Optimierung und Stabilisierung von inkompressiblen Strömungen mit Randsteuerung in M.E.S.S.

An der Fakultät für Mathematik
der Otto-von-Guericke-Universität Magdeburg
zur Erlangung des akademischen Grades
Master of Science
angefertigte

Masterarbeit

vorgelegt von
Maximilian Behr
geboren am 21.03.1990 in Schwedt/Oder,
Studiengang Mathematik,
Studienrichtung Mathematik.

19. November 2015

Betreut am Max-Plank-Institut
für Dynamik komplexer technischer Systeme von
Dr. Jan Heiland

Inhaltsverzeichnis

Abbildungsverzeichnis

Quelltextausschnitte

Symbolverzeichnis

$\mathfrak{F}(X,Y)$	$= \{f \mid f : X \to Y\}$
id	Identitätsabbildung
f^{ad}	zu f adjungierte Abbildung
$\mathfrak{L}(X,Y)$	$= \{f \mid f \in \mathfrak{F}(X,Y),\ f$ ist linear und stetig$\}$
$X \cong Y$	X ist isomorph zu Y
$\mathfrak{B}_r(y)$	$= \{x \mid \|\|x - y\|\| < r\}$
$X \subseteq\subseteq Y$	der topologische Abschluss von X ist kompakt und in Y enthalten
$\mathbb{N}$	Menge der natürlichen Zahlen
$\mathbb{N}_0$	Menge der natürlichen Zahlen mit 0
$\mathbb{Z}$	Menge der ganzen Zahlen
$\mathbb{Q}$	Menge der rationalen Zahlen
$\mathbb{R}$	Menge der reellen Zahlen
$\mathbb{R}^+$	Menge der positiven reellen Zahlen
$\mathbb{R}^-$	Menge der negativen reellen Zahlen
$\mathbb{C}$	Menge der komplexen Zahlen
$\mathbb{C}^+$	Menge der komplexen Zahlen mit positiven Realteil
$\mathbb{C}^-$	Menge der komplexen Zahlen mit negativen Realteil
$\mathbb{R}^n$	Vektorraum aller n-Tupel reeller Zahlen
$\mathbb{C}^n$	Vektorraum aller n-Tupel komplexer Zahlen
$\Re(z)$	Realteil von $z \in \mathbb{C}$
$\Im(z)$	Imaginärteil von $z \in \mathbb{C}$
$\overline{z}$	komplex Konjugierte von $z \in \mathbb{C}$
$\mathbb{K}^{n\times m}$	Menge aller $n \times m$ Matrizen über dem Körper $\mathbb{K}$
$\mathrm{GL}(n,\mathbb{K})$	Menge aller invertierbaren $n \times n$ Matrizen über dem Körper $\mathbb{K}$
$A_{i,j}$	Eintrag i,j der Matrix A
$A_{i,*}$	i-te Zeile der Matrix A
$A_{*,j}$	j-te Spalte der Matrix A
I_n	Einheitsmatrix mit n Zeilen und n Spalten
A^T	Transponierte von A
$\mathrm{tr}(A)$	$= \sum_{i=1}^{n} A_{i,i}$ Spur von A
$\Re(A)$	Realteil von $A \in \mathbb{C}^{n\times m}$
$\Im(A)$	Imaginärteil von $A \in \mathbb{C}^{n\times m}$
$\overline{A}$	komplex Konjugierte von $A \in \mathbb{C}^{n\times m}$
A^H	komplex konjugiert Transponierte von $A \in \mathbb{C}^{n\times m}$
$\ker(A)$	$= \{x \mid Ax = 0\}$
$\Lambda(A)$	$= \{\lambda \mid \exists x \neq 0 \colon Ax = \lambda x\}$
$\rho(A)$	$= \sup\{\|\mu\| \mid \mu \in \Lambda(A)\}$
$\Lambda(A,M)$	$= \{\lambda \mid \exists x \notin \ker(M) \colon Ax = \lambda Mx\}$
$\rho(A,M)$	$= \sup\{\|\mu\| \mid \mu \in \Lambda(A,M)\}$
$\mathrm{span}(\{x_1,\ldots,x_n\})$	$= \{\sum_{i=1}^{n} \lambda_i x_i \mid \lambda_i \in \mathbb{K}\}$, wobei $\mathbb{K}$ ein Körper ist

$A > 0$	A ist positiv definit
$A < 0$	A ist negativ definit
$A \geq 0$	A ist positiv semidefinit
$A \leq 0$	A ist negativ semidefinit
$A > B$	$A - B$ ist positiv definit
$A < B$	$A - B$ ist negativ definit
$A \geq B$	$A - B$ ist positiv semidefinit
$A \leq B$	$A - B$ ist negativ semidefinit
$\mathbb{S}^n$	Menge der symmetrischen Matrizen im $\mathbb{R}^{n \times n}$
$\mathbb{S}^n_+$	Menge der symmetrisch, positiv semidefiniten Matrizen im $\mathbb{R}^{n \times n}$
$\mathbb{S}^n_{++}$	Menge der symmetrisch, positiv definiten Matrizen im $\mathbb{R}^{n \times n}$
$\|\|A\|\|_2$	Spektralnorm von $A \in \mathbb{C}^{n \times n}$
$\langle A, B \rangle_F$	Frobeniusskalarprodukt zwischen $A, B \in \mathbb{C}^{n \times m}$
$\|\|A\|\|_F$	Frobeniusnorm von $A \in \mathbb{C}^{n \times n}$
$\|\|x\|\|_p$	$= \left(\sum_{i=1}^{n} \|x_i\|^p \right)^{\frac{1}{p}}$ für $1 \leq p < \infty$ und $x \in \mathbb{C}^n$
$\|\|x\|\|_\infty$	$= \max_{i=1,\ldots,n} \|x_i\|$ für $x \in \mathbb{C}^n$
D^α	$= \frac{\partial^{\|\alpha\|}}{\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}}$, wobei $\alpha \in \mathbb{N}_0^n$ und $\|\alpha\| = \sum_{i=1}^{n} \alpha_i$

1 Vorwort

Danksagung

Hiermit möchte ich mich bei Dr. Jan Heiland, Dr. Jens Saak, Dipl.-Math. techn. Heiko Weichelt und Dipl.-Math. Martin Köhler bedanken.

Motivation

Man stelle sich v_s als ein gewünschtes Geschwindigkeitsprofil einer Strömung mit zugehörigen Druck p_s in einem Becken vor. In dem Becken ist derzeit eine Strömung v mit zugehörigen Druck p. An einem Rand des Beckens befindet sich eine Düse, deren Stärke man mit u steuern kann. Wir lassen Flüssigkeit über einen anderen Rand in das Becken laufen. Das Becken besitzt auch ein Randstück, an dem die Flüssigkeit ungehindert ausströmen kann. Das ungehinderte Ausströmen ist mittels einer „natürlichen“ Ausflussbedingung modelliert. An den übrigen Randstücken fordert man eine Haftrandbedingung, d. h. $v = 0$. Ziel ist es, die Düse mit u derart zu steuern, dass die Abweichung von v zum gewünschten Profil v_s nach einer gewissen Zeit klein wird. Das bedeutet, dass dann $(v, p) \approx (v_s, p_s)$ gelten soll. Man stelle sich vor, dass v zum Zeitpunkt $t = 0$ ein wenig vom Profil v_s abweicht.
Wie soll man nun die Düse mit u steuern, damit das gewünschte Ziel erreicht wird?
Man kann versuchen in Abhängigkeit von der Abweichung die Steuerung u derart zu wählen, dass $(v, p) \approx (v_s, p_s)$ nach einem gewissen Zeitraum gilt und in dem Becken nahezu „Ruhe“ herrscht. Ziel ist es, eine Zuordnung zu finden, die jeder Abweichung, die nicht allzu groß ist, eine geeignete Steuerung u zuordnet. Wir nennen diese Zuordnung $-K^T$ und damit erhalten wir folgende Darstellung für die Steuerung der Düse in Abhängigkeit von der Abweichung v_δ

$$u = -K^T v_\delta.$$

Ziel dieser Arbeit ist es, das obige Problem mathematisch zu behandeln und an Beispielen numerische Simulationen durchzuführen.

2 Linear-quadratische Regelungsprobleme

In diesem Kapitel sollen das linear-quadratische Regelungsproblem definiert und einige Ergebnisse aus [13, 14, 40] zusammengefasst werden. Wir betrachten zunächst den Fall eines endlichen Zeithorizonts und werden sehen, dass eine autonome nichtlineare Matrix-Differentialgleichung eine wesentliche Rolle spielt.
Später wollen wir uns dem zeitasymptotischen Fall widmen. Um die Resultate aus dem Fall endlicher Zeit in den zeitasymptotischen Fall zu übertragen, sucht man geeignete stationäre Punkte der autonomen Matrix-Differentialgleichung. Hierzu untersucht man das Grenzwertverhalten. Ist die Existenz eines geeigneten Grenzwerts gesichert, werden wir eine nichtlineare Matrixgleichung erhalten.
Zur Lösung dieses nichtlinearen Nullstellenproblems werden wir ein Newton-Verfahren anwenden. Innerhalb des Newton-Verfahren müssen mehrere lineare Gleichungssysteme einer bestimmten Form gelöst werden. Es erweist sich jedoch als hilfreich, diese linearen Gleichungssysteme nicht als solche, sondern in einer kompakteren Schreibweise mit Matrizen aufzufassen. Die Lösung ist daher eine Matrix X und kein Spaltenvektor x.
In der Praxis kommt es uns sehr zur Hilfe, dass die rechte Seite dieser linearen Matrixgleichung eine spezielle Struktur hat und die Lösung X symmetrisch ist und häufig einen numerisch kleinen Rang hat. Man speichert daher X mittels $X \approx ZZ^T$ und Z hat „wenig“ Spalten. Z nennt man Niedrigrangfaktor von X.

2.1 Linear-quadratische Regelungsprobleme in endlicher Zeit

Definition 2.1 (Linear-quadratisches Regelungsproblem (LQR-Problem), [14, S. 2])
Sei $M \in \mathrm{GL}(n, \mathbb{R}), A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}, C \in \mathbb{R}^{p\times n}, Q \in \mathbb{S}^p, R \in \mathbb{S}^m_+, x_0 \in \mathbb{R}^n$ *und* $t_f \in (0, \infty]$.
Das Optimierungsproblem

$$\min_{u\in\mathcal{PC}_m} \mathcal{J}(u) := \frac{1}{2}\int_0^{t_f} y(t)^T Q y(t) + u(t)^T R u(t)\,\mathrm{d}t \tag{2.1}$$

unter der Bedingung

$$M\dot{x}(t) = Ax(t) + Bu(t) \qquad \text{für alle } t \in (0, t_f], \tag{2.2}$$
$$y(t) = Cx(t) \qquad \text{für alle } t \in [0, t_f], \tag{2.3}$$
$$x(0) = x_0, \tag{2.4}$$

wobei $\mathcal{PC}_m := \{u \in \mathfrak{F}([0, t_f], \mathbb{R}^m) \mid u \text{ ist stückweise stetig auf } [0, t_f]\}$*, nennen wir linear-quadratisches Regelungsproblem. Falls* $M \neq I_n$ *nennen wir das linear-quadratische Regelungsproblem verallgemeinertes LQR-Problem, falls* $M = I_n$ *nur LQR-Problem.* x *nennen wir Zustandsfunktion,* $x(t)$ *Zustand (zum Zeitpunkt* t*),* u *Steuerung und* $\mathcal{J}$ *Kostenfunktional.*

Wir wollen uns zunächst auf den Fall $t_f < \infty$ konzentrieren. Wir nehmen vorerst $M = I_n$ an.

Bemerkung 2.1
Falls $M \neq I_n$ können wir die Transformation

$$\dot{x}(t) = M^{-1}Ax(t) + M^{-1}Bu(t) \Leftrightarrow M\dot{x}(t) = Ax(t) + Bu(t)$$

betrachten. Die obige Transformation ändert das Kostenfunktional $\mathcal{J}$ nicht.

Notwendige Optimalitätsbedinungen für das LQR-Problem liefert folgender Satz.

Satz 2.1 (Pontryagin'sche Minimumprinzip, [13, Satz 4.2], [14, Prop. 2.2])
Gegeben sei ein LQR-Problem ($M = I_n$) mit $u_* \in \mathcal{PC}_m$ sowie x_* die zur Steuerung gehörende Zustandsfunktion. Falls $\mathcal{J}(u_*) \leq \mathcal{J}(u)$ für alle $u \in \mathcal{PC}_m$ gilt, so erfüllt u_* die notwendigen Optimalitätsbedingungen:

- *$\mathcal{H}(x_*, u_*, \mu) \leq \mathcal{H}(x, u, \mu)$ für alle $u \in \mathcal{PC}_m$ und für alle $t \in [0, t_f]$.*
- *$\mu \in \mathfrak{F}([0, t_f], \mathbb{R}^n)$ erfüllt die adjungierte Gleichung*

$$\dot{\mu}(t) = -\mathcal{H}_x, \text{ d. h. } \dot{\mu}_j(t) = -\frac{\partial \mathcal{H}}{\partial x_j} \text{ für } j = 1, \ldots, n.$$

- *$\mu(t_f) = 0$.*

$\mathcal{H}$ ist durch $\mathcal{H}(x, u, \mu) = \frac{1}{2}(x^T C^T QCx + u^T Ru) + \mu^T (Ax + Bu)$ gegeben.

Wenden wir Satz 2.1 an, so erhalten wir:

$$\dot{\mu} = -\mathcal{H}_x = -C^T QCx - A^T \mu.$$

Da an u keine Beschränktheitsbedingungen gegeben waren, gilt für ein Minimum

$$0 = \mathcal{H}_u = Ru + B^T \mu.$$

Dies liefert folgende lineare Randwertaufgabe.

Satz 2.2 (Randwertaufgabe, [13, Satz 4.5, Satz 4.7], [14, Thm. 2.3])
Gegeben sei ein LQR-Problem ($M = I_n$), $u_* \in \mathcal{PC}_m$ eine optimale Steuerung und x_* die zugehörige Zustandsfunktion. Dann gibt es $\mu \in \mathfrak{F}([0, t_f], \mathbb{R}^n)$,
sodass $\begin{bmatrix} x_*(t)^T & \mu(t)^T & u_*(t)^T \end{bmatrix}^T$ das System

$$\begin{bmatrix} I_n & 0 & 0 \\ 0 & -I_n & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_*(t) \\ \dot{\mu}(t) \\ \dot{u}_*(t) \end{bmatrix} = \begin{bmatrix} A & 0 & B \\ C^T QC & A^T & 0 \\ 0 & B^T & R \end{bmatrix} \begin{bmatrix} x_*(t) \\ \mu(t) \\ u_*(t) \end{bmatrix} \tag{2.5}$$

mit den Randbedingungen $x_*(0) = x_0$ und $\mu(t_f) = 0$ erfüllt. μ nennen wir Kozustandsfunktion. Falls zusätzlich $Q \in \mathbb{S}^p_+$ und $\begin{bmatrix} x_*(t)^T & \mu(t)^T & u_*(t)^T \end{bmatrix}^T$ (2.5) mit $x_*(0) = x_0$ und $\mu(t_f) = 0$ erfüllt, dann gilt $\mathcal{J}(u_*) \leq \mathcal{J}(u)$ für alle (x, u) mit $u \in \mathcal{PC}_m$, die (2.2), (2.3) und (2.4) erfüllen.

Falls $Q \in \mathbb{S}^p_+$ liefert Satz 2.2 ein Optimalitätszertifikat für das LQR-Problem.
Falls $R \in \mathbb{S}^m_{++}$ erhalten wir

$$u(t) = -R^{-1}B^T \mu(t).$$

Einsetzen in (2.2) ergibt

$$\dot{x}(t) = Ax(t) - BR^{-1}B^T\mu(t).$$

Damit lässt sich (2.5) zu

$$\begin{bmatrix}\dot{x}(t)\\ \dot{\mu}(t)\end{bmatrix} = \begin{bmatrix}A & -BR^{-1}B^T\\ -C^TQC & -A^T\end{bmatrix}\begin{bmatrix}x(t)\\ \mu(t)\end{bmatrix}$$

vereinfachen. Mit dem Ansatz $\mu(t) = X(t)x(t)$ erhält man

$$\begin{bmatrix}\dot{x}(t)\\ \dot{X}(t)x(t) + X(t)\dot{x}(t)\end{bmatrix} = \begin{bmatrix}A & -BR^{-1}B^T\\ -C^TQC & -A^T\end{bmatrix}\begin{bmatrix}x(t)\\ X(t)x(t)\end{bmatrix}$$

bzw.

$$(\dot{X}(t) + C^TQC + A^TX(t) + X(t)A - X(t)BR^{-1}B^TX(t))x(t) = 0 \text{ für alle } t \in [0, t_f].$$

Durch Variation von x erhalten wir die Matrix-Riccati-Differentialgleichung

$$\begin{aligned}\dot{X}(t) &= -(\mathcal{R}(X)(t)) :=\\ &\quad - (C^TQC + A^TX(t) + X(t)A - X(t)BR^{-1}B^TX(t)) \text{ für alle } t \in [0, t_f], \qquad (2.6)\\ X(t_f) &= 0. \qquad (2.7)\end{aligned}$$

Lemma 2.1
Falls die Matrix-Riccati-Differentialgleichung (2.6), (2.7) *eine eindeutige Lösung X hat, so ist diese zu allen Zeiten symmetrisch.*

Beweis. Mit $Q \in \mathbb{S}^p$ und $R \in \mathbb{S}^m$ erhalten wir

$$\begin{aligned}\dot{X^T}(t) = \dot{X}(t)^T = (-\mathcal{R}(X)(t))^T =\\ -(\mathcal{R}(X)(t))^T = -(\mathcal{R}(X)^T(t)) = -(\mathcal{R}(X^T)(t)) \text{ für alle } t \in [0, t_f].\end{aligned}$$

□

Lemma 2.2
$\mathcal{R} \in \mathfrak{F}(\mathbb{R}^{n\times n}, \mathbb{R}^{n\times n})$ *ist lokal Lipschitz-stetig.*

Beweis.

$$||\mathcal{R}(X_1) - \mathcal{R}(X_2)||_2 \leq 2||A||_2||X_1 - X_2||_2 + ||X_2BR^{-1}B^TX_2 - X_1BR^{-1}B^TX_1||_2.$$

Für den zweiten Summanden erhalten wir

$$\begin{aligned}&||X_2BR^{-1}B^TX_2 - X_1BR^{-1}B^TX_1||_2 =\\ &\tfrac{1}{2}||(X_2 - X_1)BR^{-1}B^T(X_2 + X_1) + (X_1 + X_2)BR^{-1}B^T(X_2 - X_1)||_2 \leq\\ &||X_2 - X_1||_2||BR^{-1}B^T||_2||X_1 + X_2||_2 \leq\\ &||X_2 - X_1||_2||BR^{-1}B^T||_2(||X_1||_2 + ||X_2||_2) \leq\\ &2||BR^{-1}B^T||_2(||Y||_2 + r)||(X_2 - X_1)||_2.\end{aligned}$$

Als Lipschitzkonstante wählen wir

$$L := 2(||A||_2 + ||BR^{-1}B^T||_2(||Y||_2 + r)).$$

Damit erhalten wir folgende Abschätzung

$$||\mathcal{R}(X_1) - \mathcal{R}(X_2)||_2 \leq L||X_1 - X_2||_2 \ \forall X_1, X_2 \in \mathfrak{B}_r(Y).$$

□

Der Satz von Picard-Lindelöf liefert uns lokale Existenz und Eindeutigkeit von Lösungen für die Matrix-Riccati-Differentialgleichung (2.6) und (2.7).
Für globale Existenz der Lösung auf $(-\infty, t_f]$ ist sicherzustellen, dass kein „blowup“ in endlicher Zeit auftreten kann. Hierzu sei auf [40, Kap. 10.1, Kor.10.1] verwiesen.

Lemma 2.3 ([40, Hilfssatz 10.4])
Sei X eine Lösung von (2.6)*, so gilt*

$$\dot{X}(t_f) \geq 0 \ (\dot{X}(t_f) \leq 0) \Rightarrow \dot{X}(t) \geq 0 \ (\dot{X}(t) \leq 0) \text{ für alle } t \in (-\infty, t_f].$$

Bemerkung 2.2
Lemma 2.3 *und* (2.7) *implizieren sofort eine „Monotonieeigenschaft“, denn*

$$\begin{aligned} \dot{X}(t_f) = -\mathcal{R}(X)(t_f) = -C^TQC \leq 0 \ \Rightarrow \\ \dot{X}(t) \leq 0 \text{ für alle } t \in (-\infty, t_f] \Rightarrow \\ X(t_2) - X(t_1) = \int_{t_1}^{t_2} \dot{X}(s)ds \leq 0 \text{ für alle } t_1, t_2 \in (-\infty, t_f] \text{ mit } t_1 \leq t_2. \end{aligned}$$

Damit ist X „monoton fallend“ auf $(-\infty, t_f]$ und aufgrund von (2.7) *und* Lemma 2.1 *gilt*

$$X(t) \in \mathbb{S}_+^n \text{ für alle } t \in (-\infty, t_f].$$

Darüber hinaus kann eine Darstellung für die optimalen Kosten gewonnen werden vgl. [13, Satz 4.10]. Wir fassen zusammen.

Satz 2.3 ([13, Satz 4.10], [14, Thm. 2.4])
Sei $Q \in \mathbb{S}_+^p$ und $R \in \mathbb{S}_{++}^m$. Dann hat das LQR-Problem ($M = I_n$) die eindeutige Lösung

$$u_*(t) = -R^{-1}B^TX_*(t)x(t) \text{ für alle } t \in [0, t_f],$$

wobei X_* die eindeutige Lösung der Matrix-Riccati-Differentialgleichung (2.6)*,* (2.7) *ist. Des Weiteren gilt*

$$\mathcal{J}(u_*) = \frac{1}{2}x_0^TX_*(0)x_0. \tag{2.8}$$

2.2 Linear-quadratische Regelungsprobleme mit unendlichem Zeithorizont

Nun wollen wir den Fall $t_f = \infty$ betrachten.

Definition 2.2 (Stabilisierbarkeit, [13, Def. 2.13])
Sei $A \in \mathbb{R}^{n\times n}$ und $B \in \mathbb{R}^{n\times m}$. Wir nennen das Matrixpaar (A, B) stabilisierbar, falls

$$\forall x_0 \in \mathbb{R}^n \ \exists u \in \mathcal{PC}_m \colon \lim_{t\to\infty} x(t;u) = 0,$$

wobei $x(t;u)$ die eindeutige Lösung von

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ x(0) &= x_0 \end{aligned}$$

ist.

Lemma 2.4 (Charakterisierung Stabilisierbarkeit, [14, Def. 2.1], [13, Satz 2.15])
Die folgenden Bedingungen sind äquivalent zur Stabilisierbarkeit von (A, B).

- $\operatorname{rank}(A - \lambda I_n, B) = n \quad \forall \lambda \in \mathbb{C}^+ \cup i\mathbb{R}$.
- $\exists K \in \mathbb{R}^{m\times n}$ *mit* $\Lambda(A + BK) \subset \mathbb{C}^-$.
- *Falls $z \neq 0$ ein Links-Eigenvektor von A zu einem Eigenwert λ mit $\Re(\lambda) \geq 0$ ist, so muss $z^H B \neq 0$ gelten.*

Definition 2.3 (Entdeckbarkeit, [13, Def. 2.22])
Sei $A \in \mathbb{R}^{n\times n}$ und $C \in \mathbb{R}^{p\times n}$. Wir nennen das Matrixpaar (A, C) ist entdeckbar, falls für jede Lösung z von $\dot{z}(t) = Az(t)$ und $Cz(t) \equiv 0$ folgt $\lim_{t\to\infty} z(t) = 0$.

Lemma 2.5 (Charakterisierung Entdeckbarkeit, [14, Def. 2.1])
Die folgenden Bedingungen sind äquivalent zur Entdeckbarkeit von (A, C).

- $\operatorname{rank}([A^T - \lambda I_n, C^T]^T) = n \quad \forall \lambda \in \mathbb{C}^+ \cup i\mathbb{R}$.
- $\exists K \in \mathbb{R}^{n\times p}$ *mit* $\Lambda(A + KC) \subset \mathbb{C}^-$.

Betrachten wir einen unbeschränkten Zeithorizont ($M = I_n$), so sollten wir $\mathcal{J}(u) < \infty$ fordern. Hieraus folgt aber sofort

$$\begin{aligned} \lim_{t\to\infty} u(t) &= 0, \\ \lim_{t\to\infty} x(t)^T C^T Q C x(t) &= 0. \end{aligned}$$

Setzen wir Entdeckbarkeit von $(A, C^T QC)$ voraus, so erhalten wir

$$\lim_{t\to\infty} x(t) = 0.$$

Sucht man Lösungen von autonomen Systemen, die auf ganz $\mathbb{R}$ existieren, dann gibt es nur zwei Fälle. Entweder ist die Lösung nichtkonstant und periodisch oder die Lösung ist konstant vgl. [3, Kap. 3.2, Satz 3.2.5].
Aufgrund der „Monotonieeigenschaft“ in Bemerkung 2.2 konzentriert man sich darauf, kritische Punkte von $\mathcal{R}$ zu finden.
Nehmen wir vorerst an $\mathcal{R}(X) = 0$ hätte eine eindeutige Lösung X_∞. In Rückblick auf Satz 2.3

würden wir $\Lambda(A - BR^{-1}B^T X_\infty) \subset \mathbb{C}^-$ fordern. Dies motiviert die Stabilisierbarkeit von (A, B) zu verlangen.
Um kritische Punkte von $\mathcal{R}$ zu finden, würde man untersuchen, ob $\lim_{t \to -\infty} X(t)$ existiert, wobei X die eindeutige Lösung der Matrix-Riccati-Differentialgleichung (2.6), (2.7) auf $(-\infty, t_f]$ ist. Falls der Grenzwert existiert, so ist dieser auch ein kritischer Punkt vgl. [60, XI.Autonome System (g), (h)].
Die Existenz des Grenzwertes und die Stabilisierungseigenschaft nachzuweisen ist etwas aufwändiger. Wir verweisen daher auf die ausführliche Darstellung in [40, Kap. 10.2].
Wenn der Grenzwert existiert, sieht man die Symmetrie mit Lemma 2.1 einfach, denn $(\cdot)^T \in \mathfrak{L}(\mathbb{R}^{n\times n}, \mathbb{R}^{n\times n})$.
Folgendes Lemma wird sich später als hilfreich herausstellen.

Lemma 2.6 (Eigenschaften zueinander kongruenter Matrizen)
Sei $M \in \mathrm{GL}(n, \mathbb{R})$. Wir betrachten

$$\begin{aligned} \mathcal{T}_M &\in \mathfrak{F}(\mathbb{R}^{n\times n}, \mathbb{R}^{n\times n}), \\ X &\mapsto M^{-T} X M^{-1}. \end{aligned}$$

Dann gilt:
(i) $\mathcal{T}_M \in \mathfrak{L}(\mathbb{R}^{n\times n}, \mathbb{R}^{n\times n})$.
(ii) $\mathcal{T}_M(X)$ *und* X *sind äquivalent für alle* $X \in \mathbb{R}^{n\times n}$.
(iii) $\mathrm{rank}(\mathcal{T}_M(X)) = \mathrm{rank}(X)$ *für alle* $X \in \mathbb{R}^{n\times n}$.
(iv) $\mathcal{T}_M^{-1} = \mathcal{T}_{M^{-1}} \in \mathfrak{L}(\mathbb{R}^{n\times n}, \mathbb{R}^{n\times n})$.
(v) $0 < X$ $(0 \le X) \Leftrightarrow 0 < \mathcal{T}_M(X)$ $(0 \le \mathcal{T}_M(X))$.
(vi) $0 > X$ $(0 \ge X) \Leftrightarrow 0 > \mathcal{T}_M(X)$ $(0 \ge \mathcal{T}_M(X))$.
(vii) $X_1 < X_2$ $(X_1 \le X_2) \Leftrightarrow \mathcal{T}_M(X_1) < \mathcal{T}_M(X_2)$ $(\mathcal{T}_M(X_1) \le \mathcal{T}_M(X_2))$.
(viii) $X_1 > X_2$ $(X_1 \ge X_2) \Leftrightarrow \mathcal{T}_M(X_1) > \mathcal{T}_M(X_2)$ $(\mathcal{T}_M(X_1) \ge \mathcal{T}_M(X_2))$.
(ix) $X = X^T \Leftrightarrow \mathcal{T}_M(X) = \mathcal{T}_M(X)^T$.

Beweis.

Zu (i) :
Die Linearität zeigen wir nicht.
Jede lineare Abbildung zwischen endlichdimensionalen normierten Vektorräumen ist stetig ($\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$).

Zu (ii) :
Das gilt per Definition von $\mathcal{T}_M$.

Zu (iii) :
Das folgt aus (ii).

Zu (iv) :
$\mathcal{T}_M \mathcal{T}_{M^{-1}}(X) = \mathcal{T}_M(M^T X M) = M^{-T} M^T X M M^{-1} = X$ für alle $X \in \mathbb{R}^{n\times n}$.

Zu (v), (vi) :
$0 < X \Leftrightarrow 0 < x^T X x \ \forall x \in \mathbb{R}^n \setminus \{0\} \Leftrightarrow$
$0 < (M^{-1}x)^T X M^{-1} x = x^T \mathcal{T}_M(X) x \ \forall x \in \mathbb{R}^n \setminus \{0\} \Leftrightarrow 0 < \mathcal{T}_M(X)$.
Die Fälle $\le, >, \ge$ gehen analog.

Zu (vii), $(viii)$:
Ergibt sich aus der Linearität von $\mathcal{T}_M$ sowie (v) und (vi).

Zu (ix) :

Der Fall $(\Rightarrow)$: $\mathcal{T}_M(X) = \mathcal{T}_M(X^T) = \mathcal{T}_M(X)^T$.

Der Fall $(\Leftarrow)$: $\mathcal{T}_M(X) = \mathcal{T}_M(X)^T = \mathcal{T}_M(X^T)$ und die Injektivität liefert $X = X^T$.

□

Satz 2.4 (stabilisiernde Lösung der algebraischen Riccatigleichung, [14, Thm. 2.7])
Sei $C^TQC \geq 0, BR^{-1}B^T \geq 0$, $(A, BR^{-1}B^T)$ stabilisierbar und (A, C^TQC) entdeckbar. Dann hat die algebraische Riccatigleichung

$$0 = \mathcal{R}(X) = C^TQC + A^TX + XA - XBR^{-1}B^TX \tag{2.9}$$

die eindeutige, symmetrisch positiv semidefinite, stabilisierende Lösung X_∞,
d. h. $\Lambda(A - BR^{-1}B^TX_\infty) \subset \mathbb{C}^-$.

Korollar 2.1
Sei $M \in \mathrm{GL}(n, \mathbb{R}), C^TQC \geq 0, BR^{-1}B^T \geq 0$, $(M^{-1}A, M^{-1}BR^{-1}B^T)$ stabilisierbar und $(M^{-1}A, C^TQC)$ entdeckbar. Dann hat die verallgemeinerte algebraische Riccatigleichung

$$0 = \mathcal{R}(X) = C^TQC + A^TXM + M^TXA - M^TXBR^{-1}B^TXM \tag{2.10}$$

die eindeutige, symmetrisch positiv semidefinite, stabilisierende Lösung X_∞,
d. h. $\Lambda(A - BR^{-1}B^TX_\infty M, M) \subset \mathbb{C}^-$.

Beweis. Wir führen den allgemeinen Fall auf die Situation wie in Satz 2.4 zurück.
Es gilt $0 \leq BR^{-1}B^T \Leftrightarrow 0 \leq \mathcal{T}_{M^T}(BR^{-1}B^T)$.
$(M^{-1}A, M^{-1}BR^{-1}B^T)$ ist stabilisierbar ist äquivalent zu $(M^{-1}A, M^{-1}BR^{-1}B^TM^{-T})$ ist stabilisierbar.
Anwendung von Satz 2.4 liefert nun, dass die algebraische Riccatigleichung

$$0 = C^TQC + A^TM^{-T}X + XM^{-1}A - XM^{-1}BR^{-1}B^TM^{-T}X$$

eine eindeutige, symmetrisch positiv semidefinite, stabilisierende Lösung X_∞, d. h.

$$\begin{aligned}\mathbb{C}^- \supset \Lambda(M^{-1}A - M^{-1}BR^{-1}B^TM^{-T}X_\infty) &= \Lambda(A - BR^{-1}B^TM^{-T}X_\infty M^{-1}) \\ &= \Lambda(A - BR^{-1}B^T\mathcal{T}_M(X_\infty)) \\ &= \Lambda(A - BR^{-1}B^T\mathcal{T}_M(X_\infty)M, M).\end{aligned}$$

Wir erhalten

$$\begin{aligned}&0 = C^TQC + A^TM^{-T}X_\infty + X_\infty M^{-1}A - X_\infty M^{-1}BR^{-1}B^TM^{-T}X_\infty \Leftrightarrow \\ &0 = M^{-T}(C^TQC + A^TM^{-T}X_\infty + X_\infty M^{-1}A - X_\infty M^{-1}BR^{-1}B^TM^{-T}X_\infty)M^{-1} \\ &\quad = M^{-T}C^TQCM^{-1} + M^{-T}A^T\mathcal{T}_M(X_\infty) + \mathcal{T}_M(X_\infty)AM^{-1} - \mathcal{T}_M(X_\infty)BR^{-1}B^T\mathcal{T}_M(X_\infty) \Leftrightarrow \\ &0 = C^TQC + A^T\mathcal{T}_M(X_\infty)M + M^T\mathcal{T}_M(X_\infty)A - M^T\mathcal{T}_M(X_\infty)BR^{-1}B^T\mathcal{T}_M(X_\infty)M.\end{aligned}$$

Dies bedeutet, dass $\mathcal{T}_M(X_\infty)$ die verallgemeinerte algebraische Riccatigleichung löst. Nun beachtet man die Eigenschaften (v) und (ix) aus Lemma 2.6, die Bijektivität von $\mathcal{T}_M$ und dass $\Lambda(M^{-1}A - M^{-1}BR^{-1}B^TM^{-T}X_\infty) = \Lambda(A - BR^{-1}B^T\mathcal{T}_M(X_\infty)M, M)$.
Da die algebraische Riccatigleichung eine eindeutige, symmetrisch positiv semidefinite, stabili-

sierende Lösung besitzt, hat die verallgemeinerte algebraische Riccatigleichung ebenfalls eine eindeutige, symmetrisch positiv semidefinite stabilisierende Lösung. □

Satz 2.5 (Optimale Steuerung für das LQR-Problem mit $t_f = \infty$, [14, Thm. 2.8])
Sei $Q \in \mathbb{S}^n_+$, $R \in \mathbb{S}^m_{++}$, (A, B) stabilisierbar und (A, C^TQC) entdeckbar. Dann hat das LQR-Problem eine eindeutige Lösung

$$u_\infty(t) = -R^{-1}B^TX_\infty x(t), \tag{2.11}$$

wobei X_∞ die eindeutige, symmetrisch positiv semidefinite, stabilisierende Lösung der algebraischen Riccatigleichung (2.9) *ist.*

Korollar 2.2
Sei $M \in \mathrm{GL}(n, \mathbb{R}), Q \in \mathbb{S}^n_+$, $R \in \mathbb{S}^m_{++}$, $(M^{-1}A, M^{-1}B)$ stabilisierbar und $(M^{-1}A, C^TQC)$ entdeckbar. Dann hat das verallgemeinerte LQR-Problem eine eindeutige Lösung

$$u_\infty(t) = -R^{-1}B^TX_\infty Mx(t), \tag{2.12}$$

wobei X_∞ die eindeutige, symmetrisch positiv semidefinite, stabilisierende Lösung der verallgemeinerten algebraischen Riccatigleichung (2.10) *ist.*

Beweis. Wir führen den allgemeinen Fall auf die Situation in Satz 2.5 ($M = I_n$) zurück. Nach Satz 2.5 hat das LQR-Problem ($M = I_n$) mit der Zustandsgleichung $\dot{x}(t) = M^{-1}Ax(t) + M^{-1}Bu(t)$ die eindeutige Lösung $u_\infty(t) = -R^{-1}B^TM^{-T}X_\infty x(t)$, wobei X_∞ die eindeutige, symmetrisch positiv semidefinite, stabilisierende Lösung der algebraischen Riccatigleichung

$$0 = C^TQC + A^TM^{-T}X_\infty + X_\infty M^{-1}A - X_\infty M^{-1}BR^{-1}B^TM^{-T}X_\infty$$

ist. Wie in Korrolar 2.2 ist dann $\mathcal{T}_M(X_\infty)$ die eindeutige, symmetrisch positiv semidefinite, stabilisierende Lösung der verallgemeinerten algebraischen Riccatigleichung

$$0 = C^TQC + A^T\mathcal{T}_M(X)M + M^T\mathcal{T}_M(X)A - M^T\mathcal{T}_M(X)BR^{-1}B^T\mathcal{T}_M(X)M.$$

Für die zur Steuerung u_∞ zugehörige Zustandsgleichung gilt

$$\begin{aligned}
\dot{x}(t) &= (M^{-1}A - M^{-1}BR^{-1}B^TM^{-T}X_\infty)x(t) \Leftrightarrow \\
M\dot{x}(t) &= A - BR^{-1}B^TM^{-T}X_\infty M^{-1}Mx(t) \Leftrightarrow \\
M\dot{x}(t) &= A - BR^{-1}B^T\mathcal{T}_M(X_\infty)Mx(t).
\end{aligned}$$

Damit erhalten wir für das verallgemeinerte LQR-Problem die eindeutige optimale Steuerung

$$\tilde{u}_\infty(t) = -R^{-1}B^T\tilde{X}_\infty Mx(t),$$

wobei $\tilde{X}_\infty$ die eindeutige, symmetrisch positiv semidefinite stabilisierende Lösung der verallgemeinerten algebraischen Riccatigleichung ist. □

2.3 Newton-Verfahren für die algebraische Riccatigleichung

Um Nullstellen von $\mathcal{R} \in \mathfrak{F}(\mathbb{R}^{n\times n}, \mathbb{R}^{n\times n})$ zu finden, verwenden wir ein Newton-Verfahren und berechnen dazu die Ableitung von $\mathcal{R}$ an der Stelle X. Wir betrachten den verallgemeinerten

Fall.

$$\begin{aligned}
&\lim_{t\to 0}\frac{1}{t}(\mathcal{R}(X+tN)-\mathcal{R}(X))\\
&=\lim_{t\to 0}\frac{1}{t}(A^T(X+tN)M+M^T(X+tN)A-M^T(X+tN)BR^{-1}B^T(X+tN)M)\\
&\qquad-\frac{1}{t}(A^TXM+M^TXA-M^TXBR^{-1}B^TXM)\\
&=(A-BR^{-1}B^TX^T)^TNM+M^TN(A-BR^{-1}B^TX).\\
&\frac{||\mathcal{R}(X+N)-\mathcal{R}(X)-((A-BR^{-1}B^TX^T)^TNM+M^TN(A-BR^{-1}B^TX))||_2}{||N||_2}=\\
&\frac{||MNBR^{-1}B^TNM^T||_2}{||N||_2}\leq\frac{||N||_2^2||M||_2^2||BR^{-1}B^T||_2}{||N||_2}\to 0 \text{ für } N\to 0,
\end{aligned}$$

daher ist $\mathcal{R}'(X)$ durch $N\mapsto(A-BR^{-1}B^TX^TM)^TNM+M^TN(A-BR^{-1}B^TXM)\in\mathfrak{L}(\mathbb{R}^{n\times n},\mathbb{R}^{n\times n})$ gegeben. Da wir eine symmetrische Lösung suchen, erhalten wir folgendes Verfahren.

$$\begin{aligned}
\mathcal{R}|'_{\mathbb{S}^n}(X_k)(X_{k+1}-X_k)&=-\mathcal{R}|_{\mathbb{S}^n}(X_k),\\
(A-BR^{-1}B^TX_kM)^TX_{k+1}M+M^TX_{k+1}(A-BR^{-1}B^TX_kM)&\\
=-C^TQC-M^TX_kBR^{-1}B^TX_kM.&
\end{aligned}\tag{2.13}$$

Satz 2.6 (Newton-Verfahren für algebraische Riccatigleichung, [13, Satz. 4.24], [14, Thm. 4.27])
Sei $BR^{-1}B^T\geq 0$, das Paar $(A,BR^{-1}B^T)$ stabilisierbar und es existiere die eindeutige stabilisierende Lösung $X_\infty\in\mathbb{S}^n_+$ der algebraischen Riccatigleichung. Falls $X_0\in\mathbb{S}^n$ und stabilisierend für das Paar $(A,BR^{-1}B^T)$, d. h. $\Lambda(A-BR^{-1}B^TX_0)\subset\mathbb{C}^-$, dann gelten folgende Aussagen für die durch (2.13) ($M=I_n$) generierte Folge $(X_k)_{k\in\mathbb{N}_0}$.

- $\lambda(A-BR^{-1}B^TX_k)\subset\mathbb{C}^-$ *für alle* $k\in\mathbb{N}_0$.
- $\lim\limits_{k\to\infty}X_k=X_\infty$ *existiert und ist die eindeutige stabilisierende Lösung der algebraischen Riccatigleichung.*
- $X_\infty\leq\ldots\leq X^{k+1}\leq X^k\leq\ldots\leq X_1$.
- $\exists\ \gamma>0$ *mit* $||X_{k+1}-X_\infty||\leq\gamma||X_k-X_\infty||^2$ *für alle* $k\in\mathbb{N}$ (*quadratische Konvergenz*).

Unter geeigneten Voraussetzungen sichert Satz 2.6 die Konvergenz von (2.13) gegen die eindeutige stabilisierende Lösung der algebraischen Riccatigleichung. Falls $\Lambda(A)\subset\mathbb{C}^-$, so ist $X_0=0$ zulässig.

Algorithmus 1 Newton-Verfahren für algebraische Riccatigleichungen [14, Alg. 4.26]

Eingabe: A,Q,R,C,B und Startwert X_0.
Ausgabe: X_∞ löst algebraische Riccatigleichung und $K_\infty=X_\infty BR^{-1}$.

```
K_0 = X_0 B R^{-1}
for k = 1, ... do
    Bestimme Lösung X_k von
    (A - B K_{k-1}^T)^T X_k + X_k (A - B K_{k-1}^T) = -C^T Q C - K_{k-1} R K_{k-1}^T
    K_k = X_k B R^{-1}
end for
```

Algorithmus 2 Newton-Verfahren für verallgemeinerte algebraische Riccatigleichungen [14, Alg. 4.26]

Eingabe: M, A, Q, R, C, B und Startwert X_0.
Ausgabe: X_∞ löst verallgemeinerte algebraische Riccatigleichung und $K_\infty = M^T X_\infty B R^{-1}$.
$K_0 = M^T X_0 B R^{-1}$
for $k = 1, \ldots$ **do**
Bestimme Lösung X_k von
$(A - BK_{k-1}^T)^T X_k M + M^T X_k (A - BK_{k-1}^T) = -C^T Q C - K_{k-1} R {K_{k-1}}^T$
$K_k = M^T X_k B R^{-1}$
end for

Lemma 2.7 (Sherman-Morrison-Woodburry Formel)
Sei $A \in \mathbb{C}^{n\times n}$, $U, V \in \mathbb{C}^{n\times m}$ und $A + UV^T \in \mathrm{GL}(n, \mathbb{C})$. Dann gilt

$$(A + UV^T)^{-1} = A^{-1} - A^{-1}U(I_m + V^T A^{-1} U)^{-1} V^T A^{-1}.$$

Bemerkung 2.3
Nach Lemma 2.7 *können wir lineare Gleichungssystem mit $A + UV^T$ lösen ohne die Matrix $A + UV^T$ explizit zu berechnen. Dies wird später in den Anwendungen wichtig sein. Wir werden dann annehmen, dass $m \ll n$ und n groß ist. Weiterhin ist A dünnbesetzt und U, V im Allgemeinen dichtbesetzt. Diese Bedingungen sind in der Praxis meistens gegeben. Würde man $A + UV^T$ explizit berechnen, führt dies zu hohen Speicherbedarf.*

2.4 Lyapunovgleichung und das ADI-Verfahren

Nach (2.13) sind mehrere lineare Gleichungssysteme der Form

$$FX + XF^T = -GG^T \tag{2.14}$$

zu lösen. Gleichungen dieser Art nennen wir Lyapunovgleichungen. Wir wollen uns im Folgenden mit einem numerischen Verfahren zur Lösung von Lyapunovgleichungen beschäftigen. Wir wollen folgende Annahmen treffen, $F \in \mathbb{R}^{n\times n}$ ist dünnbesetzt, $G \in \mathbb{R}^{n\times m}$ und $m \ll n$. Weiterhin soll $\Lambda(F) \subset \mathbb{C}^-$ gelten. Falls $\Lambda(F) \subset \mathbb{C}^+$, so betrachten wir $-F$.

Satz 2.7 (Existenz und Eindeutigkeit, [19, S. 762], [42, Ch. 12.3 Thm. 2, Thm. 3])
Sei $F \in \mathbb{R}^{n\times n}, \Lambda(F) \subset \mathbb{C}^-$ und $G \in \mathbb{R}^{n\times m}$. Dann besitzt die Lyapunovgleichung

$$FX + XF^T = -GG^T \tag{2.15}$$

eine eindeutige, symmetrisch positiv semidefinite Lösung. Die Lösung ist durch

$$\int_0^\infty e^{Ft} G G^T e^{F^T t} \, \mathrm{d}t \tag{2.16}$$

gegeben. Falls zusätzlich $\mathrm{rank}([F - \lambda I_n, G]) = n$ *für alle $\lambda \in \mathbb{C}$ gilt, so ist diese positiv definit.*

Korollar 2.3
Sei $M \in \mathrm{GL}(n, \mathbb{R}), F \in \mathbb{R}^{n\times n}, \Lambda(F, M) \subset \mathbb{C}^-$ und $G \in \mathbb{R}^{n\times m}$. Dann besitzt die verallgemei-

nerte Lyapunovgleichung

$$FXM^T + MXF^T = -GG^T \tag{2.17}$$

eine eindeutige, symmetrisch positiv semidefinite Lösung. Die Lösung ist durch

$$\int_0^\infty e^{M^{-1}Ft} M^{-1} G G^T M^{-T} e^{(M^{-1}F)^T t} \,\mathrm{d}t \tag{2.18}$$

gegeben. Falls zusätzlich $\operatorname{rank}([F - \lambda M, G]) = n$ *für alle* $\lambda \in \mathbb{C}$ *gilt, so ist diese positiv definit.*

Beweis. Wir wenden Satz 2.7 an.
$FXM^T + MXF^T = -GG^T \Leftrightarrow (M^{-1}F)X + X(M^{-1}F)^T = -(M^{-1}G)(M^{-1}G)^T$.
Da äquivalente Matrizen gleichen Rang haben gilt

$$\operatorname{rank}([M^{-1}F - \lambda I_n, M^{-1}G]) = \operatorname{rank}([F - \lambda M, G]) = n.$$

□

Wachspress fasste aufgrund der Struktur der Lyapunovgleichung diese als ADI-Modellproblem auf [59], wobei das ADI-Verfahren (Alternating Direction Implicit) ursprünglich von Peaceman und Rachford entwickelt wurde.

Bemerkung 2.4
Man beachte, dass (2.15) *äquivalent ist zu* $(H + V)(X) = -GG^T$, *wobei* $H(X) := FX$ *und* $V(X) := XF^T$ *ist.*

Wir fassen im Folgenden den Vektorraum $\mathbb{C}^{n\times n}$ als $\mathbb{C}$-Vektorraum auf und versehen diesen mit $\langle\cdot,\cdot\rangle_F$. Wir erhalten so einen Hilbertraum.

Lemma 2.8
Sei $F \in \mathbb{R}^{n\times n}$ *mit* $\Lambda(F) \subset \mathbb{C}^-$ *und* $H, V \in \mathfrak{L}(\mathbb{C}^{n\times n}, \mathbb{C}^{n\times n})$, *wobei* $H(X) := FX$ *und* $V(X) := XF^T$. *So gilt* $\Lambda(H) = \Lambda(F) = \Lambda(V)$.

Beweis. $\lambda \in \Lambda(H) \Rightarrow \exists X \in \mathbb{C}^{n\times n} \setminus \{0\}\colon H(X) = \lambda X \Rightarrow FX = \lambda X \neq 0 \Rightarrow \exists i \in \{1,\dots,n\}\colon X_{*,i} \neq 0$ und $FX_{*,i} = \lambda X_{*,i} \Rightarrow \lambda \in \Lambda(F)$.
$\lambda \in \Lambda(F) \Rightarrow \exists x \in \mathbb{C}^n \setminus \{0\}\colon Fx = \lambda x$. Mit $X := \begin{bmatrix} x, \dots, x \end{bmatrix} \in \mathbb{C}^{n\times n} \setminus \{0\}$ erhält man $H(X) = FX = \lambda X \neq 0. \Rightarrow \lambda \in \Lambda(H)$.
Mit $\Lambda(F) = \Lambda(F^T)$ erhält man analog $\Lambda(F) = \Lambda(V)$. □

Lemma 2.9
Sei $F \in \mathbb{R}^{n\times n}$ *mit* $\Lambda(F) \subset \mathbb{C}^-$ *und* $H, V \in \mathfrak{L}(\mathbb{C}^{n\times n}, \mathbb{C}^{n\times n})$, *wobei* $H(X) := FX$ *und* $V(X) := XF^T$. *Dann gilt:*

(i) $HV = VH$.
(ii) $(H + p_1 \operatorname{id})^{-1},\ (V + p_2 \operatorname{id})^{-1} \in \mathfrak{L}(\mathbb{C}^{n\times n}, \mathbb{C}^{n\times n})$ *für alle* $p_1, p_2 \in \mathbb{C} \setminus \Lambda(-F)$.
(iii) $(p_1 \operatorname{id} - H)^{-1},\ (p_2 \operatorname{id} - V)^{-1} \in \mathfrak{L}(\mathbb{C}^{n\times n}, \mathbb{C}^{n\times n})$ *für alle* $p_1, p_2 \in \mathbb{C} \setminus \Lambda(F)$.
(iv) Die in (ii) und (iii) aufgeführten Funktionen und (falls existent) die zugehörigen Inversen kommutieren paarweise.

Beweis.

Zu (i) :

$$(HV - VH)(X) = H(XF^T) - V(FX) = FXF^T - FXF^T = 0 \text{ für alle } X \in \mathbb{C}^{n\times n}.$$

Zu (ii), (iii) :

$$(F + pI_n)x = 0 \Leftrightarrow Fx = -px \Rightarrow \ker(F + pI_n) = \{0\} \Rightarrow$$
$$(H + p\,\mathrm{id})^{-1}(X) = (F + pI_n)^{-1}X.$$
Für $V + p\,\mathrm{id},\ p\,\mathrm{id} - H,\ p\,\mathrm{id} - V$ geht man analog vor.

Zu (iv) :

$$p_2H + p_1p_2\,\mathrm{id} - p_1V - HV = p_2H + p_2p_1\,\mathrm{id} - p_1V - VH \Leftrightarrow$$
$$(H + p_1\,\mathrm{id})(p_2\,\mathrm{id} - V) = (p_2\,\mathrm{id} - V)(H + p_1\,\mathrm{id}) \Leftrightarrow$$
$$(p_2\,\mathrm{id} - V)^{-1}(H + p_1\,\mathrm{id}) = (H + p_1\,\mathrm{id})(p_2\,\mathrm{id} - V)^{-1} \Leftrightarrow$$
$$(H + p_1\,\mathrm{id})^{-1}(p_2\,\mathrm{id} - V)^{-1} = (p_2\,\mathrm{id} - V)^{-1}(H + p_1\,\mathrm{id})^{-1}.$$
Analog lassen sich die fehlenden Fälle nachprüfen.

□

Für Fehlerabschätzungen brauchen wir eine Norm auf $\mathfrak{L}(\mathbb{C}^{n\times n}, \mathbb{C}^{n\times n})$. Hierzu wählen wir die durch $\langle\cdot,\cdot\rangle_F$ induzierte Operatornorm. Es sei an dieser Stelle daran erinnert, dass lineare Abbildungen zwischen endlichdimensionalen normierten Vektorräumen ($\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$) stetig sind, jede Norm aufgrund der umgekehrten Dreiecksungleichung stetig (Lipschitz-stetig mit $L = 1$) ist und die Einheitssphäre in endlichdimensionalen normierten Vektorräumen kompakt ist.

$$||\cdot|| : \mathfrak{L}(\mathbb{C}^{n\times n}, \mathbb{C}^{n\times n}) \to \mathbb{R}$$
$$R \mapsto \sup_{X\in\mathbb{C}^{n\times n}\setminus\{0\}} \frac{||R(X)||_F}{||X||_F} = \sup_{||X||_F=1} ||R(X)||_F = \max_{||X||_F=1} ||R(X)||_F. \tag{2.19}$$

Lemma 2.10 (Cayleytransformation und Eigenwerte [49, Kap. 6.2])
Sei $R \in \mathfrak{L}(\mathbb{C}^{n\times n}, \mathbb{C}^{n\times n})$, $p \in \mathbb{C}^-$ und $\Lambda(R) \subset \mathbb{C}^-$. Dann gibt es eine Bijektion zwischen $\Lambda(R) \subset \mathbb{C}^-$ und $\Lambda((\overline{p}\,\mathrm{id} - R)(R + p\,\mathrm{id})^{-1}) \subset \mathfrak{B}_1(0) \subset \mathbb{C}$.

Beweis. Da

$$\det\left(\begin{pmatrix} -1 & \overline{p} \\ 1 & p \end{pmatrix}\right) = -2\Re(p) \neq 0,$$

ist $z \mapsto \frac{-z+\overline{p}}{z+p}$ eine Möbiustransformation (Cayleytransformation). Weiterhin ist aus der Funktionentheorie bekannt, dass $\phi \in \mathfrak{F}(\mathbb{C}^-, \mathfrak{B}_1(0))$ mit $\phi(z) := \frac{-z+\overline{p}}{z+p} = \frac{\overline{p}-z}{z+p}$ eine Bijektion ist und $\phi^{-1}(z) = \frac{pz-\overline{p}}{-z-1} = \frac{\overline{p}-pz}{1+z}$.
Sei $\lambda \in \Lambda(R)$ beliebig und $X \in \mathbb{C}^{n\times n} \setminus \{0\}$ ein zugehöriger Eigenvektor.

$$(\overline{p}\,\mathrm{id} - R)(X) = \overline{p}X - R(X) = (\overline{p} - \lambda)X,$$
$$(R + p\,\mathrm{id})(X) = R(X) + pX = (\lambda + p)X \Leftrightarrow (R + p\,\mathrm{id})^{-1}(X) = \tfrac{1}{\lambda+p}X,$$
$$(\overline{p}\,\mathrm{id} - R)(R + p\,\mathrm{id})^{-1}(X) = \tfrac{\overline{p}-\lambda}{\lambda+p}X = \phi(\lambda)X,$$

d. h. $\phi(\Lambda(R)) \subseteq \Lambda((\overline{p}\,\mathrm{id} - R)(R + p\,\mathrm{id})^{-1})$.
Nun sei $\lambda \in \Lambda((\overline{p}\,\mathrm{id} - R)(R + p\,\mathrm{id})^{-1})$ beliebig und $X \in \mathbb{C}^{n\times n}\setminus\{0\}$ ein zugehöriger Eigenvektor. Wir setzen $Y := (R + p\,\mathrm{id})^{-1}(X) \neq 0$.

$$(\overline{p}\,\mathrm{id} - R)(R + p\,\mathrm{id})^{-1}(X) = \lambda X \Leftrightarrow$$
$$(\overline{p}\,\mathrm{id} - R)(Y) = \lambda(R + p\,\mathrm{id})(Y) \Leftrightarrow$$

$$\begin{aligned}\overline{p}Y - R(Y) &= \lambda R(Y) + \lambda pY \Leftrightarrow \\ R(Y) &= \tfrac{(\overline{p}-p\lambda)}{(1+\lambda)}Y = \phi^{-1}(\lambda)Y,\end{aligned}$$

d. h. $\phi^{-1}(\Lambda((\overline{p}\,\mathrm{id} - R)(R + p\,\mathrm{id})^{-1})) \subseteq \Lambda(R)$.
Nun erhalten wir

$$\begin{aligned}&\phi(\Lambda(R)) \subseteq \Lambda((\overline{p}\,\mathrm{id} - R)(R + p\,\mathrm{id})^{-1}) = \phi(\phi^{-1}(\Lambda((\overline{p}\,\mathrm{id} - R)(R + p\,\mathrm{id})^{-1}))) \subseteq \phi(\Lambda(R)) \Rightarrow \\ &\phi(\Lambda(R)) = \Lambda((\overline{p}\,\mathrm{id} - R)(R + p\,\mathrm{id})^{-1}).\end{aligned}$$

Aufgrund der Eigenschaften von ϕ ist die Aussage gezeigt. □

Bemerkung 2.5
Man kann Lemma 2.10 *auch wie folgt interpretieren. Falls $\lambda, p \in \mathbb{C}^-$ bzw. $\lambda, p \in \mathbb{C}^+$, so wird λ mittels ϕ in die Einheitskugel abgebildet. Falls $\lambda \in \mathbb{C}^+$ und $p \in \mathbb{C}^-$ bzw. $\lambda \in \mathbb{C}^-$ und $p \in \mathbb{C}^+$, so wird λ mittels ϕ nach $\mathbb{C} \setminus \overline{\mathfrak{B}_1(0)}$ abgebildet. Die Wahl $p \in i\mathbb{R}$ ist nicht sinnvoll, da dann ϕ konstant ist.*

Satz 2.8 (Approximationseigenschaft des Spektralradius, [2, S. 82, Satz 4.6])
Sei $(X, ||\cdot||)$ Banachraum und $R \in \mathfrak{L}(X, X)$. Dann kann man zu jedem $\varepsilon > 0$ eine äquivalente Norm $||\cdot||_\varepsilon$ auf X einführen, derart dass R in beiden Normen denselben Spektralradius hat und zusätzlich

$$\rho(R) \leq ||R||_\varepsilon \leq \rho(R) + \varepsilon$$

gilt, wobei $||R||_\varepsilon$ die Operatornorm von R in $(X, ||\cdot||_\varepsilon)$ bezeichne.

Satz 2.9 (Konvergenz des ADI-Verfahrens [58, Kap. 8.6],[4, Thm. 2.2])
Sei $F \in \mathbb{R}^{n\times n}, \Lambda(F) \subset \mathbb{C}^-, G \in \mathbb{R}^{n\times m}$ und $H, V \in \mathfrak{L}(\mathbb{C}^{n\times n}, \mathbb{C}^{n\times n})$, wobei $H(X) := FX$ und $V(X) := XF^T$. Weiterhin sei $\emptyset \neq \mathcal{P} \subset \mathbb{C}^-$ und $|\mathcal{P}| < \infty$. Dann konvergiert das Iterationsverfahren

$$X_0 = 0 \tag{2.20}$$

$$(H + p_{j+1}\,\mathrm{id})(X_{j+\frac{1}{2}}) = (p_{j+1}\,\mathrm{id} - V)(X_j) - GG^T \tag{2.21}$$

$$(V + \overline{p}_{j+1}\,\mathrm{id})(X_{j+1}) = (\overline{p}_{j+1}\,\mathrm{id} - H)(X_{j+\frac{1}{2}}) - GG^T \tag{2.22}$$

gegen die eindeutige Lösung von (2.15). *Die Shiftparameter $p_j \in \mathcal{P}$ sind zyklisch anzuwenden.*

Beweis. Nach Lemma 2.9 existieren $(V + p\,\mathrm{id})^{-1}$ und $(H + p\,\mathrm{id})^{-1}$ für alle $p \in \mathcal{P}$. Wir können (2.21) und (2.22) als Einschrittiteration formulieren

$$\begin{aligned}X_{j+1} &= (V + \overline{p}_{j+1}\,\mathrm{id})^{-1}(\overline{p}_{j+1}\,\mathrm{id} - H)(H + p_{j+1}\,\mathrm{id})^{-1}(p_{j+1}\,\mathrm{id} - V)(X_j) - \\ &\quad (V + \overline{p}_{j+1}\,\mathrm{id})^{-1}[\mathrm{id} + (\overline{p}_{j+1}\,\mathrm{id} - H)(H + p_{j+1}\,\mathrm{id})^{-1}](GG^T) \\ &:= R_{p_{j+1}}(X_j) - g_{p_{j+1}}(GG^T). \qquad (2.23) \\ g_{p_{j+1}} &= (V + \overline{p}_{j+1}\,\mathrm{id})^{-1}[\mathrm{id} + (\overline{p}_{j+1}\,\mathrm{id} - H)(H + p_{j+1}\,\mathrm{id})^{-1}] \\ &= (V + \overline{p}_{j+1}\,\mathrm{id})^{-1}[\mathrm{id} - (H - \overline{p}_{j+1}\,\mathrm{id})(H + p_{j+1}\,\mathrm{id})^{-1}] \\ &= (V + \overline{p}_{j+1}\,\mathrm{id})^{-1}[\mathrm{id} - (H + p_{j+1}\,\mathrm{id} - p_{j+1}\,\mathrm{id} - \overline{p}_{j+1}\,\mathrm{id})(H + p_{j+1}\,\mathrm{id})^{-1}] \\ &= (V + \overline{p}_{j+1}\,\mathrm{id})^{-1}[\mathrm{id} - ((H + p_{j+1}\,\mathrm{id}) - 2\Re(p_{j+1})\,\mathrm{id})(H + p_{j+1}\,\mathrm{id})^{-1}] \\ &= 2\Re(p_{j+1})(V + \overline{p}_{j+1}\,\mathrm{id})^{-1}(H + p_{j+1}\,\mathrm{id})^{-1}. \qquad (2.24)\end{aligned}$$

Offenbar erfüllt die eindeutige Lösung $X_L \in \mathbb{R}^{n\times n}$ von (2.15) die Gleichungen (2.21) und (2.22) und damit ebenso (2.23). Nun betrachten wir den Fehler in der J-ten Iteration

$$\begin{aligned} e_J &:= X_J - X_L = R_{p_J}(X_{J-1}) - R_{p_J}(X_L) \\ &= R_{p_J}(X_{J-1} - X_L) = R_{p_J}(e_{J-1}) \Rightarrow \\ e_J &= W_J(e_0) := (R_{p_J} R_{p_J-1} \cdots R_{p_1})(e_0). \end{aligned}$$

Mit Lemma 2.9 können wir die Faktoren nach den Shiftparametern sortieren

$$\begin{aligned} W_J &= \prod_{p\in\mathcal{P}} (p\,\mathrm{id} - V)^{N_p} (V + \overline{p}\,\mathrm{id})^{-N_p} (\overline{p}\,\mathrm{id} - H)^{N_p} (H + p\,\mathrm{id})^{-N_p}, \\ \sum_{p\in\mathcal{P}} N_p &= J. \end{aligned}$$

Mit Lemma 2.8 und Lemma 2.10 gilt

$$\begin{aligned} s_p &:= \rho((p\,\mathrm{id} - V)(V + \overline{p}\,\mathrm{id})^{-1}) = \rho((\overline{p}\,\mathrm{id} - H)(H + p\,\mathrm{id})^{-1}) < 1 \ \forall p \in \mathcal{P}, \\ \delta_p &:= \tfrac{1-s_p}{2}, \\ s_p + \delta_p &= \tfrac{1+s_p}{2} < \tfrac{1+1}{2} = 1 \ \forall p \in \mathcal{P}, \\ \gamma &:= \max_{p\in\mathcal{P}} s_p + \delta_p < 1. \end{aligned}$$

Satz 2.8 und die Äquivalenz von Normen auf endlichdimensionalen Räumen liefert

$$\begin{aligned} s_p &\le ||(p\,\mathrm{id} - V)(V + \overline{p}\,\mathrm{id})^{-1}||_{\delta_p} \le s_p + \delta_p \le \gamma < 1 \ \forall p \in \mathcal{P}, \\ s_p &\le ||(\overline{p}\,\mathrm{id} - H)(H + p\,\mathrm{id})^{-1}||_{\delta_p} \le s_p + \delta_p \le \gamma < 1 \ \forall p \in \mathcal{P}, \\ ||R|| &\le C_{\delta_p} ||R||_{\delta_p} \ \forall R \in \mathfrak{L}(\mathbb{C}^{n\times n}, \mathbb{C}^{n\times n}) \ \forall p \in \mathcal{P}, \\ C &:= \max_{p\in\mathcal{P}} C_{\delta_p}. \end{aligned}$$

Unter Ausnutzung der Submultiplikativität von $||\cdot||$ und $||\cdot||_{\delta_p}$ können wir den Fehler abschätzen

$$\begin{aligned} ||e_J||_F &= ||(W_J)(e_0)||_F \\ &\le || \prod_{p\in\mathcal{P}} (p\,\mathrm{id} - V)^{N_p} (V + \overline{p}\,\mathrm{id})^{-N_p} (\overline{p}\,\mathrm{id} - H)^{N_p} (H + p\,\mathrm{id})^{-N_p} || ||e_0||_F \\ &\le \prod_{p\in\mathcal{P}} (||(p\,\mathrm{id} - V)^{N_p} (V + \overline{p}\,\mathrm{id})^{-N_p} (\overline{p}\,\mathrm{id} - H)^{N_p} (H + p\,\mathrm{id})^{-N_p}||) ||e_0||_F \\ &\le \prod_{p\in\mathcal{P}} (C_{\delta_p} ||(p\,\mathrm{id} - V)^{N_p} (V + \overline{p}\,\mathrm{id})^{-N_p} (\overline{p}\,\mathrm{id} - H)^{N_p} (H + p\,\mathrm{id})^{-N_p}||_{\delta_p}) ||e_0||_F \\ &\le \prod_{p\in\mathcal{P}} (C ||(p\,\mathrm{id} - V)^{N_p} (V + \overline{p}\,\mathrm{id})^{-N_p} (\overline{p}\,\mathrm{id} - H)^{N_p} (H + p\,\mathrm{id})^{-N_p}||_{\delta_p}) ||e_0||_F \\ &\le C^{|\mathcal{P}|} \prod_{p\in\mathcal{P}} (||(p\,\mathrm{id} - V)^{N_p} (V + \overline{p}\,\mathrm{id})^{-N_p} (\overline{p}\,\mathrm{id} - H)^{N_p} (H + p\,\mathrm{id})^{-N_p}||_{\delta_p}) ||e_0||_F \end{aligned}$$

$$
\begin{aligned}
&\leq C^{|\mathcal{P}|} \prod_{p\in\mathcal{P}} (||(p\,\mathrm{id} - V)(V + \overline{p}\,\mathrm{id})^{-1}||_{\delta_p}^{N_p} ||(\overline{p}\,\mathrm{id} - H)(H + p\,\mathrm{id})^{-1}||_{\delta_p}^{N_p}) ||e_0||_F \\
&\leq C^{|\mathcal{P}|} \prod_{p\in\mathcal{P}} (\gamma^{N_p}\gamma^{N_p}) ||e_0||_F \\
&\leq C^{|\mathcal{P}|}\gamma^{2J} ||e_0||_F
\end{aligned}
$$

Da $C^{|\mathcal{P}|}$ unabhängig von J und $\gamma < 1$ ist die Konvergenz gezeigt. □

Bemerkung 2.6
Im Beweis von Satz 2.9 *haben wir nicht genutzt, dass $X_0 = 0$ gilt.*

$$
\begin{aligned}
X_{j+1} - X_{j+1}^H &= R_{p_{j+1}}(X_j) - g_{p_{j+1}}(GG^T) - (R_{p_{j+1}}(X_j) - g_{p_{j+1}}(GG^T))^H \\
&= R_{p_{j+1}}(X_j) - g_{p_{j+1}}(GG^T) - R_{p_{j+1}}(X_j)^H + g_{p_{j+1}}(GG^T)^H \\
&= R_{p_{j+1}}(X_j) - g_{p_{j+1}}(GG^T) - R_{p_{j+1}}(X_j^H) + g_{p_{j+1}}((GG^T)^H) \\
&= R_{p_{j+1}}(X_j - X_j^H).
\end{aligned}
$$

Mit $X_0 = 0$ gilt $X_j = X_j^H$ für alle $j \in \mathbb{N}_0$.

Bemerkung 2.7
In der Praxis ist man natürlich daran interessiert „gute“ Shiftparameter zu wählen um „schnell“ zu konvergieren. Dies führt auf das Problem, die Shiftparameter derart zu wählen, dass der Spektralradius minimiert wird

$$
\min_{p_1,\ldots,p_J\in\mathbb{C}^-} \max_{\lambda\in\Lambda(F)} \prod_{j=1}^{J} \frac{|p_j - \lambda|^2}{|\overline{p_j} + \lambda|^2}. \tag{2.25}
$$

Schwierigkeit ist hierbei, dass $\Lambda(F)$ apriori meistens unbekannt ist. Hierzu gibt es verschiedene Ansätze und wir verweisen auf [31, 51].

Bisher haben wir im ADI-Verfahren noch nicht ausgenutzt, dass die rechte Seite der Lyapunovgleichung von der Form $-GG^T$, wobei $G \in \mathbb{R}^{n\times m}$ und $m \ll n$. In der Praxis kann man beobachten, dass der numerische Rang der Lösung „schnell“ abfällt, falls die rechte Seite eine Niedrigrangstruktur besitzt. In der Theorie sind ebenso Abschätzungen vorhanden vgl. [5, 52, 57]. Wir wählen daher den Ansatz $X \approx ZZ^H$ und fassen die Resultate aus [15, 19, 46, 51] zusammen. Z nennen wir Niedrigrangfaktor von X.
Betrachten wir nochmal die Einschrittiteration (2.23), (2.24) mit dem Ansatz $X_{j+1} = Z_{j+1}Z_{j+1}^H$.

$$
\begin{aligned}
Z_{j+1}Z_{j+1}^H &= R_{p_{j+1}}(Z_jZ_j^H) - g_{p_{j+1}}(GG^T) \\
&= (V + \overline{p}_{j+1}\,\mathrm{id})^{-1}(\overline{p}_{j+1}\,\mathrm{id} - H)(H + p_{j+1}\,\mathrm{id})^{-1}(p_{j+1}\,\mathrm{id} - V)(Z_jZ_j^H) - \\
&\quad 2\Re(p_{j+1})(V + \overline{p}_{j+1}\,\mathrm{id})^{-1}(H + p_{j+1}\,\mathrm{id})^{-1}(GG^T) \\
&= (F + p_{j+1}I_n)^{-1}(\overline{p}_{j+1}I_n - F)Z_jZ_j^H(p_{j+1}I_n - F^T)(F^T + \overline{p}_{j+1}I_n) - \\
&\quad 2\Re(p_{j+1})(F + p_{j+1}I_n)^{-1}GG^T(F^T + \overline{p}_{j+1}I_n)^{-1} \\
&= ((F + p_{j+1}I_n)^{-1}(F - \overline{p}_{j+1}I_n)Z_j)((F + p_{j+1}I_n)^{-1}(F - \overline{p}_{j+1}I_n)Z_j)^H + \\
&\quad (\sqrt{-2\Re(p_{j+1})}(F + p_{j+1}I_n)^{-1}G)(\sqrt{-2\Re(p_{j+1})}(F + p_{j+1}I_n)^{-1}G)^H.
\end{aligned}
$$

Damit erhalten wir folgendes rekursives Schema

$$\begin{aligned}
Z_1 &= \sqrt{-2\Re(p_1)}(F + p_1 I_n)^{-1} G, \\
Z_j &= [\sqrt{-2\Re(p_j)}(F + p_j I_n)^{-1} G,\ (F + p_j I_n)^{-1}(F - \overline{p_j} I_n) Z_{j-1}], \quad j \geq 2.
\end{aligned}$$

Wir formen ein explizites Schema und richten uns nach [46].

$$\begin{aligned}
S_i :=& (F + p_i I_n)^{-1}, \\
T_i :=& (F - \overline{p} I_n), \\
Z_J :=& [S_J \sqrt{-2\Re(p_J)} G, S_J (T_J S_{J-1}) \sqrt{-2\Re(p_{J-1})} G, \ldots, S_J T_J \cdots S_2 (T_2 S_1) \sqrt{-2\Re(p_1)} G], \\
P_i :=& \frac{\sqrt{-2\Re(p_i)}}{\sqrt{-2\Re(p_{i+1})}} T_{i+1} S_i \\
=& \frac{\sqrt{-2\Re(p_i)}}{\sqrt{-2\Re(p_{i+1})}} (F - \overline{p_{i+1}} I_n)(F + p_i I_n)^{-1} \\
=& \frac{\sqrt{-2\Re(p_i)}}{\sqrt{-2\Re(p_{i+1})}} (F + p_i I_n - p_i I_n - \overline{p_{i+1}} I_n)(F + p_i I_n)^{-1} \\
=& \frac{\sqrt{-2\Re(p_i)}}{\sqrt{-2\Re(p_{i+1})}} (I_n - (p_i + \overline{p_{i+1}})(F + p_i I_n)^{-1}), \\
z_j :=& \sqrt{-2\Re(p_j)} S_j G.
\end{aligned}$$

Nun erhalten wir

$$Z_J = [z_J, P_{J-1} z_J, P_{J-2} P_{J-1} z_J, \ldots, P_1 \cdots P_{J-1} z_J].$$

Jetzt „nummeriert“ man die Shiftparameter beginnend mit J absteigend um und erhält ein ADI-Verfahren in Niedrigrangformulierung für die Lypaunovgleichung.

Algorithmus 3 Low-Rank Cholesky factor ADI iteration (LRCF-ADI), [17, Alg. 1]

Eingabe: $F \in \mathbb{R}^{n \times n}, G \in \mathbb{R}^{n \times m}$ mit $FX + XF^T = -GG^T$, $p_1, \ldots, p_{j_{max}} \in \mathbb{C}^-$
Ausgabe: $Z = Z_{j_{max}} \in \mathbb{C}^{n \times m \cdot j_{max}}$ mit $ZZ^H \approx X$

```
V_1 = sqrt(-2Re(p_1))(F + p_1 I_n)^{-1} G
Z_1 = V_1
for j = 2, ..., j_max do
    V_j = sqrt(Re(p_j)/Re(p_{j-1})) (V_{j-1} - (p_j + conj(p_{j-1}))) (F + p_j I_n)^{-1} V_{j-1}
    Z_j = [Z_{j-1}, V_j]
end for
```

Da komplexe Daten doppelten Speicheraufwand bedeuten, versucht man diese weitestgehend zu vermeiden. Hierzu wählt man $\mathcal{P} = \overline{\mathcal{P}} \subset \mathbb{C}^-$ und fordert, dass zueinander komplex konjugierte Shiftparameter nacheinander auftreten. Dann lassen sich komplexe Shiftparameter effizienter behandeln. Für Details sei auf [17] verwiesen.
Während der Iteration wäre es wünschenswert ein Abbruchkriterium zu haben. Hierzu wählt man das Residuum der Lyapunovgleichung, welches sich ebenso als Niedrigrangfaktor darstellen lässt [16].

Algorithmus 4 Low-Rank Cholesky factor ADI iteration real (LRCF-ADI real) [17, Alg. 2]

Eingabe: $F \in \mathbb{R}^{n\times n}, G \in \mathbb{R}^{n\times m}$ mit $FX + XF^T = -GG^T,\ p_1, \ldots, p_{j_{max}} \in \mathcal{P}$
Ausgabe: $Z = Z_{j_{max}} \in \mathbb{R}^{n\times m\cdot j_{max}}$ mit $ZZ^T \approx X$

for $j = 1, \ldots, j_{max}$ **do**
  if $j = 1$ **then**
    $V_1 = \sqrt{-2\Re(p_1)}(F + p_1 I_n)^{-1} G$
  else
    $\tilde{V} = \sqrt{-2\Re(p_j)}(F + p_1 I_n)^{-1} V_{j-1}$
    $V_j = \sqrt{\frac{\Re(p_j)}{\Re(p_{j-1})}}(V_{j-1} - (p_j + \overline{p_{j-1}})\tilde{V})$
  end if
  if $\Im(p_j) = 0$ **then**
    $V_j = \Re(V_j)$
    $Z_j = [Z_j, V_j]$
  else
    $\beta = 2\frac{\Re(p_j)}{\Im(p_j)}$
    $V_{j+1} = \overline{V_j} + \beta\Im(V_j)$
    $Z_{j+1} = [Z_{j-1}, \sqrt{2}\Re(V_j) + \frac{\beta}{\sqrt{2}}\Im(V_j), \sqrt{\frac{\beta^2}{2} + 2}\Im(V_j)]$
    $j = j + 1$
  end if
end for

Algorithmus 5 Low-Rank Cholesky factor ADI iteration real res. (LRCF-ADI real res.) [16, Alg. 1]

Eingabe: $F \in \mathbb{R}^{n\times n}, G \in \mathbb{R}^{n\times m}$ mit $FX + XF^T = -GG^T,\ p_1, \ldots, p_{j_{max}} \in \mathcal{P},\ 0 < \tau \ll 1$
Ausgabe: $Z = Z_{j_{max}} \in \mathbb{R}^{n\times m\cdot j_{max}}$ mit $ZZ^T \approx X$

$W_0 = G$
$Z_0 = []$
$k = 1$
while $||W_{k-1}^T W_{k-1}||_2 \geq \tau ||B^T B||_2$ **do**
  $V_k = (F + p_k I_n)^{-1} W_{k-1}$
  if $\Im(p_k) = 0$ **then**
    $W_k = W_{k-1} - 2p_k V_k$
    $Z_k = [Z_{k-1}, \sqrt{-2p_k} V_k]$
  else
    $\gamma_k = 2\sqrt{-\Re(p_k)}$
    $\delta_k = \frac{\Re(p_k)}{\Im(p_k)}$
    $W_{k+1} = W_{k-1} + \gamma_k^2(\Re(V_k) + \delta_k\Im(V_k))$
    $Z_{k+1} = [Z_{k-1}, \gamma_k(\Re(V_k) + \delta_k\Im(V_k)), \gamma_k\sqrt{\delta_k^2 + 1}\Im(V_k)]$
  end if
end while

In der Praxis trifft man häufig auf verallgemeinerte Lyapunovgleichungen

$$FXM^T + MXF^T = -GG^T, \tag{2.26}$$

wobei $M \in \mathrm{GL}(n, \mathbb{R})$ und $\Lambda(F, M) \subset \mathbb{C}^-$. Gleichung 2.26 ist offenbar äquivalent zu

$$(M^{-1}F)X + X(M^{-1}F)^T = -(M^{-1}G)(M^{-1}G)^T.$$

Die Inverse von M kann man durch implizite Behandlung vermeiden.

$$\begin{aligned}
V_1 &= \sqrt{-2\Re(p_1)}(M^{-1}F + p_1 I_n)^{-1}M^{-1}G \\
&= \sqrt{-2\Re(p_1)}(F + p_1 M)^{-1}G, \\
V_j &= \sqrt{\frac{\Re(p_j)}{\Re(p_{j-1})}}(V_{j-1} - (p_j - \overline{p_{j-1}}))(M^{-1}F + p_j I_n)^{-1}V_{j-1} \\
&= \sqrt{\frac{\Re(p_j)}{\Re(p_{j-1})}}(V_{j-1} - (p_j - \overline{p_{j-1}}))(F + p_j M)^{-1}MV_{j-1},\ j \geq 2.
\end{aligned}$$

Algorithmus 6 Generalized Low-Rank Cholesky factor ADI iteration real res. (GLRCF-ADI real res.) [16, Alg. 1]

Eingabe: $M \in \mathbb{R}^{n\times n}, F \in \mathbb{R}^{n\times n}, G \in \mathbb{R}^{n\times m}$ mit $FXM^T + MXF^T = -GG^T,\ p_1, \ldots, p_{j_{max}} \in \mathcal{P},\ 0 < \tau \ll 1$
Ausgabe: $Z = Z_{j_{max}} \in \mathbb{R}^{n\times m\cdot j_{max}}$ mit $ZZ^T \approx X$

```
W_0 = G
Z_0 = []
k = 1
while ||W_{k-1}^T W_{k-1}||_2 ≥ τ||B^T B||_2 do
  V_k = (F + p_k M)^{-1} W_{k-1}
  if ℑ(p_k) = 0 then
    W_k = W_{k-1} − 2p_k M V_k
    Z_k = [Z_{k-1}, √(−2p_k) V_k]
  else
    γ_k = 2√(−ℜ(p_k))
    δ_k = ℜ(p_k)/ℑ(p_k)
    W_{k+1} = W_{k-1} + γ_k^2 M(ℜ(V_k) + δ_k ℑ(V_k))
    Z_{k+1} = [Z_{k-1}, γ_k(ℜ(V_k) + δ_k ℑ(V_k)), γ_k √(δ_k^2 + 1) ℑ(V_k)]
  end if
  k = k + 1
end while
```

3 Stationäre Navier-Stokes Gleichungen

In diesem Abschnitt wollen wir uns mit den stationären Navier-Stokes Gleichungen für inkompressible Fluide im zweidimensionalen Fall beschäftigen. Hierzu ist es notwendig einige Grundlagen aus der linearen Funktionalanalysis zu wiederholen. Da wir keine Aussagen beweisen, wollen wir erläutern, warum die vorgestellten Konzepte wichtig sind und in welchem Zusammenhang die Begriffe stehen.
Ein nennenswerter Punkt ist beispielsweise die Regularität des Rechengebietes Ω. In der Praxis kann Ω „Ecken“ haben und die Forderung nach glatten Rändern $\partial\Omega$ ist sehr restriktiv. Man stelle sich einfach eine Triangulierung des Rechengebietes Ω vor. Dies wirft sofort die Frage auf, in welchem Sinne ein Einheitsnormalenfeld für Gebiete mit „Ecken“ existieren kann. Wir stellen hierzu die Definition des Lipschitz-Randes vor. Die Definition des Lipschitz-Randes ist etwas umfangreicher und wir wollen daher vorab einige informelle Erklärungen geben. Lipschitz-Rand heißt anschaulich, dass es zu $\Omega \cup \partial\Omega$ eine endliche Überdeckung von offenen Mengen U_i derart gibt, dass man den Rand lokal (in U_i) mit Hilfe einer Lipschitz-stetigen Funktion parametrisieren kann. Zusätzlich fordert man, dass Ω lokal stets nur auf einer Seite des Randes liegt.
„Aufgeschlitzte“ Mengen, wie z. B. $\mathbb{R}^2 \supset \mathfrak{B}_r(0) \setminus (\mathbb{R} \times \{0\})$, können wir nicht dazunehmen. Hier liegt Ω nicht auf einer Seite des Randes, man braucht diesen „Platz“ jedoch.
Den Fall, dass Ω „Ecken“ hat, können wir fast abdecken, denn $x \mapsto m|x| \in \mathfrak{F}((-1,1),\mathbb{R})$ ist Lipschitz-stetig für jedes $m \in \mathbb{R}$ ($L = |m|$).
Gebiete mit „wurzelartig“ zulaufenden Ecken, können wir leider auch nicht dazunehmen, denn $x \mapsto \sqrt{|x|} \in \mathfrak{F}((-1,1),\mathbb{R})$ ist zwar stetig, aber L wird im Nullpunkt „unendlich“ groß.
Naiv könnte man jetzt denken, dass man einfach Stetigkeit anstatt Lipschitz-stetig fordern kann. Der Preis dieses Regularitätsverlustes ist aber enorm. Wir zitieren den Satz von Rademacher in diesem Abschnitt zwar nicht, wollen ihn aber an dieser Stelle erwähnen. Jede Lipschitz-stetige Funktion ist fast überall (im Lebesgueschen Sinne) klassisch differenzierbar. Es sei jetzt daran erinnert, dass es stetige Funktionen gibt, die nirgends klassisch differenzierbar sind! Man denke an die sägezahnförmige Weierstraß-Funktion.
Da $\partial\Omega$ lokal mit Hilfe von Lipschitz-stetigen Funktionen parametrisiert ist, können wir den Satz von Rademacher anwenden, um die Existenz eines Einheitsnormalenfeldes fast überall auf $\partial\Omega$ zu sichern.
Die offene Überdeckung von Ω ist im Allgemeinen nicht disjunkt und um Eigenschaften von Funktionen auf diesen Mengen zu beweisen, konstruiert man zu der offenen Überdeckung häufig eine Zerlegung der Eins von unendlich glatten Funktionen mit kompakten Träger. Meistens beweist man Aussagen auf Gebieten mit Lipschitz-Rand, in dem man mit Hilfe einer Zerlegung der Eins die Funktion „lokalisiert“, „abschneidet“ und „glättet“. Zum „Glätten“ braucht man aber lokal am Rand etwas „Platz“, dieser steht bei „aufgeschlitzten“ Mengen nicht zur Verfügung. Die Lipschitzkonstante sichert auch, dass solche „Glättungsprozesse“ am Rand nicht vollkommen „auseinanderbrechen“. Man vergleiche hierzu die Techniken in [1, A6.7 Lemma, A6.8 Satz, A6.10 Lemma].
Ein zentrales Konzept der Funktionalanalysis ist es, Funktionen nicht als solche, sondern besser als Elemente eines Vektorraums aufzufassen. Um die Menge der Lebesgue-integrierbaren Funktionen normieren zu können, bildet man den Faktorraum und „teilt“ die Menge der Funktionen „raus“, die fast überall 0 sind. Diese Konstruktion ist notwendig, um aus der

Halbnorm eine Norm „zu machen“. Dadurch erhalten wir einen Banachraum, müssen aber mit Äquivalenzklassen arbeiten und viele Eigenschaften gelten dann nur noch fast überall in Ω.
Dies führt direkt auf ein Problem, denn partielle Differentialgleichungen komplettiert man häufig mit Randbedingungen. Da der Rand $\partial\Omega$ eine Nullmenge ist, ist es erstmal nicht klar, wie man „sinnvoll“ Randbedingungen fordern soll. Hierzu stellen wir den Spursatz vor. Für Funktionen, die auf $\overline{\Omega}$ stetig sind, ist es „sinnvoll“ von Randwerten zu sprechen. Der Spursatz liefert uns die Existenz eines eindeutigen linearen stetigen Operators S, der auf einer geeigneten dichten Menge „sinnvolle“ Randwerte liefert. Lineare stetige Operatoren lassen sich, von einer dichten Menge ausgehend, eindeutig mit Hilfe des Grenzwertes fortsetzen. Wir haben damit ein Hilfsmittel zur Verfügung um Randwertaufgaben zu stellen. Die Anwendung des Spuroperators kostet uns aber eine bzw. „eine halbe“ schwache Ableitungsordnung.
Der Begriff der schwachen Ableitung ist ein Ableitungsbegriff im Sinne des Gaußsche Integralsatzes.
Uns stehen damit ein passender Regularitätsbegriff für $\partial\Omega$, Randwerte, Einheitsnormalenfeld und Ableitungen in einem schwachen Sinne zur Verfügung. Wir möchten gerne ein schwaches Analogon zum klassischen Satz von Gauß benutzen, einerseits um partiell integrieren zu dürfen und andererseits um den Begriff der Massenerhaltung, der bei den Navier-Stokes Gleichungen zentral ist, richtig zu fassen. Wir stellen den schwachen Satz von Gauß vor. Dieser wird insbesondere benötigt um schwache Formulierungen von partiellen Differentialgeichungen herzuleiten.
Zur Motivation der schwachen Formulierung: Man stelle sich einen endlichdimensionalen Hilbertraum V vor. Sei $u \in V$ beliebig. Falls $(u, v)_V = 0 \; \forall v \in V$ gilt, dann ist u bereits 0. Das liegt daran, dass $(\cdot, \cdot)_V$ positiv definit ist.
Wir hätten u nicht zwangsweise „gegen“ alle Vektoren aus V „testen“ müssen. Sei $(v_1, \ldots, v_n)$ eine Basis von V. Falls $(u, v_i)_V = 0$ für $i = 1, \ldots, n$ gilt, dann folgt ebenfalls $u = 0$. Wir bemerken zwei Dinge. Wir können einen Vektor über dessen Wirkung auf einer Testmenge charakterisieren und die Testmenge sollte dazu eine gewisse „Qualität“ (Basis) vorweisen.
Lösungen von partiellen Differentialgleichungen sind Funktionen. Funktionenräume sind häufig unendlichdimensional. Sei V ein unendlichdimensionaler Hilbertraum. Falls $(u, v)_V = 0 \; \forall v \in V$ folgt ebenso $u = 0$. Es würde nun auch ausreichen als Testmenge eine Basis von V zu wählen. Das hat aber einen entscheidenden Nachteil. In unendlichdimensionalen, vollständigen, normierten Räumen existieren keine abzählbaren Hamelbasen. Die Hamelbasen werden zwangsweise überabzählbar. „Schuld“ daran ist der Satz von Baire, aus dem man diese Aussage folgern kann. Unter Hamelbasen verstehen wir Basen im Sinne der linearen Algebra, d. h. fast alle (alle bis auf endlich viele) Koeffizienten in der Basisdarstellung eines Vektors sind 0. In der Theorie macht dieser Umstand erstmal keine Probleme und wir fordern meistens einfach $(u, v)_V = 0 \; \forall v \in V$.
In der Praxis, z. B. bei einer Finiten-Elemente Diskretisierung, wählt man endlichdimensionale Teilräume von V um die Lösung u mit Hilfe ihrer „Wirkung“ zu charakterisieren und bezüglich einer endlichen Basis des Teilraumes zu approximieren. Es wäre daher wünschenswert, wenn bereits eine abzählbare Testmenge für die schwache Formulierung ausreichen würde. Stetigkeit von $(\cdot, \cdot)_V$ und eine abzählbar dichte Teilmenge von V als Testmenge sind ausreichend um eine Lösung u mit Hilfe der schwachen Formulierung zu „charakterisieren“. Der Begriff separabel ist in diesem Zusammenhang wesentlich. Kennt man eine (abzählbar) dichte Teilmenge mit guten Regularitätseigenschaften, dann lassen sich in Verbindung mit einem Stetigkeitsargument Beweise meistens einfacher führen, da man sich auf die dichte Teilmenge beschränken kann um die gewünschte Aussage zu beweisen.
Ein anderer Aspekt der schwachen Formulierung ist, dass man nach partieller Integration im Allgemeinen nur noch geringere Regularitätsforderungen an die Lösung stellen muss. Findet

man nun eine Lösung, die der schwache Formulierung genügt und hinreichend regulär ist, dann hat man meistens eine Lösung im klassischen Sinne.
Sind diese Grundlagen gelegt, formulieren wir im zweiten Teil des Abschnittes die Existenz- und Eindeutigkeitssätze für die stationären Navier-Stokes Gleichungen im Falle homogener und inhomogener Dirichletranddaten.
In numerischen Simulationen ist man gezwungen sich auf ein beschränktes Gebiet Ω zurückzuziehen. Ein Hilfsmittel ist hierbei eine „natürliche" Ausflussbedingung auf einem Teil des Rand vorzugeben. Wir erhalten damit ein Problem mit gemischten Randbedingungen. Leider gibt es für diesen Fall aufgrund fehlender a priori Abschätzungen keine Existenztheorie. Wir werden die theoretische Schwierigkeit skizzieren.

3.1 Grundlagen aus der linearen Funktionalanalysis

Definition 3.1 (Räume stetiger Funktionen)
Sei $\Omega \subseteq \mathbb{R}^n$ offen und $k \in \mathbb{N}$. Wir definieren

$$
\begin{aligned}
C^0(\Omega) := C(\Omega) &:= \{f \in \mathfrak{F}(\Omega, \mathbb{R}) \mid f \text{ ist stetig auf } \Omega\},\\
C^0(\overline{\Omega}) := C(\overline{\Omega}) &:= \{f \in \mathfrak{F}(\Omega, \mathbb{R}) \mid f \text{ ist stetig auf } \overline{\Omega} \text{ fortsetzbar}\},\\
C^k(\Omega) &:= \{f \in \mathfrak{F}(\Omega, \mathbb{R}) \mid \text{für } |\alpha| \leq k \ D^\alpha f \in C(\Omega)\},\\
C^k(\overline{\Omega}) &:= \{f \in \mathfrak{F}(\Omega, \mathbb{R}) \mid f \in C^k(\Omega) \text{ und für } |\alpha| \leq k \text{ ist } D^\alpha f \in C(\overline{\Omega})\},\\
C^\infty(\Omega) &:= \bigcap_{k=0}^{\infty} C^k(\Omega),\\
\operatorname{supp}(f) &:= \overline{\{x \in \Omega \mid f(x) \neq 0\}},\\
C_0^k(\Omega) &:= \{f \in C^k(\Omega) \mid \operatorname{supp}(f) \subseteq\subseteq \Omega\},\\
C_0^\infty(\Omega) &:= \{f \in C^\infty(\Omega) \mid \operatorname{supp}(f) \subseteq\subseteq \Omega\},\\
||f||_{C^k(\Omega)} &:= \sum_{0 \leq |\alpha| \leq k} \sup_{x \in \Omega} |D^\alpha f(x)|.
\end{aligned}
$$

Definition 3.2 (dicht, [30, Def. 1.15])
Sei X ein metrischer Raum. $M \subseteq X$ heißt dicht in X falls $\overline{M} = X$, d. h.

$$\forall x \in X \ \forall \varepsilon > 0 \ \exists y \in M : d_X(x, y) < \varepsilon.$$

Per Definition ist X dicht in X. „Kleinere" dichte Teilmengen sind meistens interessanter.

Definition 3.3 (separabel, [30, Def. 1.15])
Sei X ein metrischer Raum. Falls X eine abzählbar dichte Teilmenge besitzt, nennen wir X separabel.

Lemma 3.1 (Charakterisierung stetiger linearer Abbildungen, [30, Lemma 2.8])
Seien X, Y normierte Räume und $f \in \mathfrak{F}(X, Y)$ linear. Dann sind äquivalent:

- *f ist stetig.*
- *f ist im Nullpunkt stetig.*
- *Der Ausdruck $||f||_{\mathfrak{L}(X,Y)} := \sup_{x \in X \setminus \{0\}} \frac{||f(x)||_Y}{||x||_X}$ ist beschränkt.*

Lemma 3.2 ([30, Lemma 2.9])
Seien X, Y, Z normierte Räume und $f \in \mathfrak{L}(X,Y), g \in \mathfrak{L}(Y,Z)$. Dann gilt:

- $||f||_{\mathfrak{L}(X,Y)} = \sup_{x \in X, ||x||_X = 1} ||f(x)||_Y$.
- $||f(x)||_Y \leq ||f||_{\mathfrak{L}(X,Y)} ||x||_X$.
- $||gf||_{\mathfrak{L}(X,Z)} \leq ||g||_{\mathfrak{L}(Y,Z)} ||f||_{\mathfrak{L}(X,Y)}$.

Satz 3.1 ([30, Satz 2.10])
Seien X, Y normierte Räume. Dann ist $\mathfrak{L}(X,Y)$ mit $||\cdot||_{\mathfrak{L}(X,Y)}$ ein normierter Raum. Ist Y ein Banachraum, dann ist $\mathfrak{L}(X,Y)$ mit $||\cdot||_{\mathfrak{L}(X,Y)}$ ein Banachraum.

Definition 3.4 (Dualraum, [30, Def. 2.11])
Sei X ein normierter Raum über dem Körper $\mathbb{R}$. Wir nennen $X^* := \mathfrak{L}(X, \mathbb{R})$ mit $||\cdot||_{X^*} := ||\cdot||_{\mathfrak{L}(X,\mathbb{R})}$ den Dualraum von X. $f \in X^*$ nennen wir stetiges lineares Funktional.

Definition 3.5 (kompakte Abbildungen, [30, Def. 2.15])
Seien X, Y Banachräume. Wir nennen $f \in \mathfrak{F}(X,Y)$ kompakt, falls gilt:

$$M \subseteq X \text{ und } M \text{ ist beschränkt} \Rightarrow f(M) \subseteq\subseteq Y.$$

Definition 3.6 (duale Paarung, [30, Kap. 3.5 S. 51])
Sei X ein normierter Raum. Die Bilinearform $\langle \cdot, \cdot \rangle_{X^* \times X} \in \mathfrak{F}(X^* \times X, \mathbb{R})$ mit $\langle f, x \rangle_{X^* \times X} := f(x)$ nennen wir Dualitätsabbildung oder duale Paarung.

Sei X ein Hilbertraum und $x \in X$ beliebig, dann ist $(\cdot, x)_X \in X^*$. Wir fragen uns nach der Umkehrung, gibt es zu jedem $f \in X^*$ ein eindeutig bestimmtes $x \in X$ mit $f = (\cdot, x)$? Falls ja dann brauchen wir im Hilbertraum nicht „großartig“ zwischen Raum und Dualraum unterscheiden.

Satz 3.2 (Rieszscher Darstellungssatz, [30, Kap. 2 Satz 2.25])
Sei X ein Hilbertraum. Dann gibt es zu jedem $f \in X^*$ ein eindeutiges $x \in X$ mit

$$f(y) = (y, x)_X \ \forall y \in X$$

und $||x||_X = ||f||_{X^*}$. Die zugehörige Abbildung $j \in \mathfrak{F}(X, X^*)$ mit $x \mapsto (\cdot, x)_X$ ist eine bijektive, konjugiert-lineare Isometrie.

Lemma 3.3 ([25, Cor. II.3.8])
Seien V, H Hilberträume, sodass V dicht in H. Dann ist auch H^* dicht in V^*.

Lemma 3.4 ([1, 4.17 Folgerung])
Sei X ein normierter Raum und $x \in X$.

- *Ist $x \neq 0$, so gibt es ein $f \in X^*$ mit $||f||_{X^*} = 1$ und $f(x) = ||x||_X$.*
- *Ist $f(x) = 0 \ \forall f \in X^*$, so folgt $x = 0$.*

Definition 3.7 (reflexiv, [30, Def. 3.29])
Sei X ein Banachraum. Ist die Abbildung $i \in \mathfrak{F}(X, X^{**})$ mit $i(x) := \langle \cdot, x \rangle_{X^* \times X}$. surjektiv, so heißt X reflexiv.

Bemerkung 3.1
Wir fassen einige Eigenschaften von i zusammen.

- *i ist linear.*
- *i ist stetig, denn*
  $|i(x)(g)| = |\langle g, x\rangle_{X^*\times X}| = |g(x)| \leq ||g||_{X^*}||x||_X \ \forall g \in X^* \Rightarrow ||i(x)||_{X^{**}} \leq ||x||_X$
- *i ist eine Isometrie.*
  Sei $x \in X \setminus \{0\}$ beliebig. Dann existiert nach Lemma 3.4 *$f \in X^*$ mit $||f||_{X^*} = 1$ und $f(x) = ||x||_X$. Nun ist $||i(x)||_{X^{**}} = \sup_{||g||_{X^*}=1} |i(x)(g)| \geq |i(x)(f)| = |f(x)| = ||x||_X$.*
- *i ist injektiv, da i eine Isometrie ist.*
- *i ist im Allgemeinen nicht surjektiv.*

Man beachte X ist reflexiv $\Rightarrow$ X ist isometrisch isomorph zu X^{**}. Die Umkehrung ist im Allgemeinen falsch.

Definition 3.8 (Einbettung, [30, Def. 2.16])
Seien X, Y Banachräume. X heißt stetige Einbettung in Y, falls X ein Unterraum von Y ist und die Identität $\mathrm{id} \in \mathfrak{F}(X, Y)$ *stetig ist. Es gilt dann die Abschätzung*

$$||f||_Y \leq C||f||_X \ \forall f \in X.$$

Wir schreiben dann $X \hookrightarrow Y$.
Die Einbettung $X \hookrightarrow Y$ heißt kompakt, wenn id *eine kompakte lineare Abbildung ist.*
Die Einbettung $X \hookrightarrow Y$ heißt dicht, wenn $\mathrm{id}(X)$ *dicht in Y ist.*

Bemerkung 3.2 (Faktorraum, [30, Kap. 4.2])
Mit „messbar“ meinen wir stets messbar im Sinne von Lebesgue. Die auftretenden Integrale sind stets als Lebesgue-Integrale aufzufassen und mit $\mu(\cdot)$ bezeichnen wir das Lebesgue-Maß. Im folgenden identifizieren wir zwei Funktionen, falls sie außerhalb einer Nullmenge übereinstimmen. Formal geschieht dies durch die Äquivalenzrelation

$$f \sim g :\Leftrightarrow f = g \text{ fast überall (f.ü.) in } \Omega.$$

Definition 3.9 (L^p-Raum, [30, Def. 4.15])
Sei $\Omega \subseteq \mathbb{R}^n$ messbar und $1 \leq p \leq \infty$.

$$L^p(\Omega) := \{f \in \mathfrak{F}(\Omega, \mathbb{R}) \mid f \text{ ist messbar und } ||f||_{L^p(\Omega)} < \infty\} \text{ mit}$$

$$||f||_{L^p(\Omega)} := \begin{cases} (\int_\Omega |f(x)|^p \,\mathrm{d}x)^{\frac{1}{p}} & \text{falls } p < \infty \\ \inf_{\mu(N)=0} \sup_{x\in\Omega\setminus N} |f(x)| & \text{falls } p = \infty \end{cases}$$

nennen wir L^p-Raum.

Eine Verallgemeinerung der Ungleichung von Cauchy-Schwarz liefert folgendes Lemma.

Lemma 3.5 (Hölderungleichung, [30, Lemma 4.16])
Sei $1 < p, q < \infty$ mit $\frac{1}{p} + \frac{1}{q} = 1$. Falls $f \in L^p(\Omega)$ und $g \in L^q(\Omega)$ so ist $fg \in L^1(\Omega)$ und

$$||fg||_{L^1(\Omega)} \leq ||f||_{L^p(\Omega)}||g||_{L^q(\Omega)}.$$

Satz 3.3 (L^p-Räume vollständig, [30, Satz 4.17])
$L^p(\Omega)$ ist mit $||\cdot||_{L^p(\Omega)}$ ein Banachraum.

Korollar 3.1 (L^2 als Hilbertraum, [30, Korollar 4.18])
$L^2(\Omega)$ ist mit $(\cdot,\cdot)_{L^2(\Omega)}$ ein Hilbertraum, wobei

$$(f,g)_{L^2(\Omega)} := \int_\Omega fg \,\mathrm{d}x.$$

Satz 3.4 ([30, Satz 4.19])
Sei $\mu(\Omega) < \infty$ und $1 \leq p \leq q \leq \infty$, so gilt $L^q(\Omega) \hookrightarrow L^p(\Omega)$ mit

$$||f||_{L^p(\Omega)} \leq \mu(\Omega)^{\frac{1}{p}-\frac{1}{q}} ||f||_{L^q(\Omega)} \ \forall f \in L^q(\Omega).$$

Die Einbettung aus Satz 3.4 ist für Mengen unbeschränkten Maßes im Allgemeinen falsch.

Satz 3.5 (Approximationseigenschaft der L^p-Räume,[30, Satz 4.23])
$C_0^\infty(\Omega)$ ist dicht in $L^p(\Omega)$ für $1 \leq p < \infty$.

Satz 3.6 (L^p-Räume separabel, [30, Satz 4.20])
$L^p(\Omega)$ ist separabel für $1 \leq p < \infty$.

Nicht jeder L^p-Raum ist ein Hilbertraum. Können wir trotzdem den Dualraum explizit charakterisieren?

Satz 3.7 (L^p-Räume, [30, Satz 4.30])
Sei $1 \leq p < \infty$ und $g \in L^p(\Omega)^*$. Dann gibt es genau ein $f \in L^q(\Omega)$, wobei $\frac{1}{p} + \frac{1}{q} = 1$. Es gilt

$$g(h) = \int_\Omega fh \,\mathrm{d}x \ \forall h \in L^p(\Omega).$$

Weiter gilt $||f||_{L^q(\Omega)} = ||g||_{L^p(\Omega)^*}$.

Definition 3.10 (lokal integrierbare L^p-Räume, [30, Def. 4.15])
Sei $1 \leq p < \infty$. $L^p_{loc}(\Omega) := \{f \in \mathfrak{F}(\Omega,\mathbb{R}) \mid f|_K \in L^p(K) \ \forall K \subseteq\subseteq \Omega \ K$ ist offen$\}$.

Satz 3.8 (Fundamentallemma der Variationsrechnung, [30, Satz 5.1])
Sei $f \in L^1_{loc}(\Omega)$ mit

$$\int_\Omega f\phi \,\mathrm{d}x \geq 0 \ \forall \phi \in C_0^\infty(\Omega) \text{ mit } \phi \geq 0.$$

Dann ist $f \geq 0$ f.ü. in Ω.

Definition 3.11 (schwache Ableitung, [30, Def. 5.3])
Eine Funktion $f \in L^1_{loc}(\Omega)$ besitzt eine α-te schwache Ableitung in Ω, wenn es eine Funktion $f_\alpha \in L^1_{loc}(\Omega)$ gibt mit

$$\int_\Omega fD^\alpha\phi \,\mathrm{d}x = (-1)^{|\alpha|} \int_\Omega f_\alpha\phi \,\mathrm{d}x \ \forall \phi \in C_0^\infty(\Omega).$$

Die schwache Ableitung ist im Sinne des Gaußschen Integralsatzes definiert. Da die Testfunktionen kompakten Träger in Ω haben entfällt der Randterm einfach.
Wir schreiben kurz $D^\alpha f := f_\alpha$. Folgendes Lemma rechtfertigt die Schreibweise.

Lemma 3.6 (schwacher und klassischer Ableitungsbegriff, [30, Lemma 5.4])
Die schwache Ableitung ist eindeutig, sofern sie existiert. Wenn eine Funktion klassisch

differenzierbar ist, so ist sie auch schwach differenzierbar und beide Ableitungen stimmen fast überall überein.

Fordern wir für $f \in L^p(\Omega)$ zusätzlich, dass alle schwachen Ableitung bis zur Ordnung $|\alpha|$ existieren und in $L^p(\Omega)$ liegen, erhalten wir einen Raum höherer Regularität.

Definition 3.12 (Sobolev-Raum, [30, Definition 5.9])
Sei $1 \leq p \leq \infty$ und $m \in \mathbb{N}_0$. $W^{m,p}(\Omega)$ mit $||\cdot||_{W^{m,p}(\Omega)}$ nennen wir Sobolev-Raum, wobei

$$\begin{aligned} W^{m,p}(\Omega) &:= \{f \in L^p(\Omega) \mid D^\alpha f \in L^p(\Omega)\ \forall |\alpha| \leq m\}, \\ ||f||_{W^{m,p}(\Omega)} &:= (\sum_{|\alpha|\leq m} ||D^\alpha f||^p_{L^p(\Omega)})^{\frac{1}{p}}, \\ |f|_{W^{m,p}(\Omega)} &:= (\sum_{|\alpha|= m} ||D^\alpha f||^p_{L^p(\Omega)})^{\frac{1}{p}}. \end{aligned}$$

$|\cdot|_{W^{m,p}(\Omega)}$ ist eine Halbnorm. Die Definitheit ist im Allgemeinen nicht erfüllt.

Satz 3.9 ($W^{m,p}$-Räume vollständig, [30, Satz 5.10])
$W^{m,p}(\Omega)$ ist mit $||\cdot||_{W^{m,p}(\Omega)}$ ein Banachraum.

Korollar 3.2 ($W^{m,2}$-Räume als Hilberträume, [30, Kor. 5.11])
$W^{m,2}(\Omega)$ ist mit $(\cdot,\cdot)_{W^{m,2}(\Omega)}$ ein Hilbertraum, wobei

$$(f,g)_{W^{m,2}(\Omega)} := \sum_{|\alpha|\leq m} \int_\Omega D^\alpha f D^\alpha g \,\mathrm{d}x = \sum_{|\alpha|\leq m} (D^\alpha f, D^\alpha g)_{L^2(\Omega)}.$$

Satz 3.10 (Meyers und Serrin, [30, Satz 5.16])
$C^\infty(\Omega) \cap W^{m,p}(\Omega)$ ist dicht in $W^{m,p}(\Omega)$ für $1 \leq p < \infty$.

Satz 3.11 ($W^{m,p}$-Räume separabel)
$W^{m,p}(\Omega)$ ist separabel für $1 \leq p < \infty$.

Definition 3.13 (Lipschitz-Rand, [1, A 6.2])
Sei $\Omega \subset \mathbb{R}^n$ offen und beschränkt. Wir sagen Ω hat Lipschitz-Rand, falls sich $\partial\Omega$ durch endlich viele offene Mengen $U_1, \ldots, U_l$ überdecken lässt, so dass $\partial\Omega \cap U_j$ für $j = 1, \ldots, l$ der Graph einer Lipschitz-stetigen Funktion ist und $\Omega \cap U_j$ auf jeweils einer Seite dieses Graphen liegt. Damit ist folgendes gemeint:
Es gibt ein $l \in \mathbb{N}$ und für jedes $j = 1, \ldots, l$ ein euklidisches Koordinatensystem $e^j_1, \ldots, e^j_n \in \mathbb{R}^n$, einen Referenzpunkt $y^j \in \mathbb{R}^{n-1}$, Zahlen $r^j, h^j > 0$ und eine Lipschitz-stetige Funktion $g^j : \mathbb{R}^{n-1} \to \mathbb{R}$, so dass mit der Bezeichnung:

$$x^j_{,n} := (x^j_1, \ldots, x^j_{n-1}),\ \text{wenn } x = \sum_{i=1}^n x^j_i e^j_i$$

gilt

$$U_j = \{x \in \mathbb{R}^n \mid\ ||x^j_{,n} - y^j||_2 < r^j \text{ und } |x^j_n - g^j(x^j_{,n})| < h^j\}$$

und für $x \in U_j$

$$\begin{aligned} x^j_n = g^j(x^j_{,n}) &\Rightarrow x \in \partial\Omega \\ 0 < x^j_n - g^j(x^j_{,n}) < h^j &\Rightarrow x \in \Omega \\ 0 > x^j_n - g^j(x^j_{,n}) > -h^j &\Rightarrow x \notin \Omega. \end{aligned}$$

Also ist $\partial\Omega \subset \bigcup_{j=1}^{l} U_j$. *Wir können dann noch eine offene Menge* U_0 *mit* $\overline{U_0} \subset \Omega$ *derart hinzunehmen, dass* $U_0, \ldots, U_l$ *ganz* $\overline{\Omega}$ *überdecken.*

Satz 3.12 ([30, Satz 6.7])
Sei $\Omega \subset \mathbb{R}^n$ *offen, beschränkt und mit Lipschitz-Rand. Weiterhin sei* $1 \leq p < \infty$. *Dann ist* $\{f|_\Omega \mid f \in C_0^\infty(\mathbb{R}^n)\}$ *dicht in* $W^{m,p}(\Omega)$.

Definition 3.14 ([30, Def. 6.10])
Für $m \in \mathbb{N}_0$ *und* $1 \leq p < \infty$ *ist der Raum* $W_0^{m,p}(\Omega)$ *der Abschluß von* $C_0^\infty(\Omega)$ *im Raum* $W^{m,p}(\Omega)$, *d. h.*

$$W_0^{m,p}(\Omega) := \{f \in W^{m,p}(\Omega) \mid \text{es gibt eine Folge } (f_k)_{k\in\mathbb{N}} \subset C_0^\infty(\Omega) \text{ mit } f_k \to f \text{ in } W^{m,p}(\Omega)\}.$$

Satz 3.13 (Spursatz, [1, A 6.6])
Sei $\Omega \subset \mathbb{R}^n$ *offen, beschränkt und mit Lipschitz-Rand und* $1 \leq p < \infty$. *Dann gibt es genau eine stetige lineare Abbildung*

$$S : W^{1,p}(\Omega) \to L^p(\partial\Omega),$$

so dass

$$S(f) = f|_{\partial\Omega} \text{ für } \ f \in W^{1,p}(\Omega) \cap C^0(\overline{\Omega}).$$

Wir schreiben $f(x)$ *statt* $S(f)(x)$ *für* $x \in \partial\Omega$.

Satz 3.14 (schwacher Satz von Gauß, [1, A 6.8])
Sei $\Omega \subset \mathbb{R}^m$ *offen und beschränkt mit Lipschitz-Rand. Ist* $f \in W^{1,1}(\Omega)$, *so gilt für* $i = 1, \ldots m$

$$\int_\Omega \frac{\partial f}{\partial x_i}\,\mathrm{d}x = \int_{\partial\Omega} f n_i\,\mathrm{d}\sigma.$$

Sei $1 \leq p < \infty$. *Ist* $f \in W^{1,p}(\Omega)$ *und* $g \in W^{1,q}(\Omega)$ *mit* $\frac{1}{p} + \frac{1}{q} = 1$, *so gilt für* $i = 1, \ldots, m$

$$\int_\Omega f\frac{\partial g}{\partial x_i} + g\frac{\partial f}{\partial x_i}\,\mathrm{d}x = \int_{\partial\Omega} f g n_i\,\mathrm{d}\sigma.$$

n *ist dabei jeweils die äußere Einheitsnormale an* $\partial\Omega$.

Bemerkung 3.3
Das Oberflächenintegral für Mengen mit Lipschitz-Rand müsste noch genauer erklärt werden. Wir verweisen auf [1, A 6.5].

Satz 3.15 (Schwache Nullrandwerte, [1, A 6.10 Lemma])
Sei $\Omega \subset \mathbb{R}^n$ *offen und beschränkt mit Lipschitz-Rand und* $1 \leq p < \infty$. *Dann gilt mit dem Spuroperator* S *aus* Satz 3.13

$$W_0^{1,p}(\Omega) = \{f \in W^{1,p}(\Omega) \mid S(f) = 0\}.$$

Kann man eine hohe Ableitungsordnung aber kleine Integrationsordnung gegen eine geringere Ableitungsordnung aber höhere Integrationsordnung „eintauschen“?

Satz 3.16 (Sobolev-Ungleichung, [1, Satz 8.9])
Sei $\Omega \subset \mathbb{R}^n$ offen und beschränkt mit Lipschitz-Rand. Weiter sei $m_1 \geq 0$, $m_2 \geq 0$, $1 \leq p_1 < \infty$ und $1 \leq p_2 < \infty$. Ist $m_1 - \frac{n}{p_1} \geq m_2 - \frac{n}{p_2}$ und $m_1 \geq m_2$, so gilt

$$W^{m_1,p_1}(\Omega) \hookrightarrow W^{m_2,p_2}(\Omega).$$

Ist $m_1 - \frac{n}{p_1} > m_2 - \frac{n}{p_2}$ und $m_1 > m_2$, so ist $W^{m_1,p_1}(\Omega) \hookrightarrow W^{m_2,p_2}(\Omega)$ kompakt.

Man stelle sich eine Funktion f vor, die schwache Nullrandwerte besitzt und der Gradient verschwindet fast überall in Ω. Gilt dann bereits $f = 0$?

Satz 3.17 (Poincaré-Ungleichung, [55, Thm. 4.22])
Sei $\Omega \subset \mathbb{R}^n$ ein beschränktes Gebiet mit Lipschitz-Rand, $1 \leq p < \infty$ und $V \subseteq W^{1,p}(\Omega)$, sodass eine der nachfolgenden Bedingungen für alle $f \in V$ erfüllt sind.

- *$f = 0$ auf $\partial\Omega$.*
- *$\int\limits_\Omega f \,\mathrm{d}x = 0$.*
- *Für $\Gamma \subseteq \partial\Omega$ mit positiven Maß gilt $\int\limits_\Gamma f \,\mathrm{d}\sigma = 0$.*

Dann gibt es eine Konstante $C(\Omega, V)$, sodass

$$||f||_{W^{1,p}(\Omega)} \leq C(\Omega, V)|f|_{W^{1,p}(\Omega)} \ \forall f \in V.$$

Die Poincaré-Ungleichung lässt sich auch unter schwächeren Voraussetzungen an $\partial\Omega$ und stärkeren Bedingungen an V zeigen. Für unsere Zwecke benötigen wir die Bedingung das Ω Lipschitz-Rand hat und die Version ist ausreichend. Satz 3.17 besagt, dass unter geeigneten Voraussetzungen $|\cdot|_{W^{m,p}(\Omega)}$ zu einer zu $||\cdot||_{W^{m,p}(\Omega)}$ äquivalenten Norm für $1 \leq p < \infty$ wird.

3.2 Stationären Navier-Stokes Gleichungen

Wir setzen

$$\begin{aligned}
||f||_{L^p(\Omega)^n} &:= (\sum_{i=1}^n ||f_i||^p_{L^p(\Omega)})^{\frac{1}{p}}, \\
||f||_{W^{m,p}(\Omega)^n} &:= (\sum_{i=1}^n ||f_i||^p_{W^{m,p}(\Omega)})^{\frac{1}{p}}, \\
|f|_{W^{m,p}(\Omega)^n} &:= (\sum_{i=1}^n |f_i|^p_{W^{m,p}(\Omega)})^{\frac{1}{p}}, \\
(f,g)_{L^2(\Omega)^n} &:= \sum_{i=1}^n (f_i, g_i)_{L^2(\Omega)}, \\
(f,g)_{W^{m,2}(\Omega)^n} &:= \sum_{i=1}^n (f_i, g_i)_{W^{m,2}(\Omega)}.
\end{aligned}$$

Im Folgenden ist $\Omega \subset \mathbb{R}^2$ stets ein beschränktes Gebiet mit Lipschitz-Rand $\partial\Omega$ und $\nu \in \mathbb{R}^+$. Das System

$$-\nu\Delta v + (v \cdot \nabla)v + \nabla p = f \text{ in } \Omega, \tag{3.1}$$

$$\operatorname{div} v = 0 \text{ in } \Omega, \tag{3.2}$$
$$v = g \text{ auf } \partial\Omega, \tag{3.3}$$

nennen wir die inkompressiblen stationären Navier-Stokes Gleichungen in 2 Dimensionen. v beschreibt das Geschwindigkeitsfeld eines inkompressiblen Fluids und p den Druck. Inkompressibel heißt, dass Druckänderung bei konstanter Temperatur zu keiner Volumenänderung führt. $\nu \in \mathbb{R}^+$ beschreibt die Viskosität und ist ein Maß für die Zähflüssigkeit des Fluides. f beschreibt eine äußere Kraft und g die Randbedingung für v. Weiterhin gelten die folgenden Schreibweisen.

$$\begin{aligned}
\Delta v &= \begin{pmatrix} \Delta v_1 \\ \Delta v_2 \end{pmatrix} = \begin{pmatrix} \frac{\partial^2 v_1}{\partial x_1^2} + \frac{\partial^2 v_1}{\partial x_2^2} \\ \frac{\partial^2 v_2}{\partial x_1^2} + \frac{\partial^2 v_2}{\partial x_2^2} \end{pmatrix}, \\
\operatorname{div} v &= \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2}, \\
\nabla p &= \begin{pmatrix} \frac{\partial p}{\partial x_1} \\ \frac{\partial p}{\partial x_2} \end{pmatrix}, \\
(u \cdot \nabla) v &= (u_1 \frac{\partial}{\partial x_1} + u_2 \frac{\partial}{\partial x_2}) v = \begin{pmatrix} u_1 \frac{\partial v_1}{\partial x_1} + u_2 \frac{\partial v_1}{\partial x_2} \\ u_1 \frac{\partial v_2}{\partial x_1} + u_2 \frac{\partial v_2}{\partial x_2} \end{pmatrix}, \\
\nabla u : \nabla v &= \begin{pmatrix} \nabla u_1 \\ \nabla u_2 \end{pmatrix} : \begin{pmatrix} \nabla v_1 \\ \nabla v_2 \end{pmatrix} = \nabla u_1 \cdot \nabla v_1 + \nabla u_2 \cdot \nabla v_2 = \sum_{i,j=1}^{2} \frac{\partial u_i}{\partial x_j} \frac{\partial v_i}{\partial x_j}.
\end{aligned}$$

Wir wollen uns zunächst dem Fall homogener Dirichletrandbedingungen widmen, d. h. $g = 0$. Nun nehmen wir vorerst an, dass (3.1) eine klassische Lösung $v \in C^2(\Omega)^2, p \in C^1(\Omega)$ besitzt. Also ist $f \in C(\Omega)^2$. Wir multiplizieren (3.1) mit einer Testfunktion $\tilde{v} \in C_0^\infty(\Omega)^2$ und integrieren. Wir wollen den klassischen Satz von Gauß verwenden und wählen $\mathfrak{B}_r(0)$ mit $\Omega \subset \mathfrak{B}_r(0) \subset \mathbb{R}^2$. Wir setzen die Funktionen f, v und p durch 0 auf $\overline{\mathfrak{B}_r(0)}$ fort. Die auftretenden Integrale existieren aufgrund der Stetigkeit und weil wir stets über ein Kompaktum integrieren.

$$\begin{aligned}
\int_\Omega f \cdot \tilde{v} \,\mathrm{d}x &= \int_\Omega (-\nu \Delta v + (v \cdot \nabla) v + \nabla p) \cdot \tilde{v} \,\mathrm{d}x \\
&= \int_\Omega -\nu \Delta v \cdot \tilde{v} + (v \cdot \nabla) v \cdot \tilde{v} + \nabla p \cdot \tilde{v} \,\mathrm{d}x \\
&= \int_{\mathfrak{B}_r(0)} -\nu \Delta v \cdot \tilde{v} + (v \cdot \nabla) v \cdot \tilde{v} + \nabla p \cdot \tilde{v} \,\mathrm{d}x \\
&= -\nu \sum_{i,j=1}^{2} \int_{\mathfrak{B}_r(0)} \frac{\partial^2 v_i}{\partial x_j^2} \tilde{v}_i \,\mathrm{d}x + \int_{\mathfrak{B}_r(0)} (v \cdot \nabla) v \cdot \tilde{v} \,\mathrm{d}x + \sum_{i=1}^{2} \int_{\mathfrak{B}_r(0)} \frac{\partial p}{\partial x_i} \tilde{v}_i \,\mathrm{d}x \\
&= \nu \sum_{i,j=1}^{2} \int_{\mathfrak{B}_r(0)} \frac{\partial v_i}{\partial x_j} \frac{\partial \tilde{v}_i}{\partial x_j} \,\mathrm{d}x - \nu \sum_{i,j=1}^{2} \int_{\partial\mathfrak{B}_r(0)} \frac{\partial v_i}{\partial x_j} \tilde{v}_i n_j \,\mathrm{d}\sigma + \int_{\mathfrak{B}_r(0)} (v \cdot \nabla) v \cdot \tilde{v} \,\mathrm{d}x - \\
&\quad \sum_{i=1}^{2} \int_{\mathfrak{B}_r(0)} p \frac{\partial \tilde{v}_i}{\partial x_i} \,\mathrm{d}x + \sum_{i=1}^{2} \int_{\partial\mathfrak{B}_r(0)} p \tilde{v}_i n_i \,\mathrm{d}\sigma \\
&= \int_{\mathfrak{B}_r(0)} \nu \nabla v : \nabla \tilde{v} + (v \cdot \nabla) v \cdot \tilde{v} - p \operatorname{div} \tilde{v} \,\mathrm{d}x
\end{aligned}$$

$$= \int_\Omega \nu \nabla v : \nabla \tilde{v} + (v \cdot \nabla) v \cdot \tilde{v} - p \operatorname{div} \tilde{v} \, \mathrm{d}x.$$

Da $\tilde{v} \in C_0^\infty(\Omega)^2$ beliebig war, erhalten wir

$$\int_\Omega \nu \nabla v : \nabla \tilde{v} + (v \cdot \nabla) v \cdot \tilde{v} - p \operatorname{div} \tilde{v} \, \mathrm{d}x = \int_\Omega f \cdot \tilde{v} \, \mathrm{d}x \ \forall \tilde{v} \in C_0^\infty(\Omega)^2. \tag{3.4}$$

Das heißt, wenn $v \in C^2(\Omega)^2$ und $p \in C^1(\Omega)$ (3.1) erfüllen, so erfüllen v und p ebenfalls (3.4). Sei $f \in C(\Omega)^2, v \in C^2(\Omega)^2$ und $p \in C^1(\Omega)$. Falls v und p (3.4) erfüllen, so liefert Satz 3.8 und die Stetigkeit, dass (3.1) punktweise erfüllt wird.
Für die Nebenbedingung (3.2) gehen wir analog vor. Sei $v \in C^2(\Omega)^2$ beliebig. Stetigkeit und Satz 3.8 liefern

$$\operatorname{div} v = 0 \text{ punktweise in } \Omega \Leftrightarrow \int_\Omega \tilde{p} \operatorname{div} v \, \mathrm{d}x = 0 \ \forall \tilde{p} \in C_0^\infty(\Omega). \tag{3.5}$$

Im Allgemeinen haben partielle Differentialgleichungen keine Lösung im klassischen Sinne und man schwächt den Lösungsbegriff ab. (3.4) motiviert von v nur zu verlangen, dass alle schwachen Ableitungen bis zur ersten Ordnung existieren. Mit $g = 0$ und Satz 3.15 fordern wir $v \in W_0^{1,2}(\Omega)^2$.
p ist nach (3.1) nur bis auf eine additive Konstante bestimmt und wir haben keine Randbedingung für p. Da wir keine Randbedinungen für p haben, benötigen wir auch nicht den Spursatz und damit auch keine höhere Regularität. Um später ein Eindeutigkeitsresultat erhalten zu können, verlangen wir von p die Mittelwertfreiheit, d. h. $p \in L_0^2(\Omega) := \{h \in L^2(\Omega) \mid (h, 1)_{L^2(\Omega)} = 0\}$.
Wir können die Menge der Testfunktionen vergrößern vgl. Satz 3.5 und Definition 3.14.
$v \in W_0^{1,2}(\Omega)^2$, Satz 3.14 und Satz 3.15 liefern $(1, \operatorname{div} v)_{L^2(\Omega)} = 0$. Für (3.5) ergibt sich

$$\begin{aligned} c(\tilde{p}) :&= \frac{(\tilde{p}, 1)_{L^2(\Omega)}}{\mu(\Omega)} \in \mathbb{R}, \\ 0 &= (\tilde{p}, \operatorname{div} u)_{L^2(\Omega)} = (\tilde{p} - c(\tilde{p})1, \operatorname{div} u)_{L^2(\Omega)} + c(\tilde{p})(1, \operatorname{div} u)_{L^2(\Omega)} \\ &= (\tilde{p} - c(\tilde{p})1, \operatorname{div} u)_{L^2(\Omega)} \ \forall \tilde{p} \in L^2(\Omega) \Leftrightarrow \\ 0 &= (\tilde{p}, \operatorname{div} u)_{L^2(\Omega)} \ \forall \tilde{p} \in L_0^2(\Omega). \end{aligned}$$

Aufgrund der Ungleichung von Cauchy-Schwarz sind Skalarprodukte beschränkt. Bevor wir die schwache Formulierung aufstellen, müssen wir formalerweise absichern, dass $((u \cdot \nabla)v, w)_{L^2(\Omega)^2}$ wohldefiniert ist. Damit meinen wir die Beschränktheit. Diese ergibt sich aus der Linearität und der Stetigkeit. Die Stetigkeit ist nicht offensichtlich und in höheren Dimensionen steht der genutzter Einbettungssatz nicht mehr zur Verfügung.

Lemma 3.7 ([34, Ch. 4 §2 Lemma 2.1])
Die Funktion $N : W^{1,2}(\Omega)^2 \times W^{1,2}(\Omega)^2 \times W^{1,2}(\Omega)^2 \to \mathbb{R}$ mit $N(u, v, w) := ((u \cdot \nabla)v, w)_{L^2(\Omega)^2}$ ist stetig und linear in jedem Argument.

Beweis. Die Linearität zeigen wir nicht. Zur Abschätzung benutzen wir Lemma 3.5. Da $1 - \frac{2}{2} \geq 0 - \frac{2}{4}$ ist nach Satz 3.16 $W^{1,2}(\Omega) \hookrightarrow L^4(\Omega)$.

$$|((u \cdot \nabla)v, w)_{L^2(\Omega)^2}| \leq \sum_{i,j=1}^{2} ||u_i \frac{\partial v_j}{\partial x_i} w_j||_{L^1(\Omega)} \leq \sum_{i,j=1}^{2} ||u_i w_j||_{L^2(\Omega)} ||\frac{\partial v_j}{\partial x_i}||_{L^2(\Omega)}$$

$$
\begin{aligned}
&= \sum_{i,j=1}^{2} |||u_i|^2|w_j|^2||^{\frac{1}{2}}_{L^1(\Omega)}||\frac{\partial v_j}{\partial x_i}||_{L^2(\Omega)} \leq \sum_{i,j=1}^{2} |||u_i|^2||^{\frac{1}{2}}_{L^2(\Omega)}|||w_j|^2||^{\frac{1}{2}}_{L^2(\Omega)}||\frac{\partial v_j}{\partial x_i}||_{L^2(\Omega)} \\
&= \sum_{i,j=1}^{2} ||u_i||_{L^4(\Omega)}||w_j||_{L^4(\Omega)}||\frac{\partial v_j}{\partial x_i}||_{L^2(\Omega)} \leq \sum_{i,j=1}^{2} ||u_i||_{L^4(\Omega)}||w_j||_{L^4(\Omega)}||v_j||_{W^{1,2}(\Omega)} \\
&\leq C \sum_{i,j=1}^{2} ||u_i||_{W^{1,2}(\Omega)}||w_j||_{W^{1,2}(\Omega)}||v_j||_{W^{1,2}(\Omega)} \\
&\leq C \sum_{i,j=1}^{2} ||u||_{W^{1,2}(\Omega)^2}||w||_{W^{1,2}(\Omega)^2}||v||_{W^{1,2}(\Omega)^2} \\
&\leq C||u||_{W^{1,2}(\Omega)^2}||w||_{W^{1,2}(\Omega)^2}||v||_{W^{1,2}(\Omega)^2}.
\end{aligned}
$$

Die Konstante C wurde generisch verwendet. □

Wir können nun die schwache Formulierung für (3.1), (3.2), (3.3) aufschreiben und setzen $V := W_0^{1,2}(\Omega)^2$ und $Q := L_0^2(\Omega)$.

Problem 3.1 ([34, §2. (2.8)], [53, §IV.1 (1.2a), (1.2b)])
Gegeben sei $f \in V^*$. Finde $(v, p) \in V \times Q$, sodass

$$\nu(\nabla v, \nabla \tilde{v})_{L^2(\Omega)^{2\times 2}} + N(v, v, \tilde{v}) - (p, \operatorname{div} \tilde{v})_{L^2(\Omega)} = f(\tilde{v}) \ \forall \tilde{v} \in V, \tag{3.6}$$

$$(\tilde{p}, \operatorname{div} v)_{L^2(\Omega)} = 0 \qquad \forall \tilde{p} \in Q. \tag{3.7}$$

Satz 3.18 ([34, Ch. 4 §2 Thm 2.1, Thm 2.2], [53, §IV.1 Thm 1.1])
Problem 3.1 *hat eine Lösung. Falls zusätzlich*

$$\frac{\mathcal{N}}{\nu^2}||f||_* < 1$$

mit

$$
\begin{aligned}
\mathcal{N} &:= \sup_{u,v,w \in W\setminus\{0\}} \frac{N(u,v,w)}{|u|_{W^{1,2}(\Omega)^2}|v|_{W^{1,2}(\Omega)^2}|w|_{W^{1,2}(\Omega)^2}}, \\
||f||_* &:= \sup_{v \in W\setminus\{0\}} \frac{|f(v)|}{|v|_{W^{1,2}(\Omega)^2}}, \\
W &:= \{v \in V \mid (\tilde{q}, \operatorname{div} v)_{L^2(\Omega)} = 0 \ \forall \tilde{q} \in Q\},
\end{aligned}
$$

gilt, so ist die Lösung eindeutig.

Bemerkung 3.4
Eine Möglichkeit die Existenz von schwache Lösungen im Falle homogener Dirichletranddaten zu zeigen, die auch in den meisten Büchern gewählt wird, ist folgende. Man nutzt den Fixpunktsatz von Brouwer um die Existenz von Lösungen eines nichtlinearen Nullstellenproblems zu zeigen [34, Ch. 4 §1. Corollary 1.1]. *Dann untersucht man ob Lösungen in divergenzfreien endlichdimensionalen Teilräumen existieren (Galerkin-Verfahren)* [34, Ch. 4 §1. Thm. 1.2]. *Da man vorher das Nullstellenproblem studiert hat, kann man Existenz von Lösungen sichern. Eine funktionalanalytische Eigenschaften sichert die Existenz einer schwach konvergenten Teilfolge.*
Einen abstrakteren Zugang erhält man mit Mitteln der nichtlinearen Funktionalanalysis. Dieser

ist in [54, Kap. 3.2.3] *zu finden. Die Vorgehensweise ist zwar abstrakt, aber die Idee ist einfach zu verstehen.*
Betrachtet man eine Funktion $A \in \mathfrak{F}(\mathbb{R}, \mathbb{R})$ *die unbeschränkt und stetig ist, dann ist* A *surjektiv. Die Surjektivität kann man auch als ein Existenzresultat von Lösungen der womöglich nichtlinearen Gleichung* $A(x) = f$ *interpretieren.*
Zu jeder rechten Seite f *existiert mindestens ein* x*, sodass* x *die Gleichung* $A(x) = f$ *löst. Ein ähnliches Resultat in geeigneten Banachräumen und unter geeigneten Voraussetzungen an* A *liefert der Satz über pseudomonotone Operatoren* [54, 2.10 Satz].
Wir betrachten den divergenzfreien Raum

$$X := \{u \in W_0^{1,2}(\Omega)^2 \mid \operatorname{div} u = 0\} \text{ mit } |\cdot|_{W^{1,2}(\Omega)^2}.$$

Hierzu zeigt man, dass X *ein abgeschlossener Teilraum von* $W_0^{1,2}(\Omega)^2$ *ist und damit wieder ein Banachraum. Da* $W_0^{1,2}(\Omega)^2$ *separabel und reflexiv ist, ist* X *als abgeschlossener Teilraum ebenfalls separabel und reflexiv. Nun studiert man die Operatoren* $A, B \in \mathfrak{F}(X, X^*)$

$$\langle A(v), \tilde{v}\rangle_{X^*\times X} := \nu \int_\Omega \nabla v : \nabla \tilde{v} \,\mathrm{d}x,$$
$$\langle B(v), \tilde{v}\rangle_{X^*\times X} := \int_\Omega (v \cdot \nabla) v \cdot \tilde{v} \,\mathrm{d}x.$$

Diese ergeben sich aus der schwache Formulierung, wenn man divergenzfreie Testfunktionen wählt. Man zeigt, dass $A + B$ *beschränkt, pseudomonoton und koerziv ist. Für die Koerzitivität ist entscheidend, dass* $\langle B(v), v\rangle_{X^*\times X} = 0$ *gilt. Hierzu nutzt man, dass* $v \in X$ *divergenzfrei ist und nach partieller Integration der nichtlineare Term verschwindet. (Das passiert im Falle gemischter Randbedingungen leider nicht.) Der Satz über pseudomonote Operatoren liefert die Surjektivität des Operators* $A + B$*, d. h. zu jeder rechten Seite* $f \in X^*$ *gibt es ein* v*, sodass* $(A + B)(v) = f$ *gilt. Die Gleichheit ist im* X^* *Sinne zu verstehen, dass heißt* $\langle (A + B)(v), \tilde{v}\rangle_{X^*\times X} = \langle f, \tilde{v}\rangle_{X^*\times X} = f(\tilde{v})\ \forall \tilde{v} \in X$*. Das heißt aber, dass* v *die schwache Formulierung im divergenzfreien Raum erfüllt.*
Natürlich braucht man zu v *jetzt noch einen Druck* p*. Dazu wendet man* Satz 3.19 *mit* $F = A(v) + B(v) - f \in V^*$ *an.*

Bemerkung 3.5 ([34, Ch. 4 §1 Thm 1.4], [53, §IV.1 (1.4)])
Um für ein gegebenes Geschwindigkeitsfeld $u \in V$ *die Existenz eines eindeutiges Druckes* $p \in Q$ *zu sichern, ist es notwendig, dass die Räume* V *und* Q *die inf-sup Bedingung*

$$\inf_{q\in Q\setminus\{0\}} \sup_{v\in V\setminus\{0\}} \frac{(q, \operatorname{div} v)_{L^2(\Omega)}}{||q||_{L^2(\Omega)}|v|_{W^{1,2}(\Omega)^2}} \geq \beta > 0 \tag{3.8}$$

erfüllen.

Satz 3.19 (de Rham, [25, Thm. IV.2.4], [54, §3 2.35 Satz])
Sei $\Omega \subset \mathbb{R}^2$ *ein beschränktes Gebiet mit Lipschitz-Rand. Weiterhin sei* $F \in V^*$ *derart, dass für alle* $\phi \in C_0^\infty(\Omega)$ *mit* $\operatorname{div} \phi = 0$

$$F(\phi) = 0$$

gilt. Dann gibt es eine eindeutige Funktion $p \in Q$*, sodass für alle* $\phi \in V$

$$F(\phi) = \int_\Omega p \operatorname{div} \phi \,\mathrm{d}x$$

gilt.

Bemerkung 3.6
Satz 3.19 *ist für die Theorie wichtig, denn um schwache Lösungen für die Navier-Stokes Gleichungen (v, p) zu finden, reicht es zunächst die Existenz eines Geschwindigkeitsfeld v in einem geeigneten divergenzfreien Funktionenraum zu beweisen. Anwendung von* Satz 3.19 *liefert dann die Existenz eines zu v zugehörigen eindeutigen Druckes p.*

Wir wollen uns nun dem allgemeineren Fall inhomogener Dirichletranddaten widmen. Mit Satz 3.14 und $v \in V$ erhalten wir $\int_\Omega \operatorname{div} v \, \mathrm{d}x = \int_{\partial\Omega} v \cdot n \, \mathrm{d}\sigma = 0$, deshalb fordern wir

$$\int_{\partial\Omega} g \cdot n \, \mathrm{d}\sigma = 0. \tag{3.9}$$

Die schwache Formulierung können wir nun für den Fall inhomogener Dirichletranddaten aufstellen.

Problem 3.2 ([34, Ch. 4 §2 (2.20)])
Gegeben sei $f \in V^*, g \in W^{\frac{1}{2},2}(\partial\Omega)$ und g erfüllt (3.9)*. Finde $(v, p) \in W^{1,2}(\Omega)^2 \times Q$, sodass*

$$\nu(\nabla v, \nabla \tilde{v})_{L^2(\Omega)^{2\times 2}} + N(v, v, \tilde{v}) - (p, \operatorname{div} \tilde{v})_{L^2(\Omega)} = f(\tilde{v}) \ \forall \tilde{v} \in V, \tag{3.10}$$

$$(\tilde{p}, \operatorname{div} v)_{L^2(\Omega)} = 0 \qquad \forall \tilde{p} \in Q, \tag{3.11}$$

$$v = g \qquad \text{auf } \partial\Omega. \tag{3.12}$$

Satz 3.20 ([34, Ch. 4 §2 Thm. 2.3])
Problem 3.2 *hat eine Lösung.*

Bemerkung 3.7
Unter einer „Kleinheitsbedingung“ kann man für Problem 3.2 *auch ein Eindeutigkeitsresultat erhalten vgl.* [34, Ch. 4 §2 Thm. 2.4].
Der Sobolev-Raum mit $m \in \mathbb{R}^+$ ist in [34, Ch. 1 §1 Def. 1.2] *definiert. Wie sich die Sobolev-Räume mit reeler Ableitungsordnung als natürliche Erweiterung ergeben ist in einem „langen Weg zu Fuß“ in* [39, Kap. 3.1, 3.2, 3.3, 3.4, 4.2] *dargestellt. In* [25, Def. III 2.20] *wird $W^{\frac{1}{2},2}(\partial\Omega)$ als Bild von $W^{1,2}(\Omega)$ unter dem Spuroperator S definiert.*

3.3 Stationären Navier-Stokes Gleichungen mit gemischten Randbedingungen

Wir sind im Folgenden mehr an gemischten Randbedingungen interessiert. Wir folgen [27] und betrachten das folgende Problem unter genügend großen Regularitätsannahmen an v, p, f, g, Ω und $\partial\Omega$.

$$-\nu\Delta v + (v \cdot \nabla)v + \nabla p = f \quad \text{in } \Omega, \tag{3.13}$$

$$\operatorname{div} v = 0 \quad \text{in } \Omega, \tag{3.14}$$

$$v = g \quad \text{auf } \Gamma_D, \tag{3.15}$$

$$\nu \frac{\partial v}{\partial n} = pn \text{ auf } \Gamma_O, \tag{3.16}$$

wobei $\partial\Omega = \Gamma_D \dot\cup \Gamma_O$ und Γ_D sowie Γ_O haben positives Maß. Wie zuvor betrachten wir zuerst den Fall $g = 0$. Die Bedingung (3.16) beschreibt eine Ausflussbedingung. Diese ergibt sich in

natürlicher Weise durch partielle Integration, denn

$$
\begin{aligned}
&-\nu \int_\Omega \Delta v \cdot \tilde{v} + (v \cdot \nabla) v \cdot \tilde{v} + \nabla p \cdot \tilde{v} \, \mathrm{d}x = \\
&\nu \int_\Omega \nabla v : \nabla \tilde{v} \, \mathrm{d}x - \nu \int_{\partial\Omega} \nabla v n \cdot \tilde{v} \, \mathrm{d}\sigma + \int_\Omega (v \cdot \nabla) v \cdot \tilde{v} \, \mathrm{d}x - \int_\Omega p \operatorname{div} \tilde{v} \, \mathrm{d}x + \int_{\partial\Omega} p \tilde{v} \cdot n \, \mathrm{d}\sigma = \\
&\nu \int_\Omega \nabla v : \nabla \tilde{v} \, \mathrm{d}x + \int_\Omega (v \cdot \nabla) v \cdot \tilde{v} \, \mathrm{d}x - \int_\Omega p \operatorname{div} \tilde{v} \, \mathrm{d}x + \int_{\partial\Omega} (pn - \nu \frac{\partial v}{\partial n}) \cdot \tilde{v} \, \mathrm{d}\sigma.
\end{aligned}
$$

Wählen wir nun eine Testfunktion $\tilde{v}$ mit $\tilde{v}|_{\Gamma_D} = 0$ und $\operatorname{div} \tilde{v} = 0$ f.ü. in Ω so gilt

$$
\begin{aligned}
\int_{\partial\Omega} (pn - \nu \frac{\partial v}{\partial n}) \cdot \tilde{v} \, \mathrm{d}\sigma = 0, \\
\int_\Omega p \operatorname{div} \tilde{v} \, \mathrm{d}x = 0.
\end{aligned}
$$

Damit lautet die schwache Formulierung für (3.13), (3.14) (3.15), (3.16) mit $g = 0$.

Problem 3.3
Finde $v \in V$, sodass

$$
\nu(\nabla v, \nabla \tilde{v})_{L^2(\Omega)^{2\times 2}} + ((v \cdot \nabla) v, \tilde{v})_{L^2(\Omega)^2} = f(\tilde{v}) \ \forall \tilde{v} \in V, \tag{3.17}
$$

wobei V ein geeigneter Testfunktionenraum ist.

Um eine a priori Schranke herzuleiten nehmen wir an, dass (3.17) eine Lösung $v \in V$ hat und testen mit dieser. So erhalten wir

$$
\nu(\nabla v, \nabla v)_{L^2(\Omega)^{2\times 2}} + ((v \cdot \nabla) v, v)_{L^2(\Omega)^2} = f(v).
$$

Für den nichtlinearen Term gilt nun

$$
\begin{aligned}
((v \cdot \nabla) v, v)_{L^2(\Omega)^2} = \sum_{i,j=1}^{2} \int_\Omega v_i \frac{\partial v_j}{\partial x_i} v_j \, \mathrm{d}x = \sum_{i,j=1}^{2} \int_\Omega \frac{\partial v_j}{\partial x_i} v_i v_j \, \mathrm{d}x = \\
- \sum_{i,j=1}^{2} \int_\Omega v_j \frac{\partial (v_i v_j)}{\partial x_i} \, \mathrm{d}x + \sum_{i,j=1}^{2} \int_{\Gamma_O} v_j v_i v_j n_i \, \mathrm{d}\sigma = \\
- \sum_{i,j=1}^{2} \int_\Omega v_j \frac{\partial v_i}{\partial x_i} v_j \, \mathrm{d}x - \sum_{i,j=1}^{2} \int_\Omega v_j \frac{\partial v_j}{\partial x_i} v_i \, \mathrm{d}x + \sum_{i,j=1}^{2} \int_{\Gamma_O} v_j v_i v_j n_i \, \mathrm{d}\sigma = \\
- \sum_{j=1}^{2} \int_\Omega v_j v_j \operatorname{div} v \, \mathrm{d}x - ((v \cdot \nabla) v, v)_{L^2(\Omega)^2} + \int_{\Gamma_O} (v \cdot n)(v \cdot v) \, \mathrm{d}\sigma \Rightarrow \\
((v \cdot \nabla) v, v)_{L^2(\Omega)^2} = \frac{1}{2} \int_{\Gamma_O} (v \cdot n)(v \cdot v) \, \mathrm{d}\sigma.
\end{aligned}
$$

Wir erhalten eine Abschätzung

$$\nu|v|^2_{W^{1,2}(\Omega)^2} \leq ||f||_* |v|_{W^{1,2}(\Omega)^2} - \frac{1}{2}\int\limits_{\Gamma_O} (v \cdot n)(v \cdot v)\,\mathrm{d}\sigma$$

$$||f||_* = \sup_{v \in V \setminus \{0\}} \frac{|f(v)|}{|v|_{W^{1,2}(\Omega)^2}}.$$

Bemerkung 3.8
Die Schwierigkeit ist nun, dass durch den Rand Γ_O Anteile „hineinfließen“ können, d. h. $(v \cdot n) < 0$. Ziel ist es unkontrollierten Rückfluss und damit ein „aufpumpen“ der Lösung über den Ausflussrand auszuschließen und den Term mit einer geeigneten Ungleichung zu beschränken. Daher lässt sich vermutlich so keine a priori Schranke herleiten. Diese wäre aber für einen Existenzbeweis von schwachen Lösungen hilfreich.
Im Falle reiner Dirichletranddaten verschwindet der Term $(v \cdot \nabla)v, v)_{L^2(\Omega)^2}$ und macht keine weiteren „Probleme“.
Wir haben die stationären Navier-Stokes Gleichungen mit gemischten Randbedingungen unter „genügend“ großen Regularitätsannahmen an die Daten betrachtet, da eine Randbedingung an den Druck p gegeben war. Für $p \in L^2(\Omega)$ und um die Existenz von $\frac{\partial v}{\partial n}$ auf dem Rand für $v \in W^{1,2}(\Omega)$ zu sichern, steht uns die in dieser Arbeit vorgestellte Version des Spursatzes leider nicht zur Verfügung.
Die betrachtete Ausflussbedingung bringt eine „versteckte“ Randbedingung an den Druck p mit sich

$$\int\limits_{\Gamma_O} p\,\mathrm{d}\sigma = 0,$$

vgl. [32, Methods for Numerical Flow Simulation, Ch. 2.1 (2.3), (2.4), S. 290/291].
Selbst im Fall $f = 0$ und $g = 0$, weiß man nicht ob die triviale Lösung eindeutig ist vgl. [32, Methods for Numerical Flow Simulation, S. 294], [26, Ch. 14.1.2] .
Weitere Schwierigkeiten mit der betrachteten Ausflussbedingungen werden in [32, Methods for Numerical Flow Simulation, Ch. 2.1] *diskutiert.*

4 Instationäre Navier-Stokes Gleichungen

Ziel dieses Abschnitts ist es die Einführung der Bochner-Räume, um geeignete Räume für zeitabhängige partielle Differentialgleichungen zur Verfügung zu stellen. Zur Untersuchung der instationären Navier-Stokes Gleichungen arbeitet man mit drei „speziellen“ Hilberträumen, die dicht ineinander liegen. Wir fassen kurz wichtige Eigenschaften dieser Räume zusammen und stellen dann den Existenz- und Eindeutigkeitssatz für schwache Lösungen der instationären Navier-Stokes Gleichung im zweidimensionalen Fall vor. Für den Fall gemischter Randbedingungen ergeben sich ähnliche Schwierigkeiten, wie im stationären Fall. Wir verweisen daher nur auf eine bestehende Arbeit zu diesem Thema.

Definition 4.1 (Stark messbare Funktionen, [54, §2 Def. 1.1], [55, Def. 10.1])
Sei X ein Banachraum und $I \subset \mathbb{R}$ ein beschränktes Intervall. Eine einfache Funktion ist eine Abbildung $f : I \to X$ mit endlich vielen Werten, also

$$f(t) = \sum_{j=1}^{m} \chi_{E_j}(t) x_j \ \forall t \in I$$

für $m \in \mathbb{N}$. Die Funktion f ist durch m Vektoren $x_j \in X$ für $1 \leq j \leq m$ und m messbare disjunkte Teilmengen $E_j \subseteq I$ beschrieben.
Eine Funktion $f : I \to X$ heißt stark messbar, falls sie punktweise durch einfache Funktionen approximiert werden kann. Genauer muss zu f eine Folge von einfachen Funktionen $f_k : I \to X$ existieren mit $\lim_{k\to\infty} f_k(t) = f(t)$ in X für fast alle $t \in I$. Die charakteristische Funktion χ_E ist durch

$$\chi_E(s) := \begin{cases} 1 & s \in E \\ 0 & s \notin E \end{cases}$$

definiert.

Satz 4.1 (Satz von Pettis, [55, Thm. 10.2], [54, §2 Satz. 1.8])
Sei X separabel, $I \subset \mathbb{R}$ ein beschränktes Intervall und $f : I \to X$. Falls für alle $g \in X'$ die Funktion $g(f(\cdot)) : I \to \mathbb{R}$ messbar ist, so ist f stark messbar.

Definition 4.2 (Bochner-Raum, [55, Def. 10.3])
Sei X mit $||\cdot||_X$ ein Banachraum, $I \subset \mathbb{R}$ ein beschränktes Intervall und $1 \leq p \leq \infty$. Wir definieren

$$L^p(I;X) := \{f \in \mathfrak{F}(I,X) \mid \ f \text{ ist stark messbar und } ||f||_{L^p(I;X)} < \infty\}$$

$$||f||_{L^p(I;X)} := \begin{cases} (\int_I ||f(t)||_X^p \, \mathrm{d}t)^{\frac{1}{p}} & \text{für } 1 \leq p < \infty \\ \inf\limits_{\mu(N)=0} \sup\limits_{t \in I \setminus N} ||u(t)||_X & \text{für } p = \infty. \end{cases}$$

Wir identifizieren dabei Funktionen, die für fasst alle $t \in I$ übereinstimmen.

Definition 4.3 (Bochner, [55, Thm. 10.4])

Sei X ein Banachraum, $I \subset \mathbb{R}$ ein beschränktes Intervall und $1 \leq p < \infty$. Dann ist $L^p(I;X)$ ein Banachraum. Der Raum $L^\infty(I;X)$ ist ebenfalls ein Banachraum. Weiterhin gilt:

- *Zu $f \in L^p(I;X)$ gibt es einfache Funktionen $f_k : I \to X$ mit*

$$\lim_{k\to\infty} f_k(t) = f(t) \text{ in } X \text{ für fast alle } t \in I \text{ und } \lim_{k\to\infty} \int_I ||(f_k - f)(t)||_X^p \,\mathrm{d}t = 0.$$

- *Für $f \in L^p(I;X)$ ist mit Folgen wie in (i) das Bochner-Integral*

$$\int_I f_k(t)\,\mathrm{d}t := \int_I \sum_{j=1}^m \chi_{E_j}(t) x_j \,\mathrm{d}t := \sum_{j=1}^m \mu(E_j) x_j \in X$$
$$\int_I f(t)\,\mathrm{d}t := \lim_{k\to\infty} \int_I f_k(t)\,\mathrm{d}t \in X$$

  wohldefiniert, der Limes existiert und ist unabhängig von der gewählten Folge.

- *Das Integral erfüllt*

$$||\int_I f(t)\,\mathrm{d}t||_X \leq \int_I ||f(t)||_X \,\mathrm{d}t \text{ und}$$
$$\langle g, \int_I f(t)\,\mathrm{d}t\rangle_{X^*\times X} = \int_I \langle g, f(t)\rangle_{X^*\times X}\,\mathrm{d}t \;\forall g \in X^*.$$

Lemma 4.1 (Dualraum des Bochner-Raumes [55, Prop. 10.5])
Sei X ein reflexiver Banachraum, $I \subset \mathbb{R}$ ein beschränktes Intervall und $1 < p, q < \infty$ mit $\frac{1}{p} + \frac{1}{q} = 1$. Dann ist der Raum $L^p(I;X)$ reflexiv. Der Dualraum ist

$$L^p(I;X)^* = L^q(I;X^*)$$

und für $f \in L^p(I;X)$ und $g \in L^q(I;X^*)$ gilt

$$\langle g, f\rangle_{L^q(I;X^*)\times L^p(I;X)} = \int_I \langle g(t), f(t)\rangle_{X^*\times X}\,\mathrm{d}t.$$

Definition 4.4 ([25, Def. II.5.7])
Seien X, Y Banachräume, $X \subseteq Y$, $1 \leq p, q \leq \infty$ und $I \subset \mathbb{R}$ ein beschränktes Intervall. Wir sagen $f \in L^p(I;X)$ hat eine schwache Ableitung in $L^q(I;Y)$ falls eine Funktion $g \in L^q(I;Y)$ existiert, sodass

$$\int_I \phi'(t) f(t)\,\mathrm{d}t = -\int_I \phi(t) g(t)\,\mathrm{d}t \;\forall \phi \in C_0^\infty(I).$$

Falls eine schwache Ableitung existiert, so ist diese eindeutig und wir schreiben

$$\frac{\mathrm{d}f}{\mathrm{d}t} = g(t).$$

Definition 4.5 ($L^p_{loc}(I;X)$, [50, Ch. XI, 1. S.106])
Sei X ein Banachraum, $I \subseteq \mathbb{R}$ ein Intervall und $1 \leq p < \infty$.
$L^p_{loc}(I;X) := \{f \in \mathfrak{F}(I,X) \mid f|_K \in L^p(K;X)\ \forall K \subseteq\subseteq I\ K \text{ ist ein offenes Intervall}\}$.

Definition 4.6 (Gelfandscher Dreier,[30, Def. 8.18])
Für Hilberträume V, H heißt $V \hookrightarrow H \hookrightarrow V^*$, wobei beide Einbettungen dicht sind, Gelfandscher Dreier.

Wir definieren

$$\begin{aligned}\mathcal{V} &:= \{f \in C_0^\infty(\Omega)^2 \mid \operatorname{div} f = 0\}\\ H &:= \{f \in L^2(\Omega)^2 \mid \text{es gibt eine Folge } (f_k)_{k\in\mathbb{N}} \subset \mathcal{V} \text{ mit } f_k \to f \text{ in } L^2(\Omega)^2\}\\ V &:= \{f \in W_0^{1,2}(\Omega)^2 \mid \text{es gibt eine Folge } (f_k)_{k\in\mathbb{N}} \subset \mathcal{V} \text{ mit } f_k \to f \text{ in } W_0^{1,2}(\Omega)^2\}.\end{aligned}$$

- H ist der topologische Abschluss von $\mathcal{V}$ in $L^2(\Omega)^2$ bezüglich $||\cdot||_{L^2(\Omega)^2}$.
- V ist der topologische Abschluss von $\mathcal{V}$ in $W_0^{1,2}(\Omega)^2$ bezüglich $|\cdot|_{W^{1,2}(\Omega)^2}$.
- Als abgeschlossene Unterräume von Hilberträumen, sind V und H ebenfalls Hilberträume und damit reflexiv.
- Mit Satz 3.17 gilt $||f||_{L^2(\Omega)^2} \leq C|f|_{W^{1,2}(\Omega)^2}\ \forall f \in V$ und damit $V \subseteq H$, also $V \hookrightarrow H$.
- Es gilt $\mathcal{V} \subseteq V \subseteq H$ und da $\mathcal{V}$ per Definition dicht in H ist $V \hookrightarrow H$ dicht.
- Es gilt $H^* \hookrightarrow V^*$, denn sei $f \in H^*$ beliebig, dann erhält man

  $$||f||_{H^*} = \sup_{x\in H\setminus\{0\}} \frac{|f(x)|}{||x||_{L^2(\Omega)^2}} \geq \sup_{x\in V\setminus\{0\}} \frac{|f(x)|}{||x||_{L^2(\Omega)^2}} \geq \sup_{x\in V\setminus\{0\}} \frac{|f(x)|}{C|x|_{W^{1,2}(\Omega)^2}} = \frac{1}{C}||f||_{V^*} \Rightarrow$$
  $||f||_{V^*} \leq C||f||_{H^*}\ \forall f \in H^*$.

  Da $f \in H^*$ insbesondere linear auf V, erhält man $f \in V^*$ mit Lemma 3.1.
- Mit Lemma 3.3 ist $H^* \hookrightarrow V^*$ dicht.
- Mit Satz 3.7 können wir H mit H^* identifizieren.
- Nach Definition 4.6 bilden V und H mit $V \hookrightarrow H \hookrightarrow V^*$ einen Gelfandschen Dreier.

Wie im letzten Abschnitt sei im Folgenden $\Omega \subset \mathbb{R}^2$ stets ein beschränktes Gebiet mit Lipschitz-Rand $\partial\Omega, 0 < T, \nu \in \mathbb{R}^+$ und $f \in L^2((0,T), V^*)$. Wir betrachten die instationären Navier-Stokes Gleichungen in zwei Dimensionen

$$\frac{\partial v}{\partial t} - \nu\nabla v + (v\cdot\nabla)v + \nabla p = f \quad \text{in } \Omega\times(0,T), \tag{4.1}$$
$$\operatorname{div} v = 0 \quad \text{in } \Omega\times(0,T), \tag{4.2}$$
$$v = 0 \quad \text{auf } \partial\Omega\times[0,T), \tag{4.3}$$
$$v(0) = v_0 \text{ auf } \partial\Omega. \tag{4.4}$$

Da $\Omega \subset \mathbb{R}^2$ ein beschränktes Gebiet mit Lipschitz-Rand ist, erhalten wir für V eine bessere Charakterisierung [25, Lemma IV.3.4]

$$V = \{f \in W_0^{1,2}(\Omega)^2 \mid \operatorname{div} f = 0 \text{ f.ü. in } \Omega\}.$$

Wir betrachten eine schwache Formulierung von (4.1), (4.2), (4.3), (4.4) im divergenzfreien Funktionenraum.

Problem 4.1 ([25, S. V. 1.2.1])
Gegeben sei $v_0 \in H$. Finde $v \in L^2((0,T);V)$, sodass für alle $\tilde{v} \in V$ und für alle $\phi \in C_0^\infty((0,T))$

$$\int_0^T \frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega v(t) \cdot \tilde{v} \,\mathrm{d}x\phi \,\mathrm{d}t + \int_0^T \int_\Omega ((v(t) \cdot \nabla)v(t)) \cdot \tilde{v} \,\mathrm{d}x\phi \,\mathrm{d}t + \nu \int_0^T \int_\Omega \nabla v(t) : \nabla \tilde{v} \,\mathrm{d}x\phi \,\mathrm{d}t =$$
$$\int_0^T \langle f(t), \tilde{v} \rangle_{V^* \times V} \phi \,\mathrm{d}t$$

gilt. Die Bedingung $v(0) = v_0$ soll in V^* gelten.

Bemerkung 4.1
In welchem Sinne die Bedingung $v(0) = v_0$ in V^* gelten soll, ist ohne Weiteres nicht klar. Mit einem Stetigkeitsargument begründet man die Bedingung vgl. [25, Prop. V.1.3.].

Satz 4.2 ([25, Thm. III.2.24, Remark III.2.13, Thm. V.1.4],[55, Thm. 24.4])
Es gibt ein eindeutiges Paar (v,p), welches die instationären Navier-Stokes Gleichungen im schwachen Sinne löst. Es gilt

$$(v,p) \in (L^\infty((0,T);H) \cap L^2((0,T);V)) \times W^{1,1}((0,T);L_0^2(\Omega))^*.$$

Bemerkung 4.2
Im zweidimensionalen instationären Fall benötigt man im Gegensatz zum stationären Fall keine „Kleinheitsbedingungen“ an die Daten um Eindeutigkeit zu erhalten.
Im Falle gemischter Randbedingungen mit Ausflussbedingung ergeben sich ähnliche Probleme wie im stationären Fall. Dies führt dann dazu, dass man nicht weiß, ob man globale Existenz in der Zeit sichern kann [12].

5 Bernoulligleichung

Das Newton-Verfahren für die (verallgemeinerte) Riccatigleichung aus Satz 2.6 benötigt einen geeigneten Startwert X_0, falls $\Lambda(A, M)$ instabile Eigenwerte hat. Wir sind bisher nicht darauf eingegangen, wie man ein geeignetes X_0 finden kann. Eine Möglichkeit ist geeignete Lösungen der homogenen (verallgemeinerten) algebraischen Riccatigleichung finden. Diese nennt man auch (verallgemeinerte) Bernoulligleichung und wir können diese als Riccatigleichung auffassen. Wir haben ein Newton-Verfahren zur Lösung der algebraischen Riccatigleichung vorgestellt. Man kann zu einer gegebenen algebraischen Riccatigleichung eine zugehörige Hamiltonische Matrix H aufstellen, einen stabilen H-invarianten Unterraum suchen und aus diesem Unterraum seine Lösung „extrahieren“ [13, Def. 4.12, Lemma 4.13-4.17, Satz 4.18]. Satz 5.2 liefert uns dieses Resultat.
Eine beliebtes Hilfsmittel beim Studium von Systemen gewöhnlicher Differentialgleichungen erster Ordnung mit konstanten Koeffizienten ist die Matrixexponentialfunktion. Diese definiert man für gewöhnlich über die Reihendarstellung der Exponentialfunktion. Man kann für jede Funktion f, welche in einer Umgebung U_λ um jeden Eigenwert $\lambda \in \Lambda(H)$ definiert und genügend oft differenzierbar ist, $f(H)$ mit Hilfe der Jordannormalform von H definieren [37, Definition 1.2]. Wir werden sehen, dass die Signumsfunktion ein geeignetes Hilfsmittel ist, um stabilisierende Lösungen der (verallgemeinerten) algebraischen Bernoulligleichung zu finden. Hierzu stellen wir ein Verfahren vor, um das Signum einer Matrix zu berechnen. Dieses Verfahren ist nicht für große dünnbesetzte Matrizen geeignet.

Bemerkung 5.1
$\operatorname{sign} \in \mathfrak{F}(\mathbb{C} \setminus i\mathbb{R}, \{-1, 1\})$ *ist durch*

$$\operatorname{sign}(z) := \begin{cases} 1 & \textit{falls } \Re(z) > 0 \\ -1 & \textit{falls } \Re(z) < 0 \end{cases}$$

definiert.

Definition 5.1 ([37, Ch. 5 S. 107], [11, Ch. 3.1])
Sei $H \in \mathbb{C}^{n\times n}$ *mit* $\Lambda(H) \cap i\mathbb{R} = \emptyset$ *und sei*

$$H = Z \begin{bmatrix} J_- & 0 \\ 0 & J_+ \end{bmatrix} Z^{-1}$$

die Jordannormalform von H *mit* $J_- \in \mathbb{C}^{l\times l}$ *und* $J_+ \in \mathbb{C}^{(n-l)\times(n-l)}$, *wobei* $\Lambda(J_-) \subset \mathbb{C}^-$ *und* $\Lambda(J_+) \subset \mathbb{C}^+$. *Dann ist* $\operatorname{sign}(H)$ *durch*

$$\operatorname{sign}(H) := Z \begin{bmatrix} -I_l & 0 \\ 0 & I_{n-l} \end{bmatrix} Z^{-1}$$

definiert.

Bemerkung 5.2
Die Transformationsmatrix Z *aus* Definition 5.1 *ist im Allgemeinen nicht eindeutig und die Reihenfolge der Jordanblöcke ist durch* Definition 5.1 *nur teilweise festgelegt. Wir wollen*

begründen, dass sign *unabhängig von der Wahl der Transformationsmatrix ist. Andernfalls wäre* sign *durch* Definition 5.1 *nicht wohldefiniert. Seien also Z und $\tilde{Z}$ derart, dass*

$$H = Z \begin{bmatrix} J_- & 0 \\ 0 & J_+ \end{bmatrix} Z^{-1} \text{ mit } \Lambda(J_-) \subset \mathbb{C}^- \text{ und } \Lambda(J_+) \subset \mathbb{C}^+$$
$$H = \tilde{Z} \begin{bmatrix} \tilde{J}_- & 0 \\ 0 & \tilde{J}_+ \end{bmatrix} \tilde{Z}^{-1} \text{ mit } \Lambda(\tilde{J}_-) \subset \mathbb{C}^- \text{ und } \Lambda(\tilde{J}_+) \subset \mathbb{C}^+$$

Es gilt $J_-, \tilde{J}_- \in \mathbb{C}^{l\times l}$ und $J_+, \tilde{J}_+ \in \mathbb{C}^{(n-l)\times(n-l)}$. Nach [56, Thm. 5.15] *ist eine Zerlegung des Raumes $\mathbb{C}^n$ in eine direkte Summe von H-invarianten nichttrivialen Unterräumen durch die Jordannormalform nur abhängig von den Eigenwerten von H und der Dimension der zu den Eigenwerten zugehörigen H-invarianten Unterräume. Nun bilden $(Z_{*,1}, \ldots, Z_{*,l})$ und $(\tilde{Z}_{*,1}, \ldots, \tilde{Z}_{*,l})$ jeweils eine Basis eines Unteraums $L_- \subseteq \mathbb{C}^n$. Ebenfalls bilden $(Z_{*,l+1}, \ldots, Z_{*,n})$ und $(\tilde{Z}_{*,l+1}, \ldots, \tilde{Z}_{*,n})$ jeweils eine Basis eines Unterraums $L_+ \subseteq \mathbb{C}^n$. Es gilt außerdem $L_- \oplus L_+ = \mathbb{C}^n$. Weiterhin gibt es eindeutig bestimmte Matrizen $\phi_- \in \mathrm{GL}(l, \mathbb{C})$ und $\phi_+ \in \mathrm{GL}(n-l, \mathbb{C})$, sodass $\left[Z_{*,1}, \ldots, Z_{*,l}\right] \phi_- = \left[\tilde{Z}_{*,1}, \ldots, \tilde{Z}_{*,l}\right]$ und $\left[Z_{*,l+1}, \ldots, Z_{*,n}\right] \phi_+ = \left[\tilde{Z}_{*,l+1}, \ldots, \tilde{Z}_{*,n}\right]$. Wir setzen $\phi := \begin{bmatrix} \phi_- & 0 \\ 0 & \phi_+ \end{bmatrix} \in \mathrm{GL}(n, \mathbb{C})$ und damit gilt $Z\phi = \tilde{Z}$. Nun ist*

$$\begin{aligned} \operatorname{sign}(H) &= Z \begin{bmatrix} -I_l & 0 \\ 0 & I_{n-l} \end{bmatrix} Z^{-1} = Z \begin{bmatrix} -\phi_-\phi_-^{-1} & 0 \\ 0 & \phi_+\phi_+^{-1} \end{bmatrix} Z^{-1} \\ &= Z\phi \begin{bmatrix} -I_l & 0 \\ 0 & I_{n-l} \end{bmatrix} \phi^{-1} Z^{-1} = \tilde{Z} \begin{bmatrix} -I_l & 0 \\ 0 & I_{n-l} \end{bmatrix} \tilde{Z}^{-1}. \end{aligned}$$

Die Jordannormalform einer Matrix H numerisch zu berechnen und dann $\operatorname{sign}(H)$ zu bestimmen führt im Allgemeinen zu numerischen Problemen, denn kleine Störungen in H können zu große Änderungen in den Eigenwerten führen vgl. [24, Example 3.1.2].

Lemma 5.1 ([37, Thm. 5.1], [43, Prop. 22.1.1])
Sei $H \in \mathbb{C}^{n\times n}$ derart, dass $S = \operatorname{sign}(H)$ existiert. Dann gilt:

- $S^2 = I_n$.
- *S ist diagonalisierbar und $\Lambda(S) = \{-1, 1\}$.*
- $SH = HS$.
- $\operatorname{sign}(RHR^{-1}) = R \operatorname{sign}(H) R^{-1}$ *für alle* $R \in \mathrm{GL}(n, \mathbb{C})$.
- *Falls $H \in \mathbb{R}^{n\times n}$, dann ist $S \in \mathbb{R}^{n\times n}$.*
- $\frac{1}{2}(I_n - S)$ *und* $\frac{1}{2}(I_n + S)$ *sind Projektoren auf die zu den Eigenwerten in der linken bzw. rechten Halbebene zugehörigen H-invarianten Unterräume.*
- $\operatorname{sign}(H) = \operatorname{sign}(cH)$ *für alle* $c \in \mathbb{R}^+$.

Nach Lemma 5.1 erfüllt $S = \operatorname{sign}(H)$ die Gleichung $S^2 = I_n$. Die Idee ist nun ein Newton-Verfahren anzuwenden und zu zeigen, dass das Verfahren gegen S konvergiert.

$$f(H) := H^2 - I_n$$

$$\lim_{t\to 0}\frac{1}{t}(f(H+tN)-f(H)) = \lim_{t\to 0}\frac{1}{t}((H+tN)^2 - I_n - H^2 + I_n)$$
$$= \lim_{t\to 0}\frac{1}{t}(t(HN+NH)+t^2N^2) = HN+NH$$

Nun gilt

$$\frac{||f(H+N)-f(H)-(HN+NH)||_2}{||N||_2} = \frac{||N^2||_2}{||N||_2} \leq \frac{||N||_2^2}{||N||_2} = ||N||_2 \to 0 \text{ für } N\to 0.$$

Daher ist $f'(H) \in \mathfrak{L}(\mathbb{C}^{n\times n}, \mathbb{C}^{n\times n})$ durch $N \mapsto HN+NH$ gegeben.

Damit erhalten wir

$$\begin{aligned} f'(H_k)(H_{k+1}-H_k) &= -f(H_k) \Leftrightarrow \\ H_k(H_{k+1}-H_k)+(H_{k+1}-H_k)H_k &= -H_k^2 + I_n \Leftrightarrow \\ H_kH_{k+1}+H_{k+1}H_k &= H_k^2 + I_n \Rightarrow \\ H_{k+1} &= \frac{1}{2}(H_k + H_k^{-1}). \end{aligned}$$

Die Folge ist wohldefiniert, falls $\Lambda(H_0)\cap i\mathbb{R} = \emptyset$. Sei $H_k = ZJZ^{-1}$ die Jordannormalform von H_k. Da J eine obere Dreieckmatrix ist, ist J^{-1} ebenfalls eine obere Dreicksmatrix mit

$$(J^{-1})_{i,i} = \frac{1}{J_{i,i}} \text{ für } i = 1,\ldots,n.$$

Nun ist

$$\Lambda(H_{k+1}) = \Lambda(Z(J+J^{-1})Z^{-1}) = \Lambda(J+J^{-1}) = \{\lambda + \frac{1}{\lambda} \mid \lambda \in \Lambda(H_k)\}.$$

Außerdem gilt $z \in \mathbb{C}\setminus i\mathbb{R} \Rightarrow z+\frac{1}{z} \in \mathbb{C}\setminus i\mathbb{R}$ und damit $\Lambda(H_k)\cap i\mathbb{R} = \emptyset \Rightarrow \Lambda(H_{k+1})\cap i\mathbb{R} = \emptyset$.

Satz 5.1 (Konvergenz Newton-Verfahren für sign, [37, Thm. 5.6])
Sei $H \in \mathbb{C}^{n\times n}$ mit $\Lambda(H)\cap i\mathbb{R} = \emptyset$. Dann konvergiert die Iterationsvorschrift

$$H_{k+1} = \frac{1}{2}(H_k + H_k^{-1})$$

mit $H_0 = H$ quadratisch gegen $S = \operatorname{sign}(H)$. Es gilt

$$||H_{k+1}-S||_2 \leq \frac{1}{2}||H_k^{-1}||_2||H_k-S||_2^2.$$

Für $k \geq 1$ gilt $H_k = (I_n - G_0^{2^k})^{-1}(I_n + G_0^{2^k})S$ mit $G_0 := (H-S)(H+S)^{-1}$.

Bemerkung 5.3 ([37, S.119, (5.35), (5.36), (5.37)])
Man beachte, dass wir mit Lemma 5.1 *die Iteration aus* Satz 5.1 *in jedem Schritt mit $c_k \in \mathbb{R}^+$ skalieren dürfen um die Konvergenz zu beschleunigen. Wir erhalten*

$$\begin{aligned} H_{k+1} &= \frac{1}{2}(c_kH_k + c_k^{-1}H_k^{-1}) \\ H_0 &= H \end{aligned}$$

Hierzu werden in [37] *folgende Möglichkeiten vorgeschlagen.*

- $c_k = \sqrt[n]{|\det(H_k)|}$
- $c_k = \sqrt{\frac{\rho(H_k^{-1})}{\rho(H_k)}}$
- $c_k = \sqrt{\frac{||H_k^{-1}||}{||H_k||}}$

Gegeben sei $M \in \mathrm{GL}(n,\mathbb{R}), A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}$ und $R \in \mathbb{S}^m_{++}$. Wir betrachten eine homogene verallgemeinerte algebraische Riccatigleichung

$$A^T XM + M^T XA - M^T XBR^{-1}B^T XM = 0. \tag{5.1}$$

(5.1) nennen wir verallgemeinerte algebraische Bernoulligleichung. Offensichtlich ist $X = 0$ eine Lösung von (5.1), daher folgende Definition.

Definition 5.2 ([20, Def. 2.2])
Eine Lösung von (5.1) *nennen wir stabilisierend, falls* $\Lambda(A - BR^{-1}B^T XM, M) \subset \mathbb{C}^-$.

Lemma 5.2 ([11, Prop. 2],[13, Satz 3.13])
Sei $(M^{-1}A, M^{-1}B)$ *stabilisierbar und* $\Lambda(A,M) \cap i\mathbb{R} = \emptyset$*, dann hat* (5.1) *eine eindeutige stabilisierende symmetrisch positiv semidefinite Lösung* X_0*. Es gilt* $\operatorname{rank}(X_0) = \mu$*, wobei* μ *die Anzahl der instabilen Eigenwerte von* $\Lambda(A,M)$ *(mit algebraischen Vielfachheiten gezählt) ist. Für das Spektrum gilt* $\Lambda(A - BR^{-1}B^T X_0 M, M) = (\Lambda(A,M) \cap \mathbb{C}^-) \cup -(\Lambda(A,M) \cap \mathbb{C}^+)$.

Lemma 5.2 besagt, dass unter geeigneten Voraussetzungen (5.1) eine stabilisierende Lösung X_0 besitzt. Falls die Anzahl der instabilen Eigenwerte von $\Lambda(A,M)$ klein ist, können wir die Lösung mit $X_0 = ZZ^T$ und $Z \in \mathbb{R}^{n\times\mu}$ effizient „speichern“. X_0 ist ein geeigneter Startwert für die Iteration aus Satz 2.6. Die instabilen Eigenwerte werden an der imaginären Achse gespiegelt und die stabilen Eigenwerte bleiben unberührt.
Eine wichtige Beobachtung für das weitere Vorgehen ist folgende.

Bemerkung 5.4
Sei $A \in \mathbb{R}^{n\times n}, G \in \mathbb{R}^{n\times n}$*. Wir setzen* $H := \begin{bmatrix} A & G \\ 0 & -A^T \end{bmatrix}$*. Dann gilt*

$$\Lambda(H) = \Lambda(A) \cup \Lambda(-A),$$

weil

$$\det(\lambda I_{2n} - H) = \det(\lambda I_n - A)\det(\lambda I_n + A^T) = \det(\lambda I_n - A)\det(\lambda I_n + A).$$

Das bedeutet $\Lambda(A) \cap i\mathbb{R} = \emptyset \Leftrightarrow \Lambda(H) \cap i\mathbb{R} = \emptyset$ *und* $\operatorname{sign}(H)$ *existiert.*

Satz 5.2 ([23, Ch. 7.4.3 Thm. S. 173], [43, Thm. 22.2.3])
Sei $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}, R \in \mathbb{S}^m_{++}$*,* (A,B) *stabilisierbar und* $\Lambda(A) \cap i\mathbb{R} = \emptyset$*. Weiterhin sei*

$$H := \begin{bmatrix} A & -BR^{-1}B^T \\ 0 & -A^T \end{bmatrix}$$

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} := \operatorname{sign}(H)$$

mit $W_{11}, W_{12}, W_{21}, W_{22} \in \mathbb{R}^{n\times n}$. Dann ist die eindeutige, symmetrisch positiv semidefinite, stabilisierende Lösung X_0 von

$$A^T X + XA + XBR^{-1}B^T X = 0$$

ebenfalls Lösung von

$$\begin{bmatrix} W_{12} \\ W_{22} + I_n \end{bmatrix} X = - \begin{bmatrix} W_{11} + I_n \\ W_{21} \end{bmatrix}.$$

Es gilt zusätzlich

$$\operatorname{rank}\left(\begin{bmatrix} W_{12} \\ W_{22} + I_n \end{bmatrix} \right) = n$$

Bemerkung 5.5
Die Verallgemeinerung von Satz 5.2 *ergibt sich wie folgt.*
Wir fordern $\Lambda(A, M) \cap i\mathbb{R} = \emptyset$.
In [33] *wird gezeigt, wie sich das Newton-Verfahren verallgemeinern lässt. Wir fassen die Ideen zusammen.*

$$H = \begin{bmatrix} AM^{-1} & -BR^{-1}B^T \\ 0 & -(AM^{-1})^T \end{bmatrix} = \begin{bmatrix} I_n & 0 \\ 0 & M^{-T} \end{bmatrix} \begin{bmatrix} A & -BR^{-1}B^T \\ 0 & -A^T \end{bmatrix} \begin{bmatrix} M^{-1} & 0 \\ 0 & I_n \end{bmatrix} =: E_1^{-1} \tilde{H} E_2^{-1}.$$

Für $H_{k+1} = \frac{1}{2}(H_k + H_k^{-1})$ mit $H_0 = E_1^{-1}\tilde{H}E_2^{-1}$ gilt $\lim_{k\to\infty} H_{k+1} = \operatorname{sign}(H)$.
Wir setzen $Z_k = E_1 H_k E_2$ und erhalten

$$\begin{aligned} Z_{k+1} &= E_1 H_{k+1} E_2 = E_1 \frac{1}{2}(H_k + H_k^{-1}) E_2 = E_1 \frac{1}{2}(E_1^{-1} Z_k E_2^{-1} + E_2 Z_k^{-1} E_1) E_2 \\ &= \frac{1}{2}(Z_k + E_1 E_2 Z_k^{-1} E_2 E_1), \\ Z_0 &= E_1 H_0 E_2 = \tilde{H}, \\ \lim_{k\to\infty} Z_{k+1} &= E_1 \operatorname{sign}(H) E_2 = E_1 \operatorname{sign}(E_1^{-1}\tilde{H}E_2^{-1}) E_2. \end{aligned}$$

Um die Konvergenz zu beschleunigen setzt man $Z_k = E_1 H_k E_2$ direkt in die Formeln aus Bemerkung 5.3 *ein.*
Mit

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} := \operatorname{sign}(H)$$

und

$$E_1 \operatorname{sign}(H) E_2 = E_1 \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} E_2 = \begin{bmatrix} W_{11}M & W_{12} \\ M^T W_{21} M & M^T W_{22} \end{bmatrix} =: \begin{bmatrix} \tilde{W}_{11} & \tilde{W}_{12} \\ \tilde{W}_{21} & \tilde{W}_{22} \end{bmatrix}$$

erhalten wir durch Multiplikation der ersten und zweiten (zweiten) Blockzeile von rechts (links) mit M (M^T):

$$\begin{bmatrix} W_{12} \\ W_{22} + I_n \end{bmatrix} X = - \begin{bmatrix} W_{11} + I_n \\ W_{21} \end{bmatrix} \Leftrightarrow \begin{bmatrix} W_{12} \\ M^T W_{22} + M^T \end{bmatrix} XM = - \begin{bmatrix} W_{11}M + M \\ M^T W_{21} M \end{bmatrix}$$

$$\Leftrightarrow \begin{bmatrix} \tilde{W}_{12} \\ \tilde{W}_{22} + M^T \end{bmatrix} XM = - \begin{bmatrix} \tilde{W}_{11} + M \\ \tilde{W}_{21} \end{bmatrix}.$$

Korollar 5.1
Sei $M \in \mathrm{GL}(n, \mathbb{R}), A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times m}, R \in \mathbb{S}^m_{++}$, $(M^{-1}A, M^{-1}B)$ *stabilisierbar und* $\Lambda(A, M) \cap i\mathbb{R} = \emptyset$. *Weiterhin sei*

$$H := \begin{bmatrix} A & -BR^{-1}B^T \\ 0 & -A^T \end{bmatrix}$$
$$\begin{bmatrix} \tilde{W}_{11} & \tilde{W}_{12} \\ \tilde{W}_{21} & \tilde{W}_{22} \end{bmatrix} := \begin{bmatrix} I_n & 0 \\ 0 & -M^{-T} \end{bmatrix} \operatorname{sign}(H) \begin{bmatrix} M^{-1} & 0 \\ 0 & I_n \end{bmatrix}$$

mit $\tilde{W}_{11}, \tilde{W}_{12}, \tilde{W}_{21}, \tilde{W}_{22} \in \mathbb{R}^{n \times n}$. *Dann ist die eindeutige symmetrisch positiv semidefinite stabilisierende Lösung* X_0 *von*

$$A^T XM + M^T XA + M^T XBR^{-1}B^T XM = 0$$

ebenfalls Lösung von

$$\begin{bmatrix} \tilde{W}_{12} \\ \tilde{W}_{22} + M^T \end{bmatrix} XM = - \begin{bmatrix} \tilde{W}_{11} + M \\ \tilde{W}_{21} \end{bmatrix}.$$

Es gilt zusätzlich

$$\operatorname{rank}\left(\begin{bmatrix} \tilde{W}_{12} \\ \tilde{W}_{22} + M^T \end{bmatrix} \right) = n.$$

6 Index-2 Systeme

Im Abschnitt über die stationären bzw. instationären Navier-Stokes Gleichungen lag uns ein System nichtlinearer partieller Differentialgleichungen mit der Divergenzfreiheit als Nebenbedingung vor. Wir suchten abstrakt gesehen Lösungen v, sodass $v \in \ker(\mathrm{div})$ mit $\mathrm{div} \in \mathfrak{L}(W^{1,2}(\Omega), L^2(\Omega))$ erfüllt ist. Diskretisiert man diese Gleichungen erhält man ebenfalls ein System mit der Forderung das $v \in \ker(G^T)$. Wir wollen ein LQR-Problem für solche Systeme lösen, aber Systeme mit einer algebraischen Nebenbedingung an den Zustand passen nicht direkt in das Konzept. Hierzu werden wir mittels Projektion auf den diskreten divergenzfreien Unterraum eine gewöhnliche Differentialgleichung erhalten und können die Theorie beinahe anwenden.

Definition 6.1 (Index-2 System,[35, 1.Introduction])
Sei $M \in \mathbb{S}^{n_v}_{++}, A \in \mathbb{R}^{n_v \times n_v}, G \in \mathbb{R}^{n_v \times n_p}$ *mit* $\mathrm{rank}\,(G) = n_p < n_v, B \in \mathbb{R}^{n_v \times n_u}, C \in \mathbb{R}^{n_y \times n_v}$ *und* $v_0 \in \ker(G^T)$. *Das System*

$$M\dot{v}(t) = Av(t) + Gp(t) + Bu(t) \quad \textit{für } t > 0, \tag{6.1}$$
$$0 = G^T v(t) \quad \textit{für } t \geq 0, \tag{6.2}$$
$$v(0) = v_0, \tag{6.3}$$
$$y(t) = Cv(t) \quad \textit{für } t \geq 0, \tag{6.4}$$

nennen wir Index-2 System.

Bemerkung 6.1
In Definition 6.1 *kann man sich v als Geschwindigkeit und p als Druck eines inkompressiblen Fluids vorstellen. u stellt eine Wirkung (Steuerung) am Rand eines Gebietes auf das Fluid dar, die mit B realisiert wird. Die Bedingung* $v(t) \in \ker(G^T)$ *beschreibt in einem diskreten Sinn die Divergenzfreiheit des Geschwindigkeitsfeldes. Mittels C kann man die Geschwindigkeit v an einzelnen Punkten des Gebietes „messen“.*

Wir wollen die Bedingung (6.2) eliminieren und folgen dazu [35, Ch. 3]. Aus $G^T v(t) = 0$ folgt $G^T \dot{v}(t) = 0$. Multiplikation von (6.1) mit $G^T M^{-1}$ liefert

$$\begin{aligned} 0 &= G^T \dot{v}(t) = G^T M^{-1} M \dot{v}(t) = G^T M^{-1} A v(t) + G^T M^{-1} G p(t) + G^T M^{-1} B u(t) \Leftrightarrow \\ p(t) &= -(G^T M^{-1} G)^{-1} G^T M^{-1} A v(t) - (G^T M^{-1} G)^{-1} G^T M^{-1} B u(t). \end{aligned}$$

Einsetzen in (6.1) liefert

$$\begin{aligned} M\dot{v}(t) &= (I_n - G(G^T M^{-1} G)^{-1} G^T M^{-1})(Av(t) + Bu(t)) \\ &:= \Pi(Av(t) + Bu(t)). \end{aligned}$$

Damit erhalten wir folgendes System.

$$M\dot{v}(t) = \Pi A v(t) + \Pi B u(t) \text{ für } t > 0, \tag{6.5}$$
$$v(0) = v_0, \tag{6.6}$$
$$y(t) = Cv(t) \quad \text{für } t \geq 0. \tag{6.7}$$

Korollar 6.1
Sei (v, p, u) *derart, dass* (v, p, u) (6.1), (6.2), (6.3), (6.4) *erfüllt. Dann erfüllt* (v, u) (6.5), (6.6), (6.7).

Lemma 6.1 (Eigenschaften von Π, [35, Ch. 3], [21, Lemma 6.1])
Sei $M \in \mathbb{S}^{n_v}_{++}, G \in \mathbb{R}^{n_v \times n_p}$ *mit* $\operatorname{rank}(G) = n_p < n_v$. *Dann gelten für*

$$\Pi := I_n - G(G^T M^{-1} G)^{-1} G^T M^{-1}$$

folgende Eigenschaften.
(i) $\Pi^2 = \Pi$.
(ii) $\Pi M = M \Pi^T$.
(iii) $M^{-1}\Pi = \Pi^T M^{-1}$.
(iv) $\ker(\Pi) = \operatorname{span}(\{G_{*,1}, \ldots, G_{*,n_p}\})$.
(v) $\operatorname{span}(\{\Pi_{*,1}, \ldots, \Pi_{*,n}\}) = \ker(G^T M^{-1})$.
(vi) $G^T x = 0 \Leftrightarrow \Pi^T x = x$.

Beweis.

Zu (i) :

$$\begin{aligned}\Pi^2 &= (I_n - G(G^T M^{-1} G)^{-1} G^T M^{-1})^2 \\ &= I_n - 2G(G^T M^{-1} G)^{-1} G^T M^{-1} + G(G^T M^{-1} G)^{-1} G^T M^{-1} \\ &= I_n - G(G^T M^{-1} G)^{-1} G^T M^{-1} = \Pi.\end{aligned}$$

Zu (ii) :

$$\begin{aligned}\Pi M &= M - G(G^T M^{-1} G)^{-1} G^T M^{-1} M \\ &= M(I_n - M^{-1} G(G^T M^{-1} G)^{-1} G^T) \\ &= M(I_n - M^{-T} G(G^T M^{-1} G)^{-1} G^T) = M\Pi^T.\end{aligned}$$

Zu (iii) :

$$\Pi M = M\Pi^T \Rightarrow M^{-1}\Pi M M^{-1} = M^{-1} M \Pi^T M^{-1} \Rightarrow M^{-1}\Pi = \Pi^T M^{-1}.$$

Zu (iv) :

$$0 = \Pi x \Rightarrow G(G^T M^{-1} G)^{-1} G^T M^{-1} x = x \in \operatorname{span}(\{G_{*,1}, \ldots, G_{*,n_p}\}).$$

$$\Pi x = \Pi G y = Gy - G(G^T M^{-1} G)^{-1} G^T M^{-1} G y = Gy - Gy = 0 \Rightarrow x \in \ker(\Pi).$$

Zu (v) :

$$\begin{aligned}G^T M^{-1} x &= G^T M^{-1} \Pi y \\ &= (G^T M^{-1} - G^T M^{-1} G(G^T M^{-1} G)^{-1} G^T M^{-1}) y = 0 \Rightarrow x \in \ker(G^T M^{-1}).\end{aligned}$$

$$G^T M^{-1} x = 0 \Rightarrow \Pi x = x \Rightarrow x \in \operatorname{span}(\{\Pi_{*,1}, \ldots, \Pi_{*,n}\}).$$

Zu (vi) :

$$\ker(G^T) = \operatorname{span}(\{G_{*,1}, \ldots, G_{*,n_p}\})^\perp = \ker(\Pi)^\perp = \operatorname{span}(\{{\Pi^T}_{*,1}, \ldots, {\Pi^T}_{*,n_v}\}).$$

□

Lemma 6.2
Sei (v, u) *derart, dass* (v, u) (6.5), (6.6), (6.7) *erfüllt. Dann existiert ein durch* (v, u) *eindeutig bestimmtes* p, *sodass* (v, u, p) (6.1), (6.2), (6.3), (6.4) *erfüllt.*

Beweis. Sei (v, u) derart, dass (v, u) (6.5), (6.6), (6.7) erfüllt. Der Hauptsatz der Differential- und Integralrechnung liefert folgende Integraldarstellung für v.

$$v(t) = v_0 + \int_0^t M^{-1}\Pi A v(s) + M^{-1}\Pi B u(s)\,\mathrm{d}s.$$

Mit

$$\begin{aligned}\Pi^T v(t) &= \Pi^T v_0 + \int_0^t \Pi^T M^{-1}\Pi A v(s) + \Pi^T M^{-1}\Pi B u(s)\,\mathrm{d}s \\ &= v_0 + \int_0^t M^{-1}\Pi\Pi A v(s) + M^{-1}\Pi\Pi B u(s)\,\mathrm{d}s \\ &= v_0 + \int_0^t M^{-1}\Pi A v(s) + M^{-1}\Pi B u(s)\,\mathrm{d}s \\ &= v(t)\end{aligned}$$

folgt

$$G^T v(t) = 0.$$

Mit $G^T v(t) = 0 \Leftrightarrow \Pi^T v(t) = v(t) \Rightarrow \Pi^T \dot{v}(t) = \dot{v}(t)$ erhalten wir nun

$$\begin{aligned}\Pi M\dot{v}(t) = M\Pi^T\dot{v}(t) = M\dot{v}(t) &= \Pi A v(t) + \Pi B u(t) \Rightarrow \\ \Pi(M\dot{v}(t) - Av(t) - Bu(t)) &= 0 \Rightarrow \\ M\dot{v}(t) - Av(t) - Bu(t) &\in \ker(\Pi) = \operatorname{span}(\{G_{*,1}, \ldots, G_{*,n_p}\}).\end{aligned}$$

Da $\operatorname{rank}(G) = n_p$, existiert ein durch v und u eindeutig bestimmtes p mit

$$\begin{aligned}M\dot{v}(t) - Av(t) - Bu(t) &= Gp(t) \Rightarrow \\ M\dot{v}(t) &= Av(t) + Gp(t) + Bu(t).\end{aligned}$$

Man beachte die Analogie zu Satz 3.19. Den eindeutig bestimmten Druck p „findet“ man im gewissen Sinne im Residuum. □

Bemerkung 6.2
Man beachte, dass die Bedingung (6.7) *im System* (6.5), (6.6), (6.7) *äquivalent ist zu* $y(t) = C\Pi^T v(t)$*, weil wir in* Lemma 6.2 *die Divergenzfreiheit von* v *nur unter Ausnutzung von* (6.5), (6.6) *folgern konnten.*
Diese Bemerkung wird nach dem folgenden Lemma wichtig sein.

Lemma 6.3 (weitere Eigenschaften von Π, [35, Ch. 3], [21, Lemma 6.1])
Sei $M \in \mathbb{S}^{n_v}_{++}, G \in \mathbb{R}^{n_v \times n_p}$ *mit* $\operatorname{rank}(G) = n_p < n_v$ *und* $p \in \mathbb{C}\setminus\{0\}$*. Weiterhin sei* $A \in \mathbb{R}^{n_v \times n_v}$ *eine beliebige Matrix. Dann gelten für*

$$\Pi := I_n - G(G^T M^{-1} G)^{-1} G^T M^{-1}$$

folgende Eigenschaften.

(i) $\begin{bmatrix} M & G \\ G^T & 0 \end{bmatrix} \in \mathrm{GL}(n_v + n_p, \mathbb{R})$.

(ii) Falls $-p \notin \Lambda(\Pi A, M)$, *dann gilt* $\begin{bmatrix} A+pM & G \\ G^T & 0 \end{bmatrix} \in \mathrm{GL}(n_v + n_p, \mathbb{C})$.

(iii) $\Pi^T x = y \Leftrightarrow \begin{bmatrix} M & G \\ G^T & 0 \end{bmatrix} \begin{bmatrix} y \\ \star \end{bmatrix} = \begin{bmatrix} Mx \\ 0 \end{bmatrix}$.

(iv) $\Pi x = y \Leftrightarrow \begin{bmatrix} M & G \\ G^T & 0 \end{bmatrix} \begin{bmatrix} z \\ \star \end{bmatrix} = \begin{bmatrix} x \\ 0 \end{bmatrix}$ *und* $y = Mz$.

(v) $(\Pi A + pM)x = y \in \mathrm{span}(\{\Pi_{*,1}, \ldots, \Pi_{*,n_v}\}) \Leftrightarrow\ x \in \mathrm{span}(\{{\Pi^T}_{*,1}, \ldots, {\Pi^T}_{*,n_v}\})$.

(vi) $(\Pi A + pM)x = y \in \mathrm{span}(\{\Pi_{*,1}, \ldots, \Pi_{*,n_v}\}) \Rightarrow \begin{bmatrix} A+pM & G \\ G^T & 0 \end{bmatrix} \begin{bmatrix} x \\ \star \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix}$.

(vii) $\begin{bmatrix} A+pM & G \\ G^T & 0 \end{bmatrix} \begin{bmatrix} x \\ \star \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix}$ *und* $y \in \mathrm{span}(\{\Pi_{*,1}, \ldots, \Pi_{*,n_v}\}) \Rightarrow\ (\Pi A + pM)x = y$.

(viii) Falls $y \in \mathrm{span}(\{\Pi_{*,1}, \ldots, \Pi_{*,n_v}\})$ *gilt* $\begin{bmatrix} A+pM & G \\ G^T & 0 \end{bmatrix} \begin{bmatrix} x \\ \star \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix} \Leftrightarrow\ (\Pi A + pM)x = y$.

Beweis.

Zu (i) :

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} M & G \\ G^T & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \Rightarrow x \in \ker(G^T) \text{ und } 0 = Mx + Gy \Rightarrow$$

$$\mathrm{span}(\{{\Pi^T}_{*,1}, \ldots, {\Pi^T}_{*,n_v}\}) \ni x = -M^{-1}Gy \in \ker(G^T M^{-1})^\perp = \mathrm{span}(\{{\Pi^T}_{*,1}, \ldots, {\Pi^T}_{*,n_v}\})^\perp \Rightarrow$$

$$x = 0 \text{ und } 0 = M^{-1}Gy \Rightarrow x = 0 \text{ und } y = 0, \text{ da } \mathrm{rank}(G) = n_p.$$

Zu (ii) :

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} A+pM & G \\ G^T & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \Leftrightarrow$$

$$x \in \ker(G^T) = \mathrm{span}(\{{\Pi^T}_{*,1}, \ldots, {\Pi^T}_{*,n_v}\}) \text{ und } 0 = (A + pM)x + Gy \Rightarrow$$

$$0 = \Pi((A + pM)x + Gy) = \Pi A + pM \underbrace{\Pi^T x}_{=x} + \underbrace{\Pi G}_{=0} y = (\Pi A + pM)x \Leftrightarrow$$

$$\Pi A x = -pMx \Rightarrow x = 0 \text{ und } y = 0, \text{ da } \mathrm{rank}(G) = n_p.$$

Zu (iii) :

$$\begin{bmatrix} M & G \\ G^T & 0 \end{bmatrix} \begin{bmatrix} y \\ \star \end{bmatrix} = \begin{bmatrix} Mx \\ 0 \end{bmatrix} \Leftrightarrow$$

$$y \in \ker(G^T) \text{ und } Mx = My + G\star \Leftrightarrow$$

$$y \in \mathrm{span}(\{{\Pi^T}_{*,1}, \ldots, {\Pi^T}_{*,n_v}\}) \text{ und } x = \underbrace{y}_{\Pi^T y = y} + \underbrace{M^{-1}G\star}_{\in \ker(\Pi^T)} \Leftrightarrow$$

$$\Pi^T x = y.$$

Zu (iv) :

$$\Pi x = y \Leftrightarrow \Pi^T M^{-1} x = z \text{ und } Mz = y \Leftrightarrow \begin{bmatrix} M & G \\ G^T & 0 \end{bmatrix} \begin{bmatrix} z \\ 0 \end{bmatrix} = \begin{bmatrix} x \\ 0 \end{bmatrix} \text{ und } y = Mz.$$

Zu (v) :

$$(\Pi A + pM)x = y \in \mathrm{span}(\{\Pi_{*,1}, \ldots, \Pi_{*,n_v}\}) = \ker(G^T M^{-1}) \Leftrightarrow$$

$0 = G^T M^{-1}((\Pi A + pM)x) = G^T \Pi^T M^{-1} A + pG^T x = (\underbrace{\Pi G}_{=0}) M^{-1} A + pG^T x \Leftrightarrow$

$x \in \ker(G^T) = \text{span}(\{\Pi^T{}_{*,1}, \ldots, \Pi^T{}_{*,n_v}\})$.

Zu (vi) :

Mit vorigen Punkt folgt $\Pi^T x = x$.

$0 = (\Pi A + pM)x - y = (\Pi A + pM)\Pi^T x - y = \Pi((A + pM)x - y) \Rightarrow$

$(A + pM)x - y \in \ker(\Pi) \Rightarrow$

Da $\text{rank}(G) = n_p$, existiert ein eindeutig bestimmtes Element $(-\star)$ mit

$G(-\star) = ((A + pM)x - y). \Rightarrow (A + pM)x + G\star = y.$

Zu (vii) :

$0 = \Pi((A + pM)x + G \star - y) = \Pi A x + pM \underbrace{\Pi^T x}_{=x,\ G^T x = 0} + \underbrace{\Pi G}_{=0} \star - \underbrace{\Pi y}_{=y} \Rightarrow$

$y = (\Pi A + pM)x.$

Zu $(viii)$:

Ergibt sich aus (vi) und (vii).

□

Bemerkung 6.3
Die Punkte $(iii), (iv), (viii)$ *aus* Lemma 6.3 *sind wichtig, denn den Projektor* Π *in der Praxis explizit aufzustellen ist meistens nicht möglich. Mit den genannten Punkten, lässt sich der Projektor implizit anwenden.*
Insbesondere können wir mit $(viii)$*, das geshiftete System als Sattelpunktproblem lösen, falls die rechte Seite im Bild von* Π *liegt. Die rechte Seite in der ersten Iteration des Newton-ADI-Verfahrens ist* $C^T C$ *(falls* $X_0 = 0$*). Da wir nach* Bemerkung 6.2 *anstatt* C *auch* $C\Pi^T$ *verwenden können, ist diese Bedingung erfüllt.*

Wir wollen ein LQR-Problem für das System (6.5), (6.6) und $y(t) = C\Pi^T v(t)$ lösen. Wir wählen dazu $Q = I_{n_y}$ und $R = I_{n_u}$ und erhalten folgende verallgemeinerte algebraische Riccatigleichung

$$0 = (C\Pi^T)^T(C\Pi^T) + (\Pi A)^T XM + MX(\Pi A) - MX(\Pi B)(\Pi B)^T XM \tag{6.8}$$
$$= \Pi C^T C \Pi^T + A^T \Pi^T XM + MX\Pi A - MX\Pi BB^T \Pi^T XM. \tag{6.9}$$

Wir wollen das Newton-Verfahren Verfahren anwenden. Das Newton-Verfahren ergibt sich dann wie folgt.

Algorithmus 7 Newton-Verfahren für verallgemeinerte algebraische Riccatigleichungen (6.8)

Eingabe: M, A, G, B, C.
Ausgabe: X_∞ löst (6.8) und $K_\infty = MX_\infty \Pi B$.

$K_0 = []$
$\tilde{C}^T = \Pi C^T$ mit Lemma 6.3 (iv)
for $k = 1, \ldots$ **do**
  Löse mit Algorithmus 6.

$$(\Pi(A - BK_{k-1}^T))^T X_k M + MX_k(\Pi(A - BK_{k-1}^T)) = -\begin{bmatrix} \tilde{C} \\ K_{k-1}^T \end{bmatrix}^T \begin{bmatrix} \tilde{C} \\ K_{k-1}^T \end{bmatrix}$$

  $K_k = MX_kB$
end for

Bemerkung 6.4
Für $k = 1$ (in der obigen Newton-Iteration) erhalten wir

$$\begin{bmatrix} \tilde{C} \\ K_{k-1}^T \end{bmatrix}^T = \tilde{C}^T = \Pi C^T.$$

Wir müssen im Algorithmus 6 *folgendes System lösen. Da die Spalten der rechten Seite im Bild von* Π *sind, sind auch die Spalten der linken Seite im Bild von* Π. *Wir benutzen* Lemma 6.3 *(v), (viii).*

$$\begin{aligned} ((\Pi A)^T + p_1 M)V_1 = W_0 = \Pi C^T \Leftrightarrow \\ \Pi((\Pi A)^T + p_1 M)V_1 = W_0 = \Pi C^T \Leftrightarrow \\ (\Pi A^T + p_1 M)\Pi^T V_1 = W_0 = \Pi C^T \end{aligned}$$

Eine Lösung V_1 mit der Eigenschaft $\Pi^T V_1 = V_1$ erhalten wir nach Lemma 6.3 *(v), (vii) durch Lösen von folgendem System.*

$$\begin{aligned} \begin{bmatrix} A^T + p_1 M & G \\ G^T & 0 \end{bmatrix} \begin{bmatrix} V_1 \\ \star \end{bmatrix} = \begin{bmatrix} W_0 \\ 0 \end{bmatrix} \Leftrightarrow \\ \left(\begin{bmatrix} A + p_1 M & G \\ G^T & 0 \end{bmatrix} \right)^T \begin{bmatrix} V_1 \\ \star \end{bmatrix} = \begin{bmatrix} W_0 \\ 0 \end{bmatrix}. \end{aligned}$$

Wir folgen weiter Algorithmus 6. *Nach dem Lösen gilt $\Pi^T V_1 = V_1$ und wegen $MV_1 = M\Pi^T V_1 = \Pi M V_1$ ist W_1 ebenso im Bild von Π. Damit lässt sich wieder* Lemma 6.3 *(v), (viii) mit obiger Argumentation benutzen und wir sehen, dass während der Iteration stets $\Pi W_k = W_k$, $\Pi^T V_k = V_k$ und damit $\Pi^T Z_k = Z_k$ gilt.*
Nach Terminierung von Algorithmus 6, *erhalten wir $X_1 = ZZ^T$ und es gilt $\Pi^T X_1 \Pi = X_1$. Für die Feedbackmatrix in der Newton-Iteration gilt $K_1 = MX_1B = M\Pi^T X_1 \Pi B = \Pi M X_1 B$. Nun ist $k = 2$ (in der Newton-Iteration) und wir rufen wieder Algorithmus* 6 *auf.*

$$\begin{bmatrix} \tilde{C} \\ K_{k-1}^T \end{bmatrix}^T = \begin{bmatrix} \Pi C^T & K_1 \end{bmatrix} = \begin{bmatrix} \Pi C^T & \Pi M X_1 B \end{bmatrix}$$

$$= W_0.$$

Da die rechte Seite wieder im Bild von Π *ist, ist auch die linke Seite im Bild von* Π. *Wir benutzen wieder* Lemma 6.3 *(v), (viii).*

$$\begin{aligned}((\Pi(A - BK_1)^T + p_1 M)V_1 = W_0 \Leftrightarrow \\ \Pi((\Pi A - BK_1)^T + p_1 M)V_1 = W_0 \Leftrightarrow \\ (\Pi(A^T - K_1^T B^T) + p_1 M)\Pi^T V_1 = W_0\end{aligned}$$

Eine Lösung V_1 *mit der Eigenschaft* $\Pi^T V_1 = V_1$ *erhalten wir nach* Lemma 6.3 *(v), (vii) durch Lösen von folgendem System.*

$$\begin{aligned}\left(\begin{bmatrix} A^T - K_1^T B^T + p_1 M & G \\ G^T & 0 \end{bmatrix}\right)\begin{bmatrix} V_1 \\ \star \end{bmatrix} = \begin{bmatrix} W_0 \\ 0 \end{bmatrix} \Leftrightarrow \\ \left(\begin{bmatrix} A + p_1 M & G \\ G^T & 0 \end{bmatrix} - \begin{bmatrix} B \\ 0 \end{bmatrix}\begin{bmatrix} K_1 & 0 \end{bmatrix}\right)^T \begin{bmatrix} V_1 \\ \star \end{bmatrix} = \begin{bmatrix} W_0 \\ 0 \end{bmatrix}.\end{aligned}$$

Nun wiederholen sich im Wesentlichen die Schritte.

Bemerkung 6.5
Die verallgemeinerte algebraische Riccatigleichung (6.8) *hat keine eindeutige, symmetrisch positiv semidefinite, stabilisierende Lösung.*
Sei $X_\infty \in \mathbb{S}_+^{n_v}$ *eine Lösung von* (6.8) *mit der Eigenschaft, dass* $\Lambda(\Pi A - \Pi B B^T \Pi^T X_\infty M, M) \subset \mathbb{C}^-$. *Nun wähle* $Z \in \ker(\Pi^T) \setminus \{0\}$ *und setze* $\tilde{X} = ZZ^T \in \mathbb{S}_+^{n_v} \setminus \{0\}$. *Es gilt* $\Pi^T \tilde{X} = \Pi^T Z Z^T = 0$ *und ebenfalls* $\tilde{X}\Pi = ZZ^T\Pi = Z(\Pi^T Z)^T = 0$. *Die Matrix* $X_\infty + \tilde{X} \in \mathbb{S}_+^{n_v}$ *erfüllt* (6.8) *und* $\Lambda(\Pi A - \Pi B B^T \Pi^T (X_\infty + \tilde{X})M, M) = \Lambda(\Pi A - \Pi B B^T \Pi^T X_\infty M, M)$.
Ein Problem ist, dass $(M^{-1}\Pi A, M^{-1}\Pi B B^T \Pi^T)$ *nicht stabilisierbar ist.*
Sei $x \in \ker(\Pi^T) \setminus \{0\}$ *und* $y := Mx \neq 0$.

$$y^T M^{-1}\Pi A = x^T M^T M^{-1}\Pi A = x^T M M^{-1}\Pi A = x^T \Pi A = (\Pi^T x)^T A = 0 y^T.$$

Damit ist y *ein Links-Eigenvektor zum Eigenwert 0. Es gilt*

$$\begin{aligned}y^T M^{-1}\Pi B B^T \Pi^T &= x^T M^T M^{-1}\Pi B B^T \Pi^T = x^T M M^{-1}\Pi B B^T \Pi^T \\ &= x^T \Pi B B^T \Pi^T = (\Pi^T x)^T B B^T \Pi^T = 0.\end{aligned}$$

Nach Lemma 2.4 *ist die Stabilisierbarkeit nicht gegeben und* Korrolar 2.1 *eigentlich nicht anwendbar.*
Das Beispiel zeigt auch, dass es keine eindeutige symmetrische Lösung gibt. Das liegt im Wesentlichen daran, dass $\ker(\Pi^T)$ *nichttrivial ist. Wir nehmen an, dass dieser technische Umstand keine weiteren Schwierigkeiten bereitet, denn* $v \in \operatorname{span}(\{{\Pi^T}_{*,1}, \ldots, {\Pi^T}_{*,n_v}\})$ *und* $\ker(\Pi^T) \cap \operatorname{span}(\{{\Pi^T}_{*,1}, \ldots, {\Pi^T}_{*,n_v}\}) = \{0\}$.
Um ein Eindeutigkeitsresultat zu erhalten, könnte man folgendes versuchen. Wir gehen zum Faktorraum über und erhalten mit dem Homomorphiesatz

$$\mathbb{R}^{n_v}/(\ker(\Pi^T)) \cong \operatorname{span}(\{{\Pi^T}_{*,1}, \ldots, {\Pi^T}_{*,n_v}\}) \cong \mathbb{R}^{n_v - n_p}.$$

Dann könnte man versuchen die Theorie mit der Isomorphie „rüber zu retten“ und mit der Stetigkeit der Isomorphie auch Konvergenzaussagen zu übertragen. Eindeutigkeit ist dann im Sinne der Äquivalenzklasse zu verstehen und numerisch reicht es einen Vertreter einer

geeigneten Äquivalenzklasse zu finden. Für die Spalten des Niedrigrangfaktors der Lösung möchte man gerne $Z_{*,i} \in \mathbb{R}^{n_v}/(\ker(\Pi^T))$ *und man unterscheidet bis auf additive Vielfache von* $\ker(\Pi^T)$ *die Spalten des Lösungsfaktors nicht mehr.*

7 Implementierung

Für die gesamte Implementierung wurde die FEniCS Version 1.5.0 bzw. 1.6.0dev genutzt.

7.1 Generierung der Rechengebiete

Zur Generierung der Rechengebiete wurde das Python Modul mshr verwendet. mshr stellt eine Erweiterung für die Finite-Elemente Bibliothek FEniCS [48] dar. Dieses Modul wird mit den neueren FEniCS Versionen meistens mitgeliefert. Mit Hilfe der Klassen Circle und Rectangle sowie den überladenen Operatoren + (Vereinigung) und - (Mengendifferenz) lassen sich vielfältige Rechengebiete erzeugen.

7.2 Diskretisierung der Rechengebiete und Definition der Randstücke

Nachdem die Rechengebiete definiert wurden, lässt sich mit generate_mesh aus dem Modul mshr eine Instanz der Klasse Mesh erzeugen. Die Klasse Mesh gehört zum FEniCS Paket und eine Instanz dieser Klasse besteht im Wesentlichen aus Elementen (Face). Elemente bestehen aus Kanten (Edge). Kanten bestehen aus Punkten (Vertex). Eine Instanz der Klasse Mesh stellt eine Triangulierung des Rechengebietes Ω dar.
Um Randstücke zu definieren, kann man eine Unterklasse der Klasse SubDomain erstellen. Die Methode inside wird überschrieben, um das Randstück zu definieren. FEniCS ordnet eine Entität der Unterklasse zu, falls für jede Subentitäten der Entität die Methode inside True zurückgibt. In unserem Fall heißt dass, eine Kante (Edge) e gehört genau dann zum definierten Randstück, wenn für alle Punkte (Vertex) v, die inzident zu e sind, die Methode inside den Wert True zurückgibt. Durch Überschreiben der Methode snap können „gekrümmte“ Randstücke besser dargestellt werden, da bei Verfeinerung neue Randknoten, mittels der in snap implementierten Transformation, auf dem „gekrümmten“ Rand „geschoben“ werden.

```
class BallProjection(SubDomain):
"""Ball Projection, special class for projection curved boundary in refinement"""
    thresh = GAMMA_BALL_PROJECTION_THRESHOLD

def inside(self, x, on_boundary):
    r = np.sqrt((x[0] - CIRCLE["x0"]) ** 2 + (x[1] - CIRCLE["x1"]) ** 2)
    return r < self.thresh * CIRCLE["r"]

def snap(self, x):
    r = np.sqrt((x[0] - CIRCLE["x0"]) ** 2 + (x[1] - CIRCLE["x1"]) ** 2)
    if r < self.thresh * CIRCLE["r"]:
    x[0] = CIRCLE["x0"] + (CIRCLE["r"] / r) * (x[0] - CIRCLE["x0"])
    x[1] = CIRCLE["x1"] + (CIRCLE["r"] / r) * (x[1] - CIRCLE["x1"])
```

Quelltextausschnitt 7.1: Definition des Randes einer Kugel

Für jedes Randstück definiert man eine Klasse wie oben beschrieben.
Eine wichtige Klasse heißt `MeshFunction`. Instanzen der Klasse `MeshFunction` kann man sich wie Funktionen im mathematischen Sinne vorstellen. Eine Instanz der Klasse `MeshFunction` kann als „Definitionsbereich“ die Menge der Punkte, die Menge der Kanten oder die Menge der Elemente einer Instanz der Klasse `Mesh` besitzen. Für den „Wertebereich“ bestehen die Möglichkeiten `int`, `size_t`, `uint`, `double` oder `bool`. Für unsere Zwecke erzeugen wir eine Instanz der Klasse `MeshFunction`, die auf den Kanten definiert ist und als Wertebereich wählen wir `size_t`. Wir können diese Instanz benutzen um Kanten zu indizieren. Kanten die zum gleichen Randstück gehören, bekommen gleiche Funktionswerte und innere Kanten versehen wir mit einem Standardwert.
Eine nützlicher Mechanismus ist, dass man Instanzen der definierten Randstücke benutzen kann, um eine Instanz der Klasse `MeshFunction` mit den benötigten Funktionswerten zu belegen. Hierzu nutzt man die `mark` Methode der Unterklasse in der das gewollte Randstück definiert ist. Die `mark` Methode ist von der Oberklasse `SubDomain` geerbt worden.

```
def _buildboundaryfunction(self):
    """Mark boundary parts"""

    self.boundaryfunction =
    MeshFunction("size_t", self.mesh, self.mesh.topology().dim() - 1)

    # init all edges with zero
    self.boundaryfunction.set_all(self.const.GAMMA_INNER_INDICES)

    # mark edges
    GammaLeft().mark(self.boundaryfunction, self.const.GAMMA_LEFT_INDICES)
    GammaLower().mark(self.boundaryfunction, self.const.GAMMA_LOWER_INDICES)
    GammaRight().mark(self.boundaryfunction, self.const.GAMMA_RIGHT_INDICES)
    GammaUpper().mark(self.boundaryfunction, self.const.GAMMA_UPPER_INDICES)
    GammaBall().mark(self.boundaryfunction, self.const.GAMMA_BALL_INDICES)
    ...
```

Quelltextausschnitt 7.2: Definition einer `MeshFunction`

Jetzt erzeugt man mehrere Verfeinerungen der Triangulierung des Rechengebietes und berechnet die zugehörigen Instanzen der Klasse `MeshFunction`. Abschließend ist es sinnvoll die Instanzen der Klasse `Mesh` und die zugehörigen Instanzen der Klasse `MeshFunction` persistent im `xml`-Format zu speichern. Bei Bedarf kann man die Objekte wieder laden.
Die Instanzen von `Mesh` und `MeshFunction` lassen sich auch im `pvd`-Format speichern und damit kann man die Instanzen mit `Paraview` [36] visualisieren. Dies ist für Instanzen der Klasse `MeshFunction` hilfreich, da man mit Hilfe der Visualisierung kontrollieren kann, ob alle Randstücke wie gewollt definiert worden sind.

7.3 Lösung der stationären Navier-Stokes Gleichungen

Zur numerischen Lösung der stationären Navier-Stokes Gleichungen mit gemischten Randbedingungen, lädt man die persistent gespeicherte Instanz der Klasse `Mesh` und die zugehörige Instanz der Klasse `MeshFunction` aus den `xml`-Dateien.
Zur Approximation der Lösung wählen wir Taylor-Hood Elemente ($\mathcal{P}_2 - \mathcal{P}_1$). Diese erfüllen die inf-sup-Bedingung [28, §7. Finite Elemente für das Stokes-Problem, S. 161].
Man definiert nun den diskreten Funktionenraum für die Geschwindigkeit `V` und für den Druck `Q` und stellt die schwache Formulierung auf.
In einem abstrakten Sinn ist ein Nullstellenproblem gegeben. Eine Nullstelle `w_s`=(`v_s`,`p_s`)

such man im Raum V×Q. Mit Hilfe der Instanz der Klasse MeshFunction und der Klasse DirichletBC lassen sich Dirichletrandbedingungen auf den einzelnen Randstücken vorgeben. In unseren Beispielen sind stets auf einem Teilstück des Randes homogene Dirchletdaten („no-slip“), auf einem Teilstück ein Einflussprofil $g \neq 0$ und auf einem Teilstück des Randes schreiben wir die Ausflussbedingung vor.
FEniCS bietet die Klasse NonlinearVariationalProblem an um alle Problemdaten in einer Instanz dieser Klasse zu speichern und mit Hilfe der Klasse NonlinearVariationalSolver lässt sich das Problem lösen. Intern nutzt FEniCS ein Newton-Verfahren aus der Bibliothek PETSc[6, 7]. Newton-Verfahren benötigen meistens „gute“ Startwerte. Hierzu löst man erst das Problem für große Viskositäten ν mit (v0,p0)=$(0,0)$ als Startwert, speichert die Lösung (v_s,p_s) im xml-Format, verringert die Viskosität ν und benutzt die letzte Lösung (v_s,p_s) als neuen Startwert (v0,p0)=(v_s,p_s). Dies hat sich als hilfreich herausgestellt, um „gute“ Startwerte (v0,p0) zu erhalten.

```
...
# load mesh and meshfunction, define function spaces
self.mesh = Mesh(self.const.MESH_XML(ref))
self.V = VectorFunctionSpace(self.mesh, self.const.V, self.const.V_DIM)
self.Q = FunctionSpace(self.mesh, self.const.Q, self.const.Q_DIM)
self.W = self.V*self.Q
self.boundaryfunction = MeshFunction("size_t", self.mesh, const.BOUNDARY_XML(ref))
...
# build weak formulation
a1 = inner(grad(u) * u, v) * dx
a2 = self.nu * inner(grad(u), grad(v)) * dx
a3 = -1 * p * div(v) * dx
cond = -1 * div(u) * q * dx
F = a1 + a2 + a3 + cond
...
# define and collect boundary conditions
bc = [
DirichletBC(W.sub(0), const.STATIONARY_UIN, boundaryfunction, GAMMA_LEFT_INDICES),
DirichletBC(W.sub(0), noslip, boundaryfunction, GAMMA_LOWER_INDICES),
...]

# build derivative
dw = TrialFunction(self.W)
dF = derivative(F, w, dw)

# solve the problem
nsproblem = NonlinearVariationalProblem(F, w, bc, dF)
solver = NonlinearVariationalSolver(nsproblem)
solver.solve()

# split w
(u, p) = w.split(deepcopy=True)

...
```

Quelltextausschnitt 7.3: Newton-Verfahren für stationären Navier-Stokes Gleichungen

riccati gleichung

7.4 Assemblieren der Systemmatrizen für das LQR-Problem

Wir folgen [9] und fordern, dass $\partial\Omega = \Gamma_{in} \dot\cup \Gamma_h \dot\cup \Gamma_{out} \dot\cup \Gamma_{ctrl}$ und $\Gamma_{ctrl} = \dot{\bigcup_{i=1}^{n_u}} \Gamma_{ctrl_i}$. Angenommen wir haben eine Lösung (v_s, p_s) vom stationären Navier-Stokes Problem mit gemischten Randbedingungen

$$\begin{aligned}
-\nu\Delta v_s + (v_s \cdot \nabla)v_s + \nabla p_s &= 0 && \text{in } \Omega,\\
\operatorname{div} v_s &= 0 && \text{in } \Omega,\\
v_s &= g && \text{auf } \Gamma_{in},\\
v_s &= 0 && \text{auf } \Gamma_{ctrl},\\
v_s &= 0 && \text{auf } \Gamma_h,\\
\nu\frac{\partial v_s}{\partial n} &= p_s n && \text{auf } \Gamma_{out}.
\end{aligned}$$

Nun betrachten wir das System

$$\begin{aligned}
\dot v_\delta - \nu\Delta v_\delta + (v_\delta \cdot \nabla)v_s + (v_s \cdot \nabla)v_\delta + (v_\delta \cdot \nabla)v_\delta + \nabla p_\delta &= 0 && \text{in } \Omega \times (0,T),\\
\operatorname{div} v_\delta &= 0 && \text{in } \Omega \times [0,T),\\
v_\delta &= 0 && \text{auf } \Gamma_{in} \times [0,T),\\
v_\delta &= \mathcal{B}u && \text{auf } \Gamma_{ctrl} \times [0,T),\\
v_\delta &= 0 && \text{auf } \Gamma_h \times [0,T),\\
\nu\frac{\partial v_\delta}{\partial n} &= p_\delta n && \text{auf } \Gamma_{out} \times [0,T),\\
v_\delta(0) &= v_{\varepsilon_{pertub}} && \text{in } \Omega.
\end{aligned}$$

Setzt man nun $(v,p) = (v_\delta + v_s, p_\delta + p_s)$, dann erfüllt (v,u) das folgende System

$$\begin{aligned}
\underbrace{\dot v_\delta + \dot v_s}_{=\dot v,\ \dot v_s = 0} - \nu\Delta(\underbrace{v_\delta + v_s}_{=v}) + ((\underbrace{v_\delta + v_s}_{=v}) \cdot \nabla)(\underbrace{v_\delta + v_s}_{=v}) + \nabla(\underbrace{p_\delta + p_s}_{=p}) &= 0 && \text{in } \Omega \times (0,T),\\
\operatorname{div} v &= 0 && \text{in } \Omega \times [0,T),\\
v &= g && \text{auf } \Gamma_{in} \times [0,T),\\
v &= \mathcal{B}u && \text{auf } \Gamma_{ctrl} \times [0,T),\\
v &= 0 && \text{auf } \Gamma_h \times [0,T),\\
\nu\frac{\partial v}{\partial n} &= pn && \text{auf } \Gamma_{out} \times [0,T),\\
v(0) &= v_s + v_{\varepsilon_{pertub}} && \text{in } \Omega.
\end{aligned}$$

(v_s, p_s) ist das „Steuerziel“, gegen das man (v,p) mit u steuern will. Der Differenzzustand (v_δ, p_δ) soll möglichst „klein“ werden. Am Rand Γ_{in} haben wir Dirichletranddaten g. Am Rand Γ_{out} haben wir eine Ausflussbedingung. Der Rand Γ_{ctrl} ist für den Randeingriff und setzt sich aus n_u Teilrandstücken zusammen. Der Anfangswert $v_{\varepsilon_{pertub}}$ von v_δ ist als eine anfängliche Störung oder Abweichung von v zu v_s zu verstehen.
Wir stellen die schwache Formulierung für das zweite System auf, diskretisieren mit Taylor-

Hood Elementen ($\mathcal{P}_2 - \mathcal{P}_1$) und erhalten

$$\begin{aligned}
\int_\Omega v_\delta w \,\mathrm{d}x &\longrightarrow M v_\delta, \\
\nu \int_\Omega \nabla v_\delta : \nabla w \,\mathrm{d}x &\longrightarrow S v_\delta, \\
\int_\Omega (v_\delta \cdot \nabla) v_s \cdot w \,\mathrm{d}x &\longrightarrow R v_\delta, \\
\int_\Omega (v_s \cdot \nabla) v_\delta \cdot w \,\mathrm{d}x &\longrightarrow K v_\delta, \\
\int_\Omega p_\delta \operatorname{div} w \,\mathrm{d}x &\longrightarrow G p_\delta, \\
\int_\Omega v_\delta \cdot \nabla q \,\mathrm{d}x &\longrightarrow G^T v_\delta, \\
\int_\Omega (v_\delta \cdot \nabla) v_\delta \cdot w \,\mathrm{d}x &:= N(u_\delta, u_\delta).
\end{aligned}$$

Die Bedingung $v_\delta = \mathcal{B}u$ ersetzen wir durch $v_\delta = \mathcal{B}u + \varepsilon_{pen}(p_\delta n - \nu \frac{\partial v_\delta}{\partial n})$ und ε_{pen} wählen wir klein.

$$\int_{\partial\Gamma_{ctrl}} (p_\delta n - \nu \frac{\partial v_\delta}{\partial n}) \cdot w \,\mathrm{d}x = \int_{\partial\Gamma_{ctrl}} \frac{1}{\varepsilon_{pen}} (v_\delta - \mathcal{B}u) \cdot w \,\mathrm{d}x =$$
$$\int_{\partial\Gamma_{ctrl}} \frac{1}{\varepsilon_{pen}} v_\delta \cdot w \,\mathrm{d}x + \int_{\partial\Gamma_{ctrl}} -\mathcal{B}u \cdot w \,\mathrm{d}x.$$

Wir assemblieren auf jedem Rand Γ_{ctrl_i} einzeln.

$$\begin{aligned}
\int_{\partial\Gamma_{ctrl_i}} \frac{1}{\varepsilon_{pen}} v_\delta \cdot w \,\mathrm{d}x &\longrightarrow M_{\Gamma_{ctrl_i}} v_\delta \text{ für } i = 1, \ldots, n_u, \\
-\int_{\partial\Gamma_{ctrl_i}} \frac{1}{\varepsilon_{pen}} \mathcal{B}u \cdot w \,\mathrm{d}x &\longrightarrow B_{\Gamma_{ctrl_i}} u \text{ für } i = 1, \ldots, n_u.
\end{aligned}$$

Zur Assemblierung mit `FEniCS` lädt man die stationäre Lösung (`v_s`,`p_s`), setzt die Funktionenräume `V` und `Q`, stellt die schwache Formulierung auf und nutzt `assemble`. Nun setzen wir

$$\begin{aligned}
M_{ctrl} &:= \sum_{i=1}^{n_u} M_{ctrl_i}, \\
B &:= \left[B_{\Gamma_{ctrl_1}}, \ldots, B_{\Gamma_{ctrl_{n_u}}} \right], \\
A &:= -(S + R + K + M_{ctrl}).
\end{aligned}$$

Wir fassen $v_{\varepsilon_{pertub}}$ als diskret auf, vernachlässigen den nichtlinearen Term N und erhalten

$$\begin{aligned} M\dot{v}_\delta(t) &= Av_\delta(t) + Gp_\delta(t) + Bu(t), \\ 0 &= G^T v_\delta(t), \\ 0 &= v_{\varepsilon_{pertub}}, \\ y(t) &= C(p_1, \ldots, p_m)v_\delta. \end{aligned}$$

Da wir Taylor-Hood Elemente $(\mathcal{P}_2 - \mathcal{P}_1)$ gewählt haben, hat G hat vollen Rang. Die Matrix M ist symmetrisch positiv definit. In den Beispielen in Abschnitt 8.1 und Abschnitt 8.2 wollen wir zu dem obigen Index-2 System ein LQR-Problem lösen. Die Matrizen Q und R wählen für das Kostenfunktional als Einheitsmatrizen.
Für die Matrix C gehen wir wie folgt vor. Für jeden Punkt $p_i \in \mathbb{R}^2$, bestimme man das Dreieck e_{min} der Triangulierung von Ω, bei dem der Abstand vom Mittelpunkt zu p_i am kleinsten ist. Dann bestimmt man die Indices der zugehörigen Freiheitsgrade zu e_{min} und setzt in der i-ten Zeile von $C(p_1, \ldots, p_m)$ die zugehörigen Einträge auf 1.

```
def _buildC(self):

    points = [Point(*x) for x in self.const.ASSEMBLER_OBSERVER_POINTS]
    dists = dict([(p, float("inf")) for p in points])
    idxs = dict([(p, 0) for p in points])

    # find cell indices with midpoint has minimal distance
    for c in cells(self.mesh):
        for p in points:
            cdist = c.midpoint().distance(p)
            if cdist < dists[p]:
                dists[p] = cdist
                idxs[p] = c.index()

    # collect vertical dofs of cells with midpoint has minimal distance
    verticaldofs = np.empty(0, dtype=np.int64)
    for idx in idxs.values():
        verticaldofs = np.union1d(verticaldofs,
        self.V.sub(1).dofmap().cell_dofs(idx))

    # build matrix
    C = scsp.lil_matrix((verticaldofs.size, self.V.dim()))
    j = 0
    for i in verticaldofs:
        C[j, i] = 1
        j += 1

    return C.tocsc()
```

Quelltextausschnitt 7.4: Aufstellen der Matrix C

7.5 Simulation

Das Vorwärtsproblem dient der Simulation der Strömung ohne Steuerung u. Hierzu setzen wir einfach $\Gamma_{ctrl} = \emptyset$ und die Matrizen M_{ctrl} und B aus dem letzten Abschnitt entfallen. Auf dem Randstücken, die zu Γ_{ctrl} gehörten, fordern wir homogene Dirichletrandbedingungen.

Der Term N wird jetzt nicht vernachlässigt. Wir diskretisieren in der Zeit und erhalten

$$\begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v_\delta^{k+1} - v_\delta^k \\ p_\delta^{k+1} - p_\delta^k \end{bmatrix} = \Delta t \begin{bmatrix} A & G \\ G^T & 0 \end{bmatrix} \begin{bmatrix} v_\delta^{k+1} \\ p_\delta^{k+1} \end{bmatrix} - \Delta t \begin{bmatrix} N(v_\delta^{k+1}, v_\delta^{k+1}) \\ 0 \end{bmatrix} \Leftrightarrow$$
$$\begin{bmatrix} M - \Delta t A & -\Delta t G \\ -\Delta t G^T & 0 \end{bmatrix} \begin{bmatrix} v_\delta^{k+1} \\ p_\delta^{k+1} \end{bmatrix} + \Delta t \begin{bmatrix} N(v_\delta^{k+1}, v_\delta^{k+1}) \\ 0 \end{bmatrix} = \begin{bmatrix} M \\ 0 \end{bmatrix} v_\delta^k.$$

Das nichtlineare Gleichungssystem lösen wir mit einer Art Fixpunktiteration.

Setze

$$\begin{bmatrix} v_\delta^{k,0} \\ p_\delta^{k,0} \end{bmatrix} = \begin{bmatrix} v_\delta^k \\ p_\delta^k \end{bmatrix}.$$

Löse die linearen Gleichungssysteme

$$\begin{bmatrix} M - \Delta t A & -\Delta t G \\ -\Delta t G^T & 0 \end{bmatrix} \begin{bmatrix} v_\delta^{k,l+1} \\ p_\delta^{k,l+1} \end{bmatrix} = \begin{bmatrix} M \\ 0 \end{bmatrix} v_\delta^k - \Delta t \begin{bmatrix} N(v_\delta^{k,l}, v_\delta^{k,l}) \\ 0 \end{bmatrix},$$

falls

$$|| \begin{bmatrix} M - \Delta t A & -\Delta t G \\ -\Delta t G^T & 0 \end{bmatrix} \begin{bmatrix} v_\delta^{k,l+1} \\ p_\delta^{k,l+1} \end{bmatrix} + \Delta t \begin{bmatrix} N(v_\delta^{k,l+1}, v_\delta^{k,l+1}) \\ 0 \end{bmatrix} - \begin{bmatrix} M \\ 0 \end{bmatrix} v_\delta^k ||_2$$

genügend klein, setze

$$\begin{bmatrix} v_\delta^{k+1} \\ p_\delta^{k+1} \end{bmatrix} = \begin{bmatrix} v_\delta^{k,l+1} \\ p_\delta^{k,l+1} \end{bmatrix}.$$

Da wir kleine Zeitschrittweiten Δt verwenden, ist die Anzahl der Zwischenschritte meist gering und der Vektor $N(v_\delta^{k,l}, v_\delta^{k,l})$ muss nicht zu häufig assembliert werden. Die linearen Gleichungssysteme lösen wir mit einer Instanz der `LUSolver` Klasse. Im Wesentlichen nutzt `FEniCS` hierzu `PETSc`. Es hat sich auch als performanter herausgestellt, die linearen Gleichungssysteme mit `FEniCS` und der `PETSc` Schnittstelle zu lösen, anstatt die Matrizen aus `FEniCS` nach `SciPy` bzw. `NumPy` [38] zu exportieren und dann die Funktion `splu` aus `SciPy` für die LU-Zerlegung zu nutzen. `SciPy` stellt mit `splu` eine Schnittstelle zu `SuperLU` [47] her. Zur Simulation mit optimaler Steuerung verwenden wir in Verbindung mit Lemma 2.7 das gleiche Schema.

```
def assembleN(self):
    (w_test, p_test) = TestFunctions(self.W)
    (u, p) = self.up.split()
    self.N = assemble(inner(w_test, grad(u) * u)*dx)
    [bcup.apply(self.N) for bcup in self.bcups]
```

Quelltextausschnitt 7.5: Assemblieren von N

```
def correction(self, Msys_liftup):
    for i in range(self.const.LINEARIZED_SIM_C_STEPS):
        # assemble nonlinear right hand side term and solve system
        self.assembleN()
        self.solver.solve(self.up_k.vector(), -self.dt*self.N+Msys_liftup)
        self.up.assign(self.up_k)

        # compute residual
        if (i % self.const.LINEARIZED_SIM_C_RES_MOD) == 0:
            self.assembleN()
            residual = self.Msys_ode*self.up.vector()+self.dt*self.N-Msys_liftup
            res = np.linalg.norm(residual.array())

        if np.isnan(res):
            raise ValueError('nan during computation')

        if res < self.const.LINEARIZED_SIM_C_RES:
            break
```

Quelltextausschnitt 7.6: Zwischeniteration zur Lösung des nichtlinearen Gleichungssystems

7.6 Lösen der Bernoulligleichung zur Stabilisierung

Falls ν zu klein ist, hat

$$\left(\begin{bmatrix} A & G \\ G^T & 0 \end{bmatrix}, \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \right)$$

instabile Eigenwerte und um das Newton-Verfahren für die verallgemeinerte algebraische Riccatigleichung anzuwenden ist ein stabilisierendes X_0 bzw. K_0 notwendig.
Hierzu nutzt man die Funktion `eigs` aus `SciPy` um die Eigenwerte und die zugehörigen Links- und Rechtseigenvektoren mit dem größten Realteil zu bestimmen. `SciPy` stellt mit `eigs` eine Schnittstelle zu `ARPACK` [45] her. Nun konkateniert man die Links- bzw. Rechtseigenvektoren zu den Eigenwerten mit positiven Realteil spaltenweise aneinander, projiziert das obige Matrixpaar [61] und löst dann eine algebraische Bernoulligleichung mit der sign-Funktion. Zur Implementierungshilfe konnte hier auf einen am Institut vorhandenen Quelltext in `Matlab` zurückgegriffen werden. Das Verfahren musste nach `SciPy` bzw. `NumPy` übertragen werden, um es von `Python` aus direkt aufzurufen. Das Lösen der Bernoulligleichung bereite keine weiteren Probleme. Die Schwierigkeit ist die Eigenvektoren zu den Eigenwerten zu finden mit dem größten Realteil.

7.7 Lösen der verallgemeinerten algebraischen Riccatigleichung für Index-2 Systeme

Zur Lösung der algebraischen Riccatigleichung wurde die am Max-Plank-Institut Magdeburg entwickelte `C`-Bibliothek `M.E.S.S.` um das Newton-Verfahren für Index-2 Systeme erweitert. Die zur `M.E.S.S.` zugehörige `Python`-Schnittstelle wurde angepasst. Zur Berechnung der Shiftparameter wurde das Verfahren aus [18] implementiert.
Die `M.E.S.S.` Bibliothek besitzt eine Schnittstelle zu `UMFPACK`, mit der die Sattelpunktprobleme gelöst wurden. `M.E.S.S.` wurde in Verbindung mit `UMFPACK`, `METIS` und `OpenBLAS` genutzt. Die Programmierarbeiten in der `M.E.S.S.` erfolgten in `C`. Mithilfe der `Python`-Schnittstelle

konnte das Newton-Verfahren direkt aus `Python` aufgerufen werden, um die algebraische Riccatigleichung für Index-2 Systeme zu lösen.

8 Numerische Beispiele

Numerische Tests wurden an zwei Beispielen durchgeführt. Wir listen die Parameter für die Beispiele auf. Die Parameter beziehen sich jeweils auf Abschnitt 7.4. RE ist die Reynoldszahl und berechnet sich aus charakteristischer Länge, charakteristischer Geschwindigkeit und Viskosität. Da die charakteristische Länge und charakteristische Geschwindigkeit nicht einheitlich in der Literatur definiert sind, geben wir direkt eine Umrechnung zwischen Reynoldszahl und Viskosität an. Wir wollen ein LQR-Problem lösen. Wie schon in Abschnitt 7.4 erwähnt, wählen wir die Matrizen Q und R im Kostenfunktional stets als Einheitsmatrizen. In Abschnitt 8.1 wird die Wirbelstraße und in Abschnitt 8.2 das Stufengebiet behandelt.
Mit VS bezeichnen wir die Verfeinerungsstufe der Triangulierung des Rechengebietes.
Mit h_{min}/h_{max} meinen wir den minimalen/maximalen Durchmesser des Umkreises der Dreiecke des Rechengebietes.

8.1 Beispiel Wirbelstraße

8.1.1 Daten

Das Rechengebiet besteht aus einem Rechteck mit einem ausgeschnittenen Kreis. Das Rechteck hat Höhe 1 und Breite 5. Der Kreis stellt ein Hindernis dar und hat den Radius 0.1 und den Mittelpunkt $(0.5, 0.5)$.
Der linke Rand des Gebietes ist der Einflussrand Γ_{in} mit Einflussprofil g. Das Einflussprofil g zeigt in x_2 Richtung in das Gebiet Ω und geht an den Eckpunkten von Γ_{in} stetig in $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ über. Das Einflussprofil g skaliert quadratisch mit dem Abstand zur Mitte des Randstückes Γ_{in}.
Der rechte Rand ist der Ausflussrand Γ_{out} mit Ausflussbedingung.
Auf der rechten Hälfte des Kreisrandes befinden sich oben und unten auf dem Kreisrand die Ränder Γ_{ctrl_1} und Γ_{ctrl_2} für die Steuerung.
$\mathcal{B}$ wurde derart gewählt, dass $\mathcal{B}$ auf Γ_{ctrl_1} bzw. Γ_{ctrl_2} ein ins Gebiet Ω zeigendes Normalenfeld am Kreisrand darstellt und die Länge des Feldes zu den Eckpunkten von Γ_{ctrl_1} bzw. Γ_{ctrl_2} hin quadratisch mit dem Winkel skaliert, d. h. $\mathcal{B}$ hat jeweils in der „Mitte“ von Γ_{ctrl_1} bzw. Γ_{ctrl_2} Länge 1 und geht stetig in $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ an den Randpunkten von Γ_{ctrl_1} bzw. Γ_{ctrl_2} über.
Wie in Abschnitt 7.4 beschrieben assemblieren auf den Randstücken Γ_{ctrl_1} und Γ_{ctrl_2} einzeln, daher ist $n_u = 2$. Das bedeutet die Matrix B hat zwei Spalten.

Parameter	Wert
Ω	$([0,5] \times [0,1]) \setminus \mathfrak{B}_{0.1}((0.5, 0.5)) \subset \mathbb{R}^2$
Γ_{in}	$\partial\Omega \cap \{x_1 = 0\}$
Γ_{out}	$\partial\Omega \cap \{x_1 = 5\}$
Γ_{ctrl_1}	$\partial\mathfrak{B}_{0.1}((0.5, 0.5)) \cap \{(x_1, x_2) \mid 0.5 + \frac{1}{8} \cdot 0.1 < x_2 < 0.5 + \frac{6}{8} \cdot 0.1,\ 0.5 < x_1\}$
Γ_{ctrl_2}	$\partial\mathfrak{B}_{0.1}((0.5, 0.5)) \cap \{(x_1, x_2) \mid 0.5 - \frac{1}{8} \cdot 0.1 > x_2 > 0.5 - \frac{6}{8} \cdot 0.1,\ 0.5 < x_1\}$
Γ_h	$\partial\Omega \setminus \{\Gamma_{in} \cup \Gamma_{out} \cup \Gamma_{ctrl_1} \cup \Gamma_{ctrl_2}\}$
ε_{pen}	$1.0\text{e} - 06$
g	$4 \begin{pmatrix} x_2(1 - x_2) \\ 0 \end{pmatrix}$
n_u	2
RE	$\frac{0.1}{\nu}$

8.1.2 Triangulierung des Rechengebietes

Verfeinerungsstufe (VS)	Knoten	Kanten	Dreiecke	h_{min}	h_{max}
1	361	994	633	1.496070e-02	4.160083e-01
2	1355	3887	2532	7.313521e-03	2.080041e-01
3	5242	15370	10128	3.656515e-03	1.040021e-01
4	20612	61124	40512	1.828222e-03	5.200103e-02
5	81736	243784	162048	9.141068e-04	2.600052e-02

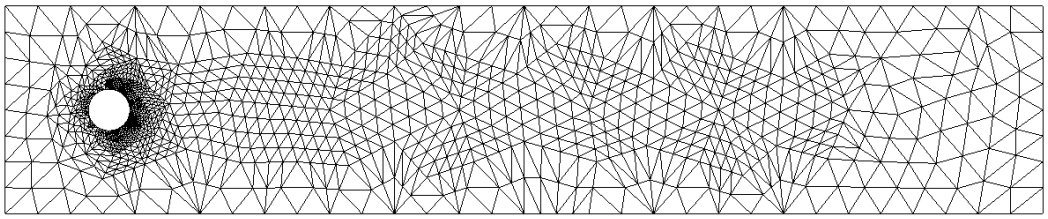

Abbildung 8.1: Gitter mit $VS = 2$

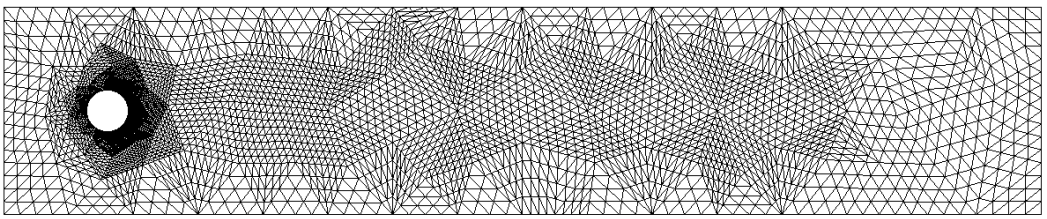

Abbildung 8.2: Gitter mit $VS = 3$

8.1.3 stationäre Lösung

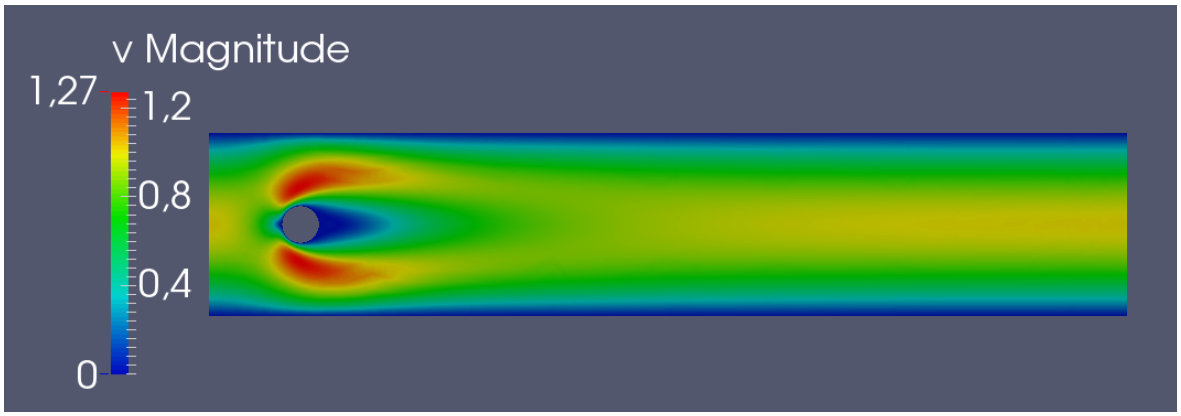

Abbildung 8.3: v_s für $RE = 10$ und $VS = 3$

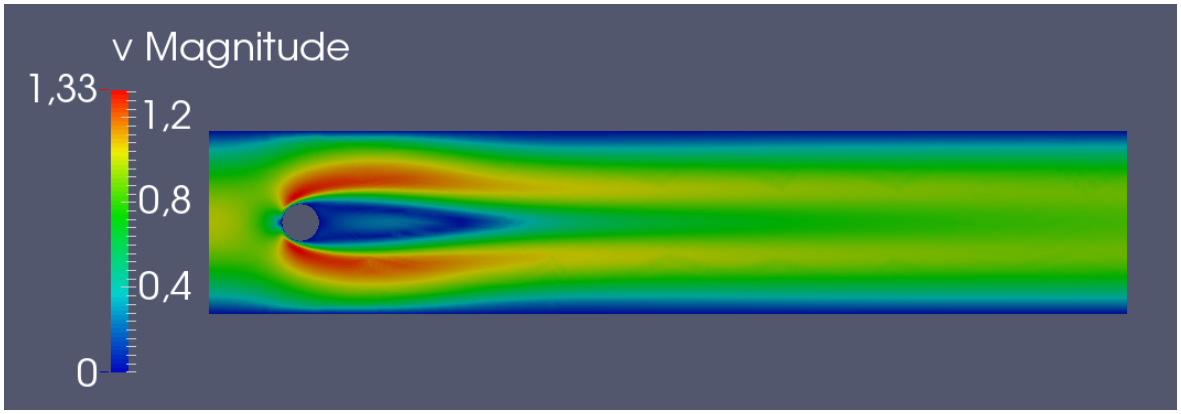

Abbildung 8.4: v_s für $RE = 60$ und $VS = 3$

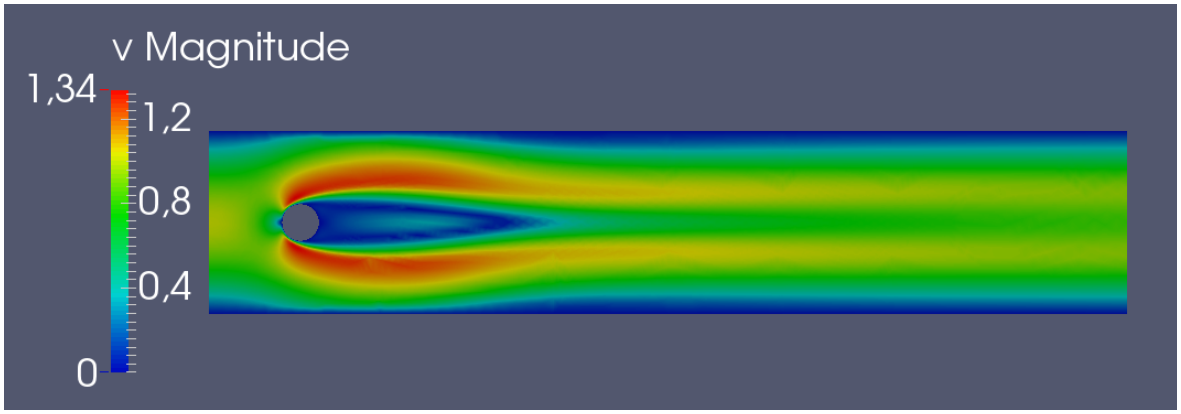

Abbildung 8.5: v_s für $RE = 90$ und $VS = 3$

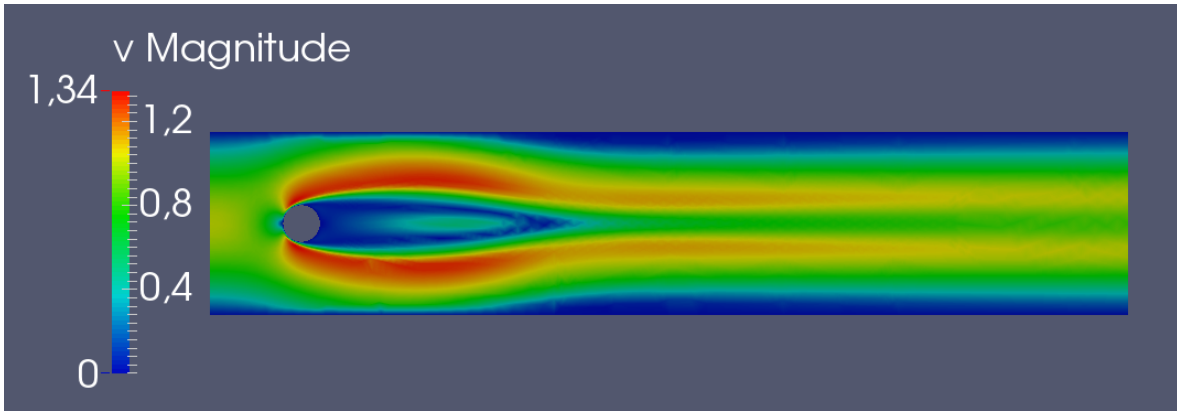

Abbildung 8.6: v_s für $RE = 160$ und $VS = 3$

8.1.4 qualitative Spektraleigenschaften des Systems und Reynoldszahl

RE	instabile Eigenwerte
10	keine
60	0.66738087-5.01595039i, 0.66738087+5.01595039i
90	1.01531692-4.982154i, 1.01531692+4.982154i
160	1.19135069-4.76179772i, 1.19135069+4.76179772i

Tabelle 8.1: Instabile Eigenwerte $VS = 2$.

Qualitativ kann man sagen, dass bei wachsenden Reynoldszahlen der Realteil der instabilen Eigenwerte größer wird. Das deckt sich mit den Beobachtungen aus [10, Fig.2.], obwohl ein anderes Rechengebiet in [10] verwendet wurde.
Das System ist für $RE \geq 60$ instabil. Für $RE = 160$ brach die Simulation ohne Steuerung bei $t \approx 12$ ab. Die algebraische Riccatigleichung ließ sich für $RE = 160$ nicht mehr lösen, da die erste ADI-Iteration nicht konvergierte. Hier kann man vermuten, dass der Eigenraum zu den instabilen Eigenwerten nicht genügend genau berechnet wurde und die Lösung der algebraischen Bernoulligleichung nicht ausreichend stabilisierend für das System war.
Eine Erklärung hierfür könnte sein, dass für große Systeme nicht das volle Eigenwertproblem gelöst werden kann und instabile Eigenwerte nahe der imaginären Achse „unentdeckt“ bleiben. Für $VS = 1$ und $RE = 160$ wurde das volle Eigenwertproblem gelöst und es gab fünf anstatt zwei instabile Eigenwerte vgl. Abbildung 8.10. Allerdings ist für $VS = 1$ das Gitter sehr grob, daher kann man nur vermuten, dass nicht alle instabilen Eigenwerte durch `eigs` gefunden wurden.
Weil

$$\left(\begin{bmatrix} A & G \\ G^T & 0 \end{bmatrix}, \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \right)$$

auch uneigentliche Eigenwerte hat, haben wir für die folgenden Visualisierungen des Spektrums das Paar

$$\left(\begin{bmatrix} A & G \\ G^T & 0 \end{bmatrix}, \begin{bmatrix} M & -0.02 \cdot G \\ -0.02 \cdot G^T & 0 \end{bmatrix} \right)$$

verwendet. Nach [21, Cor. 6.2] werden durch diese Transformation nur die uneigentlichen Eigenwerte nach $\frac{1}{-0.02}$ transformiert. Da das volle Eigenwertproblem gelöst wurde, wurde für die folgenden Grafiken $VS = 1$ gewählt. Es ist zu erkennen, dass bei größeren Reynoldszahlen sich das Spektrum mehr zum Nullpunkt konzentriert, dass könnte auch ein Grund sein warum `eigs` in Schwierigkeiten gerät und unter Umständen sehr lange iteriert.

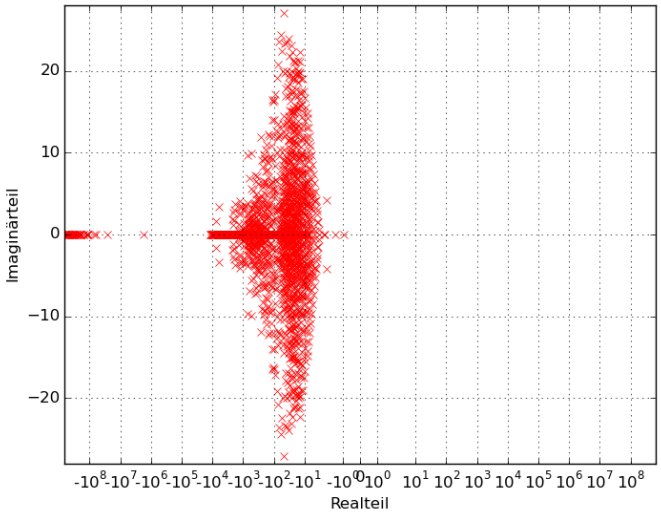

Abbildung 8.7: Eigenwerte für $RE = 10$ und $VS = 1$

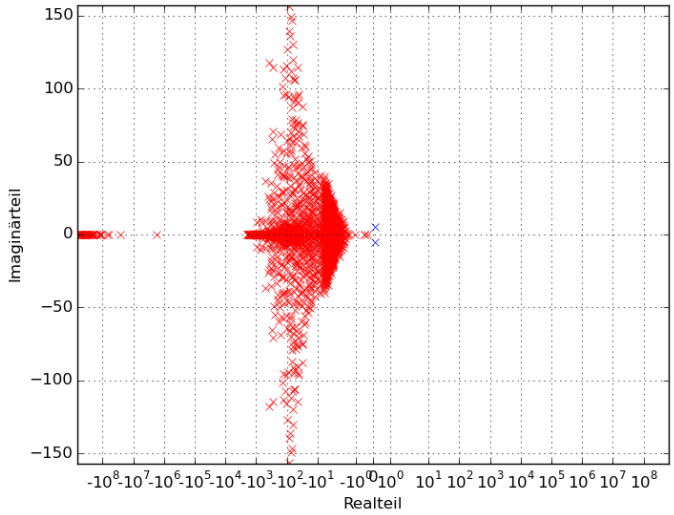

Abbildung 8.8: Eigenwerte für $RE = 60$ und $VS = 1$ ohne Stabilisierung durch algebraische Bernoulligleichung

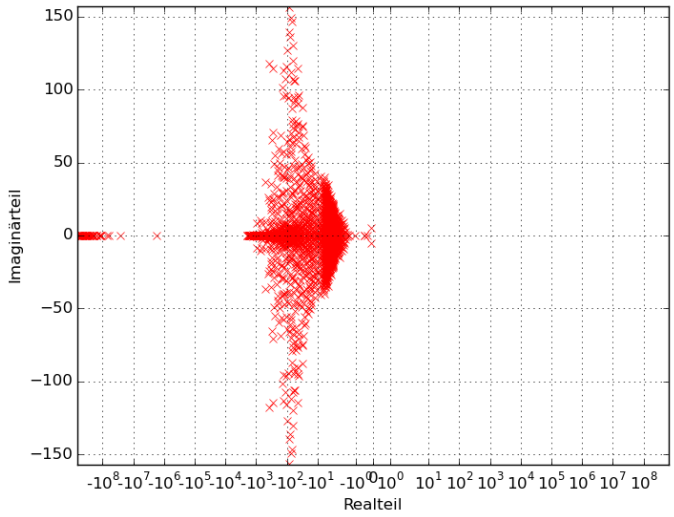

Abbildung 8.9: Eigenwerte für $RE = 60$ und $VS = 1$ mit Stabilisierung durch algebraische Bernoulligleichung

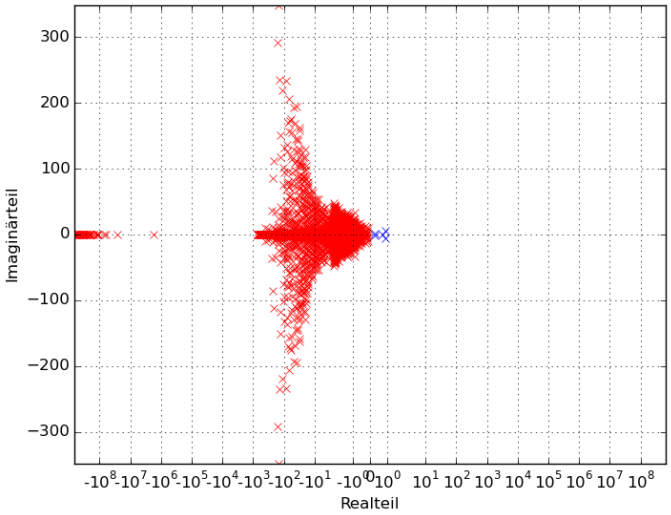

Abbildung 8.10: Eigenwerte für $RE = 160$ und $VS = 1$

8.1.5 Konvergenz von v_δ und die Wahl von C

Wir wählen für die Matrix C die Matrizen $C((2.5, 0.25), (2.5, 0.75))$, $C((4.5, 0.25), (4.5, 0.75))$ und $C((2.5, 0.5))$.
$C((2.5, 0.25), (2.5, 0.75))$ misst die Geschwindigkeit mittig im Gebiet.
Für $C((4.5, 0.25), (4.5, 0.75))$ sind die Messpunkte weiter nach hinten verschoben.
Für $C((2.5, 0.5))$ befinden sich die Messpunkte mittig im Gebiet.
Zum Vergleichen wählen wir $RE = 10, 60, 90$ und $VS = 2$.
Die Ergebnisse haben gezeigt, dass die Konvergenz im Wesentlichen unabhängig von der Wahl von C ist. Wählt man $C((2.5, 0.5))$, hat dies den Vorteil, das man weniger rechte Seiten im ADI-Verfahren hat.
Es ist aber zu erwarten, dass ungünstige Wahlen von C, zu schlechter oder keiner Konvergenz führen. Ungünstig wäre es beispielsweise die „Messpunkte“ sehr nahe am Dirichletrand zu platzieren. Solche „nicht besonders sinnvollen“ Wahlen für C wurden nicht untersucht.

$t \in [0, 60]$	$\|\|v_\delta\|\|_{L^2(\Omega)^2}$ ohne Steuerung	$\|\|v_\delta\|\|_{L^2(\Omega)^2}$ $C((2.5, 0.25), (2.5, 0.75))$	$\|\|v_\delta\|\|_{L^2(\Omega)^2}$ $C((4.5, 0.25), (4.5, 0.75))$	$\|\|v_\delta\|\|_{L^2(\Omega)^2}$ $C((2.5, 0.5))$
0.00	4.074e-01	4.074e-01	4.074e-01	4.074e-01
2.00	1.264e-02	1.264e-02	1.264e-02	1.264e-02
3.99	1.239e-03	1.239e-03	1.239e-03	1.239e-03
5.99	8.075e-05	8.075e-05	8.075e-05	8.075e-05
7.99	3.578e-06	3.578e-06	3.578e-06	3.578e-06
9.98	2.178e-07	2.178e-07	2.178e-07	2.178e-07
11.98	3.235e-08	3.235e-08	3.235e-08	3.235e-08
13.98	4.847e-09	4.847e-09	4.847e-09	4.847e-09
15.97	7.237e-10	7.237e-10	7.237e-10	7.237e-10
17.97	1.079e-10	1.079e-10	1.079e-10	1.079e-10
19.97	1.609e-11	1.609e-11	1.609e-11	1.609e-11
21.96	2.398e-12	2.398e-12	2.398e-12	2.398e-12
23.96	3.574e-13	3.574e-13	3.574e-13	3.574e-13
25.96	5.327e-14	5.327e-14	5.327e-14	5.327e-14
27.95	7.939e-15	7.939e-15	7.939e-15	7.940e-15
29.95	1.183e-15	1.183e-15	1.183e-15	1.183e-15
31.95	1.764e-16	1.764e-16	1.764e-16	1.764e-16
33.94	2.628e-17	2.628e-17	2.628e-17	2.629e-17
35.94	3.917e-18	3.917e-18	3.917e-18	3.918e-18
37.94	5.838e-19	5.838e-19	5.838e-19	5.839e-19
39.93	8.701e-20	8.701e-20	8.701e-20	8.702e-20
41.93	1.297e-20	1.297e-20	1.297e-20	1.297e-20
43.92	1.933e-21	1.933e-21	1.933e-21	1.933e-21
45.92	2.881e-22	2.881e-22	2.881e-22	2.881e-22
47.92	4.293e-23	4.293e-23	4.293e-23	4.294e-23
49.91	6.399e-24	6.399e-24	6.399e-24	6.399e-24
51.91	9.537e-25	9.537e-25	9.537e-25	9.538e-25
53.91	1.421e-25	1.421e-25	1.421e-25	1.421e-25
55.90	2.118e-26	2.118e-26	2.118e-26	2.119e-26
57.90	3.157e-27	3.157e-27	3.157e-27	3.158e-27
59.90	4.706e-28	4.706e-28	4.706e-28	4.706e-28

Tabelle 8.2: L^2-Norm des Differenzzustandes für $RE = 10$ und $VS = 2$.

$t \in [0,60]$	$\|\|v_\delta\|\|_{L^2(\Omega)^2}$ ohne Steuerung	$\|\|v_\delta\|\|_{L^2(\Omega)^2}$ $C((2.5,0.25),(2.5,0.75))$	$\|\|v_\delta\|\|_{L^2(\Omega)^2}$ $C((4.5,0.25),(4.5,0.75))$	$\|\|v_\delta\|\|_{L^2(\Omega)^2}$ $C((2.5,0.5))$
0.00	4.146e-01	4.146e-01	4.146e-01	4.146e-01
2.00	1.141e-01	1.137e-01	1.139e-01	1.140e-01
3.99	1.109e-01	8.711e-02	8.805e-02	8.674e-02
5.99	2.346e-01	6.419e-02	6.646e-02	6.286e-02
7.99	4.065e-01	2.215e-02	2.504e-02	2.038e-02
9.98	4.916e-01	6.254e-03	7.932e-03	5.770e-03
11.98	5.226e-01	2.495e-03	2.879e-03	2.442e-03
13.98	5.321e-01	9.367e-04	9.947e-04	9.369e-04
15.97	5.342e-01	3.564e-04	3.586e-04	3.600e-04
17.97	5.348e-01	1.372e-04	1.343e-04	1.392e-04
19.97	5.353e-01	5.310e-05	5.138e-05	5.401e-05
21.96	5.350e-01	2.062e-05	1.982e-05	2.098e-05
23.96	5.363e-01	8.012e-06	7.682e-06	8.158e-06
25.96	5.350e-01	3.115e-06	2.982e-06	3.172e-06
27.95	5.355e-01	1.211e-06	1.159e-06	1.234e-06
29.95	5.350e-01	4.712e-07	4.505e-07	4.797e-07
31.95	5.355e-01	1.832e-07	1.752e-07	1.865e-07
33.94	5.361e-01	7.127e-08	6.812e-08	7.255e-08
35.94	5.349e-01	2.772e-08	2.649e-08	2.821e-08
37.94	5.356e-01	1.078e-08	1.031e-08	1.097e-08
39.93	5.348e-01	4.195e-09	4.010e-09	4.268e-09
41.93	5.361e-01	1.633e-09	1.561e-09	1.661e-09
43.92	5.356e-01	6.357e-10	6.081e-10	6.465e-10
45.92	5.351e-01	2.478e-10	2.373e-10	2.519e-10
47.92	5.354e-01	9.677e-11	9.285e-11	9.829e-11
49.91	5.349e-01	3.791e-11	3.650e-11	3.846e-11
51.91	5.363e-01	1.493e-11	1.445e-11	1.512e-11
53.91	5.351e-01	5.931e-12	5.792e-12	5.990e-12
55.90	5.355e-01	2.388e-12	2.363e-12	2.401e-12
57.90	5.350e-01	9.812e-13	9.887e-13	9.803e-13
59.90	5.354e-01	4.141e-13	4.269e-13	4.104e-13

Tabelle 8.3: L^2-Norm des Differenzzustandes für $RE = 60$ und $VS = 2$.

$t \in [0, 60]$	$\|\|v_\delta\|\|_{L^2(\Omega)^2}$ ohne Steuerung	$\|\|v_\delta\|\|_{L^2(\Omega)^2}$ $C((2.5, 0.25), (2.5, 0.75))$	$\|\|v_\delta\|\|_{L^2(\Omega)^2}$ $C((4.5, 0.25), (4.5, 0.75))$	$\|\|v_\delta\|\|_{L^2(\Omega)^2}$ $C((2.5, 0.5))$
0.00	4.213e-01	4.213e-01	4.213e-01	4.213e-01
2.00	1.478e-01	1.495e-01	1.494e-01	1.497e-01
3.99	1.998e-01	1.659e-01	1.744e-01	1.625e-01
5.99	5.114e-01	1.737e-01	1.955e-01	1.625e-01
7.99	6.419e-01	1.098e-01	1.347e-01	9.720e-02
9.98	6.837e-01	3.840e-02	6.010e-02	3.059e-02
11.98	6.972e-01	8.961e-03	1.864e-02	7.448e-03
13.98	7.040e-01	2.135e-03	4.938e-03	1.814e-03
15.97	7.071e-01	7.427e-04	1.886e-03	5.325e-04
17.97	7.084e-01	3.874e-04	9.675e-04	2.265e-04
19.97	7.086e-01	2.190e-04	5.383e-04	1.368e-04
21.96	7.083e-01	1.294e-04	3.010e-04	8.575e-05
23.96	7.078e-01	7.685e-05	1.702e-04	5.571e-05
25.96	7.075e-01	4.635e-05	9.654e-05	3.600e-05
27.95	7.076e-01	2.824e-05	5.516e-05	2.331e-05
29.95	7.077e-01	1.740e-05	3.176e-05	1.512e-05
31.95	7.081e-01	1.084e-05	1.846e-05	9.809e-06
33.94	7.087e-01	6.809e-06	1.085e-05	6.372e-06
35.94	7.089e-01	4.312e-06	6.448e-06	4.143e-06
37.94	7.085e-01	2.749e-06	3.882e-06	2.696e-06
39.93	7.080e-01	1.761e-06	2.365e-06	1.755e-06
41.93	7.076e-01	1.133e-06	1.459e-06	1.143e-06
43.92	7.076e-01	7.318e-07	9.090e-07	7.451e-07
45.92	7.076e-01	4.737e-07	5.718e-07	4.857e-07
47.92	7.079e-01	3.072e-07	3.626e-07	3.168e-07
49.91	7.085e-01	1.996e-07	2.314e-07	2.066e-07
51.91	7.089e-01	1.298e-07	1.484e-07	1.348e-07
53.91	7.087e-01	8.446e-08	9.560e-08	8.793e-08
55.90	7.082e-01	5.501e-08	6.176e-08	5.738e-08
57.90	7.077e-01	3.584e-08	4.000e-08	3.744e-08
59.90	7.075e-01	2.337e-08	2.596e-08	2.444e-08

Tabelle 8.4: L^2-Norm des Differenzzustandes für $RE = 90$ und $VS = 2$.

8.1.6 Konvergenz von v_δ

Wir betrachten die Konvergenz von v_δ in Abhängigkeit von der Reynoldszahl und vergleichen den Fall mit optimaler Steurung mit dem Fall der Simulation. Da das System für $RE \geq 60$ instabil ist, erwarten wir für $RE \geq 60$ deutliche Abweichungen zwischen dem gesteuerten und ungesteuerten Fall.
Der Fall $RE = 10$ ist nach Tabelle 8.2 nicht interessant.
Wir stellen die Ergebnisse für $C((2.5, 0.25), (2.5, 0.75))$ und $VS = 2$ dar.

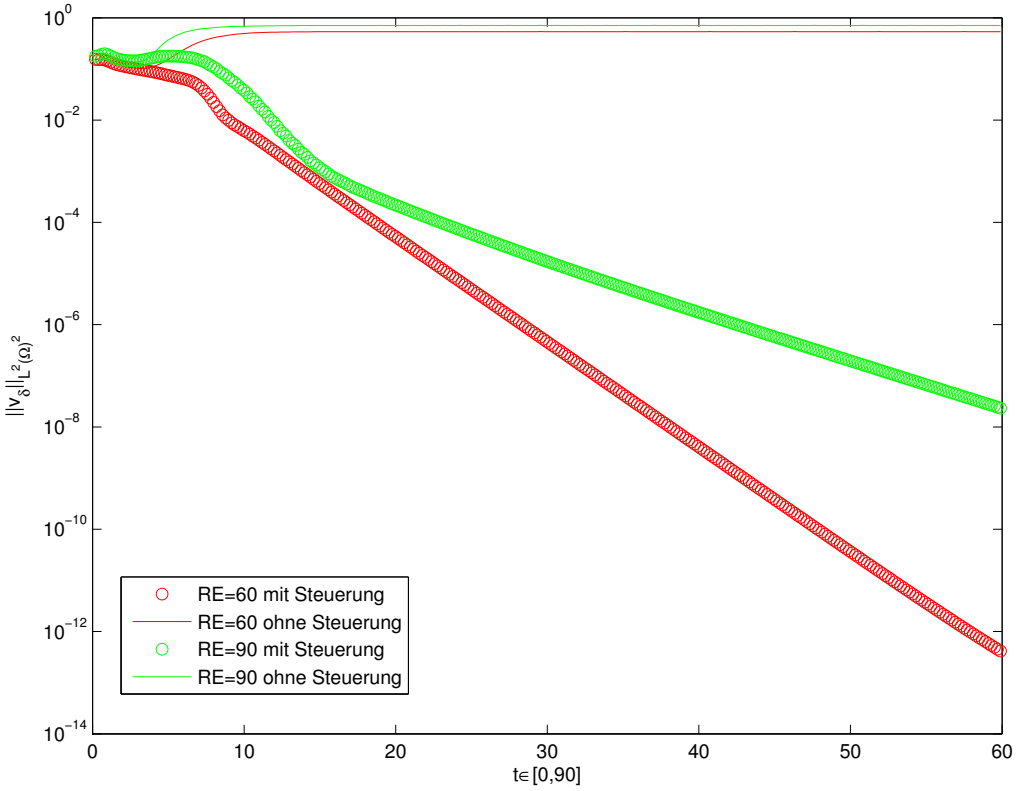

Abbildung 8.11: Konvergenzverhalten von v_δ in der L^2-Norm

8.2 Beispiel Stufengebiet

8.2.1 Daten

Das Rechengebiet besteht aus zwei vereinigten Rechtecken derart, dass eine Stufe im linken Teil des Gebietes entsteht.
Der linke Rand ist der Einflussrand mit parabelförmigen Einflussprofil. Der obere und untere Teil des Stufenrandes ist für die Steuerung. $\mathcal{B}$ zeigt für den oberen Teil des Randes in das Gebiet hinein und für den unteren Teil des Randes aus dem Gebiet raus. $\mathcal{B}$ beschreibt ein parabelförmiges Einfluss- bzw. Ausflussprofil. In Abhängigkeit von u „saugt“ eine Düse ein und die andere „bläst“ aus, daher hat B eine Spalte bzw. $n_u = 1$.

Parameter	Wert
Ω	$([0,25] \times [1,2]) \cup ([5,25] \times [0,1]) \subset \mathbb{R}^2$
Γ_{in}	$\partial\Omega \cap \{x_1 = 0.0\}$
Γ_{out}	$\partial\Omega \cap \{x_1 = 25.0\}$
Γ_{ctrl}	$\partial\Omega \cap \{(x_1, x_2) \mid x_1 = 5,\ 0.75 < x_2 < 1 \text{ oder } 0 < x_2 < 0.25\}$
Γ_h	$\partial\Omega \setminus \{\Gamma_{in} \cup \Gamma_{out} \cup \Gamma_{ctrl}\}$
ε_{pen}	1.0e−6
g	$4 \begin{pmatrix} (x_2-1)(2-x_2) \\ 0 \end{pmatrix}$
$\mathcal{B}$	$\begin{cases} \frac{1}{64} \begin{pmatrix} (1-x_2)(x_2-0.75) \\ 0 \end{pmatrix} & \text{falls } 0.75 < x_2 < 1 \\ \frac{1}{64} \begin{pmatrix} x_2(x_2-0.25) \\ 0 \end{pmatrix} & \text{falls } 0 < x_2 < 0.25 \end{cases}$
n_u	1
RE	$\frac{1}{\nu}$

8.2.2 Schwierigkeit

Eine Schwierigkeit bei dem Beispiel ist, dass das Rechengebiet sehr groß ist. Das Rechengebiet sehr groß zu wählen, hat sich beim Testen als notwendig herausgestellt, weil man sonst bei großen Reynoldszahlen nur noch einen „abgeschnitten“ Teil der Lösung beobachten kann. Durch lokales Verfeinern wurde versucht, kleinere Systemmatrizen zu erhalten. Es hat sich gezeigt, dass das im Wesentlichen kaum Vorteile bringt. Bei Vergröberung der globalen Verfeinerung lässt sich das stationäre Problem nicht mehr lösen. Daher wurde nur hinter der Stufe lokal verfeinert.
Da das Rechengebiet groß ist, kann man erwarten, dass man einen großen Zeithorizont benötigt, damit v_δ klein wird. Da man gezwungen ist kleine Zeitschritte zu nehmen, benötigen die Simulationen sehr lange. Um die stationäre Lösung zu berechnen, die Matrizen zu assemblieren, das Vorwärtsproblem zu simulieren ($T = 90$), auf instabile Eigenwerte mit `eigs` zu untersuchen, die algebraische Riccatigleichung lösen und mit optimaler Steuerung nochmals simulieren ($T = 90$) waren für $RE = 1000$ und $VS = 2$ ca. 2 Wochen nötig. Es wird sich zeigen, dass der Zeithorizont $T = 90$ zu klein gewählt wurde.
Wir können zwar die Lösung der algebraischen Riccatigleichung mit $X = ZZ^T$ „effizient“ speichern, falls aber die Shiftparameter „schlecht“ sind, resultiert das in einer langsamen Konvergenz der ADI-Iteration. Dann benötigt man viele Iterationen und der Niedrigrangfaktor Z, welcher dicht besetzt ist, wird „groß“. Die Bibliothek `M.E.S.S.` besitzt zwar Funktionen um den Lösungsfaktor wieder zu „verkleinern“ oder persistent zwischenzuspeichern, es müssen jedoch trotzdem viele Sattelpunktprobleme gelöst werden.
Man kann einerseits ausnutzen, dass die Shiftparameter zyklisch benutzt werden und für jeden Shiftparameter vorab die LU-Zerlegung mittels `UMFPACK` bestimmen. Bei dieser Strategie wird allerdings viel Hauptspeicher benötigt. Bestimmt man während der ADI-Iteration in jedem Iterationsschritt die LU-Zerlegung mit `UMFPACK` neu, kostet das wesentlich mehr Rechenzeit. In diesem Zusammenhang hat sich herausgestellt, dass es geeigneter ist, die Shiftparameter mit der Heuristik aus [18] zu bestimmen, anstatt mit der Heuristik von Penzl [51].
Während der Arbeit war auch ein Problem, dass die Funktion `mess_multidirect_umfpack` aus der `M.E.S.S.` einen Fehler enthält, der sich erst bei großen Systemen bemerkbar macht. Die Funktion `mess_multidirect_umfpack` kann genutzt werden, um für geshiftete lineare Gleichungssysteme ($A+pE$) mehrere direkte Löser mittels `UMFPACK` aufzubauen. Allerdings arbeitet `mess_multidirect_umfpack` unter der Annahme, dass das Sparsity Pattern von $A+pE$ invariant unter der Menge der Shiftparamter $p \in \mathbb{C}^-$ ist. Für größere Systeme führt diese Annahme jedoch dazu, dass `UMFPACK` nicht das korrekte Sparsity Pattern übergeben bekommt und die Systeme mit wachsender Größe „ungenauer“ gelöst werden. Den Fehler zu lokalisieren hat einige Zeit in Anspruch genommen und es mussten viele Ergebnisse neu berechnet werden.
Ursprünglich war die Idee eine „gute“ vorgesteuerte stationäre Lösung zu stabilisieren. Hier sollten geeignete Dirichletranddaten g für den stationären Fall bestimmt werden, um Rezirkulierungseffekte hinter der Stufe zu vermeiden. Hierzu werden in [29] geeignete Wahlen für mögliche Zielfunktion vorgestellt.
Aufgrund der genannten Schwierigkeiten in der praktischen Umsetzung, konnte die Idee aus Zeitgründen nicht umgesetzt werden und wir präsentieren die vorhanden Ergebnisse.

8.2.3 Triangulierung des Rechengebietes

Verfeinerungsstufe (VS)	Knoten	Kanten	Dreiecke	h_{min}	h_{max}
1	2169	6160	3992	7.855806e-02	3.179914e-01
2	8329	24296	15968	3.927903e-02	1.589957e-01
3	32625	96496	63872	1.963951e-02	7.949784e-02
4	129121	384608	255488	9.819757e-03	3.974892e-02

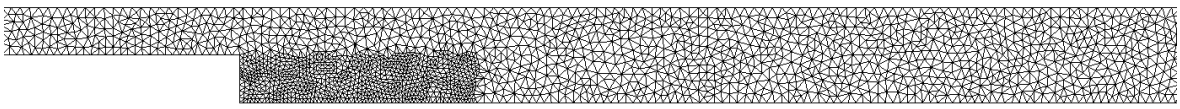

Abbildung 8.12: Gitter mit $VS = 2$

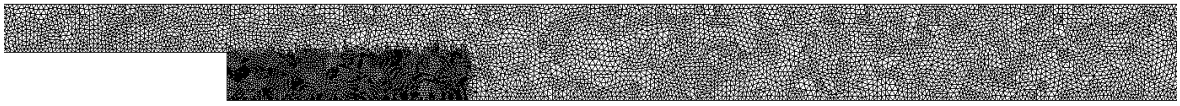

Abbildung 8.13: Gitter mit $VS = 3$

8.2.4 stationäre Lösung

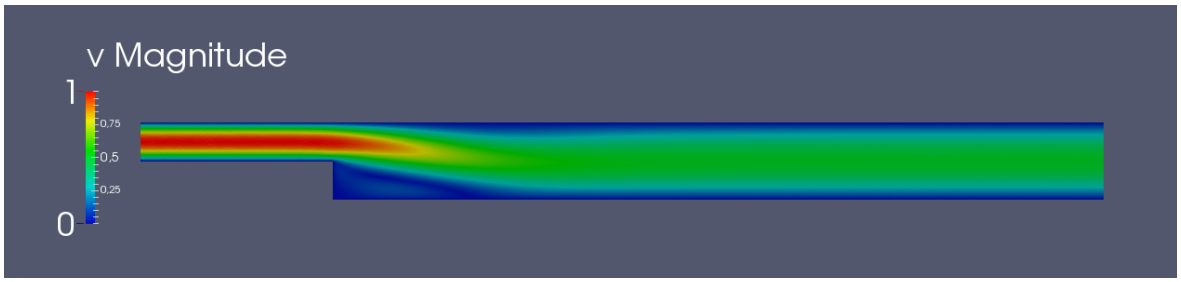

Abbildung 8.14: stationäre Lösung für $RE = 100$ und $VS = 2$

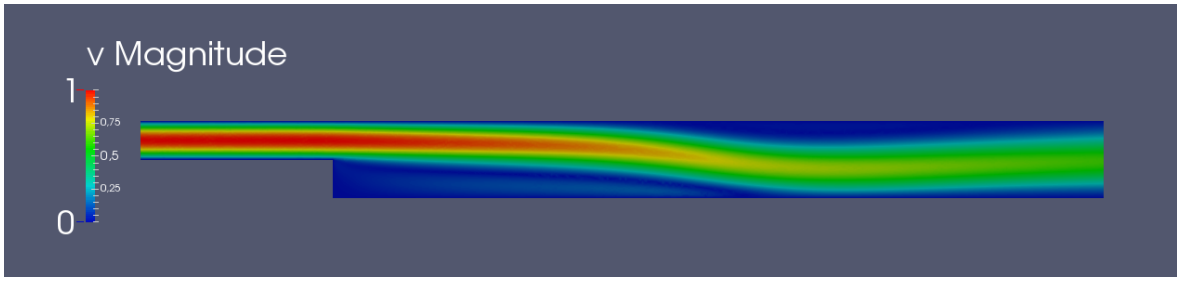

Abbildung 8.15: stationäre Lösung für $RE = 500$ und $VS = 2$

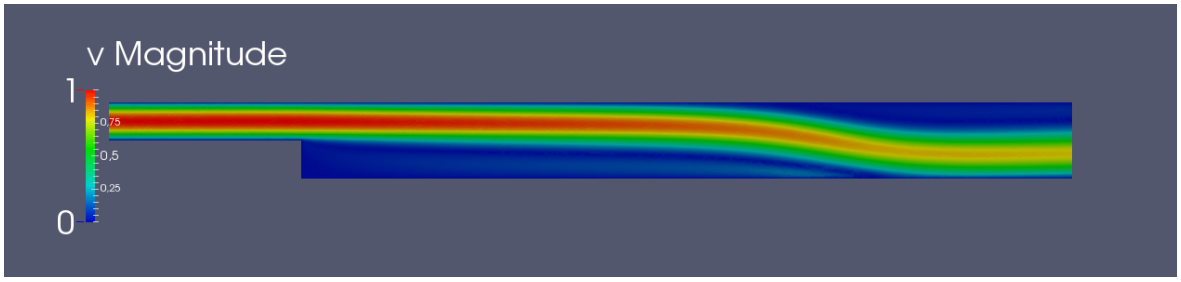

Abbildung 8.16: stationäre Lösung für $RE = 1000$ und $VS = 2$

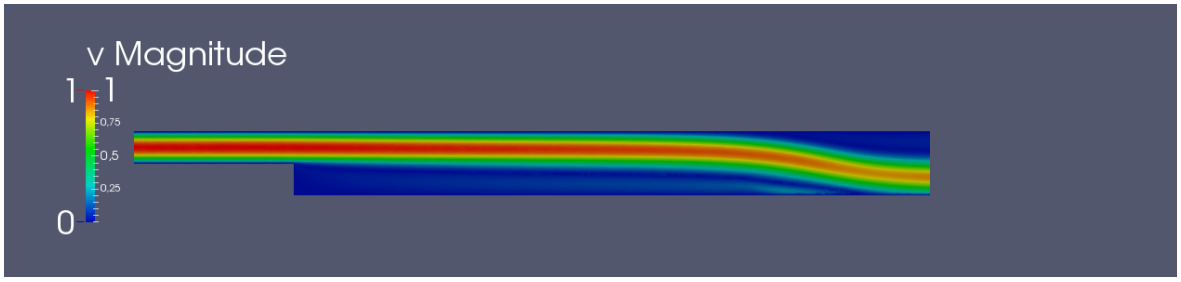

Abbildung 8.17: stationäre Lösung für $RE = 1500$ und $VS = 2$

8.2.4.1 Stufengebiet $RE = 1000$

Die Ergebnisse wurden für $C((7, 0.5), (7, 1.5))$ und $VS = 2$ berechnet. Das System ist stabil für $RE = 1000$.

Zeit $t \in [0, 90]$	$\|\|v_\delta\|\|_{L^2(\Omega)^2}$ ohne Steuerung	$\|\|v_\delta\|\|_{L^2(\Omega)^2}$ mit Steuerung
0.00	3.494e-01	8.736e-01
2.95	2.302e-01	5.702e-01
5.89	2.764e-01	6.450e-01
8.84	3.662e-01	7.612e-01
11.78	5.023e-01	9.235e-01
14.73	7.037e-01	1.116e+00
17.68	9.565e-01	1.226e+00
20.62	1.222e+00	1.289e+00
23.57	1.491e+00	1.440e+00
26.51	1.796e+00	1.595e+00
29.46	1.997e+00	1.673e+00
32.41	2.071e+00	1.712e+00
35.35	2.123e+00	1.733e+00
38.30	2.101e+00	1.597e+00
41.24	2.085e+00	1.431e+00
44.19	2.116e+00	1.450e+00
47.13	2.160e+00	1.556e+00
50.08	2.191e+00	1.601e+00
53.03	2.203e+00	1.606e+00
55.97	2.248e+00	1.594e+00
58.92	2.297e+00	1.518e+00
61.86	2.326e+00	1.437e+00
64.81	2.301e+00	1.376e+00
67.76	2.238e+00	1.339e+00
70.70	2.193e+00	1.330e+00
73.65	2.181e+00	1.363e+00
76.59	2.172e+00	1.404e+00
79.54	2.152e+00	1.416e+00
82.49	2.115e+00	1.394e+00
85.43	2.072e+00	1.308e+00
88.38	2.036e+00	1.107e+00

Tabelle 8.5: L^2-Norm des Differenzzustandes v_δ für $RE = 1000$

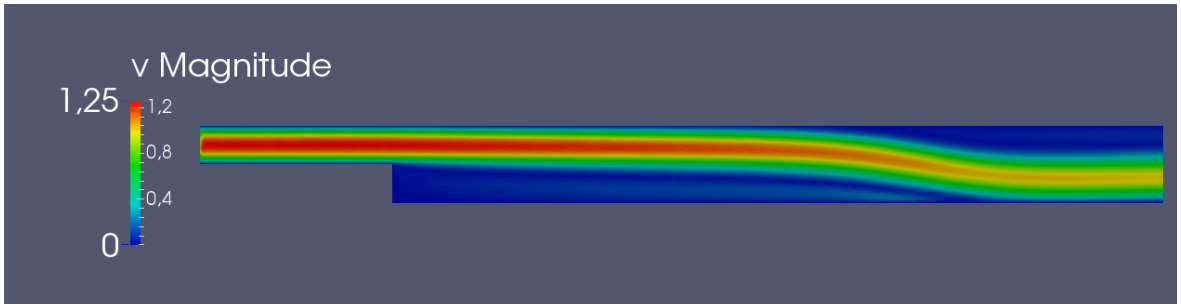

Abbildung 8.18: Simulation mit optimaler Steuerung, $v = v_\delta + v_s$ für $t = 0$ und $RE = 1000$

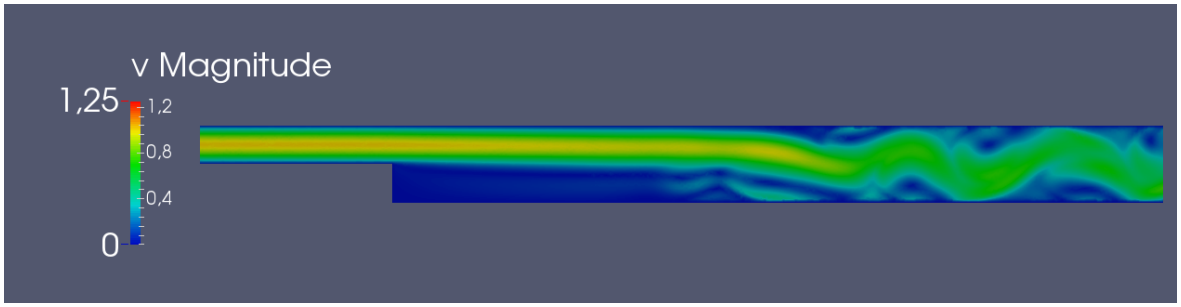

Abbildung 8.19: Simulation mit optimaler Steuerung, $v = v_\delta + v_s$ für $t = 45$ und $RE = 1000$

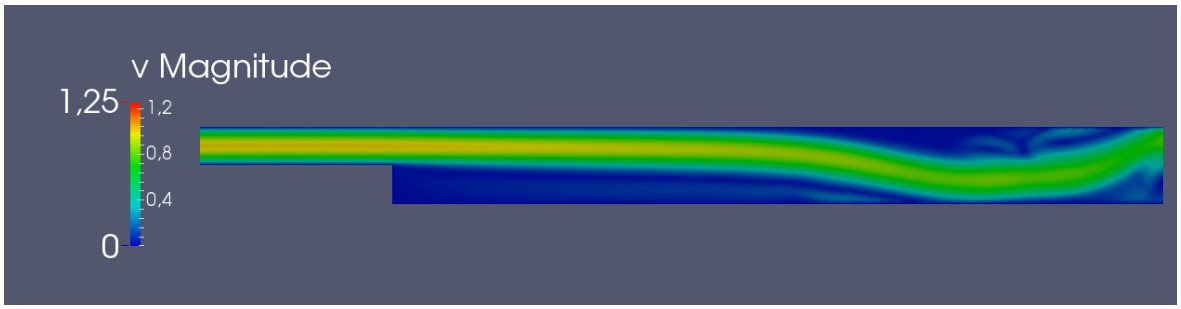

Abbildung 8.20: Simulation mit optimaler Steuerung, $v = v_\delta + v_s$ für $t = 90$ und $RE = 1000$

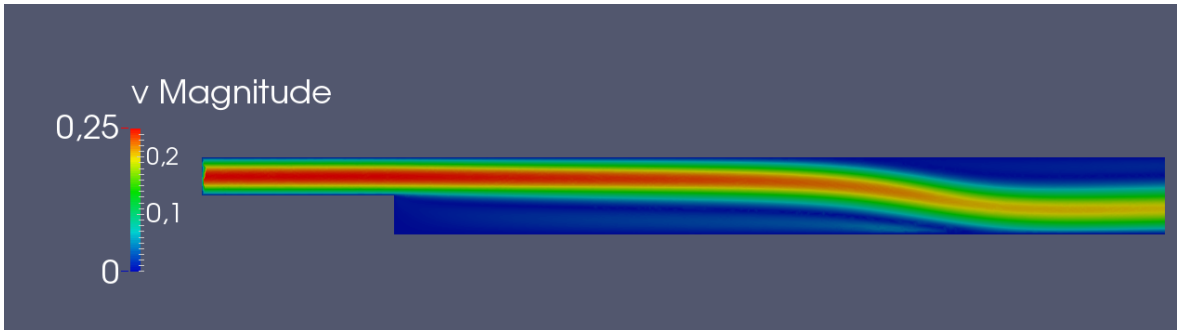

Abbildung 8.21: Simulation mit optimaler Steuerung, v_δ für $t = 0$ und $RE = 1000$

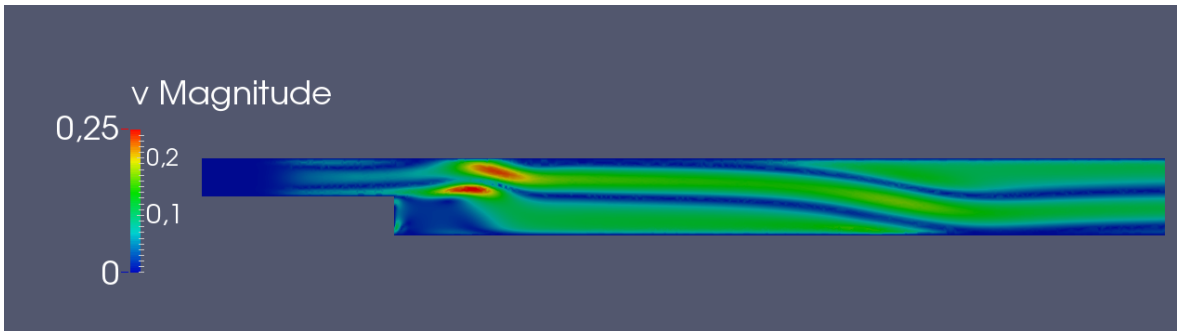

Abbildung 8.22: Simulation mit optimaler Steuerung, v_δ für $t = 5$ und $RE = 1000$, Regelungseingriff am Rand Γ_{ctrl} zu erkennen

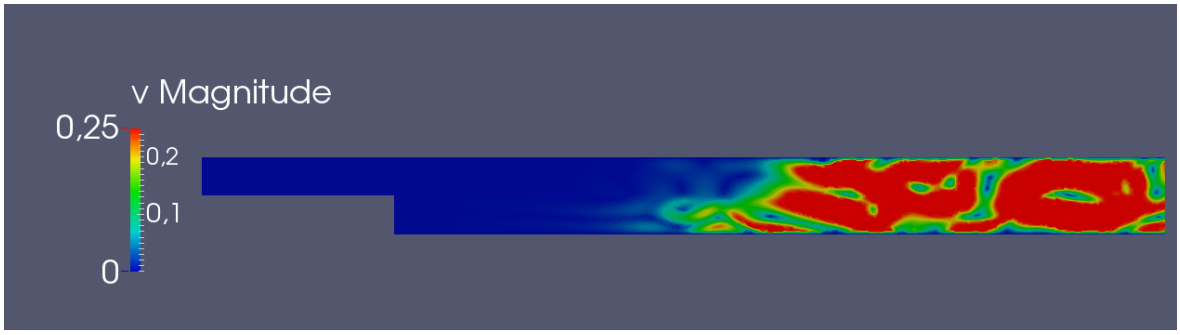

Abbildung 8.23: Simulation mit optimaler Steuerung, v_δ für $t = 45$ und $RE = 1000$

Abbildung 8.24: Simulation mit optimaler Steuerung, v_δ für $t = 90$ und $RE = 1000$

8.2.4.2 Stufengebiet $RE = 1500$

Die Ergebnisse wurden mit zweifacher Verfeinerung und $C((7, 0.5), (7, 1.5))$ berechnet. Das System ist ebenfalls stabil.

Zeit $t \in [0, 90]$	$\|\|v_\delta\|\|_{L^2(\Omega)^2}$ ohne Steuerung	$\|\|v_\delta\|\|_{L^2(\Omega)^2}$ mit Steuerung
0.00	3.560e-01	8.900e-01
2.95	2.435e-01	6.043e-01
5.89	2.970e-01	6.934e-01
8.84	4.147e-01	8.355e-01
11.78	6.197e-01	1.065e+00
14.73	9.238e-01	1.370e+00
17.68	1.285e+00	1.735e+00
20.62	1.680e+00	2.016e+00
23.57	2.062e+00	2.260e+00
26.51	2.330e+00	2.445e+00
29.46	2.411e+00	2.589e+00
32.41	2.444e+00	2.523e+00
35.35	2.442e+00	2.310e+00
38.30	2.476e+00	2.255e+00
41.24	2.545e+00	2.336e+00
44.19	2.527e+00	2.363e+00
47.13	2.414e+00	2.361e+00
50.08	2.282e+00	2.299e+00
53.03	2.260e+00	2.331e+00
55.97	2.433e+00	2.404e+00
58.92	2.535e+00	2.414e+00
61.86	2.536e+00	2.337e+00
64.81	2.468e+00	2.288e+00
67.76	2.427e+00	2.346e+00
70.70	2.455e+00	2.448e+00
73.65	2.498e+00	2.502e+00
76.59	2.506e+00	2.526e+00
79.54	2.486e+00	2.512e+00
82.49	2.509e+00	2.451e+00
85.43	2.551e+00	2.410e+00
88.38	2.612e+00	2.440e+00

Tabelle 8.6: L^2-Norm des Differenzzustandes v_δ für $RE = 1500$

8.2.4.3 Stufengebiet $RE = 2000$

Hier brach die Simulation für $t \approx 40$ ab.

8.2.5 Zusammenfassung zum Stufengebiet

Für $RE = 1000, 1500, 2000$ sind die Systeme stabil und man benötigt keine stabilisierende Lösung der algebraischen Bernoulligleichung. Die Simulationszeit ($T = 90$) war zu klein gewählt, um Konvergenz aus den Tabellen entnehmen zu können. Die Bilder lassen jedoch vermuten, dass v_δ gegen 0 konvergiert. Für $RE = 2000$ brach die Simulation ab, daher wurden keine höheren Reynoldszahlen verwendet.

9 Zusammenfassung und Ausblick

Wir haben uns im Kapitel 2 mit linear-quadratischen Regelungsproblemen befasst. Im zeitasymptotischen Fall kann man das LQR-Problem lösen, indem man die algebraische Riccatigleichung löst. Hierzu wurde das Newton-Verfahren vorgestellt und wir haben gesehen, dass während des Newton-Verfahrens mehrere Lyapunovgleichungen gelöst werden müssen. Zur numerischen Lösung der Lyapunovgleichungen haben wir das ADI-Verfahren verwendet.
In diesem Zusammenhang wäre beispielsweise eine offene Frage, wie man die Matrizen Q und R im Kostenfunktional wählen soll. Einerseits könnte man versuchen durch eine geeignete Wahl von Q und R „schnell“ zu konvergieren und andererseits könnten ungünstige Wahlen von Q und R zu „unphysikalisch“ großen Werten in der Steuerung u führen.
In Kapitel 3 und Kapitel 4 wurden die stationären und instationären Navier-Stokes Gleichungen vorgestellt. Wir haben Grundlagen aus der linearen Funktionalanalysis wiederholt und vorhandene Existenz- und Eindeutigkeitssätze im zweidimensionalen Fall vorgestellt. Wir haben eine Ausflussbedingung vorgestellt, die sich in natürlicherweise aus den Navier-Stokes Gleichungen ergibt. Die mathematische Schwierigkeit bei Wahl dieser Ausflussbedingung ist es, den „Rückfluss“ zu kontrollieren.
Hier kann man überlegen, zu anderen Randbedingungen überzugehen um „freien“ Ausfluss zu modellieren. In [27] wird eine modifizierte Ausflussbedingung vorgestellt.
In Kapitel 5 haben wir uns mit der algebraischen Bernoulligleichung beschäftigt und die Signumsfunktion als ein nützliches Hilfsmittel zum Lösen dieser aufgeführt.
Kapitel 6 befasste sich mit Index-2 Systemen. Bei diesen Systemen ist eine Nebenbedingung derart gegeben, dass die Lösung v den divergenzfreien Unterraum nicht „verlassen“ darf. Durch Projektion auf diesen Unterraum lässt sich diese Bedingung eliminieren. Wir haben dazu passend das Newton-Verfahren und ADI-Verfahren formuliert.
Es sei an dieser Stelle erwähnt, dass wir durch Projektion die Bedingung an den Zustand eliminieren konnten und dann die zugehörige Riccatigleichung lösen, allerdings lassen sich Steuerungsbeschränkungen oder Zustandbeschränkungen nicht einbauen. Beschränkungen dieser Art können zu „Bang-Bang“-Effekten in der Steuerung führen.
Wir haben in Kapitel 7 einige Implementierungsdetails besprochen und in Kapitel 8 zwei Beispiele und die numerischen Ergebnisse formuliert.
Die wesentlichen Schwierigkeiten bei der praktischen Umsetzung sind, dass beim Verwenden eines direkten Lösers viel Hauptspeicher benötigt wird. In Simulationen ist man zu kleinen Zeitschrittweiten gezwungen. Die Eigenvektoren, die zu den Eigenwerten mit positiven Realteil gehören, für große System zu berechnen.

Literatur

[1] H. W. Alt. *Lineare Funktionalanalysis: Eine anwendungsorientierte Einführung*. Springer-Verlag, 2012.

[2] J. Appell und M. Väth. *Elemente der Funktionalanalysis: Vektorräume, Operatoren und Fixpunktsätze*. Vieweg+Teubner Verlag, 2012. ISBN: 9783322802439.

[3] B. Aulbach. *Gewöhnliche Differenzialgleichungen*. Spektrum, Akad. Verlag, 2004.

[4] Z.-z. Bai u. a. „Hermitian and Skew-Hermitian Splitting Methods for Non-Hermitian Positive Definite Linear Systems“. In: *SIAM J. Matrix Anal. Appl* 24 (2001), S. 603–626.

[5] J. Baker, M. Embree und J. Sabino. *Fast singular value decay for Lyapunov solutions with nonnormal coefficients*. arXiv e-prints 1410.8741v1. math.NA. Cornell University, Okt. 2014. URL: http://arxiv.org/abs/1410.8741v1.

[6] S. Balay u. a. „Efficient Management of Parallelism in Object Oriented Numerical Software Libraries“. In: *Modern Software Tools in Scientific Computing*. Hrsg. von E. Arge, A. M. Bruaset und H. P. Langtangen. Birkhäuser Press, 1997, S. 163–202.

[7] S. Balay u. a. *PETSc Users Manual*. Techn. Ber. ANL-95/11 - Revision 3.6. Argonne National Laboratory, 2015. URL: http://www.mcs.anl.gov/petsc.

[8] S. Balay u. a. *PETSc Web page*. http://www.mcs.anl.gov/petsc. 2015. URL: http://www.mcs.anl.gov/petsc.

[9] E. Bänsch und P. Benner. „Stabilization of incompressible flow problems by Riccati-based feedback“. In: *Constrained Optimization and Optimal Control for Partial Differential Equations*. Springer, 2012, S. 5–20.

[10] E. Bänsch u. a. „Riccati-based Boundary Feedback Stabilization of Incompressible Navier–Stokes Flows“. In: *SIAM Journal on Scientific Computing* 37.2 (2015), A832–A858. DOI: 10.1137/140980016. eprint: http://dx.doi.org/10.1137/140980016. URL: http://dx.doi.org/10.1137/140980016.

[11] S. Barrachina, P. Benner und E. S. Quintana-Ortí. „Efficient algorithms for generalized algebraic Bernoulli equations based on the matrix sign function“. In: *Numerical Algorithms* 46.4 (2007), S. 351–368.

[12] M. Beneš und P. Kučera. „Solutions to the Navier-Stokes Equations with Mixed Boundary Conditions in Two-Dimensional Bounded Domains“. In: (2014). eprint: arXiv:1409.4666.

[13] Benner. *Skript zur Vorlesung mathematische System- und Regelungstheorie*. http://www3.math.tu-berlin.de/Vorlesungen/SS11/Kontrolltheorie/Benner-MathSysRegTh-2010.pdf. 2009.

[14] P. Benner. *Computational Methods for Linear Quadratic Optimization*. Berichte aus der Technomathematik, Report 98–04. Available from http://www.math.uni-bremen.de/zetem/berichte.html. 28334 Bremen (Germany): FB 3 – Mathematik und Informatik, Universität Bremen, Aug. 1998.

[15] P. Benner und J. Saak. „Numerical solution of large and sparse continuous time algebraic matrix Riccati and Lyapunov equations: a state of the art survey“. In: *GAMM Mitteilungen* 36.1 (Aug. 2013), S. 32–52. DOI: 10.1002/gamm.201310003.

[16] P. Benner, P. Kürschner und J. Saak. „A Reformulated Low-Rank ADI Iteration with Explicit Residual Factors“. In: *PAMM* 13.1 (2013), S. 585–586. ISSN: 1617-7061. DOI: 10.1002/pamm.201310273. URL: http://dx.doi.org/10.1002/pamm.201310273.

[17] P. Benner, P. Kürschner und J. Saak. „Efficient handling of complex shift parameters in the low-rank Cholesky factor ADI method“. In: *Numerical Algorithms* 62.2 (2013), S. 225–251.

[18] P. Benner, P. Kürschner und J. Saak. „Self-Generating and Efficient Shift Parameters in ADI Methods for Large Lyapunov and Sylvester Equations“. In: *Electronic Transaction on Numerical Analysis* 43 (2014), S. 142–162.

[19] P. Benner, J.-R. Li und T. Penzl. „Numerical solution of large-scale Lyapunov equations, Riccati equations, and linear-quadratic optimal control problems“. In: *Numerical Linear Algebra with Applications* 15.9 (2008), S. 755–777. ISSN: 1099-1506. DOI: 10.1002/nla.622. URL: http://dx.doi.org/10.1002/nla.622.

[20] P. Benner, E. S. Quintana-Ortı und G. Quintana-Ortı. „Solving Large-Scale Generalized Algebraic Bernoulli Equations via the Matrix Sign Function“. In: *Preprint, TU Chemnitz* (2006).

[21] P. Benner, J. Saak und M. M. Uddin. *Balancing based model reduction for structured index-2 unstable descriptor systems with application to flow control*. Preprint MPIMD/14-20. Available from http://www.mpi-magdeburg.mpg.de/preprints/. Max Planck Institute Magdeburg, Nov. 2014.

[22] P. Benner u. a. „Efficient Solution of Large-Scale Saddle Point Systems Arising in Riccati-Based Boundary Feedback Stabilization of Incompressible Stokes Flow“. In: *SIAM Journal on Scientific Computing* 35.5 (2013), S150–S170. DOI: 10.1137/120881312. eprint: http://dx.doi.org/10.1137/120881312. URL: http://dx.doi.org/10.1137/120881312.

[23] S. Bittanti, A. J. Laub und J. C. Willems. *The Riccati Equation*. Springer-Verlag Berlin Heidelberg, 1991.

[24] Å. Björck. *Numerical methods in matrix computations*. Springer, 2015.

[25] F. Boyer und P. Fabrie. *Mathematical tools for the study of the incompressible Navier-Stokes equations and related models*. Bd. 183. Springer Science & Business Media, 2012.

[26] M. Braack. *Finite Elemente Vorlesungsskript*. Jan. 2015. URL: http://www.informatik.uni-kiel.de/~mabr/lehre/skripte/fem-braack.pdf.

[27] M. Braack, P. B. Mucha und W. M. Zajaczkowski. „Directional do-nothing condition for the navier-stokes equations“. In: *Journal of Computational Mathematics* 32.5 (2014), S. 507–521.

[28] D. Braess. *Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. Springer-Verlag, 2013.

[29] H. Choi, M. Hinze und K. Kunisch. „Instantaneous control of backward-facing step flows“. In: *Applied Numerical Mathematics* 31.2 (1999), S. 133–158.

[30] M. Dobrowolski. *Angewandte Funktionalanalysis: Funktionalanalysis, Sobolev-Räume und elliptische Differentialgleichungen*. Springer-Verlag, 2006.

[31] N. S. Ellner und E. L. Wachspress. „Alternating Direction Implicit Iteration for Systems with Complex Spectra“. In: *SIAM Journal on Numerical Analysis* 28.3 (1991), S. 859–870. DOI: 10.1137/0728045.

[32] G. P. Galdi u. a. „Hemodynamical flows“. In: *Delhi Book Store* (2008).

[33] J. D. Gardiner und A. J. Laub. „A generalization of the matrix-sign-function solution for algebraic Riccati equations“. In: *International Journal of Control* 44.3 (1986), S. 823–832.

[34] V. Girault und P.-A. Raviart. *Finite element methods for Navier-Stokes equations: theory and algorithms, vol. 5 of Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, New-York, 1986.

[35] M. Heinkenschloss, D. C. Sorensen und K. Sun. „Balanced Truncation Model Reduction for a Class of Descriptor Systems with Application to the Oseen Equations“. In: *SIAM Journal on Scientific Computing* 30.2 (2008), S. 1038–1063. DOI: 10.1137/070681910. eprint: http://dx.doi.org/10.1137/070681910. URL: http://dx.doi.org/10.1137/070681910.

[36] A. Henderson und J. Ahrens. *The Paraview guide : a parallel visualization application*. New York: Kitware, Inc., 2004. ISBN: 1-930934-14-9. URL: http://opac.inria.fr/record=b1117983.

[37] N. J. Higham. *Functions of Matrices: Theory and Computation*. Philadelphia, PA, USA: Society for Industrial und Applied Mathematics, 2008, S. xx+425. ISBN: 978-0-898716-46-7.

[38] E. Jones, T. Oliphant, P. Peterson u. a. *SciPy: Open source scientific tools for Python*. [Online; accessed 2015-07-19]. 2001–. URL: http://www.scipy.org/.

[39] W. Kaballo. *Aufbaukurs Funktionalanalysis und Operatortheorie: Distributionen-lokalkonvexe Methoden-Spektraltheorie*. Springer-Verlag, 2014.

[40] H. W. Knobloch und H. Kwakernaak. *Lineare Kontrolltheorie*. In German. Berlin: Springer, 1985.

[41] H. Knobloch und F. Kappel. *Gewöhnliche Differentialgleichungen*. Mathematische Leitfäden. B. G. Teubner, 1974. URL: http://books.google.de/books?id=9UTvAAAAMAAJ.

[42] P. Lancaster und M. Tismenetsky. *The Theory of Matrices*. 2nd. Orlando: Academic Press, 1985.

[43] P. Lancaster und L. Rodman. *Algebraic riccati equations*. Oxford University Press, 1995.

[44] A. J. Laub. *Matrix analysis for scientists and engineers*. Siam, 2005.

[45] R. B. Lehoucq, D. C. Sorensen und C. Yang. *ARPACK Users Guide: Solution of Large Scale Eigenvalue Problems by Implicitly Restarted Arnoldi Methods*. 1997.

[46] J.-R. Li und J. White. „Low rank solution of Lyapunov equations“. In: *SIAM Journal on Matrix Analysis and Applications* 24.1 (2002), S. 260–280.

[47] X. S. Li. „An Overview of SuperLU: Algorithms, Implementation, and User Interface“. In: *ACM Trans. Math. Softw.* 31.3 (Sep. 2005), S. 302–325. ISSN: 0098-3500. DOI: 10.1145/1089014.1089017. URL: http://doi.acm.org/10.1145/1089014.1089017.

[48] A. Logg, K.-A. Mardal und G. Wells. *Automated solution of differential equations by the finite element method: The FEniCS book*. Bd. 84. Springer Science & Business Media, 2012.

[49] K. Meerbergen und D. Roose. „Matrix transformations for computing rightmost eigenvalues of large sparse non-symmetric eigenvalue problems“. In: *IMA J. Numer. Anal* 16 (1996), S. 297–346.

[50] J. Mikusinski. *The Bochner Integral*. Lehrbücher und Monographien aus dem Gebiete der exakten Wissenschaften. Birkhäuser Basel, 2013. ISBN: 9783034855679.

[51] T. Penzl. „A cyclic low rank Smith method for large sparse Lyapunov equations“. In: *SIAM Journal on Scientific Computing* 21.4 (2000), S. 1401–1418.

[52] T. Penzl. „Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case“. In: *System Control Letters* 40 (2 2000), S. 139–144.

[53] H.-G. Roos, M. Stynes und L. Tobiska. *Robust numerical methods for singularly perturbed differential equations: convection-diffusion-reaction and flow problems*. Bd. 24. Springer Science & Business Media, 2008.

[54] M. Ruzicka. *Nichtlineare Funktionalanalysis: Eine Einführung*. Springer-Verlag, 2006.

[55] B. Schweizer. *Partielle Differentialgleichungen: Eine anwendungsorientierte Einführung*. Springer-Verlag, 2013.

[56] I. R. Shafarevich und A. Remizov. *Linear algebra and geometry*. Springer Science & Business Media, 2012.

[57] D. C. Sorensen und Y. Zhou. *Bounds on Eigenvalue Decay Rates and Sensitivity of Solutions to Lyapunov Equations*. Techn. Ber. Houston, TX: Dept. of Comp. Appl. Math., Rice University, 2002.

[58] J. Stoer und R. Bulirsch. *Numerische Mathematik 2: Eine Einführung - unter Berücksichtigung von Vorlesungen von F.L.Bauer*. Numerische Mathematik: eine Einführung - unter Berücksichtigung von Vorlesungen von F. L. Bauer. Springer, 2005. ISBN: 9783540237778.

[59] E. L. Wachspress. *The ADI Model Problem*. Springer New York, 2013. ISBN: 978-1-4614-5121-1. DOI: 10.1007/978-1-4614-5122-8.

[60] W. Walter. *Gewöhnliche Differentialgleichungen: Eine Einführung*. Springer-Lehrbuch. Springer Berlin Heidelberg, 2000. ISBN: 9783540676423. URL: http://books.google.de/books?id=tyAdMH69NRYC.

[61] H. K. Weichelt. „Feedback-Stabilisierung von instationären, inkompressiblen Strömungen mit Riccati-Ansatz“. Diplomarbeit. D-09107 Chemnitz: Technische Universität Chemnitz, Dez. 2010.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe.

Ort, Datum, Unterschrift