (gemessen durch die L^2-Norm) der Steuerung f berücksichtigen. Die Gewichtung der Kostenminimierung wird dabei durch den Wert α festgelegt.

Der Druck p tritt nur in der ersten Zustandsgleichung und dort in Form des Gradienten auf. Das heißt für eine bestimmte Lösung p ist

$$\tilde{p}(x,t) = p(x,t) + C(t) \quad \textit{mit } C(t) \in \mathbb{R}\ \forall t \in (0,T), \tag{2.4}$$

wiederum eine Lösung. Um Eindeutigkeit für den Druck p zu gewährleisten, stellen wir an ihn die Bedingung:

$$\int_\Omega p(x,t)\,\mathrm{d}x = 0 \quad \forall t \in (0,T).$$

In diesem Fall ist nämlich der Wert der Funktion C aus (2.4) mit

$$\int_\Omega \tilde{p}(x,t)\,\mathrm{d}x = 0 \quad \forall t \in (0,T),$$

$$\Leftrightarrow$$

$$C(t) = -\frac{1}{|\Omega|}\int_\Omega p(x,t)\,\mathrm{d}x \quad \forall t \in (0,T).$$

für alle $t \in (0,T)$ eindeutig bestimmt.

2.3 Schwache Formulierung

Als Diskretisierungmethode werden wir später die Finite-Elemente-Methode verwenden. Ausgangspunkt dieser Methode ist eine sogenannte schwache Formulierung von (OP). Dazu leiten wir zunächst eine Variationsformulierung der Zustandsgleichungen (2.3) her.

2.3.1 Variationsformulierung der Zustandsgleichungen

Die Herleitung erfolgt nun in fünf Schritten:

Schritt 1: Multiplikation der beiden Gleichungen mit einer Testfunktion:

$$\int_0^T \int_\Omega (\frac{\partial v}{\partial t} - \Delta_x v + \nabla_x p) \cdot w \,\mathrm{d}x \,\mathrm{d}t = \int_0^T \int_\Omega f \cdot w \,\mathrm{d}x \,\mathrm{d}t,$$
$$\forall w \in (C^\infty((0,T) \times \overline{\Omega}))^d,$$
$$\int_0^T \int_\Omega \mathrm{div}_x \, v \, q \,\mathrm{d}x \,\mathrm{d}t = 0,$$
$$\forall q \in C^\infty((0,T) \times \overline{\Omega}),$$

$$v = 0 \quad \text{auf } \Sigma,$$
$$v(0) = 0 \quad \text{auf } \Omega.$$

Schritt 2: Partielle Integration der Terme höchster Ordnung in der ersten Gleichung bezüglich der räumlichen Variable x:

$$\int_0^T \int_\Omega \frac{\partial v}{\partial t} \cdot w \,\mathrm{d}x \,\mathrm{d}t - \int_0^T \Big(\underbrace{\int_\Gamma \nabla_x v \cdot n \cdot w \,\mathrm{d}s}_{=0 \text{ für } w=0 \text{ auf } \Sigma} - \int_\Omega \nabla_x v \cdot \nabla_x w \,\mathrm{d}x \Big) \,\mathrm{d}t$$
$$+ \int_0^T \Big(\underbrace{\int_\Gamma p \, n \cdot w \,\mathrm{d}s}_{=0 \text{ für } w=0 \text{ auf } \Sigma} - \int_\Omega p \,\mathrm{div}_x \, w \,\mathrm{d}x \Big) \,\mathrm{d}t$$
$$= \int_0^T \int_\Omega f \cdot w \,\mathrm{d}x \,\mathrm{d}t,$$
$$\forall w \in (C^\infty((0,T) \times \overline{\Omega}))^d,$$
$$\int_0^T \int_\Omega \mathrm{div}_x \, v \, q \,\mathrm{d}x \,\mathrm{d}t = 0,$$
$$\forall q \in C^\infty((0,T) \times \overline{\Omega}),$$

$$v = 0 \quad \text{auf } \Sigma,$$
$$v(0) = 0 \quad \text{auf } \Omega.$$

Schritt 3: Einbau der Randbedingungen:

Bei der Randbedingung

$$v = 0 \;\text{ auf }\; \Sigma = \Gamma \times (0,T),$$

handelt es sich offensichtlich um eine *wesentliche* Randbedingung, da sie sich nicht in die Variationsformulierung einarbeiten lässt. Sie muss also weiters explizit an die Lösung v gestellt werden und bedingt eine entsprechende homogene Bedingung

$$w = 0 \;\text{ auf }\; \Sigma,$$

an die Testfunktion w.

Es gilt also:

$$\begin{aligned} &\int_0^T \int_\Omega \frac{\partial v}{\partial t} \cdot w \,\mathrm{d}x \,\mathrm{d}t + \int_0^T \int_\Omega \nabla_x v \cdot \nabla_x w \,\mathrm{d}x \,\mathrm{d}t \\ -&\int_0^T \int_\Omega p \,\mathrm{div}_x\, w \,\mathrm{d}x \,\mathrm{d}t = \int_0^T \int_\Omega f \cdot w \,\mathrm{d}x \,\mathrm{d}t, \\ &\qquad\qquad \forall w \in (C^\infty((0,T) \times \overline{\Omega}))^d \text{ mit } w = 0 \text{ auf } \Sigma, \\ &\int_0^T \int_\Omega \mathrm{div}_x\, v \, q \,\mathrm{d}x \,\mathrm{d}t = 0, \\ &\qquad\qquad \forall q \in C^\infty((0,T) \times \overline{\Omega}), \end{aligned} \tag{2.5}$$

$$\begin{aligned} v &= 0 \quad \text{auf } \Sigma, \\ v(0) &= 0 \quad \text{auf } \Omega. \end{aligned}$$

Der Raum $W((0,T),H)$

In der Variationsformulierung (2.5) treten partielle Ableitungen nach der Zeit als auch im Ort nur im Integranden auf. Wir suchen also nach einem Arbeitsraum für die Lösungs- und Testfunktionen, dessen Elemente eine schwache Ableitung in Raum und Zeit besitzen.

Definition 2.2. *Sei H ein Hilbertraum. Eine Funktion $v : (0,T) \to H$ heißt Treppenfunktion, wenn endlich viele Elemente $v_i \in H$ und Lebesgue-messbare, paarweise disjunkte Mengen $M_i \subset (0,T)$ für $i = 1, \ldots, n$ existieren sodass*

$$\bigcup_{i=1}^{n} M_i \quad \textit{und} \quad v(t) = v_i \ \forall t \in M_i, \ \textit{für } i = 1, \ldots, n,$$

gilt.

Notation. Der Ausdruck $\dot{\forall}$, entspricht im folgenden der Aussage „für fast alle“.

Definition 2.3. *Sei H ein Hilbertraum. Eine Funktion $v : (0,T) \to H$ heißt messbar, wenn eine Folge $(v_n)_{n\in\mathbb{N}}$ von Treppenfunktionen $v_n : (0,T) \to H$ existiert, sodass gilt:*

$$\lim_{n\to\infty} v_n(t) = v \quad \dot{\forall} t \in (0,T),$$

gilt.

Definition 2.4. *Sei H ein Hilbertraum. Unter $L^2((0,T),H)$, verstehen wir den linearen Raum aller (Äquivalenzklassen von) messbaren Funktionen $v : (0,T) \to H$ mit der Eigenschaft*

$$\int_0^T ||v(t)||_H^2 \,\mathrm{d}t < \infty.$$

Versehen mit

- dem inneren Produkt

$$(v,w)_{L^2((0,T),H)} := \int_0^T (v(t),w(t))_H \mathrm{d}t$$

- und der Norm

$$||v||_{L^2((0,T),H)} := (v,v)^{\frac{1}{2}}_{L^2((0,T),H)},$$

bildet $L^2((0,T),H)$ einen Hilbertraum (siehe z.B. [7], Seite 115).

Notation. Um Schreibaufwand zu sparen, verwenden wir, falls aus dem Kontext klar folgt welcher Hilbertraum H gemeint ist, für $(\cdot,\cdot)_{L^2((0,T),H)}$ die abkürzende Schreibweise $(\cdot,\cdot)_{L^2}$.

In $L^2((0,T),H)$ werden also Funktionen, die sich nur auf einer Teilmenge von $(0,T)$ mit Lebesguemaß null unterscheiden, als gleich angesehen, da sie sich in der selben Äquivalenzklasse befinden.

Als Arbeitsraum eignet sich nun die folgende Menge:

Definition 2.5. *Sei H ein Hilbertraum. Unter $W((0,T),H)$ verstehen wir den Raum aller Funktionen $v \in L^2((0,T),H))$, die eine schwache Ableitung $\partial_t v \in L^2((0,T),H^\star)$ besitzen, das heißt*

$$\int_0^T (v(t),w)_H \,\phi(t)\mathrm{d}t = -\int_0^T \langle \partial_t v(t), w\rangle_{H^\star \times H}\,\phi(t)\mathrm{d}t \quad \forall w \in H, \forall \phi \in C_0^\infty(\Omega).$$

Also

$$W((0,T),H) := \{v \in L^2((0,T),H) \,:\, \partial_t v \in L^2((0,T),H^\star)\}.$$

Ausgestattet mit

- innerem Produkt

$$(v,w)_{W((0,T),H)} = \int_0^T (v(t),w(t))_H \,\mathrm{d}t + \int_0^T (v(t),w(t))_{H^\star}\,\mathrm{d}t,$$

- und der Norm

$$\begin{aligned} ||v||_{W((0,T),H)} &= (v,v)^{\frac{1}{2}}_{W((0,T),H)} \\ &= \left(\int_0^T (||v(t)||^2_H + ||\partial_t v(t)||^2_{H^\star})\,\mathrm{d}t\right)^{\frac{1}{2}}. \end{aligned}$$

bildet W((0,T),H) einen Hilbertraum (siehe z.B. [7], Seite 118). Für unseren Anwendungsfall gilt nun speziell

$$H = V^d \quad \text{mit} \quad V = H_0^1(\Omega).$$

Da $W((0,T),V^d) \subset L^2((0,T),V^d)$, bleibt noch die Bedeutung der Anfangsbedingung $v(0) = 0$ zu klären, denn lokal integrierbare Funktionen werden als gleich betrachtet, solange sie sich ja nur auf einer Nullmenge unterscheiden. Es lässt sich jedoch zeigen (Beweis siehe z.B. [7]):

Satz 2.6. *Jedes $v \in W((0,T),V^d)$ lässt sich durch Abänderung auf einer Menge vom Maß Null, als Funktion aus $C([0,T],(L^2(\Omega))^d)$ darstellen. In diesem Sinne gilt die Einbettung*

$$W((0,T),V^d) \hookrightarrow C([0,T],(L^2(\Omega))^d)$$

und diese ist stetig, das heißt

$$\exists c > 0 \quad ||v||_{C([0,T],(L^2(\Omega))^d)} \leq c||v||_{W((0,T),V^d)} \quad \forall v \in W((0,T),V^d).$$

Somit kann jede Funktion $v \in W((0,T),V^d)$ durch ihren stetigen Repräsentanten $\tilde{v} \in C([0,T],(L^2(\Omega))^d)$ identifiziert werden. Die Vorgabe einer Anfangsbedingung $v(0) = v_0 \in (L^2(\Omega))^d$ ist also im Sinne

$$\tilde{v}(0) = v_0,$$

zu verstehen. Weiters erhält man aus der Stetigkeit der Einbettung die Abgeschlossenheit des Unterraumes

$$W_0((0,T),V^d) := \{w \in W((0,T),V^d) : \ w(0) = 0\}.$$

Endgültige Variationsformulierung

Schritt 4: Wahl der Funktionenräume:

Für eine korrekte Variationsformulierung ist es notwendig nachzuweisen, dass alle auftretenden Integrale wohldefiniert, also einen endlich Wert besitzen. Wir wählen nun

- Lösungen:

$$\begin{aligned} v &\in W_0((0,T),V^d), \\ f &\in L^2((0,T),(L^2(\Omega))^d), \\ p &\in L^2((0,T),Q), \end{aligned}$$

- Testfunktionen:

$$\begin{aligned} w &\in W((0,T),V^d), \\ q &\in L^2((0,T),Q), \end{aligned}$$

- Daten:

$$v_d \in W((0,T),V^d),$$

wobei

$$V = H_0^1(\Omega),$$
$$Q = L_0^2(\Omega) := \{p \in L^2(\Omega) : \int_\Omega p(x) = 0 \,\mathrm{d}x\}.$$

Für diese Wahl lässt sich mit Hilfe der Cauchy-Ungleichung die Wohldefiniertheit aller Integrale nachweisen.

Um Eindeutigkeit für den Druck p zu erhalten, haben wir seinen Lösungsraum von $L^2((0,T),L^2(\Omega))$ auf $L^2((0,T),Q)$ eingeschränkt. Diese Einschränkung wurde auch für die Menge der Testfunktionen q vorgenommen, da die Lösung der Variationsformulierung dadurch nicht beeinflusst wird:

Jede Funktion $q \in L^2((0,T),L^2(\Omega))$ lässt sich darstellen in der Form

$$q = q_1 + q_2,$$

mit $q_1 \in L^2((0,T),L_0^2(\Omega))$ und $q_2 \in L^2((0,T),\mathbb{R})$, sodass gilt

$$(\mathrm{div}_x\, v, q)_{L^2} = (\mathrm{div}_x\, v, q_1)_{L^2} + (\mathrm{div}_x\, v, q_2)_{L^2}.$$

Es bleibt zu zeigen

$$(\mathrm{div}_x\, v, q)_{L^2((0,T),L^2(\Omega))} = 0 \quad \forall q \in L^2((0,T),\mathbb{R}).$$

Per Definition der schwachen Ableitung gilt für $v \in W((0,T),V^d)$

$$(\mathrm{div}_x\, v, \phi)_{L^2} = -(v, \nabla_x \phi)_{L^2} \quad \forall \phi \in C_0^\infty((0,T)\times\Omega)$$

und somit weiters für $\phi \in C_0^\infty((0,T),\mathbb{R})$

$$(\mathrm{div}_x\, v, \phi)_{L^2} = 0.$$

Aus der Dichtheit von $C_0^\infty((0,T),\mathbb{R})$ in $L^2((0,T),\mathbb{R})$ (siehe z.B. ([22], Seite 42) erhalten wir schließlich

$$(\mathrm{div}_x\, v, q)_{L^2} = 0 \quad \forall q \in L^2((0,T),\mathbb{R}).$$

Schritt 5: Endgültige Gestalt der Variationsformulierung:

Finde:

$$(v, f, p) \in W_0((0,T), V^d) \times L^2((0,T), (L^2(\Omega))^d) \times L^2((0,T), Q),$$

sodass gilt:

$$\begin{aligned} \int_0^T \int_\Omega \langle \partial_t v, w \rangle \,\mathrm{d}x\,\mathrm{d}t + \int_0^T \int_\Omega \nabla_x v \cdot \nabla_x w \,\mathrm{d}x\,\mathrm{d}t \\ - \int_0^T \int_\Omega p \,\mathrm{div}_x\, w \,\mathrm{d}x\,\mathrm{d}t = \int_0^T \int_\Omega f \cdot w \,\mathrm{d}x\,\mathrm{d}t, \\ \forall w \in W((0,T), V^d), \\ \int_0^T \int_\Omega \mathrm{div}_x\, v\, q \,\mathrm{d}x\,\mathrm{d}t = 0, \\ \forall q \in L^2((0,T), Q). \end{aligned} \tag{2.6}$$

2.3.2 Schwache Formulierung des Kontrollproblems

Im Unterschied zur klassischen Formulierung (OP) erfolgt bei der schwachen Formulierung die Minimierung der Zielfunktion (2.2) über die Lösungen der Variationsformulierung (2.6), welche auch als schwache Lösungen der Zustandsgleichungen (2.3) bezeichnet werden:

(SOP) Schwache Formulierung des optimalen Kontrollproblems

Gegeben

- $v_d \in W((0,T), V^d)$,
- $\alpha > 0$.

Finde

- $v \in W_0((0,T), V^d)$,
- $f \in L^2((0,T), (L^2(\Omega))^d)$,
- $p \in L^2((0,T), Q)$,

welche die **Zielfunktion**

$$J(v,f) := \frac{1}{2} ||v - v_d||^2_{L^2((0,T),(L^2(\Omega))^d)} + \frac{\alpha}{2} ||f||^2_{L^2((0,T),(L^2(\Omega))^d)}, \tag{2.7}$$

minimiert, sodass die **Zustandsgleichungen**

$$\begin{aligned} (\partial_t v, w)_{(L^2((0,T)\times\Omega))^d} + (\nabla_x v, \nabla_x w)_{(L^2((0,T)\times\Omega))^{d\times d}} \\ + (\mathrm{div}_x\, w, p)_{(L^2((0,T)\times\Omega))^d} = (f, w)_{(L^2((0,T)\times\Omega))^d}, \\ \forall w \in W((0,T), V^d), \\ (\mathrm{div}_x\, v, q)_{(L^2((0,T)\times\Omega))^d} = 0, \\ \forall q \in L^2((0,T), Q), \end{aligned} \tag{2.8}$$

erfüllt sind.

2.4 Zusammenhang: Klassische und Schwache Formulierung

Es lässt sich zeigen, dass jede Lösung der klassischen Formulierungen (OP), auch Lösungen der schwachen Formulierung (SOP) ist.

Auch die Umkehrung, kann unter bestimmten Differenzierbarkeitsbedingungen an die Lösungen von (SOP) nachgewiesen werden.

Unter bestimmten Voraussetzungen lassen sich also gezeigte Aussagen für (OP) oder (SOP) auf die jeweilige andere Formulierung übertragen. Dieser Tatsache bewusst, erfolgt nun im nächsten Kapitel die Untersuchung bezüglich der Existenz und Eindeutigkeit einer Lösung an Hand der schwachen Formulierung (SOP).

Kapitel 3

Existenz, Eindeutigkeit und Charakterisierung der Lösung

In diesem Kapitel analysieren wir die schwache Formulierung (SOP) bezüglich der Existenz, Eindeutigkeit und Charakterisierung von Lösungen.

3.1 Existenz und Eindeutigkeit der Lösung

Wir betrachten zunächst ein Optimierungsproblem mit Nebenbedingungen in der abstrakten Gestalt

$$\min_{x \in C} F(x), \tag{3.1}$$

wobei H ein Hilbertraum, $C \subset H$ und $F : H \to \mathbb{R}$ ein Funktional ist. Speziell für unser Optimierungsproblem (SOP), geschrieben in der Form (3.1), gilt:

1.
$$H := W_0((0,T), V^d) \times L^2((0,T), (L^2(\Omega))^d) \times L^2((0,T), Q). \tag{3.2}$$

2.
$$\begin{aligned} C &:= \big\{(v,f,p) \in H : (v,f,p) \text{ Lösung der Zustandsgleichung (2.8)}\big\} \\ &= \big\{(v,f,p) \in H : \ S(v,p) = \begin{pmatrix} f \\ 0 \end{pmatrix}\big\}, \end{aligned} \tag{3.3}$$

mit

$$S : W_0((0,T), V^d) \times L^2((0,T), Q) \to W((0,T), V^d)^\star \times L^2((0,T), Q)^\star,$$

$$\langle S(v,p), (w,q) \rangle := \begin{pmatrix} (\partial_t w, \lambda)_{L^2} + (\nabla_x v, \nabla_x w)_{L^2} + (\operatorname{div}_x w, p)_{L^2} \\ (\operatorname{div}_x v, q)_{L^2} \end{pmatrix}, \tag{3.4}$$

$$\forall v, w \in W((0,T), V^d), \ \forall p, q \in L^2((0,T), Q),$$

und

$$\langle f, w\rangle := (f, w)_{L^2} \quad \forall w \in W((0,T), V^d).$$

3.

$$\begin{aligned} F &: H \longrightarrow \mathbb{R}, \\ F(x) &:= J(v, f), \quad \forall x = (v, f, p) \in H, \end{aligned} \tag{3.5}$$

Für das abstrakte Problem (3.1), lassen sich nun unter bestimmten Voraussetzungen an die Menge C und das Funktional F Aussagen über die Lösung nachweisen. Dazu bedarf es der Einführung der folgenden Begriffe:

Definition 3.1. *Sei X ein Vektorraum und $M \subseteq X$. Dann bezeichnen wir M als konvexe Teilmenge von X, falls gilt*

$$\forall x, y \in M \, \forall \lambda \in (0,1): \quad \lambda x + (1-\lambda) y \in M.$$

Beispiel 3.2. Sei $Z = X \times Y$ wobei X, Y Hilberträume und $L_1 : X \to Y^\star, L_2 : Y \to Y^\star$ lineare Operatoren, dann ist

$$M := \{(x, y) \in Z \mid L_1 x = L_2 y\},$$

eine konvexe Teilmenge von Z: Für $\lambda \in (0,1)$ und $(x_1, y_1), (x_2, y_2) \in Z$ gilt:

$$L_1 x_i = L_2 y_i \quad \text{für } i \in \{1, 2\}.$$

Aus der Linearität von L_1 und L_2 erhält man nun,

$$\begin{aligned} L_1(\lambda x_1 + (1-\lambda) x_2) =& \lambda L_1 x_1 + (1-\lambda) L_2 x_2 \\ =& \lambda L_2 y_1 + (1-\lambda) L_2 y_2 \\ =& L_2(\lambda y_1 + (1-\lambda) y_2). \end{aligned}$$

Also ist $\lambda(x_1, y_1) + (1-\lambda)(x_2, y_2) \in Z$ und somit M konvex.

Definition 3.3. *Sei X ein Vektorraum, $M \subset X$ konvex und $F : X \to \mathbb{R}$.*

- *F heißt konvex auf M, falls*

$$\forall x, y \in M \, \forall \lambda \in (0,1) : F(\lambda x + (1-\lambda) y) \leq \lambda F(x) + (1-\lambda) F(y),$$

- *F heißt strikt konvex auf M, falls*

$$\forall x, y \in M \text{ mit } x \neq y \, \forall \lambda \in (0,1) : F(\lambda x + (1-\lambda) y) < \lambda F(x) + (1-\lambda) F(y).$$

Beispiel 3.4. Sei $X = L^2(\Omega)$, $M \subset X$ konvex, $w \in L^2(\Omega)$, dann ist die Funktion

$$F(v) := ||v - w||_{L^2}^2,$$

auf M strikt konvex: Für $v_1, v_2 \in M$ und $\lambda \in (0,1)$ gilt:

$$\begin{aligned} F(\lambda v_1 + (1-\lambda)v_2) &= ||\lambda v_1 + (1-\lambda)v_2 - (\lambda + (1-\lambda))w||_{L^2}^2 \\ &= ||\lambda(v_1 - w) + (1-\lambda)(v_2 - w)||_{L^2}^2 \\ &< \lambda||v_1 - w||_{L^2}^2 + (1-\lambda)||v_1 - w||_{L^2}^2 \\ &= \lambda F(v_1) + (1-\lambda)F(v_2), \\ &\Leftrightarrow \\ 0 < \underbrace{\lambda(1-\lambda)}_{>0}&\big(||v_1 - w||_{L^2} + ||v_2 - w||_{L^2}\big)^2, \\ &\Leftrightarrow \\ v_1 &\neq v_2. \end{aligned}$$

Definition 3.5. *Sei X ein normierter Raum, $M \subset X$ und $F : X \to \mathbb{R}$. Dann heißt F radial unbeschränkt auf M, falls*

$$\forall (x_n)_{n\in\mathbb{N}} \subset M : \quad ||x_n||_X \to \infty \Rightarrow F(x_n) \to \infty.$$

Beispiel 3.6. Sei $X = \mathbb{R}^2$, $M = \{(x,y) \in X \mid y = 0\}$ und $((x_n, y_n))_{n\in\mathbb{N}}$ eine Folge M mit $\lim_{n\to\infty} ||(x_n, y_n)||_{\ell^2}$, dann ist die Funktion $F(x,y) := |x|$ auf der konvexen Menge M radial unbeschränkt, denn aus der Tatsache $y_n = 0$ für alle $n \in \mathbb{N}$ erhält man

$$\lim_{n\to\infty} ||(x_n, y_n)||_{\ell^2} = \lim_{n\to\infty} |x_n| = \infty.$$

Mit Hilfe dieser Begriffe lässt sich jetzt für (3.1), eine Aussage über die Existenz und Eindeutigkeit einer Lösung nachweisen (Beweis siehe z.B. [7], Seite 16).

Satz 3.7. *Sei*

- *H ein Hilbertraum,*
- *$C \subset X$ nichtleer, konvex und abgeschlossen,*
- *$F : X \to \mathbb{R}$ stetig, radial unbeschränkt auf C und strikt konvex auf C.*

Dann besitzt das Optimierungsproblem (3.1)

$$\min_{x\in C} F(x), \tag{3.6}$$

genau eine Lösung.

Nun sind aber alle Voraussetzungen von Satz 3.7 für unser Problem (SOP) erfüllt:

1. Der Raum H bildet mit dem inneren Produkt

$$\begin{aligned}(x,y)_H :=&(v,e)_{W((0,T),V^d)} + (f,g)_{L^2((0,T),(L^2(\Omega))^d)} \\ &+ (p,q)_{L^2((0,T),Q)}\end{aligned}$$

$$\forall x =(v,f,p), y = (w,g,q) \in H,$$

als Kreuzprodukt von zwei Hilberträumen wiederum einen Hilbertraum.

2. Der Lösungsoperator S besitzt die folgenden Eigenschaften (Beweis siehe [19], Seite 46):

Satz 3.8. *Es gilt:*

Der Operator S aus (3.4) bildet einen Isomorphismus, das heißt S ist linear, stetig und besitzt eine stetige Inverse S^{-1}.

Die variationelle Formulierung der Zustandsgleichung besitzt also für eine beliebige rechte Seite $f \in L^2((0,T),(L^2(\Omega))^d)$ eine eindeutige Lösung $(v,p) \in W((0,T),V^d) \times L^2((0,T),Q)$, welche wiederum stetig von der rechten Seite f abhängt. Daraus folgt weiters für die Menge C:

Satz 3.9. *Es gilt:*

Die in (3.3) definierte Menge C ist nicht leer, konvex und abgeschlossen.

Beweis. C nicht leer und konvex, folgt aus der Tatsache, dass $(0,0,0) \in C$ und Beispiel 3.2.

C ist abgeschlossen: Sei $((v_n,f_n,p_n))_{n\in\mathbb{N}}$ eine Folge in H mit

$$(v_n,f_n,p_n) \xrightarrow[n\to\infty]{} (v,f,p).$$

Aus der Stetigkeit des Operators S folgt nun

$$f = \lim_{n\to\infty} f_n = \lim_{n\to\infty} S(v_n,p_n) = S(\lim_{n\to\infty}(v_n,p_n)) = S(v,p),$$

das heißt $(v,f,p) \in C$ und somit ist C abgeschlossen. □

3. Das Funktional F besitzt die folgenden Eigenschaften:

Satz 3.10. *Sei $\alpha > 0$, $v_d \in W((0,T),V^d)$, H der Hilbertraum aus (3.2) und C die in (3.3) definierte Menge. Dann gilt:*

Das Funktional F aus (3.5) ist stetig auf H, strikt konvex auf C und radial unbeschränkt auf C.

Beweis. F ist stetig: Sei $x = (v, f, p) \in H$ beliebig aber fix und $(x_n)_{n\in\mathbb{N}}$ mit $x_n = (v_n, f_n, p_n) \in H$ eine gegen x konvergente Folge, dass heißt

$$\begin{aligned}\lim_{n\to\infty} ||x_n - x||_X^2 = \lim_{n\to\infty} (&||v_n - v||_{L^2}^2 + ||\partial_t v_n - \partial_t v||_{L^2}^2 \\ &+ ||f_n - f||_{L^2}^2 + ||p_n - p||_{L^2}^2) \\ =&0.\end{aligned}$$

Also

$$\lim_{n\to\infty} v_n = v \quad \text{und} \quad \lim_{n\to\infty} f_n = f,$$

in L^2. Daraus folgt

$$\begin{aligned}\lim_{n\to\infty} |F(x_n) - F(x)| =&|\frac{1}{2}(||v_n - v_d||_{L^2}^2 - ||v - v_d||_{L^2}^2) \\ &+ \frac{\alpha}{2}(||f_n||_{L^2}^2 - ||f||_{L^2}^2)| \\ =&0\end{aligned} \tag{3.7}$$

und somit ist F in x stetig. Wegen der Beliebigkeit von $x \in H$ ist F auf ganz H stetig.

F ist strikt konvex auf C: Definiert man zunächst

$$j(v) := ||v - v_d||_{L^2}^2 \quad \text{und} \quad l(f) := ||f||_{L^2}^2,$$

dann lässt sich F schreiben als

$$F(v, f, p) = \frac{1}{2} j(v) + \frac{\alpha}{2} l(f).$$

Nach Beispiel 3.4 sind die Funktionen j und l strikt konvex auf $L^2((0,T), (L^2(\Omega))^d)$, woraus die Konvexität von F folgt. Für die strikte Konvexität von F, gilt es jetzt noch zu zeigen, dass für (v, f, p) und (w, g, q) in H gilt:

$$p \neq q \Rightarrow (f \neq g \vee v \neq w).$$

Sei nun $f = g$ und $v = w$, dann folgt aus

$$0 = f - g = S(v, p) - S(w, q) = S(0, p - q),$$

und der Injektivität von S, $p = q$, womit die Implikation gezeigt wäre.

F ist radial unbeschränkt auf C: Sei also $(x_n)_{n\in\mathbb{N}}$ wobei $x_n = (v_n, f_n, p_n)$ eine Folge in C, mit

$$\lim_{n\to\infty} ||x_n||_H = \infty,$$

dann gilt

$$\lim_{n\to\infty} ||v_n||_{W((0,T),V^d)} = \infty \vee \lim_{n\to\infty} ||f_n||_{L^2((0,T),(L^2(\Omega))^d)} = \infty \\ \vee \lim_{n\to\infty} ||p_n||_{L^2((0,T),Q)} = \infty. \tag{3.8}$$

Aus der Beschränktheit des Operators S^{-1},

$$(||v_n||^2_{W((0,T),V^d)} + ||p_n||^2_{L^2((0,T),Q)})^{\frac{1}{2}} \leq ||S^{-1}||||f_n||_{L^2((0,T),Q)} \quad \forall n \in \mathbb{N},$$

und der Tatsache

$$\frac{\alpha}{2}||f_n||^2_{L^2((0,T),Q)} \leq |J(x_n)| \quad \forall n \in \mathbb{N},$$

folgt aus (3.8):

$$\lim_{n\to\infty} ||x_n||_H = \infty \Rightarrow \lim_{n\to\infty} ||f_n||_{L^2((0,T),(L^2(\Omega))^d)} = \infty \Rightarrow \lim_{n\to\infty} |J(x_n)| = \infty.$$

□

Somit erhalten wir aus Satz 3.7 für (SOP):

Satz 3.11. *Es gilt:*

Das Optimierungsproblem (SOP) besitzt genau eine Lösung.

3.2 Charakterisierung der Lösung

Weiters lassen sich für das abstrakte Problem (3.1)

$$\min_{x\in C} F(x).$$

notwendige und hinreichenden Optimalitätsbedingungen herleiten, dazu benötigten wir die folgenden beiden verallgemeinerten Ableitungsbegriffe:

Definition 3.12. *Sei*

- X *und* Y *ein Banachraum,*
- $M \subset X$ *offen,*
- $F : X \to Y$.

Existiert für $x \in M$, $h \in X$ *der Grenzwert*

$$F'(x,h) := \lim_{t\to 0} \frac{1}{t}(F(x+th) - F(x))$$

in Y, *so bezeichnet* $F'(x,h)$ *die Richtungsableitung von* F *an der Stelle* x *in Richtung* h.

Definition 3.13. *Sei*

- X *und* Y *ein Banachraum,*
- $M \subset X$ *offen,*
- $F : X \to Y$.

Existiere weiters für $x \in M$ *die Richtungsableitung in alle Richtungen* $h \in X$ *und ein linearer stetiger Operator* $L : X \to Y$, *sodass*

$$F'(x, h) = Lh$$

für alle h *aus* X, *dann heißt* F *an der Stelle* x *Gâteaux-differenzierbar und* L *Gâteaux-Ableitung von* F *and der Stelle* x. *Wir schreiben* $L = F'(x)$. *Weiters heißt* F *Gâteaux-differenzierbar auf* M, *falls* F *für alle* $x \in M$ *Gâteaux-differenzierbar ist.*

Stetige und lineare Funktionale auf Hilberträumen, lassen sich dabei eindeutig bei einem Element aus dem Hilbertraum identifizieren (Beweis siehe z.B. [8],Seite 162):

Satz 3.14 (Der Rieszsche Darstellungssatz)**.** *Sei* H *ein Hilbertraum. Dann exisiert eine isomorphe Abildung*

$$\mathcal{J} : H^\star \longrightarrow H, \quad \textit{(Riesz-Isomorphismus)}$$

für welche gilt:

$$\langle F, v\rangle_{H^\star \times H} = (\mathcal{J}(F), v)_H \quad \forall F \in H^\star, \ \forall v \in H.$$

Diese Tatsache motiviert die Einführung des folgenden Begriffes:

Definition 3.15. *Sei*

- H *ein Hilbertraum,*
- $M \subset H$ *offen,*
- $x \in M$,
- $F : H \longrightarrow \mathbb{R}$ *linear, stetig und in* x *Gâteux-differenzierbar.*

Weiters sei $z \in H$ *jenes nach dem Rieszschen Darstellungssatz existierende Element für das gilt*

$$\langle F'(x), h\rangle_{H^\star \times H} = (z, h)_H \quad \forall h \in H,$$

dann bezeichnen wir z *als den (Gâteaux-)Gradient von* F *an der Stelle* x. *Wir schreiben* $z = \nabla F(x)$.

Beispiel 3.16. 1. Lineare Abildung: Seien U, V Banachräume und $A : U \to V$, und

$$\begin{aligned} j &: U \to V \\ j(x) &:= Ax \quad \forall x \in H, \end{aligned}$$

dann gilt für die Gâteaux-Ableitung

$$\begin{aligned} \lim_{t\to 0} \frac{1}{t}(j(x+th) - j(x)) =& \lim_{t\to 0} \frac{1}{t}(A(x+th) - A(x)) \\ =& \lim_{t\to 0} \frac{1}{t}(Ax + tAh - Ax) \\ =& Ah \end{aligned}$$

also

$$j'(x, h) = Ah$$

und somit

$$j'(x) = A.$$

2. Normquadrat im Hilbertraum: Sei H ein Hilbertraum, $y \in H$ und

$$\begin{aligned} j &: H \to \mathbb{R} \\ j(x) &:= (x - y, x - y)_H = ||x - y||_H^2 \quad \forall x \in H. \end{aligned}$$

Für die Gâteaux-Ableitung von j folgt:

$$\begin{aligned} \lim_{t\to 0} \frac{1}{t}(j(x+th) - j(x)) =& \lim_{t\to 0} \frac{1}{t}(||x + th - y||^2 - ||x - y||_h^2) \\ =& \lim_{t\to 0} \frac{1}{t}((x - y, x - y)_H + 2t(x - y, h)_H \\ & + t^2(h, h)_H - (x, x)_H) \\ =& 2(x - y, h)_H, \end{aligned}$$

also

$$\langle f'(x), h\rangle_{H^\star \times H} = 2(x - y, h)_H \quad \text{und} \quad \nabla f(x) = 2(x - y).$$

Unter Verwendung der zuvor definierten Ableitungsbegriffe lassen sich die Lösungen von (3.1) auf folgende Weise charakterisieren:

Satz 3.17. *Sei*

- X *und* Y *ein Hilbertraum,*
- $C \subset X$ *konvex,*

- $F : X \to \mathbb{R}$ *auf* C *Gâteaux-differenzierbar,*
- $x \in C$.

Dann gilt:

1. *(Notwendige Bedingung 1.Ordnung) Falls* x *Lösung von (3.1), dann ist die folgende Variationsungleichung erfüllt*

$$F'(x)(y - x) \geq 0 \quad \forall y \in C. \tag{3.9}$$

2. *(Hinreichende Bedingung 1.Ordnung) Erfüllt* x *die Bedingung (3.9) und ist zusätzlich* F *konvex, dann löst* x *das Problem (3.1).*

Beweis. (siehe auch [7], Seite 50) Sei $x \in C$ eine Lösung von (3.1), dann gilt:

$$F(y) - F(x) \geq 0 \quad \forall y \in C. \tag{3.10}$$

Für ein beliebig gewähltes $y \in C$ betrachten wir nun die konvexe Linearkombination

$$y(t) = x + t(y - x),$$

mit $t \in (0, 1)$. Die Konvexität von C sichert $y(t) \in C$. Aus (3.10) folgt weiters

$$\frac{1}{t}(F(x + t(y - x)) - F(x)) \geq 0.$$

Nach dem Grenzübergang $t \to 0$ erhält man

$$F'(x)(y - x) \geq 0,$$

die Variationsungleichung (3.9).

Für die Umkehrung, nehmen wir an, dass für $x \in C$ die Variationsungleichung (3.9) erfüllt ist, dann folgt weiters aus der Konvexität von F auf C

$$\begin{aligned} 0 \leq & F'(x)(y - x) \\ = & \lim_{t \to 0} \frac{F(x + t(y - x)) - F(x)}{t} \\ \leq & \lim_{t \to 0} \frac{t(F(y) - F(x))}{t} \\ = & F(y) - F(x) \quad \forall y \in C, \end{aligned}$$

also ist x Lösung von (3.1). □

Für unser Problem (SOP), ist jedoch auch unter der Annahme, dass alle Voraussetzungen von Satz 3.17 erfüllt sind, die Überprüfung der Bedingung (3.9) praktisch unmöglich, da die Menge

$$C := \{(v, f, p) \in H : (v, f, p) \text{ Lösung der Zustandgleichung (2.8)}\}$$

unendlich viele Elemente besitzt.

Um praktische anwendungsfreundlichere Optimalitätsbedingungen zu erhalten, schränken wir uns zunächst auf Optimierungsprobleme mit linearen Nebenbedingungen ein. Diese besitzen abstrakt die Gestalt:

$$\begin{aligned} &\min_{x \in C} F(x) \\ &C := \{x \in H : \ Lx = 0\}, \end{aligned} \tag{3.11}$$

wobei H, Y Hilberträume und $F : H \to \mathbb{R}$ ein Funktional und $L : H \longrightarrow Y^\star$ ein linearer Operator sind.

Offensichtlich handelt es sich auch bei unserem optimalen Kontrollproblem (SOP), um ein Optimierungsproblem mit linearen Nebenbedingungen. Unter Verwendung des in (3.2) definierten Hilbertraums H und dem Funktional F aus (3.5), lässt sich dann (SOP) in der Form (3.11) schreiben, wobei für Y und L gilt:

1.

$$Y := W((0,T), V^d) \times L^2((0,T), Q), \tag{3.12}$$

2.

$$L : H \to Y^\star,$$

$$\langle Lx, y\rangle_{Y^\star \times Y} := \left\langle S(v,p) - \begin{pmatrix} f \\ 0 \end{pmatrix}, (w,q) \right\rangle_{Y^\star \times Y}, \tag{3.13}$$

$$\forall x = (v, f, p) \in H, \forall y = (w, q) \in Y.$$

dabei bezeichnet S den Operator aus (3.4).

Optimalitätsbedingungen für (3.11) können nur unter Annahme bestimmter Abbildungseigenschaften für den Operator L gezeigt werden. Um diese zu beschreiben, definieren wir zunächst ein paar Begriffe:

Definition 3.18. *Sei*

- *X und Y ein Hilbertraum,*
- *$T : X \longrightarrow Y^\star$ linear und stetig.*

Dann bezeichnen wir mit $T^\star : Y \longrightarrow X^\star$ den zu T adjungierten Operator, falls gilt:

$$\langle Tx, y\rangle_{Y^\star \times Y} = \langle x, T^\star y\rangle_{X \times X^\star} \quad \forall x \in X, \ \forall y \in Y. \tag{3.14}$$

Bezüglich der Existenz und Eindeutigkeit des adjungierten Operators, lässt sich die folgende Aussage zeigen:

Satz 3.19. *Sei*

- *X und Y ein Hilbertraum,*
- *$T : X \longrightarrow Y^\star$ linear und stetig.*

Dann existiert ein eindeutiger zu T adjungierter Operator $T^\star$, welcher linear und stetig ist. Weiters gilt

$$||T||_{L(X,Y^\star)} = ||T^\star||_{L(Y,X^\star)}. \tag{3.15}$$

Beweis. Existenz: Sei

$$T^\star : Y \longrightarrow X^\star$$
$$\langle T^\star y, x\rangle_{X^\star \times X} := (\mathcal{J}(Tx), y)_Y,$$

wobei $\mathcal{J}$ den *Riesz-Isomorphismus* aus Satz bezeichnet, dann ist $L^\star$ linear und erfüllt die Bedinung aus (3.14). Weiters folgt aus

$$\begin{aligned} ||T||_{L(X,Y^\star)} :=& \sup_{x\in X}\sup_{y\in Y} \frac{\langle Tx, y\rangle_{Y^\star\times Y}}{||x||_X||y||_Y} \\ =& \sup_{y\in Y}\sup_{x\in X} \frac{\langle T^\star y, x\rangle_{X^\star\times X}}{||x||_X||y||_Y} \\ =& ||T^\star||_{L(X,Y^\star)}, \end{aligned}$$

die Normgleichheit (3.15) und die Beschränktheit von $T^\star$.

Eindeutigkeit: Seien T_1 und T_2 jeweils zu T adjungierte Operatoren, dann gilt:

$$\begin{aligned} \langle x, (T_1^\star - T_2^\star)y\rangle_{X\times X^\star} =& \langle x, T_1^\star y\rangle_{X\times X^\star} - \langle x, T^\star y\rangle_{X\times X^\star} \\ =& \langle Tx, y\rangle_{Y^\star\times Y} - \langle Tx, y\rangle_{Y^\star\times Y} \\ =& 0 \\ \forall x \in X, \forall y \in Y,& \\ \Leftrightarrow& \\ (T_1^\star - T_2^\star)y =& 0 \ \forall y \in Y, \\ \Leftrightarrow& \\ T_1^\star =& T_2^\star. \end{aligned}$$

□

Definition 3.20. *Sei*

- X *und* Y *ein Banachraum,*
- $S \subseteq X$
- $T : X \longrightarrow Y$ *linear und stetig.*

Dann bezeichnen wir mit

$$\mathrm{im}(T) := \{y \in Y \ : \ \exists x \in X : \ Tx = y\} \text{ das Bild von } T,$$

$$\mathrm{kern}(T) := \{x \in X : \ Tx = 0\} \text{ den Kern von } T,$$

$$S^\circ := \{y \in X^\star \ : \ \langle y, x\rangle_{X^\star \times X} = 0 \ \forall x \in S\} \text{ den Annihilator von } S.$$

Für den Bild- und Kernbereich eines Operators und seiner Adjungierten, gelten dabei die folgenden Zusammenhänge (Beweis siehe z.B. [17], Seite 159):

Satz 3.21 (Satz vom abgeschlossenen Bild). *Sei*

- X *und* Y *ein Banachraum,*
- $T : X \longrightarrow Y$ *linear und stetig.*

Dann sind folgende Aussagen äquivalent:

1. *Das Bild von* T *ist abgeschlossen, d.h.* $\mathrm{im}(T) = \overline{\mathrm{im}(T)}$,
2. $\mathrm{im}(T) = (\mathrm{kern}(T^\star))^\circ$,
3. *Das Bild von* $T^\star$ *ist abgeschlossen, d.h.* $\mathrm{im}(T^\star) = \overline{\mathrm{im}(T^\star)}$,
4. $\mathrm{im}(T^\star) = (\mathrm{kern}(T))^\circ$.

Mit Hilfe der letzten Aussage, lassen sich nun die Lösungen von (3.11) folgend charakterisieren:

Satz 3.22. *Sei*

- H *und* Y *ein Hilbertraum,*
- $F : H \to \mathbb{R}$ *auf* X *Gâteaux-differenzierbar und konvex,*
- $L : H \to Y^\star$ *ein linearer, stetiger und surjektiver Operator,*
- $x \in C$.

Dann gilt:

$$\begin{gathered} x \text{ ist eine Lösung von } (3.11), \\ \Longleftrightarrow \\ \exists \lambda \in Y \text{ sodass gilt: } \nabla \mathcal{L}(x, \lambda) = 0, \end{gathered}$$

wobei

$$\mathcal{L}(x, \lambda) := F(x) - \langle Lx, \lambda \rangle_{Y^\star \times Y}$$

das Lagrange-Funktional von (3.11) bezeichnet.

Beweis. Aus der Linearität des Operators L erhalten wir die Konvexität der Menge C, den für $\lambda \in (0,1)$ und $x, y \in H$ gilt:

$$L(\lambda x + (1-\lambda)y) = \lambda Lx + (1-\lambda)Ly = 0,$$

also $\lambda x + (1-\lambda)y \in C$.

Weiters folgt aus der Surjetivität von L und Abgeschlossenheit des Hilbertraumes $Y^\star$, $\mathrm{im}(L) = Y^\star = \overline{\mathrm{im}(L)}$. Unter Verwendung der Aussagen von Satz 3.17 und Satz 3.21 lässt sich nun zeigen:

$$\begin{gathered} x \text{ ist eine Lösung von } (3.11), \\ \Longleftrightarrow \\ x \in C \ \wedge \ F'(x)(y-x) \geq 0 \ \ \forall y \in C, \\ \Longleftrightarrow \\ Lx = 0 \wedge F'(x)y = 0 \ \ \forall y \in \mathrm{kern}(L), \\ \Longleftrightarrow \\ Lx = 0 \ \wedge \ F'(x) \in \mathrm{kern(L)}^\circ = \mathrm{im}(L^\star), \\ \Longleftrightarrow \\ Lx = 0 \ \wedge \ \exists \lambda \in Y : \ L^\star \lambda = F'(x), \\ \Longleftrightarrow \\ \exists \lambda \in Y : \nabla \mathcal{L}(x, \lambda) = 0. \end{gathered}$$

□

Wir werden jetzt zeigen, dass unser Optimierungproblem (SOP) alle Voraussetzungen von Satz 3.22 erfüllt:

1. Mit der analogen Argumentation wie zuvor für H gilt: Y ausgestattet mit dem inneren Produkt

$$(x, y)_Y := (v, w)_{W((0,T),V^d)} + (p, q)_{L^2((0,T),Q)},$$

$$\forall x = (v, p), y = (w, q) \in Y,$$

bildet als direktes Produkt von zwei Hilberträumen, wiederum einen Hilbertraum.

2. Mit Hilfe der Beispiele aus (3.16) lässt sich für F zeigen:

Satz 3.23. *Sei $\alpha > 0$, $v_d \in W((0,T), V^d)$ und H der Hilbertraum aus (3.2). Dann gilt:*

Das Funktional F aus (3.5) ist auf H Gâteaux-Differenzierbar mit der Ableitung

$$\begin{aligned}\langle F'(x), h\rangle =&(v - v_d, w)_{L^2} + \alpha(f, q)_{L^2}\\ =&(\nabla F(x), h)_{L^2},\end{aligned}$$

$$\forall x = (v, f, p),\ h = (w, g, q) \in H,$$

und Gradienten:

$$\nabla F(x) = \begin{pmatrix} v - v_d \\ \alpha f \\ 0 \end{pmatrix},$$

$$\forall x =(v, f, p) \in H.$$

Beweis. Mittels Beispiel 3.16 gilt also für die Gâteaux-Ableitung von F:

$$\begin{aligned}&\lim_{t\to 0} \frac{1}{t}(F(x + th) - F(x))\\ &= \lim_{t\to 0} \frac{1}{t}\left(\frac{||v + tw||^2_{L^2} - ||v||^2_{L^2}}{2} + \frac{\alpha}{2}(||f + tg||^2_{L^2} - ||f||^2_{L^2})\right)\\ &= \frac{1}{2}\lim_{t\to 0} \frac{||v + tw||^2_{L^2} - ||v||^2_{L^2}}{t} + \frac{\alpha}{2}\lim_{t\to 0} \frac{||f + tg||^2_{L^2} - ||f||^2_{L^2}}{t}\\ &= (v - v_d, w)_{L^2} + \alpha(f, g)_{L^2},\end{aligned}$$

$$\forall x = (v, f, p),\ h = (w, g, q) \in H,$$

also

$$\begin{aligned}\langle F'(x), h\rangle =&(v - v_d, w)_{L^2} + \alpha(f, g)_{L^2}\\ =&\left(\begin{pmatrix} v - v_d \\ \alpha f \\ 0 \end{pmatrix}, \begin{pmatrix} w \\ g \\ q \end{pmatrix}\right)_{L^2}\\ =&(\nabla F(x), h)_{L^2},\end{aligned}$$

$$\forall x = (v, f, p),\ h = (w, g, q) \in H.$$

□

3. Der Operator L besitzt die folgenden Eigenschaften:

Satz 3.24. *Seien H und Y die in (3.2) und (3.12) definierten Hilberträume. Dann gilt:*

Der in (3.13) definierte Operator L ist linear und stetig.

Beweis. Die Linearität von L erhalten wir aus der Linearität des Operators S. Weiters folgt aus der Beschränktheit von S und der Cauchy-Ungleichung für $\mathbb{R}^2$,

$$\begin{aligned} ||L(v,f,p)||_{Y^\star} &\leq (||S||||(v,p)||_{W((0,T),V^d)\times L^2((0,T),Q)} + ||f||_{L^2((0,T),(L^2(\Omega))^d)}) \\ &\leq (1+||S||)^{\frac{1}{2}}||(v,f,p)||_H \quad \forall (v,f,p) \in H \end{aligned}$$

also L ist beschränkt. Lineare und beschränkte Operatoren sind aber stetig, womit die Aussage gezeigt ist. □

Aus Satz 3.22 erhalten wir somit die folgende Charakterisierung der Lösungen von (SOP):

Satz 3.25. *Es gilt:*

$$\begin{gathered} (v,f,p) \text{ ist eine Lösung von } (SOP), \\ \Longleftrightarrow \\ \exists (\lambda,\mu) \in W((0,T),V^d) \times L^2((0,T),Q) \text{ sodass gilt:} \\ \\ \nabla \mathcal{L}(v,f,p,\lambda,\mu) = 0, \end{gathered}$$

wobei $\mathcal{L}(v,f,p,\lambda,\mu)$ das Lagrange-Funktional von (SOP) bezeichnet.

Fassen wir die Aussagen von Satz (3.11) und Satz (3.25) zusammen, so haben wir für die Lösung unseres optimalen Kontrollproblems (SOP) gezeigt:

Satz 3.26. *Es gilt:*

Das optimale Kontrollproblem (SOP) besitzt eine eindeutige Lösung

$$(v,f,p) \in W_0((0,T),V^d) \times L^2((0,T),(L^2(\Omega))^d) \times L^2((0,T),Q),$$

für welche gilt:

$$\begin{gathered} \exists (\lambda,\mu) \in W((0,T),V^d) \times L^2((0,T),Q) \text{ sodass} \\ \\ \nabla \mathcal{L}(v,f,p,\lambda,\mu) = 0, \end{gathered} \tag{3.16}$$

wobei $\mathcal{L}(v,f,p,\lambda,\mu)$ das Lagrange-Funktional von (SOP) bezeichnet.

Sei im folgenden

$$\begin{aligned} X :=& W((0,T),V^d) \times L^2((0,T),(L^2(\Omega))^d) \times L^2((0,T),Q) \\ & \times W((0,T),V^d) \times L^2((0,T),Q). \end{aligned} \tag{3.17}$$

Motiviert bei Satz 3.26 suchen wir also nach

$$(v,f,p,\lambda,\mu) \in X \text{ mit } v(0)=0,$$

welche die Optimalitätsbedingungen (3.16)

$$\nabla \mathcal{L}(v,f,p,\lambda,\mu) = 0.$$

erfüllen. In der Literatur wird die Bedingung (3.16) oft mit dem Namen ihres Erfinders, als **Karish-Kuhn-Tucker-Bedingung** bezeichnet.

3.2.1 Das Lagrange Funktional

Das Lagrange Funktional von (SOP) besitzt dabei die folgende Form:

$$\mathcal{L} : X \to \mathbb{R},$$

wobei

$$\begin{aligned} \mathcal{L}(v,f,p,\lambda,\mu) =& J(v,f) + (f,\lambda)_{L^2} - (\partial_t v,\lambda)_{L^2} - (\nabla_x v, \nabla_x \lambda)_{L^2} \\ & - (\mathrm{div}_x\, \lambda, p)_{L^2} - (\mathrm{div}_x\, v, \mu)_{L^2} \\ =& \frac{1}{2}||v-v_d||^2_{L^2} + \frac{\alpha}{2}||f||^2_{L^2} + (f,\lambda)_{L^2} - (\partial_t v,\lambda)_{L^2} \\ & - (\nabla_x v, \nabla_x \lambda)_{L^2} - (\mathrm{div}_x\, \lambda, p)_{L^2} - (\mathrm{div}_x\, v, \mu)_{L^2}, \end{aligned} \tag{3.18}$$

mit Lagrange-Multiplikatoren

$$\begin{aligned} \lambda &\in W((0,T),V^d), \\ \mu &\in L^2((0,T),Q). \end{aligned}$$

Mit Hilfe des für $W((0,T),V^d)$-Funktionen geltenden Satz der partiellen Integration (Beweis siehe [7], Seite 119)

Satz 3.27. *Sei $v,w \in W((0,T),V^d)$. Dann gilt*

$$\int_0^T (\partial_t v(t), w(t))_{L^2}\, \mathrm{d}t = (v(T),w(T))_{L^2} - (v(0),w(0))_{L^2} - \int_0^T (\partial_t w(t), v(t))_{L^2}\, \mathrm{d}t.$$

folgt dann weiters

$$\begin{aligned} \mathcal{L}(v,f,p,\lambda,\mu) =& \frac{1}{2}||v-v_d||^2_{L^2} + \frac{\alpha}{2}||f||^2_{L^2} + (f,\lambda)_{L^2} \\ & - (v(T),\lambda(T))_{L^2} + (v(0),\lambda(0))_{L^2} + (v,\partial_t\lambda)_{L^2} \\ & - (\nabla_x v, \nabla_x \lambda)_{L^2} - (\mathrm{div}_x\, \lambda, p)_{L^2} - (\mathrm{div}_x\, v, \mu)_{L^2}, \end{aligned}$$

sodass gilt:

$$\begin{aligned}\mathcal{L}(v,f,p,\lambda,\mu) =& \frac{1}{2}||v-v_d||^2_{L^2} + \frac{\alpha}{2}||f||^2_{L^2} + (f,\lambda)_{L^2} \\ & - (v(T),\lambda(T))_{L^2} + (v,\partial_t\lambda)_{L^2} \\ & - (\nabla_x v, \nabla_x \lambda)_{L^2} - (\mathrm{div}_x\,\lambda, p)_{L^2} \\ & - (\mathrm{div}_x\, v, \mu)_{L^2},\end{aligned} \tag{3.19}$$

$$\forall (v,f,p,\lambda,\mu) \in X \text{ mit } v(0) = 0.$$

3.2.2 Das Optimalitätssystem

Wir werden nun für unser Problem (SOP) die Optimalitätsbedingung (3.16) genauer analysieren.

Es gilt für $x = (v,f,p,\lambda,\mu) \in X$ mit $v(0) = 0$:

$$\begin{aligned}\nabla\mathcal{L}(x) &= 0 \quad \text{in } X, \\ &\Leftrightarrow \\ \langle \mathcal{L}'(x), y\rangle_{X^\star \times X} &= 0 \quad \forall y \in X, \\ &\Leftrightarrow \\ \langle \partial_v \mathcal{L}(x), w\rangle &= 0 \quad \forall w \in W((0,T),V^d), \\ \langle \partial_f \mathcal{L}(x), g\rangle &= 0 \quad \forall g \in L^2((0,T),(L^2(\Omega))^d), \\ \langle \partial_p \mathcal{L}(x), q\rangle &= 0 \quad \forall q \in L^2((0,T),Q), \\ \langle \partial_\lambda \mathcal{L}(x), w\rangle &= 0 \quad \forall w \in W((0,T),V^d), \\ \langle \partial_\mu \mathcal{L}(x), q\rangle &= 0 \quad \forall q \in L^2((0,T),Q).\end{aligned} \tag{3.20}$$

Hinter der Optimalitätsbedingung (3.16) verbirgt sich also mit (3.20) ein gekoppeltes System von Variationsgleichungen, welches wir im Weiteren als **Optimalitätssystem** bezeichnen.

Mit Hilfe von Beispiel 3.16 und den Darstellungen des Lagrange-Funktionals $\mathcal{L}$ aus (3.18) und (3.19), lassen sich nun die partiellen Gâteaux-Ableitungen aus (3.20) berechnen:

1. $\partial_v \mathcal{L} = 0$:

$$\begin{aligned}
\langle \partial_v \mathcal{L}(v, f, p, \lambda, \mu), w \rangle &= 0 \quad \forall w \in W((0,T), V^d), \\
&\Leftrightarrow \\
-(\partial_t \lambda, w)_{L^2} + (w(T), \lambda(T))_{L^2} & \\
+(\nabla_x \lambda, \nabla_x w)_{L^2} & \\
+(\operatorname{div}_x w, \mu)_{L^2} &= (v - v_d, w)_{L^2}, \\
\forall w \in W((0,T), V^d), & \\
&\Leftrightarrow \\
-(\partial_t \lambda, w)_{L^2} + (\nabla_x \lambda, \nabla_x w)_{L^2} & \\
+(\operatorname{div}_x w, \mu)_{L^2} &= (v - v_d, w)_{L^2}, \\
\forall w \in W((0,T), V^d), & \\
\lambda(T) &= 0,
\end{aligned}$$

2. $\partial_f \mathcal{L} = 0$:

$$\begin{aligned}
\langle \partial_f \mathcal{L}(v, f, p, \lambda, \mu), g \rangle &= 0 \quad \forall g \in L^2((0,T), (L^2(\Omega))^d), \\
&\Leftrightarrow \\
\alpha(f, g)_{L^2} + (\lambda, g)_{L^2} &= 0, \\
\forall g \in L^2((0,T), (L^2(\Omega))^d), &
\end{aligned}$$

3. $\partial_p \mathcal{L} = 0$:

$$\begin{aligned}
\langle \partial_p \mathcal{L}(v, f, p, \lambda, \mu), q \rangle &= 0 \quad \forall q \in L^2((0,T), Q), \\
&\Leftrightarrow \\
(\operatorname{div}_x \lambda, q)_{L^2} &= 0, \\
\forall q \in L^2((0,T), Q), &
\end{aligned}$$

4. $\partial_\lambda \mathcal{L} = 0$:

$$\begin{aligned}
\langle \partial_\lambda \mathcal{L}(v, f, p, \lambda, \mu), w \rangle &= 0 \quad \forall w \in W((0,T), V^d), \\
&\Leftrightarrow \\
(\partial_t v, w)_{L^2} + (\nabla_x v, \nabla_x w)_{L^2} & \\
+(\operatorname{div}_x w, p)_{L^2} &= (f, w)_{L^2}, \\
\forall w \in W((0,T), V^d), &
\end{aligned}$$

5. $\partial_\mu \mathcal{L} = 0$:

$$\begin{aligned}
\langle \partial_\mu \mathcal{L}(v, f, p, \lambda, \mu), q \rangle &= 0 \quad \forall q \in L^2((0,T), Q), \\
&\Leftrightarrow \\
(\operatorname{div}_x v, q)_{L^2} &= 0, \\
\forall q \in L^2((0,T), Q). &
\end{aligned}$$

Aus der zweiten Gleichung folgt

$$\begin{aligned} f &= \frac{1}{\alpha}\lambda \ \text{ in } L^2((0,T),(L^2(\Omega))^d), \\ &\Leftrightarrow \\ (f,\cdot)_{L^2} &= (\alpha^{-1}\lambda,\cdot)_{L^2} \ \text{ in } W((0,T),V^d)^\star, \end{aligned} \tag{3.21}$$

dass heißt für die weitere Untersuchung dürfen wir annehmen

$$f \in W((0,T),V^d).$$

Einsetzen des Ausdrucks (3.21) in die anderen Gleichungen führt auf das reduzierte Optimalitätssystem:

(OS) Optimalitätssystem für das optimale Kontrollproblem

Gegeben:

- $v_d \in W((0,T),V^d)$,
- $\alpha > 0$.

Gesucht: $(v,f,p,\lambda,\mu) \in X$ sodass gilt:

1.

$$\begin{aligned} -(\partial_t\lambda,w)_{L^2} + (\nabla_x\lambda,\nabla_x w)_{L^2} & \\ +(\operatorname{div}_x w,\mu)_{L^2} &= (v-v_d,w)_{L^2}, \\ \forall w \in W((0,T),V^d), & \\ (\operatorname{div}_x \lambda,q)_{L^2} &= 0, \\ \forall q \in L^2((0,T),Q), & \end{aligned} \tag{3.22}$$

2.

$$\begin{aligned} (\partial_t v,w)_{L^2} + (\nabla_x v,\nabla_x w)_{L^2} & \\ +(\operatorname{div}_x w,p)_{L^2} &= -\alpha^{-1}(\lambda,w)_{L^2}, \\ \forall w \in W((0,T),V^d), & \\ (\operatorname{div}_x v,q)_{L^2} &= 0, \\ \forall q \in L^2((0,T),Q), & \end{aligned} \tag{3.23}$$

mit

$$\begin{aligned} v(0) &= 0, \\ \lambda(T) &= 0, \\ f &= \alpha^{-1}\lambda. \end{aligned}$$

Das Optimalitätssystem von (SOP) entspricht also der schwachen Formulierung zweier gekoppelter nicht stationäre Stokes-Gleichungen.

Kapitel 4

Diskretisierung des Problems

Die Berechnung einer nummerischen Näherungslösung des Optimalitätssystems (OP) erfordert eine Diskretisierung bezüglich des Raumes und der Zeit.

4.1 Räumliche Diskretisierung

Wir diskretisieren zuerst nach der räumlichen Variable. Diese Variante wird auch als die *(vertikale) Linienmethode* bezeichnet.

4.1.1 Anwendung einer stetigen Galerkin-Methode

Für $v \in W((0,T),V^d)$ und $p \in L_2((0,T),Q)$ gilt

$$\begin{aligned} v^i(t) &\in V = H_0^1(\Omega) \quad i \in \{1,\dots,d\}, \dot{\forall} t \in [0,T], \\ p(t) &\in Q = L_0^2(\Omega) \quad \dot{\forall} t \in [0,T]. \end{aligned}$$

Die Diskretisierung erfolgt nun an Hand der **stetigen Galerkin-Methode** durch auswählen der endlich-dimensionalen Teilräume

$$V_h \subset V, U_h \subset L^2(\Omega) \text{ und } Q_h = U_h \cap Q = U_h \cap L_0^2(\Omega),$$

wobei gelte

$$\dim(V_h) = n_h, \dim(Q_h) = m_h \text{ und } \dim(U_h) = r_h$$

für $n_h, m_h, r_h \in \mathbb{N}$, mit $r_h \geq m_h$.

Bemerkung 4.1. *Ziel ist es schlussendlich eine Näherungslösung des räumlich und zeitlich diskretisierten Optimalitätssystems mit Hilfe des Computers zu berechnen. Dieser ist aber in der Speicherkapazität und Rechenleistung begrenzt, weshalb wir uns bei der Wahl auf endlichdimensionale Teilräume beschränken.*

Aus der endlichen Dimension folgt nun die Abgeschlossenheit von V_h, U_h und Q_h, sodass

$$(V_h, (\cdot,\cdot)_{H_1(\Omega)}), (U_h, (\cdot,\cdot)_{L^2(\Omega)}), \text{ und } (Q_h, (\cdot,\cdot)_{L^2(\Omega)}),$$

weiterhin Hilberträume bilden. Daraus folgt wiederum die Wohldefiniertheit der Hilberträume

$$\begin{aligned} W((0,T),(V_h)^d) &\subset W((0,T),V^d), \\ L^2((0,T),(U_h)^d) &\subset L^2((0,T),(L^2(\Omega))^d), \\ L^2((0,T),Q_h) &\subset L^2((0,T),Q). \end{aligned} \tag{4.1}$$

Für die Dimension r_h von U_h gilt genauer

$$r_h = m_h \vee r_h = m_h + 1.$$

Für den Nachweis dieser Tatsache unterscheiden wir zwei Fälle: Sei $U_h \subset Q$, dann gilt offensichtlich $r_h = m_h$. Im Fall $U_h \not\subset Q$ erhält man

$$\operatorname{span}(\{\phi_1^\star, \ldots, \phi_{r_h-1}^\star\}) \subset Q \cap L_0^2(\Omega),$$

mit linear unabhängigen Funktionen

$$\phi_i^\star = \phi_i - \frac{||\phi_i||_{L^1(\Omega)}}{||\phi_{r_h}||_{L^1(\Omega)}}\phi_{r_h} \quad \textit{für} \quad i = 1, \ldots, r_h - 1,$$

sodass gilt $m_h \geq r_h - 1$ woraus mit $U_h \not\subset Q$ folgt $m_h = r_h - 1$.

4.1.2 Existenz einer Basis der diskretisierten Räume

Da ein Computer mit Operatoren, Funktionen und Biliniarformen nichts anzufangen weiß, gilt es diese Ausdrücke so umzuformulieren sodass er arbeiten kann.

Dazu benötigen wir eine Basis für V_h, U_h und Q_h. Es gelte:

$$\begin{aligned} &\{\psi_1, \psi_2, \ldots, \psi_{n_h}\} \text{ sei eine Basis von } V_h, \\ &\{\phi_1, \phi_2, \ldots, \phi_{m_h}\} \text{ sei eine Basis von } Q_h, \\ &\{\phi_1, \phi_2, \ldots, \phi_{m_h}, \phi_{m_h+1}, \ldots, \phi_{r_h}\} \text{ sei eine Basis von } U_h. \end{aligned}$$

Durch das Vorhandensein einer Basis ergeben sich nun die folgenden Zusammenhänge:

Zusammenhang: Funktion-Vektor:

Jedes Element lässt sich eindeutig durch ihren Koeffizientenvektor repräsentieren:

1.

$$\begin{gathered} v_h = (v_h^i)_{i=1,\dots,d} \in V_h^d, \\ \Leftrightarrow \\ \exists! \underline{v} = (\underline{v}^i)_{i=1,\dots,d} \text{ mit } \underline{v}^i = (v_k^i)_{k=1,\dots,n_h} \in \mathbb{R}^{n_h} : \\ v_h^i = \sum_{k=1}^{n_h} v_k^i \, \psi_k, \end{gathered}$$

2.

$$\begin{gathered} p_h \in Q_h, \\ \Leftrightarrow \\ \exists! \ \underline{p} = (p_k)_{k=1,\dots,m_h} \in \mathbb{R}^{m_h} : \quad p_h = \sum_{k=1}^{m_h} p_k \, \phi_k. \end{gathered}$$

Zusammenhang: Operator-Matrix:

Diese eindeutige Identifizierung einer Funktion durch ihren Koeffizientenvektor ermöglicht es Operatoren durch Matrizen zu ersetzen:

1.

$$\begin{aligned} (\nabla_x v_h, \nabla_x w_h)_{(L^2(\Omega))^{d\times d}} &= \sum_{i=1}^{d}\sum_{j=1}^{d} (\partial_{x_j} v_h^i, \partial_{x_j} w_h^i)_{L^2(\Omega)} \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} (\partial_{x_j} v_h^i, \partial_{x_j} w_h^i)_{L^2(\Omega)} \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{n_h}\sum_{l=1}^{n_h} v_l^i w_k^i (\partial_{x_j}\psi_k, \partial_{x_j}\psi_l)_{L^2(\Omega)} \\ &= (A_h \underline{v}, \underline{w})_{\ell^2} \quad \forall v_h, w_h \in V_h^d, \end{aligned} \tag{4.2}$$

wobei

$$A_h = \begin{pmatrix} K_h & 0 & \dots & 0 \\ 0 & K_h & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & K_h \end{pmatrix} = \operatorname{diag}(\underbrace{K_h \ \dots \ K_h}_{d \text{ mal}}) \in \mathbb{R}^{dn_h \times dn_h},$$

mit

$$\begin{aligned} K_h &= \sum_{i=1}^{d} K_h^i = K_h^1 + \cdots + K_h^d, \\ K_h^i &= (K_{kl}^i)_{k,l=1,\ldots,n_h} \\ &= ((\partial_{x_i}\psi_k, \partial_{x_i}\psi_l)_{L^2(\Omega)})_{k,l=1,\ldots,n_h} \\ &= \left(\int_\Omega \partial_{x_i}\psi_l(x)\partial_{x_i}\psi_k(x)\mathrm{d}x \right)_{k,l=1,\ldots,n_h} \qquad \text{für } i = 1,\ldots,d. \end{aligned}$$

2.

$$\begin{aligned} (\mathrm{div}_x\, v_h, p_h)_{L_2(\Omega)} &= \sum_{i=1}^{d} (\partial_{x_i} v_h^i, p_h)_{L^2(\Omega)} \\ &= \sum_{i=1}^{d}\sum_{k=1}^{n_h}\sum_{l=1}^{m_h} v_k^i p_l (\partial_{x_i}\psi_k, \phi_l)_{L^2(\Omega)} \\ &= (B_h \underline{v}, \underline{p})_{\ell^2} \quad \forall v_h \in V_h^d, \forall p_h \in Q_h, \end{aligned} \tag{4.3}$$

wobei

$$B_h = \begin{pmatrix} D_h^1 & \ldots & D_h^d \end{pmatrix} \in \mathbb{R}^{dn_h \times m_h},$$

mit

$$\begin{aligned} D_h^i &= (D_{kl}^i)_{k=1,\ldots,m_h,l=1,\ldots,n_h} \\ &= ((\partial_{x_i}\psi_k, \phi_l)_{L^2(\Omega)})_{k=1,\ldots,m_h,l=1,\ldots,n_h} \\ &= \left(\int_\Omega \partial_{x_i}\psi_l(x)\phi_k(x)\mathrm{d}x \right)_{k=1,\ldots,m_h,l=1,\ldots,n_h} \qquad \text{für } i = 1,\ldots,d. \end{aligned}$$

3.

$$\begin{aligned} (v_h, w_h)_{(L_2(\Omega))^d} &= \sum_{i=1}^{d} (v_h^i, w_h^i)_{L^2(\Omega)} \\ &= \sum_{i=1}^{d}\sum_{k=1}^{n_h}\sum_{l=1}^{n_h} v_k^i w_l^i (\psi_k, \psi_l)_{L^2(\Omega)} \\ &= (C_h \underline{v}, \underline{w})_{\ell^2} \quad \forall v_h, w_h \in V_h^d, \end{aligned} \tag{4.4}$$

wobei

$$C_h = \begin{pmatrix} M_h & 0 & \ldots & 0 \\ 0 & M_h & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & M_h \end{pmatrix} = \mathrm{diag}(\underbrace{M_h \ \ldots \ M_h}_{d \text{ mal}}) \in \mathbb{R}^{dn_h \times dn_h},$$

mit

$$\begin{aligned} M_h &= (M_{kl})_{k,l=1,\dots,n_h} \\ &= ((\psi_k, \psi_l)_{L^2(\Omega)})_{k,l=1,\dots,n_h} \\ &= \left(\int_\Omega \psi_l(x)\psi_k(x)\mathrm{d}x \right)_{k,l=1,\dots,n_h} . \end{aligned}$$

4.1.3 Konstruktion der diskreten Räume mittels der Finite-Element-Methode

Unter einer Finite-Elemente-Methode versteht man spezielle Techniken zur Konstruktion von endlich-dimensionalen Teilräumen.

Wir werden nun überblicksmäßig die Anwendung der Finite-Elemente-Methode für die Konstruktion der Räume V_h, Q_h und Q_h erläutern. Für eine ausführliche Diskussion der Finite-Elemente-Methode siehe ([9]):

Zerlegung

Ausgangspunkt für die Finite Elemente Methode ist eine Zerlegung der Menge Ω:

Definition 4.2. *Sei $\mathcal{R}_h \subset \mathbb{N}$ eine Indexmenge. Die Menge $\mathcal{T}_h(\Omega) = \{\delta_r \mid r \in \mathcal{R}_h\}$ heißt eine Zerlegung der Menge Ω, falls gilt:*

1. $$\bigcup_{r\in\mathcal{R}_h} \overline{\delta_r} = \overline{\Omega},$$

2. $$\delta_r \cap \delta_s = \emptyset \quad \forall r, s \in \mathcal{R}_h \text{ mit } r \neq s.$$

Dabei bezeichne für $r \in \mathcal{R}_h$,

$$h_r := \max_{x,y\in\overline{\delta_r}} |x - y|,$$

den Durchmesser von δ_r und

$$h = \max_{r\in\mathcal{R}_h} h_r,$$

die Feinheit der Zerlegung $\mathcal{T}_h$.

Wahl der endlich dimensionalen Räume

Es seien mit $\mathcal{T}_H(\Omega)$ und $\mathcal{T}_h(\Omega)$ zwei Zerlegungen der Menge Ω gegeben, wobei wir annehmen, dass man $\mathcal{T}_h(\Omega)$ durch Verfeinerung der Zerlegung $\mathcal{T}_H(\Omega)$ erhält. Folglich existiert eine Konstante $c \geq 1$, mit

$$H = ch.$$

Die Räume V_h, U_h und Q_h, werden dann wie folgt gewählt:

$$\begin{aligned} V_h :=& \{v \in H_0^1(\Omega) : v|_{\delta_r} \in \mathcal{F}_1(\delta r) \subset C^1(\delta_r) \ \ \forall r \in \mathcal{R}_h\} \subset V, \\ U_h :=& \{p \in L^2(\Omega) : p|_\delta \in \mathcal{F}_2(\delta_r) \subset C(\delta_r) \ \ \forall \ \ \forall r \in \mathcal{R}_H\} \subset U, \\ Q_h :=& U_h \cap Q \subset Q. \end{aligned}$$

Dabei bezeichnet $\mathcal{F}(\delta_r)$ den Raum der *Formfunktionen* auf δ_r, einen endllich-dimensionalen Raum für welchen eine Basis $\{p^{r,\beta} \mid \beta \in A_r \subset N\}$ gegeben ist. Typische Wahlen für $\mathcal{F}(\delta_r)$ sind:

$$\begin{aligned} \mathbb{P}_k :=& \{\sum_{|\alpha| \leq k} c_\alpha x^\alpha \mid c_\alpha \in \mathbb{R}\}, \quad \text{(Polynome mit Gesamtgrad } \leq k) \\ \mathbb{Q}_k :=& \{\sum_{\alpha_i \leq k, i=1,\ldots,d} c_\alpha x^\alpha \mid c_\alpha \in \mathbb{R}\}, \quad \text{(Polynome mit komponentenweisen Grad } \leq k) \end{aligned}$$

wobei $k \in \mathbb{N}_0$, $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ und $x^\alpha := x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_d^{\alpha_d}$.

Für die Charakterisierung von stückweise differenzierbaren Funktionen in $V = H_0^1(\Omega)$, existiert dabei das folgende Resultat (Beweis siehe [9], Seite 59):

Satz 4.3. *Sei eine Zerlegung von Ω und für v gelte, $v \in C^1(\delta_r)$ für alle $r \in \mathcal{R}_h$ Dann gilt:*

$$v \in H_0^1(\Omega) \Leftrightarrow v \in C_0(\overline{\Omega}).$$

Der Raum V_h kann also weiters in der Form

$$V_h = \{v \in C_0(\overline{\Omega}) : v|_{\delta_r} \in \mathcal{F}_1(\delta r) \subset C^1(\delta_r) \ \ \forall r \in \mathcal{R}_h\},$$

geschrieben werden.

Konstruktion einer Basis

Die Konstruktion einer globalen Basis für die diskreten Räume erfolgt unter Verwendung der gegebenen lokalen Basen $\{p^{r,\beta} \mid \beta \in A_r \subset N\}$ für die einzelnen Formfunktionsräume $\mathcal{F}(\delta_r)$. Dabei wird bei der Konstruktion darauf geachtet, dass die Basisfunktionen einen **lokalen Träger** besitzen, d.h. bis auf ein paar Elemente δ_r sollen die Basisfunktionen verschwinden.

Beispiel 4.4. Definiert man für $r \in \mathcal{R}_{h_2}$ und $\beta \in A_r$ die Funktion

$$p_h^{r,\beta}(x) := \begin{cases} p^{r,\beta}(x) & \text{für } x \in \delta_r, \\ 0 & \text{sonst,} \end{cases}$$

dann bildet die Menge $\{p_h^{r,\beta} \mid r\mathcal{R}_{h_2} \wedge \beta \in A_r\}$ eine Basis von U_h gegeben.

Bemerkung 4.5. *Die Basisfunktionen der diskreten Räume V_h, U_h und Q_h besitzen also einen lokalen Träger, sodass die meisten Einträge in A_h, B_h und C_h verschwinden. Die Matrizen A_h, B_h und C_h sind also dünnbesetzt.*

Für die Dimension der diskreten Räume gilt dabei:

$$h > 0 \Leftrightarrow n_h, m_h \in \mathbb{N}. \tag{4.5}$$

Eigenschaften der Matrizen A_h, B_h und C_h

Wir führen zunächst die folgende Begriffe ein:

Definition 4.6. *Sei $A \in \mathbb{R}^{n\times n}$. Dann heißt A ...*

- *... positiv (semi-) definit, falls gilt*

$$(Av, v)_{\ell^2} > 0 \quad \big((Av, v)_{\ell^2} \geq 0\big) \quad \forall v \in \mathbb{R}^n.$$

- *... negativ (semi-) definit, falls gilt*

$$(Av, v)_{\ell^2} < 0 \quad \big((Av, v)_{\ell^2} \leq 0\big) \quad \forall v \in \mathbb{R}^n.$$

- *... indefinit, falls $v_1, v_2 \in \mathbb{R}^n$ existieren, sodass*

$$\Big((Av_1, v_1)_{\ell^2} > 0 \quad \wedge \quad (Av_2, v_2)_{\ell^2} < 0\Big).$$

Für symmetrische Matrizen, lassen sich diese Begriffe aber auch mit Hilfe der Eigenwerte ausdrücken: Diese Tatsache beruht auf der folgenden Aussage über die Diagonalisierbarkeit von symmetrischen Matrizen (Beweis siehe z.B. [21], Seite 289):

Satz 4.7 (Spektralsatz für Matrizen)**.** *Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische Matrix. Dann existiert eine orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$ und eine Diagonalmatrix*

$$D = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix} \in \mathbb{R}^{n\times n},$$

mit

$$A = QDQ^T,$$

wobei $\lambda_1, \dots, \lambda_n$ die reellen Eigenwerte von A bezeichnen.

Als Folgerung des *Spektralsatzes* erhält man:

Satz 4.8. *Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische, wobei λ_i für $i \in \{1,\dots,n\}$ ihre Eigenwerte bezeichne. Dann gilt: A ist genau dann ...*

- *... positiv (semi-) definit, wenn*

$$\lambda_i > 0 \quad (\lambda_i \geq 0) \quad \forall i \in \{1,\dots,n\}.$$

- *... negativ (semi-) definit, wenn*

$$\lambda_i < 0 \quad (\lambda_i \leq 0) \quad \forall i \in \{1,\dots,n\}.$$

- *... indefinit, wenn $i,j \in \{1,\dots,n\}$ existieren, sodass*

$$\lambda_i > 0 \quad \wedge \quad \lambda_j < 0.$$

Die Matrizen A_h und C_h besitzen die folgenden Eigenschaften:

Satz 4.9. *Sei $h > 0$. Dann gilt:*

> *Die Matrizen A_h und C_h aus (4.2) und (4.4) sind symmetrisch und positiv definit (s.p.d.).*

Beweis. Die Symmetrie der Matrizen K_h und M_h, erhält man aus der Symmetrie des inneren Produktes $(\cdot,\cdot)_{L^2}$.

z.z. K_h ist positiv definit: Sei $0 \neq \underline{v} \in \mathbb{R}^{n_h}$ beliebig aber fix, dann ist $v_h = \sum_{j=1}^{n_h} v_j\phi_j \in V_h = H_0^1(\Omega)$ also nicht konstant. Demnach muss ein Index $i \in \{1,\dots,d\}$ existieren, sodass $\partial_{x_i} v_h$ ungleich der Nullfunktion ist. Woraus folgt:

$$(K_h\underline{v},\underline{v})_{\ell^2} = \sum_{j=1}^{d}(\partial_{x_j}v_h,\partial_{x_j}v_h)_{L^2(\Omega)} \geq (\partial_{x_i}v_h,\partial_{x_i}v_h)_{L^2(\Omega)} = ||\partial_{x_i}v_h||^2_{L^2(\Omega)} > 0.$$

z.z. M_h ist positiv definit: Sei $0 \neq \underline{v} \in \mathbb{R}^{n_h}$ beliebig aber fix, dann ist $v_h = \sum_{j=1}^{n_h} v_j\phi_j \in V_h = H_0^1(\Omega)$ ungleich der Nullfunktion und somit:

$$(M_h\underline{v},\underline{v})_{\ell^2} = ||v_h||^2_{L^2(\Omega)} > 0.$$

□

Für die Matrix B_h treffen wir die folgende Annahme:

Annahme 1. Die Matrix B_h aus (4.3) besitzt vollen Rang.

Die Annahme 1 ist von der Wahl der Räume V_h und Q_h abhängig.

Definition 4.10. *Wir bezeichnen das Vektorraumpaar V_h und Q_h als stabiles Element, falls es eine sogenannte diskrete* inf-sup*-Bedingung erfüllt, das heißt*

$$\exists C > 0 \quad \inf_{0 \neq p_h \in Q_h} \sup_{0 \neq v_h \in V_h^d} \frac{(\operatorname{div} v_h, p_h)_{L^2(\Omega)}}{||v_h||_{V^d} ||p_h||_Q} > C, \tag{4.6}$$

mit einer von der Feinheit h unabhängigen Konstante C.

Beispiel 4.11. 1. **Modifiziertes $P_1 - P_0$ Element**: Sei $\mathcal{T}_{2h}$ eine Triangulation von $\Omega \subset \mathbb{R}^2$ und weiters $\mathcal{T}_h$ die verfeinerte Triangulation, welche man bei gleichförmiger Zerlegung erhält: Jedes Dreieck $T \in \mathcal{T}_{2h}$ wird zerteilt in vier kongruente Dreiecke (siehe Abbildung 4.1). Auf diesen beiden Triangulationen seien nun die folgenden Räume definiert:

$$V_h = \{v \in C_0(\overline{\Omega}) : v|_T \in P_1 \ \forall T \in \mathcal{T}_h\} \subset V,$$
$$Q_h = \{q \in L_0^2(\Omega) : q|_T \in P_1 \ \forall T \in \mathcal{T}_H\} \subset Q,$$

wobei $H = 2h$.

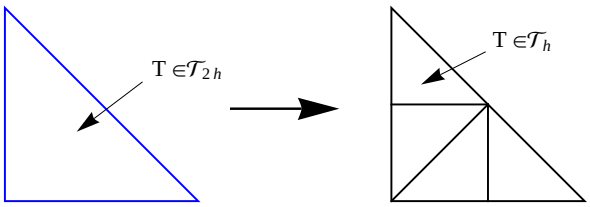

Abbildung 4.1: Gleichförmige Dreieckszerlegung

2. $P_2 - P_1$ **(Taylor-Hood) Element**: Dabei wählen wir auf einer Triangulation $\mathcal{T}_h$ von $\Omega \subset \mathbb{R}^2$ die Räume:

$$V_h = \{v \in C_0(\overline{\Omega}) : v|_T \in P_2 \ \forall T \in \mathcal{T}_h\} \subset V,$$
$$Q_h = \{q \in C(\overline{\Omega}) \cap L_0^2(\Omega) : q|_T \in P_1 \ \forall T \in \mathcal{T}_h\} \subset Q.$$

Für den Beweis der diskreten inf-sup-Bedingung und weiteren Beispielen für stetige Elemente siehe [1].

Für stabile Elemente lässt sich nun die Annahme 1 zeigen:

Satz 4.12. *Sei $h > 0$ und für das Raumpaar (V_h, Q_h) die diskrete* inf-sup*-Bedingung (4.6) erfüllt. Dann gilt:*

Die Matrix B_h aus (4.3) besitzt vollen Rang.

Beweis. Aus der Äquivalenz von Normen auf endlich dimensionalen Räumen erhält man

$$\exists C_1 > 0 \; \forall v_h \in V_h^d : \quad C_1 ||\underline{v}||_{\ell^2}^2 \leq ||v_h||_V^2 = (M_h \underline{v}, \underline{v})_{\ell^2}.$$

Weiters folgt aus der diskreten inf-sup-Bedingung (4.6),

$$\begin{aligned}
0 < C \leq & \inf_{0 \neq p_h \in Q_h} \sup_{0 \neq v_h \in V_h^d} \frac{(\operatorname{div} v_h, p_h)_{L^2(\Omega)}}{||v_h||_{V^d} ||p_h||_Q} \\
= & \inf_{0 \neq p_h \in Q_h} \sup_{0 \neq v_h \in V_h^d} \frac{(B_h \underline{v}, \underline{p})_{\ell^2}}{||v_h||_{V^d} ||p_h||_Q} \\
\leq & \inf_{0 \neq p_h \in Q_h} \sup_{0 \neq \underline{v} \in \mathbb{R}^{dn_h}} \frac{(\underline{v}, B_h^T \underline{p})_{\ell^2}}{C_1 ||\underline{v}||_{\mathbb{R}^{n_h}} ||p_h||_Q} \\
= & \inf_{0 \neq p_h \in Q_h} \frac{||B_h^T \underline{q}||_{\ell^2}}{||p_h||_Q} \\
& \Leftrightarrow \\
\forall 0 \neq \underline{p} \in \mathbb{R}^{m_h} : \; & ||B_h^T \underline{p}||_{\ell^2} > 0 \\
& \Leftrightarrow \\
\operatorname{kern}(B_h^T) = & \{0\}, \\
& \Leftrightarrow \\
\operatorname{rank}(B_h) = & \, m_h.
\end{aligned}$$

□

4.1.4 Das räumlich diskretisierte Optimalitätssystem

Sei im folgenden

$$\begin{aligned}
X_h := & W((0,T), V_h^d) \times L^2((0,T), U_h^d) \times L^2((0,T), Q_h) \\
& \times W((0,T), V_h^d) \times L^2((0,T), Q_h).
\end{aligned}$$

Das räumlich diskretisierte Optimalitätssystem (OS_h) erhält man nun durch Ersetzung von V bei V_h, Q bei Q_h und X bei X_h in (OS):

(OS_h) Räumlich diskretisiertes Optimalitätssystem

Gegeben:

- $v_d \in W((0,T), V^d)$,
- $\alpha > 0$.

Gesucht: $(v_h, f_h, p_h, \lambda_h, \mu_h) \in X_h$ sodass gilt:

1.

$$\begin{aligned} -(\partial_t \lambda_h, w_h)_{L^2} + (\nabla_x \lambda_h, \nabla_x w_h)_{L^2} & \\ +(\mathrm{div}_x\, w_h, \mu_h)_{L^2} &= (v_h - v_d, w_h)_{L^2}, \\ \forall w_h \in W((0,T), V_h^d), & \\ (\mathrm{div}_x\, \lambda_h, q_h)_{L^2} &= 0, \\ \forall q_h \in L^2((0,T), Q_h), & \end{aligned} \tag{4.7}$$

2.

$$\begin{aligned} (\partial_t v_h, w_h)_{L^2} + (\nabla_x v_h, \nabla_x w_h)_{L^2} & \\ +(\mathrm{div}_x\, w_h, p_h)_{L^2} &= -\alpha^{-1}(\lambda_h, w_h)_{L^2}, \\ \forall w_h \in W((0,T), V_h^d), & \\ (\mathrm{div}_x\, v_h, q_h)_{L^2} &= 0, \\ \forall q_h \in L^2((0,T), Q_h). & \end{aligned}$$

mit

$$\begin{aligned} v_h(0) &= 0, \\ \lambda_h(T) &= 0, \\ f_h &= \alpha^{-1}\lambda_h. \end{aligned}$$

4.2 Zeitliche Diskretisierung

Für die Diskretisierung nach der Zeit wollen wir eine unstetige Galerkin-Methode verwenden. Der folgende Abschnitt basierend auf der Literatur *„Nummerik gewöhnlicher Differentialgleichungen [10], Unterkapitel 2.10 , Seite 64 - 70“*, widmet sich dabei der Analyse dieser Verfahren.

4.2.1 Kurzeinführung in die unstetigen Galerkin-Methoden

Dieser Abschnitt soll neben der Analyse auch als Motivation gelten, für parabolische Probleme auch nach unstetigen Näherungs-Lösungen (bezüglich der Zeit) zu suchen. Die Verwendung dieser unstetigen Ansätze ermöglicht eine größere Anzahl von Freiheitsgraden. Außerdem stellt das unstetige Galerkin-Verfahren einen systematischen Zugang dar, um ausgehend von einer Finite Element Diskretisierung Verfahren höherer Ordnung für die Zeitdiskretisierung zu erhalten.

Wir betrachten dazu im Folgenden das Anfangswertproblem

Gesucht ist $v_h \in W(I, V_h^d)$ sodass gilt:

$$\begin{aligned} v_h'(t) &= f(t, v_h(t)) \ \dot{\forall} t \in I, \\ v_h(0) &= v_0, \end{aligned} \tag{4.8}$$

mit Zeitintervall $I = (0, T)$, V_h ein endlichdimensionaler Teilraum von $V = H_0^1(\Omega)$,

$$f \in L^2(I \times V_h^d, (V_h^d)^\star) \text{ und } v_0 \in V_h^d.$$

Beweistechnisch werden die folgenden beiden Aussagen über die variationelle Formulierung von Gleichungen von großer Bedeutung sein:

Satz 4.13 (Fundamental Lemma of Variational Calculus). *Sei $a < b \in \mathbb{R}$ und $f \in L^2([a, b])$, sodass*

$$\int_a^b f(t)\,\phi(t)\,\mathrm{d}t = 0,$$

für alle $\phi \in H^1([a, b])$ mit $\phi(a) = \phi(b) = 0$. Dann gilt:

$$f(t) = 0,$$

für fast alle $t \in [a, b]$.

Beweis. (siehe auch [11], Seite 1) Annahme die Folgerung ist falsch. Dann existiert mindestens ein Intervall der Form $[t_1, t_2] \subset [a, b]$, sodass o.B.d.A.

$$f(t) > 0 \quad \text{für fast alle } t \in [t_1, t_2].$$

Wählen wir nun die Funktion ϕ definiert bei

$$\phi(t) := \begin{cases} (t - t_1)(t_2 - t) & t \in [t_1, t_2], \\ 0 & \text{sonst}, \end{cases}$$

so ist also $\phi \in H^1([a, b])$ mit $\phi(a) = \phi(b) = 0$. Es gilt

$$\int_a^b f(t)\,\phi(t)\,\mathrm{d}t = \int_{t_1}^{t_2} (t - t_1)(t_2 - t) f(t)\,\mathrm{d}t > 0, \tag{4.9}$$

weil der Integrand positiv ist für fast alle $t \in [a, b]$. Aber (4.9) ist ein Widerspruch zu Annahme und so folgt die Behauptung. □

Korollar 4.14. *Sei $a < b \in \mathbb{R}$, V_n ein n-dimensionaler Hilbertraum mit $n \in \mathbb{N}$ und $f \in L^2([a, b], V_n^\star)$, sodass*

$$\int_a^b \langle f(t), \phi(t) \rangle\,\mathrm{d}t = 0,$$

für alle $\phi \in W([a, b], V_n)$ mit $\phi(a) = \phi(b) = 0$. Dann gilt:

$$f(t) = 0,$$

für fast alle $t \in [a, b]$.

Beweis. Da V_n zu $\mathbb{R}^n$ isomorph ist, genügt es die Aussage für $V_n = \mathbb{R}^n$ zu zeigen. Weiters lässt sich für $t \in [a,b]$ das Funktional $f(t) \in V_n$ wegen des *Riezschen Darstellungssatzes* (siehe Satz 3.14) durch seinen Repräsanten $F(t) \in V_n$ ersetzen, sodass

$$\begin{aligned}\int_a^b \langle f(t), \phi(t)\rangle \,\mathrm{d}t &= \int_a^b (F(t), \phi(t))_{\mathbb{R}^n} \,\mathrm{d}t \\ &= \sum_{i=1}^n \int_a^b F_i(t)\phi_i(t) \,\mathrm{d}t.\end{aligned} \tag{4.10}$$

Sei $i \in \{1,\dots,n\}$ beliebig, $\phi_\star \in H^1([a,b]) = W([a,b],\mathbb{R})$ mit $\phi(a) = \phi(b) = 0$ und $\phi_j = \delta_{ij}\phi_\star$ für $j = 1,\dots,n$, dann folgt aus (4.10),

$$\int_a^b \langle f(t), \phi(t)\rangle \,\mathrm{d}t = \int_a^b F_i(t)\phi_\star(t) \,\mathrm{d}t = 0.$$

Daraus erhalten wir weiters mit Satz 4.13, $F_i = 0$ fast überall in $[a,b]$. Womit wegen der Beliebigkeit von i, der Repräsentant F und somit f auf $[a,b]$ fast überall verschwindet. □

Variationelle Formulierung

Mit Hilfe des *Fundamental Lemma of Variational Calculus* (Satz 4.13) lässt sich nun das Anfangswertproblem (4.8) auch als *Variationsproblem* schreiben:

Gesucht ist $v_h \in W(I, V_h^d)$ mit $v_h(0) = v_0$ sodass gilt

$$\int_I \langle v_h'(t) - f(t, v_h(t)), \phi(t))\rangle \,\mathrm{d}t = 0, \quad \forall \phi \in W(I, V_h^d). \tag{4.11}$$

Wir nehmen nun eine Unterteilung des Zeitintervalls $I = (0,T)$,

$$\Delta_\tau \;:\; 0 = t_0 < t_1 < \cdots < t_N = T,$$

in nicht notwendig äquidistante Teilintervalle $I_k = (t_{k-1}, t_k)$ vor. Dabei bezeichne $\tau_k = t_k - t_{k-1}$ für $k = 1,\dots N$ die Länge der Teilintervalle und $\tau_{\max} = \max\limits_{k=1,\dots,N} \tau_k$ die maximale Teilintervalllänge. Auf der Unterteilung Δ_τ, definieren wir jetzt den Raum

$$W(\Delta_\tau, V_h^d) := \{v_h : I \to V_h^d \mid v_h|_{I_k} \in W(I_k, V_h^d),\ k = 1,\dots,N\}.$$

Führen wir weiters für $v_h \in W(\Delta_\tau, V_h^d)$ noch die Notation

$$v_{h,k} := v_h(t_k),$$

$$v_{h,k}^+ := \lim_{\substack{s\to 0\\ s>0}} v_h(t_k + s), \quad v_{h,k}^- := \lim_{\substack{s\to 0\\ s<0}} v_h(t_k + s), \quad [v_h]_k = v_{h,k}^+ - v_{h,k}^-,$$

mit $k = 0,\dots,N$ ein, dann lässt sich die folgende Aussage zeigen:

Satz 4.15. *Sei $(\cdot,\cdot)_{V_h}$ ein inneres Produkt auf V_h. Dann gilt: Das Anfangswertproblem (4.8) und das Variationsproblem*

Gesucht ist $v_h \in W(\Delta_\tau, V_h^d)$ mit $v_h(0) = v_{h,0}^- = v_0$ sodass gilt

$$\sum_{k=1}^{N}\left(\int_{I_k} \langle v_h'(t) - f(t, v_h(t)), \phi(t)\rangle \,\mathrm{d}t + ([v_h]_{k-1}, \phi_{k-1}^+)_{V_h^d}\right) = 0 \quad \forall \phi \in W(\Delta_\tau, V_h^d), \tag{4.12}$$

sind äquivalent.

Beweis. Wir zeigen eine Implikation in beide Richtungen:

- „⇒“ Falls v_h Lösung vom (4.8) ist, dann gilt $v_h'(t) - f(t, v_h(t)) = 0$ fast überall auf I. Aus der Grenzwertdefinition für $W(I, V_h^d)$-Funktionen folgt sofort $[v_h]_{k-1} = 0$ für $k \in \{1, \ldots, N\}$ und somit

$$\sum_{k=1}^{N}\left(\int_{I_k} \langle v_h'(t) - f(t, v_h(t)), \phi(t)\rangle \,\mathrm{d}t + ([v_h]_{k-1}, \phi_{k-1}^+)_{V_h^d}\right) = 0 \quad \forall \phi \in W(\Delta_\tau, V_h^d).$$

- „⇐“ Sei v_h Lösung von (4.12) und $k \in \{1, \ldots N\}$ beliebig aber fix, dann gilt:

$$\int_{I_k} \langle v_h'(t) - f(t, v_h(t)), \phi(t)\rangle \,\mathrm{d}t = 0,$$

für alle $\phi \in W(\Delta_\tau, V_h^d)$ mit $\phi(t) = 0$ für fast alle $t \in I \backslash (t_{k-1}, t_k)$. Aus Satz (4.13) folgt somit $v_h'(t) = f(t, v(t))$ auf I_k fast überall. Wählen wir weiters jene Testfunktionen aus $W(\Delta_\tau, (V_h)^d)$ der Form $\phi = \chi_{I_k} w$ mit $w \in V_h^d$, so erhält man aus (4.12)

$$([v_h]_{k-1}, \phi_{k-1}^+)_{V_h^d} = ([v_h]_{k-1}, w)_{V_h^d} = 0 \quad \forall w \in V_h^d,$$
$$\Leftrightarrow$$
$$v_{h,k}^+ = \lim_{\substack{s\to 0\\ s>0}} v_h(t_k + s) = \lim_{\substack{s\to 0\\ s<0}} v_h(t_k + s) = v_{h,k}^-,$$

womit auch die Gleichheit von v_h an den Teilintervallgrenzen gezeigt wäre.

□

Galerkin-Approximation

Wegen der in Satz (4.15) gezeigten Äquivalenz ist es also durchaus sinnvoll als Näherung für die Lösung des Anfangsrandwertproblems (4.8)

Gesucht ist $v_h \in W(I, V_h^d)$ sodass gilt:

$$v_h'(t) = f(t, v_h(t)) \; \dot{\forall} t \in I,$$
$$v_h(0) = v_0,$$

eine Lösung des diskretisierten Variationsproblems (4.12) zu bestimmen.

Bei der Diskretisierung approximiert man v_h durch stückweise polynomiale Funktionen, das heißt, durch Funktionen aus den Räumen

$$W^r(\Delta_\tau, V_h^d) := \{\phi \in W(\Delta_\tau, V_h^d) \mid \phi|_{I_k} \in \mathbb{P}_r(I_k, V_h^d), \; k = 1, \ldots, N \; \},$$

wobei $\mathbb{P}_r(I_k, V_h^d)$ die Polynome vom Grad kleiner oder gleich r auf I_k mit Werte in V_h^d bezeichnet.

Als **Galerkin-Approximation** für die Lösung von (4.12) sucht man also eine Funktion v_h, welche das folgende *Variationsproblem* löst

Gesucht ist $v_h \in W^r(\Delta_\tau, V_h^d)$ mit $v_h(t_0) = v_{h,0}^- = v_0$ sodass gilt

$$\sum_{k=1}^{N} \left(\int_{I_k} \langle v_h'(t) - f(t, v_h(t)), \phi(t) \rangle \, \mathrm{d}t + ([v_h]_{k-1}, \phi_{k-1}^+)_{V_h^d} \right) = 0 \quad \forall \phi \in W^r(\Delta_\tau, V_h^d). \tag{4.13}$$

Das Variationsproblem (4.13) kann als sukzessives Zeitschrittverfahren geschrieben werden,

$$\int_{I_k} \langle v_h'(t), \phi(t) \rangle \, \mathrm{d}t + (v_{h,k-1}^+, \phi_{k-1}^+)_{V_h^d} = \int_{I_k} \langle f(t, v_h(t))), \phi(t) \rangle \, \mathrm{d}t + (v_{h,k-1}^-, \phi_{k-1}^+)_{V_h^d},$$
$$\forall \phi \in \mathbb{P}_r(I_k, V_h^d), \quad k = 1, \ldots, N, \tag{4.14}$$

wobei $v_{h,0}^- = v_{h,0} = v_0$. Man bezeichnet (4.13) und (4.14) als **unstetiges Galerkin-Verfahren**.

Im Gegensatz zu den **stetigen Galerkin-Verfahren** wo die Näherungslösung aus demselben Raum wie die exakte Lösung gewählt wird, ist dies bei den **unstetigen Galerkin-Verfahren** nicht mehr der Fall, denn für $r \in \mathbb{N}_0$ gilt:

$$W^r(\Delta_\tau, V_h^d) \not\subset W((0, T), V^d).$$

Je nach Wahl des Polynomgrades r erhält man eine andere Variante eines unstetigen Galerkin-Verfahrens:

Beispiel 4.16. (*Implizites Euler-Verfahren*) Fall: $r = 0$. Setzt man $v_{h,k} = v_{h,k}^-$ konstant auf $I_k = (t_{k-1}, t_k]$, dann gilt, da $v_h'(t) = 0$ auf I_k:

$$\sum_{k=1}^{N}\Bigg(\underbrace{\int_{I_k}\langle v_h'(t),\phi(t)\rangle\,\mathrm{d}t}_{=0 \text{ da } v_h|_{I_k}=\text{const}}+(v_{h,k-1}^+,\phi_{k-1}^+)_{V_h^d}\Bigg)=\sum_{k=1}^{N}\Bigg(\int_{I_k}\langle f(t,v_h(t)),\phi(t)\rangle\,\mathrm{d}t + (v_{h,k-1}^-,\phi_{k-1}^+)_{V_h^d}\Bigg),$$
$$\forall\phi\in W^0(\Delta_\tau,V_h^d),$$

$$\Leftrightarrow$$

$$(v_{h,k-1}^+,\phi_{k-1}^+)_{V_h^d}=\int_{I_k}\langle f(t,v_h(t)),\phi(t)\rangle\,\mathrm{d}t+(v_{h,k-1}^-,\phi_{k-1}^+)_{V_h^d},$$
$$\forall\phi\in\mathbb{P}_0(I_k,V_h^d),\quad k=1,\dots,N,$$

$$\Leftrightarrow$$

$$(v_{h,k},\phi_k)_{V_h^d}=\int_{I_k}\langle f(t,v_{h,k}),\phi(t_k)\rangle\,\mathrm{d}t+(v_{h,k-1},\phi_k)_{V_h^d},$$
$$\forall\phi\in\mathbb{P}_0(I_k,V_h^d),\quad k=1,\dots,N,$$

$$\Leftrightarrow$$

$$(v_{h,k}-v_{h,k-1},v)_{V_h^d}=\int_{I_k}\langle f(t,v_{h,k}),w\rangle\,\mathrm{d}t$$
$$\forall w\in V_h,\ k=1,\dots,N,$$

$$\Leftrightarrow$$

$$v_{h,k}-v_{h,k-1}=\int_{I_k}\langle f(t,v_{h,k}),\cdot\rangle\,\mathrm{d}t$$
$$k=1,\dots,N,$$

wobei $v_{h,0}=v_0^-=v_0$. Besitzt f die spezielle Form

$$f(t,v_h(t))=f(v_h(t)),$$

dann vereinfacht sich das *implizite Euler-Verfahren* zu

$$\frac{v_{h,k} - v_{h,k-1}}{\tau_k} = \langle f(v_{h,k}), \cdot \rangle \quad k = 1, \ldots, N,$$

mit $v_{h,0} = v_0$.
Damit sind wir am Ende der kurzen Einführung in die unstetigen Galerkin-Verfahren angelangt. Für weitere Beispiele und a priori bzw. a posteriori Fehlerabschätzungen verweisen wir an dieser Stelle auf [10].

Anwendung einer unstetigen Galerkin-Methode

Für die zeitliche Diskretisierung soll nun das implizite Euler-Verfahren als Beispiel eines unstetigen Galerkin-Verfahrens verwendet werden, das heißt wir suchen nach zeitlich stückweise konstanten Lösungen. Dazu nehmen wir eine Unterteilung des Zeitintervalls $I = (0, T)$

$$\Delta_\tau \ : \ 0 = t_0 < t_1 = \frac{1}{N} < t_2 = \frac{2}{N} < \cdots < t_N = T,$$

in äquidistante Teilintervalle $I_k = (t_{k-1}, t_k] = (\frac{k-1}{N}, \frac{k}{N}]$ vor. In unserem Fall ist also die Schrittweite τ_k konstant, sodass gilt:

$$\tau = \tau_k = \frac{1}{N} \text{ für } k = 1, \ldots N.$$

Für das räumlich und zeitlich diskretisierte Optimalitätssystem wenden wir jetzt das implizite Euler-Verfahrens auf (OP_h) an. Die Durchführung des impliziten Euler-Verfahrens wurde aber nur an Hand von Anfangswertprobleme diskutiert, deshalb transformieren zunächst das Endwertproblem (4.7) auf ein Anfangswertproblem:

1.

$$\begin{aligned}
(\partial_t \gamma_h, w_h)_{L^2} + (\nabla_x \gamma_h, \nabla_x w_h)_{L^2} & \\
+(\mathrm{div}_x\, w_h, \eta_h)_{L^2} &= (v_h(T - \cdot) - v_d(T - \cdot)), w_h)_{L^2}, \\
\forall w_h &\in W((0,T), V_h^d), \\
(\mathrm{div}_x\, \gamma_h, q_h)_{L^2} &= 0, \\
\forall q_h &\in L^2((0,T), Q_h), \\
\text{mit } \gamma_h(0) &= 0,
\end{aligned} \tag{4.15}$$

2.

$$\begin{aligned}
(\partial_t v_h, w_h)_{L^2} + (\nabla_x v_h, \nabla_x w_h)_{L^2} & \\
+(\mathrm{div}_x\, w_h, p_h)_{L^2} &= -\alpha^{-1}(\gamma_h(T - \cdot), w_h)_{L^2}, \\
\forall w_h &\in W((0,T), V_h^d), \\
(\mathrm{div}_x\, v_h, q_h)_{L^2} &= 0, \\
\forall q_h &\in L^2((0,T), Q_h), \\
\text{mit } v_h(0) &= 0,
\end{aligned} \tag{4.16}$$

mit

$$\begin{aligned} f_h &= \alpha^{-1}\lambda_h, \\ \gamma_h(t) &= \lambda(T-t), \quad \dot\forall t \in I, \\ \eta_h(t) &= \mu(T-t) \quad \dot\forall t \in I. \end{aligned}$$

Als nächstes erfolgt die zeitliche Diskretisierung mit dem impliziten Euler-Verfahren an Hand von (4.15):

Schritt 1: Formulierung als Variationsporblem der Form (4.8):

$$\begin{aligned} \gamma_h'(t) =& f_1(t,\gamma_h(t)), \quad \dot\forall t \in I, \\ 0 =& f_2(t), \quad \dot\forall t \in I, \\ \gamma_h(0) =& 0, \end{aligned} \tag{4.17}$$

mit

$$\begin{aligned} \langle f_1(t,\gamma_h(t)), w_h\rangle :=& (v_h(T-t) - v_d(T-t), w_h)_{L^2} - (\nabla_x u_1, \nabla_x w_h)_{L^2} \\ & - (\operatorname{div}_x w_h, \eta_h(t))_{L^2} \quad \forall q_h \in (V_h^d) \\ \langle f_2(t), q_h\rangle :=& (\operatorname{div}_x \gamma_h(t), q_h)_{L^2} \quad \forall q_h \in Q_h. \end{aligned}$$

Schritt 2: Wahl von $(\cdot,\cdot)_{V_h}$:

Wir setzen

$$(\cdot,\cdot)_{V_h} = (\cdot,\cdot)_{L^2},$$

da $V_h \subset V = H_0^1(\Omega)$ ausgestattet mit $(\cdot,\cdot)_{L^2}$ wegen der endlichen Dimension von V_h wiederum einen Hilbertraum bildet.

Schritt 3: Anwendung des implizites Euler-Verfahren auf (4.17):

$$\begin{aligned} \sum_{k=1}^{N}\left(\int_{I_k} \langle \gamma'(t) - f_1(t,\gamma(t)), \phi(t)\rangle \,\mathrm{d}t + ([\gamma]_{k-1}, \phi_{k-1}^+)_{L^2}\right) & \\ =0, & \\ \forall \phi \in W_0(\Delta_\tau, V_h^d), & \\ \sum_{k=1}^{N}\left(\int_{I_k} \langle f_2(t,\gamma(t)), \phi(t)\rangle \,\mathrm{d}t\right) & \\ =0, & \\ \forall \phi \in W_0(\Delta_\tau, Q_h), & \\ \gamma_h(0) =0, & \end{aligned}$$

$$\Leftrightarrow$$

$$\begin{aligned}
\frac{\gamma_{h,k}-\gamma_{h,k-1}}{\tau_k} &= \int_{I_k} \langle f_1(t,\gamma_k(t)),\cdot\rangle \,\mathrm{d}t,\\
k &= 1,\dots,N,\\
0 &= \int_{I_k} \langle f_2(t),\cdot\rangle \,\mathrm{d}t,\\
k &= 1,\dots,N,\\
\gamma_{h,0} &= 0,
\end{aligned}$$

$$\Leftrightarrow$$

$$\begin{aligned}
\frac{(\gamma_{h,k}-\gamma_{h,k-1},w_h)_{L^2}}{\tau_k} &= \int_{I_k} \langle f_1(t,\gamma_k(t)),w_h\rangle \,\mathrm{d}t,\\
\forall w_h &\in V_h^d, \quad k=1,\dots,N,\\
0 &= \int_{I_k} \langle f_2(t),q_h\rangle \,\mathrm{d}t,\\
\forall q_h &\in Q_h, \quad k=1,\dots,N,\\
\gamma_{h,0} &= 0,
\end{aligned}$$

$$\Leftrightarrow$$

$$\begin{aligned}
\frac{(\gamma_{h,k}-\gamma_{h,k-1},w_h)_{L^2}}{\tau_k} &= (v_{h,N-(k-1)} - \boldsymbol{v_{d,h}}(t_{N-(k-1)}),w_h)_{L^2}\\
&\quad - (\nabla_x \gamma_{h,k},\nabla_x w_h)_{L^2} - (\operatorname{div}_x w_h,\eta_{h,k})_{L^2}\\
\forall w_h &\in (V_h)^d, \quad k=1,\dots,N,\\
(\operatorname{div}_x \gamma_{h,k},q_h) &= 0,\\
\forall q_h &\in Q_h, \quad k=1,\dots,N,\\
\gamma_{h,0} &= 0,
\end{aligned} \tag{4.18}$$

$$\Leftrightarrow$$

$$\begin{aligned}
\frac{(C_h(\underline{\gamma}_k-\underline{\gamma}_{k-1}),\underline{w})_{\ell^2}}{\tau_k} &= (C_h(\underline{v}_{N-(k-1)} - \underline{v}_{d,N-(k-1)}),\underline{w})_{\ell^2}\\
&\quad - (A_h\underline{\gamma},\underline{w})_{\ell^2} - (B_h\underline{w},\underline{\eta})_{\ell^2}\\
\forall \underline{w} &\in \mathbb{R}^{dn_h}, \quad k=1,\dots,N,\\
(B_h\underline{\gamma}_k,\underline{w})_{\ell^2} &= 0,\\
\forall \underline{q} &\in \mathbb{R}^{m_h}, \quad k=1,\dots,N,\\
\underline{\gamma}_0 &= 0,
\end{aligned}$$

$$\Leftrightarrow$$

$$\begin{aligned} -\frac{1}{\tau}C_h\underline{\gamma}_{k-1} + (\frac{1}{\tau}C_h + A_h)\underline{\gamma}_k + B_h^T\underline{\eta}_k - C_h\underline{v}_{N-(k-1)} &= -C_h\underline{v}_{d,N-(k-1)}, \\ k &= 1,\dots,N, \\ B_h\underline{\gamma}_k &= 0, \\ k &= 1,\dots,N, \\ \underline{\gamma}_0 &= 0. \end{aligned} \tag{4.19}$$

gesetzt.

Um mit v_d das letzte Objekt aus einem unendlich dimensionalen Raum aus dem Optimalitätssystem zu entfernen, wurde dieses in (4.18) durch $v_{d,h} \in W^0(\Delta_\tau, V_h^d)$ mit

$$(v_{d,h}, w)_{L^2} = (v_d, w)_{L^2} \quad \forall w \in W^0(\Delta_\tau, V_h^d), \tag{4.20}$$

ersetzt. Die Existenz und Eindeutigkeit von $v_{d,h}$ als Lösung von (4.20), folgt wiederum mit dem *Rieszschen Darstellungssatz* (siehe Satz 3.14) und der Tatsache, dass für den endlichdimensionalen Raum $V_h \subset V \subset L^2(\Omega)$, $W^0(\Delta_\tau, V_h^d)$ mit $(\cdot,\cdot)_{L^2}$ einen Hilbertraum bildet.

Analog lässt sich das implizite Euler-Verfahren auf die Gleichung (4.16) anwenden:

$$\begin{aligned} -\frac{1}{\tau}C_h\underline{v}_{k-1} + (\frac{1}{\tau}C_h + A_h)\underline{v}_k + B_h^T\underline{p}_k & \\ + \alpha^{-1}C_h\underline{\gamma}_{N-(k-1)} &= 0, \\ B_h\underline{v}_k &= 0, \\ k &= 1,\dots,N, \\ \underline{v}_0 &= 0. \end{aligned} \tag{4.21}$$

4.2.2 Das räumlich und zeitlich diskretisierte Optimalitätssystem

Das räumlich und zeitlich diskretisierte Optimalitätssystem besteht also aus den beiden gekoppelten linearen Systeme (4.19) und (4.21). Geschrieben in Form eines linearen Gleichungssystems, erhält man:

($\mathbf{OS}_{h,\tau}$) Räumlich und zeitlich diskretisiertes Optimalitätssystem

Gegeben:

- $v_{d,h} \in W^0(\Delta_\tau, V_h^d)$,
- $\alpha > 0$,

- $\underline{v}_0 = 0, \underline{\lambda}_N = 0.$

Gesucht: $\underline{x} \in \mathbb{R}^{2N(dn_h+m_h)}$ sodass gilt:

$$\begin{pmatrix} -C & 0 & K^T & B^T \\ 0 & 0 & B & 0 \\ K & B^T & \frac{1}{\alpha}C & 0 \\ B & 0 & 0 & 0 \end{pmatrix} \underline{x} = \underline{b}, \tag{4.22}$$

wobei

- $$K = \begin{pmatrix} \frac{1}{\tau}C_h + A_h & 0 & 0 & \dots & 0 & 0 \\ -\frac{1}{\tau}C_h & \frac{1}{\tau}C_h + A_h & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -\frac{1}{\tau}C_h & \frac{1}{\tau}C_h + A_h \end{pmatrix} \in \mathbb{R}^{Ndn_h \times Ndn_h},$$

  $$C = \begin{pmatrix} C_h & 0 & \dots & 0 \\ 0 & C_h & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & C_h \end{pmatrix} \in \mathbb{R}^{Nn_h \times Ndn_h} \quad \text{und} \quad B = \begin{pmatrix} B_h & 0 & \dots & 0 \\ 0 & B_h & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & B_h \end{pmatrix} \in \mathbb{R}^{Nm_h \times Ndn_h}. \tag{4.23}$$

- $$\underline{x} = \begin{pmatrix} \underline{x}_1 \\ \underline{x}_2 \end{pmatrix} \in \mathbb{R}^{2N(dn_h+m_h)},$$

  mit

  $$\underline{x}_1 = \begin{pmatrix} \underline{v}_1 \\ \vdots \\ \underline{v}_N \\ \underline{p}_0 \\ \vdots \\ \underline{p}_N \end{pmatrix} \quad \text{und} \quad \underline{x}_2 = \begin{pmatrix} \underline{\gamma}_N \\ \vdots \\ \underline{\gamma}_1 \\ \underline{\eta}_N \\ \vdots \\ \underline{\eta}_0 \end{pmatrix} = \begin{pmatrix} \underline{\lambda}_0 \\ \vdots \\ \underline{\lambda}_{N-1} \\ \underline{\mu}_0 \\ \vdots \\ \underline{\mu}_N \end{pmatrix}.$$

- $$\underline{b} = \begin{pmatrix} \underline{b}_1 \\ \underline{b}_2 \end{pmatrix} \in \mathbb{R}^{2N(dn_h+m_h)},$$

mit

$$\underline{b}_1 = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -C_h \underline{v}_{d,1} \\ \vdots \\ -C_h \underline{v}_{d,N} \end{pmatrix} \quad \text{und} \quad \underline{b}_2 = 0.$$

Wir nutzen nun die Möglichkeit, dass sich die Matrix aus (4.22), in eine sogenannte **Sattelpunktsform**

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}, \tag{4.24}$$

transformieren lässt. Dazu führen wir folgende Matrixoperationen an (4.22) durch:

1. Vertauschen der 1. Zeile mit der 2. Zeile:

$$\begin{pmatrix} 0 & 0 & B & 0 \\ -C & 0 & K^T & B^T \\ K & B^T & \frac{1}{\alpha} C & 0 \\ B & 0 & 0 & 0 \end{pmatrix}.$$

2. Vertauschen der 1. Spalte mit der 2. Spalte:

$$\begin{pmatrix} 0 & 0 & B & 0 \\ 0 & -C & K^T & B^T \\ B^T & K & \frac{1}{\alpha} C & 0 \\ 0 & B & 0 & 0 \end{pmatrix}.$$

3. Vertauschen der 1. Zeile mit der 3. Zeile:

$$\begin{pmatrix} B^T & K & \frac{1}{\alpha} C & 0 \\ 0 & -C & K^T & B^T \\ 0 & 0 & B & 0 \\ 0 & B & 0 & 0 \end{pmatrix}.$$

4. Vertauschen der 1. Spalte mit der 3. Spalte:

$$\begin{pmatrix} \frac{1}{\alpha} C & K & B^T & 0 \\ K^T & -C & 0 & B^T \\ B & 0 & 0 & 0 \\ 0 & B & 0 & 0 \end{pmatrix}.$$

5. Multiplikation der 1. Zeile und 1. Spalte mit Faktor $\sqrt{\alpha}$:

$$\begin{pmatrix} C & \sqrt{\alpha}K & \sqrt{\alpha}B^T & 0 \\ \sqrt{\alpha}K^T & -C & 0 & B^T \\ \sqrt{\alpha}B & 0 & 0 & 0 \\ 0 & B & 0 & 0 \end{pmatrix}.$$

6. Division der 3. Zeile und 3. Spalte durch den Faktor $\sqrt{\alpha}$:

$$\begin{pmatrix} C & \sqrt{\alpha}K & B^T & 0 \\ \sqrt{\alpha}K^T & -C & 0 & B^T \\ B & 0 & 0 & 0 \\ 0 & B & 0 & 0 \end{pmatrix}.$$

Werden die Transformationen auch am Lösungsvektor $\underline{x}$ und der rechten Seite $\underline{b}$ aus (4.22) berücksichtigt, so lässt sich $(\mathrm{OP})_{h,\tau}$ weiters schreiben:

$(\mathbf{OS}_{h,\tau})$

Gegeben:

- $v_{d,h} \in W^0(\Delta_\tau, V_h^d)$,
- $\alpha > 0$,
- $\underline{v}_0 = 0, \underline{\lambda}_N = 0$.

Gesucht: $\underline{x} \in \mathbb{R}^{2N(dn_h+m_h)}$ sodass gilt:

$$\mathcal{A}\underline{x} = \underline{b}, \tag{4.25}$$

wobei

- $$\mathcal{A} := \begin{pmatrix} A_1 & B_1^T \\ B_1 & 0 \end{pmatrix}, \tag{4.26}$$

  mit

  $$A_1 = \begin{pmatrix} C & \sqrt{\alpha}K \\ \sqrt{\alpha}K^T & -C \end{pmatrix} \in \mathbb{R}^{Ndn_h \times Ndn_h}, \quad B_1 = \begin{pmatrix} B & 0 \\ 0 & B \end{pmatrix} \in \mathbb{R}^{Nm_h \times Ndn_h} \tag{4.27}$$

  und K, C und B aus (4.23).

- $$\underline{x} = \begin{pmatrix} \underline{x}_1 \\ \underline{x}_2 \end{pmatrix} \in \mathbb{R}^{2N(dn_h+m_h)},$$

  mit

  $$\underline{x}_1 = \begin{pmatrix} \alpha^{-1}\underline{v}_1 \\ \vdots \\ \alpha^{-1}\underline{v}_N \\ \underline{p}_0 \\ \vdots \\ \underline{p}_N \end{pmatrix} \quad \text{und} \quad \underline{x}_2 = \begin{pmatrix} \alpha\underline{\lambda}_0 \\ \vdots \\ \alpha\underline{\lambda}_{N-1} \\ \underline{\mu}_0 \\ \vdots \\ \underline{\mu}_N \end{pmatrix}.$$

- $$\underline{b} = \begin{pmatrix} \underline{b}_1 \\ \underline{b}_2 \end{pmatrix} \in \mathbb{R}^{2N(dn_h+m_h)}$$

  mit

  $$\underline{b}_1 = \begin{pmatrix} -C_h\underline{v}_{d,1} \\ \vdots \\ -C_h\underline{v}_{d,N} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{und} \quad \underline{b}_2 = 0.$$

4.2.3 Eigenschaften von $\mathcal{A}$

Für die Matrix $\mathcal{A}$ lassen sich nun die folgenden Eigenschaften zeigen:

Satz 4.17. *Sei $\alpha > 0, \tau > 0, h > 0$ und Annahme 1 erfüllt. Dann gilt für die Matrix $\mathcal{A}$ aus (4.26) und ihre Bestandteile aus (4.23) und (4.27):*

1. *Die Matrix K ist regulär, die Matrix B besitzt vollen Rang und die Matrix C ist symmetrisch und positiv definit.*
2. *Die Matrix A_1 ist regulär, symmetrisch und indefinit und die Matrix B_1 besitzt vollen Rang.*
3. *Die Matrix $\mathcal{A}$ ist regulär, symmetrisch und indefinit.*

Der Beweis dieser Aussage benötigt jedoch etwas Vorbereitung:

Wir starten mit einer Aussage über den Zusammenhang zwischen den Eigenwerten *zueinander kongruenter* Matrizen:

Satz 4.18 (Trägheitssatz von Sylvester). *Sei $n \in \mathbb{N}$ und $\mathcal{S}, \mathcal{T} \in \mathbb{R}^{n\times n}$ symmetrische und zueinander kongurente Matrizen, das heißt*

$$\exists \mathcal{X} \in \mathbb{R}^{n\times n} \text{ regulär}: \quad \mathcal{T} = \mathcal{X}\mathcal{S}\mathcal{X}^T.$$

Dann gilt:

> *Die Matrizen $\mathcal{S}$ und $\mathcal{T}$ besitzen die selbe Anzahl von positiven und negativen Eigenwerten.*

Für den Beweis siehe zum Beispiel [12]. Es sei an dieser Stelle noch bemerkt, dass symmetrische Matrizen nur reelle Eigenwerte besitzen, also die Aussage des letzten Satzes wohl definiert ist.

Der nächste Satz gibt uns Auskunft über die Kongruenz zwischen Blockmatrizen der Gestalt (4.24) und Matrizen mit Diagonal-Blockstruktur.

Satz 4.19. *Sei*

- $n, m \in \mathbb{N}$,
- $A \in \mathbb{R}^{n\times n}$ *symmetrisch und regulär,*
- $B \in \mathbb{R}^{m\times n}$,
- $C \in \mathbb{R}^{m\times m}$ *symmetrisch,*

dann gilt:

> *Die Matrix*
>
> $$\mathcal{T} = \begin{pmatrix} A & B^T \\ B & -C \end{pmatrix}$$
>
> *ist symmetrisch und kongruent zu*
>
> $$\begin{pmatrix} A & 0 \\ 0 & -S \end{pmatrix} \quad \text{wobei} \quad S = C + BA^{-1}B^T.$$

Beweis. Die Symmetrie von $\mathcal{S}$ erhält man unter Verwendung der Symmetrie von A und C durch nachrechnen.

Weiters lässt sich durch nachrechnen zeigen, dass die Matrix $\mathcal{T}$ dargestellt werden kann als

$$\mathcal{T} = \mathcal{X} \begin{pmatrix} A & 0 \\ 0 & -S \end{pmatrix} \mathcal{X}^T,$$

mit

$$S = C + BA^{-1}B^T,$$

und

$$\mathcal{X} = \begin{pmatrix} I & 0 \\ BA^{-1} & I \end{pmatrix}.$$

Aus der Regularität von $\mathcal{X}$ folgt dann die Behauptung. □

Als Folgerungen der letzten beiden Sätze erhalten wir:

Korollar 4.20. *Es seien die Voraussetzung von Satz 4.19 für die Matrizen A, B und C erfüllt und weiters gelte*

- *$n = m$,*
- *A sei positiv definit,*
- *B besitze vollen Rang,*
- *C sei positiv semi-definit.*

Dann gilt:

Die Matrix

$$\mathcal{T} = \begin{pmatrix} A & B^T \\ B & -C \end{pmatrix}$$

ist regulär, symmetrisch und indefinit.

Beweis. Offensichtlich ist die Matrix $\mathcal{T}$ symmetrisch. Unter den gegebenen Voraussetzungen ist die Matrix $S = C + BA^{-1}B^T$ symmetrisch und positiv definit. Nach Satz 4.19 ist die Matrix $\mathcal{T}$ kongruent zu einer regulären, indefiniten Matrix und wegen dem *Trägheitssatz von Sylvester* (Satz 4.18), deshalb selbst regulär und indefinit. □

Korollar 4.21. *Es seien die Voraussetzung von Satz 4.19 für die Matrizen A, B und C erfüllt und weiters gelte*

- *A sei indefinit,*
- *B besitze vollen Rang,*

Dann gilt:

Die Matrix

$$\mathcal{T} = \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}$$

ist regulär, symmetrisch und indefinit.

Beweis. Der Nachweis der Symmetrie und Indefinitness von $\mathcal{T}$ erfolgt analog zum vorhergehenden Korrollar.

Die Regularität von $\mathcal{T}$ werden wir nun unter Verwendung des Satzes von Brezzi (Satz 7.12, siehe Seite 99) zeigen: Dazu wählen wir die Hilberträume $X = \mathbb{R}^n$ und $Y = \mathbb{R}^m$ ausgesstattet mit dem ℓ_2-inneren Produkt und definieren weiters die linearen Operatoren

$$\langle Av, w\rangle := (Av, w)_{\ell^2} \quad \text{und} \quad \langle Bw, q\rangle := (Bw, q)_{\ell^2} \quad \forall v, w \in \mathbb{R}^n, \forall q \in \mathbb{R}^m. \tag{4.28}$$

Wegen

$$\begin{aligned} A \text{ regulär und symmetrisch} \Leftrightarrow \exists \alpha_1 > 0 : \inf_{v\in X} \sup_{w\in X} \frac{(Av, w)_{\ell^2}}{||v||_X||w||_Y} \\ = \inf_{v\in X} \sup_{w\in X} \frac{(v, Aw)_{\ell^2}}{||v||_X||w||_Y} \geq \alpha_1, \\ B \text{ besitzt vollen Rang} \Leftrightarrow \exists \beta_1 > 0 : \inf_{q\in Y} \sup_{w\in X} \frac{(Bw, q)_{\ell^2}}{||w||_X||q||_Y} \geq \beta_1, \end{aligned}$$

und

$$\begin{aligned} |\langle Av, w\rangle| \leq & \alpha_2||v||_X||w||_Y \quad \forall v, w \in \mathbb{R}^n, \text{ mit } \alpha_2 = ||A||_{\ell_2}, \\ |\langle Bw, q\rangle| \leq & \beta_2||w||_X||q||_Y \quad \forall w \in \mathbb{R}^n, \forall q \in \mathbb{R}^m, \text{ mit } \beta_2 = ||B||_{\ell_2}, \end{aligned}$$

sind dann alle Voraussetzungen vom Satz von Brezzi für die linearen Operatoren A und B aus (4.28) erfüllt. Die lineare Gleichung

$$\mathcal{T}x = b,$$

besitzt also für eine beliebige rechte Seite $b \in \mathbb{R}^{n+m}$ eine eindeutige Lösung $x \in \mathbb{R}^{n+m}$, das heißt die Matrix $\mathcal{T}$ ist regulär. □

Mit Hilfe der letzten beiden Resultate sind wir nun im Stande Satz 4.17 zu beweisen:

Beweis von Satz 4.17: Aus Satz 4.9 und Annahme 1 erhalten wir,

$$A_h, C_h \text{ sind s.p.d. und } B_h \text{ besitzt vollen Rang.} \tag{4.29}$$

Nun gilt:

1. Die Aussagen für die Matrizen C und B folgen sofort aus (4.29) und der Diagonal-Blockstruktur. Die Matrix K besitzt eine untere Dreiecks-Blockstruktur wobei die Blöcke in der Diagonale

$$\frac{1}{\tau}C_h + A_h,$$

wegen (4.29) symmetrisch und positiv definit sind. Aus

$$\det(K) = \underbrace{(\det(\frac{1}{\tau}C_h + A_h))^N}_{>0} > 0,$$

folgt nun die Regularität von K.

2. Der volle Rang der Matrix B_1 folgt aus dem vollen Rang der Matrix B und der Diagonal-Blockstruktur. Die Symmetrie und Indefinitness der Matrix A_1 folgt aus Korollar 4.20, wobei die Voraussetzungen wegen 1. und $\alpha > 0$ erfüllt sind.

3. Die Behauptung für die Matrix $\mathcal{A}$ folgt schließlich aus den Aussagen von 2. und Korollar 4.21.

□

Die Berechnung einer Lösung des diskretisierten Optimalitätssystem ($\mathrm{OP}_{h,\tau}$), erfordert also die Lösung des linearen Gleichungssystems (4.25)

$$\mathcal{A}\underline{x} = \underline{b},$$

mit **großer, dünnbesetzter, regulärer, symmetrisch und indefiniter Blockmatrix** $\mathcal{A}$ in **Sattelpunktsform**.

Kapitel 5

Der zeitharmonische Fall

5.1 Problemstellung und Problemformulierung

In diesem Kapitel widmen wir uns dem zeitharmonischen Fall, das heißt wir betrachten wiederum die Problemstellung aus 2.1, jedoch nun mit der zusätzlichen Information, dass die Erregung des Fluids, zum Beispiel durch ein magnetisches Feld, zeitharmonisch erfolgt.

Im Unterschied zum zeitabhängigen Fall, wird hier **keine Anfangsbedinung** für das gesuchte optimale Geschwindigkeitsfeld v vorgegeben.

Starke Formulierung:

Mathematisch lässt sich unser Problem wiederum in Form eines optimalen Kontrollproblems unbeschränkter Steuerung schreiben:

($\mathbf{OP}^{\omega}$) Optimales Kontrollproblem im zeitharmonischen Fall

Gegeben

- Ortsgebiet $\Omega \subset \mathbb{R}^d$ mit $d \in \{1, 2, 3\}$,
- Zeitintervall $I = (0, T)$,
- Geschwindigkeitsfeld $v_d \in C^1([0, T], (C^2(\Omega))^d)$ mit $v_d(0) = v_d(T)$,
- **Kontrollparameter** $\alpha > 0$.

Finde

- **Zustand** $v \in C^1([0, T], (C^2(\overline{\Omega}))^d)$ mit $v(0) = v(T)$ (Geschwindigkeitsfeld),
- **Zustand** $p \in C((0, T), C^1(\overline{\Omega}))$ mit $p(0) = p(T)$ (Druck),

- **Steuerung** $f \in C((0,T),(C(\overline{\Omega}))^d)$ mit $f(0) = f(T)$ (Volumenkraftdichte),

welche die **Zielfunktion**

$$\begin{aligned} J(v,f) :=& \frac{1}{2}\int_0^T \int_\Omega |v(x,t) - v_d(x,t)|^2 \,\mathrm{d}x\,\mathrm{d}t + \frac{\alpha}{2}\int_0^T \int_\Omega |f(x,t)|^2 \,\mathrm{d}x\,\mathrm{d}t \\ =& \frac{1}{2}||v - v_d||^2_{L^2((0,T),(L^2(\Omega))^d)} + \frac{\alpha}{2}||f||^2_{L^2((0,T),(L^2(\Omega))^d)}, \end{aligned}$$

minimiert, sodass die **Zustandsgleichungen**

$$\begin{aligned} \frac{\partial v}{\partial t} - \Delta_x v + \nabla_x p &= f \quad \text{in } Q = \Omega \times (0,T), \\ \operatorname{div}_x v &= 0 \quad \text{in } Q = \Omega \times (0,T), \\ v &= 0 \quad \text{auf } \Sigma = \Gamma \times (0,T), \end{aligned} \tag{5.1}$$

erfüllt sind.

Bemerkung 5.1. *Für den weiteren Verlauf seien die in Bemerkung 2.1 gestellten Annahmen auch für das Problem (OP^ω) erfüllt.*

Unter der Voraussetzung das $v, f, p \in C([0,T], \dots)$ eine beschränkte Variation besitzen, können diese punktweise in eine Fourierreihe entwickelt werden (siehe [13], Seite 141), das heißt z.B. $v \in C^1([0,T],(C^2(\overline{\Omega}))^d)$ lässt sich schreiben in der Form

$$v(x,t) = \sum_{i=0}^{\infty} (v_c^i(x)\cos(\omega_i t) + v_s^i(x)\sin(\omega_i t)),$$

*mit $v_c^i, v_s^i \in (C^2(\Omega))^d$ für $i \in \mathbb{N}_0$ und $N \in \mathbb{N}$. Als weitere Forderung an (OP^ω), nehmen wir an, dass die Fourierreihendarstellung für v, f, p und v_d **endlich** ist, also z.B. gelte für $v \in C^1([0,T],(C^2(\overline{\Omega}))^d)$*

$$v(x,t) = \sum_{i=0}^{N} (v_c^i(x)\cos(\omega_i t) + v_s^i(x)\sin(\omega_i t))$$

mit $v_c^i, v_s^i \in (C^2(\Omega))^d$ für $i = 0, \dots, N$.

Schwache Formulierung

Für die schwache Formulierung wird analog zum zeitabhängigen Problem eine Variationsformulierung der Zustandsgleichung (5.1) durchgeführt. Als Arbeitsraum eignet sich die Menge:

Definition 5.2. *Sei H ein Hilbertraum und $N \in \mathbb{N}$. Dann definieren wir den Raum*

$$\begin{aligned} W_N^\omega((0,T),H) := \{ & v \in C([0,T],H) : \\ & v(t) := \sum_{i=0}^{N} (v_c^i \cos(\omega_i t) + v_s^i \sin(\omega_i t)) \\ & \textit{mit } v_c^i, v_s^i \in V, \omega_i = \frac{2\pi i}{T} \textit{ for } i = 0, \dots, N\}, \end{aligned}$$

Dabei gilt offensichtlich

$$W_N^\omega((0,T),H) \subset W((0,T),H) \tag{5.2}$$

und weiters lässt sich zeigen:

Satz 5.3. *Sei $N \in \mathbb{N}$ und H ein Hilbertraum. Dann bildet der Raum*

$$W_N^\omega((0,T),H)$$

ausgestattet mit dem inneren Produkt

$$(\cdot,\cdot)_{W((0,T),H)}$$

wiederum einen Hilbertraum.

Beweis. Wegen (5.2) bildet $W_N^\omega((0,T),H)$ ausgestattet mit dem inneren Produkt $(\cdot,\cdot)_{W((0,T),H)}$ einen Prähilbertraum. Es bleibt also nur mehr die Abgeschlossenheit von $W_N^\omega((0,T),H)$ zu zeigen:

Sei also $(v_n)_{n\in\mathbb{N}}$ eine Folge in $W_N^\omega((0,T),H)$ mit

$$\lim_{n\to\infty} v_n = v \in W_N^\omega((0,T),H),$$

dann folgt aus der der endlichen Summendarstellung

$$v_n(t) = \sum_{i=0}^{N} (v_{n,c}^i \cos(\omega_i t) + v_{n,s}^i \sin(\omega_i t))$$

mit $v_{n,c}^i, v_{n,s}^i \in H$ für alle $i \in \{0, \dots, N\}$ und $n \in \mathbb{N}$,

$$\lim_{n\to\infty} v_n = \sum_{i=0}^{N} (\underbrace{\lim_{n\to\infty} v_{n,c}^i}_{\in H} \cos(\omega_i \cdot) + \underbrace{\lim_{n\to\infty} v_{n,s}^i}_{\in H} \sin(\omega_i \cdot)) \in W_N^\omega((0,T),H).$$

□

Mit der Wahl

$$V = H_0^1(\Omega),$$
$$Q = L_0^2(\Omega) := \{p \in L^2(\Omega) : \int_\Omega p(x) = 0 \, \mathrm{d}x\},$$

erhält man dann für die schwache Formulierung:

(SOP$^\omega$) Schwache Formulierung des optimalen Kontrollproblems im zeitharmonischen Fall:

Gegeben

- $v_d \in W_N^\omega((0,T), V^d)$,
- $\alpha > 0$.

Finde

- $v \in W_N^\omega((0,T), V^d)$,
- $f \in W_N^\omega((0,T), (L^2(\Omega))^d)$,
- $p \in W_N^\omega((0,T), Q)$,

welche die **Zielfunktion**

$$J(v,f) := \frac{1}{2}||v - v_d||_{L^2}^2 + \frac{\alpha}{2}||f||_{L^2}^2,$$

minimiert, sodass die **Zustandsgleichungen**

$$\begin{aligned} (\partial_t v, w)_{L^2} + (\nabla_x v, \nabla_x w)_{L^2} & \\ +(\mathrm{div}_x\, w, p)_{L^2} &= (f, w)_{L^2}, \\ \forall w \in W_N^\omega((0,T), V^d), & \\ (\mathrm{div}_x\, v, q)_{L^2} &= 0, \\ \forall q \in W_N^\omega((0,T), Q), & \end{aligned} \tag{5.3}$$

erfüllt sind.

5.2 Existenz, Eindeutigkeit und Charakterisierung einer Lösung

Der zeitharmonische Ansatz

Wir nutzen nun die Möglichkeit, dass sich die gesuchten Funktionen $v_h, f_h, p_h, \lambda_h, \mu_h$ und die Daten $v_{d,h}$ in der Form

$$\begin{aligned}
v(x,t) &= \sum_{i=0}^{N} (v_c^i(x)\cos(\omega_i t) + v_s^i(x)\sin(\omega_i t)) \\
&\quad \text{mit} \quad v_c^i, v_s^i \in V^d, \quad \text{für } i = 1,\dots,N, \\
f(x,t) &= \sum_{i=0}^{N} (f_c^i(x)\cos(\omega_i t) + f_s^i(x)\sin(\omega_i t)) \\
&\quad \text{mit} \quad f_c^i, f_s^i \in U^d, \quad \text{für } i = 1,\dots,N, \\
p(x,t) &= \sum_{i=0}^{N} (p_c^i(x)\cos(\omega_i t) + p_s^i(x)\sin(\omega_i t)) \\
&\quad \text{mit} \quad p_c^i, p_s^i \in Q, \quad \text{für } i = 1,\dots,N,
\end{aligned} \tag{5.4}$$

$$\begin{aligned}
v_d(x,t) &= \sum_{i=0}^{N} (v_{d,c}^i(x)\cos(\omega_i t) + v_{d,s}^i(x)\sin(\omega_i t)) \\
&\quad \text{mit} \quad v_{d,c}^i, v_{d,s}^i \in V^d, \quad \text{für } i = 1,\dots,N,
\end{aligned}$$

schreiben lassen. Anhand der ersten Gleichung der beiden Zustandsgleichungen (5.3) werden wir nun zeigen, dass sich durch Einsetzen der harmonischen Ansätze aus (5.4) der Zeitparameter t eliminieren lässt:

$$\begin{aligned}
-(\partial_t v, w)_{L^2} + (\nabla_x v, \nabla_x w)_{L^2} + (\operatorname{div}_x w, p)_{L^2} &= (f, w)_{L^2}, \\
\forall w \in W_N^\omega((0,T), V^d),&
\end{aligned}$$

$$\Leftrightarrow$$

$$\begin{aligned}
-\int_I (\partial_t v(t), w(t))_{L^2}\,\mathrm{d}t + \int_I \underbrace{(\nabla_x v(t), \nabla_x w(t))_{L^2} + (\operatorname{div}_x w(t), p(t))_{L^2}}_{:=a(v,w)}\,\mathrm{d}t& \\
= \int_I (f(t), w)_{L^2}\,\mathrm{d}t& \\
\forall w \in W_N^\omega((0,T), V^d),&
\end{aligned}$$

$$\Leftrightarrow$$

$$
\begin{aligned}
-\int_I (\partial_t v, \cos(\omega_i t) w)_{L^2}\,\mathrm{d}t + \int_I a(v, \cos(\omega_i t) w)\,\mathrm{d}t \\
= \int_I (f(t), \cos(\omega_i t) w)_{L^2}\,\mathrm{d}t, \\
\wedge \\
-\int_I (\partial_t v, \sin(\omega_i t) w)_{L^2}\,\mathrm{d}t + \int_I a(v, \sin(\omega_i t) w)\,\mathrm{d}t \\
= \int_I (f(t), \sin(\omega_i t) w)_{L^2}\,\mathrm{d}t,, \\
\forall w \in V^d,\, i = 0, \ldots, N,
\end{aligned}
$$

$$\Leftrightarrow$$

$$
\begin{aligned}
&-\sum_{j=0}^{N} \omega_j \underbrace{(-\sin(\omega_j t), \cos(\omega_i t))_{L^2(0,T)}}_{=0} (v_c^j, w)_{L^2(\Omega)^d} - \sum_{j=0}^{N} \omega_j \underbrace{(\cos(\omega_j t), \cos(\omega_i t))_{L^2(0,T)}}_{=\delta_{ij} T/2} (v_s^j, w)_{L^2(\Omega)^d} \\
&\quad + \sum_{j=0}^{N} \underbrace{(\cos(\omega_j t), \cos(\omega_i t))_{L^2(0,T)}}_{=\delta_{ij} T/2} a(v_s^j, w)_{L^2(\Omega)^d} \\
&\quad + \sum_{j=0}^{N} \underbrace{(\sin(\omega_j t), \cos(\omega_i t))_{L^2(0,T)}}_{=0} (v_s^j, w)_{L^2(\Omega)^d} \\
&\quad = -\sum_{j=0}^{N} \underbrace{(\cos(\omega_j t), \cos(\omega_i t))_{L^2(0,T)}}_{=\delta_{ij} T/2} (f_c^j, w)_{L^2(\Omega)^d} \\
&\quad + \sum_{j=0}^{N} \underbrace{(\sin(\omega_j t), \cos(\omega_i t))_{L^2(0,T)}}_{=0} (f_s^j, w)_{L^2(\Omega)^d}, \\
&\quad \wedge \\
&-\sum_{j=0}^{N} \omega_j \underbrace{(-\sin(\omega_j t), \sin(\omega_i t))_{L^2(0,T)}}_{=\delta_{ij} T/2} (v_c^j, w)_{L^2(\Omega)^d} - \ldots \\
&\qquad \forall w \in V^d,\ i = 0, \ldots, N,
\end{aligned}
$$

$$\Leftrightarrow$$

$$\begin{aligned}
\omega_i(v_c^i, w)_{L^2} + (\nabla_x v_s^i, \nabla_x w)_{L^2} & \\
+(\operatorname{div}_x w, p_s^i)_{L^2} &= (f_s^i, w)_{L^2}, \\
-\omega_i(v_s^i, w)_{L^2} + (\nabla_x v_c^i, \nabla_x w)_{L^2} & \\
+(\operatorname{div}_x w, p_c^i)_{L^2} &= (f_c^i, w)_{L^2}, \\
\forall w \in V^d,\ i = 0, \ldots, N.&
\end{aligned}$$

Eine analoge Transformierung lässt sich auch, unter Einsetzung des zeitharmonischen Ansatzes, für die zweite der beiden Zustandsgleichung (5.3) durchführen, sodass insgesamt gilt:

$$\begin{aligned}
-(\partial_t v, w)_{L^2} + (\nabla_x v, \nabla_x w)_{L^2} + (\operatorname{div}_x w, p)_{L^2} &= (f, w)_{L^2}, \\
\forall w \in W_N^\omega((0,T), V^d), & \\
(\operatorname{div}_x v, q)_{L^2} &= 0, \\
\forall q \in W_N^\omega((0,T), Q), &
\end{aligned}$$

$$\Leftrightarrow$$

$$\begin{aligned}
\omega_i(v_c^i, w)_{L^2} + (\nabla_x v_s^i, \nabla_x w)_{L^2} & \\
+(\operatorname{div}_x w, p_s^i)_{L^2} &= (f_s^i, w)_{L^2}, \\
-\omega_i(v_s^i, w)_{L^2} + (\nabla_x v_c^i, \nabla_x w)_{L^2} & \\
+(\operatorname{div}_x w, p_c^i)_{L^2} &= (f_c^i, w)_{L^2}, \\
\forall w \in V^d, (\operatorname{div}_x v_c^i, q)_{L^2} &= 0, \\
(\operatorname{div}_x v_s^i, q)_{L^2} &= 0, \\
\forall q \in Q, & \\
\text{für } \ i = 0, \ldots, N. &
\end{aligned}$$

Der Nachweis der Existenz, Eindeutigkeit und Charakterisierung einer Lösung von (SOP^ω) lässt sich nun völlig analog wie für das zeitabhängige Problem (SOP) in Kapitel 3 durchführen:

Dies folgt einerseits aus der Tatsache, dass die Räume

$$\begin{aligned}
W_N^\omega((0,T), V^d) &\subset W((0,T), V^d), \\
W_N^\omega((0,T), Q) &\subset L^2((0,T), Q), \\
W_N^\omega((0,T), (L^2(\Omega))^d) &\subset L^2((0,T), (L^2(\Omega))^d),
\end{aligned} \tag{5.5}$$

wiederum Hilberträume bilden und sich andererseits für den Lösungsoperator

$$S^\omega : W_N^\omega((0,T),V^d) \times W_N^\omega((0,T),Q) \to W_N^\omega((0,T),V^d)^\star \times W_N^\omega((0,T),Q)^\star,$$

$$\begin{aligned}\langle S^\omega(v,p),(w,q)\rangle :=& \begin{pmatrix}(\partial_t w,\lambda)_{L^2} + (\nabla_x v,\nabla_x w)_{L^2} + (\operatorname{div}_x w,p)_{L^2}\\ (\operatorname{div}_x v,q)_{L^2}\end{pmatrix}\\ =&\langle S(v,p),(w,q)\rangle,\end{aligned} \tag{5.6}$$

$$\forall v,w \in W_N^\omega((0,T),V^d),\ \forall p,q \in W_N^\omega((0,T),Q).$$

wobei S den Isomorphismus aus (3.4) bezeichnet, die folgende Aussage zeigen lässt:

Satz 5.4. *Es gilt:*

Der in (5.6) definierte Operator bildet einen Isomorphismus.

Beweis. Um die Bijektivität von S^ω zu zeigen, wählen wir zunächst $(F,G) \in W_N^\omega((0,T),V^d)^\star \times W_N^\omega((0,T),Q)^\star$ beliebig aber fix. Dann gilt unter Verwendung des zeitharmonischen Ansatzes:

$$\begin{aligned}S^\omega(v,p) &= (F,G)^T,\\ &\Leftrightarrow\\ (\partial_t v,w)_{L^2} + (\nabla_x v,\nabla_x w)_{L^2}&\\ +(\operatorname{div}_x w,p)_{L^2} &= \langle F,w\rangle,\\ \forall w &\in W_N^\omega((0,T),V^d),\\ (\operatorname{div}_x v,q)_{L^2} &= \langle G,q\rangle,\\ \forall q &\in W_N^\omega((0,T),Q),\\ &\Leftrightarrow\\ -\omega_i(v_c^i,w)_{L^2} + (\nabla_x v_s^i,\nabla_x w)_{L^2}&\\ +(\operatorname{div}_x w,p_s^i)_{L^2} &= \langle F,w\rangle,\\ \omega_i(v_s^i,w)_{L^2} + (\nabla_x v_c^i,\nabla_x w)_{L^2}&\\ +(\operatorname{div}_x w,p_c^i)_{L^2} &= \langle F,w\rangle,\\ \forall w &\in V^d,\\ (\operatorname{div}_x v_c^i,q)_{L^2} &= \langle G,q\rangle,\\ (\operatorname{div}_x v_s^i,q)_{L^2} &= \langle G,q\rangle,\\ \forall q &\in Q,\\ \text{für}\ \ i &= 0,\dots,N.\end{aligned} \tag{5.7}$$

Aus der Bijektivität von S folgt nun, dass die Gleichung $S^\omega(v,p) = (F,G)^T$ eine eindeutige Lösung besitzt, welche sich wegen (5.7) in $W_N^\omega((0,T),V^d) \times W_N^\omega((0,T),Q)$ befindet. Die Linearität und Beschränktheit von S^ω und $(S^\omega)^{-1}$ folgt wiederum aus der Linearität und Beschränktheit von S und S^{-1}. □

Ersetzt man jetzt

$$\begin{aligned} W((0,T),V^d) &\mapsto W_N^\omega((0,T),V^d),\\ L^2((0,T),Q) &\mapsto W_N^\omega((0,T),Q),\\ L^2((0,T),(L^2(\Omega))^d) &\mapsto W_N^\omega((0,T),(L^2(\Omega))^d),\\ S &\mapsto S^\omega, \end{aligned}$$

dann besitzen die Aussagen aus Kapitel 3 wegen (5.5) und Satz 5.4 weiterhin ihre Gültigkeit. Aus Satz 3.26 erhalten wir somit für (SOP$^\omega$):

Satz 5.5. *Es gilt:*

Das optimale Kontrollproblem (SOP$^\omega$) besitzt eine eindeutige Lösung

$$(v,f,p) \in W_N^\omega((0,T),V^d) \times W_N^\omega((0,T),(L^2(\Omega))^d) \times W_N^\omega((0,T),Q),$$

für welche gilt:

$$\begin{gathered} \exists(\lambda,\mu) \in W_N^\omega((0,T),V^d) \times W_N^\omega((0,T),Q) \text{ sodass} \\ \nabla\mathcal{L}(v,f,p,\lambda,\mu) = 0, \end{gathered} \tag{5.8}$$

wobei $\mathcal{L}(v,f,p,\lambda,\mu)$ das Lagrange-Funktional von (SOP$^\omega$) bezeichnet.

5.2.1 Das Lagrange Funktional

Im Folgenden sei

$$\begin{aligned} X^\omega :=& W_N^\omega((0,T),V^d) \times W_N^\omega((0,T),(L^2(\Omega))^d) \times W_N^\omega((0,T),Q)\\ &\times W_N^\omega((0,T),V^d) \times W_N^\omega((0,T),(L^2(\Omega))^d) \subset X, \end{aligned}$$

mit X aus (3.17). Das Lagrange Funktional von (SOP$^\omega$) besitzt mit

$$\mathcal{L} : X^\omega \to \mathbb{R},$$

wobei

$$\begin{aligned} \mathcal{L}(v,f,p,\lambda,\mu) =& \frac{1}{2}||v-v_d||_{L^2}^2 + \frac{\alpha}{2}||f||_{L^2}^2 + (f,\lambda)_{L^2} - (\partial_t v,\lambda)_{L^2}\\ &- (\nabla_x v,\nabla_x\lambda)_{L^2} - (\operatorname{div}_x\lambda,p)_{L^2} - (\operatorname{div}_x v,\mu)_{L^2}, \end{aligned} \tag{5.9}$$

die selbe Gestalt wie für (SOP). Unter Anwendung des Satzes der partiellen Integration erhält man weiters

$$\begin{aligned} \mathcal{L}(v,f,p,\lambda,\mu) =& \frac{1}{2}||v-v_d||_{L^2}^2 + \frac{\alpha}{2}||f||_{L^2}^2 + (f,\lambda)_{L^2}\\ &- (v(T),\lambda(T))_{L^2} + (v(0),\lambda(0))_{L^2} + (v,\partial_t\lambda)_{L^2}\\ &- (\nabla_x v,\nabla_x\lambda)_{L^2} - (\operatorname{div}_x\lambda,p)_{L^2} - (\operatorname{div}_x v,\mu)_{L^2}\\ =& \frac{1}{2}||v-v_d||_{L^2}^2 + \frac{\alpha}{2}||f||_{L^2}^2 + (f,\lambda)_{L^2} + (v,\partial_t\lambda)_{L^2}\\ &- (\nabla_x v,\nabla_x\lambda)_{L^2} - (\operatorname{div}_x\lambda,p)_{L^2} - (\operatorname{div}_x v,\mu)_{L^2}, \end{aligned} \tag{5.10}$$

wobei der Term

$$-(v(T),\lambda(T))_{L^2}+(v(0),\lambda(0))_{L^2},$$

wegen der zeitlichen Periodizität von v auf $[0,T]$, verschwindet.

5.2.2 Das Optimalitätssystem

Für eine Lösung $x=(v,f,p,\lambda,\mu)\in X^\omega$ der Optimalitätsbedinung in (5.8) gilt dabei:

$$\begin{aligned}
\nabla\mathcal{L}(x)&=0 \ \text{ in } X,\\
&\Leftrightarrow\\
\langle\partial_v\mathcal{L}(x),w\rangle&=0 \ \ \forall w\in W_N^\omega((0,T),V^d),\\
\langle\partial_f\mathcal{L}(x),g\rangle&=0 \ \ \forall g\in W_N^\omega((0,T),(L^2(\Omega))^d),\\
\langle\partial_p\mathcal{L}(x),q\rangle&=0 \ \ \forall q\in W_N^\omega((0,T),Q),\\
\langle\partial_\lambda\mathcal{L}(x),w\rangle&=0 \ \ \forall w\in W_N^\omega((0,T),V^d),\\
\langle\partial_\mu\mathcal{L}(x),q\rangle&=0 \ \ \forall q\in W_N^\omega((0,T),Q).
\end{aligned}\tag{5.11}$$

Die Berechnung der partiellen Gâteaux-Ableitungen erfolgt dabei analog zum zeitabhängigen Problem mit Hilfe von Beispiel 3.16 und den beiden Darstellungen (5.9) und (5.10) für $\mathcal{L}$. Für das Optimalitätssystem (5.11) erhält man dann die folgende Darstellung:

(OS$^\omega$) Optimalitätssystem für das optimale Kontrollproblem im zeitharmonischen Fall:

Finde $(v,f,p,\lambda,\mu)\in X^\omega$ sodass gilt:

1.

$$\begin{aligned}
-(\partial_t\lambda,w)_{L^2}+(\nabla_x\lambda,\nabla_x w)_{L^2}&\\
+(\operatorname{div}_x w,\mu)_{L^2}&=(v-v_d,w)_{L^2},\\
\forall w\in W_N^\omega((0,T),V^d),&\\
(\operatorname{div}_x\lambda,q)_{L^2}&=0,\\
\forall q\in W_N^\omega((0,T),Q),&\\
\overset{(a)}{\Leftrightarrow}&\\
\omega_i(\lambda_c^i,w)_{L^2}+(\nabla_x\lambda_s^i,\nabla_x w)_{L^2}&\\
+(\operatorname{div}_x w,\mu_s^i)_{L^2}&=(v_s^i-v_{d,s}^i,w)_{L^2},\\
-\omega_i(\lambda_s^i,w)_{L^2}+(\nabla_x\lambda_c^i,\nabla_x w)_{L^2}&\\
+(\operatorname{div}_x w,\mu_c^i)_{L^2}&=(v_c^i-v_{d,c}^i,w)_{L^2},\\
\forall w\in V^d,&\\
(\operatorname{div}_x\lambda_c^i,q)_{L^2}&=0,\\
(\operatorname{div}_x\lambda_s^i,q)_{L^2}&=0,\\
\forall q\in Q,&
\end{aligned}\tag{5.12}$$

2.

$$
\begin{aligned}
(\partial_t v, w)_{L^2} + (\nabla_x v, \nabla_x w)_{L^2} & \\
+(\mathrm{div}_x\, w, p)_{L^2} &= -\alpha^{-1}(\lambda, w)_{L^2}, \\
\forall w \in W_N^\omega((0,T), V^d), & \\
(\mathrm{div}_x\, v, q)_{L^2} &= 0, \\
\forall q \in W_N^\omega((0,T), Q), & \\
\overset{(a)}{\Leftrightarrow} & \\
-\omega_i(v_c^i, w)_{L^2} + (\nabla_x v_s^i, \nabla_x w)_{L^2} & \\
+(\mathrm{div}_x\, w, p_s^i)_{L^2} &= -\alpha^{-1}(\lambda_s^i, w)_{L^2}, \\
\omega_i(v_s^i, w)_{L^2} + (\nabla_x v_c^i, \nabla_x w)_{L^2} & \\
+(\mathrm{div}_x\, w, p_c^i)_{L^2} &= -\alpha^{-1}(\lambda_c^i, w)_{L^2}, \\
\forall w \in V^d, & \\
(\mathrm{div}_x\, v_c^i, q)_{L^2} &= 0, \\
(\mathrm{div}_x\, v_s^i, q)_{L^2} &= 0, \\
\forall q \in Q &
\end{aligned}
\tag{5.13}
$$

mit

$$f = \alpha^{-1}\lambda$$

und

$$
\begin{aligned}
v_d \in & W_N^\omega((0,T), V^d), \\
\alpha > & 0,
\end{aligned}
$$

Die Äquivalenz in (a) ergibt sich wiederum unter Verwendung des zeitharmonischen Ansatzes.

Das Optimalitätssystem für den zeitharmonischen Fall lässt also durch $N+1$ entkoppelte **stationäre** Systeme beschreiben.

5.3 Diskretisierung

Für die räumliche Diskretisierung des Optimalitätssystem, konstruieren wir analog zum zeitabhängige Problem in Abschnitt 4.1.3, die Teilräume

$$V_h \subset V, U_h \subset U \text{ und } Q_h = U_h \cap Q,$$

mittels einer Finiten-Element-Methode.

5.3.1 Das räumlich diskretisierte Optimalitätssystem

Im folgenden sei

$$\begin{aligned} X_h^\omega :=& W_N^\omega((0,T),V_h^d) \times W_N^\omega((0,T),U_h^d) \times W_N^\omega((0,T),Q_h) \\ & \times W_N^\omega((0,T),V_h^d) \times W_N^\omega((0,T),Q_h). \end{aligned}$$

Das räumlich diskretisierte Optimalitätssystem (OS_h) erhält nun durch Ersetzung von V bei V_h, Q bei Q_h und X bei X_h^ω:

(OS_h^ω) Räumlich diskretisiertes Optimalitätssystem im zeitharmonischen Fall:

Gegeben:

- $\boldsymbol{v_{d,h}} \in W_N^\omega((0,T),V_h^d)$,
- $\alpha > 0$.

Finde $(v_h, f_h, p_h, \lambda_h, \mu_h) \in X_h^\omega$ sodass gilt:

1.

$$\begin{aligned} -(\partial_t \lambda_h, w_h)_{L^2} + (\nabla_x \lambda_h, \nabla_x w_h)_{L^2} & \\ +(\mathrm{div}_x\, w_h, \mu_h)_{L^2} &= (v_h - \boldsymbol{v_{d,h}}, w)_{L^2}, \\ \forall w_h &\in W_N^\omega((0,T),V_h^d), \\ (\mathrm{div}_x\, \lambda, q_h)_{L^2} &= 0, \\ \forall q_h &\in W_N^\omega((0,T),Q_h), \end{aligned}$$

$$\Leftrightarrow$$

$$\begin{aligned} \omega_i(\lambda_{c,h}^i, w_h)_{L^2} + (\nabla_x \lambda_{s,h}^i, \nabla_x w_h)_{L^2} & \\ +(\mathrm{div}_x\, w_h, \mu_{s,h}^i)_{L^2} &= (v_{s,h}^i - v_{d,s,h}^i, w_h)_{L^2}, \\ -\omega_i(\lambda_{s,h}^i, w_h)_{L^2} + (\nabla_x \lambda_{c,h}^i, \nabla_x w_h)_{L^2} & \\ +(\mathrm{div}_x\, w_h, \mu_{c,h}^i)_{L^2} &= (v_{c,h}^i - v_{d,c,h}^i, w_h)_{L^2}, \\ \forall w_h &\in V^d, \\ (\mathrm{div}_x\, \lambda_{c,h}^i, q_h)_{L^2} &= 0, \\ (\mathrm{div}_x\, \lambda_{s,h}^i, q_h)_{L^2} &= 0, \\ \forall q_h &\in Q, \end{aligned}$$

$$\overset{(b)}{\Leftrightarrow}$$

$$\begin{aligned}
\omega_i C_h \underline{\lambda}_c^i + A_h \underline{\lambda}_s^i + B_h^T \mu_s^i &= C_h(\underline{v}_s^i - \underline{v}_{d,s}^i),\\
-\omega_i C_h \underline{\lambda}_s^i + A_h \underline{\lambda}_c^i + B_h^T \mu_c^i &= C_h(\underline{v}_c^i - \underline{v}_{d,c}^i),\\
B_h \underline{\lambda}_c^i &= 0,\\
B_h \underline{\lambda}_s^i &= 0,
\end{aligned}$$

2.

$$\begin{aligned}
(\partial_t v, w)_{L^2} + (\nabla_x v, \nabla_x w)_{L^2} &\\
+(\mathrm{div}_x\, w, p)_{L^2} &= -\alpha^{-1}(\lambda, w)_{L^2},\\
\forall w \in W_N^\omega((0,T), V^d),&\\
(\mathrm{div}_x\, v, q)_{L^2} &= 0,\\
\forall q \in W_N^\omega((0,T), Q),&
\end{aligned}$$

$$\Leftrightarrow$$

$$\begin{aligned}
-\omega_i(v_{c,h}^i, w_h)_{L^2} + (\nabla_x v_{s,h}^i, \nabla_x w_h)_{L^2} &\\
+(\mathrm{div}_x\, w_h, p_{s,h}^i)_{L^2} &= -\alpha^{-1}(\lambda_{s,h}^i, w_h)_{L^2},\\
\omega_i(v_{s,h}^i, w_h)_{L^2} + (\nabla_x v_{c,h}^i, \nabla_x w_h)_{L^2} &\\
+(\mathrm{div}_x\, w_h, p_{c,h}^i)_{L^2} &= -\alpha^{-1}(\lambda_{c,h}^i, w)_{L^2},\\
\forall w_h \in V^d,&\\
(\mathrm{div}_x\, v_{c,h}^i, q_h)_{L^2} &= 0,\\
(\mathrm{div}_x\, v_{s,h}^i, q_h)_{L^2} &= 0,\\
\forall q_h \in Q.&
\end{aligned}$$

$$\overset{(b)}{\Leftrightarrow}$$

$$\begin{aligned}
-\omega_i C_h \underline{v}_c^i + A_h \underline{v}_s^i + B_h^T \underline{p}_s^i &= -\alpha^{-1} C_h \underline{\lambda}_s^i,\\
\omega_i C_h \underline{v}_s^i + A_h \underline{v}_c^i + B_h^T \underline{p}_c^i &= -\alpha^{-1} C_h \underline{\lambda}_c^i,\\
B_h \underline{v}_c^i &= 0,\\
B_h \underline{v}_s^i &= 0.
\end{aligned}$$

mit

$$f_h = \alpha^{-1}\lambda_h.$$

Mit der Ersetzung von v_d durch $v_{d,h} \in W_N^\omega((0,T), V_h^d)$, wobei gelte

$$(v_{d,h}, w)_{L^2} = (v_d, w)_{L^2} \quad \forall w \in W_N^\omega((0,T), V_h^d),$$

wird das letzte Objekt eines unendlich dimensionalen Raumes aus dem Optimalitätssystem entfernt. Die Existenz und Eindeutigkeit von $v_{d,h}$ erhält man wiederum aus dem *Rieszschen Darstellungssatz* (siehe Satz 3.14) und der Tatsache, dass der endlich dimensionale Raum $W_N^\omega(\Delta_\tau, V_h^d)$ ausgestattet mit $(\cdot,\cdot)_{L^2}$) einen Hilbertraum bildet.

Die Äquivalenz in (b) folgt aus den in den Abschnitt 4.1.2 gezeigten Zusammenhängen

$$\begin{aligned}(\nabla_x v_h, \nabla_x w_h)_{L^2} &= (A_h \underline{v}, \underline{w})_{\ell^2} \quad \forall v_h, w_h \in V_h^d,\\ (\operatorname{div}_x v_h, q_h)_{L^2} &= (B_h \underline{v}, \underline{w})_{\ell^2} \quad \forall v_h \in, \forall q_h \in, Q_h,\\ (v_h, w_h)_{L^2} &= (C_h \underline{v}, \underline{w})_{\ell^2} \quad \forall v_h, w_h \in V_h^d.\end{aligned}$$

Geschrieben in Form eines linearen Gleichungssystems, erhält man:

(OS$_h^\omega$)

Gegeben:

- $v_d \in W_N^\omega((0,T), V_h^d)$,
- $\alpha > 0$.

Gesucht: $x = (\underline{x}^i)_{i=0,\dots,N} \in \mathbb{R}^{4(N+1)(dn_h+m_h)}$ sodass gilt:

$$\begin{pmatrix} -C_h & 0 & 0 & 0 & A_h & -\omega_i C_h & B_h^T & 0 \\ 0 & -C_h & 0 & 0 & \omega_i C_h & A_h & 0 & B_h^T \\ 0 & 0 & 0 & 0 & B_h & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & B_h & 0 & 0 \\ A_h & \omega_i C_h & B_h^T & 0 & \frac{1}{\alpha} C_h & 0 & 0 & 0 \\ -\omega_i C_h & A_h & 0 & B_h^T & 0 & \frac{1}{\alpha} C_h & 0 & 0 \\ B_h & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & B_h & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \underline{v}_c^i \\ \underline{v}_s^i \\ \underline{p}_c^i \\ \underline{p}_s^i \\ \underline{\lambda}_c^i \\ \underline{\lambda}_s^i \\ \underline{\mu}_c^i \\ \underline{\mu}_s^i \end{pmatrix} = \begin{pmatrix} -C_h \underline{v}_{d,c}^i \\ -C_h \underline{v}_{d,s}^i \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \tag{5.14}$$

für $i = 0, \dots, N$.

Wie im zeitabhängigen Fall, lässt sich auch hier die Matrix aus (5.14) in die **Sattelpunktsform** (4.24)

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix},$$

transformieren. Dazu führen wir die folgenden Matrixoperationen durch:

1. Vertauschen der 3. und 4. Zeile mit der 5. und 6. Zeile:

$$\begin{pmatrix} -C_h & 0 & 0 & 0 & A_h & -\omega_i C_h & B_h^T & 0 \\ 0 & -C_h & 0 & 0 & \omega_i C_h & A_h & 0 & B_h^T \\ A_h & \omega_i C_h & B_h^T & 0 & \frac{1}{\alpha} C_h & 0 & 0 & 0 \\ -\omega_i C_h & A_h & 0 & B_h^T & 0 & \frac{1}{\alpha} C_h & 0 & 0 \\ 0 & 0 & 0 & 0 & B_h & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & B_h & 0 & 0 \\ B_h & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & B_h & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

2. Vertauschen die 3. und 4. Spalte mit der 5. und 6. Spalte:

$$\begin{pmatrix} -C_h & 0 & A_h & -\omega_i C_h & 0 & 0 & B_h^T & 0 \\ 0 & -C_h & \omega_i C_h & A_h & 0 & 0 & 0 & B_h^T \\ A_h & \omega_i C_h & \frac{1}{\alpha} C_h & 0 & B_h^T & 0 & 0 & 0 \\ -\omega_i C_h & A_h & 0 & \frac{1}{\alpha} C_h & 0 & B_h^T & 0 & 0 \\ 0 & 0 & B_h & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & B_h & 0 & 0 & 0 & 0 \\ B_h & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & B_h & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

3. Multiplikation der 1. und 2. Zeile mit dem Faktor -1:

$$\begin{pmatrix} C_h & 0 & -A_h & \omega_i C_h & 0 & 0 & -B_h^T & 0 \\ 0 & C_h & -\omega_i C_h & -A_h & 0 & 0 & 0 & -B_h^T \\ A_h & \omega_i C_h & \frac{1}{\alpha} C_h & 0 & B_h^T & 0 & 0 & 0 \\ -\omega_i C_h & A_h & 0 & \frac{1}{\alpha} C_h & 0 & B_h^T & 0 & 0 \\ 0 & 0 & B_h & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & B_h & 0 & 0 & 0 & 0 \\ B_h & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & B_h & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

4. Multiplikation der 3. Spalte und der 4. Zeile mit dem Faktor -1:

$$\begin{pmatrix} C_h & 0 & A_h & \omega_i C_h & 0 & 0 & -B_h^T & 0 \\ 0 & C_h & \omega_i C_h & -A_h & 0 & 0 & 0 & -B_h^T \\ A_h & \omega_i C_h & -\frac{1}{\alpha} C_h & 0 & B_h^T & 0 & 0 & 0 \\ \omega_i C_h & -A_h & 0 & -\frac{1}{\alpha} C_h & 0 & -B_h^T & 0 & 0 \\ 0 & 0 & -B_h & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & B_h & 0 & 0 & 0 & 0 \\ B_h & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & B_h & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

5. Multiplikation der Zeilen und Spalten 3,4 mit dem Faktor $\sqrt{\alpha}$:

$$\begin{pmatrix} C_h & 0 & \sqrt{\alpha}A_h & \sqrt{\alpha}\omega_i C_h & 0 & 0 & -B_h^T & 0 \\ 0 & C_h & \sqrt{\alpha}\omega_i C_h & -\sqrt{\alpha}A_h & 0 & 0 & 0 & -B_h^T \\ \sqrt{\alpha}A_h & \sqrt{\alpha}\omega_i C_h & -C_h & 0 & \sqrt{\alpha}B_h^T & 0 & 0 & 0 \\ \sqrt{\alpha}\omega_i C_h & -\sqrt{\alpha}A_h & 0 & -C_h & 0 & -\sqrt{\alpha}B_h^T & 0 & 0 \\ 0 & 0 & -\sqrt{\alpha}B_h & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sqrt{\alpha}B_h & 0 & 0 & 0 & 0 \\ B_h & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & B_h & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

6. Multiplikation der Zeilen und Spalten 5,6 mit dem Faktor $\frac{1}{\sqrt{\alpha}}$:

$$\begin{pmatrix} C_h & 0 & \sqrt{\alpha}A_h & \sqrt{\alpha}\omega_i C_h & 0 & 0 & -B_h^T & 0 \\ 0 & C_h & \sqrt{\alpha}\omega_i C_h & -\sqrt{\alpha}A_h & 0 & 0 & 0 & -B_h^T \\ \sqrt{\alpha}A_h & \sqrt{\alpha}\omega_i C_h & -C_h & 0 & B_h^T & 0 & 0 & 0 \\ \sqrt{\alpha}\omega_i C_h & -\sqrt{\alpha}A_h & 0 & -C_h & 0 & -B_h^T & 0 & 0 \\ 0 & 0 & -B_h & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & B_h & 0 & 0 & 0 & 0 \\ B_h & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & B_h & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

7. Multiplikation der Zeilen 6 bis 8, und der 5. Spalte mit dem Faktor -1.

$$\begin{pmatrix} C_h & 0 & \sqrt{\alpha}A_h & \sqrt{\alpha}\omega_i C_h & 0 & 0 & -B_h^T & 0 \\ 0 & C_h & \sqrt{\alpha}\omega_i C_h & -\sqrt{\alpha}A_h & 0 & 0 & 0 & -B_h^T \\ \sqrt{\alpha}A_h & \sqrt{\alpha}\omega_i C_h & -C_h & 0 & -B_h^T & 0 & 0 & 0 \\ \sqrt{\alpha}\omega_i C_h & -\sqrt{\alpha}A_h & 0 & -C_h & 0 & -B_h^T & 0 & 0 \\ 0 & 0 & -B_h & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -B_h & 0 & 0 & 0 & 0 \\ -B_h & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -B_h & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \quad (5.15)$$

Bemerkung 5.6. *Aus der Darstellung (5.15), lassen sich nun einige Eigenschaften ableiten, die für die weitere Untersuchung und insbesonders bei der Konstruktion eines robusten Präkonditioniers in Kapitel 7 von großer Bedeutung sein werden. Für die Idee und Durchführung der Transformation von der Matrix aus (5.14) in die Gestalt (5.15) siehe [5].*

Werden die Transformationen auch am Lösungsvektor $\underline{x}^i$ und auf der rechten Seite $\underline{b}^i$ aus (5.14) berücksichtigt, so lässt sich das (OP_h^ω) weiters schreiben:

$(\mathbf{OS}_h^\omega)$

Gegeben:

- $v_d \in W_N^\omega((0,T), V_h^d)$,
- $\alpha > 0$.

Gesucht: $x = (\underline{x}^i)_{i=0,\dots,N} \in \mathbb{R}^{4(N+1)(dn_h+m_h)}$ sodass gilt:

$$\mathcal{A}^{\omega_i}\underline{x}^i = \underline{b}^i \quad \text{für} \quad i = 0,\dots,N, \tag{5.16}$$

wobei

- $$\mathcal{A}^\omega := \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}, \tag{5.17}$$

  mit

  $$A = \begin{pmatrix} A_1 & B_1^T \\ B_1 & -A_1 \end{pmatrix} \in \mathbb{R}^{4dn_h \times 4dn_h},$$

  wobei (5.18)

  $$A_1 = \begin{pmatrix} C_h & 0 \\ 0 & C_h \end{pmatrix} \quad \text{und} \quad B_1 = \begin{pmatrix} \alpha^{\frac{1}{2}} A_h & \alpha^{\frac{1}{2}}\omega C_h \\ \alpha^{\frac{1}{2}}\omega C_h & -\alpha^{\frac{1}{2}} A_h \end{pmatrix}$$

  und

  $$B = \begin{pmatrix} 0 & 0 & -B_h & 0 \\ 0 & 0 & 0 & -B_h \\ -B_h & 0 & 0 & 0 \\ 0 & -B_h & 0 & 0 \end{pmatrix} \in \mathbb{R}^{4m_h \times 4dn_h}. \tag{5.19}$$

- $$\underline{x}^i = \begin{pmatrix} \underline{x}_1^i \\ \underline{x}_2^i \end{pmatrix} \in \mathbb{R}^{4(dn_h+m_h)},$$

  mit

  $$\underline{x}_1^i = \begin{pmatrix} \underline{v}_c^i \\ \underline{v}_s^i \\ \alpha^{-\frac{1}{2}}\underline{\lambda}_c^i \\ \alpha^{-\frac{1}{2}}\underline{\lambda}_s^i \end{pmatrix} \in \mathbb{R}^{4dn_h} \quad \text{und} \quad \underline{x}_2^i = \begin{pmatrix} \alpha^{\frac{1}{2}}\underline{p}_c^i \\ \alpha^{\frac{1}{2}}\underline{p}_s^i \\ \underline{\mu}_c^i \\ \underline{\mu}_s^i \end{pmatrix}.$$

- $$\underline{b} = \begin{pmatrix} \underline{b}_1^i \\ \underline{b}_2^i \end{pmatrix} \in \mathbb{R}^{4(dn_h+m_h)}$$

  mit

  $$\underline{b}_1^i = \begin{pmatrix} -C_h\underline{v}_{d,c}^i \\ -C_h\underline{v}_{d,s}^i \\ 0 \\ 0 \end{pmatrix} \quad \text{und} \quad \underline{b}_2^i = 0.$$

5.3.2 Eigenschaften von $\mathcal{A}^\omega$

Die Matrix $\mathcal{A}^\omega$ besitzt nun die folgenden Eigenschaften:

Satz 5.7. *Sei $\alpha > 0, \omega > 0, h > 0$ und die Annahme 1 erfüllt. Dann gilt für die Matrix $\mathcal{A}$ aus (5.17) und ihre Bestandteile aus (5.18) und (5.19):*

1. *Die Matrix A_1 ist symmetrisch und positiv definit und die Matrix B_1 ist symmetrisch und regulär.*
2. *Die Matrix A ist regulär, symmetrisch und indefinit und die Matrix B besitzt vollen Rang.*
3. *Die Matrix $\mathcal{A}^\omega$ ist regulär, symmetrisch und indefinit.*

Beweis. Aus Satz 4.9 und Annahme 1 folgt

$$A_h, C_h \text{ sind s.p.d. und } B_h \text{ besitzt vollen Rang.} \tag{5.20}$$

Nun gilt:

1. Die Aussage für die Matrix A_1 erhält man sofort aus (5.20) und der Diagonal-Blockstruktur. Weiters sind wegen (5.20) mit $\omega > 0$ und $\alpha > 0$ alle Voraussetzungen von 4.20 für B_1 erfüllt, sodass gilt: B_1 ist regulär, symmetrisch und indefinit.
2. Der volle Rang der Matrix B folgt aus dem vollen Rang der Matrix B_h und der Blockstruktur. Die Regularität, Symmetrie und Indefinitness der Matrix A folgt dann aus Korollar 4.20, wobei die Voraussetzungen wegen 1. erfüllt sind.
3. Die Behauptung für die Matrix $\mathcal{A}^\omega$ folgt schließlich aus den Aussagen von 2. und Korollar 4.21.

□

Die Berechnung des diskretisierten Optimalitätssystems im zeit-harmonischen Fall (OP_h^ω) bedarf also der Lösung von linearen Gleichungssystemen der allgemeinen Form

$$\mathcal{A}^\omega \underline{x} = \underline{b} \quad \text{mit} \quad \omega > 0.$$

Dabei handelt es sich bei $\mathcal{A}^\omega$ wie im zeitabhängigen Fall wiederum um eine **große, dünnbesetzte, reguläre, symmetrische und indefinite Blockmatrix** in **Sattelpunktsform**.

Kapitel 6

Das MINRES-Verfahren

Sowohl im zeitabhängigen als auch im zeitharmonischen Fall erfordert die Lösung des diskretisierten Optimalitätssystem die Berechnung einer Lösung eines linearen Gleichungssystems

$$Ax = b,$$

mit **großer, dünnbesetzter, regulärer, symmetrisch und indefiniter Blockmatrix** A in **Sattelpunktsform**.

Eine Möglichkeit zu Lösung solcher Systeme ist das MINRES-Verfahren. Der folgende Abschnitt basierend auf der Literatur *„ Nummerik linearer Gleichungssysteme: Direkte und iterative Verfahren [14], Kapitel 6, Seite 221 - 255“*, dient der Vorstellung und näheren Untersuchung.

6.1 Das Verfahren

Das MINRES (Minimal Residual)-Verfahren gehört zur Klasse der Allgemeinen Minimum-Residuum-Verfahren und dient zur Lösung von linearen Gleichungssystemen

$$Ax = b, \tag{6.1}$$

mit $x, b \in \mathbb{R}^n$ und $A \in \mathbb{R}^{n \times n}$ **symmetrisch** und regulär.

Charakteristisch für die Allgemeinen Minimum-Residuum-Verfahren ist die Wahl der k-ten Iterierten als

$$x_k = \arg \min_{x \in x^0 + \mathcal{K}_k(A, r^0)} \frac{1}{2} ||b - Ax||_{\ell^2}, \tag{6.2}$$

das heißt x_k minimiert das Residuum in $\mathcal{K}_k(A, r_0)$, wobei

$$\mathcal{K}_k(A, r_0) := \text{span}(\{r_0, Ar_0, \ldots, A^{k-1} r_0\}),$$

den Krylov-Raum k-ter Ordnung, $x_0 \in \mathbb{R}^n$ den Startwert und $r_0 := b - Ax_0$ das Anfangsresiduum bezeichnet.
Wir werden nun für das MINRES-Verfahren grob die Vorgehensweise zur Lösung des Minimierungsproblems (6.2) im k-ten Iterationsschritt beschreiben:

1. Konstruktion einer Orthogonalbasis von $\mathcal{K}_k(A, r_0)$: Für symmetrische reguläre Matrizen kann eine orthogonale Basis $\{r_0, v_1, \ldots, v_{k-1}\}$ des Krylov-Raumes $\mathcal{K}_k(A, r_0)$, mit Hilfe des *Lanczos-Algorithmus* (siehe [14], Seite 241), in einer drei Term Rekurrenz-Relation der Form

$$Av_k = v_{k+1}t_{k+1k} + v_k t_{kk} + v_{k-1}t_{k-1k} \tag{6.3}$$

konstruiert werden. Dabei lässt sich die Rekurrenz (6.3) auch als Matrixform schreiben

$$AV_k = V_{k+1}T_k, \tag{6.4}$$

wobei $T_k \in \mathbb{R}^{k+1\times k}$ eine Tridiagonalmatrix ist und die Matrix $V_k \in \mathbb{R}^{n\times k+1}$ als i-ten Spalteneintrag den i-ten Basisvektor v_i besitzt.

2. Da die Spalten der Matrix V_k eine Basis von $\mathcal{K}_k(A, r_0)$ bilden, kann für $x_k \in \mathcal{K}_k(A, r_0)$ der Ansatz verwendet werden

$$x_k = V_k y \quad \text{mit} \quad y \in \mathbb{R}^k.$$

Unter Verwendung dieses Ansatzes und (6.4) gilt

$$||Ax_i - b||_{\ell^2} = ||AV_k y - b||_{\ell^2} = ||V_{k+1}T_k y - b||_{\ell^2}.$$

Sei weiters

$$D_{k+1} = \operatorname{diag}(||r_0||_{\ell^2}, ||v_1||_{\ell^2}, \ldots, ||v_k||_{\ell^2}),$$

dann ist die Matrix $V_{k+1}D_{k+1}^{-1}$ orthogonal, woraus für die Darstellung des Residuums folgt:

$$\begin{aligned} ||Ax_i - b||_{\ell^2} =& ||V_{k+1}T_k y - b||_{\ell^2} \\ =& ||V_{k+1}D_{k+1}^{-1}D_{k+1}T_k y - b||_{\ell^2} \\ =& ||D_{k+1}T_k y - ||r_0||_{\ell^2}e_1||_{\ell^2}, \end{aligned}$$

wobei $e_1 \in \mathbb{R}^{k+1}$ den ersten Einheitsvektor bezeichnet.

3. Eliminiert man jetzt noch durch eine einfache Givens-Rotation $G \in \mathbb{R}^{k+1\times k+1}$ den Wert an der Position $(k+1, k)$ von der Tridiagonalenmatrix T_k (die anderen Werte der unteren Nebendiagonale der Matrix T_k wurden bereits in den vohergenden Schritten ausgelöscht), so gilt:

$$GT_k = \begin{pmatrix} R \\ 0 \end{pmatrix}, \tag{6.5}$$

mit $R \in \mathbb{R}^{k\times k}$ welche nur Nichtnulleinträge in der Haupt- und oberen Nebendiagonale besitzt. Mit

$$z = ||r_0||_{\ell^2} G e_1 \in \mathbb{R}^{k+1}, \tag{6.6}$$

und (6.5) kann das Residuum geschrieben werden als

$$\begin{aligned} ||Ax_i - b||_{\ell^2}^2 =& ||D_{k+1} T_k y_k - ||r_0||_{\ell^2} e_1||_{\ell^2}^2 \\ =& ||Ry_k - (z_i)_{i=1,\dots k}||_{\ell^2}^2 + ||z_{k+1}||_{\ell^2}^2, \end{aligned}$$

sodass der Koeffizientenvektor y einfach durch lösen des linearen Gleichungssystems

$$Ry = (z_i)_{i=1,\dots k},$$

mit der Rücksubstitutionsmethode berechnet werden kann.

4. Berechnung der Lösung:

$$x_k = Ry.$$

6.2 Die Konvergenzanalyse

Es lässt sich folgendes Konvergenzresultat für das MINRES-Verfahren zeigen (Beweis siehe z.B. [15]):

Satz 6.1. *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und regulär, und sei x^k die durch das MINRES-Verfahren erzeugte k-te Iterierte, dann gilt*

$$||r^{2m}||_{\ell^2} \leq \frac{2q^m}{1+q^{2m}} ||r^0||_{\ell^2},$$

wobei

$$\begin{aligned} r^k =& b - Ax^k \quad \textit{für} \quad k \geq 0, \\ q =& \frac{\kappa(A)-1}{\kappa(A)+1}. \end{aligned}$$

Die Anzahl der benötigten Iterationen hängt also von der Größe der Konditionszahl $\kappa(A)$ ab.

In unserem Fall, also $A = \mathcal{A}$ oder $A = \mathcal{A}^\omega$ hängt die Konditionszahl von den Modellparametern und der Schrittweite h ab. Nun ist es der Fall, dass für die Konditionszahl für bestimmte Parameterwahlen $\alpha > 0, \omega > 0$ und $h > 0$, sehr große Werte annimmt, was eine sehr langsame Konvergenzgeschwindigkeit und damit unakzeptable Wartezeiten zur Folge hat.

Die Kondtionszahl lässt sich allerdings durch Präkonditionierung beeinflussen:

6.3 Das präkonditionierte MINRES-Verfahren

Wir sind also besonders interessiert in der Anwendung des MINRES-Verfahren auf das präkonditionierte System

$$P^{-1}Ax = P^{-1}b, \tag{6.7}$$

wobei $P \in \mathbb{R}^{n\times n}$ eine reguläre Matrix ist. Die Matrix $P^{-1}A$ ist aber im Allgemeinen nicht mehr symmetrisch, sodass das MINRES-Verfahren nicht direkt angewendet werden kann.

Sei nun im folgenden $(\cdot,\cdot)_H$ ein beliebiges inneres Produkt auf $\mathbb{R}^n$ mit induzierter Norm $||\cdot||_H := (\cdot,\cdot)_H^{\frac{1}{2}}$. Ersetzen wir nun das innere Produkt $(\cdot,\cdot)_{\ell^2}$ und die Norm $||\cdot||_{\ell^2}$ im MINRES-Algorithmus durch $(\cdot,\cdot)_H$ und $||\cdot||_H$, dann lässt sich zeigen, dass sich das MINRES-Verfahren auf das lineare Gleichungssystem aus (6.1) anwenden lässt, falls die Matrix A bezüglich dem inneren Produkt $(\cdot,\cdot)_H$ *selbstadjungiert* ist (siehe [14], Seite 250 - 255).

Definition 6.2. *Sei $A \in \mathbb{R}^{n\times n}$ und $(\cdot,\cdot)_H$ ein inneres Produkt auf $\mathbb{R}^n$. Dann heißt A bezüglich $(\cdot,\cdot)$ selbstadjugiert, falls gilt:*

$$(Av, w)_H = (v, Aw)_H \quad \forall v, w \in \mathbb{R}^n.$$

Für $(\cdot,\cdot)_H := (\cdot,\cdot)_P$ mit

$$(v, w)_P := (Pv, w)_{\ell^2},$$

folgt aus

$$\begin{aligned}(P^{-1}Av, v)_P =& (PP^{-1}Av, v)_{\ell^2}\\ =& (v, A^Tv)_{\ell^2},\\ =& (v, PP^{-1}Av)_{\ell^2}\\ =& (v, P^{-1}Av)_P \quad \forall v \in \mathbb{R}^n,\end{aligned}$$

die Selbstadjungiertheit der Matrix $P^{-1}A$ bezüglich $(\cdot,\cdot)_P$, sodass es eine Anwendung des MINRES-Verfahren auf das präkonditionierte System (6.7) wiederum möglich ist.

Eine Konvergenzaussage für das präkonditionierte MINRES-Verfahren, erhält man nun, in dem man in der Aussage von Satz 6.1 die ℓ^2-Norm durch $||\cdot||_P$ ersetzt (Beweis siehe z.B. [15]):

Satz 6.3 (Greenbaum). *Sei*

- *$A \in \mathbb{R}^{n\times n}$ symmetrisch und regulär,*
- *$P \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit,*

und sei x^k die durch das PMINRES-Verfahren erzeugte k-te Iterierte, dann gilt

$$||r^{2m}||_{P^{-1}} \leq \frac{2q^m}{1+q^{2m}}||r^0||_{P^{-1}},$$

wobei

$$\begin{aligned} r^k =& b - Ax^k \quad \textit{für} \quad k \geq 0, \\ q =& \frac{\kappa(P^{-1}A)-1}{\kappa(P^{-1}A)+1}, \end{aligned}$$

mit

$$\kappa(P^{-1}A) := ||P^{-1}A||_P||(P^{-1}A)^{-1}||_P.$$

6.4 Die Implementierung

Algorithmus 1 zeigt eine Möglichkeit der Implementierung des präkonditionierten MINRES-Verfahrens (siehe auch [16]):

Algorithmus 1 Präkonditioniertes MINRES-Verfahren

Input: Startvektor $x_0 \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$ symmetrisch und regulär, $P \in \mathbb{R}^{n\times n}$ symmetrisch und postiv definite Matrix, $b \in \mathbb{R}^n$ rechte Seite, $maxIter \in \mathbb{N}$ maximale Anzahl von Iterationen, $\varepsilon > 0$ Toleranz.

Output: Näherungslösung $x \in \mathbb{R}^n$, $iter \in \mathbb{N}$ Anzahl der benötigten Iterationen, $res \in \mathbb{R}$ mit $||r_{iter}||_{P^{-1}}$ und $flag \in \{0, 1\}$, wobei $flag = 0$ falls der Algorithmus konvergiert mit $iter <= maxIter$ und $flag = 1$ sonst.

Setze $v_1 = b - Ax_0$ und bestimme z_1 aus $Pz_1 = v_1$
Setze $\gamma_0 = 1$ und $\gamma_1 = \sqrt{(z_1, v_1)_{\ell^2}}$
Setze $\eta = \gamma_1$, $s_0 = s_1 = 0$, und $c_0 = c_1 = 1$
Setze $x = 0$, $iter = 0$, $res = 0$ und $flag = 1$
for $j = 1$ **to** $maxIter$ **step** 1 **do**
  $z_j = \gamma_j^{-1} z_j$
  $\delta_j = (Az_j, z_j)_{\ell^2}$
  $v_{j+1} = Az_j - \frac{\delta_j}{\gamma_j} v_j - \frac{\gamma_j}{\gamma_{j-1}} v_{j-1}$
  Löse $Pz_{j+1} = v_{j+1}$
  $\gamma_{j+1} = \sqrt{(z_{j+1}, v_{j+1})_{\ell^2}}$
  $\alpha_0 = c_j \delta_j - c_{j-1} s_j \gamma_j$
  $\alpha_1 = \sqrt{\alpha_0^2 + \gamma_{j+1}^2}$
  $\alpha_2 = s_j \delta_j + c_{j-1} s_j \gamma_j$
  $\alpha_3 = s_{j-1} \gamma_j$
  $c_{j+1} = \frac{\alpha_0}{\alpha_1}$; $s_{j+1} = \frac{\gamma_{j+1}}{\alpha_1}$
  $w_{j+1} = \alpha_1^{-1}(z_j - \alpha_3 w_{j-1} - \alpha_2 w_j)$
  $x_j = x_{j-1} + c_{j+1} \eta w_{j+1}$
  $\eta = -s_{j+1} \eta$
  if $\gamma_{j+1} \leq \varepsilon$ **then**
    $x = x_j$
    $iter = j$
    $res = \gamma_j$
    $flag = 0$
    break
  end if
end for
return $x, iter, res, flag$

Damit sind wir am Ende dieses Abschnittes angelangt und möchten an dieser Stelle für eine noch ausführlichere Untersuchung des MINRES- und PMINRES-Verfahrens auf [14] verweisen.

Kapitel 7

Robuste Präkonditionierung für den zeitharmonischen Fall

Wir sind nun beim Hauptresultat dieser Diplomarbeit angelangt, der Konstrunktion eines robusten Präkonditioniers für den zeitharmonischen Fall.

Für das gesamte Kapitel gelten die folgenden Annahmen:

$$\alpha > 0, \omega > 0, h > 0,$$

und die Annahme 1 (siehe Seite 39).

7.1 Das Ziel

Unter Verwendung der MINRES-Methode soll nun das präkonditionierte lineare System

$$P^{-1}\mathcal{A}^{\omega}x = P^{-1}b,$$

mit P **symmetrisch und positiv definit**, gelöst werden.

Dabei hängt die Anzahl der benötigten MINRES-Iterationen, um das Anfangsresidual $r_0 = b - Ax_0$ unter eine gegebene Schranke ε zu verkleinern, nach Satz 6.1 von der Konditionszahl $\kappa(P^{-1}\mathcal{A}^{\omega})$ ab.

Das **Ziel** ist nun:

> Konstruktion einer symmetrisch, positiv definiten Präkonditionierungsmatrix $\mathcal{P}$ in Block-Diagonalform, welche gegenüber den Parametern α, ω und der Feinheit h **robust** ist, d.h.:
>
> $$\exists C > 0 : \quad \kappa(\mathcal{P}^{-1}\mathcal{A}^{\omega}) := ||\mathcal{P}^{-1}\mathcal{A}^{\omega}||_{\mathcal{P}}||(\mathcal{P}^{-1}\mathcal{A}^{\omega})^{-1}||_{\mathcal{P}} < C, \qquad (7.1)$$
>
> wobei C eine von α, ω und h **unabhängige** Konstante ist.

Wir wollen nun zeigen, dass sich die Konditionszahl $\kappa(\mathcal{P}^{-1}\mathcal{A})$ auch mit Hilfe der Eigenwerte der Matrix $\mathcal{P}^{-1}\mathcal{A}$ ausdrücken lässt, dazu nutzen wir die Möglichkeit, dass die Konditionszahl $\kappa(\mathcal{P}^{-1}\mathcal{A})$ weiters in der Form

$$\begin{aligned}\kappa(\mathcal{P}^{-1}\mathcal{A}^\omega) :=& ||\mathcal{P}^{-1}\mathcal{A}^\omega||_{\mathcal{P}}||(\mathcal{P}^{-1}\mathcal{A}^\omega)^{-1}||_{\mathcal{P}} \\ =& ||\mathcal{P}^{\frac{1}{2}}(\mathcal{P}^{-1}\mathcal{A}^\omega)\mathcal{P}^{-\frac{1}{2}}||_{\ell_2}||(\mathcal{P}^{\frac{1}{2}}(\mathcal{P}^{-1}\mathcal{A}^\omega)\mathcal{P}^{-\frac{1}{2}})^{-1}||_{\ell_2},\end{aligned} \tag{7.2}$$

geschrieben werden kann.

Notation. Für eine symmetrisch und positiv definite Matrix $A \in \mathbb{R}^{n\times n}$ bezeichne $A^{\frac{1}{2}}$ jene existente symmetrisch und positiv definite Matrix, für welche gilt: $A = A^{\frac{1}{2}}A^{\frac{1}{2}}$ (siehe auch Defition 7.5).

Zunächst führen wir den Begriff der *Ähnlichkeit* von Matrizen ein:

Definition 7.1. *Sei $\mathcal{S}, \mathcal{T} \in \mathbb{R}^{n\times n}$. Dann bezeichnen wir die Matrizen $\mathcal{S}$ und $\mathcal{T}$ als zueinander ähnlich, falls gilt:*

$$\exists \mathcal{X} \in \mathbb{R}^{n\times n} \text{ regulär}: \quad \mathcal{T} = \mathcal{X}\mathcal{S}\mathcal{X}^{-1}.$$

In unserem Fall ist die Matrix $\mathcal{A}$ offensichtlich ähnlich zur Matrix $\mathcal{P}^{\frac{1}{2}}\mathcal{A}^\omega\mathcal{P}^{-\frac{1}{2}}$. Für ähnliche Matrizen, lässt sich dabei die folgende Aussage bezüglich der Eigenwerte zeigen:

Satz 7.2. *Sei $\mathcal{S}, \mathcal{T} \in \mathbb{R}^{n\times n}$ zueinander ähnliche Matrizen, das heißt*

$$\exists \mathcal{X} \in \mathbb{R}^{n\times n} \text{ regulär}: \quad \mathcal{T} = \mathcal{X}\mathcal{S}\mathcal{X}^{-1}.$$

Dann gilt:

Die Matrizen $\mathcal{S}$ und $\mathcal{T}$ besitzen dieselben Eigenwerte.

Beweis. (siehe auch [20], Seite 171) Sei $v \in \mathbb{R}^n$ Eigenvektor von $\mathcal{S}$ mit Eigenwert λ, dann gilt

$$\begin{aligned}\mathcal{S}v &= \lambda v, \\ &\Leftrightarrow \\ \mathcal{S}X^{-1}Xv &= \lambda X^{-1}Xv, \\ &\Leftrightarrow \\ X\mathcal{S}X^{-1}Xv &= \lambda Xv, \\ &\Leftrightarrow \\ \mathcal{T}Xv &= \lambda Xv.\end{aligned}$$

Also ist λ ein Eigenwert von $\mathcal{S}$ genau dann, falls λ Eigenwert von $\mathcal{T}$ ist. □

Unter Verwendung der letzten Aussage, folgt aus (7.2) für die Darstellung der Konditionszahl $\kappa(\mathcal{P}^{-1}\mathcal{A}^{\omega})$,

$$\begin{aligned}\kappa(\mathcal{P}^{-1}\mathcal{A}^{\omega}) =& ||\mathcal{P}^{\frac{1}{2}}(\mathcal{P}^{-1}\mathcal{A}^{\omega})\mathcal{P}^{-\frac{1}{2}}||_{\ell_2}||(\mathcal{P}^{\frac{1}{2}}(\mathcal{P}^{-1}\mathcal{A}^{\omega})\mathcal{P}^{-\frac{1}{2}})^{-1}||_{\ell_2}\\ =& ||\mathcal{P}^{-1}\mathcal{A}^{\omega}||_{\ell_2}||(\mathcal{P}^{-1}\mathcal{A}^{\omega})^{-1}||_{\ell_2}\\ =& \frac{|\lambda|_{\max}(\mathcal{P}^{-1}\mathcal{A}^{\omega})}{|\lambda|_{\min}(\mathcal{P}^{-1}\mathcal{A}^{\omega})}.\end{aligned}$$

Notation. Für eine Matrix $A \in \mathbb{R}^{n\times n}$ bezeichne dabei $|\lambda|_{\max}(A)$ den betragsgrößten und $|\lambda|_{\min}(A)$ den betragskleinsten Eigenwert von A.

7.2 Strategie für die Konstruktion des robusten Präkonditionierers für $\mathcal{A}^{\omega}$

Bei der Konstruktion eines robusten Präkonditionierer $\mathcal{P}$ für die Matrix $\mathcal{A}$ gehen wir nun folgend vor:

Zunächst konzentrieren wir uns nur auf den 1×1-Block A der Matrix $\mathcal{A}^{\omega}$: Dabei werden wir für die Matrix A einen **robusten, symmetrischen und positiv definiten** Präkonditionierer $\mathcal{Q}$ konstruieren.

Als Präkonditionierer für das Gesamtsystem $\mathcal{A}^{\omega}$ wählen wir dann:

$$\mathcal{P} = \begin{pmatrix}\mathcal{Q} & 0\\ 0 & B\mathcal{Q}^{-1}B^T\end{pmatrix},$$

wobei die Matrix $B\mathcal{Q}^{-1}B^T$ auch als *inexaktes Schur-Komplement* bezeichnet wird.

Der Nachweis der Tatsache, dass $\mathcal{P}$ einen **robusten** Präkonditionierer für die Matrix $\mathcal{A}^{\omega}$ bildet, erfolgt dann im letzten Schritt.

7.3 Berechnung der Präkonditionierungsmatrix Q

Die Matrix A besitzt die Blockgestalt

$$A = \begin{pmatrix}A_1 & B_1^T\\ B_1 & -A_1\end{pmatrix}, \tag{7.3}$$

mit

$$A_1 = \begin{pmatrix}C_h & 0\\ 0 & C_h\end{pmatrix} \quad \text{und} \quad B_1 = \begin{pmatrix}\sqrt{\alpha}A_h & \sqrt{\alpha}\omega C_h\\ \sqrt{\alpha}\omega C_h & -\sqrt{\alpha}A_h\end{pmatrix}.$$

und nach Satz 5.7 die Eigenschaften:

- $A_1 \in \mathbb{R}^{dn_h\times dn_h}$ ist symmetrisch und positiv definit,
- $B_1 \in \mathbb{R}^{m_h\times dn_h}$ ist symmetrisch und regulär.

7.3.1 Robuste Präkonditionierung von Sattelpunktsproblemen

Für Blockmatrizen in Sattelpunktsform (7.3), sind uns die folgenden Aussagen bezüglich der robusten Präkonditionierung bekannt (siehe [3]):

Satz 7.3. *Sei $m \leq n \in \mathbb{N}$,*

- *$A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit,*
- *$C \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semi-definit,*
- *$B \in \mathbb{R}^{m\times n}$ mit $\operatorname{rank}(B) = m$.*

Falls die Matrix

$$\mathcal{A} := \begin{pmatrix} A & B^T \\ B & -C \end{pmatrix}$$

präkonditioniert ist mit

$$\mathcal{P} = \begin{pmatrix} A & 0 \\ 0 & S \end{pmatrix} \quad \textit{und} \quad S = C + BA^{-1}B^T,$$

dann gilt für die präkonditionierte Matrix $\mathcal{P}^{-1}\mathcal{A}$:

$$\frac{\sqrt{5}-1}{2} \leq |\lambda|_{\min}(\mathcal{P}^{-1}\mathcal{A}) \quad \textit{und} \quad |\lambda|_{\max}(\mathcal{P}^{-1}\mathcal{A}) \leq \frac{\sqrt{5}+1}{2}.$$

Beweis. Wir betrachten das folgende Eigenwertproblem:

$$\begin{gathered} \mathcal{P}^{-1}\mathcal{A}\begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix} \\ \Leftrightarrow \\ \begin{pmatrix} I & A^{-1}B^T \\ S^{-1}B & -S^{-1}C \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix}, \\ \Leftrightarrow \\ \begin{aligned} &1.) \quad x + A^{-1}B^T y = \lambda x, \\ &2.) \quad S^{-1}Bx - S^{-1}Cy = \lambda y. \end{aligned} \end{gathered}$$

Aus 1.) folgt

$$(\lambda - 1)x = A^{-1}B^T y. \tag{7.4}$$

Wir unterscheiden nun zwei Fälle:

Fall $\lambda = 1$: Aus dem vollen Rang von B^T folgt $y = 0$. Falls $m < n$, wählen wir $0 \neq x \in \ker B$, somit ist auch 2.) erfüllt. Der Vektor $(x, 0)^T$ bildet dann einen Eigenvektor von $P^{-1}\mathcal{A}$ zum Eigenwert $\lambda = 1$. Für $m = n$ ist B regulär und somit $\lambda = 1$ kein Eigenwert von $\mathcal{P}^{-1}\mathcal{A}$, weil die Eigenwertgleichung nur für $x = y = 0$ erfüllt ist.

Fall $\lambda \neq 1$: Umformen von (7.4) führt auf

$$x = \frac{1}{\lambda - 1} A^{-1} B^T y.$$

Eingesetzt in 2.) mit $0 \neq y \in \mathbb{R}^m$ erhalten wir für λ

$$\begin{aligned}
\frac{1}{\lambda - 1} S^{-1} B A^{-1} B^T y - S^{-1} C y &= \lambda y, \\
&\Leftrightarrow \\
-\lambda C y &= (\lambda^2 - \lambda - 1) S y, \\
&\Leftrightarrow \\
-\lambda y^T C y &= (\lambda^2 - \lambda - 1) y^T S y, \\
&\Leftrightarrow \\
-\lambda y^T C y &= (\lambda^2 - \lambda - 1)(y^T C y + y^T B A^{-1} B^T y), \\
&\Leftrightarrow \\
\underbrace{y^T C y}_{\geq 0} &= -\frac{\lambda^2 - \lambda - 1}{\lambda} (\underbrace{y^T C_1 y}_{\geq 0} + \underbrace{y^T B A^{-1} B^T y}_{>0}).
\end{aligned}$$

Also

$$\frac{-\lambda^2 + \lambda + 1}{\lambda} = \rho,$$

wobei $\rho \in [0, 1)$. Durch Lösen der quadratischen Gleichung ergibt sich

$$\lambda(\rho) := -\frac{(\rho - 1) \pm \sqrt{(1 - \rho)^2 + 4}}{2}. \tag{7.5}$$

Für $\rho \in [0, 1)$ ist (7.5) jedoch monoton fallend, sodass gilt

$$\lambda \in (-1, \frac{1 - \sqrt{5}}{2}] \cup (1, \frac{1 + \sqrt{5}}{2}].$$

Die Eigenwerte von $\mathcal{P}^{-1} A$ liegen also in der Menge $(-1, \frac{1-\sqrt{5}}{2}] \cup [1, \frac{1+\sqrt{5}}{2}]$, woraus die Aussage folgt. □

Satz 7.4. *Sei $m \geq n \in \mathbb{N}$,*

- $A \in \mathbb{R}^{n \times n}$ *symmetrisch und positiv semi-definit,*
- $C \in \mathbb{R}^{n \times n}$ *symmetrisch und positiv definit,*
- $B \in \mathbb{R}^{m \times n}$ *mit* $\operatorname{rank}(B) = n$.

Falls die Matrix

$$\mathcal{A} := \begin{pmatrix} A & B^T \\ B & -C \end{pmatrix}$$

präkonditioniert ist mit

$$\mathcal{P} = \begin{pmatrix} R & 0 \\ 0 & C \end{pmatrix} \quad \textit{und} \quad R = A + B^T C^{-1} B,$$

dann gilt für die präkonditionierte Matrix $\mathcal{P}^{-1}A$*:*

$$\frac{\sqrt{5}-1}{2} \leq |\lambda|_{\min}(\mathcal{P}^{-1}A) \quad \textit{und} \quad |\lambda|_{\max}(\mathcal{P}^{-1}A) \leq \frac{\sqrt{5}+1}{2}. \tag{7.6}$$

Beweis. Durch Nachrechnen, lässt sich zeigen

$$\mathcal{A} = X\mathcal{T}X^{-1} \text{ und } \mathcal{P} = X\mathcal{S}X^{-1},$$

mit

$$X = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}, \quad \mathcal{T} = \begin{pmatrix} C & B^T \\ B & -A \end{pmatrix} \text{ und } \mathcal{S} = \begin{pmatrix} A + BC^{-1}B^T & 0 \\ 0 & C \end{pmatrix},$$

zeigen. Daraus folgt:

$$\mathcal{P}^{-1}\mathcal{A} = X\mathcal{S}^{-1}X^{-1}X\mathcal{T}X^{-1} = X\mathcal{S}^{-1}\mathcal{T}X^{-1},$$

das heißt die präkonditionierte Matrix $\mathcal{P}^{-1}\mathcal{A}$ ist ähnlich zur Matrix $\mathcal{S}^{-1}\mathcal{T}$ und besitzt somit nach Satz 7.2 die gleichen Eigenwerte. Für $\mathcal{S}^{-1}\mathcal{T}$ sind aber die Eigenwertabschätzungen aus (7.6) wegen Satz 7.3 erfüllt. □

7.3.2 Zwei Kandidaten Q_1 und Q_2

Auf die Matrix A lassen sich nun die Sätze 7.3 und 7.4 anwenden, sodass wir mit

$$Q_1 = \begin{pmatrix} A_1 & 0 \\ 0 & S_1 \end{pmatrix} \text{ und } Q_2 = \begin{pmatrix} R_1 & 0 \\ 0 & A_1 \end{pmatrix} = \begin{pmatrix} S_1 & 0 \\ 0 & A_1 \end{pmatrix}, \tag{7.7}$$

wobei

$$\begin{aligned} S_1 =& A_1 + B_1 A_1^{-1} B_1 \\ =& \begin{pmatrix} C_h & 0 \\ 0 & C_h \end{pmatrix} + \begin{pmatrix} \sqrt{\alpha}A_h & \sqrt{\alpha}\omega C_h \\ \sqrt{\alpha}\omega C_h & -\sqrt{\alpha}A_h \end{pmatrix} \\ & \begin{pmatrix} C_h^{-1} & 0 \\ 0 & C_h^{-1} \end{pmatrix} \begin{pmatrix} \sqrt{\alpha}A_h & \sqrt{\alpha}\omega C_h \\ \sqrt{\alpha}\omega C_h & -\sqrt{\alpha}A_h \end{pmatrix} \\ =& ((1+\alpha\omega^2)C_h + \alpha A_h C_h^{-1} A_h) \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}, \end{aligned} \tag{7.8}$$

und

$$\kappa(Q_i^{-1}A) = \frac{|\lambda|_{\max}(Q_i^{-1}A)}{|\lambda|_{\min}(Q_i^{-1}A)} \leq \frac{\sqrt{5}+1}{\sqrt{5}-1} < 4 \quad \text{für } i = 1, 2,$$

gleich zwei Kandidaten für einen **robusten, symmetrischen und positiv definiten** Präkonditionierer in Block-Diagonalform der Matrix A.

Die Matrizen Q_1 und Q_2 beinhalten dabei mit

$$A_h C_h^{-1} A_h$$

einen Term, welcher der Diskretisierung eines Differentialoperators der 4. Ordnung entspricht. Im Optimalitätssystem (OP) treten jedoch keine Differentialoperatoren mit höherer Ordnung als zwei auf, sodass man annehmen könnte, dass auch eine zum diskretisierten System dazugehörige Präkonditionierungsmatrix existiert, welche keine Terme mit höherer Ordnung enthält.

Tatsächlich lässt sich durch Interpolation von Q_1 und Q_2 ein weiterer Präkonditionierungskandidat der Matrix A konstruieren, dessen Terme nur der Diskretisierung eines Differentialoperators 2. Ordnung entsprechen.

7.3.3 Interpolation von Q_1 und Q_2

Wir werden in diesem Abschnitt nur das Hauptresultat der Interpolation im $\mathbb{R}^n$, den sogenannten *Interpolationssatz* angeben und anwenden. Für eine ausführliche Einführung in die Interpolationstheorie verweisen wir an dieser Stelle auf [2].

Für das Verständnis des *Interpolationsatzes* ist es zunächst notwendig, den Begriff der reellen Potenz einer symmetrisch und positiv definiten Matrix einzuführen:

Definition 7.5. *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, $\theta \in \mathbb{R}$ und $Q, D \in \mathbb{R}^{n\times n}$ die nach dem Spektralsatz (siehe Satz 4.7) existierenden Matrizen, sodass gilt*

$$A = QDQ^T.$$

Dann definieren wir die θ-te Potenz von A als

$$A^\theta = QD^\theta Q^T,$$

wobei

$$D^\theta = \begin{pmatrix} \lambda_1^\theta & 0 & \dots & 0 \\ 0 & \lambda_2^\theta & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n^\theta \end{pmatrix}.$$

Wir sind nun in der Lage das Hauptresultat der Interpolation im $\mathbb{R}^n$, zu formulieren (Beweis siehe z.B. [2]):

Satz 7.6 (Interplationssatz für $\mathbb{R}^n$). *Es seien $M_i, N_i \in \mathbb{R}^{n\times n}$ symmetrische und positiv definite Matrizen für $i \in \{0,1\}$, $0 < \theta < 1$ und*

$$T : \mathbb{R}^n \longrightarrow \mathbb{R}^n,$$

mit

$$||Tx||_{N_0} \leq c_0||x||_{M_0} \quad mit \quad ||Tx||_{N_1} \leq c_1||x||_{M_1} \quad \forall x \in \mathbb{R}^n,$$

wobei die Normen $||\cdot||_{M_i}$ und $||\cdot||_{N_i}$ induziert werden durch die inneren Produkte

$$(x,y)_{M_i} := (M_i x, y)_{\mathbb{R}^n} \quad und \quad (x,y)_{N_i} := (N_i x, y)_{\mathbb{R}^n} \quad für \quad i = 0,1.$$

Dann gilt

$$||Tx||_{N_\theta} \leq c_0^{1-\theta} c_1^{\theta} ||x||_{M_\theta},$$

wobei die Normen $||\cdot||_{M_\theta}$ und $||\cdot||_{N_\theta}$ induziert werden durch die inneren Produkte

$$\begin{aligned}(x,y)_{M_\theta} &:= (M_\theta x, y)_{\mathbb{R}^n} \quad mit \quad M_\theta = M_0^{\frac{1}{2}}(M_0^{-\frac{1}{2}} M_1 M_0^{-\frac{1}{2}})^\theta M_0^{\frac{1}{2}},\\ (x,y)_{N_\theta} &= (N_\theta x, y)_{\mathbb{R}^n} \quad mit \quad N_\theta = N_0^{\frac{1}{2}}(N_0^{-\frac{1}{2}} N_1 N_0^{-\frac{1}{2}})^\theta N_0^{\frac{1}{2}}.\end{aligned}$$

induziert werden. Weiters gelten für $\theta = \frac{1}{2}$ die Zusammenhänge:

$$\begin{aligned} M_0^{\frac{1}{2}}(M_0^{-\frac{1}{2}} M_1 M_0^{-\frac{1}{2}})^\theta M_0^{\frac{1}{2}} &= M_1^{\frac{1}{2}}(M_1^{-\frac{1}{2}} M_0 M_1^{-\frac{1}{2}})^{\frac{1}{2}} M_1^{\frac{1}{2}} \\ &\quad und \\ N_0^{\frac{1}{2}}(N_0^{-\frac{1}{2}} N_1 N_0^{-\frac{1}{2}})^\theta N_0^{\frac{1}{2}} &= N_1^{\frac{1}{2}}(N_1^{-\frac{1}{2}} N_0 N_1^{-\frac{1}{2}})^{\frac{1}{2}} N_1^{\frac{1}{2}}. \end{aligned} \tag{7.9}$$

Als nächstes erfolgt die Anwendung des Interpolationssatzes auf die beiden Kandidaten Q_1 und Q_2 aus (7.7) an:

Sei

$$\begin{aligned} M_0 = Q_1, \quad M_1 = Q_2, \\ N_0 = Q_1^{-1} \quad N_1 = Q_2^{-1}, \end{aligned} \tag{7.10}$$

dann gilt für $z \in \mathbb{R}^{4dn_h}$:

$$\begin{aligned} ||Az||_{Q_1^{-1}} &= ||Q_1^{-1}A||_{Q_1} \leq c_1||z||_{Q_1}, \\ ||Az||_{Q_2^{-1}} &= ||Q_2^{-1}A||_{Q_2} \leq c_1||z||_{Q_2}, \end{aligned}$$

$$\begin{aligned} ||A^{-1}y||_{Q_1} &\leq c_0||y||_{Q_1^{-1}} = c_0||Q_1^{-1}y||_{Q_1}, \\ ||A^{-1}y||_{Q_2} &\leq c_0||y||_{Q_2^{-1}} = c_0||Q_2^{-1}y||_{Q_2}, \end{aligned} \tag{7.11}$$

mit

$$c_0 = \frac{2}{\sqrt{5}-1} \quad \text{und} \quad c_1 = \frac{1+\sqrt{5}}{2}. \tag{7.12}$$

Mit (7.10) - (7.12) sind also für

$$T_0(x) := Ax \quad \text{and} \quad T_1(x) := A^{-1}x,$$

und der Wahl

$$\theta = \frac{1}{2},$$

alle Voraussetzungen des *Interpolationssatzes* erfüllt, sodass gilt:

1.

$$\begin{aligned} ||T_0 z||_{N_{1/2}} &= ||Az||_{N_{1/2}} = ||Az||_{M_{1/2}^{-1}} \\ &= ||M_{1/2}^{-1} Az||_{M_{1/2}} \\ &\leq \left(\frac{\sqrt{5}+1}{\sqrt{5}-1}\right)^{\frac{1}{2}} ||z||_{M_{1/2}} \quad \forall z \in \mathbb{R}^{4dn_h}, \\ &\Leftrightarrow \\ ||M_{1/2}^{-1} A||_{M_{1/2}} &\leq \left(\frac{\sqrt{5}+1}{\sqrt{5}-1}\right)^{\frac{1}{2}}, \end{aligned} \tag{7.13}$$

2.

$$\begin{aligned} ||T_1 z||_{N_{1/2}^{-1}} &= ||A^{-1} z||_{N_{1/2}^{-1}} = ||A^{-1} z||_{M_{1/2}} \\ &\leq \left(\frac{\sqrt{5}+1}{\sqrt{5}-1}\right)^{\frac{1}{2}} ||z||_{M_{1/2}^{-1}} \\ &\leq \left(\frac{\sqrt{5}+1}{\sqrt{5}-1}\right)^{\frac{1}{2}} ||M_{1/2} z||_{M_{1/2}} \quad \forall z \in \mathbb{R}^{4dn_h}, \\ &\Leftrightarrow \\ ||(M_{1/2}^{-1} A)^{-1} z||_{M_{1/2}} &\leq \left(\frac{\sqrt{5}+1}{\sqrt{5}-1}\right)^{\frac{1}{2}} ||z||_{M_{1/2}} \quad \forall z \in \mathbb{R}^{4dn_h}, \\ &\Leftrightarrow \\ ||(M_{1/2}^{-1} A)^{-1}||_{M_{1/2}} &\leq \left(\frac{\sqrt{5}+1}{\sqrt{5}-1}\right)^{\frac{1}{2}}, \end{aligned} \tag{7.14}$$

mit

$$M_{1/2} = Q_1^{-\frac{1}{2}} (Q_1^{\frac{1}{2}} Q_2^{-1} Q_1^{\frac{1}{2}})^{\frac{1}{2}} Q_1^{-\frac{1}{2}} \quad \text{und} \quad N_{1/2} = M_{1/2}^{-1}.$$

Aus (7.13) und (7.14) folgt dabei, sofort die Robustheit von $M_{1/2}$:

$$\kappa(M_{1/2}^{-1}A) := ||M_{1/2}^{-1}A||_{M_{1/2}}||(M_{1/2}^{-1}A)^{-1}||_{M_{1/2}} \leq \left(\frac{\sqrt{5}+1}{\sqrt{5}-1}\right).$$

Für die Darstellung von $M_{1/2}$ gilt einerseits

$$\begin{aligned} M_{1/2} =& Q_1^{-\frac{1}{2}}(Q_1^{\frac{1}{2}}Q_2^{-1}Q_1^{\frac{1}{2}})^{\frac{1}{2}}Q_1^{-\frac{1}{2}} \\ =& \begin{pmatrix} A_1^{\frac{1}{2}} & 0 \\ 0 & S_1^{\frac{1}{2}} \end{pmatrix} \left(\left(\begin{pmatrix} A_1^{-\frac{1}{2}} & 0 \\ 0 & S_1^{-\frac{1}{2}} \end{pmatrix} \begin{pmatrix} R_1 & 0 \\ 0 & C_1 \end{pmatrix} \right. \right. \\ & \left. \left. \begin{pmatrix} A_1^{-\frac{1}{2}} & 0 \\ 0 & S_1^{-\frac{1}{2}} \end{pmatrix} \right)^{\frac{1}{2}} \begin{pmatrix} A_1^{\frac{1}{2}} & 0 \\ 0 & S_1^{\frac{1}{2}} \end{pmatrix} \right. \\ =& \begin{pmatrix} C_1^{\frac{1}{2}}(C_1^{-\frac{1}{2}}S_1C_1^{-\frac{1}{2}})^{\frac{1}{2}}C_1^{\frac{1}{2}} & 0 \\ 0 & S_1^{\frac{1}{2}}(S_1^{-\frac{1}{2}}C_1S_1^{-\frac{1}{2}})^{\frac{1}{2}}S_1^{\frac{1}{2}} \end{pmatrix}. \end{aligned} \tag{7.15}$$

und andererseits folgt für $\theta = \frac{1}{2}$ aus (7.9)

$$\begin{aligned} M_{1/2} =& Q_2^{-1/2}(Q_2^{1/2}Q_1^{-1}Q_2^{1/2})Q_2^{-1/2} \\ =& \begin{pmatrix} S_1^{\frac{1}{2}}(S_1^{-\frac{1}{2}}C_1S_1^{-\frac{1}{2}})^{\frac{1}{2}}S_1^{\frac{1}{2}} & 0 \\ 0 & C_1^{\frac{1}{2}}(C_1^{-\frac{1}{2}}S_1C_1^{-\frac{1}{2}})^{\frac{1}{2}}C_1^{\frac{1}{2}} \end{pmatrix}. \end{aligned} \tag{7.16}$$

Setzen wir die Darstellungen aus (7.15) und (7.16) gleich, so erhält man mit der Darstellung für die Matrix S_1 aus (7.8):

$$\begin{aligned} M_{1/2} =& \left(C_1^{\frac{1}{2}}(C_1^{-\frac{1}{2}}S_1C_1^{-\frac{1}{2}})^{\frac{1}{2}}C_1^{\frac{1}{2}}\right) \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \\ =& \Big(C_h^{\frac{1}{2}}\Big(C_h^{-\frac{1}{2}}\Big((1+\alpha\omega^2)C_h \\ & + \alpha A_h C_h^{-1}A_h\Big)C_h^{-\frac{1}{2}}\Big)^{\frac{1}{2}}C_h^{\frac{1}{2}}\Big) \begin{pmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{pmatrix} \\ =& \begin{pmatrix} I_{1/2} & 0 & 0 & 0 \\ 0 & I_{1/2} & 0 & 0 \\ 0 & 0 & I_{1/2} & 0 \\ 0 & 0 & 0 & I_{1/2} \end{pmatrix}, \end{aligned} \tag{7.17}$$

wobei

$$I_{1/2} = C_h^{\frac{1}{2}}\Big((1+\alpha\omega^2)I + \alpha C_h^{-\frac{1}{2}}A_hC_h^{-\frac{1}{2}}C_h^{-\frac{1}{2}}A_hC_h^{-\frac{1}{2}}\Big)^{\frac{1}{2}}C_h^{\frac{1}{2}}. \tag{7.18}$$

Die Matrix $M_{1/2}$ bildet also einen weiteren Kandidaten eines robusten, symmetrischen und positiv definiten Präkonditioniers in Block-Diagonalform. Dabei tritt mit

$$C_h^{-\frac{1}{2}} A_h C_h^{-1} A_h C_h^{-\frac{1}{2}} = (C_h^{-\frac{1}{2}} A_h C_h^{-\frac{1}{2}})^2,$$

wiederum ein Term auf, welcher der Diskretisierung eines Differentialoperators der 4. Ordnung entspricht, jedoch nun unter der „Wurzel". Ziel ist es, den „Wurzel"-Ausdruck in $I_{1/2}$ weiter zu vereinfachen, ohne die Eigenschaften der Robustheit zu verlieren, dazu definieren wir zunächst den Begriff der spektralen Äquivalenz:

Definition 7.7. *Seien $A, C \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definite Matrizen. Dann heißen die Matrizen A und C zueinander spektral äquivalent, falls gilt:*

$$\exists \gamma_0, \gamma_1 > 0: \quad \gamma_0 (Cv, v)_{\ell^2} \leq (Av, v)_{\ell^2} \leq \gamma_1 (Cv, v)_{\ell^2} \quad \forall v \in \mathbb{R}^n.$$

Die Konstanten γ_0, γ_1 werden auch als Spektralkonstanten bezeichnet. Weiters verwenden wir die Schreibweisen

$$\gamma_0 C \leq A \leq \gamma_1 C$$

oder kurz $A \sim C$.

Dabei folgt aus $A \sim B$ offensichtlich $B \sim A$. Besonders hilfreich zur Auflösung des „Wurzel"-Ausdrucks in $I_{1/2}$ ist nun die folgende Aussage:

Satz 7.8. *Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrisch und positiv definite Matrix. Dann gilt:*

$$\frac{1}{\sqrt{2}}(I + A^{\frac{1}{2}}) \leq (I + A)^{\frac{1}{2}} \leq I + A^{\frac{1}{2}}.$$

Beweis. Wir zeigen zuerst

$$\frac{1}{\sqrt{2}}(1 + \sqrt{x}) \leq \sqrt{1 + x} \leq 1 + \sqrt{x} \quad \forall x \geq 0. \tag{7.19}$$

Sei $x \geq 0$, $v_1 = \begin{pmatrix}\sqrt{x} & 1\end{pmatrix}^T$ und $v_2 = \frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\end{pmatrix}^T$, dann folgt aus der Cauchy Ungleichung:

$$|(v_1, v_2)_{\ell^2}| = \frac{1}{\sqrt{2}}(1 + \sqrt{x}) \leq ||v_1||_{\ell^2} ||v_2||_{\ell^2} = \sqrt{1 + x},$$

und weiters

$$\sqrt{1 + x} \leq \sqrt{1 + 2\sqrt{x} + x} = \sqrt{(1 + \sqrt{x})^2} = 1 + \sqrt{x},$$

womit die Ungleichung (7.19) gezeigt wäre.

Aus dem *Spektralsatz* erhält man zunächst für die Matrix A die Darstellung

$$A = QDQ^T,$$

wobei Q orthogonal und D jene Diagonalmatrix bezeichnet, welche die Eigenwerte von A enthält. Mit

$$\begin{aligned} I + A^{\frac{1}{2}} =& QQ^T + QD^{\frac{1}{2}}Q^T \\ =& Q\,\mathrm{diag}(1+\sqrt{\lambda_1}, 1+\sqrt{\lambda_2}, \ldots, 1+\sqrt{\lambda_n})Q^T \end{aligned}$$

und

$$\begin{aligned} (I + A)^{\frac{1}{2}} =& (QQ^T + QDQ^T)^{\frac{1}{2}} \\ =& Q(I+D)^{\frac{1}{2}}Q^T \\ =& Q\,\mathrm{diag}(\sqrt{1+\lambda_1}, \sqrt{1+\lambda_2}, \ldots, \sqrt{1+\lambda_n})Q^T \end{aligned}$$

folgt weiters

$$\begin{aligned} \frac{1}{\sqrt{2}}((I + A^{\frac{1}{2}})v, v)_{\ell^2} &= (\mathrm{diag}(1+\sqrt{\lambda_1}, 1+\sqrt{\lambda_2}, \ldots, 1+\sqrt{\lambda_n})Qv, Qv)_{\ell^2} \\ &\overset{(7.19)}{\leq} (\mathrm{diag}(\sqrt{1+\lambda_1}, \sqrt{1+\lambda_2}, \ldots, \sqrt{1+\lambda_n})Qv, Qv)_{\ell^2} \\ &= ((I + A)^{\frac{1}{2}})v, v)_{\ell^2} \\ &\overset{(7.19)}{\leq} (\mathrm{diag}(1+\sqrt{\lambda_1}, 1+\sqrt{\lambda_2}, \ldots, 1+\sqrt{\lambda_n})Qv, Qv)_{\ell^2} \\ &= ((I + A^{\frac{1}{2}})v, v)_{\ell^2} \quad \forall v \in \mathbb{R}^n, \end{aligned}$$

womit die Aussage gezeigt wäre. □

Unter Anwendung des letzten Satzes erhalten wir für

$$I_{1/2} = C_h^{\frac{1}{2}} \Big(\underbrace{(1+\alpha\omega^2)I + \alpha C_h^{-\frac{1}{2}} A_h C_h^{-\frac{1}{2}} C_h^{-\frac{1}{2}} A_h C_h^{-\frac{1}{2}}}_{\tilde{I}_{1/2}} \Big)^{\frac{1}{2}} C_h^{\frac{1}{2}},$$

aus (7.18):

- $$\begin{aligned} \frac{1}{\sqrt{2}}(I_{1/2}v, v)_{\ell^2} &= \frac{1}{\sqrt{2}}(C_h^{\frac{1}{2}}\tilde{I}_{1/2}v, C_h^{\frac{1}{2}}v)_{\ell^2} \\ &\overset{(7.8)}{\leq} (((1+\alpha\omega^2)^{\frac{1}{2}}I + \sqrt{\alpha}C_h^{-\frac{1}{2}}A_h C_h^{-\frac{1}{2}})C_h^{\frac{1}{2}}v, C_h^{\frac{1}{2}}v)_{\ell^2} \\ &= (((1+\alpha\omega^2)^{\frac{1}{2}}C_h + \sqrt{\alpha}A_h)v, v)_{\ell^2} \\ &\overset{(7.8)}{\leq} (I_{1/2}v, v)_{\ell^2} \quad \forall v \in \mathbb{R}^{n_h}, \end{aligned}$$

- $$\begin{aligned}\frac{1}{\sqrt{2}}(((1+\sqrt{\alpha}\omega)C_h+\sqrt{\alpha}A_h)v,v)_{\ell^2}&\\ &= \frac{1}{\sqrt{2}}(1+\sqrt{\alpha}\omega)(C_h^{\frac{1}{2}}v,C_h^{\frac{1}{2}}v)_{\ell^2}+\sqrt{\frac{\alpha}{2}}(A_hv,v)_{\ell^2}\\ &\overset{(7.8)}{\leq} (((1+\alpha\omega^2)^{\frac{1}{2}}C_h+\sqrt{\alpha}A_h)v,v)_{\ell^2}\\ &= (1+\alpha\omega^2)^{\frac{1}{2}}(C_h^{\frac{1}{2}}v,C_h^{\frac{1}{2}}v)_{\ell^2}+\sqrt{\alpha}(A_hv,v)_{\ell^2}\\ &\overset{(7.8)}{\leq} (((1+\sqrt{\alpha}\omega)C_h+\sqrt{\alpha}A_h)v,v)_{\ell^2}\quad \forall v\in\mathbb{R}^{n_h},\end{aligned}$$

sodass

$$\begin{aligned}\frac{1}{2}(I_{1/2}v,v)_{\ell^2} &\leq (((1+\sqrt{\alpha}\omega)C_h+\sqrt{\alpha}A_h)v,v)_{\ell^2}\\ &\leq (I_{1/2}v,v)_{\ell^2}\quad \forall v\in\mathbb{R}^{n_h}.\end{aligned} \tag{7.20}$$

Die Wahl von Q

Sei

$$Q=\big((1+\sqrt{\alpha}\omega)C_h+\sqrt{\alpha}A_h\big)\begin{pmatrix}I&0&0&0\\0&I&0&0\\0&0&I&0\\0&0&0&I\end{pmatrix}, \tag{7.21}$$

dann gilt offensichtlich Q ist symmetrisch positiv definit und weiters folgt aus (7.20)

$$\frac{1}{2}Q\leq M_{1/2}\leq Q. \tag{7.22}$$

Die Matrix Q besitzt also nur noch Terme die der Diskretisierung eines Differentialoperators der 2. Ordnung entsprechen. Um sie für die Präkonditionierung von A zu verwenden, bleibt allerdings noch ihre Robustheit zu zeigen.

Robustheit von Q

Präkonditioniert man eine Matrix, bei einer zu ihr spektral äquivalenten Matrix so lässt sich für die Konditionszahl die folgende Abschätzung zeigen:

Satz 7.9. *Seien $A,C\in\mathbb{R}^{n\times n}$ symmetrisch positiv definite Matrizen, welche zueinander spektral äquivalent sind, also*

$$\gamma_0 C\leq A\leq\gamma_1 C.$$

Dann gilt:

$$\kappa(C^{-1}A):=||C^{-1}A||_{\ell^2}||(A^{-1}C)^{-1}||_{\ell^2}=\frac{\lambda_{\max}(A^{-1}C)}{\lambda_{\min}(A^{-1}C)}\leq\frac{\gamma_1}{\gamma_0}.$$

Beweis. Aus

$$\gamma_0(Cv,v)_{\ell^2} \leq (Av,v)_{\ell^2} \leq \gamma_1(Cv,v)_{\ell^2} \quad \forall v \in \mathbb{R}^n$$
$$\Leftrightarrow$$
$$\gamma_0(v,v)_{\ell^2} \leq (C^{\frac{1}{2}}(C^{-1}A)C^{-\frac{1}{2}}v,v)_{\ell^2} \leq \gamma_1(v,v)_{\ell^2} \quad \forall v \in \mathbb{R}^n$$

und der Gleichheit der Eigenwerte von ähnlichen Matrizen (Satz 7.2), erhält man

$$\gamma_0 \leq \lambda_{\min}(C^{\frac{1}{2}}C^{-1}AC^{-\frac{1}{2}}) = \lambda_{\min}(C^{-1}A)$$
$$\text{und}$$
$$\lambda_{\max}(C^{\frac{1}{2}}C^{-1}AC^{-\frac{1}{2}}) = \lambda_{\max}(C^{-1}A) \leq \gamma_1,$$

womit die Aussage gezeigt wäre. □

Somit gilt für die Matrix Q:

$$\begin{aligned}
\kappa(Q^{-1}A) &= ||Q^{-1}A||_Q||(Q^{-1}A)^{-1}||_Q \\
&= ||Q^{-1}A||_{\ell^2}||(Q^{-1}A)^{-1}||_{\ell^2} \\
&\leq ||Q^{-1}M_{1/2}||_{\ell^2}||(Q^{-1}M_{1/2})||_{\ell^2}||M_{1/2}^{-1}A||_{\ell^2}||(M_{1/2}^{-1}A)^{-1}||_{\ell^2} \\
&= \kappa(Q^{-1}M_{1/2})\kappa(M^{-1/2}A) \\
&\overset{(7.9),\ (7.22)}{\leq} \frac{1}{\frac{1}{2}}\frac{\sqrt{5}+1}{\sqrt{5}-1} \leq 6.
\end{aligned}$$

Zusammenfassend wurde also gezeigt:

Satz 7.10. *Es gilt:*

> *Die Matrix Q definiert in (7.21) bildet einen robusten, symmetrisch und positiv definiten Präkonditioner für A.*

Damit wäre der 1. Schritt in unserer Optimierungsstrategie abgeschlossen.

7.4 Wahl der Präkonditionierungsmatrix $\mathcal{P}$

Nach der Ersetzung von A durch Q, besitzt $\mathcal{P}$ also die folgende Gestalt:

$$\mathcal{P} = \begin{pmatrix} \boldsymbol{Q} & 0 \\ 0 & B\boldsymbol{Q}^{-1}B^T \end{pmatrix}, \tag{7.23}$$

mit

$$Q = \begin{pmatrix} I_h & 0 & 0 & 0 \\ 0 & I_h & 0 & 0 \\ 0 & 0 & I_h & 0 \\ 0 & 0 & 0 & I_h \end{pmatrix},$$

und

$$I_h = (1 + \sqrt{\alpha}\omega)C_h + \sqrt{\alpha}A_h.$$

Aus Satz 7.10 und der Eigenschaft des vollen Ranges von B folgt:

Satz 7.11. *Es gilt:*

Die Matrix $\mathcal{P}$ definiert in (7.23) ist symmetrisch und positiv definit.

Damit wäre auch der zweite Schritt unser Strategie erledigt und wir können und dem Nachweis der Robustheit von $\mathcal{P}$ zuwenden.

7.5 Robustheit der Präkonditionierungsmatrix $\mathcal{P}$

7.5.1 Das äquivalente Operatorproblem

Zunächst definieren wir den linearen Operator

$$\begin{aligned} L :& X \times Y \to X^\star \times Y^\star, \\ \langle Lu, w\rangle :=& (\mathcal{A}^\omega u, w)_{\ell_2} = \left(\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} u, w \right)_{\ell_2}, \end{aligned} \tag{7.24}$$

mit

- $X = \mathbb{R}^{4dn_h}$,

$$(u, v)_X := (Qu, v)_{\ell_2} \quad \forall u, v \in X,$$

- $Y = \mathbb{R}^{4n_h}$,

$$(p, q)_Y := (BQ^{-1}B^T p, q)_{\ell_2} \quad \forall p, q \in Y,$$

- $X \times Y = \mathbb{R}^{4dn_h} \times \mathbb{R}^{4n_h}$,

$$\begin{aligned} (w, z)_{X\times Y} :=& (\mathcal{P}w, z)_{\ell_2} = (u, x)_X + (v, y)_Y, \\ \forall w =& (u, v), z = (x, y) \in X \times Y. \end{aligned}$$

Aus

$$\begin{aligned} ||w||_{X\times Y} &= \sqrt{(w, w)_{X\times Y}} \\ &= \sqrt{(\mathcal{P}w, w)_{\ell_2}} \\ &= ||w||_{\mathcal{P}} \quad \forall w \in X \times Y, \end{aligned}$$

und

$$\begin{aligned}||Lw||_{X^\star\times Y^\star} &= \sup_{0\neq z\in X\times Y}\frac{\langle Lw,z\rangle}{||z||_{X\times Y}}\\ &= \sup_{0\neq z\in X\times Y}\frac{(\mathcal{A}^\omega w,z)_{\ell_2}}{||z||_{\mathcal{P}}}\\ &= \sup_{0\neq z\in X\times Y}\frac{(\mathcal{P}(\mathcal{P}^{-1}\mathcal{A}^\omega)w,z)_{\ell_2}}{\sqrt{(\mathcal{P}z,z)_{\ell_2}}}\\ &=||\mathcal{P}^{-1}\mathcal{A}^\omega w||_{\mathcal{P}}\quad\forall w\in X\times Y,\end{aligned}$$

folgt dann für $C_1, C_2 > 0$:

$$\begin{aligned}&C_1||w||_{X\times Y}\leq||Lw||_{X^\star\times Y^\star}\leq C_2||w||_{X\times Y}\quad\forall w\in X\times Y,\\ &\qquad\qquad\Leftrightarrow\\ &C_1||w||_{\mathcal{P}}\leq||\mathcal{P}^{-1}\mathcal{A}^\omega w||_{\mathcal{P}}\leq C_2||w||_{\mathcal{P}}\quad\forall w\in X\times Y,\\ &\qquad\qquad\Downarrow\\ &\kappa(\mathcal{P}^{-1}\mathcal{A}^\omega):=||\mathcal{P}^{-1}\mathcal{A}^\omega||_{\mathcal{P}}||(\mathcal{P}^{-1}\mathcal{A}^\omega)^{-1}||_{\mathcal{P}}\leq\frac{C_2}{C_1}.\end{aligned}\tag{7.25}$$

Gelingt es uns also zeigen, dass der in (7.24) definierte Operator L einen Isomorphimus mit **robusten** Konstanten C_1 und C_2 bildet, d.h. L ist bijektiv und

$$C_1||w||_{X\times Y}\leq||Lw||_{X^\star\times Y^\star}\leq C_2||w||_{X\times Y}\quad\forall w\in X\times Y.\tag{7.26}$$

wobei $C_1, C_2 > 0$ unabhängig von α, ω und h, dann würde aus (7.25) sofort die Robustheit von $\mathcal{P}$ folgen.

Eine Aussage über die Existenz und Stetigkeit einer Inversen für lineare Operatoren der Gestalt (7.24), liefert nun der folgende Satz (Beweis siehe [23], Seite 42).

7.5.2 Satz von Brezzi und verbesserte Version

Satz 7.12 (Brezzi). *Seien*

- X, Y *Hilberträume,*
- $F \in X^\star$, $G \in Y^\star$,
- $A : X \to X^\star$, $B : X \to Y^\star$ *lineare Operatoren.*

Weiters gelte

1. $\exists\alpha_2 > 0$*:*

$$\begin{aligned}&||Au||_{Y^\star}\leq\alpha_2||u||_X\quad\forall u\in X,\\ &\qquad\Leftrightarrow\\ &|\langle Au,v\rangle_{X^\star\times X}|\leq\alpha_2||u||_X||v||_X\quad\forall u,v\in X,\end{aligned}$$

2. $\exists \beta_2 > 0$*:*

$$\begin{gathered} ||Bu||_{Y^\star} \leq \beta_2 ||u||_X \quad \forall u \in X, \\ \Leftrightarrow \\ |\langle Bu, v\rangle_{Y^\star \times Y}| \leq \beta_2 ||u||_X ||v||_Y \quad \forall u \in X, \forall v \in Y, \end{gathered}$$

3. $\exists \alpha_1 > 0$*:*

$$\inf_{0 \neq w \in X_0} \sup_{0 \neq v \in X_0} \frac{\langle Av, w\rangle_{X \times X^\star}}{||v||_X ||w||_X} \geq \alpha_1 \quad \wedge \quad \inf_{0 \neq v \in X_0} \sup_{0 \neq w \in X_0} \frac{\langle A^\star w, v\rangle_{X^\star \times X}}{||v||_X ||w||_X} \geq \alpha_1,$$

mit $X_0 = \operatorname{kern} B := \{v \in X \mid \langle Bv, q\rangle_{Y^\star \times Y} = 0 \quad \forall q \in Y\}$.

4. $\exists \beta_1 > 0$*:*

$$\begin{gathered} ||B^\star q||_{X^\star} \geq \beta_1 ||q||_Y \quad \forall q \in X, \\ \Leftrightarrow \\ \inf_{0 \neq q \in Y} \sup_{0 \neq v \in X} \frac{\langle Bv, q\rangle_{Y^\star \times Y}}{||v||_X ||q||_Y} \geq \beta_1. \end{gathered}$$

Dann besitzt die Operatorgleichung

$$L \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} F \\ G \end{pmatrix},$$

mit

$$L : X \times Y \longrightarrow X^\star \times Y^\star$$

$$L := \begin{pmatrix} A & B^\star \\ B & 0 \end{pmatrix},$$

eine eindeutige Lösung $(u, p) \in X \times Y$*, wobei gilt:*

$$\begin{aligned} ||u||_X &\leq \frac{1}{\alpha_1} ||F||_{X^\star} + \frac{1}{\beta_1} \left(\mathbf{1} + \frac{\alpha_2}{\alpha_1} \right) ||G||_{Y^\star}, \\ ||p||_Y &\leq \frac{1}{\beta_1} \left(\mathbf{1} + \frac{\alpha_2}{\alpha_1} \right) ||F||_{X^\star} + \frac{\alpha_2}{\beta_1^2} \left(\mathbf{1} + \frac{\alpha_2}{\alpha_1} \right) ||G||_{Y^\star}. \end{aligned} \tag{7.27}$$

Die Abschätzungen aus (7.27) vom Satz von Brezzi können jedoch verbessert werden:

Genauer lässt sich zeigen, dass die durch Fettdruck in (7.27) gekennzeichneten Einsen vernachlässigbar sind.

Satz 7.13 (Zulehner 2011). *Es gelten die Voraussetzungen vom Satz von Brezzi. Dann besitzt die Operatorgleichung*

$$L \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} F \\ G \end{pmatrix},$$

mit

$$L : X \times Y \longrightarrow X^\star \times Y^\star$$

$$L := \begin{pmatrix} A & B^\star \\ B & 0 \end{pmatrix},$$

eine eindeutige Lösung $(u, p) \in X \times Y$*, wobei gilt:*

$$||u||_X \leq \frac{1}{\alpha_1}||F||_{X^\star} + \frac{\alpha_2}{\beta_1\alpha_1}||G||_{Y^\star},$$
$$||p||_Y \leq \frac{\alpha_2}{\beta_1\alpha_1}||F||_{X^\star} + \frac{\alpha_2^2}{\beta_1^2\alpha_1}||G||_{Y^\star}.$$

Als wichtiges Hilfsmittel für den Beweis erweist sich die folgende Aussage über die Charakterisierung von einem Isomorphismus (Beweis siehe z.B. [1], Seite 33):

Satz 7.14 (Babuška und Aziz). *Seien* X *und* Y *Hilberträume und*

$$\mathcal{T} : X \to Y^\star,$$

ein linearer und stetiger Operator. Dann ist $\mathcal{T}$ *ein Isomorphismus (das heißt* $\mathcal{T}$ *ist ein linearer und stetiger Operator welcher zusätzlich eine stetige Inverse besitzt), genau dann wenn*

1. $\exists \mu_2 > 0$ *mit*

$$||\mathcal{T}x||_{Y^\star} \leq \mu_2||x||_X \quad \forall x \in X,$$
$$\Leftrightarrow$$
$$|\langle \mathcal{T}x, y\rangle_{Y^\star \times Y}| \leq \mu_2||x||_X||y||_Y \quad \forall x \in X, \forall y \in Y,$$

2. $\exists \mu_1 > 0$ *mit*

$$||\mathcal{T}x||_{Y^\star} \geq \mu_1||x||_X \quad \forall x \in X,$$
$$\Leftrightarrow$$
$$\inf_{0 \neq x \in X} \sup_{0 \neq y \in Y} \frac{\langle \mathcal{T}x, y\rangle_{Y^\star \times Y}}{||x||_X||y||_Y} \geq \mu_1,$$

3.

$$\begin{gathered}\operatorname{kern}\mathcal{T}^\star = \{0\},\\ \Leftrightarrow\\ \forall y \in Y \; mit \; y \neq 0 \, \exists x \in X : \langle \mathcal{T}x, y\rangle \neq 0.\end{gathered}$$

Nun zum Beweis von Satz 7.13:

Beweis. (siehe [4]) Als erstes führen wir eine Aufspaltung des Raumes X in

$$X = X_0 \oplus X_1,$$

mit $X_0 := \operatorname{kern} B$ und $X_1 := X_0^\perp$ durch, sodass sich jedes $u \in X$ eindeutig schreiben lässt als

$$u = u_0 + u_1,$$

mit $u_0 \in X_0$ und $u_1 \in X_1$.

Definieren wir weiters für $i \in \{0, 1\}$ die Operatoren

$$\begin{aligned}A_{ij} &: X_j \longrightarrow X_i^\star,\\ B_i &: X_i \longrightarrow Y^\star,\\ F_i &: X_i \longrightarrow \mathbb{R},\end{aligned}$$

mit

$$\begin{aligned}\langle A_{ij}v, u\rangle_{X_i^\star \times X_i} &:= \langle Av, u\rangle_{X^\star \times X} \quad \forall v \in X_j, \forall u \in X_i,\\ \langle B_i v, p\rangle_{Y^\star \times Y} &:= \langle Bv, p\rangle_{Y^\star \times Y} \quad \forall v \in X_i, \forall p \in Y,\\ \langle F_i, v\rangle_{X_i^\star} &:= \langle F, v\rangle_{X^\star} \quad \forall v \in X_j,\end{aligned}$$

dann lassen sich die Operatoren A, B und F darstellen als:

$$A = \begin{pmatrix} A_{00} & A_{01} \\ A_{10} & A_{11}\end{pmatrix}, \quad F = \begin{pmatrix} F_0 \\ F_1\end{pmatrix},$$

und

$$B = (0, B_1),$$

wobei wegen

$$\langle B_0 v, p\rangle_{Y^\star \times Y} := \langle Bv, p\rangle_{Y^\star \times Y} = 0 \quad \forall v \in X_0 \; \forall p \in Y,$$

$B_0 = 0$ gilt.

Wegen

- $$\begin{aligned}\alpha_2 \geq ||A||_{L(X,X^\star)} &:= \sup_{0\neq v\in X}\sup_{0\neq w\in X}\frac{\langle Av,w\rangle_{X^\star\times X}}{||v||_X||w||_X}\\ &\geq \sup_{0\neq v\in X_j}\sup_{0\neq w\in X_i}\frac{\langle Av,w\rangle_{X^\star\times X}}{||v||_X||w||_X}\\ &= \sup_{0\neq v\in X_j}\sup_{0\neq w\in X_i}\frac{\langle A_{ij}v,w\rangle_{X_i^\star\times X_i}}{||v||_X||w||_X}\\ &=||A_{ij}||_{L(X_j,X_i^\star)}\\ &=||A_{ij}^\star||_{L(X_i^\star,X_j^\star)} \text{ für } i,j\in\{0,1\},\end{aligned}$$

- $$\begin{aligned}\inf_{0\neq u\in X_0}||A_{00}u||_X &= \inf_{0\neq v\in X_0}\sup_{0\neq w\in X_0}\frac{\langle A_{0,0}v,w\rangle_{X_0^\star\times X_0}}{||v||_X||w||_X}\\ &= \inf_{0\neq v\in X_0}\sup_{0\neq w\in X_0}\frac{\langle Av,w\rangle_{X_0^\star\times X_0}}{||v||_X||w||_X}\\ &= \inf_{0\neq v\in X_0}\sup_{0\neq w\in X_0}\frac{\langle Aw,w\rangle_{X^\star\times X}}{||v||_X||w||_X}\\ &\geq \alpha_1 > 0,\end{aligned}$$

- $$\inf_{0\neq v\in X_0}\sup_{0\neq w\in X_0}\frac{\langle A^\star w,v\rangle_{X^\star\times X}}{||v||_X||w||_X} = \inf_{0\neq v\in X_0}\sup_{0\neq w\in X_0}\frac{\langle A_{00}^\star w,v\rangle_{X_0^\star\times X_0}}{||v||_X||w||_X}\geq\alpha_1,$$

sind für A_{00} alle Voraussetzungen von Satz 7.14 erfüllt. Der Operator A_{00} bildet also einen Isomorphismus, wobei gilt:

$$||A_{00}||_{L(X_0\times X_0^\star)}\leq\alpha_2 \quad\wedge\quad ||A_{00}^{-1}||_{L(X_0^\star\times X_0)}\leq\frac{1}{\alpha_1}.$$

Aus

- $$\begin{aligned}\beta_2 \geq ||B||_{L(X,Y^\star)} &:= \sup_{0\neq v\in X}\sup_{0\neq p\in Y}\frac{\langle Bv,p\rangle_{Y^\star\times Y}}{||u||_X||v||_Y}\\ &\geq \sup_{0\neq u\in X_1}\sup_{0\neq v\in Y}\frac{\langle Bv,p\rangle_{Y^\star\times Y}}{||v||_X||p||_Y}\\ &= \sup_{0\neq v\in X_1}\sup_{0\neq p\in Y}\frac{\langle v,B_1^\star p\rangle_{X_1\times X_1^\star}}{||v||_X||p||_Y}\\ &=||B_1^\star||_{L(Y,X_1^\star)},\end{aligned}$$

- $$\begin{aligned}\inf_{0\neq p\in Y}||B_1^\star p||_X &= \inf_{0\neq p\in Y}\sup_{0\neq u\in X_1}\frac{\langle B_1^\star p,u\rangle_{X_1^\star\times X_1}}{||u||_X||p||_Y}\\ &= \inf_{0\neq p\in Y}\sup_{0\neq u\in X_1}\frac{b(u,p)}{||u||_X||p||_Y}\\ &= \inf_{0\neq p\in Y}\sup_{0\neq u\in X}\frac{b(u,p)}{||u||_X||p||_Y}\\ &\geq \beta_1 > 0,\end{aligned}$$

- $$\operatorname{kern} B_1 = \{0\},$$

folgt wiederum mit Satz 7.14 die Isomorphie für $B_1^\star$. Dabei gilt:

$$||B_1^\star||_{L(Y,X_1^\star)} = ||B_1||_{L(Y,X_1^\star)} \leq \beta_2 \quad \wedge \quad ||(B_1^\star)^{-1}||_{L(X_1^\star,Y)} = ||(B_1)^{-1}||_{L(X_1^\star,Y)} \leq \frac{1}{\beta_1}.$$

Wir erhalten für die Operatorgleichung die folgende Form:

$$L\begin{pmatrix}v\\p\end{pmatrix} = \begin{pmatrix}F\\G\end{pmatrix},$$
$$\Leftrightarrow$$
$$\begin{pmatrix}A_{0,0} & A_{0,1} & 0\\ A_{1,0} & A_{1,1} & B_1^\star\\ 0 & B_1 & 0\end{pmatrix}\begin{pmatrix}v_0\\v_1\\p\end{pmatrix} = \begin{pmatrix}F_0\\F_1\\G\end{pmatrix}.$$

Aus der Invertierbarkeit der Operatoren $A_{0,0}$, B_1 und $B_1^\star$ folgt nun weiters die Invertierbarkeit von L:

$$L^{-1} = \begin{pmatrix}A_{00}^{-1} & 0 & -A_{00}^{-1}A_{01}B_1^{-1}\\ 0 & 0 & B_1^{-1}\\ -(B_1^\star)^{-1}A_{10}A_{00}^{-1} & (B_1^\star)^{-1} & (B_1^\star)^{-1}(A_{10}A_{00}^{-1}A_{01}-A_{11})B_1^{-1}\end{pmatrix}, \tag{7.28}$$

Bezüglich der Idee den Operator L in Form 7.28 zu schreiben und der Berechnung seine Inverse L^{-1} zu berechnen, siehe [6]. Die Lösung lässt sich jetzt darstellen, als

$$\begin{pmatrix}v_0\\v_1\\p\end{pmatrix} = L^{-1}\begin{pmatrix}F_0\\F_1\\G\end{pmatrix},$$
$$\Leftrightarrow$$
$$\begin{pmatrix}v_0\\v_1\end{pmatrix} = \begin{pmatrix}A_{00}^{-1} & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}F_0\\F_1\end{pmatrix} + \begin{pmatrix}-A_{00}^{-1}A_{01}\\ I\end{pmatrix}B_1^{-1}G,$$
$$\wedge$$
$$p = (B_1^\star)^{-1}\begin{pmatrix}-A_{10}A_{00}^{-1} & I\end{pmatrix}\begin{pmatrix}F_0\\F_1\end{pmatrix} + (B_1^\star)^{-1}(A_{10}A_{00}^{-1}A_{01}-A_{11})B_1^{-1}G,$$

wodurch sich folgende Abschätzungen ergeben:

$$\begin{aligned}
||u||_X &\leq \left|\left|\begin{pmatrix} A_{00}^{-1} & 0 \\ 0 & 0 \end{pmatrix}\right|\right|_{L(X^\star,X)} ||F||_{X^\star} \\
&\quad + \left|\left|\begin{pmatrix} -A_{00}^{-1}A_{01} \\ I \end{pmatrix}\right|\right|_{L(X_1,X)} ||B_1^{-1}||_{L(Y^\star,X_1)}||G||_{Y^\star} \\
&\leq \frac{1}{\alpha_1}||F||_{X^\star} + \frac{\alpha_2}{\alpha_1\beta_1}||G||_{Y^\star},
\end{aligned}$$

$$\begin{aligned}
||p||_Y &\leq ||(B_1^\star)^{-1}||_{L(X_1^\star,Y)}|| \begin{pmatrix} -A_{10}A_{00}^{-1} & I \end{pmatrix} ||_{L(X^\star,X_0^\star)}||F||_{X^\star} \\
&\quad + ||(B_1^\star)^{-1}||_{L(X_1^\star,Y)}||A_{10}A_{00}^{-1}A_{01} - A_{11}||_{L(X_1,X_1^\star)}||B_1^{-1}||_{L(Y^\star,X_1)}||G||_{Y^\star}, \\
&\leq ||(B_1^\star)^{-1}||_{L(X_1^\star,Y)}|| \begin{pmatrix} -A_{10}A_{00}^{-1} & I \end{pmatrix} ||_{L(X^\star,X_0^\star)}||F||_{X^\star} \\
&\quad + ||(B_1^\star)^{-1}||_{L(X_1^\star,Y)}|| \begin{pmatrix} A_{10} & -A_{11} \end{pmatrix} ||_{L(X,X_1^\star)} \\
&\quad \left|\left|\begin{pmatrix} A_{00}^{-1}A_{01} \\ I \end{pmatrix}\right|\right|_{L(X_1,X)} ||B_1^{-1}||_{L(Y^\star,X_1)}||G||_{Y^\star} \\
&\leq \frac{\alpha_2}{\alpha_1\beta_1}||F||_{X^\star} + \frac{\alpha_2^2}{\alpha_1\beta_1^2}||G||_{Y^\star}.
\end{aligned}$$

□

Mit Hilfe des letzten Satzes lässt sich nun die folgende Aussage bezüglich der Existenz und Größe der Konstanten C_1 und C_2 aus (7.26) für lineare Operatoren der Gestalt (7.24) nachweisen:

Korollar 7.15. *Es gelten die Voraussetzungen vom Satz von Brezzi. Dann ist*

$$L : X \times Y \longrightarrow X^\star \times Y^\star,$$

mit

$$L := \begin{pmatrix} A & B^\star \\ B & 0 \end{pmatrix}.$$

ein Ismorphismus, wobei gilt:

$$C_1||w||_{X\times Y} \leq ||Lw||_{X^\star\times Y^\star} \leq C_2||w||_{X\times Y} \quad \forall w \in X \times Y, \tag{7.29}$$

mit

$$\begin{aligned}
C_1 &:= \frac{\alpha_1}{1 + \left(\frac{\alpha_2}{\beta_1}\right)^2}, \\
C_2 &:= \frac{\alpha_2 + \sqrt{\alpha_2^2 + 4\beta_2^2}}{2}.
\end{aligned}$$

und

$$||(v,q)^T||_{X\times Y} := \sqrt{||v||_X^2 + ||q||_Y^2} \quad \forall (v,q)^T \in X \times Y.$$

Beweis. Nach dem Satz von Brezzi, besitzt für $(F,G)^T \in X^\star \times Y^\star$ beliebig aber fix, die Operatorgleichung

$$L\begin{pmatrix} v \\ q \end{pmatrix} = \begin{pmatrix} F \\ G \end{pmatrix},$$

eindeutige Lösung ($v_\star, q_\star \in X \times Y$), womit die Bijektivität von L gezeigt wäre.

Für den weiteren Beweisverlauf wählen wir $(v,q)^T \in X \times Y$ beliebig aber fix. Dabei gilt:

$$\begin{aligned}\left|\left|L\begin{pmatrix} v \\ q \end{pmatrix}\right|\right|_{X^\star\times Y^\star} :=& \sup_{0\neq(u,p)^T\in X\times Y} \frac{\left\langle L\begin{pmatrix} v \\ q \end{pmatrix}, \begin{pmatrix} u \\ p \end{pmatrix}\right\rangle_{(X\times Y)\times(X^\star\times Y^\star)}}{||(u,p)^T||_{X\times Y}} \\ =& \sqrt{||Av + B^\star q||_{X^\star}^2 + ||Bv||_{Y^\star}^2},\end{aligned}$$

Ermittlung der Konstanten C_1 und C_2 aus (7.29):

- C_2:

  Es gilt:

$$\begin{aligned}||L\begin{pmatrix} v \\ q \end{pmatrix}||_{X^\star\times Y^\star} &= \sqrt{||Av + B^\star q||_{X^\star}^2 + ||Bv||_{Y^\star}^2} \\ &\leq \sqrt{\left(||Av||_{X^\star} + ||B^\star q||_{X^\star}\right)^2 + ||Bv||_{Y^\star}^2} \\ &\leq \sqrt{\left(\alpha_2||v||_X + \beta_2||q||_Y\right)^2 + \beta_2||v||_X^2} \\ &= \left|\left|\begin{pmatrix} \alpha_2 & \beta_2 \\ \beta_2 & 0 \end{pmatrix}\begin{pmatrix} ||v||_X \\ ||q||_Y \end{pmatrix}\right|\right|_{\ell_2} \\ &\leq \underbrace{\left|\left|\begin{pmatrix} \alpha_2 & \beta_2 \\ \beta_2 & 0 \end{pmatrix}\right|\right|_{\ell_2}}_{K_2:=} \sqrt{||v||_X^2 + ||q||_Y^2} \\ &\leq |\lambda|_{\max}(K_2)||(v,q)^T||_{X\times Y},\end{aligned}$$

  wobei sich die Eigenwerte der Matrix K_2 als Nullstellen des charakteristischen Polynoms

$$c_{(K_2)}(\lambda) := \det(K_2 - \lambda I_2) = -(\alpha_2 - \lambda)\lambda - \beta_2^2 = \lambda^2 - \alpha_2\lambda + \beta_2^2,$$

berechnen lassen:

$$\begin{aligned} c_{(K_2)}(\lambda) &= 0 \\ &\Leftrightarrow \\ \lambda^2 - \alpha_2\lambda + \beta_2^2 &= 0 \\ &\Leftrightarrow \\ \lambda &= \frac{\alpha_2}{2} \pm \frac{\sqrt{\alpha_2^2 + 4\beta_2^2}}{2}. \end{aligned}$$

Aus

$$\begin{aligned} |\lambda|_{\max}(K_2) &= \max(\{|\frac{\alpha_2}{2} + \frac{\sqrt{\alpha_2^2 + 4\beta_2^2}}{2}|, |\frac{\alpha_2}{2} - \frac{\sqrt{\alpha_2^2 + 4\beta_2^2}}{2}|\}) \\ &= \frac{\alpha_2}{2} + \frac{\sqrt{\alpha_2^2 + 4\beta_2^2}}{2}, \end{aligned}$$

folgt dann

$$\begin{aligned} ||L \begin{pmatrix} v \\ q \end{pmatrix} ||_{X^\star \times Y^\star} \leq & C_2 ||(v,q)^T||_{X \times Y}, \\ \text{mit } \; C_2 := & \frac{\alpha_2 + \sqrt{\alpha_2^2 + 4\beta_2^2}}{2}. \end{aligned}$$

- C_1:

  Aus den Abschätzungen von Satz 7.13

$$\begin{aligned} ||v||_X &\leq \frac{1}{\alpha_1}||Av + B^\star q||_{X^\star} + \frac{\alpha_2}{\beta_1\alpha_1}||Bv||_{Y^\star}, \\ ||q||_Y &\leq \frac{\alpha_2}{\beta_1\alpha_1}||Av + B^\star q||_{X^\star} + \frac{\alpha_2^2}{\beta_1^2\alpha_1}||Bv||_{Y^\star}. \end{aligned}$$

folgt

$$\begin{aligned}
||(v,q)^T||_{X\times Y} =& \sqrt{||v||_X^2 + ||q||_Y^2} \\
\leq& \left(\left(\frac{1}{\alpha_1} ||Av + B^\star q||_{X^\star} + \frac{\alpha_2}{\beta_1\alpha_1} ||Bv||_{Y^\star} \right)^2 \right. \\
&\left. + \left(\frac{\alpha_2}{\beta 1 \alpha_1} ||Av + B^\star q||_{X^\star} + \frac{\alpha_2^2}{\beta_1^2\alpha_1} ||Bv||_{Y^\star} \right)^2 \right)^{\frac{1}{2}} \\
=& \left|\left| \begin{pmatrix} \frac{1}{\alpha_1} & \frac{\alpha_2}{\alpha_1\beta_1} \\ \frac{\alpha_2}{\alpha_1\beta_1} & \frac{\alpha^2}{\alpha_1\beta_1^2} \end{pmatrix} \begin{pmatrix} ||Av + B^\star q||_{X^\star} \\ ||Bq||_{Y^\star} \end{pmatrix} \right|\right|_{\ell_2} \\
\leq& \frac{1}{\alpha_1} \left|\left| \underbrace{\begin{pmatrix} 1 & \frac{\alpha_2}{\beta_1} \\ \frac{\alpha_2}{\beta_1} & \frac{\alpha^2}{\beta_1^2} \end{pmatrix}}_{K_1:=} \right|\right|_{\ell_2} \sqrt{||Av + B^\star q||_{X^\star}^2 + ||Bv||_{Y^\star}^2} \\
\leq& \frac{|\lambda|_{\max}(K_1)}{\alpha_1} \left|\left| L \begin{pmatrix} v \\ q \end{pmatrix} \right|\right|_{X^\star\times Y^\star}.
\end{aligned}$$

Setzen wir $\gamma = \frac{\alpha_2}{\beta_1}$, so besitzt die Matrix

$$K_1 = \begin{pmatrix} 1 & \frac{\alpha_2}{\beta_1} \\ \frac{\alpha_2}{\beta_1} & \frac{\alpha^2}{\beta_1^2} \end{pmatrix} = \begin{pmatrix} 1 & \gamma \\ \gamma & \gamma^2 \end{pmatrix},$$

das charakteristische Polynom

$$\begin{aligned}
c_{(K_1)}(\lambda) :=& \det(K_1 - \lambda I_2) = (1-\lambda)(\gamma^2 - \lambda) - \gamma^2 \\
=& \lambda^2 - \lambda(\gamma^2 + 1).
\end{aligned}$$

Die Eigenwerte der Matrix K_1 ergeben sich wiederum durch Berechnung der Nullstellen des charakteristischen Polynoms:

$$\begin{aligned}
c_{(K_1)}(\lambda) &= 0, \\
&\Leftrightarrow \\
\lambda^2 - \lambda(\gamma^2+1) &= 0, \\
&\Leftrightarrow \\
\lambda^2 &= \lambda(\gamma^2+1), \\
&\Leftrightarrow \\
\lambda = 0 \quad \vee \quad \lambda &= 1 + \gamma^2 = 1 + \left(\frac{\alpha_2}{\beta_1}\right)^2.
\end{aligned}$$

Also

$$\begin{aligned}
|\lambda|_{\max}(K) &= \max(\{|1 + (\tfrac{\alpha_2}{\beta_1})^2|, 0\}) \\
&= 1 + \left(\frac{\alpha_2}{\beta_1}\right)^2
\end{aligned}$$

und somit:

$$\begin{aligned} ||(v,q)^T||_{X\times Y} &\leq \frac{1+\left(\frac{\alpha_2}{\beta_1}\right)^2}{\alpha_1}\left|\left|L\begin{pmatrix}v\\q\end{pmatrix}\right|\right|_{X^\star\times Y^\star}, \\ &\Leftrightarrow \\ ||(v,q)^T||_{X\times Y}\ C_1 &\leq \left|\left|L\begin{pmatrix}v\\q\end{pmatrix}\right|\right|_{X^\star\times Y^\star}, \\ \text{mit}\quad C_1 &:= \frac{\alpha_1}{1+\left(\frac{\alpha_2}{\beta_1}\right)^2}. \end{aligned}$$

□

7.5.3 Nachweis der Voraussetzungen vom Satz von Brezzi

Es bleibt noch zu zeigen, dass für den in (7.24) definierten linearen Operator L, alle Voraussetzungen des *Satz von Brezzi*, mit gegenüber den Parametern α, ω_i und der Schrittweite h robusten Konstanten

$$\alpha_1, \beta_1, \alpha_2, \beta_2,$$

erfüllt sind. In diesem Fall würde nämlich sofort aus Korrollar 7.15, die Robustheit der Konstanten C_1 und C_2 aus 7.26 für L folgen.

Überprüfung der Voraussetzungen vom *Satz von Brezzi*:

- Die Matrizen Q und BQB^T sind symmetrisch und positiv definit, sodass

  $$\begin{aligned} (v,w)_X &:= (Qu,v)_{\ell_2} \quad \forall v,w \in X, \\ (p,q)_Y &:= (BQB^Tp,q)_{\ell_2} \quad \forall p,q \in Y, \end{aligned}$$

  Biliniarformen auf X und Y bilden. Die Abgeschlossenheit folgt aus der endlichen Dimension von X und Y.

- A und B bilden lineare Operatoren:

  In unserem Fall gilt:

  $$A : X \longrightarrow X^\star,$$

  mit

  $$\langle Av, w\rangle_{X^\star\times X} :=(Au,v)_{\ell_2} \quad \forall v \in X, \forall w \in X,$$

  und

  $$B : X \longrightarrow Y^\star,$$

mit

$$\langle Bv, p\rangle_{Y^\star \times Y} := (Bv, p)_{\ell_2} \quad \forall v \in X, \forall p \in Y,$$

Die Linearität von A und B folgt aus der Linearitätseigenschaft des inneren Produktes.

1. A ist beschränkt:

$$\begin{aligned}
|\langle Av, w\rangle_{X^\star \times X}| =& |(Av, w)_{\ell_2}| \\
=& \Bigg| \begin{pmatrix} v_1 & v_2 & -v_3 & -v_4 \end{pmatrix} \begin{pmatrix} C_h & 0 & 0 & 0 \\ 0 & C_h & 0 & 0 \\ 0 & 0 & C_h & 0 \\ 0 & 0 & 0 & C_h \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{pmatrix} \\
& + \sqrt{\alpha}\omega \begin{pmatrix} v_4 & v_3 & v_2 & v_1 \end{pmatrix} \begin{pmatrix} C_h & 0 & 0 & 0 \\ 0 & C_h & 0 & 0 \\ 0 & 0 & C_h & 0 \\ 0 & 0 & 0 & C_h \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{pmatrix} \\
& + \sqrt{\alpha} \begin{pmatrix} v_3 & -v_4 & v_1 & -v_2 \end{pmatrix} \begin{pmatrix} A_h & 0 & 0 & 0 \\ 0 & A_h & 0 & 0 \\ 0 & 0 & A_h & 0 \\ 0 & 0 & 0 & A_h \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{pmatrix} \Bigg| \\
\overset{(i)}{\leq} & (1 + \sqrt{\alpha}\omega) \sum_{i=1}^{4} ||v_i||_{C_h} ||w_i||_{C_h} + \sqrt{\alpha} \sum_{i=1}^{4} ||v_i||_{A_h} ||w_i||_{A_h} \\
\overset{(ii)}{\leq} & \left((1 + \sqrt{\alpha}\omega) \sum_{i=1}^{4} ||v_i||^2_{C_h} + \sqrt{\alpha} \sum_{i=1}^{4} ||v_i||^2_{A_h} \right)^{\frac{1}{2}} \\
& \left((1 + \sqrt{\alpha}\omega) \sum_{i=1}^{4} ||w_i||^2_{C_h} + \sqrt{\alpha} \sum_{i=1}^{4} ||w_i||^2_{A_h} \right)^{\frac{1}{2}} \\
=& \left(\sum_{i=1}^{4} (I_h v_i, v_i)_{\ell_2} \right)^{\frac{1}{2}} \left(\sum_{i=1}^{4} (I_h w_i, w_i)_{\ell_2} \right)^{\frac{1}{2}} \\
=& \sqrt{(Qv, v)_{\ell^2}} \sqrt{(Qw, w)_{\ell^2}} = ||v||_X ||w||_X \quad \forall v, w \in X,
\end{aligned}$$

$$\Rightarrow \alpha_2 = 1.$$

In (i) und (ii) wurde die Cauchy-Schwarz-Ungleichung, bezüglich der inneren Produkte $(\cdot,\cdot)_{A_h}$, $(\cdot,\cdot)_{C_h}$ (in (i)) und $(\cdot,\cdot)_{\ell^2}$ (in (ii)) verwendet.

2. B ist ist beschränkt:

$$\begin{aligned}
|\langle Bv, p\rangle_{Y^\star \times Y}| =& |(Bv, p)_{\ell_2}| \\
=& |(v, B^T p)_{\ell_2}| \\
\overset{(i)}{=}& |(Q^{-1}Qv, B^T p)_{\ell_2}| \\
\leq& \sqrt{(Q^{-1}Qv, Qv)_{\ell_2}}\sqrt{(Q^{-1}B^T p, B^T p)_{\ell_2}} \\
\leq& \sqrt{(v, Qv)_{\ell_2}}\sqrt{(BQ^{-1}B^T p, p)_{\ell_2}} \\
=& ||v||_X ||p||_Y \quad \forall v \in X, \forall p \in Y.
\end{aligned}$$

$$\Rightarrow \beta_2 = 1.$$

In (i) wurde die Idendität $v = Q^{-1}Qv$ für $v \in X = \mathbb{R}^{4dn_h}$ benützt.

3. A_0 und $A_0^\star$ erfüllen eine inf-sup-Bedingung:

Aus der speziellen Block-Form der Matrix

$$B = \begin{pmatrix} 0 & 0 & -B_h & 0 \\ 0 & 0 & 0 & -B_h \\ -B_h & 0 & 0 & 0 \\ 0 & -B_h & 0 & 0 \end{pmatrix}$$

erhält man für

$$\begin{aligned}
X_0 := \text{kern}(B) =& \{v \in X : \langle Bv, p\rangle_{Y^\star \times y} = 0 \ \ \forall p \in Y\} \\
=& \{v \in X : (Bv, p)_{\ell^2} = 0\} \\
=& \{v \in X : Bv = 0\},
\end{aligned}$$

zunächst die folgende Charakterisierung

$$\begin{aligned}
& w = (w_1, w_2, w_3, w_4)^T \in X_0 \\
& \Leftrightarrow \\
& w_i \in \text{kern}(B_h) \ \text{ für } i = 1, \ldots, 4.
\end{aligned} \tag{7.30}$$

Wir wählen

$$w = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{pmatrix} \in X_0.$$

Als nächstes wollen wir Vektoren

$$v_1^{(w)}, v_2^{(w)}, v_3^{(w)} \in X_0,$$

konstruieren, welche die Eigenschaften

$$\begin{aligned}
&(a)\ \ ||v_i^{(w)}||_X^2 = ||w||_X^2 \ \text{ für } i = 1,2,3,\\
&(b)\ \ (v_i^{(w)}, v_j^{(w)})_X = 0 \ \text{ für } \ i \neq j = 1,2,3,\\
&(c)\ \ \langle Av_1^{(w)}, w\rangle_{X^\star \times X} = \sum_{i=1}^{4} (C_h w_i, w_i)_{\ell_2},\\
&(d)\ \ \langle Av_2^{(w)}, w\rangle_{X^\star \times X} = \sqrt{\alpha}\omega \sum_{i=1}^{4} (C_h w_i, w_i)_{\ell_2}\\
&(e)\ \ \langle Av_3^{(w)}, w\rangle_{X^\star \times X} = \sqrt{\alpha} \sum_{i=1}^{4} (A_h w_i, w_i)_{\ell_2}.
\end{aligned} \tag{7.31}$$

erfüllen.

Aus der Darstellung

$$\begin{aligned}
\langle Av, w\rangle_{X^\star \times X} =& (Av, w)_{\ell^2}\\
=& \begin{pmatrix} v_1 & v_2 & -v_3 & -v_4 \end{pmatrix} \begin{pmatrix} C_h & 0 & 0 & 0\\ 0 & C_h & 0 & 0\\ 0 & 0 & -C_h & 0\\ 0 & 0 & 0 & -C_h \end{pmatrix} \begin{pmatrix} w_1\\ w_2\\ w_3\\ w_4 \end{pmatrix}\\
&+ \sqrt{\alpha}\omega \begin{pmatrix} v_4 & v_3 & v_2 & v_1 \end{pmatrix} \begin{pmatrix} C_h & 0 & 0 & 0\\ 0 & C_h & 0 & 0\\ 0 & 0 & C_h & 0\\ 0 & 0 & 0 & C_h \end{pmatrix} \begin{pmatrix} w_1\\ w_2\\ w_3\\ w_4 \end{pmatrix}\\
&+ \sqrt{\alpha} \begin{pmatrix} v_3 & -v_4 & v_1 & -v_2 \end{pmatrix} \begin{pmatrix} A_h & 0 & 0 & 0\\ 0 & A_h & 0 & 0\\ 0 & 0 & A_h & 0\\ 0 & 0 & 0 & A_h \end{pmatrix} \begin{pmatrix} w_1\\ w_2\\ w_3\\ w_4 \end{pmatrix},
\end{aligned}$$

erhalten wir dann:

$$\begin{aligned}
v_1^{(w)} = \begin{pmatrix} w_1 & w_2 & -w_3 & -w_4 \end{pmatrix}^T &\Rightarrow v_1^{(w)} \in M \wedge \text{erfüllt } (c),\\
v_2^{(w)} = \begin{pmatrix} w_4 & w_3 & w_2 & w_1 \end{pmatrix}^T &\Rightarrow v_2^{(w)} \in M \wedge \text{erfüllt } (d),\\
v_3^{(w)} = \begin{pmatrix} -w_3 & w_4 & -w_1 & w_2 \end{pmatrix}^T &\Rightarrow v_3^{(w)} \in M \wedge \text{erfüllt } (e).
\end{aligned}$$

Weiters sind die Vektoren

$$v_1^{(w)} = \begin{pmatrix} w_1\\ w_2\\ -w_3\\ -w_4 \end{pmatrix}, \quad v_2^{(w)} = \begin{pmatrix} w_4\\ w_3\\ w_2\\ w_1 \end{pmatrix}, \quad v_3^{(w)} = \begin{pmatrix} -w_3\\ w_4\\ -w_1\\ w_2 \end{pmatrix} \in X_0, \tag{7.32}$$

bezüglich $(\cdot,\cdot)_X$ orthogonal, sodass also alle Eigenschaften $(a)-(e)$ erfüllt sind.

Für

$$v^{(w)} := v_1^{(w)} + v_2^{(w)} + v_3^{(w)}, \tag{7.33}$$

mit

$$\begin{aligned}\langle Av^{(w)}, w\rangle_{X^\star\times X} :=& (Av^{(w)}, w)_{\ell_2}\\ =& (Av_1^{(w)}, w)_{\ell_2} + (Av_2^{(w)}, w)_{\ell_2} + (Av_3^{(w)}, w)_{\ell_2}\\ =& \sum_{i=1}^{4}((1+\sqrt{\alpha}\omega)C_h + \sqrt{\alpha}A_h w_i, w_i)_{\ell_2}\\ =& \sum_{i=1}^{4}(I_h w_i, w_i)_{\ell_2} = (Qw, w)_{\ell_2} = ||w||_X^2,\end{aligned} \tag{7.34}$$

$$\begin{aligned}||v^{(w)}||_X^2 =& ||v_1^{(w)}||_X^2 + ||v_2^{(w)}||_X^2 + ||v_3^{(w)}||_X^2\\ =& 3||w||_X^2,\\ \Leftrightarrow&\\ ||v^{(w)}||_X =& \sqrt{3}||w||_X,\end{aligned}$$

lässt sich nun zeigen:

$$\begin{aligned}\inf_{0\neq w\in X_0}\sup_{0\neq v\in X_0}\frac{\langle Aw, v\rangle_{X^\star\times X}}{||w||_X||v||_X} =& \inf_{0\neq w\in X_0}\sup_{0\neq v\in X_0}\frac{(Aw, v)_{\ell_2}}{||w||_X||v||_X}\\ \overset{(i)}{\geq}& \inf_{0\neq w\in X_0}\frac{(Aw, v^{(w)})_{\ell_2}}{||w||_X||v^{(w)}||_X}\\ \overset{(7.34)}{=}& \inf_{0\neq v\in X_0}\frac{||w||_X^2}{\sqrt{3}||w||_X||w||_X}\\ =& \frac{1}{\sqrt{3}},\end{aligned}$$

$$\Rightarrow \alpha_1 = \frac{1}{\sqrt{3}}.$$

In (i) wurde der Ausdruck $\sup_{0\neq v\in X_0}$ durch das konkrete Element $v(w)\in X_0$ aus (7.33) ersetzt. Weiters erhalten wir aus der Tatsache $A = A^\star$, dass gilt:

$$\inf_{0\neq v\in X_0}\sup_{0\neq w\in X_0}\frac{\langle A^\star w, v\rangle_{X^\star\times X}}{||w||_X||v||_X} = \inf_{0\neq w\in X_0}\sup_{0\neq v\in X_0}\frac{\langle Aw, v\rangle_{X^\star\times X}}{||w||_X||v||_X} \geq \alpha_1.$$

4. $B^\star$ erfüllt eine inf-sup-Bedingung:

$$\begin{aligned}\inf_{0\neq q\in Y}\sup_{0\neq v\in X}\frac{\langle Bv,q\rangle_{Y^\star\times Y}}{||v||_X||q||_Y} &:= \inf_{0\neq q\in Y}\sup_{0\neq v\in X}\frac{(Bv,q)_{\ell_2}}{\sqrt{(Qv,v)_{\ell_2}}\sqrt{(BQ^{-1}B^Tq,q)_{\ell_2}}}\\ &\overset{(i)}{\geq}\inf_{0\neq q\in Y}\frac{(BQ^{-1}B^Tq,q)_{\ell_2}}{\sqrt{(QQ^{-1}B^Tq,Q^{-1}B^Tq)_{\ell_2}}||q||_Y}\\ &=\inf_{0\neq q\in Y}\frac{(BQ^{-1}B^Tq,q)_{\ell_2}}{\sqrt{(q,BQ^{-1}B^Tq)_{\ell_2}}||q||_Y}\\ &=\inf_{0\neq q\in Y}\frac{||q||_Y^2}{||q||_Y^2}\\ &=1\end{aligned}$$

$$\Rightarrow \beta_1 = 1.$$

In (i) wurde der Ausdruck $\sup_{0\neq v\in X}$ durch das konkrete Element $v(q) := Q^{-1}B^Tq \in X$ ersetzt.

7.6 Das Resultat

Da alle Voraussetzungen von Korollar 7.15 erfüllt sind, erhalten wir:

$$\begin{gathered}\frac{1}{2\sqrt{3}}||w||_{X\times Y} \leq ||Lw||_{X^\star\times Y^\star} \leq \frac{1+\sqrt{5}}{2}||w||_{X\times Y} \quad \forall w\in X\times Y,\\ \Leftrightarrow (7.25)\\ \frac{1}{2\sqrt{3}}||w||_{\mathcal{P}} \leq ||\mathcal{P}^{-1}\mathcal{A}^\omega w||_{\mathcal{P}} \leq \frac{1+\sqrt{5}}{2}||w||_{\mathcal{P}}, \quad \forall w\in X\times Y,\\ \Downarrow\\ \kappa(\mathcal{P}^{-1}\mathcal{A}^\omega) := ||\mathcal{P}^{-1}\mathcal{A}^\omega||_{\mathcal{P}}||(\mathcal{P}^{-1}\mathcal{A}^\omega)^{-1}||_{\mathcal{P}} \leq C,\end{gathered}$$

$$\text{mt } C = \sqrt{3}(1+\sqrt{5}) = 5.605\ldots.$$

Fassen wir also alle Resultate zusammen, so haben wir gezeigt:

Satz 7.16. *Es gilt:*

Die Matrix $\mathcal{P}$ definiert in (7.23) bildet einen robusten, symmetrisch und positiv definiten Präkonditionier in Block-Diagonalform von $\mathcal{A}^i$, mit

$$\kappa(\mathcal{P}^{-1}\mathcal{A}^\omega) \leq \sqrt{3}(1+\sqrt{5}) = 5.605\ldots.$$

Kapitel 8

Nummerische Resultate

In diesem Kapitel wollen wir die Robustheit der Präkonditionierungsmatrix $\mathcal{P}$ an Hand von nummerischen Beispielen illustrieren.

Dazu ermitteln wir die Anzahl der benötigten Iterationen für das präkonditionierten MINRES-Verfahren zur Lösung des linearen Gleichungssystems

$$\mathcal{P}^{-1}\mathcal{A}^{\omega}x = \mathcal{P}^{-1}b, \tag{8.1}$$

für unterschiedliche Parameter α, ω, h.

Bei der Wahl der Parameter und der endlich dimensionalen Teilräume V_h und Q_h muss darauf geachtet werden, dass alle zu Beginn des letzten Kapitels getroffenen Annahmen

$$\alpha > 0, \omega > 0, h > 0,$$

und die Annahme 1 (siehe Seite 39), erfüllt sind.

8.1 Eingabedaten

Es gelte:

- Ortsgebiet:

$$\Omega = (0,1)^2,$$

- PMINRES-Methode Parameter:
  - Toleranz: tol $= 10^{-8}$,
  - T $= 1$,
  - Startvektor: $x_0 = (0, \ldots, 0)^T$,

- Wahl der rechten Seite b:

$$b = \frac{b_\star}{||b_\star||_{\mathcal{P}^{-1}}},$$

wobei die Komponenten des Vektors $b_\star$ gleichverteilt aus dem Intervall $(0, 1)$ gewählt werden.

8.2 Theoretische Konvergenz

Für das System (8.1), erhalten wir aus der Konvergenzaussage für das präkonditionierte MINRES-Verfahren (Satz 6.3):

$$||r_k||_{\mathcal{P}^{-1}} \leq \frac{2q^k}{1+q^{2k}}||r_0||_{\mathcal{P}^{-1}},$$

wobei

$$r_k = b - \mathcal{A}^\omega x_k \quad \text{and} \quad q = \frac{\kappa(\mathcal{P}^{-1}\mathcal{A}^\omega) - 1}{\kappa(\mathcal{P}^{-1}\mathcal{A}^\omega) + 1},$$

mit

- $\text{tol} = 10^{-8}$,
- $\kappa(\mathcal{P}^{-1}\mathcal{A}^i) < 5.605$,
-

$$\begin{aligned} ||r0||_{\mathcal{P}^{-1}} &= ||b - \mathcal{A}^\omega x_0||_{\mathcal{P}^{-1}} \\ &= ||b||_{\mathcal{P}^{-1}} = 1, \end{aligned}$$

für die theoretische Schranke der Anzahl der benötigten Iterationen:

$$||r_k||_{\mathcal{P}^{-1}} < 10^{-8} \Leftrightarrow \frac{2q^{k/2}}{1+q^k} < 10^{-8} \Leftrightarrow k \geq 106.$$

8.3 Wahl der finiten Räume

Auf einer gleichmäßigen Triangulation $\mathcal{T}_h$ in rechtwinklige Dreiecke mit Kathetenlänge $\tilde{h} = h/\sqrt{2}$ (siehe auch Abbildung 8.1),

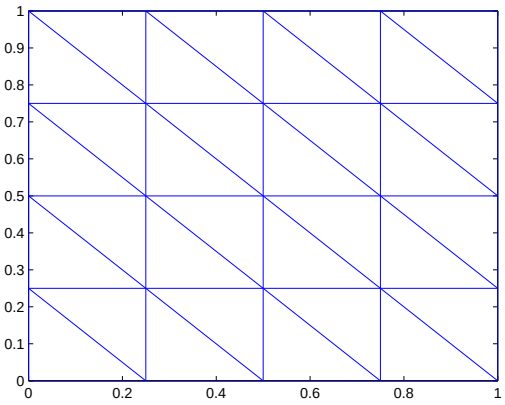

Abbildung 8.1: Gleichmäßige Triangulation mit $\tilde{h} = 0.25$

wählen wir das Taylor-Hood-Element:

$$V_h = \{v \in C_0(\overline{\Omega}) : v|_T \in P_2 \ \forall T \in \mathcal{T}_h\} \subset V = H_0^1(\Omega),$$
$$Q_h = \{q \in C(\overline{\Omega}) \cap L_0^2(\Omega) : q|_T \in P_1 \ \forall T \in \mathcal{T}_h\} \subset Q = L_0^2(\Omega),$$

Dieses Element ist stabil und erfüllt deshalb auch die Annahme 1 (siehe auch Bem. 4.11).

8.4 Ergebnisse

Im Folgenden bezeichnet $n(h)$ die Dimension der Systemmatrix $\mathcal{A}^\omega$, wobei diese von der Schrittweite h abhängt.

Fall 1: Schrittweite h fix

Wahl: $h = \frac{\sqrt{2}}{8} \approx 0.1768$ (Matrixdimension: n(h) = 2120),

	α						
ω	10^{-8}	10^{-4}	10^{-2}	1	10^2	10^4	10^8
0	40	40	36	24	18	12	10
10^{-8}	40	40	36	24	18	12	10
10^{-4}	40	40	36	24	18	12	10
10^{-2}	40	40	36	24	18	12	12
1	40	40	36	26	24	22	22
10^2	40	44	40	40	40	40	40
10^4	30	38	38	38	38	38	38
10^8	11	10	10	10	10	10	10

Tabelle 8.1: PMINRES-Iterationen

Wahl: $h = \frac{\sqrt{2}}{16} \approx 0.0883$ (Matrixdimension: n(h) = 8840),

	α						
ω	10^{-8}	10^{-4}	10^{-2}	1	10^2	10^4	10^8
0	44	40	34	22	18	12	10
10^{-8}	44	40	34	22	10	12	10
10^{-4}	44	40	34	22	18	12	10
10^{-2}	44	40	34	22	18	12	12
1	44	40	36	26	22	22	22
10^2	44	45	44	44	44	44	44
10^4	36	42	42	42	42	42	42
10^8	13	11	11	11	11	11	11

Tabelle 8.2: PMINRES-Iterationen

Fall 2: Kontroll-Parameter α fix

Wahl: $\alpha = 1$,

		ω							
h	$n(h)$	0	10^{-8}	10^{-4}	10^{-2}	1	10^2	10^4	10^8
0.7071	105	24	24	24	24	26	30	22	10
0.3536	488	24	24	24	24	26	36	30	10
0.2357	1160	24	24	24	24	26	38	35	10
0.1768	2120	24	24	24	24	26	40	38	10
0.1414	3368	24	24	24	24	26	40	40	10
0.1179	4904	22	22	22	23	26	40	40	10
0.0884	8840	22	22	22	22	26	40	42	11

Tabelle 8.3: PMINRES-Iterationen

Fall 3: Frequenz ω fix

Wahl: $\omega = 1$,

		α						
h	$n(h)$	10^{-8}	10^{-4}	10^{-2}	1	10^2	10^4	10^8
0.7071	105	24	32	34	26	24	22	23
0.3536	488	34	38	38	26	24	2	23
0.2357	1160	24	32	34	26	24	22	23
0.1768	2120	40	40	36	26	24	22	22
0.1414	3368	42	40	36	26	22	22	22
0.1179	4904	42	40	36	26	22	22	22
0.0884	8840	44	40	36	26	22	22	22

Tabelle 8.4: PMINRES-Iterationen

8.4.1 Interpretation der Ergebnisse

Mit $k_{\max} = 44$ benötigten Iterationen liegt die bei den Tests höchstens auftretende Iterationsanzahl weit unterhalb der theoetrischen Grenze $k_0 = 106$. Die nummerischen Resultate für den zwei-dimensionalen Fall illustrieren also das theoretische Ergebnis!

Kapitel 9

Konklusion

Zusammenfassung

Zusammenfassend haben wir das optimale Kontrollproblem mit instationären Stokes-Gleichungen als Nebenbedingung gezeigt:

- zeitabhängiger und zeit-harmonischer Fall:
  - Existenz, Eindeutigkeit und Charakterisierung einer schwachen Lösung.
  - Formulierung des Optimalitätssystems.
  - Diskretisierung des Optimalitätssystems wobei für die zeitliche Diskretisierung eine unstetigen Galerkin-Methode angewendet wurde.
  - Formulierung des diskretisierten Optimalitätssystem in Form eines linearen Gleichungssystem, mit dünnbesetzter, regulärer, symmetrisch und indefiniter Blockmatrix in Sattelpunktsform.
- zeit-harmonischer Fall:
  - Konstruktion eines robusten Präkonditioniers in Block-Diagonalform.
  - Robuste Konvergenz für das PMINRES-Verfahren.

Mögliche nächste Schritte:

- Effiziente Implementierung von $\mathcal{P}$:

  Die Anwendung von $\mathcal{P}^{-1}$ entspricht wegen der Block-Diagonalform, der Lösung von linearen Gleichungssystemen der Form:

  $$I_h x = b \quad \text{und} \quad B_h I_h^{-1} B_h^T x = b,$$

  mit I_h symmetrisch und positiv definit. Das heißt für eine effiziente Implementierung des Präkonditionierers $\mathcal{P}$ ist es notwendig, auch für die Matrizen I_h und $B_h I_h^{-1} B_h^T$ einen robusten Präkonditionierer zu konstruieren.

- Konstruktion eines effizienten Lösers für den zeitabhängigen Fall:

  Wegen der ähnlichen Struktur von $\mathcal{A}$ und $\mathcal{A}^\omega$ lässt sich die in Kapitel 7 vorgestellte Strategie zur Konstruktion eines robusten Präkonditioniers, auch für das zeitabhängige Problem anwenden.

- Untersuchung des zeitabhängigen und zeitharmonischen Problems unter Hinzunahme von Einschränkungen an die Steuerung f, in der Form

  $$f_a(x) \leq f(x) \leq f_b(x) \quad \forall x \in \Omega.$$

  für gegebene Funktionen f_a und f_b.

Literaturverzeichnis

[1] W. Zulehner. *Numerical Methods for Continuum Mechanics 1.* Vorlesungsskriptum, Institut für Numerische Mathematik, Linz, 2006.

[2] W. Zulehner. *Interpolation in* $\mathbb{R}^n$. Seminar-Report, Institut für Numerische Mathematik, Linz, 2010.

[3] W. Zulehner. *Non-standard norms and robust estimates for saddle point problems.* Seminar-Report, Institut für Numerische Mathematik, Linz, 2010.

[4] W. Zulehner. *persönliche Mitteilung.* 2011.

[5] V. Simoncini. *persönliche Mitteilung.* 2011.

[6] D. Arnold. *Discretization by Finite Elements of a Model Parameter dependent Problem.* Department of Mathematics and Institute for Physical Science and Technology, Maryland, 1981

1980.

[7] F. Tröltzsch. *Optimale Steuerung partieller Differentialgleichungen.* 2. Auflage, Vieweg und Teubner, 2009.

[8] H.W. Alt. *Lineare Funktionalanalysis.* 5. Auflage, Springer, Berlin, 2006.

[9] D. Braess. *Finite Elemente - Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie.* 4. Auflage, Springer, Berlin, 2007.

[10] H.J. Reinhardt. *Nummerik gewöhnlicher Differentialgleichungen.* 1. Auflage, Walter de Gruyter, Berlin, 2008.

[11] A. Gupta. *Lecture Course On Variational Calculus.* Academic Publishers,Indien, 2009.

[12] G.O. Michler und H.J. Kowalsky. *Lineare Algebra.* 11. Auflage, Walter de Gruyter, Berlin, 1998.

[13] H. Heuser. *Lehrbuch der Analysis 2.* 13. Auflage, Teubner, Wiesbaden, 2004.

[14] Z. Kanzow. *Nummerik linearer Gleichungssysteme: Direkte und iterative Verfahren.* 1. Auflage, Springer, Berlin, 2005.

[15] A. Greenbaum. *Iterative Methods for solving linear Systems.* 17. Auflage, Frontiers in Applied Mathematics, Philadelphia, 1997.

[16] H. Elman, D. Silvester und A. Wathen. *Finite Elements and Fast Iterative Solvers.* 2. Auflage, Oxford University Press, Philadelphia, 2006.

[17] D. Werner. *Funktionalanalysis.* 6. Auflage, Springer, Berlin, 2007.

[18] A. Walter. *Nummerische Verfahren der konvexen, nicht glatten Optimierung.* 1. Auflage, Teubner, Wiesbaden, 2004.

[19] S. Gross und A. Reusken. *Numerical Methods for Two-Phase incompressible Flows.* 1. Auflage, Springer, Berlin, 2011.

[20] A. Fischer, K. Vetters und W.Schirotzek. *Lineare Algebra - Eine Einführung für Ingenieure und Naturwissenschaftler.* 1. Auflage, Teubner, Wiesbaden, 2003.

[21] M. Antin. *Algebra.* 1. Auflage, Birkhauser, Basel, 1993.

[22] S. Sauter und C. Schwab. *Randelement Methoden - Analyse, Numerik und Implementierung schneller Algorithmen.* 1. Auflage, Teubner, Wiesbaden, Juni 2004.

[23] F. Brezzi und M. Fortin. *Mixed and Hybrid Finite Element Methods.* 1. Auflage, Springer, New York, 1991.

[24] M.F. Murphy, G.H. Golub. und A.J. Wathen. *A note on preconditioning for indefinite linear systems.* Report, SIAM Computer Sience, Oxford, 1999.

Eidesstattliche Erklärung

Ich, Krendl Wolfgang, erkläre an Eides statt, dass ich die vorliegende Diplomarbeit selbständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Linz, Oktober 2011

_____

Krendl Wolfgang