Adaptive Verfahren zur numerischen Berechnung von Reaktions-Diffusions-Systemen

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

genehmigt durch die Fakultät für Mathematik

der Otto-von-Guericke-Universität Magdeburg

von Dipl.-Math. Wolfram Heineken

geb. am 25. Februar 1972 in Dresden

Gutachter:

Prof. Dr. Klaus Deckelnick

Prof. Dr. Stefan C. Müller

Prof. Dr. Gerald Warnecke

Prof. Dr. Rüdiger Weiner

Eingereicht am 1. September 2004

Verteidigung am 10. Mai 2005

„Jeder, der sich die Fähigkeit
erhält, Schönes zu erkennen,
wird nie alt werden.“

Franz Kafka ⟨1883 – 1924⟩

Vorwort

Die vorliegende Arbeit entstand während meiner Tätigkeit am Institut für Analysis und Numerik der Otto-von-Guericke-Universität Magdeburg unter der Betreuung von Herrn Prof. Gerald Warnecke, dem ich an dieser Stelle für seine Unterstützung der Arbeit, manche nützlichen Hinweise und nicht zuletzt für sein Engagement in der DFG-Forschergruppe „Grenzflächendynamik bei Strukturbildungsprozessen“, welches mich an die Problematik der erregbaren Medien herangeführt hat, herzlich Dank sagen möchte. Meinem Kollegen Dr. Matthias Kunik danke ich für die Durchsicht von Teilen des Manuskripts und einige Verbesserungsvorschläge. Mein Dank gilt weiterhin Herrn Dr. Niklas Manz, mit dem ich in dem Projekt „Erregungsfronten in der Cyclohexandion-BZ-Reaktion auf gekrümmten Oberflächen“ zusammengearbeitet habe. Herrn Dr. Walfred Grambow danke ich für seinen Einsatz, wenn es galt, verschiedene Probleme am Rechner zu beheben. Schließlich danke ich meinen jetzigen und ehemaligen Kollegen am Institut für Analysis und Numerik, denen ich mich freundschaftlich verbunden fühle und die zu einer angenehmen Arbeitsatmosphäre beigetragen haben. Neben vielen anderen möchte ich hier besonders Dr. Nikolai Andrianov, Dr. Yousef Zahaykah und Dr. Matthias Kunik nennen.

Magdeburg, im Juni 2004 Wolfram Heineken

Inhaltsverzeichnis

Kapitel 1

Einleitung

Reaktions-Diffusions-Gleichungen sind eine spezielle Klasse partieller Differentialgleichungen parabolischen Typs. Gleichungen dieser Art ergeben sich bei der Modellierung verschiedener in den Naturwissenschaften betrachteter Vorgänge. Der Name der Gleichungen stammt von der Beschreibung chemischer Reaktionen. Dabei werden zwei wesentliche Prozesse, die Stoffumwandlung und der Konzentrationsausgleich durch Diffusion, mit Hilfe der Reaktions-Diffusions-Gleichungen ausgedrückt. Auch in anderen Gebieten, etwa in der Biologie und der Physik, treten Differentialgleichungen vom Reaktions-Diffusions-Typ auf. Kapitel 2 enthält eine kurze Zusammenstellung einiger klassischer Reaktions-Diffusions-Systeme.

Die Lösung von Reaktions-Diffusions-Systemen ist oft auf analytischem Wege nicht mehr möglich. Nur in einigen einfachen Fällen kann man die Lösung exakt angeben. In der überwiegenden Mehrheit der Fälle muß daher auf numerische Näherungsverfahren zurückgegriffen werden. Der erste Schritt zur Bestimmung einer Näherungslösung ist eine Diskretisierung des Problems.

Die unabhängigen Veränderlichen in einem Reaktions-Diffusions-System sind die Variable $\mathbf{x} \in \mathbb{R}^n$, die in einem Gebiet Ω liegt, und die Variable $t \in \mathbb{R}$. In den meisten Anwendungen beschreibt $\mathbf{x}$ den Ort und t die Zeit. Auch die von uns betrachtete Form der Diskretisierung, die Linienmethode, folgt dieser Einteilung: Zuerst wird die Reaktions-Diffusions-Gleichung bezüglich der Ortsvariablen $\mathbf{x}$ diskretisiert. Im Ergebnis erhält man ein System gewöhnlicher Differentialgleichungen in der Zeitvariablen t. Die Diskretisierung bezüglich t liefert schließlich ein algebraisches Gleichungssystem, aus dessen Lösung die Näherungslösung der Reaktions-Diffusions-Gleichung konstruiert wird. Wir betrachten in dieser Arbeit lediglich Probleme, in denen das Gebiet Ω ein- oder zweidimensional ist. Die Ortsdiskretisierung nehmen wir mit Hilfe der Methode der finiten Elemente vor, die Zeitdiskretisierung mit speziellen linear-impliziten Runge-Kutta-Verfahren, den sogenannten W-Methoden. Diese haben den Vorteil daß das resultierende Gleichungssystem linear ist. Orts- und Zeitdiskretisierung werden in Kapitel 3 und 5 beschrieben. Kapitel 6 ist der Lösung der im Ergebnis der Diskretisierung auftretenden linearen Gleichungssysteme gewidmet.

Die Menge Ω ist in vielen Fällen ein Gebiet in $\mathbb{R}^n$. Vor dem Hintergrund einiger Anwendungen sind aber auch Systeme interessant, bei denen Ω ein Gebiet auf einer gekrümmten Fläche ist. In Kapitel 10 werden wir bei der Modellierung erregbarer Medien auf gekrümmten Flächen

derartige Probleme betrachten. Die Methode der finiten Elemente läßt sich in eleganter Weise auch auf glatten gekrümmten Flächen formulieren. Von Dziuk [54] wurde ein Verfahren zur Diskretisierung der Laplace-Beltrami-Gleichung auf gekrümmten Flächen angegeben. Auf der Grundlage dieses Verfahrens formulieren wir eine Methode zur Ortsdiskretisierung von Reaktions-Diffusions-Systemen.

Bei der Methode der finiten Elemente wird das Gebiet Ω zunächst mit einer Triangulierung überzogen. Wir beschränken uns auf die Triangulierung zweidimensionaler Gebiete durch Dreiecksgitter. Ist der Diskretisierungsfehler nicht gleichmäßig über Ω verteilt, so erhöht sich die Effizienz des Näherungsverfahrens, wenn das Gitter entsprechend angepaßt wird. In den Bereichen, in denen ein hoher Fehler vorliegt, sollte die Triangulierung besonders fein sein. Ein solches Vorgehen wird als Gitteradaption bezeichnet. In Kapitel 4 beschreiben wir einige von uns entwickelte Strategien zur Steuerung der Gitteradaption.

In vielen Fällen wird durch die Gitteradaption eine Gleichverteilung des Fehlers über das Gebiet Ω angestrebt. Wir stellen hingegen eine Variante der Adaption vor, die eine differenziertere Einflußnahme auf das zu erzeugende Gitter erlaubt. Dabei kann jedem Wert des Fehlers eine gewünschte Gitterfeinheit zugeordnet werden. Auf diese Weise kann beispielsweise die Umgebung von Fronten der Lösung noch stärker aufgelöst werden, als es eine Gleichverteilung des Fehlers erfordern würde. Wir stellen ein numerisches Beispiel vor, bei dem die Gleichverteilung des Fehlers nicht die effizienteste Variante darstellt. Durch stärkere Verfeinerung der Umgebung der Front kann die Effizienz hier deutlich gesteigert werden. Auch für die Wahl der Zeitpunkte, in denen eine Gitteradaption vorgenommen wird, werden verschiedene Strategien vorgestellt.

Für die bereits oben angesprochenen Probleme auf einer gekrümmten Fläche muß diese Fläche trianguliert werden. In Abschnitt 4.5 entwickeln wir Methoden zur adaptiven Triangulierung einer Sphäre und eines Ellipsoids durch Projektion eines ebenen Gitters.

Zur Lösung der im Zuge der Diskretisierung entstehenden linearen Gleichungssysteme sind iterative Verfahren besonders geeignet. Wir konzentrieren uns auf zwei Vertreter – das BiCGstab-Verfahren von van der Vorst [166] und den multiplen Arnoldi-Prozeß von Schmitt und Weiner [143], der auf dem Verfahren von Arnoldi [8] basiert. Für eine effiziente Lösung ist der rechtzeitige Abbruch dieser Iterationsverfahren von entscheidender Bedeutung. Wir übertragen ein Abbruchkriterium, das von Blom, Verwer und Trompert für iterative Gleichungslöser in impliziten BDF-Verfahren entwickelt wurde, auf das BiCGstab-Verfahren. Im Falle des multiplen Arnoldi-Prozesses nutzen wir die Abbruchbedingung, die von Schmitt und Weiner [143] angegeben wurde. In beiden Fällen wird die Toleranz für das Residuum des Gleichungslösers an die Toleranz für den zeitlichen lokalen Fehler, die in der Zeitschrittsteuerung der W-Methode verwendet wird, gekoppelt. Ein Kopplungsfaktor muß jedoch durch numerische Testrechnungen noch bestimmt werden. In Kapitel 9 wird für drei ausgewählte Modellprobleme eine Einstellung dieses Faktors vorgenommen.

Eine Schwierigkeit bei der Lösung von Reaktions-Diffusions-Systemen stellt die mögliche Steifheit der Probleme dar. Ursache der Steifheit kann einerseits der Diffusionsterm im Zusammenhang mit starker Gitterverfeinerung sein, andererseits aber auch steile Gradienten im Reaktionsterm. Steifheit ist ein Grund für die Verwendung stabiler impliziter Zeitdiskretisierungs-Verfahren. In vielen Fällen tritt Steifheit lokal sehr unterschiedlich auf, beispielsweise nur an einer stark verfeinerten Front. In einem solchen Falle kann ein lokales Partitionierungs-

Verfahren möglicherweise zu einer erheblichen Steigerung der Effizienz beitragen. In Kapitel 8 stellen wir verschiedene Varianten lokaler Partitionierung dar. Ein lokales Partitionierungs-Verfahren, das auf einer W-Methode basiert, wurde 1993 von Weiner, Arnold, Rentrop und Strehmel [167] angegeben. In Kapitel 8 stellen wir eine von uns entwickelte Modifikation dieses Verfahrens vor. Zum einen wurde die Erkennung steifer Komponenten etwas abgeändert, zum anderen besteht bei unserem Verfahren die Möglichkeit, Diffusions- und Reaktionsterm getrennt auf Steifheit zu untersuchen und zu partitionieren. Auf diese Weise können Probleme, bei denen die Steifheit nur von einem der beiden Terme ausgeht, geeignet behandelt werden.

Eine weitere Möglichkeit, auf Steifheit zu reagieren, ist der Einsatz spezieller linearer Löser, die den dominanten und daher für die Steifheit verantwortlichen Eigenraum besonders schnell approximieren. Der multiple Arnoldi-Prozeß von Schmitt und Weiner ist ein solches Verfahren. Das Arnoldi-Verfahren wurde in einer Reihe von Arbeiten [33, 143, 34, 169, 144, 168] stets zur Lösung unpartitionierter Systeme verwendet. Wir benutzen diesen Löser auch in unseren lokalen Partitionierungs-Verfahren. Damit werden die beiden für lokal steife Probleme entwickelten Ansätze – lokale Partitionierung und ein den dominanten Eigenraum schnell approximierender Löser – kombiniert.

In Kapitel 9 werden umfangreiche numerische Untersuchungen mit verschiedenen Varianten von Partitionierungs-Verfahren präsentiert. Es wurden drei Reaktions-Diffusions-Probleme ausgewählt, die in unterschiedlichem Maße zur Partitionierung geeignet sind. In den Lösungen aller drei Probleme treten bewegte Fronten auf. Eine Differentialgleichung mit ortsabhängigem Diffusionskoeffizient wurde derart konstruiert, daß die Steifheit lokal eng begrenzt gehalten wird. Bei diesem Problem zeigt sich der Erfolg lokaler Partitionierung erwartungsgemäß am deutlichsten. Die Suche nach einem effizienten Lösungsverfahren für das System von Krinsky et al. ist für die in Kapitel 10 dargestellten umfangreichen Langzeitberechnungen dieses Problems von großer Bedeutung.

Um einen Effizienzvergleich der einzelnen Verfahren zu ermöglichen, mußte zunächst für jedes Verfahren eine möglichst günstige Abbruchbedingung für den iterativen linearen Löser gefunden werden. Zu diesem Zwecke wurden umfangreiche numerische Untersuchungen an den drei Testbeispielen durchgeführt. Eine weitere Studie befaßt sich mit dem Einfluß von Gitterfeinheit und Toleranz des zeitleich lokalen Fehlers auf die Genauigkeit der Lösung. In einer abschließenden Untersuchung wird die Effizienz verschiedener Verfahren mit und ohne Partitionierung durch eine Gegenüberstellung von Fehler und Rechenaufwand verglichen. Die bei den drei Problemen unterschiedlichen Ursachen für Steifheit werden erläutert.

Die Simulation erregbarer Medien ist ein interessanter Anwendungsfall für Reaktions-Diffusions-Systeme. Erregbare Medien besitzen einen Gleichgewichtszustand, jedoch reicht bereits eine relativ geringe Störung aus, damit sie die Gleichgewichtslage verlassen und einen erregten Zustand einnehmen, der erst nach längerer Zeit wieder abklingt. Der erregte Bereich kann sich dabei in der Art einer Welle durch den Raum bewegen. Bei einer Reihe chemischer Reaktionen – etwa der Belousov-Zhabotinsky-Reaktion [21, 176] zwischen Malonsäure und Bromat-Ionen – können die beteiligten Stoffe als erregbares Medium aufgefaßt werden. Hierbei bildet beispielsweise eine niedrige Stoffkonzentration die Gleichgewichtslage und eine hohe Konzentration den erregten Zustand. Die Ausbreitung von Wellen kann im Experiment optisch beobachtet werden.

Derartige Erregungswellen treten auch in anderen Bereichen auf. Wellenbewegungen im Herzmuskel, die bereits 1946 von WIENER und ROSENBLUETH [172] untersucht wurden, sind von besonderem medizinischen Interesse, da sie eine Ursache von Herzrhytmusstörungen darstellen [46].

In Kapitel 10 präsentieren wir eine Reihe von Ergebnissen der Simulation erregbarer Medien. Wir konzentrieren uns dabei besonders auf das Modellproblem von KRINSKY et al.. Im räumlich eindimensionalen Fall untersuchen wir, wie sich Gitterfeinheit und Toleranz des zeitlichen Fehlers auf die Genauigkeit der Lösung auswirken. Ferner studieren wir den Einfluß gewisser Systemparameter auf die Geschwindigkeit der auftretenden bewegten Wellen. Interessanter ist der Fall eines räumlich mehrdimensionalen Mediums. Wir beschränken uns hier auf den zweidimensionalen Fall. Das Medium liegt in einem Gebiet Ω auf einer Fläche S. Wir betrachten Fälle, in denen die Fläche S eine Ebene, eine Sphäre oder ein Ellipsoid ist. Bei diesen räumlich zweidimensionalen erregbaren Medien können, in Abhängigkeit von dem zugrundeliegenden Modell und dessen Parametern, rotierende Spiralwellen auftreten. Form und Bewegung dieser Wellen können mit Hilfe der kinematischen Theorie [172, 180, 29] näherungsweise beschrieben werden. Bei geeignet gewählten Parametern rotieren die Spiralwellen um einen *festen* Punkt, eine Bewegung, die wir als stationäre Rotation bezeichnen.

Ein wichtiges Merkmal der Bewegung der Welle ist die Bahn, die die Wellenspitze beschreibt. Bei einer stationären Rotation ist die Bahnkurve ein Kreis. Eine Störung der stationären Rotation tritt bei besonders kurzen Wellen in der Ebene und bei Wellen auf einer nichtgleichmäßig gekrümmten Fläche auf. Dabei bewegt sich das Rotationszentrum in eine bestimmte Richtung, ein Vorgang, der als Drift bezeichnet wird. In beiden Fällen wird eine solche Drift durch unsere numerischen Untersuchungen bestätigt.

Durch Auswertung einer umfangreichen numerischen Studie zur Wellendrift auf verschiedenen Ellipsoiden ermitteln wir einen Zusammenhang zwischen der Gaußschen Krümmung der Fläche, dem Erregungsparameter in den Modellgleichungen und der Wellendrift. Überraschenderweise ergibt sich hier ein Widerspruch zwischen unseren numerischen Ergebnissen und den Aussagen der kinematischen Theorie in [50], der bisher nicht geklärt werden konnte.

Zusammenfassend läßt sich die vorliegende Arbeit in der folgenden Weise umreißen: Der eine Schwerpunkt liegt in der Entwicklung effizienter Lösungsverfahren für Reaktions-Diffusions-Systeme. Dabei soll lokal steifen Problemen durch den Einsatz lokaler Partitionierungs-Verfahren in der Zeitdiskretisierung Rechnung getragen werden. Bei der Simulation erregbarer Medien kommen diese Diskretisierungsverfahren zum Einsatz. Umfangreiche numerische Studien – insbesondere zur Wellendrift auf Ellipsoiden – bilden einen zweiten Schwerpunkt der Arbeit.

Kapitel 2

Reaktions-Diffusions-Systeme

2.1 Ein Überblick

Es gibt eine Reihe von physikalischen, chemischen oder auch biologischen Prozessen, die durch **Reaktions-Diffusions-Systeme** mathematisch modelliert werden können. Ein Reaktions-Diffusions-System ist ein System partieller Differentialgleichungen, welches in der Form

$$\frac{\partial u_k}{\partial t}(\mathbf{x},t) = \text{div } \mathcal{D}_k(\mathbf{u}, \nabla\mathbf{u}, \mathbf{x}, t) + \mathcal{R}_k(\mathbf{u}, \mathbf{x}, t), \qquad k = 1, \ldots, m \tag{2.1}$$

dargestellt werden kann. Die Gleichungen gelten für alle reellen Zahlen t in einem gewissen Intervall $[t_0, t_e]$, sowie für alle Vektoren $\mathbf{x}$ aus einem Gebiet Ω. In den meisten Anwendungen beschreibt der Vektor $\mathbf{x}$ den Ort und die Variable t die Zeit. Die Lösungen $u_k(\mathbf{x}, t)$ des Systems fassen wir in dem Lösungsvektor $\mathbf{u}(\mathbf{x}, t) = (u_1(\mathbf{x}, t), \ldots, u_m(\mathbf{x}, t)) \in \mathbb{R}^m$ zusammen.

Wir betrachten in dieser Arbeit die folgenden beiden Fälle:

1. Es sei Ω ein Gebiet in $\mathbb{R}^n$. In diesem Falle sind ∇ und div Gradient- bzw. Divergenzoperator in $\mathbb{R}^n$, die Anwendung auf $\mathbf{u}$ geschieht komponentenweise. Der Divergenzoperator kann auch in der Form $\nabla\cdot$ geschrieben werden.

2. Es sei Ω ein Gebiet auf einer differenzierbaren Mannigfaltigkeit $S \subset \mathbb{R}^n$ mit der Dimension dim $(S) < n$. In diesem Falle bezeichnet ∇ den tangentialen Gradienten und div die tangentiale Divergenz *bezüglich der Mannigfaltigkeit*[1] S.

Die Symbole $\mathcal{D}_k$ und $\mathcal{R}_k$ stehen für gewisse Funktionen

$$\mathcal{D}_k : \mathbb{R}^m \times \mathbb{R}^{mn} \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n, \qquad k = 1, \ldots, m$$

und

$$\mathcal{R}_k : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}, \qquad k = 1, \ldots, m.$$

Die Funktion $\mathcal{D}_k$ ist bezüglich ihrer ersten drei Variablen differenzierbar.

[1] Siehe dazu Abschnitt A im Anhang.

Ein klassisches Anwendungsgebiet von Systemen der Form (2.1), die Modellierung chemischer Reaktionen, hat den Reaktions-Diffusions-Systemen den Namen gegeben. Der Vektor $\mathbf{u}$ beschreibt in diesem Falle die Konzentrationen von m beteiligten Stoffen. Dabei werden Diffusionsprozesse durch den Term div $\mathcal{D}_k$ und die spezifische Reaktionskinetik durch den Operator $\mathcal{R}_k$ modelliert.

Reaktions-Diffusions-Systeme sind eine spezielle Klasse der **Reaktions-Konvektions-Diffusions-Systeme**

$$\frac{\partial u_k}{\partial t}(\mathbf{x},t) = \text{div } [\mathcal{D}_k(\mathbf{u},\nabla\mathbf{u},\mathbf{x},t) + \mathcal{C}_k(\mathbf{u},\mathbf{x},t)] + \mathcal{R}_k(\mathbf{u},\mathbf{x},t), \qquad k = 1,\ldots,m.$$

Mit Hilfe des Konvektionstermes div $\mathcal{C}_k(\mathbf{u},\mathbf{x},t)$ können zusätzlich Strömungsprozesse berücksichtigt werden. Einige Probleme aus den Naturwissenschaften, die auf gewisse Reaktions-Konvektions-Diffusions-Systeme führen, sind beispielsweise bei Amann [7] dargestellt. Darunter sind

- Probleme der Populationsdynamik,
- Konvektionsprozesse in porösen Medien,
- Diffusion in Polymeren,
- elektrolytische Prozesse,
- Phasenübergänge (Schmelzen, Erstarren, das sogenannte Stefan-Problem).

Wir wenden uns nun wieder den Reaktions-Diffusions-Systemen zu.

Um die Lösung eines Reaktions-Diffusions-Systems eindeutig festzulegen, müssen gewisse Bedingungen an die Lösung gestellt werden. In vielen Fällen ist es sinnvoll, **Anfangs-** und **Randbedingungen** zu formulieren. Die Anfangsbedingung beschreibt die Lösung zum Zeitpunkt t_0. Sie ist von der Gestalt

$$\mathbf{u}(\mathbf{x},t_0) = \mathbf{u}_0(\mathbf{x}),$$

wobei $\mathbf{u}_0$ eine fest vorgegebene Funktion ist. Randbedingungen sind Vorgaben der Lösung auf dem Rand des Gebietes Ω. Die vor dem Hintergrund praktischer Anwendungen am häufigsten verwendeten Randbedingungen sind die **Dirichlet-Randbedingungen**

$$\mathbf{u}(\mathbf{x},t) = \mathbf{g}_{\text{Dir}}(\mathbf{x},t) \qquad \text{für } \mathbf{x} \in \partial\Omega$$

oder die **Neumann-Randbedingungen**

$$\mathcal{D}_k(\mathbf{u},\nabla\mathbf{u},\mathbf{x},t)\cdot\mathbf{n}_{\partial\Omega} = g_{\text{Neu},k}(\mathbf{x},t) \qquad \text{für } \mathbf{x} \in \partial\Omega,\ k = 1,\ldots,m.$$

Hierbei ist $\mathbf{n}_{\partial\Omega}$ der nach außen gerichtete Normalvektor auf dem Rand $\partial\Omega$. Die Funktionen $\mathbf{g}_{\text{Dir}}$ bzw. $g_{\text{Neu},k}$ müssen vorgegeben werden. Gilt $\mathbf{g}_{\text{Dir}}(\mathbf{x},t) = 0$ bzw. $g_{\text{Neu},k}(\mathbf{x},t) = 0$ für alle $\mathbf{x} \in \partial\Omega$ und alle $t \in [t_0,t_e]$, so spricht man von **homogenen Dirichlet-** bzw. **homogenen Neumann-Randbedingungen**.

Durch gewisse Bedingungen an die einzelnen in (2.1) auftretenden Terme erhält man die im folgenden aufgeführten Klassen von Reaktions-Diffusions-Gleichungen, siehe etwa Knabner/Angermann [94, Abschnitt 0.4]:

Definition 2.1. Quasilineare Reaktions-Diffusions-Systeme sind solche, bei denen $\mathcal{D}_k$ *linear* von $\nabla\mathbf{u}$ abhängt. Schreiben wir $\nabla\mathbf{u}$ in der Form

$$\nabla\mathbf{u} = \begin{pmatrix} \nabla u_1 \\ \vdots \\ \nabla u_m \end{pmatrix} \in \mathbb{R}^{mn},$$

dann existiert in diesem Falle eine Matrix $\mathbf{D}_k(\mathbf{u}, \mathbf{x}, t) \in \mathbb{R}^{n \times mn}$, so daß $\mathcal{D}_k = \mathbf{D}_k \nabla\mathbf{u}$ gilt.

Semilineare Reaktions-Diffusions-Systeme sind quasilinear, und es gilt zusätzlich, daß $\mathcal{D}_k$ bzw. die Matrix $\mathbf{D}_k$ *nicht von* $\mathbf{u}$ *abhängt*.

Lineare Reaktions-Diffusions-Systeme sind semilineare, bei denen $\mathcal{R}_k$ *nur linear* von $\mathbf{u}$ abhängt. □

In vielen Anwendungen treten semilineare Reaktions-Diffusions-Systeme mit dem vereinfachten Diffusionsterm div $\mathcal{D}_k$ = div $(d_k(\mathbf{x}, t)\nabla u_k)$ auf, wobei $d_k(\mathbf{x}, t)$ eine *skalare* Funktion ist. Wir werden uns im Rahmen dieser Arbeit auf derartige Probleme beschränken. Sie können in der folgenden Form angegeben werden:

Semilineares Reaktions-Diffusions-System

$$\begin{aligned} \frac{\partial u_k}{\partial t}(\mathbf{x}, t) &= \text{div } (d_k(\mathbf{x})\nabla u_k(\mathbf{x}, t)) + r_k(\mathbf{x})u_k(\mathbf{x}, t) \\ &\quad + p_k(u_1(\mathbf{x}, t), \ldots, u_m(\mathbf{x}, t)) + q_k(\mathbf{x}, t), \qquad k = 1, \ldots, m,\ \mathbf{x} \in \Omega \subset \mathbb{R}^n, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad t \in [t_0, t_e], \\ u(\mathbf{x}, t_0) &= u_0(\mathbf{x}), \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \mathbf{x} \in \Omega \subset \mathbb{R}^n \end{aligned} \tag{2.2}$$

versehen mit **Dirichlet-Randbedingungen**

$$u_k(\mathbf{x}, t) = g_{\text{Dir},k}(\mathbf{x}, t), \qquad k = 1, \ldots, m, \quad \mathbf{x} \in \partial\Omega$$

oder **Neumann-Randbedingungen**

$$d_k(\mathbf{x}, t)\nabla u_k \cdot \mathbf{n}_{\partial\Omega} = g_{\text{Neu},k}(\mathbf{x}, t), \qquad k = 1, \ldots, m, \quad \mathbf{x} \in \partial\Omega$$

Ist zusätzlich der Diffusionskoeffizient d_k nicht von $\mathbf{x}$ abhängig, so hat der Diffusionsterm die Form div $\mathcal{D}_k = d_k \Delta u_k$. Falls Ω ein Gebiet in $\mathbb{R}^n$ ist, steht Δ für den Laplace-Operator; falls Ω ein Gebiet auf der Mannigfaltigkeit S ist, so bezeichnet Δ den Laplace-Beltrami-Operator. Einige Beispiele semilinearer Reaktions-Diffusions-Gleichungen sollen im folgenden Abschnitt angegeben werden.

Bemerkung 2.2. In den Kapiteln 3 bis 9 und im Anhang werden die verschiedenen Bedeutungen von ∇, div und Δ auch durch unterschiedliche Bezeichnung deutlich gemacht. Wir schreiben dann

- ∇ für den Gradienten bezüglich $\mathbb{R}^n$,
- ∇_S für den tangentialen Gradienten auf einer Mannigfaltigkeit S des $\mathbb{R}^n$,
- $\nabla\cdot$ für die Divergenz bezüglich $\mathbb{R}^n$,
- div_S für die tangentiale Divergenz auf einer Mannigfaltigkeit S des $\mathbb{R}^n$,
- Δ für den Laplace-Operator bezüglich $\mathbb{R}^n$,
- Δ_S für den Laplace-Beltrami-Operator auf einer Mannigfaltigkeit S des $\mathbb{R}^n$.

Tangentialer Gradient, tangentiale Divergenz und Laplace-Beltrami-Operator sind im Anhang in Abschnitt A.1 definiert.

2.2 Beispiele für semilineare Reaktions-Diffusions-Gleichungen

Wir beginnen die Zusammenstellung mit zwei skalaren semilinearen Reaktions-Diffusions-Gleichungen.

Die Fisher-Gleichung

Die Gleichung wurde von Fisher zu populationsgenetischen Untersuchungen herangezogen und von Kolmogorov, Petrovsky und Piskunov [96] (1937) analytisch untersucht. Sie ist von der Form

$$u_t = \Delta u + f(u),$$

wobei $f(u)$ eine Funktion mit zwei Nullstellen ist.

Die bistabile Diffusions-Gleichung

Diese skalare Gleichung wurde u.a. in Verbindung mit Verbrennungsmodellen von Kanel [91] (1962) und im Bereich der Populationsgenetik von Aronson und Weinberger [9] (1978) untersucht. Sie ist von der Form

$$u_t = \Delta u + r(u),$$

wobei die Reaktionsfunktion r drei einfache Nullstellen u_1, u_2, u_3 hat und $r'(u_1) < 0$, $r'(u_2) > 0$ und $r'(u_3) < 0$ gilt.

Die van der Pol'schen Gleichungen mit Diffusion

Dieses System von Reaktions-Diffusions-Gleichungen hat die Form

$$\begin{aligned} u_t &= d_1\Delta u + v, \\ v_t &= d_2\Delta v + \varepsilon(1-u^2)v - u. \end{aligned}$$

Die Gleichungen wurden, ohne den Diffusionsterm, 1926 von VAN DER POL [131] zur Beschreibung elektrischer Stromkreise angegeben. Wegen ihrer Steifheit ist diese Gleichung, auch in der Form mit Diffusionsterm, ein beliebtes Testbeispiel für verschiedene numerische Verfahren, siehe etwa JACKSON und SEWARD [86], SHAMPINE [148], HAIRER und WANNER [78] oder WEINER et al. [169].

Die folgenden drei Beispiele semilinearer Reaktions-Diffusions-Systeme finden sich beispielsweise bei SMOLLER [150].

Die Hodgkin-Huxley-Gleichungen

Die Gleichungen wurden 1952 von HODGKIN und HUXLEY [84] zur Modellierung der Signalübertragung in Nervenbahnen aufgestellt.[2] Das vierkomponentige System hat die Form

$$\begin{aligned} cu_t &= R^{-1}u_{xx} + g(u,v,w,z), \\ v_t &= \varepsilon_1 v_{xx} + g_1(u)(h_1(u) - v), \\ w_t &= \varepsilon_2 w_{xx} + g_2(u)(h_2(u) - w), \\ z_t &= \varepsilon_3 z_{xx} + g_3(u)(h_3(u) - z), \end{aligned} \tag{2.3}$$

wobei $g(u,v,w,z) = k_1 v^3 w(c_1 - u) + k_2 z^4(c_2 - u) + k_3(c_3 - u)$ ist und die Bedingungen $c, R, k_i > 0$, $c_1 > c_3 > 0 > c_2$, $\varepsilon_i \geq 0$, $g_i(u) > 0$, $1 > h_i(0) > 0$ gelten.

Die FitzHugh-Nagumo-Gleichungen

Diese Gleichungen stellen eine Vereinfachung des oben angegebenen HODGKIN-HUXLEY-Systems dar. Sie wurden von FITZHUGH (1961) [66] und von NAGUMO, ARIMOTO und YOSHIZAWA [124] (1964) angegeben. Das System hat die Form

$$\begin{aligned} u_t &= u_{xx} + f(u) - v, \\ v_t &= \varepsilon v_{xx} + (\sigma u - \gamma v), \end{aligned} \tag{2.4}$$

wobei $\sigma, \gamma > 0$, $\varepsilon \geq 0$ ist und $f(u)$ qualitativ von der Form eines kubischen Polynoms $-u(u-a)(u-b)$ mit $0 < a < b$ ist.

Das Brusselator-Modell

Die Gleichungen wurde 1971 von LEFEVER und NICOLIS [107] aufgestellt. Sie gelten als das klassische Beispiel für oszillierende chemische Systeme und haben die Gestalt

$$\begin{aligned} u_t &= d_1 \Delta u + a + u^2 v - (b+1)u, \\ v_t &= d_2 \Delta v + bu - u^2 v. \end{aligned}$$

Dabei sind d_1, d_2, a und b positive Parameter. Typische Werte sind $d_1 = d_2 = 0{,}02$, $a = 1$, $b = 3$, siehe etwa HAIRER/WANNER [78].

[2] Für ihre Arbeiten auf diesem Gebiet erhielten HODGKIN und HUXLEY im Jahre 1963 den Nobelpreis für Physiologie und Medizin.

Die folgenden vier Modelle beschreiben die Wellendynamik in erregbaren Medien. Ein klassisches Beispiel für derartige Prozesse ist die Belousov-Zhabotinsky-Reaktion, eine katalytische Reaktion von Bromat-Ionen und Malonsäure zu Brommalonsäure. Diese chemische Reaktion führt zur Herausbildung interessanter Muster, insbesondere zu rotierenden Spiralwellen. Die Gleichungen erregbarer Medien stellen ein wichtiges Anwendungsgebiet der in dieser Arbeit untersuchten Verfahren dar. Sie werden daher in Kapitel 10 gesondert behandelt.

Die Gleichungen von Krinsky, Pertsov und Reshetilov

Dieses Reaktions-Diffusions-System mit stückweise linearer Reaktionsfunktion wurde 1972 von Krinsky, Pertsov und Reshetilov [98] zur Beschreibung erregbarer Medien angegeben:

$$\begin{aligned} u_t &= d_1\Delta u + f(u,v), \\ v_t &= d_2\Delta v + g(u,v), \\ f(u,v) &= \begin{cases} -k_1 u - v, & u < \sigma, \\ k_f(u-a) - v, & \sigma < u < 1-\sigma, \\ k_2(1-u) - v, & 1-\sigma < u, \end{cases} \\ g(u,v) &= \begin{cases} k_g u - v, & k_g u \geq v, \\ k_\varepsilon(k_g u - v), & k_g u < v. \end{cases} \end{aligned}$$

Erregbare Medien, die mit diesem System modelliert werden, haben eine Reihe interessanter Eigenschaften, siehe etwa Davydov, Zykov und Michailov [49]. Typische Parameterwerte sind: $d_1 = 1$, $d_2 \in [0,1]$, $k_f = 1{,}7$, $k_g = 2$, $k_\varepsilon = 6$, $a = 0{,}1$, $\sigma = 0{,}01$, $\varepsilon \in [0{,}1, 0{,}5]$. Die Parameter k_1 und k_2 werden stets so gesetzt, daß f stetig ist.

Das dreikomponentige Oregonator-Modell (Field-Noyes-Gleichungen)

Field und Noyes [65] (1974) beschrieben die Belousov-Zhabotinsky-Reaktion durch drei Prozesse, die durch das folgenden System ausgedrückt werden:

$$\begin{aligned} u_t &= d_1\Delta u + \frac{1}{\varepsilon_1}((1-u)u + (q-u)w), \\ v_t &= d_2\Delta v + u - v, \\ w_t &= d_3\Delta w + \frac{1}{\varepsilon_2}(fv - (q+u)w). \end{aligned} \tag{2.5}$$

Die Parameter $\varepsilon_1 = 0{,}01$, $\varepsilon_1 = 10^{-4}$, $q = 2 \cdot 10^{-4}$, $f = 3$ ergeben sich aus typischen Stoffkonzentrationen bei der Belousov-Zhabotinsky-Reaktion sowie aus den von Tyson [162] bestimmten Reaktionsraten.

Das zweikomponentige Oregonator-Modell (Tyson-Fife-Keener-Gleichungen)

Eine Vereinfachung des dreikomponentigen Modells durch $w_t = 0$, $\Delta w = 0$ führt auf das zweikomponentige System

$$\begin{aligned} u_t &= d_1 \Delta u + \frac{1}{\varepsilon_1}\left((1-u)u + fv\frac{q-u}{q+u}\right), \\ v_t &= d_2 \Delta v + u - v. \end{aligned} \tag{2.6}$$

Dieses Modell wurde ohne den Diffusionsteil von Tyson und Fife [161, 163] (1979/80) aufgestellt und später von Keener und Tyson [93] (1986) um den Diffusionsanteil erweitert. Als ein Beispiel für mögliche Parameter geben wir die von Pardhanani und Carey [129] zu numerischen Simulationen verwendeten Werte $d_1 = 1, d_2 = 0{,}6, \varepsilon = 0{,}01, q = 0{,}002, f = 3$ an.

Die Gleichungen von Barkley, Kness und Tuckerman

Barkley, Kness und Tuckerman [18] entwickelten 1990 das folgende Modell zur Beschreibung von Erregungswellen:

$$\begin{aligned} u_t &= \Delta u + \frac{1}{\varepsilon}\left(u(1-u)\left(u - \frac{v+b}{a}\right)\right), \\ v_t &= \Delta v + u - v. \end{aligned}$$

In einer numerischen Simulation in der gleichen Arbeit wurden die Parameter $a = 0{,}3$, $b = 0{,}01$ und $\varepsilon = 2{,}5 \cdot 10^{-3}$ verwendet.

2.3 Einige analytische Aussagen zur Lösbarkeit

Die analytische Theorie parabolischer Differentialgleichungen sichert Existenz und Eindeutigkeit von Lösungen, wenn das System gewisse Bedingungen erfüllt. Auch wenn diese Bedingungen für viele in der Praxis auftretende Probleme zu restriktiv sind, so können doch für einige Systeme relevante Aussagen gewonnen werden. Die folgende Auswahl analytischer Aussagen zu Reaktions-Diffusions-Systemen wurde der Darstellung von Smoller [150, Kap. 14] entnommen. Zunächst wird für räumlich eindimensionale Probleme eine Integraldarstellung der Lösung präsentiert; es folgen Aussagen zu lokaler und globaler Existenz und Eindeutigkeit der Lösung.

Im folgenden betrachten wir, in Übereinstimmung mit Smoller [150], lediglich *räumlich eindimensionale* Reaktions-Diffusions-Systeme. Wir benötigen die folgende Definition von Funktionenräumen:

Definition 2.3. Es sei $BC(\mathbb{R})$ der mit einer geeigneten Norm $\|\cdot\|_{BC}$ versehene Banachraum der beschränkten und gleichmäßig stetigen Funktionen auf $\mathbb{R}$ mit Werten in $\mathbb{R}^m$. Wir definieren ferner mit $C([t_0, t_e], BC(\mathbb{R}))$ den Raum stetiger Funktionen $\mathbf{u} : [t_0, t_e] \to BC(\mathbb{R})$, der mit der Norm

$$\|\mathbf{u}\| = \sup_{t \in [t_0, t_e]} \|\mathbf{u}(t)\|_{BC}$$

versehen wird. Dieser Raum ist ebenfalls ein Banachraum. □

Wir betrachten das folgende Problem: Gesucht sind Lösungen $\mathbf{u} \in C([t_0, t_e], BC(\mathbb{R}))$ des Systems

$$\begin{aligned} \mathbf{u}_t &= \mathbf{D}\mathbf{u}_{xx} + \mathbf{f}(\mathbf{u}), \qquad x \in \mathbb{R},\ t \in [t_0, t_e], \\ \mathbf{u}(x,t) &= \mathbf{u}_0(x) \end{aligned} \tag{2.7}$$

mit der Diffusions-Matrix $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_m),\ d_i \geq 0,\ i = 1, \ldots, m$ und einer hinreichend glatten Funktion $\mathbf{f}$.

Der folgende Satz gibt eine Integraldarstellung der Lösung von (2.7) an.

Satz 2.4. *Gegeben sei das System (2.7) mit $\mathbf{u}_0 \in BC(\mathbb{R})$. Eine Funktion $\mathbf{u} \in C([t_0, t_e], BC(\mathbb{R}))$ ist genau dann Lösung von (2.7), wenn*

$$\mathbf{u}(x,t) = \int_{\mathbb{R}} \mathbf{G}(x-y,t)\ \mathbf{u}_0(y)\ dy + \int_{t_0}^{t_e} \int_{\mathbb{R}} \mathbf{G}(x-y, t-s)\ \mathbf{f}(\mathbf{u}(y,s))\ dy\ ds$$

gilt, wobei $\mathbf{G}(x,t) = diag(g_1(x,t), \ldots, g_m(x,t))$ mit

$$g_i(x,t) = \frac{1}{\sqrt{4\pi d_i t}} e^{-\frac{x^2}{4 d_i t}}, \qquad i = 1, \ldots, m$$

ist. Die Funktion $\mathbf{G}(x,t)$ wird als Fundamentallösung des Systems (2.7) bezeichnet.

Beweis. Siehe Smoller [150, 14.A.]. □

Mit Hilfe dieser Darstellung läßt sich die **lokale Existenz und Eindeutigkeit** einer Lösung von (2.7) zeigen.

Satz 2.5. *Gegeben sei das System (2.7) mit $\mathbf{u}_0 \in BC(\mathbb{R})$. Es gelte $\mathbf{f}(\mathbf{0}) = \mathbf{0}$. Dann gibt es ein $t_1 \in]t_0, t_e]$, so daß (2.7) eine eindeutig bestimmte Lösung in $C([t_0, t_1], BC(\mathbb{R}))$ besitzt. t_1 hängt dabei nur von $\mathbf{f}$ und $\|\mathbf{u}_0\|_\infty$ ab.*

Beweis. Siehe Smoller [150, Theorem 14.2.]. □

Falls $\mathbf{f}$ linear ist, d.h. falls $\mathbf{f}(\mathbf{u}) = \mathbf{F}\mathbf{u}$ mit einer Matrix $\mathbf{F}$ gilt, so existiert die Lösung sogar global:

Satz 2.6. *Gegeben sei das System*

$$\begin{aligned} \mathbf{u}_t &= \mathbf{D}\mathbf{u}_{xx} + \mathbf{F}\mathbf{u}, \qquad x \in \mathbb{R},\ t \in [t_0, t_e], \\ \mathbf{u}(x,t) &= \mathbf{u}_0(x) \end{aligned} \tag{2.8}$$

mit $\mathbf{u}_0 \in BC(\mathbb{R})$. Dann existiert eine eindeutig bestimmte Lösung von (2.8) in $C([t_0, \infty[, BC(\mathbb{R}))$.

Beweis. Siehe Smoller [150, Bemerkung S. 198]. □

Wenn die Lösung von (2.7) in einem Intervall $[t_0, t_e]$ a-priori beschränkt ist, so existiert sie dort sogar global:

Satz 2.7. *Gegeben sei das System (2.7) mit* $\mathbf{u}_0 \in BC(\mathbb{R})$*. Für eine beliebige Lösung* $\mathbf{u}$ *dieses Systems gelte*

$$\max_{x\in\mathbb{R}} |\mathbf{u}(x,t)| < K < \infty, \qquad K > 0 \tag{2.9}$$

für alle $t \in [t_0, t_e]$*, in denen* $\mathbf{u}$ *definiert ist. Die Konstante* K *hängt nicht von* t *ab. Dann existiert die Lösung* $\mathbf{u}$ *für alle* $t \in [t_0, t_e]$ *und ist eindeutig bestimmt.*

Beweis. Siehe Smoller [150, Theorem 14.4.]. □

Die Zahl K in (2.9) wird als **a-priori-Schranke** der Lösung bezeichnet.

2.4 Invariante Bereiche

Um die globale Existenz einer Lösung mit Hilfe von Satz 2.7 zeigen zu können, benötigt man eine a-priori-Schranke für die Lösung. Eine derartige Beschränktheit ist insbesondere dann gegeben, wenn sich die Lösung $\mathbf{u}$ in einem invarianten Bereich der Differentialgleichung befindet. Die Darstellung der Theorie invarianter Bereiche in diesem Abschnitt orientiert sich erneut an Smoller [150, Abschnitt 14.B.], wo man noch eine Reihe weiterer Aussagen zu dieser Problematik findet. Wir betrachten in diesem Abschnitt das folgende Problem:

Problem 2.8. Es sei $\Omega \subset \mathbb{R}$ ein offenes Intervall und X ein Raum glatter Funktionen auf Ω mit Werten in $\mathbb{R}^m$. Gesucht sind Lösungen $\mathbf{u} \in C([t_0, t_e], X)$ des Systems

$$\begin{aligned} \mathbf{u}_t &= \mathbf{D}\mathbf{u}_{xx} + \mathbf{f}(\mathbf{u}), \qquad x \in \Omega,\ t \in [t_0, t_e], \\ \mathbf{u}(x,t) &= \mathbf{u}_0(x) \end{aligned} \tag{2.10}$$

mit der Diffusions-Matrix $\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_m),\ d_i \geq 0,\ i = 1, \ldots, m$ und einer hinreichend glatten Funktion $\mathbf{f}$. Die Lösung $\mathbf{u}$ erfülle Dirichlet- oder Neumann-Randbedingungen.

Für dieses Problem definieren wir den invarianten Bereich wie folgt:

Definition 2.9. Gegeben sei das Problem 2.8. Es sei $\Sigma \subset \mathbb{R}^n$ eine abgeschlossene Menge, für die aus $\mathbf{u}_0(x) \in \Sigma, \forall x \in \Omega$ stets $\mathbf{u}(x,t) \in \Sigma, \forall x \in \overline{\Omega}, \forall t \in [t_0, t_e]$ folgt. Dann ist Σ ein **invarianter Bereich** des Problems 2.8. □

Auch wenn wir hier nur räumlich eindimensionale Probleme betrachten, so läßt sich die Theorie der invarianten Bereiche auch auf räumlich mehrdimensionale Systeme übertragen, siehe dazu Chueh, Conley und Smoller [40].

Die folgende Bedingung wird von Smoller [150] als „Bedingung K“ bezeichnet:

Bedingung K. Falls $\mathbf{u} \in X$ ist, dann existiert eine kompakte Menge $K \subset \Omega$, so daß aus $x \notin K$ stets $\mathbf{u}(x) \in \operatorname{int}(\Sigma)$ folgt.

Bemerkung 2.10. Falls Ω ein beschränktes Intervall ist, so ist Bedingung K immer erfüllt. □

Der folgende Satz gibt Bedingungen an, unter denen ein verallgemeinertes Rechteck ein invarianter Bereich ist.

Satz 2.11. *Gegeben sei das Problem 2.8. Es gelte die Bedingung K. Sei*

$$\Sigma = \{\mathbf{u} = (u_1, \ldots, u_n) \in \mathbb{R}^n : a_i \leq u_i \leq b_i, i = 1, \ldots, n\}$$

ein verallgemeinertes Rechteck, gegeben durch die Zahlen a_i und b_i. Das Vektorfeld $\mathbf{f}(\mathbf{u})$ zeige für alle $\mathbf{u} \in \partial\Sigma$ in das Gebiet Σ hinein. (Der Fall, daß $\mathbf{f}(\mathbf{u})$ tangential zu $\partial\Sigma$ steht, ist dabei nicht zugelassen.) Dann ist $\mathbf{f}$ ein invarianter Bereich des Problems.

Beweis. Siehe Smoller [150, Corollary 14.8. (a)]. □

Aus Satz 2.7 folgt unmittelbar der folgende Satz.

Satz 2.12. *Gegeben sei das Problem 2.8 mit $X = BC(\mathbb{R})$. Falls das Problem einen beschränkten invarianten Bereich Σ besitzt und $\mathbf{u}_0(x) \in \Sigma\ \forall x \in \mathbb{R}$ ist, so existiert eine eindeutig bestimmte Lösung für $t \in [t_0, t_e]$.*

Beweis. Siehe Smoller [150, Corollary 14.9.].

In Smoller [150, S. 208ff.] werden für die in (2.3), (2.4) und (2.5) angegebenen Systeme beschränkte invariante Bereiche angegeben. Damit wird für entsprechende Anfangswerte $\mathbf{u}_0$ die globale Existenz eindeutig bestimmter Lösungen dieser Probleme gezeigt. Wir wollen im folgenden einen invarianten Bereich für das zweikomponentige Oregonator-Modell angeben, das der in Kapitel 10 näher beschriebenen Modellierung der Belousov-Zhabotinsky-Reaktion dient.

Das Modell ist durch die bereits in (2.6) angegebenen Gleichungen

$$\begin{aligned} u_t &= d_1 \Delta u + \frac{1}{\varepsilon}\left((1-u)u + fv\frac{q-u}{q+u}\right), \\ v_t &= d_2 \Delta v + u - v \end{aligned}$$

für $x \in \mathbb{R}$ und $t > 0$ gegeben. Die Anfangsbedingungen seien mit $u(x,0) = u_0(x)$ und $v(x,0) = v_0(x)$ bezeichnet. Eine möglich Wahl für die enthaltenen Parameter ist $d_1 = 1, d_2 = 0{,}6, \varepsilon = 0{,}01, q = 0{,}002, f = 3$, siehe etwa Pardhanani und Carey [129]. Wir setzen

$$F(u,v) = \frac{1}{\varepsilon}\left((1-u)u + fv\frac{q-u}{q+u}\right) \qquad \text{und} \qquad G(u,v) = u - v.$$

Abbildung 2.1 zeigt die Kurven $F(u,v) = 0$ und $G(u,v) = 0$ und das Rechteck

$$\Sigma = \{(u,v) : q \leq u \leq 1{,}1,\ 0 \leq v \leq 1{,}2\}.$$

Wie aus der Abbildung hervorgeht, zeigt das Vektorfeld $\mathbf{f} = (F, G)$ auf dem Rand von Σ stets in das Gebiet Σ hinein. Nach Satz 2.11 ist demnach Σ ein invarianter Bereich der Differentialgleichung. Falls die Anfangswerte $\mathbf{u}_0(x) = (u_0(x), v_0(x))$ für alle $x \in \mathbb{R}$ in Σ liegen, so existiert nach Satz 2.12 global für $t \geq 0$ eine eindeutig bestimmte Lösung.

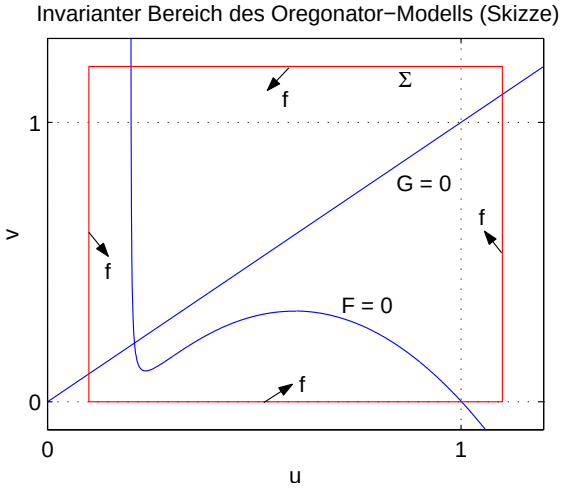


Abbildung 2.1: Invarianter Bereich des zweikomponentigen Oregonator-Modells (Skizze nicht maßstabsgerecht)

2.5 Schwache Formulierung skalarer semilinearer Reaktions-Diffusions-Gleichungen

Um räumlich mehrdimensionale Reaktions-Diffusions-Gleichungen numerisch zu lösen, wollen wir die in Abschnitt 3.1 vorgestellte Methode der finiten Elemente verwenden. Diese Methode basiert auf einer Umformulierung der Differentialgleichung in eine Integralgleichung, die als schwache Formulierung des Problems bezeichnet wird. Die Lösungen der schwachen Formulierung sind Elemente gewisser Funktionenräume, die wir zunächst in Abschnitt 2.5.1 definieren werden. Anschließend stellen wir in Abschnitt 2.5.2 für skalare semilineare Reaktions-Diffusions-Gleichungen die schwache Formulierung auf.

2.5.1 Funktionenräume

Wir stellen in diesem Abschnitt nur die Definitionen und einige wenige Eigenschaften von Lebesgue- und Sobolev-Räumen dar. Eine umfassendere Darstellung der Theorie dieser Räume findet sich beispielsweise bei Adams [2]. Zunächst führen wir unter den meßbaren Funktionen, die auf einer offenen Menge $\Omega \subset \mathbb{R}^n$ definiert sind, die folgende Äquivalenzrelation ein: Zwei derartige Funktionen u und v nennen wir äquivalent, wenn sie in fast allen Punkten $\mathbf{x} \in \Omega$ übereinstimmen[3]. Wir schreiben dafür $u, v \in [u] = [v]$, wobei $[u] = [v]$ die entsprechende Äquivalenzklasse meßbarer Funktionen bezeichnet.

Definition 2.13. Es sei $\Omega \subset \mathbb{R}^n$ eine offene Menge. Die linearen Räume

$$L^p(\Omega) = \left\{ [v] : v : \Omega \to \mathbb{R},\ v \text{ ist meßbar}, \int\limits_\Omega |v(\mathbf{x})|^p \, d\mathbf{x} < \infty \right\}, \qquad 1 \le p < \infty$$

[3] d.h. wenn $\int\limits_\Omega |u - v| \, d\mathbf{x} = 0$ ist

und

$$L^\infty(\Omega) = \{[v] : v : \Omega \to \mathbb{R},\ v \text{ ist meßbar, } \operatorname{ess\,sup}_{\mathbf{x}\in\Omega}|v(\mathbf{x})| < \infty\},$$

versehen mit den durch

$$\|v\|_{L^p(\Omega)} := \left(\int_\Omega |v|^p d\mathbf{x}\right)^{1/p}, \qquad 1 \le p < \infty$$

und

$$\|v\|_{L^\infty(\Omega)} := \operatorname{ess\,sup}_{\mathbf{x}\in\Omega}|v(\mathbf{x})|$$

definierten Normen werden als **Lebesgue-Räume** bezeichnet. □

Auch wenn die Lebesgue-Räume nach dieser Definition Räume von Äquivalenzklassen sind, so schreibt man für $[v] \in L^p(\Omega)$ oft vereinfachend $v \in L^p(\Omega)$. Wir werden diese Schreibweise im folgenden auch verwenden. Faßt man die Lebesgue-Räume in dieser Art als Funktionenräume auf, so ist dabei stets zu beachten, daß zwei äquivalente Funktionen in $L^p(\Omega)$ miteinander identifiziert werden.

Satz 2.14. *Sei $\Omega \in \mathbb{R}^n$ ein Gebiet. Die Räume $L^p(\Omega)$ sind für $1 \le p \le \infty$ Banachräume. Der Raum $L^2(\Omega)$ ist ein Hilbertraum mit dem Skalarprodukt*

$$\langle u, v\rangle_{L^2(\Omega)} := \int_\Omega u(\mathbf{x})v(\mathbf{x})\ d\mathbf{x},$$

für $u, v \in L^2(\Omega)$.

Beweis. Der Beweis findet sich in vielen Lehrbüchern der Funktionalanalysis, beispielsweise bei Adams [2, Theorem 2.10]. □

Für die nun folgende Definition der Sobolev-Räume werden nicht nur Bedingungen an eine Funktion v, sondern auch an deren distributionelle Ableitungen gestellt[4]. Höhere Ableitungen können in der Multiindex-Schreibweise formal dargestellt werden.

Definition 2.15. Es sei $\alpha = (\alpha_1, \ldots, \alpha_n)$ ein Multiindex mit Komponenten $\alpha_i \in \mathbb{N}_0,\ i = 1, \ldots, n$ sowie $v : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ eine Funktion, die hinreichend oft distributionell differenzierbar ist. Dann führen wir die Bezeichnung

$$\partial^\alpha v := \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}} v$$

ein. Auf der rechten Seite stehen dabei distributionelle Ableitungen nach den einzelnen Komponenten von $\mathbf{x}$. Ferner definieren wir $|\alpha| = \sum_{i=1}^n \alpha_i$. □

Beispiel 2.16. Für $\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$ und eine glatte Funktion $v : \mathbb{R}^3 \to \mathbb{R}$ gilt beispielsweise

$$\Delta v = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2} = \partial^{(2,0,0)} v + \partial^{(0,2,0)} v + \partial^{(0,0,2)} v.$$

□

[4] Eine Darstellung der Distributionentheorie wird hier nicht gegeben, siehe dafür etwa Triebel [158].

Ausgerüstet mit diesem Formalismus, können wir nun die folgenden Räume definieren.

Definition 2.17. Es sei $\Omega \subset \mathbb{R}^n$ ein Gebiet. Die linearen Räume

$$H^k(\Omega) = \{v \in L^2(\Omega) : \partial^\alpha v \in L^2(\Omega), \text{ falls } |\alpha| \leq k\}, \qquad k \in \mathbb{N}_0$$

werden mit den Normen

$$\|v\|_{H^k(\Omega)} = \left(\int_\Omega \sum_{|\alpha| \leq k} |\partial^\alpha v|^2 d\mathbf{x} \right)^{1/2}$$

versehen und als **Sobolev-Räume** bezeichnet. □

Satz 2.18. *Es sei Ω ein Gebiet in $\mathbb{R}^n$. Die Sobolevräume $H^k(\Omega)$, $k \in \mathbb{N}_0$ sind Hilberträume mit dem Skalarprodukt*

$$\langle u, v \rangle_{H^k(\Omega)} := \sum_{|\alpha| \leq k} \langle \partial^\alpha u, \partial^\alpha v \rangle_{L^2(\Omega)}$$

für $u, v \in H^k(\Omega)$.

Beweis. Siehe etwa Werner [170, Satz V.1.12]. □

Zusätzlich definiert man für die Räume $H^k(\Omega)$ die folgenden sogenannten **Halbnormen**:

$$|v|_{H^k(\Omega)} := \left(\int_\Omega \sum_{|\alpha| = k} |\partial^\alpha v|^2 d\mathbf{x} \right)^{1/2}.$$

Für die Lebesgue- und Sobolevräume gelten die Inklusionen $L^q(\Omega) \subset L^p(\Omega)$, falls $p \leq q$ und $|\Omega| := \int_\Omega d\mathbf{x} < \infty$ ist, sowie $H^l(\Omega) \subset H^k(\Omega)$, falls $k \leq l$ ist.

Definition 2.19. Für zeitabhängige Funktionen $v : \Omega \times (t_0, t_e) \to \mathbb{R}$ und einen normierten Funktionenraum V definieren wir ferner den Raum

$$L^p(]t_0, t_e[, V) = \{v \ : \ v(\cdot, t) \in V, \ \forall t \in]t_0, t_e[\text{ und } F \in L^p(]t_0, t_e[) \text{ mit } F(t) = \|v(\cdot, t)\|_V\}$$

und versehen ihn mit der Norm $\|v\|_{L^p(]t_0,t_e[,V)} := \|F\|_{L^p(]t_0,t_e[)}$. □

Schließlich führen wir einen weiteren Raum ein:

Definition 2.20. Es sei Ω ein beschränktes Gebiet in $\mathbb{R}^n$ mit Lipschitz-stetigem Rand und $\mathrm{tr}_{\partial\Omega} : H^1(\Omega) \to L^2(\partial\Omega)$ die **Spurabbildung**[5] bezüglich Ω. Wir definieren den Raum

$$H_0^1(\Omega) := \{v \in H^1(\Omega) : \mathrm{tr}_{\partial\Omega} v = 0 \text{ fast überall auf } \partial\Omega\}.$$

[5] Die Spurabbildung ist eine Verallgemeinerung der Einschränkung $v|_{\partial\Omega}$ auf Funktionen v, die bei $\partial\Omega$ nicht glatt sind. Siehe dazu etwa Knabner/Angermann [94, Satz 3.5] oder Alt [6, S. 249 ff.].

2.5.2 Die schwache Formulierung

Es sei Ω ein beschränktes Gebiet in $\mathbb{R}^n$ mit Lipschitz-stetigem Rand. Wir betrachten in diesem Abschnitt die *skalare* semilineare Reaktions-Diffusions-Gleichung von der Form

$$\begin{aligned} \frac{\partial u}{\partial t}(\mathbf{x},t) &= \nabla \cdot (d(\mathbf{x})\nabla u(\mathbf{x},t)) + r(\mathbf{x})u(\mathbf{x},t) + p(u(\mathbf{x},t)) + q(\mathbf{x},t), \\ u(\mathbf{x},t_0) &= u_0(\mathbf{x}) \end{aligned} \tag{2.11}$$

entweder mit Dirichlet-Randbedingung

$$u(\mathbf{x},t) = g_{\mathrm{Dir}}(\mathbf{x},t), \qquad \mathbf{x} \in \partial\Omega \tag{2.12}$$

oder mit Neumann-Randbedingung

$$d(\mathbf{x})\nabla u \cdot \mathbf{n}_{\partial\Omega} = g_{\mathrm{Neu}}(\mathbf{x},t), \qquad \mathbf{x} \in \partial\Omega. \tag{2.13}$$

Die **schwache Formulierung** entsteht aus der Differentialgleichung (2.11) durch Multiplikation mit einer geeigneten Testfunktion v und Integration über Ω. Die Wahl der Testfunktion und der Funktionenraum, in dem sich die Lösung befindet, hängen dabei vom Typ der Randbedingung ab.

Schwache Formulierung für Dirichlet-Randbedingung

Gegeben sei $u_0 \in L^2(\Omega)$. Finde eine Funktion $u \in L^2(]t_0,t_e[, H^1(\Omega))$ mit $u_t \in L^2(]t_0,t_e[, L^2(\Omega))$, so daß die folgenden drei Aussagen gelten:

1. Für alle $v \in H_0^1(\Omega)$ und alle $t \in]t_0,t_e[$ gilt

$$\begin{aligned} \frac{\partial}{\partial t}\int_\Omega u\ v\ d\mathbf{x} &= -\int_\Omega d(\mathbf{x})\ \nabla u \cdot \nabla v\ d\mathbf{x} + \int_\Omega r(\mathbf{x})\ u\ v\ d\mathbf{x} \\ &\quad + \int_\Omega p(u(\mathbf{x},t))\ v\ d\mathbf{x} + \int_\Omega q(\mathbf{x},t)\ v\ d\mathbf{x}. \end{aligned} \tag{2.14}$$

2. Für fast alle $\mathbf{x} \in \Omega$ gilt $u(\mathbf{x},t_0) = u_0(\mathbf{x})$.

3. Für alle $t \in [t_0,t_e]$ und fast alle $\mathbf{x} \in \partial\Omega$ gilt $\mathrm{tr}_{\partial\Omega} u(\mathbf{x},t) = g_{\mathrm{Dir}}(\mathbf{x},t)$.

Die Zeitableitung von u wird hierbei im distributionellen Sinne verstanden.

Schwache Formulierung für Neumann-Randbedingung

Gegeben sei $u_0 \in L^2(\Omega)$. Finde eine Funktion $u \in L^2(]t_0, t_e[, H^1(\Omega))$ mit $u_t \in L^2(]t_0, t_e[, L^2(\Omega))$, so daß die folgenden beiden Aussagen gelten:

1. Für alle $v \in H^1(\Omega)$ und alle $t \in]t_0, t_e[$ gilt

$$\begin{aligned} \frac{\partial}{\partial t} \int_\Omega u\ v\ d\mathbf{x} &= -\int_\Omega d(\mathbf{x})\ \nabla u \cdot \nabla v\ d\mathbf{x} + \int_\Omega r(\mathbf{x})\ u\ v\ d\mathbf{x} \\ &\quad + \int_\Omega p(u(\mathbf{x},t))\ v\ d\mathbf{x} + \int_\Omega q(\mathbf{x},t)\ v\ d\mathbf{x} \\ &\quad + \int_{\partial\Omega} g_{\text{Neu}}\ v\ ds, \end{aligned} \tag{2.15}$$

   wobei ds das Bogenelement auf $\partial\Omega$ ist.

2. Für fast alle $\mathbf{x} \in \Omega$ gilt $u(\mathbf{x}, t_0) = u_0(\mathbf{x})$.

Für die *lineare* Reaktions-Diffusions-Gleichung, d.h. für $p \equiv 0$, existiert unter einer Reihe von Regularitätsbedingungen eine eindeutig bestimmte Lösung der schwachen Formulierung. Wir verweisen den interessierten Leser auf Knabner/Angermann [94, Abschnitt 6.1].

2.6 Numerische Lösung durch Diskretisierung

Die Lösung von Reaktions-Diffusions-Systemen, wie dem in (2.1) angegebenen, kann in den meisten Fällen nicht mehr auf analytischem Wege erfolgen. Nur für einige einfache Spezialfälle dieser Gleichungen kann eine exakte Lösung angegeben werden. Man ist daher in der Regel auf numerische Verfahren angewiesen, mit deren Hilfe eine Näherungslösung berechnet werden kann. Grundlage der numerischen Lösung des Problems ist zunächst eine Diskretisierung der Gleichungen. Die Differentialgleichung wird dabei durch ein endlichdimensionales Gleichungssystem approximiert, welches numerisch gelöst werden kann, und dessen Lösung eine Näherung der exakten Lösung der Differentialgleichung darstellt.

Die Diskretisierung der hier betrachteten Reaktions-Diffusions-Systeme erfolgt in zwei grundlegenden Schritten, der **Orts-** und der **Zeitdiskretisierung**. Dabei wird die Ortsvariable $\mathbf{x}$ und die Zeitvariable t durch eine endliche Anzahl diskreter Werte ersetzt. Das Zeitintervall $[t_0, t_e]$ wird durch eine Folge von Zeitpunkten $t_0 < t_1 < \cdots < t_e$ aufgeteilt. Die Länge des i-ten Zeitschritts bezeichnen wir mit $\tau_i := t_{i+1} - t_i$. Zu jedem Zeitpunkt t_i wird das Gebiet Ω mit einem Gitter G_i überzogen, dessen Gitterknoten die Punkte $\mathbf{x}_{ij} \in \Omega$ seien. Wird zur Ortsdiskretisierung die in Kapitel 3 beschriebene Methode linearer finiter Elemente benutzt, so werden in jedem Zeitschritt in den Gitterknoten Näherungswerte für die exakte Lösung $u(\mathbf{x}_{ij}, t_i)$ berechnet.

Die Reihenfolge von Orts- und Zeitdiskretisierung kann unterschiedlich gewählt werden. Das

läßt sich am bestem am Beispiel einer einfachen parabolischen Differentialgleichung, etwa der Wärmeleitungsgleichung

$$u_t(\mathbf{x}, t) = \Delta u(\mathbf{x}, t), \tag{2.16}$$

erläutern. Bei der sogenannten **Linienmethode** erfolgt zuerst die Orts- und dann die Zeitdiskretisierung. Im Ergebnis der Ortsdiskretisierung ersetzt man die Lösung $u(\mathbf{x}, t)$ durch einen Vektor $\mathbf{u}(t)$, dessen Komponenten Näherungswerte in den Gitterpunkten darstellen. Der Laplace-Operator wird durch die Steifigkeits-Matrix $\mathbf{S}$ approximiert. Die partielle Differentialgleichung (2.16) geht in das semidiskrete Problem

$$\mathbf{M}\mathbf{u}_t(t) = \mathbf{S}\mathbf{u}(t)$$

über. Reduziert man die Massenmatrix $\mathbf{M}$, wie in Abschnitt 3.1.5 beschrieben, auf eine Diagonalmatrix $\mathbf{L}$, so erhält man das System gewöhnlicher Differentialgleichungen

$$\mathbf{u}_t(t) = \mathbf{A}\mathbf{u}(t), \tag{2.17}$$

wobei $\mathbf{A} = \mathbf{L}^{-1}\mathbf{S}$ ist. Dieses System wird dann mit einem numerischen Zeitintegrationsverfahren gelöst. Verwendet man das in Abschnitt 5.3 beschriebene implizite Euler-Verfahren, so ergibt sich im i-ten Zeitschritt als diskretes Problem das lineare Gleichungssystem

$$(I - \tau_i \mathbf{A})\mathbf{u}_{i+1} = \mathbf{u}_i. \tag{2.18}$$

Dabei sind $\mathbf{u}_i$ und $\mathbf{u}_{i+1}$ die Näherungslösungen zu den Zeitpunkten t_i bzw. t_{i+1}.

Eine andere Reihenfolge der Diskretisierungen wird bei der **Rothe-Methode** gewählt. Hier erfolgt zuerst die Zeitdiskretisierung. Verwenden wir das in Abschnitt 5.3 beschriebene implizite Euler-Verfahren, so geht die parabolische Differentialgleichung (2.16) in das semidiskrete Problem

$$-\tau_i \Delta u_{i+1}(\mathbf{x}) + u_{i+1}(\mathbf{x}) = u_i(\mathbf{x}) \tag{2.19}$$

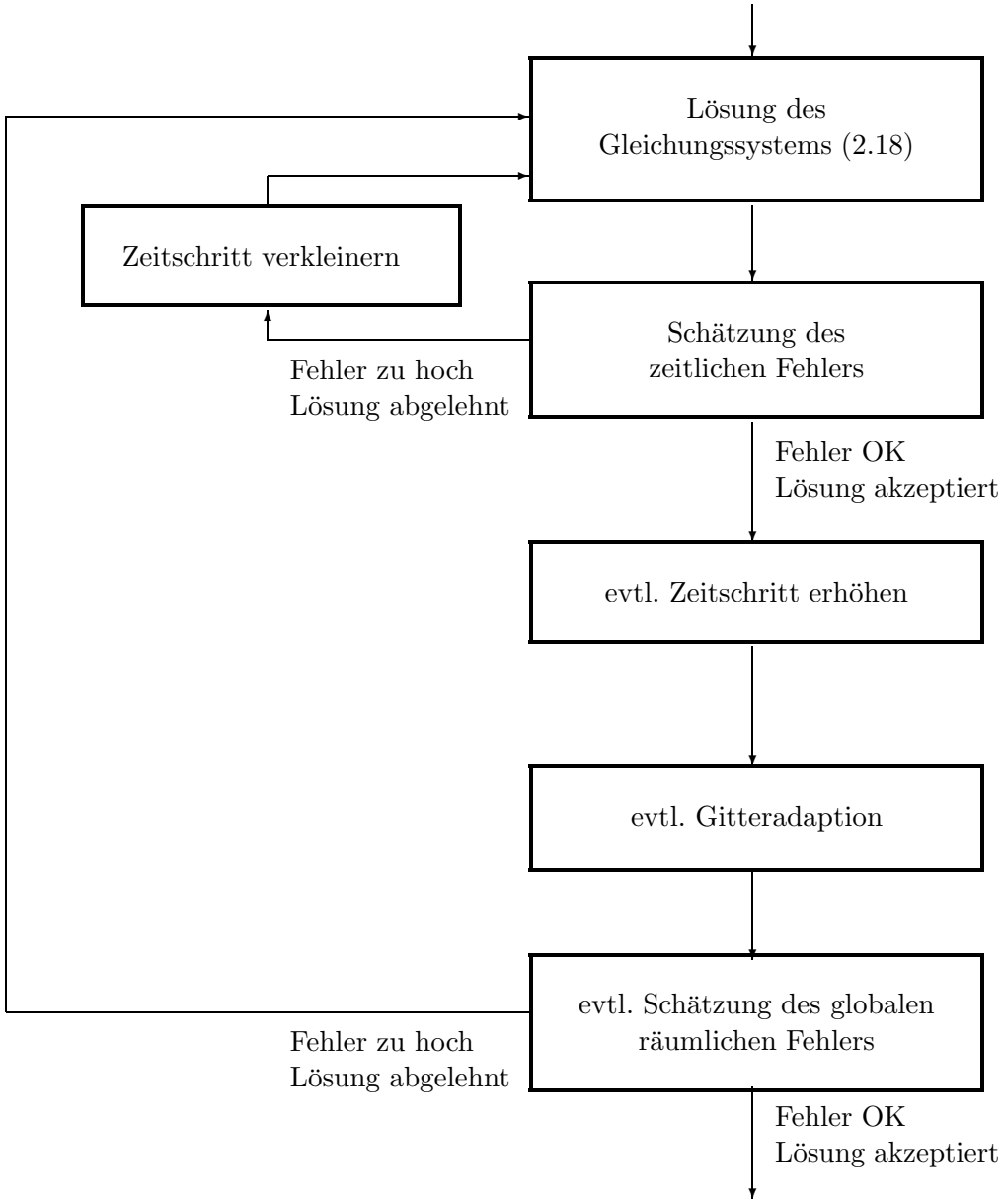
über. Dabei sind $u_i(\mathbf{x})$ und $u_{i+1}(\mathbf{x})$ Näherungslösungen zu den Zeitpunkten t_i bzw. t_{i+1}. Die Geichung (2.19) ist eine elliptische Differentialgleichung bezüglich der Unbekannten u_{i+1}, die beispielsweise mit linearen finiten Elementen und reduzierter Massenmatrix gelöst werden kann. Im Ergebnis erhält man als diskretes Problem wieder das Gleichungssystem (2.18).

Bei vielen Problemen ist es sinnvoll, sowohl die Länge der Zeitschritte τ_i als auch die lokale Feinheit der Gitter G_i geeignet anzupassen um in Raum und Zeit lokal auftretende Diskretisierungsfehler zu dämpfen. Dieser Vorgang wird als **Adaption** bezeichnet. Sowohl Linien- als auch Rothe-Methode führen im Endeffekt auf das gleiche lineare Gleichungssystem. Beide Methoden sind also identisch, wenn die numerische Berechnung *ohne Adaption* vorgenommen wird. Gewisse Unterschiede ergeben sich erst, wenn bezüglich Zeit und Raum adaptive Verfahren zum Einsatz kommen[6].

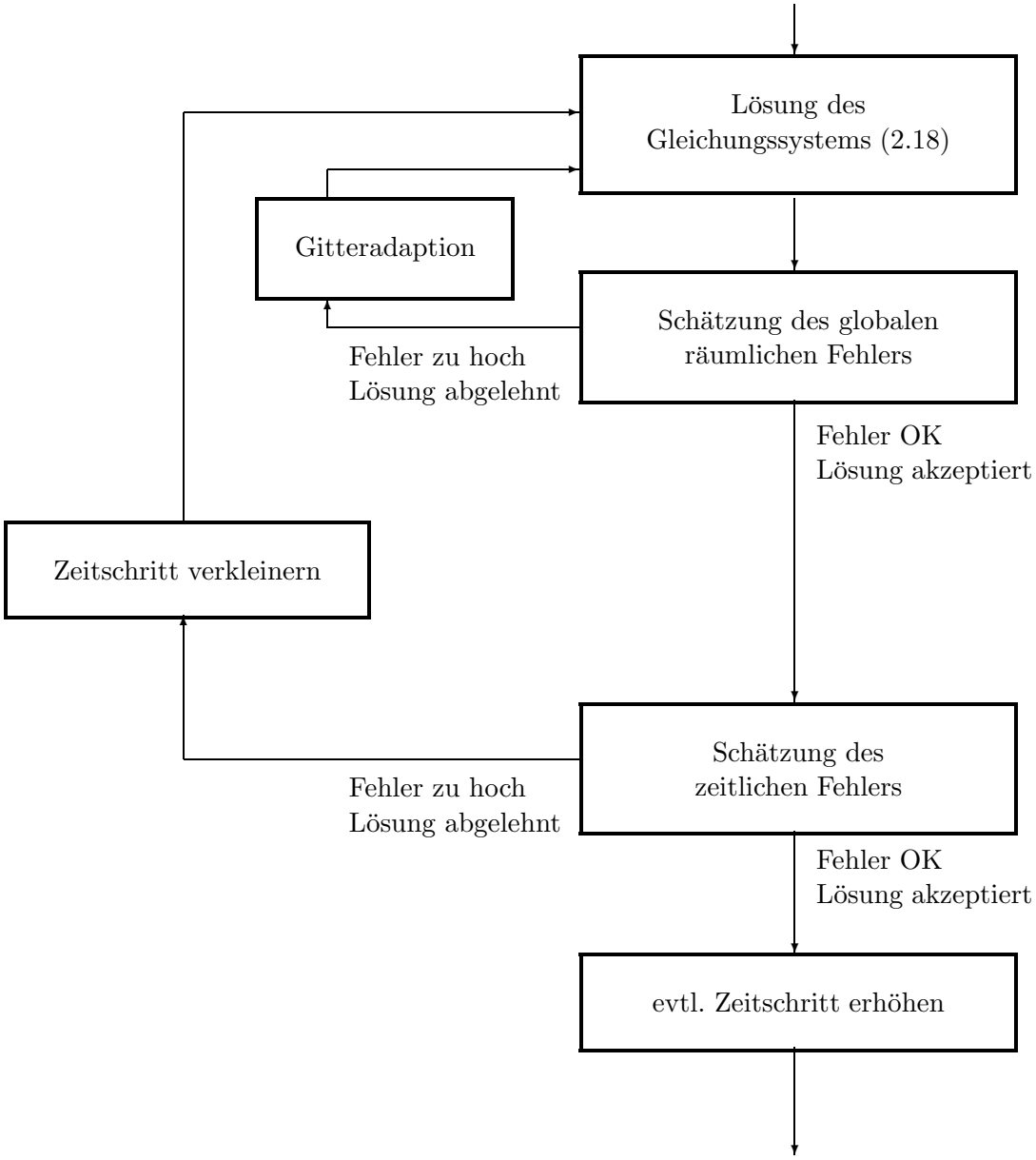
Bei der Linienmethode steht zunächst das semidiskrete Problem (2.17) im Vordergrund. Dieses zeitabhängige Problem wird durch ein Verfahren mit Zeitschrittsteuerung gelöst. Erst danach nimmt man die Anpassung des räumlichen Gitters vor. Bei der Rothe-Methode erfolgt dieser Prozeß in umgekehrter Reihenfolge. Zunächst wird das semidiskrete Problem (2.19) betrachtet und die Gitteradaption durchgeführt. Im Anschluß erfolgt die Zeitschrittsteuerung. Die folgenden beiden Ablaufpläne verdeutlichen den Unterschied.

[6] In Lang [102] werden die Unterschiede zwischen Linien- und Rothe-Methode bei adaptiven Verfahren kurz dargestellt.

Ein Zeitschritt der Linienmethode



Ein Zeitschritt der Rothe-Methode



Für die numerischen Berechnungen in dieser Arbeit gehen wir nach der Linienmethode vor, wobei auf die Schätzung des globalen räumlichen Fehlers verzichtet wird und auch die Gitteradaption nicht in jedem Zeitschritt erfolgt. In den folgenden Kapiteln werden wir Orts- und Zeitdiskretisierung näher erläutern. Ein Ablaufplan, der den in dieser Arbeit verwendeten Diskretisierungsverfahren für Reaktions-Diffusions-Systeme zugrunde liegt, ist im Anhang in Abschnitt C zu finden.

Kapitel 3

Ortsdiskretisierung semilinearer Reaktions-Diffusions-Gleichungen

Wie in Abschnitt 2.6 erläutert, erfolgt die Diskretisierung eines Reaktions-Diffusions-Systems nach der Linienmethode in zwei Schritten. In diesem Kapitel wollen wir uns mit dem ersten Schritt, also der Diskretisierung bezüglich des Ortes, befassen. Wir beschränken uns auf semilineare Reaktions-Diffusions-Systeme auf einem ein- oder zweidimensionalen Gebiet Ω. Zunächst wird nur die Diskretisierung skalarer Reaktions-Diffusions-Gleichungen betrachtet. Für $\Omega \subset \mathbb{R}^2$ beschreiben wir in Abschnitt 3.1 die Diskretisierung mit der Methode der finiten Elemente. Abschnitt 3.2 befaßt sich mit dem Fall, daß das Gebiet Ω Teil einer gekrümmten zweidimensionalen Fläche ist. Dieser Fall ist für eine in den Abschnitten 10.7 und 10.9 beschriebene Anwendung interessant, bei der die Differentialgleichungen erregbarer Medien auf gekrümmten Flächen berechnet werden. Schließlich werden wir in Abschnitt 3.4 die entsprechenden Diskretisierungen für Systeme von Reaktions-Diffusions-Gleichungen angeben.

3.1 Ortsdiskretisierung semilinearer Reaktions-Diffusions-Gleichungen in der Ebene

Wir betrachten in diesem Abschnitt eine skalare semilineare Reaktions-Diffusions-Gleichung

$$\begin{aligned} \frac{\partial u}{\partial t}(\mathbf{x},t) &= \nabla \cdot (d(\mathbf{x})\nabla u(\mathbf{x},t)) + r(\mathbf{x})u(\mathbf{x},t) \\ &\quad + p(u(\mathbf{x},t)) + q(\mathbf{x},t), && \mathbf{x} \in \Omega, \quad t \in [t_0, t_e], \\ u(\mathbf{x},t_0) &= u_0(\mathbf{x}), && \mathbf{x} \in \Omega \end{aligned} \tag{3.1}$$

versehen mit **Dirichlet-Randbedingungen**

$$u(\mathbf{x},t) = u_{\mathrm{Dir}}(\mathbf{x},t), \qquad \mathbf{x} \in \partial\Omega \tag{3.2}$$

oder **Neumann-Randbedingungen**

$$d(\mathbf{x},t)\nabla u \cdot \mathbf{n}_{\partial\Omega} = g_{\mathrm{Neu}}(\mathbf{x},t), \qquad \mathbf{x} \in \partial\Omega, \tag{3.3}$$

wobei Ω ein Gebiet in $\mathbb{R}^2$ ist. Die schwache Formulierung dieses Problems ist in (2.14), (2.15) angegeben. Die Ortsdiskretisierung derartiger Gleichungen kann mit Hilfe eines Differenzenverfahrens oder mit der **Methode der finiten Elemente** vorgenommen werden. Wir wollen uns hier auf das letztere Verfahren konzentrieren. Finite Elemente erzeugen eine Diskretisierung von Differentialausdrücken auf unstrukturierten Gittern. Damit können Gebiete beliebiger Geometrie approximiert werden. Außerdem kann eine adaptive Gitterverfeinerung in einfacher Weise gehandhabt werden. Wie in Abschnitt 3.2 vorgestellt wird, können finite Elemente auch auf gekrümmten Flächen eingesetzt werden. Ein Nachteil der Methode der finiten Elemente ist der – im Vergleich zum Differenzenverfahren – hohe Aufwand, den die in Abschnitt 3.1.6 beschriebene Berechnung der Matrixelemente erfordert.

Wir beschränken uns bei der Darstellung der Methode der finiten Elemente auf den einfachen Fall einer stückweise linearen stetigen Approximation auf einem Dreiecksgitter. Für die zahlreichen weiteren Varianten finiter Elemente verweisen wir auf Knabner/Angermann [94, Abschnitt 3.3] oder Grossmann/Roos [75, Kapitel 4].

3.1.1 Zur Geschichte der Methode der finiten Elemente

Es ist nicht einfach, den Ursprung der Methode der finiten Elemente eindeutig anzugeben. Die Approximation von Variationsproblemen auf Dreiecksgittern ist älter als die Diskretisierung partieller Differentialgleichungen. Bereits im Jahre 1851 gab Schellbach [142] zur Lösung eines Minimalflächenproblems ein Verfahren an, das eine spezielle Finite-Element-Technik darstellt. Im Jahre 1943 betrachtete Courant [42] die stückweise lineare Approximation einer elliptischen Differentialgleichung auf einem Dreiecksgitter. Von vielen Mathematikern wird in dieser Arbeit die Geburtsstunde der Methode der finiten Elemente gesehen.

Aus der Sicht des Anwenders ist jedoch nicht nur die stückweise polynomiale Approximation einer Differentialgleichung, sondern auch die Verwendung einer effizienten Assemblierungs-Strategie zum Aufbau der benötigten Matrizen ein essentieller Bestandteil der Methode der finiten Elemente. Unter diesem Gesichtspunkt trug die Arbeit von Turner, Clough, Martin und Topp [160] aus dem Jahre 1956 wesentlich zur Entwicklung der Methode der finiten Elemente bei. Im Jahre 1960 wurde von Clough [41] erstmals der Name „finite Elemente“ verwendet. In den darauffolgenden Jahren wurde eine umfangreiche mathematische Theorie zur Finite-Elemente-Methode entwickelt. Ein historischer Abriß, dem auch die hier zitierten Beispiele entnommen sind, findet sich bei Oden [127].

3.1.2 Triangulierung des Gebietes

Die Vernetzung des Gebietes Ω kann durch verschiedene Arten von Gittern vorgenommen werden. Häufig kommen Gitter aus Dreiecken oder Vierecken zum Einsatz. Wir wollen hier lediglich auf Dreiecksgitter eingehen. Liegt ein Gebiet $\Omega \subset \mathbb{R}^2$ mit krummlinigem Rand vor, so approximiert man Ω zunächst durch ein polygonal berandetes Gebiet Ω_h. Die Ecken des Randes von Ω_h liegen dabei auf dem Rand von Ω. Der Einfachheit halber betrachten wir in Abschnitt 3.1 jedoch stets ein polygonal berandetes Gebiet Ω, so daß Ω und Ω_h zusammenfallen. Wir können daher stets Ω statt Ω_h schreiben.

Es sei $\mathcal{T}_h$ eine Zerlegung des polygonal berandeten Gebietes Ω in abgeschlossene Dreiecke mit

den folgenden Eigenschaften.

(T1) Es gilt $\overline{\Omega} = \cup_{T \in \mathcal{T}_h} T$.

(T2) Für $T_1, T_2 \in \mathcal{T}_h$, $T_1 \neq T_2$ ist int $(T_1) \cap \text{int}(T_2) = \emptyset$.

(T3) Ist für $T_1, T_2 \in \mathcal{T}_h$, $T_1 \neq T_2$ der Durchschnitt $T_1 \cap T_2$ nicht leer, so ist $T_1 \cap T_2$ entweder ein Punkt oder eine gemeinsame Kante von T_1 und T_2.

Eine Zerlegung $\mathcal{T}_h$, die diese Bedingungen erfüllt, wird als **Triangulierung** bezeichnet.

Bemerkung 3.1. Bedingung (T3) schließt sogenannte hängende Knoten aus; das sind Punkte, die Eckpunkt eines Dreiecks sind, gleichzeitig aber im Inneren einer Dreiecksseite eines andern Dreiecks liegen. □

Die Feinheit einer Triangulierung $\mathcal{T}_h$ wird global durch die Länge der längsten auftretenden Dreiecksseite beschrieben. Wir bezeichnen diese Größe mit h. Die Gitterpunkte der Triangulierung werden numeriert und mit $\mathbf{x}_i$ bezeichnet. Ferner bezeichne

$$\begin{array}{ll} \mathcal{B} = \{i : \mathbf{x}_i \in \partial\Omega\} & \text{die Menge der Randindizes,} \\ \mathcal{I} = \{i : \mathbf{x}_i \in \text{int}(\Omega)\} & \text{die Menge der inneren Indizes und} \\ \mathcal{A} = \mathcal{I} \cup \mathcal{B} & \text{die Menge sämtlicher Indizes} \end{array} \tag{3.4}$$

der Knoten der Triangulierung. Eine Eckenindex-Funktion C wird wie folgt definiert: Hat ein Dreieck T die Eckpunkte $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k$ mit $i < j < k$, so sagen wir $C_1(T) = i,\ C_2(T) = j,\ C_3(T) = k$. Außerdem bezeichnen wir mit ω_i die Vereinigung aller Dreiecke, die $\mathbf{x}_i$ als Eckpunkt besitzen. Die Menge der Dreiecksseiten, deren beide Endpunkte auf $\partial\Omega$ liegen, wird mit $\partial\mathcal{T}_h$ bezeichnet. Analog zur Eckenindex-Funktion definieren wir die Randindex-Funktion c: Hat eine Randkante $E \in \partial\mathcal{T}_h$ die Endpunkte $\mathbf{x}_i$ und $\mathbf{x}_j$, $i < j$, so sei $c_1(E) = i,\ c_2(E) = j$.

Wir definieren die folgenden Funktionenräume:

Definition 3.2. Es sei $\mathcal{T}_h$ eine Triangulierung des Gebietes Ω. Dann sei

- $V_h \subset H^1(\Omega)$ der Raum der stetigen und stückweise – d.h. auf jedem Dreieck $T \in \mathcal{T}_h$ linearen Funktionen auf Ω sowie
- $V_{h,0} \subset H_0^1(\Omega)$ der Raum stückweise linearer stetiger Funktionen, die zusätzlich auf dem Rand von Ω den Wert 0 annehmen.

□

Der Raum V_h besitzt eine durch die Funktionen

$$\varphi_i \in V_h, \qquad \varphi_i(\mathbf{x}_j) = \delta_{ij}$$

gegebene Basis. Dabei ist δ_{ij} das Kronecker-Symbol.

Zusätzlich zu den oben angegebenen Bedingungen (T1), (T2) und (T3) fordert man in der Regel noch die sogenannte **Maximalwinkelbedingung**, die besagt, daß alle Innenwinkel der Dreiecke einer Triangulierung durch eine Konstante $\alpha < \pi$ nach oben beschränkt sind.

Definition 3.3. Es sei $v : \Omega \to \mathbb{R}$ eine beliebige stetige Funktion und $\mathcal{T}_h$ eine Triangulierung auf Ω mit den Knotenpunkten $\mathbf{x}_i$, $i \in \mathcal{A}$. Dann ist der lineare Interpolationsoperator I_h definiert durch

$$I_h(v) \in V_h, \qquad I_h(v)(\mathbf{x}_i) = v(\mathbf{x}_i) \quad \forall i \in \mathcal{A}.$$

□

Die Maximalwinkelbedingung garantiert, daß die Interpolierende $I_h(v)$ einer Funktion v für $h \to 0$ in der H_1-Norm gegen v konvergiert, eine Eigenschaft, die für die Konvergenz der mittels finiter Elemente gewonnenen Näherungslösung von Bedeutung ist.

Satz 3.4. *Es existiert eine Konstante $C > 0$, so daß für beliebige Triangulierungen $\mathcal{T}_h$, die die Maximalwinkelbedingung erfüllen, die Ungleichung*

$$\|v - I_h(v)\|_1 \leq Ch|v|_2 \quad \forall v \in H^2(\Omega_h)$$

gilt.

Beweis. Siehe Knabner/Angermann [94], Satz 3.35. □

3.1.3 Quadratur auf Dreiecksgittern

Bei der Ortsdiskretisierung semilinearer Reaktions-Diffusions-Gleichungen spielt die numerische Integration gewisser Terme eine wichtige Rolle, wie wir in den folgenden Abschnitten sehen werden. Vor diesem Hintergrund soll hier ein einfaches Quadraturverfahren, nämlich die Trapezregel, dargestellt sowie dessen Konvergenzordnung abgeschätzt werden. Die Trapezregel wird in der folgenden Weise definiert:

Definition 3.5. Es sei $\Omega \subset \mathbb{R}^2$ ein polygonal berandetes Gebiet und $f : \Omega \to \mathbb{R}$ eine stetige Funktion. Dann liefert die Trapezregel

$$Q_\Omega^{\mathcal{T}_h}(f) := \int_\Omega I_h(f(\mathbf{x}))\, d\mathbf{x}$$

eine Näherung für das Integral

$$\int_\Omega f(\mathbf{x})\, d\mathbf{x}.$$

Dabei ist I_h der in Definition 3.3 eingeführte lineare Interpolationsoperator. □

Wenn T ein Dreieck aus $\mathcal{T}_h$ mit den Eckpunkten $\mathbf{x}_i$, $i = 1, 2, 3$ ist, dann liefert die Trapezregel gerade

$$Q_T^{\mathcal{T}_h}(f) = \frac{f(\mathbf{x}_1) + f(\mathbf{x}_2) + f(\mathbf{x}_3)}{3}|T|.$$

Um die Konvergenzordnung der Trapezregel anzugeben, müssen wir zunächst ein Maß für die Uniformität von Triangulierungen einführen.

Definition 3.6. Eine Triangulierung $\mathcal{T}_h$ heißt (C_1, C_2)**-uniform**, wenn es zwei Konstanten C_1 und C_2 mit $0 < C_1 < C_2$ gibt, so daß jedes Element $T \in \mathcal{T}_h$ einen Kreis vom Radius $C_1 h$ enthält und in einem Kreis vom Radius $C_2 h$ enthalten ist. Die Größe h sei wie oben erwähnt das Maximum der Längen aller in $\mathcal{T}_h$ vorkommenden Dreiecksseiten. □

Auf einer Familie $(\mathcal{T}_h)_{h\in\mathbb{R}_+}$ von (C_1, C_2)-uniformen Triangulierungen ist die Trapezregel eine Quadraturformel erster Ordnung in h. Das geht aus dem folgenden Satz hervor.

Satz 3.7. *Sei $\Omega \subset \mathbb{R}^2$ ein polygonal berandetes Gebiet. Es seien C_1, C_2 und $h_{\max}$ gewisse Konstanten, so daß $0 < C_1 < C_2$, $h_{\max} > 0$ gilt und für beliebiges $h \in]0, h_{\max}]$ eine (C_1, C_2)-uniforme Triangulierung $\mathcal{T}_h$ existiert. Dann gilt für den Fehler der Quadratur*

$$\left| Q_\Omega^{\mathcal{T}_h}(f) - \int_\Omega f(\mathbf{x})\ d\mathbf{x} \right| \le Ch|f|_{H^2(\Omega)}$$

für eine beliebige Funktion $f \in H^2(\Omega)$ und beliebiges $h \in]0, h_{\max}]$. Die Konstante C hängt dabei nur von Ω, C_1 und C_2, nicht aber von f und h ab.

Beweis. Für ein beliebiges Dreieck $T \in \mathcal{T}_h$ und eine Funktion $f \in H^2(T)$ gilt die Abschätzung

$$\left| Q_T^{\mathcal{T}_h}(f) - \int_T f(\mathbf{x})\ d\mathbf{x} \right| \le C_3 h^2 |f|_{H^2(T)}.$$

Diese Ungleichung läst sich aus dem Bramble-Hilbert-Lemma herleiten; sie ist etwa in Grossmann/Roos [75, Seite 313] angegeben. Mit der Dreiecksungleichung folgt dann

$$\begin{aligned} \left| Q_\Omega^{\mathcal{T}_h}(f) - \int_\Omega f(\mathbf{x})\ d\mathbf{x} \right| &= \left| \sum_{T\in\mathcal{T}_h} Q_T^{\mathcal{T}_h}(f) - \sum_{T\in\mathcal{T}_h} \int_T f(\mathbf{x})\ d\mathbf{x} \right| \\ &\le \sum_{T\in\mathcal{T}_h} \left| Q_T^{\mathcal{T}_h}(f) - \int_T f(\mathbf{x})\ d\mathbf{x} \right| \le C_3 h^2 \sum_{T\in\mathcal{T}_h} |f|_{H^2(T)}. \end{aligned}$$

Es gilt nun die elementare Ungleichung $\left(\sum_{k=1}^n a_k\right)^2 \le n \sum_{k=1}^n a_k^2$, siehe etwa Heuser [83, Teil I, Kap. 12, Aufgabe 6]. Folglich ist

$$\sum_{T\in\mathcal{T}_h} |f|_{H^2(T)} \le \sqrt{|\mathcal{T}_h| \sum_{T\in\mathcal{T}_h} |f|^2_{H^2(T)}} = \sqrt{|\mathcal{T}_h|}|f|_{H^2(\Omega)}.$$

Da $\mathcal{T}_h$ eine (C_1, C_2)-uniforme Triangulierung ist, folgt $\sqrt{|\mathcal{T}_h|} \le C_4/h$. Insgesamt ergibt sich

$$\left| Q_\Omega^{\mathcal{T}_h}(f) - \int_\Omega f(\mathbf{x})\ d\mathbf{x} \right| \le C_3 C_4 h |f|_{H^2(\Omega)},$$

mit $C = C_3 C_4$ folgt die Behauptung. □

3.1.4 Ortsdiskretisierung der semilinearen Reaktions-Diffusions-Gleichung

Wir beschreiben in diesem Abschnitt die Ortsdiskretisierung der semilinearen Reaktions-Diffusions-Gleichung (3.1) mit Randbedingungen (3.2) oder (3.3), deren schwache Formulierung je nach Randbedungung in (2.14) bzw. (2.15) angegeben ist. Zur Diskretisierung verwenden wir lineare finite Elemente auf einer Triangulierung $\mathcal{T}_h$ des Gebietes Ω, die die Maximalwinkelbedingung erfüllt. Die diskrete Form geht aus der schwachen Formulierung hervor, wenn in der letzteren die Räume $H^1(\Omega)$ und $H_0^1(\Omega)$ durch die endlichdimensionalen Räume V_h und $V_{h,0}$ approximiert werden. Wir geben hier die diskreten Probleme für Dirichlet- und Neumann-Randbedingung an.

Ortsdiskretisierung für Dirichlet-Randbedingung

Gegeben sei eine Approximation $u_{0,h} \in V_h$ der Anfangsbedingung u_0. Finde eine Funktion $u_h \in L^2(]t_0, t_e[, V_{h,Dir})$ mit $\partial u_h/\partial t \in L^2(]t_0, t_e[, L^2(\Omega_h))$, die die folgenden drei Bedingungen erfüllt:

1. Für alle $v_h \in V_{h,0}$ und alle $t \in]t_0, t_e[$ gilt

$$\begin{aligned}\frac{\partial}{\partial t}\int\limits_\Omega u_h\ v_h\ d\mathbf{x} &= -\int\limits_\Omega d(\mathbf{x})\ \nabla u_h \cdot \nabla v_h\ d\mathbf{x} + \int\limits_\Omega r(\mathbf{x})\ u_h\ v_h\ d\mathbf{x} \\ &\quad + \int\limits_\Omega p(u_h)\ v_h\ d\mathbf{x} + \int\limits_\Omega q(\mathbf{x},t)\ v_h\ d\mathbf{x}.\end{aligned}$$

2. Für alle $\mathbf{x} \in \Omega$ gilt $u_h(\mathbf{x}, t_0) = u_{0,h}(\mathbf{x})$.

3. Für alle $t \in [t_0, t_e]$ und alle $\mathbf{x} \in \partial\Omega$ gilt $u_h(\mathbf{x}, t) = g_{\mathrm{Dir}}(\mathbf{x}, t)$.

Anstatt die Bedingung 1. für alle $v_h \in V_{h,0}$ zu fordern, reicht es aus, nur die Basiselemente φ_i des Raumes $V_{h,0}$ als Testfunktionen einzusetzen. Man setzt $u_i = u(\mathbf{x}_i)$ und $u_{0,i} = u_0(\mathbf{x}_i)$, so daß die Basisdarstellungen $u_h = \sum_{i\in\mathcal{A}} u_i\ \varphi_i$ und $u_{0,h} = \sum_{i\in\mathcal{A}} u_{0,i}\ \varphi_i$ erfüllt sind. Damit ist die Ortsdiskretisierung äquivalent zu einem System gewöhnlicher Differentialgleichungen der Form

$$\begin{aligned}\sum_{i\in\mathcal{A}} \frac{\partial u_i}{\partial t}\int\limits_\Omega \varphi_i\ \varphi_j\ d\mathbf{x} = &-\sum_{i\in\mathcal{A}} u_i \int\limits_\Omega d(\mathbf{x})\ \nabla\varphi_i \cdot \nabla\varphi_j\ d\mathbf{x} + \sum_{i\in\mathcal{A}} u_i \int\limits_\Omega r(\mathbf{x})\ \varphi_i\ \varphi_j\ d\mathbf{x} \qquad (3.5)\\ &+ \int\limits_\Omega p\left(\sum_{i\in\mathcal{A}} u_i\ \varphi_i\right)\varphi_j\ d\mathbf{x} + \int\limits_\Omega q(\mathbf{x},t)\ \varphi_j\ d\mathbf{x}, \qquad j \in \mathcal{I},\\ u_i(t) = g_{\mathrm{Dir}}(\mathbf{x}_i, t), &\qquad i \in \mathcal{B},\\ u_i(t_0) = u_{0,i}, &\qquad i \in \mathcal{A}.\end{aligned}$$

Der nichtlineare Term

$$\int_\Omega p\left(\sum_{i\in\mathcal{A}} u_i\ \varphi_i\right)\varphi_j\ d\mathbf{x}, \tag{3.6}$$

kann vereinfacht werden, indem das darin auftretende Integral mit einer Quadraturformel approximiert wird. Wir verwenden dafür die in Definition 3.5 angegebene Trapezregel $Q_\Omega^{\mathcal{T}_h}$. Eine derartige Approximation wird auch von Knabner und Angermann [94, Abschnitt 7.3] angegeben. Es ergibt sich

$$\begin{aligned}\int_\Omega p\left(\sum_{i\in\mathcal{A}} u_i\ \varphi_i\right)\varphi_j\ d\mathbf{x} \quad &\approx \quad Q_\Omega^{\mathcal{T}_h}\left(p\left(\sum_{i\in\mathcal{A}} u_i\ \varphi_i\right)\varphi_j\right)\\ &= \int_\Omega I_h\left(p\left(\sum_{i\in\mathcal{A}} u_i\ \varphi_i\right)\varphi_j\ d\mathbf{x}\right) = \frac{1}{3}|\omega_j|p(u_j).\end{aligned}$$

Das System (3.5) nimmt mit dieser Approximation die Form

$$\begin{aligned}\sum_{i\in\mathcal{A}}\frac{\partial u_i}{\partial t}\int_\Omega \varphi_i\ \varphi_j\ d\mathbf{x} = &-\sum_{i\in\mathcal{A}} u_i\int_\Omega d(\mathbf{x})\ \nabla\varphi_i\cdot\nabla\varphi_j\ d\mathbf{x} + \sum_{i\in\mathcal{A}} u_i\int_\Omega r(\mathbf{x})\ \varphi_i\ \varphi_j\ d\mathbf{x} \qquad (3.7)\\ &+\frac{1}{3}|\omega_j|p(u_j) + \int_\Omega q(\mathbf{x},t)\ \varphi_j\ d\mathbf{x}, && j\in\mathcal{I},\\ u_i(t) = g_{\mathrm{Dir}}(\mathbf{x}_i,t), && i\in\mathcal{B},\\ u_i(t_0) = u_{0,i}, && i\in\mathcal{A}.\end{aligned}$$

an. Um dieses System in Matrixschreibweise darzustellen, führen wir die folgenden Matrizen und Vektoren ein:

$$\begin{aligned}&\mathbf{M} = \left(\int_{\Omega_h}\varphi_i\ \varphi_j\ d\mathbf{x}\right)_{i\in\mathcal{A},j\in\mathcal{A}}, \qquad \mathbf{S} = \left(\int_{\Omega_h} d(\mathbf{x})\ \nabla\varphi_i\cdot\nabla\varphi_j\ d\mathbf{x}\right)_{i\in\mathcal{A},j\in\mathcal{A}}, \qquad (3.8)\\ &\widehat{\mathbf{R}} = \left(\int_{\Omega_h} r(\mathbf{x})\ \varphi_i\ \varphi_j\ d\mathbf{x}\right)_{i\in\mathcal{A},j\in\mathcal{A}}, \qquad \mathbf{R} = \mathrm{diag}\ (r(\mathbf{x}_i))_{i\in\mathcal{A}},\\ &\mathbf{L} = \frac{1}{3}\mathrm{diag}\ (|\omega_i|)_{i\in\mathcal{A}}, \qquad \widehat{\mathbf{q}} = \left(\int_{\Omega_h} q(\mathbf{x},t)\ \varphi_i\ d\mathbf{x}\right)_{i\in\mathcal{A}},\\ &\mathbf{q} = (q(\mathbf{x}_i,t))_{i\in\mathcal{A}}, \qquad \mathbf{u} = (u_i)_{i\in\mathcal{A}}.\end{aligned}$$

Ferner seien Untermatrizen in der folgenden Weise bezeichnet: $\mathbf{M}_{\mathcal{I},\mathcal{A}}$ ist die Untermatrix von $\mathbf{M}$, die aus den Zeilen $i\in\mathcal{I}$ und den Spalten $j\in\mathcal{A}$ gebildet wird, usw.

Ausgehend von gewissen Anwendungen aus der Mechanik wird $\mathbf{M}$ als **Massenmatrix**, $\mathbf{S}$ als **Steifigkeitsmatrix** und $\widehat{\mathbf{q}}$ als **Lastvektor** bezeichnet. $\widehat{\mathbf{q}}$ kann durch exakte Integration

gewonnen werden. Häufig approximiert man jedoch $q(\cdot,t)$ durch eine stückweise lineare Funktion $q_h(\cdot,t) \in V_h$. Dann ist $\widehat{\mathbf{q}}$ gerade durch die Beziehung $\widehat{\mathbf{q}} = \mathbf{M}\mathbf{q}$ gegeben, und das System (3.7) ist äquivalent zu

$$\mathbf{M}_{\mathcal{I},\mathcal{A}}\mathbf{u}_t = -\mathbf{S}_{\mathcal{I},\mathcal{A}}\mathbf{u} + \widehat{\mathbf{R}}_{\mathcal{I},\mathcal{A}}\mathbf{u} + \mathbf{L}_{\mathcal{I},\mathcal{I}}p(\mathbf{u}_\mathcal{I}) + \mathbf{M}_{\mathcal{I},\mathcal{A}}\mathbf{q}.$$

Durch Elimination der Randkomponenten erhält man schließlich das zu lösende System gewöhnlicher Differentialgleichungen

$$\begin{aligned}\mathbf{M}_{\mathcal{I},\mathcal{I}}\,\frac{\partial \mathbf{u}_\mathcal{I}}{\partial t} &= -\mathbf{S}_{\mathcal{I},\mathcal{I}}\;\mathbf{u}_\mathcal{I} + \widehat{\mathbf{R}}_{\mathcal{I},\mathcal{I}}\;\mathbf{u}_\mathcal{I} + \mathbf{L}_{\mathcal{I},\mathcal{I}}p(\mathbf{u}_\mathcal{I}) + \mathbf{M}_{\mathcal{I},\mathcal{A}}\;\mathbf{q} \\ &\quad -\mathbf{M}_{\mathcal{I},\mathcal{B}}\;\frac{\partial \mathbf{u}_\mathcal{B}}{\partial t} - \mathbf{S}_{\mathcal{I},\mathcal{B}}\;\mathbf{u}_\mathcal{B} + \widehat{\mathbf{R}}_{\mathcal{I},\mathcal{B}}\;\mathbf{u}_\mathcal{B}.\end{aligned} \tag{3.9}$$

in den Unbekannten $\mathbf{u}_\mathcal{I}$. Die Funktion $p(\mathbf{u}_\mathcal{I})$ wird komponentenweise verstanden. Die Vektoren $\mathbf{u}_\mathcal{B}$ und $\frac{\partial \mathbf{u}_\mathcal{B}}{\partial t}$ ergeben sich aus der Dirichlet-Randbedingung. Anfangsbedingungen werden entsprechend (3.5) gesetzt.

Ortsdiskretisierung für Neumann-Randbedingung

Das diskrete Problem für Neumann-Randbedingungen lautet:

Gegeben sei eine Approximation $u_{0,h} \in V_h$ der Anfangsbedingung. Finde eine Funktion $u_h \in L^2(]t_0,t_e[,V_h)$ mit $\partial u_h/\partial t \in L^2(]t_0,t_e[,L^2(\Omega))$, so daß die folgenden beiden Bedingungen erfüllt sind:

1. Für alle $v_h \in V_h$ und alle $t \in]t_0,t_e[$ gilt

$$\begin{aligned}\frac{\partial}{\partial t}\int_\Omega u_h\; v_h\; d\mathbf{x} &= -\int_\Omega d(\mathbf{x})\;\nabla u_h \cdot \nabla v_h\; d\mathbf{x} + \int_\Omega r(\mathbf{x})\; u_h\; v_h\; d\mathbf{x} \\ &\quad + \int_\Omega p(u_h)\; v_h\; d\mathbf{x} + \int_\Omega q(\mathbf{x},t)\; v_h\; d\mathbf{x} + \int_{\partial\Omega} g_{\text{Neu}}\; v_h\; ds,\end{aligned} \tag{3.10}$$

   wobei ds das Bogenelement auf $\partial\Omega$ ist.

2. Für alle $\mathbf{x} \in \Omega$ gilt $u_h(\mathbf{x},t_0) = u_{0,h}(\mathbf{x})$.

Man setzt

$$g_i = \begin{cases} g_{\text{Neu}}(\mathbf{x}_i), & i \in \mathcal{B}, \\ 0, & i \in \mathcal{I}, \end{cases} \qquad \mathbf{g}_{\text{Neu}} = (g_i)_{i\in\mathcal{A}} \tag{3.11}$$

und erzeugt damit die auf ganz Ω_h definierte Funktion $\widetilde{g}_{\text{Neu}} = \sum_{i\in\mathcal{A}} g_i\varphi_i$. Offenbar ist $\widetilde{g}_{\text{Neu}}|_{\partial\Omega_h}$ eine Approximation an g_{Neu}. Wie im Falle Dirichletscher Randbedingungen testet

man (3.10) nur mit den Basisfunktionen φ_j und approximiert $q(\cdot,t)$ durch $q_h(\cdot,t) \in V_h$ sowie $\int_\Omega p(u_h)\ v_h\ d\mathbf{x}$ mit der Trapezregel durch $Q_\Omega^{\mathcal{T}_h}(p(u_h)\ v_h)$. Es ergibt sich das System

$$\begin{aligned}\sum_{i\in\mathcal{A}} \frac{\partial u_i}{\partial t} \int_\Omega \varphi_i\ \varphi_j\ d\mathbf{x} = &- \sum_{i\in\mathcal{A}} u_i \int_\Omega d(\mathbf{x})\ \nabla\varphi_i \cdot \nabla\varphi_j\ d\mathbf{x} + \sum_{i\in\mathcal{A}} u_i \int_\Omega r(\mathbf{x})\ \varphi_i\ \varphi_j\ d\mathbf{x} \qquad (3.12)\\ &+ \frac{1}{3}|\omega_j| p(u_j) + \int_\Omega q_h(\mathbf{x},t)\ \varphi_j\ d\mathbf{x} + \sum_{i\in\mathcal{A}} \int_\Omega g_i\ \varphi_i\ \varphi_j\ ds, \qquad j\in\mathcal{A},\\ u_i(t_0) = u_{0,i}, &\qquad i\in\mathcal{A},\end{aligned}$$

welches in Matrixform die Gestalt

$$\boxed{\mathbf{M}\frac{\partial \mathbf{u}}{\partial t} = -\mathbf{S}\mathbf{u} + \widehat{\mathbf{R}}\mathbf{u} + \mathbf{L}p(\mathbf{u}) + \mathbf{M}\mathbf{q} + \mathbf{B}\mathbf{g}_{\text{Neu}}} \qquad (3.13)$$

annimmt. Die Größen $\mathbf{M}$, $\mathbf{S}$, $\widehat{\mathbf{R}}$, $\mathbf{L}$, $\mathbf{q}$ und $\mathbf{u}$ sind wie in (3.8) definiert. Die Matrix $\mathbf{B}$ ist durch

$$\mathbf{B} = \left(\int_{\partial\Omega} \varphi_i\ \varphi_j\ d\mathbf{x} \right)_{i,j\in\mathcal{A}} \qquad (3.14)$$

gegeben.

Für *lineare* Reaktions-Diffusions-Gleichungen existiert eine eindeutige Lösung der diskreten Probleme (3.9) und (3.13), wenn gewisse Regularitätsbedingungen an die Koeffizienten erfüllt sind, siehe Knabner/Angermann [94, Satz 6.6].

3.1.5 Reduktion der Massenmatrix

Eine Vereinfachung der Systeme (3.9) und (3.13) ergibt sich, wenn die in den Matrizen $\mathbf{M}$ und $\widehat{\mathbf{R}}$ auftretenden Integrale durch die in Definition 3.5 angegebene Trapezregel approximiert werden. Ein solches Vorgehen wird als **Reduktion der Massenmatrix** bezeichnet[1]. Für die in der Massenmatrix $\mathbf{M}$ auftretenden Integrale ergibt sich

$$\int_\Omega \varphi_i\varphi_j\ d\mathbf{x} \approx Q_\Omega^{\mathcal{T}_h}(\varphi_i\varphi_j) = \begin{cases} \frac{1}{3}|\omega_i|, & \text{falls } i=j,\\ 0, & \text{sonst.}\end{cases}$$

Die Matrix $\mathbf{M}$ wird also durch die Matrix

$$\frac{1}{3}\ \text{diag}\ (|\omega_i|)_{i\in\mathcal{A}}$$

approximiert. Das ist jedoch gerade die in (3.8) definierte Matrix $\mathbf{L}$. In analoger Weise approximiert man

$$\int_{\Omega_h} r(\mathbf{x})\varphi_i\varphi_j\ d\mathbf{x} \approx Q_{\Omega_h}(r(\mathbf{x})\varphi_i\varphi_j) = \begin{cases} \frac{1}{3}|\omega_i| r(\mathbf{x}_i), & \text{falls } i=j,\\ 0, & \text{sonst.}\end{cases}$$

[1] Häufig findet sich in der Literatur die englische Bezeichnug „Lumping“, siehe Bemerkung 3.8.

und erhält als Näherung für die Matrix $\widehat{\mathbf{R}}$ den Ausdruck

$$\frac{1}{3}\ \text{diag}\ (|\omega_i| r(\mathbf{x}_i))_{i\in\mathcal{A}} = \mathbf{LR}.$$

Ersetzt man in dem System (3.13) $\mathbf{M}$ durch $\mathbf{L}$ und $\widehat{\mathbf{R}}$ durch $\mathbf{LR}$ und multipliziert anschließend mit $\mathbf{L}^{-1}$, so ergibt sich das

System mit reduzierter Massenmatrix für Neumann-Randbedingungen

$$\frac{\partial \mathbf{u}}{\partial t} = -\mathbf{L}^{-1}\mathbf{S}\mathbf{u} + \mathbf{R}\mathbf{u} + p(\mathbf{u}) + \mathbf{q} + \mathbf{L}^{-1}\mathbf{B}\mathbf{g}_{\text{Neu}}. \tag{3.15}$$

Man beachte jedoch, daß die Symmetrie der Steifigkeitsmatrix beim Übergang von $\mathbf{S}$ nach $\mathbf{L}^{-1}\mathbf{S}$ verloren geht.

In analoger Weise erhält man das

System mit reduzierter Massenmatrix für Dirichlet-Randbedingung

$$\frac{\partial \mathbf{u}_{\mathcal{I}}}{\partial t} = -\mathbf{L}_{\mathcal{I},\mathcal{I}}^{-1}\mathbf{S}_{\mathcal{I},\mathcal{I}}\ \mathbf{u}_{\mathcal{I}} + \mathbf{R}_{\mathcal{I},\mathcal{I}}\mathbf{u}_{\mathcal{I}} + p(\mathbf{u}_{\mathcal{I}}) + \mathbf{q}_{\mathcal{I}} - \mathbf{L}_{\mathcal{I},\mathcal{I}}^{-1}\mathbf{S}_{\mathcal{I},\mathcal{B}}\ \mathbf{u}_{\mathcal{B}} \tag{3.16}$$

Bemerkung 3.8. Die Matrix $\mathbf{L}$ ist nach ihrer Definition gerade die Diagonalmatrix, die die Zeilensummen der Massenmatrix $\mathbf{M}$ als Diagonalelemente enthält. Daher leitet sich die englische Bezeichnung „Lumping“ ab. „Lumping“ bedeutet „etwas zusammenballen“. In diesem Sinne werden die Außerdiagonalelemente der Massenmatrix in der Diagonale „zusammengeballt“, d.h. aufaddiert. □

3.1.6 Elementweise Berechnung der Matrizen

Zur Aufstellung der Differentialgleichungs-Systeme (3.9), (3.13), (3.16) und (3.15) benötigt man u.a. die Matrizen $\mathbf{M}$, $\mathbf{S}$, $\widehat{\mathbf{R}}$, $\mathbf{L}$ und $\mathbf{B}$. Diese Matrizen enthalten Integrale der lokalen Basisfunktionen φ_i und φ_j. Die Matrizen sind dünn besetzt, da die entsprechenden Integrale nur dann von 0 verschieden sind, wenn i und j gleich sind oder benachbarten Knoten der Triangulierung entsprechen. Die Matrizen $\mathbf{M}$, $\mathbf{S}$, $\widehat{\mathbf{R}}$ und $\mathbf{L}$ setzen sich additiv aus den Beiträgen der einzelnen Dreieckselemente, den sogenannten **Elementmatrizen** zusammen. Sei $T \in \mathcal{T}$ ein Element der Triangulierung, welches die Ecken mit den Indizes i, j, k habe. Dann sind die zu T gehörigen Elementmatrizen durch

$$\mathbf{M}_T = \left(\int_T \varphi_l\ \varphi_m\ d\mathbf{x} \right)_{l,m\in\{i,j,k\}},$$

$$\mathbf{S}_T = \left(\int_T d(\mathbf{x})\ \nabla\ \varphi_l \cdot \nabla\ \varphi_m\ d\mathbf{x} \right)_{l,m\in\{i,j,k\}}, \tag{3.17}$$

$$\widehat{\mathbf{R}}_T = \left(\int_T r(\mathbf{x})\ \varphi_l\ \varphi_m\ d\mathbf{x} \right)_{l,m\in\{i,j,k\}},$$

und

$$\mathbf{L}_T = \left(Q_T^{\mathcal{T}_h}(\varphi_l\ \varphi_m) \right)_{l,m\in\{i,j,k\}}$$

gegeben. Der folgende Algorithmus beschreibt am Beispiel der Massenmatrix $\mathbf{M}$, wie $\mathbf{M}$ aus den einzelnen Elementmatrizen $\mathbf{M}_T$ zusammengesetzt wird. Dieser Vorgang wird als **Assemblierung** bezeichnet und verläuft völlig analog für die Matrizen $\mathbf{S}$, $\widehat{\mathbf{R}}$ und $\mathbf{L}$.

Algorithmus 3.9.

$\mathbf{M} := \mathbf{0}$

for $T \in \mathcal{T}_h$

  for $i = 1, 2, 3$

    for $j = 1, 2, 3$

      $(\mathbf{M})_{C_i(T),C_j(T)} := (\mathbf{M}_T)_{ij}$

    end

  end

end

□

Hier bezeichnet $(\mathbf{M}_T)_{ij}$ das Element in Zeile i, Spalte j der Matrix $\mathbf{M}_T$. Eine entsprechende Schreibweise werden wir auch im folgenden verwenden. Die Eckenindex-Funktion C ist die in Abschnitt 3.1.2 definierte.

Die Berechnung der Elementmatrizen wird im folgenden präzisiert.

Berechnung von $\mathbf{M}_T$

Die Matrix $\mathbf{M}_T$ hängt nur vom Flächeninhalt des Dreiecks T ab; durch exakte Integration erhält man nämlich (siehe etwa Schwarz [146])

$$\mathbf{M}_T = \frac{|T|}{12} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

Berechnung von $\mathbf{S}_T$

Die Berechnung der Matrix $\mathbf{S}_T$ kann durch exakte Integration erfolgen, wenn die auftretenden Terme integrierbar sind. Eine andere Möglichkeit besteht darin, die Funktion d durch eine stückweise lineare Funktion $d_h \in V_h$ zu approximieren. Dann können die Integrale

$\int_T d_h(\mathbf{x})\ \nabla\varphi_l \cdot \nabla\varphi_m\ d\mathbf{x}$ exakt berechnet werden. Wir setzen $d_i = d(\mathbf{x}_i)$, $i \in \mathcal{A}$. Die Eckpunkte des Dreiecks T seien $\mathbf{x}_i$, $\mathbf{x}_j$ und $\mathbf{x}_k$. Dann berechnet sich das Diagonalelement $(\mathbf{S}_T)_{11}$ gemäß

$$(\mathbf{S}_T)_{11} = \int_T d_h(\mathbf{x})\ \nabla\ \varphi_i \cdot \nabla\ \varphi_i\ d\mathbf{x} = \frac{d_i + d_j + d_k}{12|T|}\|\mathbf{x}_j - \mathbf{x}_k\|^2.$$

Die beiden übrigen Diagonalelemente erhält man analog durch entsprechende Vertauschung von $i,\ j$ und k. Das Nicht-Diagonalelement $(\mathbf{S}_T)_{12}$ ist durch

$$(\mathbf{S}_T)_{12} = \int_T d_h(\mathbf{x})\ \nabla\ \varphi_i \cdot \nabla\ \varphi_j\ d\mathbf{x} = \frac{d_i + d_j + d_k}{12|T|}(\mathbf{x}_i - \mathbf{x}_k) \cdot (\mathbf{x}_k - \mathbf{x}_j)$$

gegeben, die übrigen Nicht-Diagonalelemente erhält man wieder analog. Setzt man $\sigma_{l,m,n} := (\mathbf{x}_l - \mathbf{x}_m) \cdot (\mathbf{x}_m - \mathbf{x}_n)$ für $l, m, n \in \{i, j, k\}$, so ergibt sich die Matrix $\mathbf{S}_T$ zu

$$\mathbf{S}_T = \frac{d_i + d_j + d_k}{12|T|} \begin{pmatrix} -\sigma_{j,k,j} & \sigma_{i,k,j} & \sigma_{i,j,k} \\ \sigma_{j,k,i} & -\sigma_{k,i,k} & \sigma_{j,i,k} \\ \sigma_{k,j,i} & \sigma_{k,i,j} & -\sigma_{i,j,i} \end{pmatrix}. \tag{3.18}$$

Berechnung von $\widehat{\mathbf{R}}_T$

Es sei wieder T das Dreieck mit den Eckpunkten $\mathbf{x}_i$, $\mathbf{x}_j$ und $\mathbf{x}_k$. Approximiert man r durch eine Funktion $r_h \in V_h$ und setzt $r_i = r(\mathbf{x}_i)$ für alle $i \in \mathcal{A}$, so berechnet sich das Diagonalelement $(\widehat{\mathbf{R}}_T)_{11}$ gemäß

$$(\widehat{\mathbf{R}}_T)_{11} = \int_T r_h(\mathbf{x})\ \varphi_i\ \varphi_i\ d\mathbf{x} = \left(\frac{r_i}{10} + \frac{r_j + r_k}{30}\right)|T|,$$

die übrigen Diagonalelemente $(\widehat{\mathbf{R}}_T)_{22}$ und $(\widehat{\mathbf{R}}_T)_{33}$ entsprechend. Das Nicht-Diagonalelement $(\widehat{\mathbf{R}}_T)_{12}$ ergibt sich zu

$$(\widehat{\mathbf{R}}_T)_{12} = \int_T r_h(\mathbf{x})\ \varphi_i\ \varphi_j\ d\mathbf{x} = \left(\frac{r_i + r_j}{30} + \frac{r_k}{60}\right)|T|,$$

die übrigen Nicht-Diagonalelemente erhält man in analoger Weise. Setzt man $\varrho_{l,m,n} := 6r_l + 2(r_m + r_n)$ und $\varkappa_{l,m,n} := 2(r_l + r_m) + r_n$ für $l, m, n \in \{i, j, k\}$, so ist $\widehat{\mathbf{R}}_T$ durch

$$\widehat{\mathbf{R}}_T = \frac{|T|}{60} \begin{pmatrix} \varrho_{i,j,k} & \varkappa_{i,j,k} & \varkappa_{i,k,j} \\ \varkappa_{j,i,k} & \varrho_{j,k,i} & \varkappa_{j,k,i} \\ \varkappa_{k,i,j} & \varkappa_{k,j,i} & \varrho_{k,i,j} \end{pmatrix}$$

gegeben.

Berechnung von $\mathbf{L}_T$

Die Elementmatrix $\mathbf{L}_T$ ist mit

$$\mathbf{L}_T = \frac{|T|}{3}\mathbf{I}$$

gegeben, wobei $\mathbf{I}$ die 3×3-Einheitsmatrix ist.

Die Berechnung von $\mathbf{B}$ ist nur für inhomogene Neumannsche Randbedingungen, d.h. für $g_{\text{Neu}} \neq 0$ vonnöten. $\mathbf{B}$ wird aus den Beiträgen der Randkanten von Ω zusammengesetzt. Sei E eine solche Randkante mit den Endpunkten $\mathbf{x}_i$ und $\mathbf{x}_j$. Dann berechnet sich die zugehörige Kantenmatrix

$$\mathbf{B}_E = \begin{pmatrix} \int\limits_E \varphi_i \, \varphi_i & \int\limits_E \varphi_i \, \varphi_j \\ \int\limits_E \varphi_j \, \varphi_i & \int\limits_E \varphi_j \, \varphi_j \end{pmatrix}$$

wie folgt:

Berechnung von $\mathbf{B}_E$

Die Matrix $\mathbf{B}_E$ hängt nur von der Länge der Randkante E ab. Durch Ausrechnen der Integrale erhält man

$$\mathbf{B}_E = \frac{|E|}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Die Matrix $\mathbf{B}$ entsteht jetzt wieder durch Addition der einzelnen Kantenmatrizen $\mathbf{B}_E$, wie der folgende Algorithmus zeigt.

Algorithmus 3.10.

$\mathbf{B} := \mathbf{0}$

for $E \in \partial\mathcal{T}_h$

  for $i = 1, 2$

    for $j = 1, 2$

      $(\mathbf{B})_{c_i(E), c_j(E)} := (\mathbf{B}_E)_{ij}$

    end

  end

end

□

3.2 Ortsdiskretisierung semilinearer Reaktions-Diffusions-Gleichungen auf gekrümmten Flächen

3.2.1 Zwei Varianten der Diskretisierung

In Abschnitt 3.1 war Ω stets ein Gebiet in $\mathbb{R}^2$. In einigen Anwendungen spielen jedoch Differentialgleichungen eine Rolle, bei denen Ω ein Gebiet auf einer Mannigfaltigkeit S des $\mathbb{R}^n$ ist, d.h. $\Omega \subset S \subset \mathbb{R}^n$, $\dim S < n$. Ein Beispiel hierfür soll uns in Abschnitt 10.9 beschäftigen.

Dort werden die Reaktions-Diffusions-Gleichungen von KRINSKY et al., die zur Modellierung erregbarer Medien dienen, auf einer gekrümmten Fläche in $\mathbb{R}^3$ betrachtet. Im folgenden wollen wir uns der Diskretisierung derartiger Probleme zuwenden. Gekrümmte Flächen hinreichender Glattheit werden in der Differentialgeometrie mit dem Konzept der Riemannschen Mannigfaltigkeit beschrieben. Eine kurze Zusammenstellung einiger Aussagen zur Theorie Riemannscher Mannigfaltigkeiten findet sich in Anhang A.

Der Einfachheit halber betrachten wir hier nur den Fall einer zweidimensionalen Mannigfaltigkeit S, die in den $\mathbb{R}^3$ eingebettet ist. Auf S sei ein Gebiet $\Omega \subset S$ gegeben. Eine semilineare Reaktions-Diffusions-Gleichung auf Ω läßt sich in der folgenden Form angeben, vgl (2.2):

$$\begin{aligned} \frac{\partial u}{\partial t}(\mathbf{x},t) &= \operatorname{div}_S(d(\mathbf{x})\nabla u(\mathbf{x},t)) + r(\mathbf{x})u(\mathbf{x},t) \\ &\quad + p(u(\mathbf{x},t)) + q(\mathbf{x},t), \qquad \mathbf{x}\in\Omega, \quad t\in[t_0,t_e], \\ u(\mathbf{x},t_0) &= u_0(\mathbf{x}), \qquad \mathbf{x}\in\Omega. \end{aligned} \tag{3.19}$$

Falls Ω einen Rand besitzt, stellen wir zusätzlich Dirichlet- oder Neumann-Randbedingungen

$$u(\mathbf{x},t) = g_{\mathrm{Dir}}(\mathbf{x},t), \qquad \mathbf{x}\in\partial\Omega$$

bzw.

$$d(\mathbf{x},t)\nabla u\cdot\mathbf{n}_{\partial\Omega} = g_{\mathrm{Neu}}(\mathbf{x},t), \qquad \mathbf{x}\in\partial\Omega.$$

Die Symbole ∇_S und div_S bezeichnen den **tangentialen Gradienten** bzw. die **tangentiale Divergenz** auf der Fläche S. Falls $d(x) = d = \text{const.}$ ist, so ist (3.19) äquivalent zu der Gleichung

$$u_t = d\Delta_S u + r(\mathbf{x})\ u + p(u(\mathbf{x},t)) + q(\mathbf{x},t),$$

wobei Δ_S der **Laplace-Beltrami-Operator** ist. Die Operatoren ∇_S, div_S und Δ_S sind in Anhang A definiert. Es gibt generell zwei Möglichkeiten, eine Differentialgleichung auf einer Mannigfaltigkeit zu diskretisieren. Zum einen kann man die zweidimensionale Fläche S geeignet parametrisieren; man ordnet dabei jedem Punkt auf S ein Parameterpaar (φ,ϑ) zu. Die Operatoren ∇_S, div_S und Δ_S besitzen eine Darstellung in diesen Parametern. Man diskretisiert dann die Differentialgleichung im Parameterraum, beispielsweise mit einem Differenzenverfahren. Das folgende Beispiel soll diese Vorgehensweise illustrieren:

Beispiel 3.11. Wir betrachten die Wärmeleitungsgleichung

$$u_t = \Delta_S u$$

auf der Einheitssphäre, die durch geographische Länge φ und geographische Breite ϑ parametrisiert wurde:

$$x = \cos\varphi\cos\vartheta, \qquad y = \sin\varphi\cos\vartheta, \qquad z = \sin\vartheta$$

In Anhang A wird der Laplace-Beltrami-Operator

$$\Delta_S u = \frac{1}{\cos^2\vartheta}\frac{\partial^2 u}{\partial\varphi^2} - \tan\vartheta\frac{\partial u}{\partial\vartheta} + \frac{\partial^2 u}{\partial\vartheta^2}$$

hergeleitet. Man diskretisiert nun die resultierende Differentialgleichung

$$u_t = \frac{1}{\cos^2\vartheta}\frac{\partial^2 u}{\partial\varphi^2} - \tan\vartheta\frac{\partial u}{\partial\vartheta} + \frac{\partial^2 u}{\partial\vartheta^2},$$

die eine Konvektions-Diffusions-Gleichung in den Variablen φ und ϑ darstellt. Das kann durch ein Differenzenverfahren oder mittels finiter Elemente geschehen. □

Eine zweite Möglichkeit besteht darin, die Fläche S als in den Raum $\mathbb{R}^3$ eingebettet zu betrachten, d.h. jeder Punkt auf S ist durch die $\mathbb{R}^3$-Koordinaten (x, y, z) festgelegt. Auf derartigen eingebetteten Flächen kann man die Diskretisierung der Differentialgleichung direkt mit der Methode der finiten Elemente vornehmen, ohne einer Parametrisierung zu bedürfen. Zur Erläuterung dieser Methode werden einige Hilfsmittel benötigt. Wir werden im folgenden Abschnitt näher auf diese Vorgehensweise eingehen.

3.2.2 Finite Elemente auf Mannigfaltigkeiten

Die hier betrachtete Diskretisierung partieller Differentialgleichungen auf Mannigfaltigkeiten fußt auf einer von Dziuk [54] (1988) angegebenen Methode zur Diskretisierung des Laplace-Beltrami-Operators. Wie in Abschnitt 3.1 beschreiben wir der Einfachheit halber wieder nur die Diskretisierung mit linearen Dreieckselementen. Das auf der zweidimensionalen Mannigfaltigkeit S liegende Gebiet Ω wird durch ein Polyeder, also eine aus Dreiecken stetig zusammengesetzte Fläche Ω_h approximiert. Die Eckpunkte $\mathbf{x}_i$ von Ω_h liegen dabei in Ω, die Randpunkte $\mathbf{x}_i \in \partial\Omega_h$ liegen auf $\partial\Omega$, falls Ω berandet ist. Die Menge der Dreiecke, aus denen sich Ω_h zusammensetzt, wird wieder als Triangulierung $\mathcal{T}_h$ bezeichnet. Die Indexmengen $\mathcal{I}$, $\mathcal{B}$ und $\mathcal{A}$ sind wie im ebenen Fall definiert, siehe (3.4). Lebesgue- und Sobolev-Räume können in ähnlicher Weise wie im $\mathbb{R}^n$ auch auf hinreichend glatten Mannigfaltigkeiten eingeführt werden. Durch Multiplikation der Differentialgleichung (3.19) mit einer Funktion $v \in H^1(S)$, Integration und Anwendung der Greenschen Formel, siehe (A.2), erhält man die schwache Formulierung, vgl. (2.14), (2.15).

Schwache Formulierung für Dirichlet-Randbedingung

Gegeben sei $u_0 \in L^2(\Omega)$. Finde eine Funktion $u \in L^2(]t_0, t_e[, H^1(\Omega))$ mit $u_t \in L^2(]t_0, t_e[, L^2(\Omega))$, die den folgenden drei Bedingungen genügt:

1. Für alle $v \in H_0^1(\Omega)$ und alle $t \in]t_0, t_e[$ gilt

$$\begin{aligned} \frac{\partial}{\partial t} \int_\Omega u\ v\ d\mathbf{x} \;&=\; -\int_\Omega d(\mathbf{x})\ \nabla_S u \cdot \nabla_S v\ d\mathbf{x} + \int_\Omega r(\mathbf{x})\ u\ v\ d\mathbf{x} \\ &\quad + \int_\Omega p(u(\mathbf{x},t))\ v\ d\mathbf{x} + \int_\Omega q(\mathbf{x},t)\ v\ d\mathbf{x}. \end{aligned} \tag{3.20}$$

2. Für fast alle $\mathbf{x} \in \Omega$ gilt $u(\mathbf{x}, t_0) = u_0(\mathbf{x})$.

3. Für alle $t \in [t_0, t_e]$ und fast alle $\mathbf{x} \in \partial\Omega$ gilt $\mathrm{tr}_{\partial\Omega} u(\mathbf{x}, t) = g_{\mathrm{Dir}}(\mathbf{x}, t)$.

Schwache Formulierung für Neumann-Randbedingung

Gegeben sei $u_0 \in L^2(\Omega)$. Finde eine Funktion $u \in L^2(]t_0, t_e[, H^1(\Omega))$ mit $u_t \in L^2(]t_0, t_e[, L^2(\Omega))$, so daß die folgenden beiden Aussagen gelten:

1. Für alle $v \in H^1(\Omega)$ und alle $t \in]t_0, t_e[$ gilt

$$\begin{aligned} \frac{\partial}{\partial t} \int_\Omega u\ v\ d\mathbf{x} &= -\int_\Omega d(\mathbf{x})\ \nabla_S u \cdot \nabla_S v\ d\mathbf{x} + \int_\Omega r(\mathbf{x})\ u\ v\ d\mathbf{x} \\ &\quad + \int_\Omega p(u(\mathbf{x},t))\ v\ d\mathbf{x} + \int_\Omega q(\mathbf{x},t)\ v\ d\mathbf{x} \\ &\quad + \int_{\partial\Omega} g_{\text{Neu}}\ v\ ds, \end{aligned} \tag{3.21}$$

   wobei ds das Bogenelement auf $\partial\Omega$ ist.

2. Für fast alle $\mathbf{x} \in \Omega$ gilt $u(\mathbf{x}, t_0) = u_0(\mathbf{x})$.

Die schwache Formulierung einer randlosen Fläche entspricht der für Neumann-Randbedingungen, wobei der Randterm $\int_{\partial\Omega} g_{\text{Neu}}\ v\ ds$ entfällt.

Die diskreten Räume V_h und $V_{h,0}$ werden wie in Definition 3.2 eingeführt, wobei Ω durch Ω_h zu ersetzen ist. Man approximiert in der schwachen Formulierung $H^1(\Omega)$ durch V_h und $H_0^1(\Omega)$ durch $V_{h,0}$. Im Ergebnis erhält man die in (3.9), (3.13), (3.16) und (3.15) angegebenen Matrixformulierungen der Ortsdiskretisierung. Für ein randloses Gebiet Ω entspricht das System dem für Neumann-Randbedingungen, jedoch ohne den Term $\mathbf{B}\mathbf{g}_{\text{Neu}}$.

Die in der Matrixformulierung auftretenden Matrizen und Vektoren werden analog zu (3.8), (3.11) und (3.14) definiert, wobei hier wieder Ω durch Ω_h und ∇ durch ∇_{Ω_h} ersetzt werden muß. Die Element- und Kantenmatrizen $\mathbf{M}_T$, $\widehat{\mathbf{R}}_T$, $\mathbf{R}_T$, $\mathbf{L}$, $\widehat{\mathbf{q}}$, $\mathbf{q}$, $\mathbf{g}_{\text{Neu}}$ und $\mathbf{B}_E$ werden daher *auf die gleiche Weise wie im ebenen Fall*, d.h. wie in Abschnitt 3.1.6 angegeben, berechnet. In die Matrix $\mathbf{S}_T$ gehen jedoch die tangentialen Gradienten ∇_{Ω_h} ein. Seien $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k \in \mathbb{R}^3$ die Eckpunkte des Dreiecks T, so gilt

$$\mathbf{S}_T = \left(\int_T d(\mathbf{x})\ \nabla_T \varphi_l \cdot \nabla_T \varphi_m\ d\mathbf{x} \right)_{l,m \in \{i,j,k\}}, \tag{3.22}$$

etc.. Der folgende Satz gibt an, wie die Element-Steifigkeitsmatrix $\mathbf{S}_T$ berechnet werden kann.

Satz 3.12. *Für ein Dreieck $T \subset \Omega_h$ mit den Eckpunkten $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k \in \mathbb{R}^3$ seien $\varphi_i, \varphi_j, \varphi_k$ die lokalen Basisfunktionen. Mit $d_j := d(\mathbf{x}_l) = d_h(\mathbf{x}_l)$, $l \in \{i,j,k\}$ werde der Diffusionskoeffizient in den Eckpunkten bezeichnet. Für $l, m, n \in \{i,j,k\}$ definieren wir die Skalarprodukte*

$$\sigma_{l,m,n} := (\mathbf{x}_l - \mathbf{x}_m) \cdot (\mathbf{x}_m - \mathbf{x}_n).$$

Dann gilt

$$\mathbf{S}_T = \frac{d_i + d_j + d_k}{12|T|} \begin{pmatrix} -\sigma_{j,k,j} & \sigma_{i,k,j} & \sigma_{i,j,k} \\ \sigma_{j,k,i} & -\sigma_{k,i,k} & \sigma_{j,i,k} \\ \sigma_{k,j,i} & \sigma_{k,i,j} & -\sigma_{i,j,i} \end{pmatrix}. \tag{3.23}$$

Bemerkung 3.13. Zwischen der Aussage dieses Satzes und der in (3.18) angegebenen Beziehung herrscht eine bemerkenswerte Übereinstimmung. Die Elements-Steifigkeitsmatrix $\mathbf{S}_T$ berechnet sich also *in der gleichen Weise* wie im ebenen Fall aus den Skalarprodukten der Vektoren $\mathbf{x}_i$.

Der Beweis des Satzes 3.12 bedarf zunächst einiger Aussagen aus der linearen Algebra, siehe etwa KOECHER [95]. Wir betrachten lineare Abbildungen im $\mathbb{R}^3$, die die Abstände beliebiger Punkte unverändert lassen, d.h. Isometrien bezüglich der Euklidischen Metrik sind. Diese Abbildungen werden auch als **Bewegungen** bezeichnet. Jede Bewegung läßt offensichtlich die Gestalt eines Dreiecks unverändert. Es gilt aber darüberhinaus, daß bei Vorgabe zweier kongruenter Dreiecke T und $\widetilde{T}$ stets eine Bewegung existiert, die T in $\widetilde{T}$ überführt. Jede Bewegung $b : \mathbb{R}^3 \to \mathbb{R}^3$ läßt sich als Nacheinanderausführung einer **orthogonalen Abbildung** und einer **Translation** schreiben:

$$b(\mathbf{x}) = \mathbf{Q}\mathbf{x} + \mathbf{t}.$$

Dabei ist $\mathbf{Q}$ eine orthogonale Matrix, d.h. $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$. Wir zeigen nun drei Lemmata.

Lemma 3.14. *Es seien $v, \widetilde{v} : \mathbb{R}^3 \to \mathbb{R}$ zwei Funktionen. Die Funktion $\widetilde{v}$ gehe durch eine Bewegung der Koordinaten aus v hervor, d.h. es gilt $\widetilde{v}(\mathbf{Q}\mathbf{x}+\mathbf{t}) = v(\mathbf{x})$, wobei $\mathbf{Q}$ eine orthogonale 3×3-Matrix ist. Dann hat der Gradient die Darstellung*

$$\nabla\widetilde{v}(\mathbf{Q}\mathbf{x} + \mathbf{t}) = \mathbf{Q}\nabla v(\mathbf{x}).$$

Beweis. Wegen $\mathbf{Q}^{-1} = \mathbf{Q}^T$ gilt $\widetilde{v}(\mathbf{x}) = v(\mathbf{Q}^T(\mathbf{x} - \mathbf{t}))$. Die partiellen Ableitungen von $\widetilde{v}$ berechnen sich nach der Kettenregel:

$$\frac{\partial\widetilde{v}}{\partial\mathbf{x}_i}(\mathbf{x}) = (\mathbf{Q}^T\mathbf{e}_i) \cdot \nabla v(\mathbf{Q}^T(\mathbf{x} - \mathbf{t})), \qquad i = 1, 2, 3$$

Der Gradient ergibt sich also zu

$$\nabla\widetilde{v}(\mathbf{x}) = \mathbf{Q}\nabla v(\mathbf{Q}^T(\mathbf{x} - \mathbf{t})),$$

woraus nach Substitution $\mathbf{x} \mapsto \mathbf{Q}\mathbf{x} + \mathbf{t}$ die Behauptung folgt. □

Eine analoge Aussage gilt auch für den tangentialen Gradienten auf einem Dreieck in $\mathbb{R}^3$:

Lemma 3.15. *Es sei T ein Dreieck im $\mathbb{R}^3$ und $\mathbf{Q}$ eine orthogonale 3×3-Matrix. Gegeben sei ferner eine Funktion $v : \mathbb{R}^3 \to \mathbb{R}$. Das Dreieck*

$$\widetilde{T} = \mathbf{Q}T + \mathbf{t}$$

entsteht durch eine Bewegung von T. Dann gilt für den tangentialen Gradienten die Beziehung

$$\nabla_{\widetilde{T}}\ \widetilde{v}(\mathbf{Q}\mathbf{x} + \mathbf{t}) = \mathbf{Q}\nabla_T\ v(\mathbf{x}).$$

Beweis. Wir bezeichnen mit $\mathbf{n}_T$ den Flächennormalvektor des Dreiecks T. Der Vektor $\mathbf{n}_{\widetilde{T}} = \mathbf{Q}\mathbf{n}_T$ steht offensichtlich senkrecht auf dem Dreieck $\widetilde{T}$ und hat den Betrag 1, ist also Flächennormalvektor von $\widetilde{T}$. Nach Lemma A.3 haben die tangentialen Gradienten von v und $\widetilde{v}$ die Darstellung

$$\begin{aligned} \nabla_T\, v &= \nabla v - (\nabla v \cdot \mathbf{n}_T)\mathbf{n}_T, \\ \nabla_{\widetilde{T}}\, \widetilde{v} &= \nabla \widetilde{v} - (\nabla \widetilde{v} \cdot \mathbf{n}_{\widetilde{T}})\mathbf{n}_{\widetilde{T}}. \end{aligned}$$

Wegen $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ und Lemma 3.14 folgt

$$\begin{aligned} \nabla_{\widetilde{T}}\, \widetilde{v}(\mathbf{Q}\mathbf{x}+\mathbf{t}) &= \nabla \widetilde{v}(\mathbf{Q}\mathbf{x}+\mathbf{t}) - (\nabla \widetilde{v}(\mathbf{Q}\mathbf{x}+\mathbf{t}) \cdot \mathbf{n}_{\widetilde{T}})\mathbf{n}_{\widetilde{T}} \\ &= \mathbf{Q}\nabla v(\mathbf{x}) - ((\mathbf{Q}\nabla v(\mathbf{x})) \cdot (\mathbf{Q}\mathbf{n}_T))\mathbf{Q}\mathbf{n}_T \\ &= \mathbf{Q}\nabla v(\mathbf{x}) - \mathbf{Q}(\nabla v(\mathbf{x}) \cdot \mathbf{n}_T)\mathbf{n}_T \;=\; \mathbf{Q}\nabla_T\, v(\mathbf{x}). \end{aligned}$$

□

Lemma 3.16. *Die Einträge $\int_T \nabla_T\, \varphi_l(\mathbf{x}) \cdot \nabla_T\, \varphi_m(\mathbf{x})\; d\mathbf{x},\; l,m \in \{i,j,k\}$ der in (3.23) gegebenen Element-Steifigkeitsmatrix bleiben bei einer Bewegung $\widetilde{T} = \mathbf{Q}T + \mathbf{t}$ des Dreiecks T invariant, d.h. mit $\widetilde{\varphi}_l(\mathbf{Q}\mathbf{x}+\mathbf{t}) = \varphi_l(\mathbf{x}),\; l \in \{i,j,k\}$ folgt*

$$\int_{\widetilde{T}} \nabla_{\widetilde{T}}\, \widetilde{\varphi}_l(\mathbf{x}) \cdot \nabla_{\widetilde{T}}\, \widetilde{\varphi}_m(\mathbf{x})\; d\mathbf{x} = \int_T \nabla_T\, \varphi_l(\mathbf{x}) \cdot \nabla_T\, \varphi_m(\mathbf{x})\; d\mathbf{x}.$$

Beweis. Aus Lemma 3.15 folgt

$$\begin{aligned} \nabla_{\widetilde{T}}\, \widetilde{\varphi}_l(\mathbf{x}) \cdot \nabla_{\widetilde{T}}\, \widetilde{\varphi}_m(\mathbf{x}) &= \mathbf{Q}\nabla_T\, \varphi_l(\mathbf{Q}^T(\mathbf{x}-\mathbf{t})) \cdot \mathbf{Q}\nabla_T\, \varphi_m(\mathbf{Q}^T(\mathbf{x}-\mathbf{t})) \\ &= \nabla_T\, \varphi_l(\mathbf{Q}^T(\mathbf{x}-\mathbf{t})) \cdot \nabla_T\, \varphi_m(\mathbf{Q}^T(\mathbf{x}-\mathbf{t})) \end{aligned}$$

wegen der Orthogonalität von $\mathbf{Q}$. Integriert man über $\widetilde{T}$, so erhält man

$$\int_{\widetilde{T}} \nabla_{\widetilde{T}}\, \widetilde{\varphi}_l(\mathbf{x}) \cdot \nabla_{\widetilde{T}}\, \widetilde{\varphi}_m(\mathbf{x})\; d\mathbf{x} = \int_{\widetilde{T}} \nabla_T\, \varphi_l(\mathbf{Q}^T(\mathbf{x}-\mathbf{t})) \cdot \nabla_T\, \varphi_m(\mathbf{Q}^T(\mathbf{x}-\mathbf{t}))\; d\mathbf{x}.$$

Da $|\det \mathbf{Q}| = 1$ ist, folgt mit linearer Substitution die Behauptung. □

Bemerkung 3.17. Die hier verwendete Kurzschreibweise der Integrale in der Form $\int_T w(\mathbf{x})\, d\mathbf{x}$ macht keine Angabe über die Parametrisierung von T. Damit geht das Vorzeichen dieser Integrale aus der Schreibweise nicht hervor. Wir definieren daher das Integral so, daß $\int_T d\mathbf{x} = |T| > 0$ ist. Für diese Definition sind die Elementmatrizen richtig angegeben. Eine solche Definition der Integrale hat Auswirkung auf die Formel zur linearen Substitution: Mit einer regulären konstanten Matrix B und einem konstanten Vektor $\mathbf{b}$ folgt

$$\int_T w(B\mathbf{x}+\mathbf{b})\; d\mathbf{x} = \frac{1}{|\det B|} \int_{\widetilde{T}} w(\mathbf{x})\; d\mathbf{x},$$

wobei $\widetilde{T} = BT + \mathbf{b}$ ist. □

Ausgerüstet mit diesen Werkzeugen, läßt sich nun Satz 3.12 leicht beweisen:

Beweis des Satzes 3.12.

Wir betrachten zunächst den Fall, daß T in der (x, y)-Ebene liegt. In diesem Fall sind die Matrizen $\mathbf{S}_T$ in (3.17) und (3.22) auf die gleiche Weise definiert, da der Gradient $\nabla\varphi_i$ in (3.17) gleich dem tangentialen Gradienten $\nabla_T\varphi_i$ in (3.22) ist. Auch die rechten Seiten der Gleichungen (3.18) und (3.23) stimmen in diesem Fall überein. Demnach folgt aus der Gültigkeit der Beziehung (3.18) auch die der Beziehung (3.23).

Aus Lemma 3.16 folgt die Invarianz von $\mathbf{S}_T$ bei einer beliebigen Bewegung des Dreiecks T in $\mathbb{R}^3$. Zu zeigen bleibt noch, daß auch die rechte Seite von (3.23) bei einer Bewegung $\mathbf{Q}T+\mathbf{t}$ des Dreiecks T unverändert bleibt. Durch eine solche Bewegung geht $\sigma_{l,m,n} = (\mathbf{x}_l-\mathbf{x}_n)\cdot(\mathbf{x}_n-\mathbf{x}_m)$ in $\widetilde{\sigma}_{l,m,n} = ((\mathbf{Q}\mathbf{x}_l+\mathbf{t})-(\mathbf{Q}\mathbf{x}_n+\mathbf{t}))\cdot((\mathbf{Q}\mathbf{x}_n+\mathbf{t})-(\mathbf{Q}\mathbf{x}_m+\mathbf{t}))$ über. Wegen der Orthogonalität von $\mathbf{Q}$ gilt jedoch

$$\begin{aligned}\widetilde{\sigma}_{l,m,n} &= ((\mathbf{Q}\mathbf{x}_l+\mathbf{t})-(\mathbf{Q}\mathbf{x}_n+\mathbf{t}))\cdot((\mathbf{Q}\mathbf{x}_n+\mathbf{t})-(\mathbf{Q}\mathbf{x}_m+\mathbf{t})) = \mathbf{Q}(\mathbf{x}_l-\mathbf{x}_n)\cdot\mathbf{Q}(\mathbf{x}_n-\mathbf{x}_m)\\ &= (\mathbf{x}_l-\mathbf{x}_n)\cdot(\mathbf{x}_n-\mathbf{x}_m) \;=\; \sigma_{l,m,n}.\end{aligned}$$

Der Faktor $(d_i+d_j+d_k)/(12|T|)$ wird ebenfalls bei einer Bewegung nicht verändert. Folglich ist die rechte Seite von (3.23) bewegungsinvariant. □

Untersuchung 3.18 (Konvergenz der Näherungslösung der Poisson-Gleichung auf der Einheitssphäre). Es sei $S = \{(x, y, z) : x^2 + y^2 + z^2 = 1\}$ die Einheitssphäre. Zur Untersuchung der Konvergenz des hier dargestellten Diskretisierungsverfahrens berechnen wir numerisch die Lösung der Poisson-Gleichung

$$-\Delta_S\, u = 2(x + y + z)$$

auf der nördlichen Hemisphäre $\Omega = \{(x, y, z) \in S : z > 0\}$. Wie in Anhang A.4 gezeigt wird, ist die exakte Lösung durch $u = x + y + z$ gegeben. Entsprechende Dirichlet-Randbedingungen werden verwendet. Die Berechnung wird auf einem uniformen Dreiecksgitter durchgeführt. Die folgende Graphik veranschaulicht den Fehler in der L^2-Norm in Abhängigkeit von der mittleren Seitenlänge der Dreiecke sowie die numerisch berechnete Konvergenzordnung.

Ergebnis. Es zeigt sich, daß näherungsweise Konvergenz zweiter Ordnung erreicht wird. Man erhält damit auf der Mannigfaltigkeit die gleiche Konvergenzordnung wie für lineare finite Elemente im $\mathbb{R}^2$. Ein derartiges numerisches Resultat wurde auch von Dziuk [54] gezeigt. □

3.3 Probleme mit räumlicher Spiegelsymmetrie

Ist ein Reaktions-Diffusions-Problem in der Ortsvariablen $\mathbf{x}$ spiegelsymmetrisch, so muß die numerische Simulation nur auf einer Hälfte des Gebietes Ω vorgenommen werden. Zunächst erläutern wir, was wir unter einem spiegelsymmetrischen Problem verstehen, das wir hier in Kurzform als symmetrisches Problem bezeichnen wollen. Wir beschränken uns in der Darstellung wieder auf semilineare Reaktions-Diffusions-Gleichungen. Eine Übertragung auf allgemeinere parabolische Systeme ist ohne Schwierigkeiten möglich.

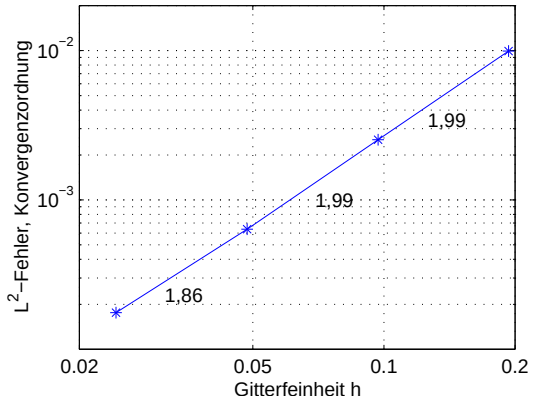


Abbildung 3.1: L^2-Fehler und numerische Konvergenzordnung der Poisson-Gleichung auf einer Halbsphäre

3.3.1 Symmetrisches Problem in der Ebene

Ausgangspunkt sei die in (3.1) gegebene Differentialgleichung

$$\begin{aligned} \frac{\partial u}{\partial t}(\mathbf{x},t) &= \nabla\cdot(d(\mathbf{x})\nabla u(\mathbf{x},t)) + r(\mathbf{x})u(\mathbf{x},t) \\ &\quad + p(u(\mathbf{x},t)) + q(\mathbf{x},t), \qquad \mathbf{x}\in\Omega, \quad t\in[t_0,t_e] \\ u(\mathbf{x},t_0) &= u_0(\mathbf{x}), \qquad \mathbf{x}\in\Omega, \end{aligned} \tag{3.24}$$

wobei Ω ein Gebiet in $\mathbb{R}^2$ ist. Das Problem sei mit Dirichlet-

$$u(\mathbf{x},t) = g_{\mathrm{Dir}}(\mathbf{x},t), \qquad \mathbf{x}\in\partial\Omega$$

oder Neumann-Randbedingungen

$$d(\mathbf{x},t)\nabla u\cdot\mathbf{n}_{\partial\Omega} = g_{\mathrm{Neu}}(\mathbf{x},t), \qquad \mathbf{x}\in\partial\Omega$$

versehen. Durch die rechte Seite von (3.24) ist der Differentialoperator

$$\Phi(u,\mathbf{x},t) := \nabla\cdot(d(\mathbf{x})\nabla u(\mathbf{x},t)) + r(\mathbf{x})u(\mathbf{x},t) + p(u(\mathbf{x},t)) + q(\mathbf{x},t)$$

definiert.

Definition 3.19. Das Problem (3.24) wird als symmetrisch bezeichnet, wenn die folgenden Aussagen gelten:

- Durch eine Gerade g in $\mathbb{R}^2$ ist die Symmetrieabbildung $\sigma_g : \mathbb{R}^2 \to \mathbb{R}^2$ gemäß

  $$\sigma_g(\mathbf{x}) = 2\mathbf{x}_g - \mathbf{x} \tag{3.25}$$

  definiert, wobei $\mathbf{x}_g \in g$ die orthogonale Projektion von $\mathbf{x}$ auf g bezeichnet.

- Das Gebiet Ω ist symmetrisch bezüglich g, d.h. $\sigma_g(\Omega) = \Omega$. Die Gerade g teilt Ω in die beiden Teilgebiete Ω_1 und Ω_2, deren gemeinsames Randstück wir mit $\gamma := \Omega \cap g$ bezeichnen.

- Die Anfangsbedingungen sind symmetrisch, d.h. $u_0(\mathbf{x}) = u_0(\sigma_g(\mathbf{x})),\ \forall \mathbf{x} \in \Omega$.
- Die Randbedingungen sind symmetrisch, d.h. es gilt

$$g_{\mathrm{Dir}}(\mathbf{x}, t) = g_{\mathrm{Dir}}(\sigma_g(\mathbf{x}), t) \qquad \forall \mathbf{x} \in \Omega, \forall t \in]t_0, t_e[$$

  bzw.

$$g_{\mathrm{Neu}}(\mathbf{x}, t) = g_{\mathrm{Neu}}(\sigma_g(\mathbf{x}), t) \qquad \forall \mathbf{x} \in \Omega, \forall t \in]t_0, t_e[.$$

- Der Operator Φ ist symmetrisch, d.h. es gilt

$$\Phi(v, \mathbf{x}, t) = \Phi(v, \sigma_g(\mathbf{x}), t) \qquad \forall \mathbf{x} \in \Omega, \forall t \in]t_0, t_e[,$$

  falls $v(\mathbf{x}, t) = v(\sigma_g(\mathbf{x}), t), \forall \mathbf{x} \in \Omega, \forall t \in]t_0, t_e[$ ist. Das ist der Fall, wenn die Koeffizienten-Funktionen d, r und q symmetrisch sind.

3.3.2 Symmetrisches Problem auf der Mannigfaltigkeit

In ähnlicher Weise definieren wir ein symmetrisches Problem auf einer Mannigfaltigkeit. Wir gehen von der in (3.1) gegebenen Differentialgleichung

$$\begin{aligned} \frac{\partial u}{\partial t}(\mathbf{x}, t) &= \mathrm{div}_S(d(\mathbf{x})\nabla u(\mathbf{x}, t)) + r(\mathbf{x})u(\mathbf{x}, t) \\ &\quad + p(u(\mathbf{x}, t)) + q(\mathbf{x}, t), \qquad \mathbf{x} \in \Omega, \quad t \in [t_0, t_e] \\ u(\mathbf{x}, t_0) &= u_0(\mathbf{x}), \qquad \mathbf{x} \in \Omega \end{aligned} \tag{3.26}$$

aus. Nun sei Ω ein Gebiet auf der zweidimensionalen Mannigfaltigkeit S. Es seien Dirichlet-

$$u(\mathbf{x}, t) = g_{\mathrm{Dir}}(\mathbf{x}, t), \qquad \mathbf{x} \in \partial\Omega$$

oder Neumann-Randbedingungen

$$d(\mathbf{x}, t)\nabla_S u \cdot \mathbf{n}_{\partial\Omega} = g_{\mathrm{Neu}}(\mathbf{x}, t), \qquad \mathbf{x} \in \partial\Omega$$

vorgegeben. Die rechte Seite von 3.26 definiert den Differentialoperator

$$\Phi(u, \mathbf{x}, t) := \mathrm{div}_S(d(\mathbf{x})\nabla_S u(\mathbf{x}, t)) + r(\mathbf{x})u(\mathbf{x}, t) + p(u(\mathbf{x}, t)) + q(\mathbf{x}, t).$$

Definition 3.20. Das Problem (3.26) wird als symmetrisch bezeichnet, wenn die folgenden Aussagen gelten:

- Durch eine Ebene E in $\mathbb{R}^3$ ist die Symmetrieabbildung $\sigma_E : \mathbb{R}^3 \to \mathbb{R}^3$ gemäß

$$\sigma_E(\mathbf{x}) = 2\mathbf{x}_E - \mathbf{x} \tag{3.27}$$

  definiert, wobei $\mathbf{x}_E$ den Fußpunkt des Lotes von $\mathbf{x}$ auf E bezeichnet.
- Das Gebiet Ω ist symmetrisch bezüglich E, d.h. $\sigma_E(\Omega) = \Omega$. Die Ebene E teilt Ω in die beiden Teilgebiete Ω_1 und Ω_2, deren gemeinsames Randstück wir mit γ bezeichnen.
- Die Anfangsbedingungen sind symmetrisch, d.h. $u_0(\mathbf{x}) = u_0(\sigma_E(\mathbf{x})),\ \forall \mathbf{x} \in \Omega$.

- Die Randbedingungen sind symmetrisch, d.h. es gilt

$$g_{\mathrm{Dir}}(\mathbf{x},t) = g_{\mathrm{Dir}}(\sigma_E(\mathbf{x}),t) \qquad \forall \mathbf{x} \in \Omega, \forall t \in]t_0,t_e[$$

  bzw.

$$g_{\mathrm{Neu}}(\mathbf{x},t) = g_{\mathrm{Neu}}(\sigma_E(\mathbf{x}),t) \qquad \forall \mathbf{x} \in \Omega, \forall t \in]t_0,t_e[.$$

- Der Operator Φ ist symmetrisch, d.h. es gilt

$$\Phi(v,\mathbf{x},t) = \Phi(v,\sigma_E(\mathbf{x}),t) \qquad \forall \mathbf{x} \in \Omega, \forall t \in]t_0,t_e[$$

  falls $v(\mathbf{x},t) = v(\sigma_E(\mathbf{x}),t), \forall \mathbf{x} \in \Omega, \forall t \in]t_0,t_e[$ ist. Das ist der Fall, wenn die Koeffizienten-Funktionen d, r und q symmetrisch sind.

3.3.3 Randbedingungen an der Symmetrielinie

Wenn ein symmetrisches Problem eine eindeutige Lösung u besitzt, so ist diese symmetrisch, d.h. für die Lösung von (3.24) gilt

$$u(\mathbf{x},t) = u(\sigma_g(\mathbf{x})),\ \forall \mathbf{x} \in \Omega, \forall t \in]t_0,t_e[$$

bzw. für die Lösung von (3.26)

$$u(\mathbf{x},t) = u(\sigma_E(\mathbf{x})),\ \forall \mathbf{x} \in \Omega, \forall t \in]t_0,t_e[.$$

Infolgedessen reicht es aus, die Lösung auf dem Teilgebiet Ω_1 zu berechnen und symmetrisch fortzusetzen. Die verbleibende Frage ist dann, welche Randbedingungen auf γ zu setzen sind.

Satz 3.21. *Es sei $\Omega \subset \mathbb{R}^2$ ein bezüglich der Gerade $g \subset \mathbb{R}^2$ symmetrisches Gebiet, d.h. es gelte $\sigma_g(\Omega) = \Omega$, wobei σ_g die in (3.25) definierte Symmetrieabbildung ist. Sei ferner $u : \Omega \to \mathbb{R}$ eine bezüglich der Geraden g symmetrische Funktion, d.h. $u(\mathbf{x}) = u(\sigma_g(\mathbf{x}))$ für alle $\mathbf{x} \in \Omega$. Sei u in einer Umgebung von $\gamma := \Omega \cap g$ differenzierbar. Die Gerade g teile Ω in die beiden Teilgebiete Ω_1 und Ω_2. Dann erfüllt die Einschränkung $u|_{\Omega_1}$ auf γ homogene Neumannsche Randbedingungen*

$$\nabla u \cdot \mathbf{n}_{\partial\Omega_1} = 0.$$

Beweis. Es sei $\mathbf{y} \in \gamma$. Wir bezeichnen mit $\mathbf{n}_{\partial\Omega_1}(\mathbf{y})$ den äußeren Normalvektor an den Rand von Ω_1 im Punkte $\mathbf{y}$. Da u in einer Umgebung von γ differenzierbar ist, gilt

$$\begin{aligned} \nabla u(\mathbf{y}) \cdot \mathbf{n}_{\partial\Omega_1}(\mathbf{y}) &= \lim_{\varepsilon \to 0+0} \frac{u(\mathbf{y} + \varepsilon \mathbf{n}_{\partial\Omega_1}(\mathbf{y})) - u(\mathbf{y} - \varepsilon \mathbf{n}_{\partial\Omega_1}(\mathbf{y}))}{2\varepsilon} \\ &= \lim_{\varepsilon \to 0+0} \frac{u(\sigma_g(\mathbf{y} - \varepsilon \mathbf{n}_{\partial\Omega_1}(\mathbf{y}))) - u(\mathbf{y} - \varepsilon \mathbf{n}_{\partial\Omega_1}(\mathbf{y}))}{2\varepsilon} = 0, \end{aligned}$$

weil u symmetrisch ist. □

Wenn Ω ein symmetrisches Gebiet auf einer Mannigfaltigkeit S ist, gilt ein analoger Satz:

Satz 3.22. *Es sei S eine zweidimensionale differenzierbare Mannigfaltigkeit in $\mathbb{R}^3$ und $\Omega \subset S$ ein bezüglich der Ebene $E \subset \mathbb{R}^3$ symmetrisches Gebiet, d.h. es gelte $\sigma_E(\Omega) = \Omega$, wobei σ_E die in (3.27) definierte Symmetrieabbildung ist. Sei ferner $u : \Omega \to \mathbb{R}$ eine bezüglich der Ebene E symmetrische Funktion, d.h. $u(\mathbf{x}) = u(\sigma_E(\mathbf{x}))$ für alle $\mathbf{x} \in \Omega$. Sei u in einer Umgebung von $\gamma := \Omega \cap E$ differenzierbar. Die Ebene E teile Ω in die beiden Teilgebiete Ω_1 und Ω_2. Dann erfüllt die Einschränkung $u|_{\Omega_1}$ auf γ homogene Neumannsche Randbedingungen*

$$\nabla u \cdot \mathbf{n}_{\partial\Omega_1} = 0.$$

Beweis. Der Satz 3.22 läßt sich analog zu Satz 3.21 beweisen. □

Wenn nun u die Lösung eines symmetrischen Anfangs-Randwert-Problems auf einem symmetrischen Gebiet Ω ist und Ω_1, Ω_2 und γ wie oben definiert sind, dann ist die Einschränkung $u|_{\Omega_1}$ die eindeutig bestimmte Lösung des entsprechenden Anfangs-Randwert-Problems auf Ω_1 mit homogenen Neumannschen Randbedingungen auf γ. Wird das symmetrische Problem in Ω auf einem symmetrischen Gitter diskretisiert, so erhält man das gleiche Gleichungssystem wie bei einer Diskretisierung in Ω_1 mit homogenen Neumannschen Randbedingungen auf γ. Wir wollen diese Tatsache am Beispiel der Wärmeleitungsgleichung erläutern, das Resultat läßt sich jedoch auch auf allgemeine Probleme übertragen.

Wir betrachten die Wärmeleitungsgleichung $u_t = \Delta u$ auf einem bezüglich einer Geraden g symmetrischen Gebiet Ω mit symmetrischen Anfangsbedingungen und homogenen Neumannschen Randbedingungen $\partial u/\partial \mathbf{n}_{\partial\Omega} = 0$ auf $\partial\Omega$. Wir triangulieren das Teilgebiet Ω_1 durch ein Dreiecksgitter $\mathcal{T}_1$, dessen Knoten wir mit $\mathbf{x}_i$ bezeichnen. Es sei $\mathcal{G}$ die Indexmenge der Knoten $\mathbf{x}_i$, die auf γ liegen und $\mathcal{R}_1$ die Indexmenge der übrigen Knoten in $\overline{\Omega_1}$. Wir wählen die Numerierung der Knoten derart, daß $\mathcal{R}_1 = \{1, \ldots, M\}$ und $\mathcal{G} = \{M+1, \ldots, N\}$ ist.

Spiegelt man die Triangulierung $\mathcal{T}_1$ an γ, so erhält man eine Triangulierung $\mathcal{T}_2$ von Ω_2. Die Gitterpunkte von $\mathcal{T}_2$ werden so numeriert, daß $\mathbf{x}_{N+k}$ gerade das Spiegelbild von $\mathbf{x}_k$, $k = 1, \ldots, M$ ist. Die Indexmenge der Knoten von $\mathcal{T}_2$, die in $\overline{\Omega_2}\backslash\gamma$ liegen, bezeichnen wir mit $\mathcal{R}_2 = \{N+1, \ldots, N+M\}$.

Wir betrachten zunächst das symmetrische Problem $u_t = \Delta u$ *auf ganz* Ω mit der Triangulierung $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2$. Die Diskretisierung mit linearen finiten Elementen führt auf ein System gewöhnlicher Differentialgleichungen

$$\mathbf{M}\mathbf{u}_t = -\mathbf{S}\mathbf{u}, \tag{3.28}$$

siehe (3.13). Dabei haben, wegen der Symmetrie des Gitters $\mathcal{T}$, Massen- und Steifigkeitsmatrix die Block-Gestalt

$$\mathbf{M} = \left(\begin{array}{c|c|c} \mathbf{M}_1 & \mathbf{M}_2 & \mathbf{0} \\ \hline \mathbf{M}_3 & \mathbf{M}_4 & \mathbf{M}_3 \\ \hline \mathbf{0} & \mathbf{M}_2 & \mathbf{M}_1 \end{array}\right) \begin{array}{l} \} \text{ Zeilen mit Indizes in } \mathcal{R}_1 \\ \} \text{ Zeilen mit Indizes in } \mathcal{G} \\ \} \text{ Zeilen mit Indizes in } \mathcal{R}_2 \end{array}$$

und

$$\mathbf{S} = \left(\begin{array}{c|c|c} \mathbf{S}_1 & \mathbf{S}_2 & \mathbf{0} \\ \hline \mathbf{S}_3 & \mathbf{S}_4 & \mathbf{S}_3 \\ \hline \mathbf{0} & \mathbf{S}_2 & \mathbf{S}_1 \end{array}\right) \begin{array}{l} \} \mathcal{R}_1 \\ \} \mathcal{G} \\ \} \mathcal{R}_2 \end{array} .$$

Da auch die Lösung u symmetrisch ist, so ist der Vektor $\mathbf{u}$ ebenfalls von der Form

$$\mathbf{u} = \left(\begin{array}{c} \mathbf{u}_1 \\ \hline \mathbf{u}_2 \\ \hline \mathbf{u}_1 \end{array}\right) \begin{array}{l} \} \, \mathcal{R}_1 \\ \} \, \mathcal{G} \\ \} \, \mathcal{R}_2 \end{array} .$$

Folglich reduziert sich (3.28) auf das System gewöhnlicher Differentialgleichungen

$$\left(\begin{array}{c|c} \mathbf{M}_1 & \mathbf{M}_2 \\ \hline 2\mathbf{M}_3 & \mathbf{M}_4 \end{array}\right) \frac{d}{dt} \left(\begin{array}{c} \mathbf{u}_1 \\ \hline \mathbf{u}_2 \end{array}\right) = - \left(\begin{array}{c|c} \mathbf{S}_1 & \mathbf{S}_2 \\ \hline 2\mathbf{S}_3 & \mathbf{S}_4 \end{array}\right) \left(\begin{array}{c} \mathbf{u}_1 \\ \hline \mathbf{u}_2 \end{array}\right), \tag{3.29}$$

welches gelöst werden muß.

Wir betrachten nun das Problem $u_t = \Delta u$ nur auf dem Teilgebiet Ω_1, versehen mit der Triangulierung $\mathcal{T}_1$, und setzen auf γ homogene Neumann-Randbedingungen

$$\partial u / \partial \mathbf{n}_{\partial\Omega_1} = 0 \tag{3.30}$$

voraus. Dann ergibt sich nach der Diskretisierung durch lineare finite Elemente das System

$$\widetilde{\mathbf{M}} \widetilde{\mathbf{u}}_t = -\widetilde{\mathbf{S}} \widetilde{\mathbf{u}}. \tag{3.31}$$

In $\widetilde{\mathbf{M}}$, $\widetilde{\mathbf{S}}$ und $\widetilde{\mathbf{u}}$ existieren Blöcke, die zu $\mathcal{R}_1$ und solche, die zu $\mathcal{G}$ gehören:

$$\widetilde{\mathbf{M}} = \left(\begin{array}{c|c} \widetilde{\mathbf{M}}_1 & \widetilde{\mathbf{M}}_2 \\ \hline \widetilde{\mathbf{M}}_3 & \widetilde{\mathbf{M}}_4 \end{array}\right) \begin{array}{l} \} \, \mathcal{R}_1 \\ \} \, \mathcal{G} \end{array} , \qquad \widetilde{\mathbf{S}} = \left(\begin{array}{c|c} \widetilde{\mathbf{S}}_1 & \widetilde{\mathbf{S}}_2 \\ \hline \widetilde{\mathbf{S}}_3 & \widetilde{\mathbf{S}}_4 \end{array}\right) \begin{array}{l} \} \, \mathcal{R}_1 \\ \} \, \mathcal{G} \end{array} , \qquad \widetilde{\mathbf{u}} = \left(\begin{array}{c} \widetilde{\mathbf{u}}_1 \\ \hline \widetilde{\mathbf{u}}_2 \end{array}\right) \begin{array}{l} \} \, \mathcal{R}_1 \\ \} \, \mathcal{G} \end{array} .$$

Für die Elemente von $\mathbf{M} = (m_{ij})$ und $\widetilde{\mathbf{M}} = (\widetilde{m}_{ij})$ gilt

$$m_{ij} = \int_{\Omega} \varphi_i \; \varphi_j \; d\mathbf{x} = \int_{\Omega_1} \varphi_i \; \varphi_j \; d\mathbf{x} = \widetilde{m}_{ij}$$

für alle i, j mit ($i \in \mathcal{R}_1$ und $j \in \mathcal{G}$) oder ($i \in \mathcal{G}$ und $j \in \mathcal{R}_1$). Aus der Symmetrie des Gitters $\mathcal{T}$ folgt

$$m_{ij} = \int_{\Omega} \varphi_i \; \varphi_j \; d\mathbf{x} = 2 \int_{\Omega_1} \varphi_i \; \varphi_j \; d\mathbf{x} = 2\widetilde{m}_{ij} \qquad \forall i, j \in \mathcal{G}.$$

Deshalb gilt $\mathbf{M}_1 = \widetilde{\mathbf{M}}_1$, $\mathbf{M}_2 = \widetilde{\mathbf{M}}_2$, $\mathbf{M}_3 = \widetilde{\mathbf{M}}_3$, $\mathbf{M}_4 = 2\widetilde{\mathbf{M}}_4$. In analoger Weise zeigt man für die Steifigkeits-Untermatrizen $\mathbf{S}_1 = \widetilde{\mathbf{S}}_1$, $\mathbf{S}_2 = \widetilde{\mathbf{S}}_2$, $\mathbf{S}_3 = \widetilde{\mathbf{S}}_3$, $\mathbf{S}_4 = 2\widetilde{\mathbf{S}}_4$. Damit hat das System (3.31) die Form

$$\left(\begin{array}{c|c} \mathbf{M}_1 & \mathbf{M}_2 \\ \hline \mathbf{M}_3 & \frac{1}{2}\mathbf{M}_4 \end{array}\right) \frac{d}{dt} \left(\begin{array}{c} \mathbf{u}_1 \\ \hline \mathbf{u}_2 \end{array}\right) = - \left(\begin{array}{c|c} \mathbf{S}_1 & \mathbf{S}_2 \\ \hline \mathbf{S}_3 & \frac{1}{2}\mathbf{S}_4 \end{array}\right) \left(\begin{array}{c} \mathbf{u}_1 \\ \hline \mathbf{u}_2 \end{array}\right).$$

Dieses System ist jedoch äquivalent zu (3.29).

Damit haben wir gezeigt, daß die in (3.30) vorausgesetzten homogenen Neumannschen Randbedingungen gerade der symmetrischen Fortsetzung der Lösung entsprechen. Dieser Sachverhalt, der hier nur am Beispiel der Wärmeleitungsgleichung dargestellt wurde, läßt sich unschwer auch bei allgemeineren symmetrischen Problemen einsehen.

3.4 Ortsdiskretisierung semilinearer Reaktions-Diffusions-Systeme

Ist ein System semilinearer Reaktions-Diffusions-Gleichungen der Form

$$\begin{aligned}\frac{\partial u_i}{\partial t}(\mathbf{x},t) &= \nabla\cdot(d_i(\mathbf{x})\nabla u_i(\mathbf{x},t)) + r_i(\mathbf{x})u_i(\mathbf{x},t)\\ &\quad + p_i(u_1(\mathbf{x},t),\ldots,u_m(\mathbf{x},t)) + q_i(\mathbf{x},t), \qquad i=1,\ldots,m\\ u_i(\mathbf{x},t_0) &= u_{i,0}(\mathbf{x})\end{aligned}$$

gegeben, siehe (2.2), so diskretisiert man komponentenweise. Wir erläutern das kurz an einem Beispiel.

Beispiel 3.23. Es werden Neumannsche Randbedingungen

$$d_i(\mathbf{x},t)\nabla u_i\cdot\mathbf{n}_{\partial\Omega} = g_{\mathrm{Neu},i}(\mathbf{x},t), \qquad \mathbf{x}\in\partial\Omega, \quad i=1,\ldots,m$$

angenommen und die Reduktion der Massenmatrix durchgeführt, siehe Abschnitt 3.1.5. Im Ergebnis erhält man die m Systeme

$$\begin{aligned}\frac{\partial \mathbf{u}_i}{\partial t} &= -\mathbf{L}^{-1}\mathbf{S}_i\mathbf{u}_i + \mathbf{R}_i\mathbf{u}_i + p_i(\mathbf{u}_1,\ldots,\mathbf{u}_m) + \mathbf{q}_i + \mathbf{L}^{-1}\mathbf{B}\mathbf{g}_{\mathrm{Neu},i} \qquad (3.32)\\ &=: \mathbf{f}_i(\mathbf{u}_1,\ldots,\mathbf{u}_m), \qquad i=1,\ldots,m,\end{aligned}$$

siehe (3.15). Dabei sind $\mathbf{u}_i(t)$ die diskreten Näherungswerte von $u_i(\mathbf{x},t)$. Die Matrizen und Vektoren $\mathbf{S}_i$, $\mathbf{R}_i$ und $\mathbf{q}_i$ sind wie in (3.8) definiert, wobei die dort auftretenden Koeffizienten $d(\mathbf{x})$, $r(\mathbf{x})$ und $q(\mathbf{x})$ jeweils durch $d_i(\mathbf{x})$, $r_i(\mathbf{x})$ und $q_i(\mathbf{x})$ zu ersetzen sind. Die Funktionen p_i werden wieder komponentenweise verstanden. Die in (3.8) definierte Matrix $\mathbf{L}$ und die in (3.14) definierte Matrix $\mathbf{B}$ hängen nur von der Triangulierung ab und sind daher für alle Komponenten gleich. Der Vektor $\mathbf{g}_{\mathrm{Neu,i}}$ ist wie in (3.11) definiert, wobei dort die Funktion g_{Neu} durch $g_{\mathrm{Neu,i}}$ ersetzt werden muß.

Setzt man $\mathbf{u} := (\mathbf{u}_1,\ldots,\mathbf{u}_m)$ und $\mathbf{f}(\mathbf{u}) := (\mathbf{f}_1(\mathbf{u}_1,\ldots,\mathbf{u}_m),\ldots,\mathbf{f}_m(\mathbf{u}_1,\ldots,\mathbf{u}_m))$, so erhält man ein System gewöhnlicher Differentialgleichungen

$$\frac{\partial \mathbf{u}}{\partial t} = \mathbf{f}(\mathbf{u}), \qquad (3.33)$$

welches durch eines der in den Kapiteln 5, 7 und 8 beschriebenen Verfahren gelöst werden kann. □

Kapitel 4

Gitteradaption

Bei vielen parabolischen Problemen ist der räumliche Dikretisierungsfehler bei der Verwendung eines uniformen Gitters lokal von sehr unterschiedlicher Größenordnung. Insbesondere in der Nähe von Fronten treten Fehler auf, die um ein Vielfaches höher sind als in Gebieten, in denen sich die Lösung in einer Gleichgewichtslage befindet. Uniforme Gitter benötigen deshalb eine extrem große Anzahl von Elementen, um den globalen Fehler unter einer gewünschten Toleranz zu halten. Einen Ausweg aus diesem Problem liefert die adaptive Gitterverfeinerung, bei der Elemente unterschiedlicher Größenordnungen zugelassen werden. Häufig wünscht man eine näherungsweise Gleichverteilung des räumlichen Diskretisierungsfehlers über das Gebiet Ω. Auf diese Weise werden Fronten wesentlich feiner aufgelöst als Teile der Lösung mit einem geringeren Gradienten. Benötigt wird dafür eine lokale Abschätzung des räumlichen Diskretisierungsfehlers. Hierfür werden in der Regel **a-posteriori-Fehlerschätzer** verwendet, d.h. solche, die eine Information über den Fehler direkt aus der numerischen Näherungslösung beziehen.

4.1 Räumliche a-posteriori-Fehlerschätzung

Grundsätzlich unterscheidet man zwischen den eigentlichen Fehler*schätzern* und sogenannten Fehler*indikatoren*. Fehlerschätzer geben eine untere und obere Schranke des Fehlers an, während Indikatoren Hinweise auf lokale Fehler geben. Wir betrachten hier die Reaktions-Diffusions-Gleichung

$$\begin{aligned} \frac{\partial u}{\partial t}(\mathbf{x},t) &= \operatorname{div}(d(\mathbf{x})\nabla u(\mathbf{x},t)) + r(\mathbf{x})u(\mathbf{x},t) \\ &\quad + p(u(\mathbf{x},t)) + q(\mathbf{x},t), && \mathbf{x}\in\Omega, \quad t\in[t_0,t_e], \\ u(\mathbf{x},t_0) &= u_0(\mathbf{x}), && \mathbf{x}\in\Omega \end{aligned} \tag{4.1}$$

mit der Lösung $u(\mathbf{x},t)$. Die Menge Ω kann ein Gebiet in $\mathbb{R}^n$ oder ein Gebiet auf einer glatten Fläche $S\subset\mathbb{R}^n$ sein. Im letzteren Falle muß in diesem Kapitel stets div durch div_S und ∇ durch ∇_S ersetzt werden. Die Diskretisierung erfolge mit der Linienmethode. Dabei werde das Zeitintervall $[t_0,t_e]$ gemäß $t_0<t_1<\cdots<t_{N-1}<t_N=t_e$ unterteilt. In jedem Zeitpunkt t_i existiert eine Triangulierung $\mathcal{T}_h$. Es sei V_h der in Definition 3.2 eingeführte Raum stetiger

und stückweise linearer Funktionen auf Ω bezüglich der Triangulierung $\mathcal{T}_h$.[1] Die Funktion $u_h^\tau : \Omega \times [t_0, t_e] \to \mathbb{R}$ sei eine Näherungslösung; es gelte $u_h^\tau(\cdot, t_i) \in V_h,\ i = 0, \ldots, N$. Es sei $\widetilde{u}_i$ die exakte Lösung der Differentialgleichung (4.1) *mit der Anfangsbedingung* $\widetilde{u}_i(\mathbf{x}, t_i) = u_h^\tau(\mathbf{x}, t_i)$. Ein **globaler räumlicher a-posteriori-Fehlerschätzer** ist eine Funktion $\epsilon : V_h \to \mathbb{R}_+$ mit den folgenden Eigenschaften:

(F1) **Zuverlässigkeit**: Es gelte

$$\|\widetilde{u}_{i-1}(\cdot, t_i) - u_h^\tau(\cdot, t_i)\|_\Omega \le C_1 \epsilon(u_h^\tau(\cdot, t_i))$$

für alle $i = 1, \ldots, N$.

(F2) **Effizienz**: Es gelte

$$\epsilon(u_h^\tau(\cdot, t_i)) \le C_2 \|\widetilde{u}_{i-1}(\cdot, t_i) - u_h^\tau(\cdot, t_i)\|_\Omega$$

für alle $i = 1, \ldots, N$.

Dabei muß in (F1) und (F2) *die gleiche Norm* $\|\cdot\|_\Omega$ verwendet werden. Der Fehlerschätzer heißt **asymptotisch exakt**, wenn für ein beliebiges $\varepsilon > 0$ bei hinreichend starker Gitterverfeinerung $|C_1 - 1| < \varepsilon$ und $|C_2 - 1| < \varepsilon$ ist.

Ein **lokaler räumlicher Fehlerindikator** ist eine Funktion $\eta : V_h \times \mathcal{T}_h \to \mathbb{R}_+$, die eine Ungleichung der Form

$$\eta(u_h^\tau(\cdot, t_i), T) \le C_3 \|\widetilde{u}_{i-1}(\cdot, t_i) - u_h^\tau(\cdot, t_i)\|_{U(T)}$$

für alle $i = 1, \ldots, N$ und alle $T \in \mathcal{T}_h$ erfüllt, wobei $U(T)$ eine gewisse lokal begrenzte Umgebung des Dreiecks T ist. In der Praxis sind besonders die Fehlerschätzer interessant, die eine Zerlegung in lokale Fehlerindikatoren erlauben, etwa in der Form

$$\epsilon(u_h^\tau(\cdot, t_i))^2 = \sum_{T \in \mathcal{T}_h} \eta(u_h^\tau(\cdot, t_i), T)^2. \tag{4.2}$$

Für elliptische Probleme gibt es mittlerweile eine Vielzahl von Arbeiten, die sich mit a-posteriori-Fehlerschätzung befassen. Erste Untersuchungen zu Fehlerschätzern stammen von Babuška und Rheinboldt [12] aus dem Jahre 1978, später kamen u.a. Arbeiten von Bank und Weiser [16] (1985), Eriksson und Johnson [57] (1988) hinzu, siehe auch die Übersicht von Verfürth [165] (1996).

Die Analysis von Fehlerschätzern bei parabolischen Problemen hängt wesentlich davon ab, welche Diskretisierungsmethode betrachtet wird. Fehlerschätzer im Kontext der Linienmethode wurden beispielsweise von Bieterman und Babuška [23] (1986), Adjerid und Flaherty [3] (1988), Moore [120] (1994) und Babuška, Feistauer und Šolín [11] (2001) entwickelt und untersucht. Bornemann [26] (1992), Lang und Walter [104] (1992) und Lang [102] (1998) analysierten Fehlerschätzer für die Rothe-Methode. Eine weitere Möglichkeit besteht in der Verwendung von finiten Raum-Zeit-Elementen, ein Weg, der etwa von Eriksson und Johnson [58] (1995) sowie Verfürth [164] (1998) eingeschlagen wurde. Unter den Fehlerindikatoren ist vor allem der 1987 von Zienkiewicz und Zhu [177] vorgestellte sogenannte

[1] Die Triangulierung $\mathcal{T}_h$ und der Raum V_h hängen natürlich von dem aktuellen Zeitpunkt t_i ab. Der Einfachheit halber verzichten wir hier jedoch auf einen Index i.

Z^2-Indikator zu nennen, der sowohl für elliptische als auch für parabolische Probleme eingesetzt werden kann, siehe etwa PAPASTAVROU [128].

Zu den Vorteilen des Z^2-Indikators zählt es, daß er unabhängig vom Modellproblem formuliert werden kann und einfach zu implementieren ist, während die oben erwähnten Fehlerschätzer in der Regel Bedingungen an die Differentialgleichung oder an die verwendeten Verfahren stellen, die für viele praktische Probleme nicht erfüllt sind. Für die stationäre lineare Reaktions-Diffusions-Gleichung $-\Delta u + \alpha u = f(x)$ wurde von VERFÜRTH [165] 1996 gezeigt, daß der Z^2-Indikator bei der entsprechend (4.2) vorgenommenen quadratischen Aufsummierung sogar einen Fehlerschätzer liefert.

Wir werden in den in dieser Arbeit dargestellten numerischen Rechnungen den Z^2-Indikator zur Gitteradaption verwenden. Deshalb soll dieser Fehlerindikator im folgenden näher beschrieben werden. Wir orientieren uns dabei stark an der Darstellung von VERFÜRTH [165].

4.1.1 Der Z^2-Fehlerindikator

Wie oben seien $t_i,\ i = 0, \ldots, N$, $\mathcal{T}_h$, V_h, u_h^τ und $\widetilde{u}_i,\ i = 0, \ldots, N$ gegeben. Wir bezeichnen mit W_h den Raum der auf $\mathcal{T}_h$ stückweise linearen, aber nicht notwendigerweise stetigen Funktionen. Offenbar ist dann $\nabla u_h^\tau \in W_h$. Der Raum V_h ist ein Untervektorraum von W_h. Sei $T \in \mathcal{T}_h$ ein Dreieck mit den Eckpunkten $\mathbf{x}_1,\ \mathbf{x}_2$ und $\mathbf{x}_3$. Für eine Funktion $\varphi \in W_h$ setzen wir $\varphi|_T$ stetig auf ∂T fort und definieren auf diese Weise $\varphi|_T(\mathbf{x}_i),\ i = 1, 2, 3$ als den entsprechenden Funktionswert der Fortsetzung. Die Abbildung $(\cdot,\cdot)_h : W_h \times W_h \to \mathbb{R}$, definiert durch

$$(\varphi, \psi)_h := \sum_{T \in \mathcal{T}_h} \frac{|T|}{3} \left(\sum_{i=1}^{3} \varphi|_T(\mathbf{x}_i)\ \psi|_T(\mathbf{x}_i) \right)$$

ist dann ein Skalarprodunkt auf W_h. Wir bezeichnen das Bild der Orthogonalprojektion von ∇u_h^τ auf V_h mit $G(u_h^\tau)$, d.h. es gelte

$$(G(u_h^\tau), \varphi)_h = (\nabla u_h^\tau, \varphi)_h$$

für alle $\varphi \in V_h$. Die Größe $G(u_h^\tau)$ ist demnach ein „geglätteter Gradient“ der Näherungslösung.

Wenn die Lösung $\widetilde{u}_{i-1}(\cdot, t_i)$ hinreichend glatt ist und $u_h^\tau(\cdot, t_i)$ eine ausreichend gute Näherung an $\widetilde{u}_{i-1}(\cdot, t_i)$ darstellt, dann kann man davon ausgehen, daß die stetige Funktion $G(u_h^\tau(\cdot, t_i))$ den Gradienten $\nabla\widetilde{u}_{i-1}(\cdot, t_i)$ besser approximiert als die stückweise konstante Funktion
$\nabla u_h^\tau(\cdot, t_i)$, daß also eine Ungleichung der Form

$$\|\nabla\widetilde{u}_{i-1}(\cdot, t_i) - G(u_h^\tau(\cdot, t_i))\|_{L^2(T)} \le \alpha\|\nabla\widetilde{u}_{i-1}(\cdot, t_i) - \nabla u_h^\tau(\cdot, t_i)\|_{L^2(T)}, \qquad \alpha < 1 \tag{4.3}$$

gilt. Es folgt

$$\begin{aligned} \frac{1}{1+\alpha}\|\nabla u_h^\tau(\cdot, t_i) - G(u_h^\tau(\cdot, t_i))\|_{L^2(T)} &\le \|\nabla\widetilde{u}_{i-1}(\cdot, t_i) - \nabla u_h^\tau(\cdot, t_i)\|_{L^2(T)} \\ &\le \frac{1}{1-\alpha}\|\nabla u_h^\tau(\cdot, t_i) - G(u_h^\tau(\cdot, t_i))\|_{L^2(T)}. \end{aligned} \tag{4.4}$$

Der Einfachheit halber schreiben wir im folgenden für $u_h^\tau(\cdot, t_i)$ nur u_h^τ. Wie aus (4.4) folgt, erfüllt die Größe

$$\eta^Z(u_h^\tau, T) := \|\nabla u_h^\tau - G(u_h^\tau)\|_{L^2(T)}$$

die Funktion eines lokalen Fehlerindikators, der die L^2-Norm des Fehlergradienten abschätzt.

Man beachte insbesondere, daß der Z^2-Indikator – wie in der Definition eines Fehlerindikators verlangt – den *zeitlich lokalen* Fehler $\|\nabla \widetilde{u}_{i-1}(\cdot, t_i) - \nabla u_h^\tau(\cdot, t_i)\|_{L^2(T)}$ abschätzt und *nicht* etwa den zeitlich globalen Fehler $\|\nabla u(\cdot, t_i) - \nabla u_h^\tau(\cdot, t_i)\|_{L^2(T)}$. Das liegt daran, daß der Gradient der Näherungslösung ∇u_h^τ nur $\nabla \widetilde{u}_{i-1}$, nicht aber ∇u hinreichend gut approximiert. Daher ist auch der geglättete Gradient $G(u_h^\tau)$ in der Regel nur eine gute Approximation an $\nabla \widetilde{u}_{i-1}$, nicht aber an ∇u. Abbildung 4.1 verdeutlicht diesen Sachverhalt.

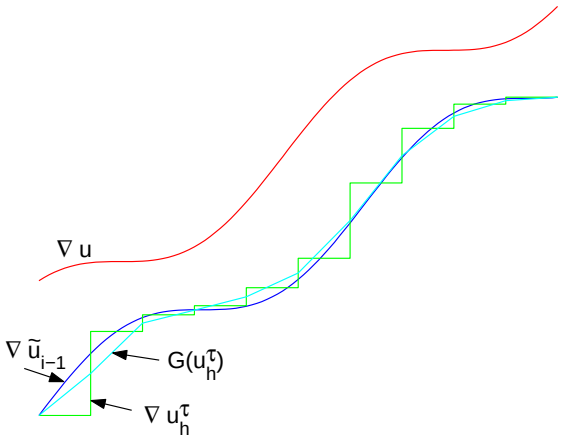


Abbildung 4.1: Approximation von $\nabla \widetilde{u}_{i-1}$ durch ∇u_h^τ und $G(u_h^\tau)$ zum Zeitpunkt t_i. Offenbar ist $G(u_h^\tau)$ die bessere Approximation, d.h. es gilt die Ungleichung (4.3).

Die Größe $G(u_h^\tau)$ kann in einfacher Weise aus u_h^τ gewonnen werden, wie der folgende Satz zeigt:

Satz 4.1. *Es sei ω_i die Vereinigung aller Dreiecke der Triangulierung $\mathcal{T}_h$, die $\mathbf{x}_i$ als Eckpunkt enthalten. Der Wert $\nabla u_h^\tau|_T(\mathbf{x}_i)$ sei wie oben als Funktionswert der stetigen Fortsetzung von $\nabla u_h^\tau|_T$ auf ∂T definiert. Dann ist $G(u_h^\tau)$ gerade durch die Beziehung*

$$G(u_h^\tau) \in V_h, \qquad G(u_h^\tau)(\mathbf{x}_i) = \frac{1}{|\omega_i|} \sum_{T \subset \omega_i} |T| \ \nabla u_h^\tau|_T(\mathbf{x}_i)$$

gegeben.

Beweis. Zum Beweis des Satzes sei auf das Buch von Verfürth [165] verwiesen. □

Der Vektor $G(u_h^\tau)(\mathbf{x}_i)$ ist demnach das gewichtete Mittel der Gradienten von u_h^τ in ω_i.

4.2 Verfeinerung und Vergröberung des Gitters

Liegt das Resultat des Fehlerschätzers vor, so kann das Dreiecksgitter entsprechend angepaßt werden, ein Vorgang der als **Gitteradaption** bezeichnet wird. Unter Auswertung des Fehler-

schätzers markiert man alle Elemente, die zur Verfeinerung vorgesehen sind und ebenso alle, die vergröbert werden sollen. Die von uns verwendete Markierungsstrategie wird in Abschnitt 4.3 beschrieben. Weitere Strategien sind etwa in VERFÜRTH [165] angegeben. Für die hier vorgestellten numerischen Berechnungen wurde der in dem Programmpaket UG [20] integrierte Gittergenerator benutzt. Dieser Gittergenerator basiert auf einem von BANK, SHERMAN und WEISER [15] 1983 entwickelten Algorithmus zur Erzeugung adaptiver Gitter, der auch in dem Programm PLTMG zur Lösung elliptischer Differentialgleichungen von BANK [14] verwendet wird. Für eine ausführlichere Beschreibung der Gitteradaption sei auf BASTIAN et al. [20] verwiesen.

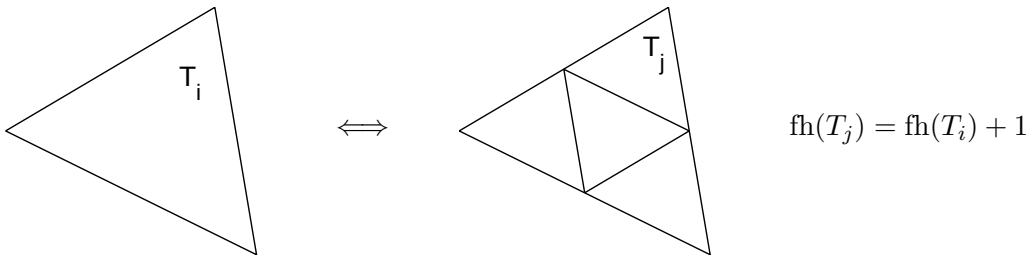
Zu Beginn der numerischen Berechnung, also zum Zeitpunkt $t = t_0$, erzeugt man ein möglichst uniformes Grundgitter $\mathcal{T}_{h,0}$. Durch sukzessive Verfeinerung bzw. Vergröberung von Elementen entstehen daraus weitere Gitter $\mathcal{T}_{h,1}, \mathcal{T}_{h,2}, \ldots$. Im i-ten Adaptionsschritt wird das Gitter $\mathcal{T}_{h,i-1}$ in das Gitter $\mathcal{T}_{h,i}$ überführt. Eine Verfeinerung eines Dreiecks bedeutet, daß das Dreieck (das sogenannte „Vater"-Element) durch zwei oder vier kleinere Dreiecke (die „Sohn"-Elemente) ersetzt wird. Man unterscheidet **reguläre** und **irreguläre Verfeinerung** eines Dreiecks. Bei regulärer Verfeinerung entstehen aus einem Vater-Element vier zu diesem kongruente Sohn-Elemente; bei irregulärer Verfeinerung entstehen zwei Sohn-Elemente, die nicht zum Vater-Element kongruent sind. Bei Vergröberung werden Sohn-Elemente wieder durch ihr Vater-Element ersetzt. Vergröberung ist bei dem hier verwandten Algorithmus daher nur möglich, wenn alle Sohn-Elemente eines Vater-Elements zur Vergröberung markiert worden sind. Durch irreguläre Verfeinerung erzeugte Elemente dürfen nicht weiter verfeinert werden. Wird ein solches Element zur Verfeinerung markiert, dann muß es zunächst vergröbert und anschließend regulär verfeinert werden. Außerdem ist die Gitteradaption so durchzuführen, daß keine hängenden Knoten entstehen und alle zur Verfeinerung markierten Elemente auch tatsächlich verfeinert werden. Um das zu gewährleisten, werden meist auch einige Elemente verfeinert, die nicht dafür markiert worden sind.

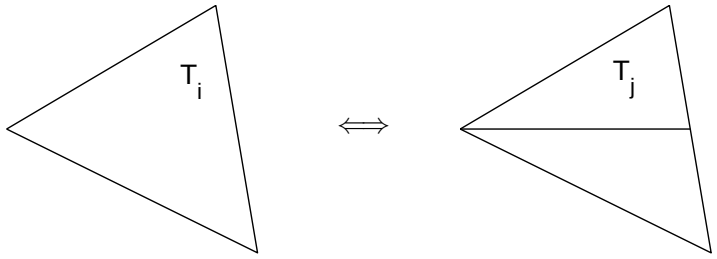
Wenn T_1 ein Sohn-Element von T_0 ist, so schreiben wir $T_1 = S(T_0)$ und $T_0 = V(T_1)$. Die Funktion $\mathrm{fh} : \bigcup_{i=1,2,\ldots} \mathcal{T}_{h,i} \to \mathbb{N}_0$ soll in folgender Weise jedem Dreieck eine natürliche Zahl zuordnen:

- $\mathrm{fh}(T) = 0$ für alle $T \in \mathcal{T}_{h,0}$,
- $\mathrm{fh}(S(T)) = \mathrm{fh}(T) + 1$.

Der Funktionswert $\mathrm{fh}(T)$ wird als **Feinheit** des Elements T bezeichnet; er gibt an, wieviele Verfeinerungsschritte nötig wären, um ein Dreieck T *direkt* aus der Ausgangstriangulierung $\mathcal{T}_{h,0}$ zu erzeugen.

In der folgenden Skizze sind reguläre und irreguläre Verfeinerung eines Dreiecks dargestellt.





$$\Longleftrightarrow \qquad \mathrm{fh}(T_j) = \mathrm{fh}(T_i) + 1$$

Wir bezeichnen die nach dem i-ten Adaptionsschritt vorliegende Triangulierung $\mathcal{T}_{h,i}$ im im folgenden nur noch mit $\mathcal{T}_h$. Die Dreiecke aus $\mathcal{T}_h$, die durch ausschließlich reguläre Verfeinerung aus dem Grundgitter hervorgehen, fassen wir in der Menge $\mathcal{T}_h^{\mathrm{reg}}$ zusammen. Startet man mit einem uniformen Gitter $\mathcal{T}_{h,0}$ auf dem Gebiet Ω, d.h. mit einem Gitter, dessen Dreiecke näherungsweise gleichlange Seiten der Länge h_0 haben, so gilt in guter Näherung

$$h_{\min}(T) \approx 2^{-\mathrm{fh}(T)} h_0 \quad \forall T \in \mathcal{T}_h \qquad \text{und} \qquad |T| \approx 2^{-2\ \mathrm{fh}(T)} \frac{|\Omega|}{|\mathcal{T}_{h,0}|} \quad \forall T \in \mathcal{T}_h^{\mathrm{reg}}. \tag{4.5}$$

Dabei ist $h_{\min}$ die Länge der kleinsten Dreiecksseite, $|T|$ die Fläche des Dreiecks T und $|\mathcal{T}_{h,0}|$ die Anzahl der Dreiecke in $\mathcal{T}_{h,0}$. Enthält das Grundgitter nur Dreiecke gleichen Flächeninhalts, so gilt die rechte Näherung sogar exakt:

$$|T| = 2^{-2\ \mathrm{fh}(T)} \frac{|\Omega|}{|\mathcal{T}_{h,0}|} \quad \forall T \in \mathcal{T}_h^{\mathrm{reg}}. \tag{4.6}$$

Nach einer Verfeinerung werden die Lösungswerte in den neu entstandenen Knoten durch lineare Interpolation gewonnen.

4.3 Steuerung der Gitterstruktur

4.3.1 Die Zielfeinheits-Funktion

Die verbleibende Frage ist nun, welche Elemente zur Verfeinerung bzw. Vergröberung ausgewählt werden sollen. Mitunter strebt man durch die Gitteradaption eine Gleichverteilung des Fehlers auf die einzelnen Dreiecke der Triangulierung $\mathcal{T}_h$ an. Das ist jedoch nicht die einzige – und nicht immer die effizienteste – Möglichkeit. Mitunter möchte man beispielsweise eine besonders sensible Front stärker auflösen, während man sich abseits der Front größere Fehler erlauben kann. Bei manchen Problemen will man hingegen eine bewegte Front nicht bis zur Gleichverteilung des Fehlers verfeinern, etwa um nicht zu oft eine Adaption vornehmen zu müssen. In Abschnitt 4.4.3 stellen wir ein numerisches Beispiel vor, bei dem eine Gleichverteilung des Fehlers *nicht* die effizienteste Möglichkeit zur Lösung darstellt.

In diesem Abschnitt wird ein Algorithmus vorgestellt, der eine weitreichende Einflußnahme des Nutzers auf die Struktur des Gitters erlaubt. Durch Variation gewisser Parameter können verschiedene Abstufungen zwischen einem uniformen und einem lokal stark verfeinerten Gitter erzeugt werden. Zunächst benötigen wir einen Verfeinerungs-Indikator φ. Dieser wird aus dem Fehlerindikator η gewonnen. Ein Fehlerindikator, der – wie etwa der Z^2-Indikator – die L^2-Norm des Gradienten des Fehlers auf einem Dreieck T abschätzt, wird bei einer Verfeinerung

des Dreiecks *kleiner*. Ist beispielsweise zum Zeitpunkt t_i der Fehler $\widetilde{u}_{i-1}(\cdot,t_i) - u_h^\tau(\cdot,t_i)$ auf dem Dreieck T eine lineare Funktion und $S(T)$ ein durch reguläre Verfeinerung erzeugtes Sohn-Element von T, so gilt

$$\|\nabla\widetilde{u}_{i-1}(\cdot,t_i) - \nabla u_h^\tau(\cdot,t_i)\|_{L^2(S(T))} = \frac{1}{2}\|\nabla\widetilde{u}_{i-1}(\cdot,t_i) - \nabla u_h^\tau(\cdot,t_i)\|_{L^2(T)},$$

also wird auch für den Fehlerindikator η näherungsweise

$$\eta(u_h^\tau, S(T)) \approx \frac{1}{2}\eta(u_h^\tau, T)$$

gelten. Der Verfeinerungs-Indikator φ soll jedoch bei einer Verfeinerung des Dreiecks T *näherungsweise konstant* bleiben. Wir wählen daher

$$\varphi(u_h^\tau, T) := \frac{\eta(u_h^\tau, T)}{\sqrt{|T|}} \tag{4.7}$$

als Verfeinerungs-Indikator.

Unser Ziel ist es, eine Beziehung zwischen dem Verfeinerungs-Indikator φ und der benötigten Gitterfeinheit herzustellen. Dazu wollen wir eine Funktion $\mathrm{zfh} : \mathbb{R}_+ \to \mathbb{N}_0$ vorgeben, die dem Wert des Verfeinerungs-Indikators die gewünschte Gitterfeinheit zuordnet. Die Funktion zfh wird als **Zielfeinheits-Funktion** bezeichnet. Das Ziel der Verfeinerung ist erreicht, wenn die Beziehung

$$\mathrm{zfh}(\varphi(u_h^\tau, T)) = \mathrm{fh}(T) \qquad \forall T \in \mathcal{T}_h \tag{4.8}$$

gilt. Wir bezeichnen eine Triangulierung $\mathcal{T}_h$, für die diese Beziehung gilt, als bezüglich der Zielfeinheits-Funktion zfh **ideale Triangulierung**.

Bei vielen praktischen Problemen existiert jedoch keine ideale Triangulierung, nämlich dann, wenn die in Abschnitt 3.1.2 aufgeführten Eigenschaften einer Triangulierung mit der Forderung (4.8) nicht vereinbar sind. Das ist der Fall, wenn (4.8) die unmittelbare Nachbarschaft sehr großer und sehr kleiner Dreiecke vorschreibt. Man kann dann lediglich eine Triangulierung finden, die der Beziehung (4.8) möglichst nahe kommt. Ein in Abschnitt 4.4.1 betrachtetes numerisches Beispiel verdeutlicht diesen Sachverhalt.

Die Zielfeinheits-Funktion zfh ist eine monoton wachsende Treppenfunktion, die durch Vorgabe ihrer Sprungstellen vollständig festgelegt werden soll. Wir bezeichnen die Sprungstellen mit $\mu(k)$. Es gelte $0 \le \mu(0) \le \cdots \le \mu(M-1)$. Die Zielfeinheits-Funktion ist dann durch

$$\mathrm{zfh}(\varphi) = \begin{cases} 0, & 0 \le \varphi < \mu(0), \\ k, & \mu(k-1) \le \varphi < \mu(k),\ k = 1, \ldots, M-1, \\ M, & \mu(M-1) < \varphi \end{cases} \tag{4.9}$$

gegeben. Die Zahl $M \in \mathbb{N}_0$ ist die maximale Feinheit des Gitters. Verschiedene Möglichkeiten zur Wahl der Zielfeinheits-Funktion werden anhand eines numerischen Beispiels in Abschnitt 4.4.1 diskutiert.

4.3.2 Algorithmen zur Gitteradaption

Die Gitteradaption erfolgt nach dem folgenden Algorithmus:

Algorithmus 4.2 (Gitteradaption).

gegeben: Zielfeinheits-Funktion zfh durch die Sprungstellen $\mu(0), \ldots, \mu(M-1)$

for $T \in \mathcal{T}_h$

if $\varphi(u_h^\tau, T) > \mu(\mathrm{fh}(T))$

markiere T zur Verfeinerung

end

if $\mathrm{fh}(T) > 0$ and $\varphi(u_h^\tau, T) < \mu(\mathrm{fh}(T) - 1)$

markiere T zur Vergröberung

end

end

□

Da wir diesen Algorithmus im Rahmen der Diskretisierung parabolischer Probleme mit der Linienmethode einsetzen wollen, stellt sich die Frage, wie oft eine Gitteradaption vorgenommen werden soll. Häufig stellt die für die Gitteradaption benötigte Rechenzeit einen nicht unerheblichen Anteil an der insgesamt benötigten Zeit des Verfahrens dar. Ein Aufruf von Algorithmus 4.2 in jedem Zeitschritt ist daher oftmals nicht die effizienteste Variante, insbesondere dann, wenn nur geringfügige Änderungen des Gitters erforderlich sind. Implementiert wurden dazu die folgenden drei Strategien:

Strategie 1. Algorithmus 4.2 wird in festen Zeitabständen τ_{adapt} aufgerufen.

Strategie 2. Algorithmus 4.2 wird jeweils nach einer fest vorgegebenen Anzahl k_{adapt} akzeptierter Zeitschritte aufgerufen.

Strategie 3. Algorithmus 4.2 wird nach k akzeptierten Zeitschritten aufgerufen. Zu Beginn des Zeitschrittverfahrens wird $k = 1$ gesetzt. Die Zahl k wird dem aktuellen Bedürfnis einer Gitterverfeinerung angepaßt und *bei jeder Gitteradaption neu* berechnet. Es sei

- n_{ref} die Anzahl der zur Verfeinerung markierten Elemente,
- n_{tot} die Anzahl aller Elemente,
- $\alpha < 1$ ein vorgegebener Faktor,
- M die maximal zulässige Feinheit der Elemente,
- $k_{\max}$ die vorgegebene maximale Anzahl von Zeitschritten, nach denen adaptiert wird.

In die Berechnung der Zahl k gehen die folgenden Überlegungen ein:

1. Wenn die maximal erlaubte Feinheit M noch nicht erreicht ist, soll sofort wieder verfeinert werden: $k = 1$.

2. Wenn nur ein geringer Teil der Elemente zur Verfeinerung markiert wurden, so kann in Zukunft seltener verfeinert werden. Offenbar befindet sich die Lösung in einer Phase nur geringer Veränderung.

3. Wenn 1. und 2. nicht zutreffen, dann kann öfter verfeinert werden.

Das drückt sich in dem folgenden Algorithmus zur Berechnung von k aus.

Algorithmus 4.3 (Häufigkeit der Adaption nach Strategie 3).

gegeben: M, $k_{\max}$, α

if $\max_{T\in\mathcal{T}_h} \mathrm{fh}(T) < M$

    $k = 1$

else

    if $n_{\mathrm{ref}} < \alpha\ n_{\mathrm{tot}}$
        $k = \min\{k+1, k_{\max}\}$
    else
        $k = \max\{k-1, 1\}$
    end

end

□

In den numerischen Berechnungen, die in Kapitel 9 vorgestellt werden, verwenden wir die Parameter $k_{\max} = 10$ und $\alpha = 0{,}05$.

Die in Algorithmus 4.2 markierten Elemente werden anschließend verfeinert bzw. vergröbert, wobei Vergröberung nur dann erfolgen kann, wenn alle Sohn-Elemente eines Vater-Elements zur Vergröberung markiert wurden.

4.3.3 Gleichverteilung des Fehlers

Durch die Wahl der Zahlenfolge $(\mu(i))_{i=0,\dots,M-1}$ kann die Struktur des Gitters gesteuert werden. So wird beispielsweise für $\mu(0) = \cdots = \mu(M-1) = 0$ ein uniformes Gitter der Feinheit M erzeugt. Mitunter wird das Ziel einer Gleichverteilung des geschätzten Fehlers angestrebt. Der folgende Satz gibt Schranken für die Zahlenfolge $(\mu(i))$ an, die für eine Gleichverteilung des Fehlers notwendig sind.

Satz 4.4. *Es sei*

- $\mathcal{T}_{h,0}$ *ein uniformes Grundgitter, das nur Dreiecke gleichen Flächeninhalts enthält,*
- η *ein Fehlerindikator, der die* L^2*-Norm des Gradienten des zeitlich und räumlich lokalen Fehlers abschätzt, d.h. zum Zeitpunkt* t_i *gilt* $\eta(u_h^\tau, T) \approx \|\nabla\widetilde{u}_{i-1}(\cdot,t_i) - \nabla u_h^\tau(\cdot,t_i)\|_{L^2(T)}$,

- η^* *ein vorgegebener Zielwert für den Fehlerindikator* η,
- φ *der zu* η *gehörige Verfeinerungs-Indikator,*
- $M \in \mathbb{N}_0$ *die maximal zulässige Gitterfeinheit,*
- zfh *die Zielfeinheits-Funktion, gegeben durch die Sprungstellen* $\mu(k)$, $k = 0, \ldots, M-1$,
- $\mathcal{T}_h$ *eine bezüglich der Zielfeinheit ideale Triangulierung und*
- $\mathcal{T}_h^{\mathrm{reg}} \subset \mathcal{T}_h$ *die Menge aller Dreiecke aus* $\mathcal{T}_h$*, die durch ausschließlich reguläre Verfeinerung aus den Dreiecken des Grundgitters* $\mathcal{T}_{h,0}$ *hervorgehen.*

Für jedes $k = 0, \ldots, M$ *existiere mindestens ein Dreieck* $T \in \mathcal{T}_h^{\mathrm{reg}}$ *mit* $fh(T) = k$*. Eine notwendige Bedingung für die Gleichverteilung*

$$\eta(u_h^\tau, T) = \eta^* \qquad \forall T \in \mathcal{T}_h^{\mathrm{reg}} \tag{4.10}$$

des lokalen Fehlerindikators über $\mathcal{T}_h^{\mathrm{reg}}$ *ist durch die Ungleichungen*

$$2^k \sqrt{\frac{|\mathcal{T}_{h,0}|}{|\Omega|}}\ \eta^* < \mu(k) \le 2^{k+1} \sqrt{\frac{|\mathcal{T}_{h,0}|}{|\Omega|}}\ \eta^*, \qquad k = 0, \ldots, M-1 \tag{4.11}$$

gegeben.

Beweis. Zur Vereinfachung der Schreibweise vereinbaren wir $\mu(-1) := 0$ und $\mu(M) := \infty$. Da $\mathcal{T}_h$ eine ideale Triangulierung ist, gilt $\mathrm{zfh}(\varphi(u_h^\tau, T)) = \mathrm{fh}(T)$ für alle $T \in \mathcal{T}_h$, siehe (4.8). Das ist wegen (4.9) gleichbedeutend mit

$$\mu(\mathrm{fh}(T) - 1) \le \varphi(u_h^\tau, T) < \mu(\mathrm{fh}(T)) \qquad \forall T \in \mathcal{T}_h. \tag{4.12}$$

Es gilt $\eta(u_h^\tau, T) = \sqrt{|T|}\ \varphi(u_h^\tau, T)$, siehe (4.7). Zusammen mit (4.6) und (4.10) ergibt sich

$$\eta^* = 2^{-\mathrm{fh}(T)} \sqrt{\frac{|\Omega|}{|\mathcal{T}_{h,0}|}}\ \varphi(u_h^\tau, T) \qquad \forall T \in \mathcal{T}_h^{\mathrm{reg}}. \tag{4.13}$$

Mit Hilfe der Beziehung (4.13) folgt aus (4.12)

$$\mu(\mathrm{fh}(T) - 1) \le 2^{\mathrm{fh}(T)} \sqrt{\frac{|\mathcal{T}_{h,0}|}{|\Omega|}}\ \eta^* < \mu(\mathrm{fh}(T)) \qquad \forall T \in \mathcal{T}_h^{\mathrm{reg}}. \tag{4.14}$$

Da diese Beziehung für alle $T \in \mathcal{T}_h^{\mathrm{reg}}$ gilt und Dreiecke aller Feinheiten $0, \ldots, M$ in $\mathcal{T}_h^{\mathrm{reg}}$ vertreten sind, folgt

$$\mu(k-1) \le 2^k \sqrt{\frac{|\mathcal{T}_{h,0}|}{|\Omega|}}\ \eta^* < \mu(k) \qquad \forall k = 0, \ldots, M,$$

also auch

$$2^k \sqrt{\frac{|\mathcal{T}_{h,0}|}{|\Omega|}}\ \eta^* < \mu(k) \le 2^{k+1} \sqrt{\frac{|\mathcal{T}_{h,0}|}{|\Omega|}}\ \eta^* \qquad \forall k = 0, \ldots, M-1.$$

□

Unter den Bedingungen des Satzes folgt

$$\mathrm{fh}(T) = \log_2 \left(\frac{\varphi(u_h^\tau, T)}{\eta^*} \sqrt{\frac{|\Omega|}{|\mathcal{T}_{h,0}|}} \right) \qquad \forall T \in \mathcal{T}_h^{\mathrm{reg}}$$

durch Umformung von (4.13). Wir definieren nun eine stetige Funktion $\zeta : \mathbb{R}_+ \to \mathbb{R}$ durch

$$\zeta(\varphi) := \log_2 \left(\frac{\varphi}{\eta^*} \sqrt{\frac{|\Omega|}{|\mathcal{T}_{h,0}|}} \right).$$

Die Funktion ζ wird als **Gleichverteilungs-Funktion** bezeichnet. Die Treppenfunktion zfh sollte eine Approximation an ζ darstellen, wenn eine Gleichverteilung des Fehlerindikators mit dem Wert η^* angestrebt wird.

Eine exakte Gleichverteilung des Fehlerindikators η, wie in (4.10) angegeben, ist in der Praxis natürlich nicht erfüllbar. Will man eine näherungsweise Gleichverteilung des Fehlerindikators erzielen, so erscheint es naheliegend, $\mu(k)$ gleich dem geometrischen Mittel der in (4.11) angegebenen Grenzen zu wählen. In diesem Falle approximiert die Zielfeinheits-Funktion zfh die Gleichverteilungs-Funktion ζ in der bestmöglichen Weise, siehe Abbildung 4.2. Der folgende Satz gibt bei dieser Wahl von $\mu(k)$ Schranken für $\eta(u_h^\tau, T)$ an.

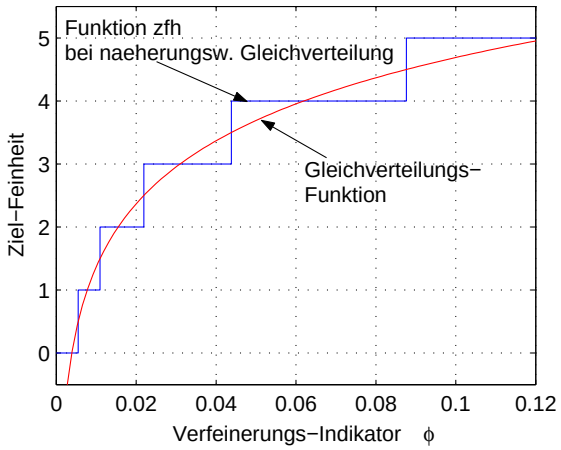


Abbildung 4.2: Zielfeinheits-Funktion zfh und Gleichverteilungs-Funktion ζ. Parameter: $|\Omega| = 100,\ |\mathcal{T}_{h,0}| = 60,\ M = 5,\ \eta^* = 0{,}005$

Satz 4.5. *Es sei $\mathcal{T}_{h,0}$ ein uniformes Grundgitter. Für ein $M \in \mathbb{N}_0$ und ein $\eta^* > 0$ seien*

$$\mu(k) := 2^{k+1/2} \sqrt{\frac{|\mathcal{T}_{h,0}|}{|\Omega|}}\ \eta^* \qquad \forall k = 0, \ldots, M-1 \tag{4.15}$$

die Sprungstellen der Zielfeinheits-Funktion zfh. *Es sei $\mathcal{T}_h$ eine bezüglich der Zielfeinheits-Funktion* zfh *ideale Triangulierung und $T \in \mathcal{T}_h$ ein Dreieck, für das die Beziehung (4.6) gilt.*

Dann genügt ein Fehlerindikator η, der die L^2-Norm des Gradienten des zeitlich und räumlich lokalen Fehlers abschätzt, der Ungleichung

$$\frac{1}{\sqrt{2}}\eta^* \leq \eta(u_h^\tau, T) < \sqrt{2}\eta^*.$$

Beweis. Setzt man in (4.15) $k = \mathrm{fh}(T)$, so ergibt sich

$$\mu(\mathrm{fh}(T)) = 2^{\mathrm{fh}(T)+1/2}\sqrt{\frac{|\mathcal{T}_{h,0}|}{|\Omega|}}\ \eta^*,$$

setzt man hingegen $k = \mathrm{fh}(T) - 1$, so erhält man

$$\mu(\mathrm{fh}(T)-1) = 2^{\mathrm{fh}(T)-1/2}\sqrt{\frac{|\mathcal{T}_{h,0}|}{|\Omega|}}\ \eta^*.$$

Aus beidem folgt, zusammen mit (4.12),

$$2^{\mathrm{fh}(T)-1/2}\sqrt{\frac{|\mathcal{T}_{h,0}|}{|\Omega|}}\ \eta^* \leq \varphi(u_h^\tau, T) < 2^{\mathrm{fh}(T)+1/2}\sqrt{\frac{|\mathcal{T}_{h,0}|}{|\Omega|}}\ \eta^*.$$

Wegen (4.7) ergibt sich daraus

$$2^{\mathrm{fh}(T)-1/2}\sqrt{\frac{|\mathcal{T}_{h,0}||T|}{|\Omega|}}\ \eta^* \leq \eta(u_h^\tau, T) < 2^{\mathrm{fh}(T)+1/2}\sqrt{\frac{|\mathcal{T}_{h,0}||T|}{|\Omega|}}\ \eta^*.$$

Aus (4.6) folgt

$$\sqrt{|T|} = 2^{-\mathrm{fh}(T)}\sqrt{\frac{|\Omega|}{|\mathcal{T}_{h,0}|}}$$

und damit die Behauptung. □

4.4 Ein numerisches Beispiel

Wir stellen hier ein numerisches Beispiel vor, bei dem die in Abschnitt 4.3.3 diskutierte Gleichverteilung des Fehlerindikators über die Dreiecke der Triangulierung *nicht* die effizienteste Variante darstellt, wie wir in Abschnitt 4.4.3 zeigen werden. In den folgenden beiden Abschnitten veranschaulichen wir zunächst die Möglichkeiten der Gittersteuerung und untersuchen die Güte des Z^2-Fehlerindikators.

Wir betrachten die Reaktions-Diffusions-Gleichung

$$u_t = \Delta u + r(1-u^2) + 2q^2(u-u^3) \tag{4.16}$$

mit der exakten Lösung

$$u(x,y,t) = \tanh((3x-12)+3t)$$

auf dem Gebiet $\Omega =]-5,5[^2$ im Zeitintervall $t \in [0,5]$. Anfangsbedingung und Dirichlet-Randbedingung seien entsprechend der exakten Lösung vorgegeben. Wir verwenden ein uniformes Grundgitter $\mathcal{T}_{h,0}$ aus 60 Dreiecken. Die maximal zulässige Gitterfeinheit sei $M = 5$. Wir verwenden lineare finite Elemente und die in Beispiel 5.22 beschriebene W-Methode mit der Fehlertoleranz $TOL_t = 10^{-4}$ zur Diskretisierung.

Die Lösung der Differentialgleichung bildet eine gerade Front, die sich bei $t = 0$ auf der Linie $x = 4$ befindet und sich mit gleichmäßiger Geschwindigkeit vom Betrag 1 in negativer x-Richtung bewegt.

Das Problem ist im Grunde ein eindimensionales, denn die Lösung ist nicht von y abhängig und Δu ist somit gleich u_{xx}. NOWAK [125] und LANG [103] verwendeten diese Differentialgleichung zu numerischen Untersuchungen. Eine Modifikation des Problems (4.16) wird in Abschnitt 9.1.1 vorgestellt und ebenfalls zu numerischen Testrechnungen herangezogen.

4.4.1 Güte des Z^2-Fehlerindikators

Untersuchung 4.6. Zunächst streben wir die Gleichverteilung des Fehlers an. Wir wählen $\eta^* = 0{,}005$ und berechnen die Sprungstellen $\mu(k),\ k = 0,\ldots,M-1$ der Zielfeinheits-Funktion entsprechend Satz 4.5. Daraus resultiert die Zielfeinheits-Funktion[2]

$$\mathrm{zfh}_0(\varphi) = \begin{cases} 0, & 0 \le \varphi \le 0{,}0055, \\ 1, & 0{,}0055 < \varphi \le 0{,}0110, \\ 2, & 0{,}0110 < \varphi \le 0{,}0219, \\ 3, & 0{,}0219 < \varphi \le 0{,}0438, \\ 4, & 0{,}0438 < \varphi \le 0{,}0876, \\ 5, & 0{,}0438 < \varphi. \end{cases} \tag{4.17}$$

Abbildung 4.2 zeigt die Funktion zfh_0 und die Gleichverteilungs-Funktion für dieses Problem.

Zum Zeitpunkt $t_0 = 0$ verfeinern wir das Gitter $\mathcal{T}_{h,0}$ entsprechend der Zielfeinheits-Funktion, diskretisieren mittels linearer finiter Elemente, führen *einen* Zeitschritt, $\tau = 0{,}01$ mit der in Beispiel 5.22 beschriebenen W-Methode aus und verfeinern zum Zeitpunkt $t_1 = \tau$ erneut. In Abbildung 4.3 stellen wir den Fehler $\|\nabla u_h^\tau(\mathbf{x},t_1) - \nabla\widetilde{u}_0(\mathbf{x},t_1)\|_{L^2(T)}$, den Z^2-Fehlerindikator $\eta^{\mathrm{Z}}(u_h^\tau, T)$ sowie das Gitter $\mathcal{T}_h$ graphisch dar.

Bei der Betrachtung der Abbildung 4.3 erkennt man, daß die Gleichverteilung des Fehlers hier ein etwas utopisches Ziel darstellt. Fehler und Z^2-Indikator befinden sich jedoch in recht guter Übereinstimmung. Direkt an der Front wird die Zielvorgabe η^* gut angenähert. Etwas abseits der Front ist das Gitter oftmals feiner, als es benötigt würde. Der Grund dafür ist, daß der Gittergenerator keine zu abrupten Übergänge von dem feinen zu dem groben Gitter erlaubt. Es werden 8453 Elemente benötigt.

Wie oben bereits angesprochen wurde, kann man auch andere Zielfeinheits-Funktionen verwenden, die nicht die Gleichverteilung des Fehlers anstreben. Wir stellen in Abbildung 4.4

- die oben definierte Funktion zfh_0 (Gleichverteilung),

[2] Zur Unterscheidung von einigen später definierten Zielfeinheits-Funktionen bezeichnen wir diese mit zfh_0.

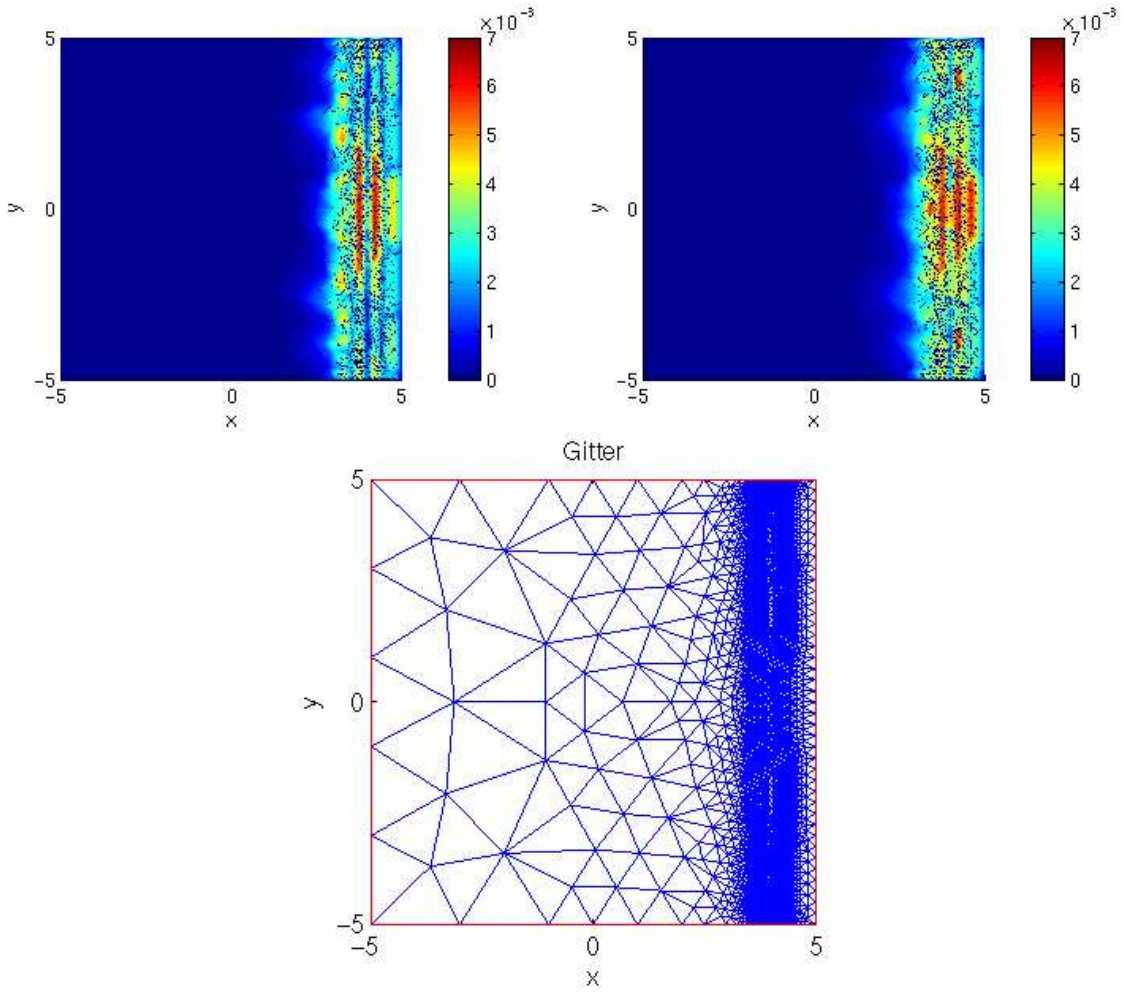


Abbildung 4.3: Problem (4.16) mit Zielfeinheits-Funktion zfh_0 nach einem Zeitschritt. **links oben:** Fehler $\|\nabla u_h^\tau(\mathbf{x},t_1) - \nabla \widetilde{u}_0(\mathbf{x},t_1)\|_{L^2(T)}$, **rechts oben:** Z^2-Fehlerindikator $\eta^{\mathrm{Z}}(u_h^\tau, T)$, **unten:** Gitter

- die Zielfeinheits-Funktion

$$\mathrm{zfh}_1(\varphi) = \begin{cases} 0, & 0 \le \varphi \le 0{,}0483, \\ 5, & 0{,}0483 < \varphi \end{cases} \tag{4.18}$$

  für eine besonders starke Frontverfeinerung und

- die Zielfeinheits-Funktion

$$\mathrm{zfh}_2(\varphi) = \begin{cases} 2, & 0 \le \varphi \le 0{,}05, \\ 3, & 0{,}05 < \varphi \end{cases} \tag{4.19}$$

  für eine nur schwache Frontverfeinerung

vor. Abbildung 4.5 zeigt Fehler und Z^2-Indikator bei Verwendung von zfh_1. Man erkennt hier wegen der hohen Frontauflösung eine sehr gute Übereinstimmung von Fehler und Indikator;

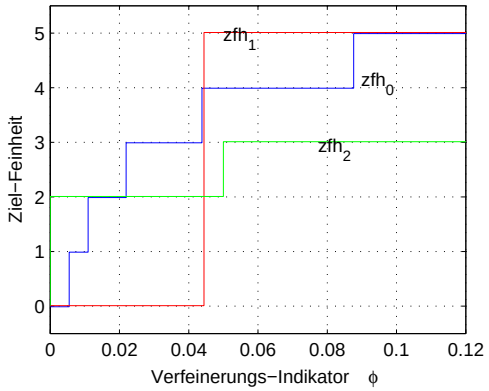


Abbildung 4.4: Zielfeinheits-Funktionen zfh_0, zfh_1 und zfh_2

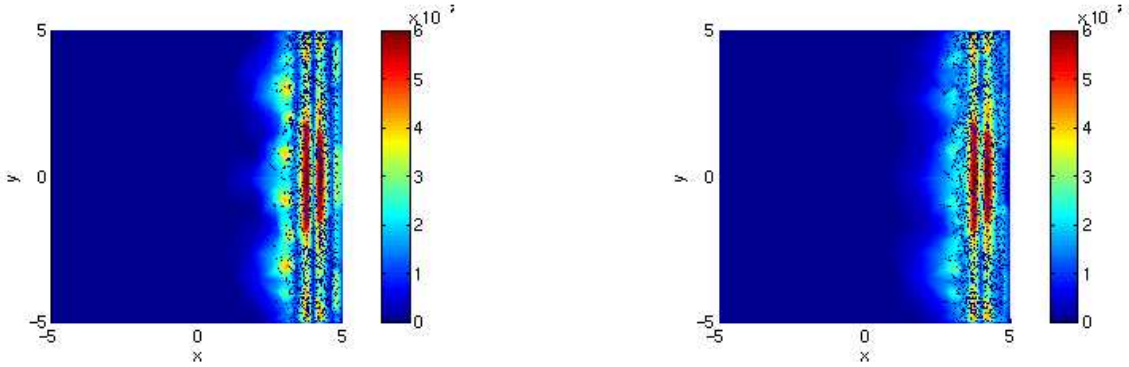


Abbildung 4.5: Fehler und Z^2-Indikator bei Verwendung von zfh_1

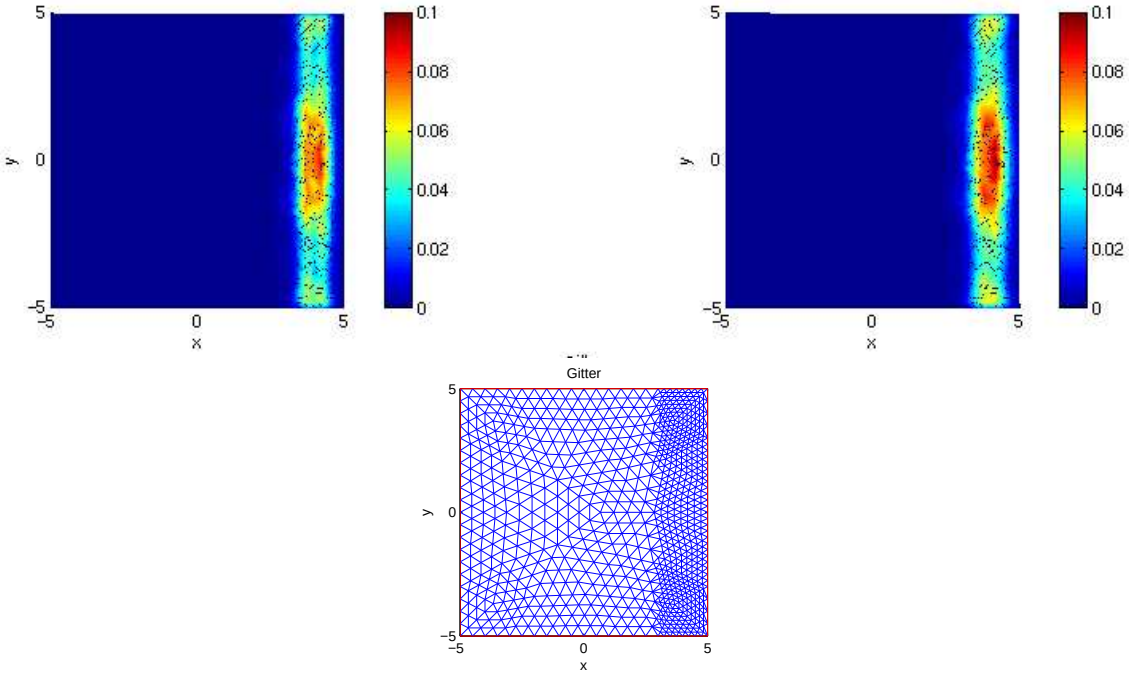


Abbildung 4.6: Fehler, Z^2-Indikator und Gitter bei Verwendung von zfh_2

sie ist besser als mit zfh_0, siehe Abbildung 4.3. Es werden 10209 Elemente benötigt, deutlich mehr als bei Verwendung von zfh_0. Aus der Abbildung wird deutlich, daß der Fehler besonders *an den Rändern* der Front auftritt, nämlich dort, wo die Krümmung der Lösung besonders hoch ist.

In Abbildung 4.6 sind Fehler, Z^2-Indikator und das Gitter bei Verwendung von zfh_2 dargestellt. Die Front ist hier nur um eine Stufe höher aufgelöst als die frontfernen Bereiche. Wegen der geringen Frontauflösung sind die Werte von Fehler und Indikator an der Front sehr hoch, aber die Übereinstimmung zwischen Fehler und Indikator ist immer noch gut. Es werden hier nur 1596 Elemente benötigt.

Ergebnisse der Untersuchung

- Durch verschiedene Zielfeinheits-Funktionen läßt sich die Struktur des Gitters steuern.
- Eine Gleichverteilung des Fehlers kann auch bei Verwendung von zfh_0 nicht erreicht werden, weil die dafür benötigte ideale Triangulierung nicht existiert.
- Fehler und Indikator stimmen gut überein.

□

4.4.2 Steuerung des Fehlers bei angestrebter Gleichverteilung

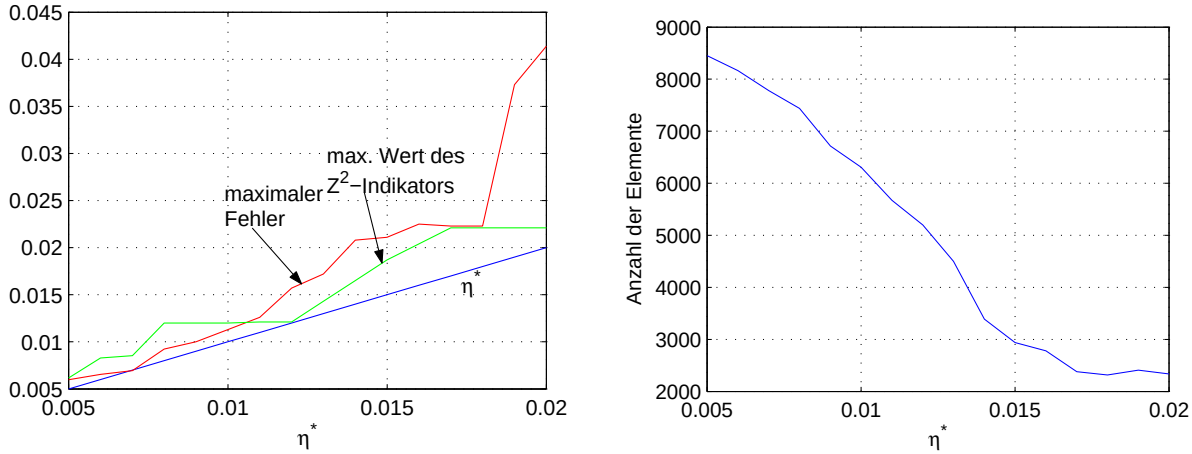


Abbildung 4.7: Einfluß von η^* auf Fehler und Fehlerindikator **(links)** und Anzahl der Elemente von $\mathcal{T}_h$ **(rechts)**

Untersuchung 4.7. In dieser numerischen Untersuchung wollen wir herausfinden, wie gut sich der Fehler durch die Vorgabe η^* steuern läßt. Wir betrachten dazu das obige Problem mit Zielfeinheits-Funktion zfh_0, lassen aber nun η^* nacheinander die Werte 0,005, 0,006, ..., 0,02 durchlaufen. Abbildung 4.7 zeigt links den maximalen Fehler $\max_{T\in\mathcal{T}_h} \|\nabla u_h^\tau(\mathbf{x},t_1) - \nabla\widetilde{u}_0(\mathbf{x},t_1)\|_{L^2(T)}$ und den maximalen Fehlerindikator $\max_{T\in\mathcal{T}_h} \eta^Z(u_h^\tau, T)$ in Abhängigkeit von η^*. Rechts stellen wir den Einfluß von η^* auf die Anzahl der Elemente von $\mathcal{T}_h$ dar.

Ergebnisse der Untersuchung

- Das Maximum des Fehlerindikators ist nur geringfügig größer als die Vorgabe η^*, was der Aussage des Satzes 4.5 entspricht.

- Für kleine Werte von η^*, also für ein feines Gitter, stimmen Fehler und Indikator gut überein. Lediglich für sehr große η^* ist die Übereinstimmung schlecht. Das entspricht der in Abschnitt 4.1.1 formulierten theoretischen Aussage, wonach der Z^2-Indikator vor allem dann den Fehler gut abschätzt, wenn der Fehler klein ist, also auf dem feinen Gitter.

□

4.4.3 Effizienzuntersuchung

Anhand des Problems (4.16) läßt sich zeigen, daß eine Gitteradaption mit dem Ziel der Gleichverteilung des Fehlerindikators nicht immer einer effizienten Lösung des Problems dient. Die folgende numerische Untersuchung verdeutlicht diese Tatsache.

Untersuchung 4.8. Wir führen numerische Berechnungen des Problems (4.16) mit dem oben angegebenen Verfahren und den in (4.17) und (4.18) definierten Zielfeinheits-Funktionen zfh_0 und zfh_1 durch. Die Häufigkeit der Gitteradaption legen wir dabei nach der in Abschnitt 4.3.2 beschriebenen Strategie 1 fest. Die dort auftretende Größe τ_{adapt} hat einen starken Einfluß auf die Effizienz des Verfahrens. Um das zu berücksichtigen, variieren wir τ_{adapt} gemäß $\tau_{\mathrm{adapt}} = 0{,}01, 0{,}02, \ldots, 0{,}20$. Aus Abbildung 4.9 links geht hervor, daß – wie erwartet – das Verfahren mit zfh_1 mehr Gitterknoten erzeugt als das Verfahren mit zfh_0. In Abbildung 4.8 stellen wir die gemessene Rechenzeit über dem zeitlich gemittelten L^2-Fehler dar. Die Graphik zeigt eine erheblich höhere Effizienz des Verfahrens mit der Zielfeinheits-Funktion zfh_1, also des Verfahrens, bei dem *keine* Gleichverteilung des Fehlers angestrebt wird.

Die höhere Genauigkeit des Verfahrens mit zfh_1 läßt sich mit der stärkeren Gitterverfeinerung erklären. Überraschend ist jedoch, daß dieses Verfahren bei kleinen Werten von τ_{adapt} *weniger Rechenzeit* als das Verfahren mit zfh_0 benötigt. Um dieses Verhalten verständlich zu machen, zeigen wir in Abbildung 4.9 rechts jeweils den mittleren Zeitschritt. Man erkennt bei dem Verfahren mit zfh_0 einen starken Einbruch des Zeitschritts für die kleinen Werte von τ_{adapt}, also gerade dann, wenn durch häufige Verfeinerung eigentlich ein *besseres* Gitter vorliegen sollte. Eine mögliche Erklärung für dieses Phänomen wäre, daß durch häufige Gitterverfeinerung direkt an der sensiblen Front Fehler entstehen – beispielsweise infolge einer möglicherweise unzureichenden Interpolation der Lösungswerte auf neue Gitterpunkte. Die extrem kleinen Zeitschritte des Verfahrens mit zfh_0 für kleine Werte von τ_{adapt} sind jedenfalls für die in diesem Falle sehr hohe Rechenzeit verantwortlich.

Für $\tau_{\mathrm{adapt}} = 0{,}01$ sind in Abbildung 4.10 für beide Zielfeinheits-Funktionen die Zeitschritte und der Fehler der Frontposition über der Zeit aufgetragen. Die Abbildungen 4.11 zeigen jeweils einen Ausschnitt des Gitters für $\tau_{\mathrm{adapt}} = 0{,}01$ zur Zeit $t = 2{,}5$. Man erkennt die stärkere Auflösung der Front, wenn die Zielfeinheits-Funktion zfh_1 verwendet wird.

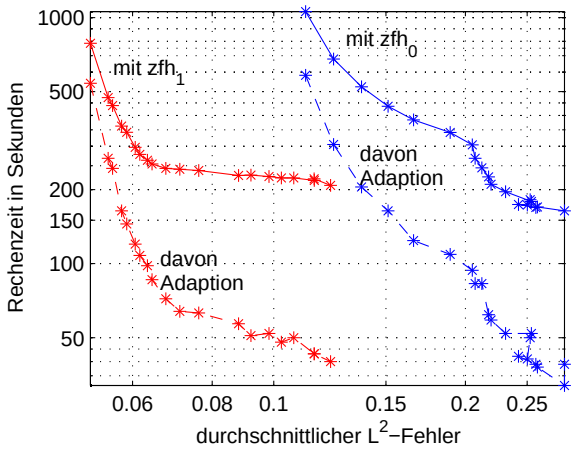


Abbildung 4.8: Vergleich der Effizienz bei Verwendung von zfh_0 und zfh_1. Auf jeder Kurve variiert τ_{adapt} von 0,01 (links oben) bis 0,20 (rechts unten). Die gestrichelte Linie gibt an, wieviel Rechenzeit davon die Gitteradaption benötigt.

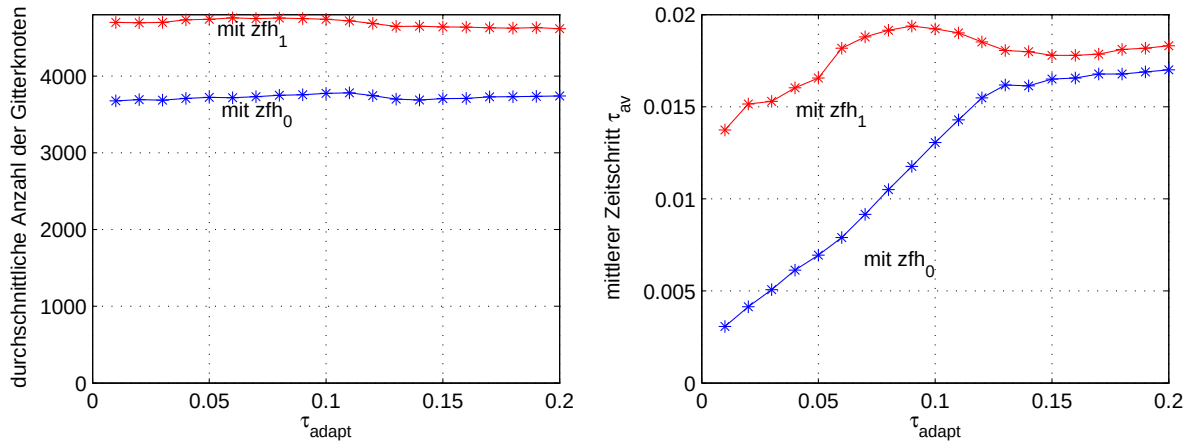


Abbildung 4.9: **links:** Anzahl der Gitterknoten, **rechts:** mittlerer Zeitschritt, beides in Abhängigkeit von τ_{adapt}

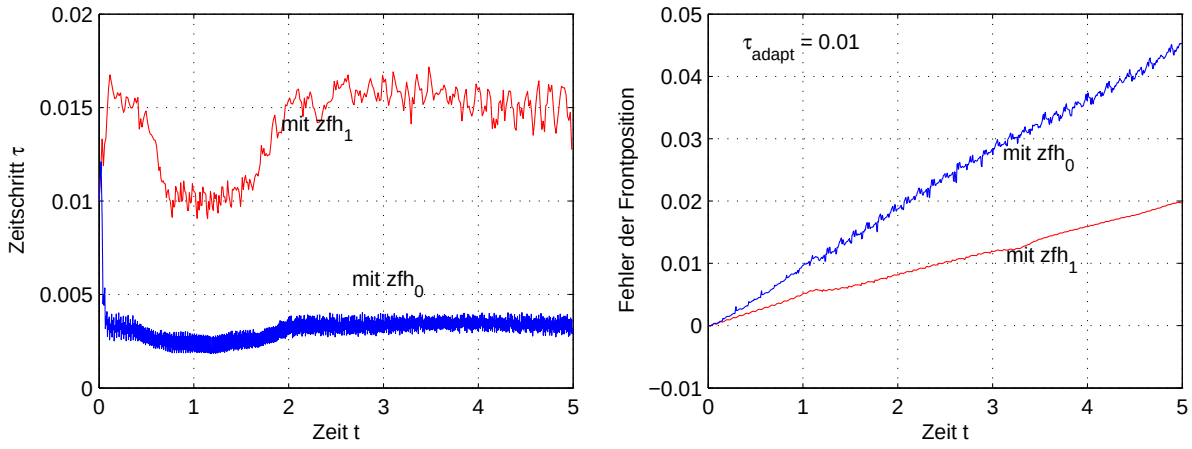


Abbildung 4.10: **links:** Zeitschritte, **rechts:** Fehler der Frontposition, beides für $\tau_{\mathrm{adapt}} = 0{,}01$

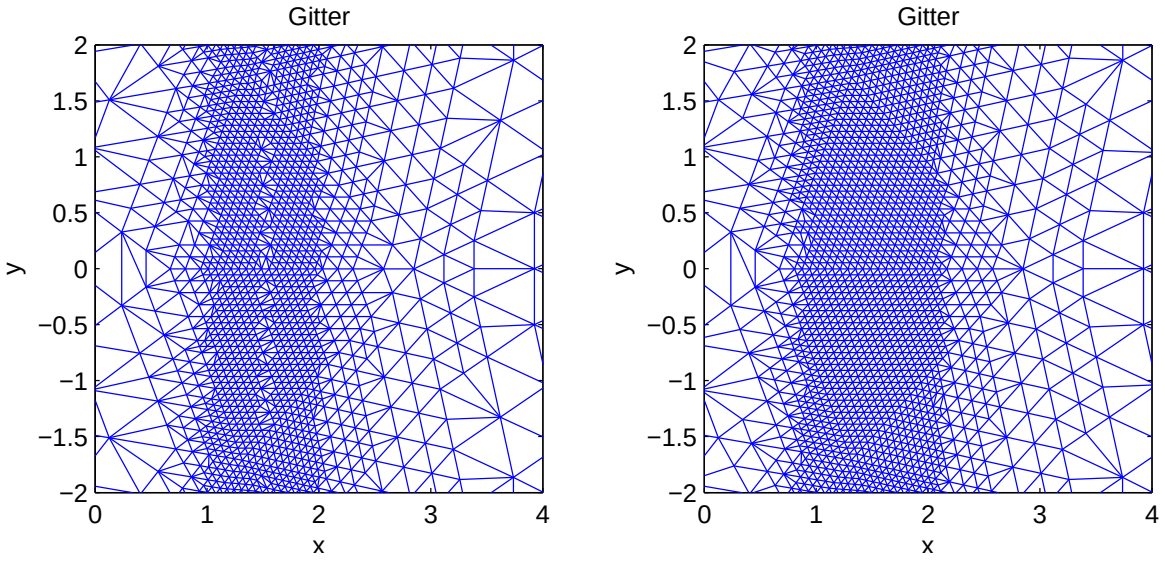


Abbildung 4.11: Gitter (Ausschnitt) für $\tau_{\text{adapt}} = 0{,}01$ zur Zeit $t = 2{,}5$, **links:** Verfahren mit zfh_0, **rechts:** Verfahren mit zfh_1

Ergebnisse der Untersuchung

- Die Gitterverfeinerung mit dem Ziel der Gleichverteilung des Fehlers ist nicht so effizient wie die Adaption bei Verwendung der Zielfeinheits-Funktion zfh_1.
- Bei häufiger Gitteradaption direkt an der Front treten hohe numerische Fehler auf.

□

Aus dem betrachteten Beispiel läßt sich schließen, daß es mitunter für eine effiziente Lösung außerordentlich sinnvoll ist, eine bewegte Front in einer gewissen Umgebung uniform zu verfeinern, auch wenn das dem Grundsatz der Gleichverteilung des Fehlers über das Gitter widerspricht.

4.5 Gittererzeugung auf gekrümmten Flächen

Für die in Abschnitt 3.2 beschriebene Methode der finiten Elemente auf einer gekrümmten Fläche S im $\mathbb{R}^3$ benötigen wir die Approximation der Fläche S durch eine Polyederfläche S_h, deren Eckpunkte sämtlich auf S liegen. Diese Polyederfläche wird in der folgenden Weise erzeugt:

- Die Fläche S wird durch eine Projektion P in eine Ebene E abgebildet.
- In dem Bildgebiet $P(S)$ wird durch den Gittergenerator aus dem Programmpaket UG eine Triangulierung $\mathcal{T}_h$ erzeugt.

- Die Gitterpunkte dieser Triangulierung werden durch die Umkehrabbildung P^{-1} wieder auf die Fläche S projiziert.
- Wenn drei Gitterpunkte $\mathbf{x}_i$, $\mathbf{x}_j$ und $\mathbf{x}_k$ Eckpunkte eines Dreiecks $T \in \mathcal{T}_h$ sind, dann sind die Bildpunkte $P^{-1}(\mathbf{x}_i)$, $P^{-1}(\mathbf{x}_j)$ und $P^{-1}(\mathbf{x}_k)$ Eckpunkte eines ebenen Dreiecks $\widetilde{T}$ in $\mathbb{R}^3$. Die Gesamtheit aller derartiger Dreiecke $\widetilde{T}$ bildet die Polyederfläche S_h.

Der verwendete UG-Gittergenerator vermeidet automatisch Dreiecke mit sehr großen Innenwinkeln, d.h. er beachtet die in Abschnitt 3.1.2 angegebene Maximalwinkel-Bedingung. Diese Eigenschaft sollte auch für die Dreiecke $\widetilde{T} \subset S_h$ gewahrt bleiben. Es ist daher sinnvoll, eine möglichst winkeltreue Projektion P zu verwenden.

Wir beschreiben im folgenden die Gittererzeugung, wenn die Fläche S eine Sphäre oder ein Ellipsoid ist, da diese beiden Fälle für die in den Abschnitten 10.7 und 10.9 dargestellten Probleme relevant sind.

4.5.1 Gittererzeugung auf der Sphäre

Ist die Fläche S eine Sphäre, so existiert mit der **stereographischen Projektion** eine winkeltreue Abbildung. Die stereographische Projektion ist die Zentralprojektion aus einem Punkt der Sphäre auf die gegenüberliegende Tangentialebene. Wir betrachten etwa die durch die Gleichung

$$x^2 + y^2 + z^2 = r^2$$

gegebene Sphäre S. Die in einem Punkte $\mathbf{x} \in S$ angeheftete Tangentialebene von S werde mit $T_{\mathbf{x}}(S)$ bezeichnet. Die Einheitsvektoren in Richtung der kartesischen Koordinatenachsen seien $\mathbf{i}$, $\mathbf{j}$ und $\mathbf{k}$. Die stereographische Projektion $P_{\mathbf{a}} : S \to T_{-\mathbf{a}}(S)$ aus dem Punkte $\mathbf{a}$ wird für $\mathbf{a} \neq \pm r\mathbf{i}$ durch die Beziehungen

$$\mathbf{f} = \frac{\mathbf{i} \times \mathbf{a}}{|\mathbf{i} \times \mathbf{a}|}, \qquad \mathbf{e} = \frac{\mathbf{a} \times \mathbf{f}}{r}, \qquad P_{\mathbf{a}}(\mathbf{x}) = \mathbf{a} + \sigma(\mathbf{x} - \mathbf{a}),$$

$$P_{\mathbf{a}}(\mathbf{x}) = -\mathbf{a} + X\mathbf{e} + Y\mathbf{f}, \qquad X, Y, \sigma \in \mathbb{R}$$

vollständig beschrieben. Dabei sind $\mathbf{e}$ und $\mathbf{f}$ orthonormale Vektoren, die die Tangentialebene $T_{-\mathbf{a}}(S)$ aufspannen. Die stereographische Projektion auf einer Sphäre wird in Abbildung 4.12 dargestellt. Die Zahl σ wird als Streckungsfaktor bezeichnet, da ein infinitesimal kleines geometrisches Objekt, das auf S liegt und den Punkt $\mathbf{x}$ enthält, unter der Projektion $P_{\mathbf{a}}$ eine Ähnlichkeitstransformation erfährt, bei der es um diesen Faktor gestreckt wird. Dieser Sachverhalt wird in Satz 4.9 gezeigt. Der Streckungsfaktor σ kann auf die folgende Weise anschaulich dargestellt werden: Es sei a die Entfernung des Bildpunktes $P_{\mathbf{a}}(\mathbf{x})$ vom Projektionszentrum $\mathbf{a}$ und b die Entfernung des Punktes $\mathbf{x}$ von $\mathbf{a}$. Dann ist der Streckungsfaktor $\sigma(\mathbf{x})$ gerade gleich dem Verhältnis a/b.

In unserem Falle sind wir zumeist nicht an dem Vektor $P_{\mathbf{a}}(\mathbf{x})$ interessiert, sondern lediglich an den Koordinaten X und Y von $P_{\mathbf{a}}(\mathbf{x})$ bezüglich der Basis $\{\mathbf{e}, \mathbf{f}\}$. Daher führen wir die Abbildung $\widetilde{P}_{\mathbf{a}} : S \to \mathbb{R}^2$ ein, die durch $\widetilde{P}_{\mathbf{a}}(\mathbf{x}) = (X, Y)$ mit

$$X = \sigma\mathbf{x} \cdot \mathbf{e}, \qquad Y = \sigma\mathbf{x} \cdot \mathbf{f}, \qquad \sigma = \frac{2r^2}{r^2 - \mathbf{a} \cdot \mathbf{x}}, \qquad \mathbf{f} = \frac{\mathbf{i} \times \mathbf{a}}{|\mathbf{i} \times \mathbf{a}|}, \qquad \mathbf{e} = \frac{\mathbf{a} \times \mathbf{f}}{r} \tag{4.20}$$

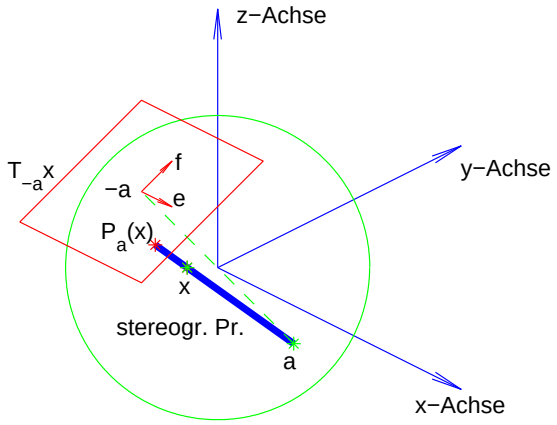


Abbildung 4.12: Stereographische Projektion

definiert ist. Oft wird vereinfachend auch von $\widetilde{P}_{\mathbf{a}}$ als stereographischer Projektion gesprochen. Wir wollen diese Bezeichnung hier auch verwenden.

Die stereographische Projektion $\widetilde{P}_{\mathbf{a}}$ bildet S auf den gesamten $\mathbb{R}^2$ ab. Da der Gittergenerator jedoch nur zur Triangulierung beschränkter Gebiete geeignet ist, so bilden wir mit der stereographischen Projektion lediglich eine Halbsphäre in den $\mathbb{R}^2$ ab. Die Bildmenge ist dann ein Kreis mit dem Radius $2r$. Die andere Halbsphäre wird durch die Projektion aus dem gegenüberliegenden Punkt abgebildet. Wir wählen beispielsweise $\widetilde{P}_{-r\mathbf{k}} : S_1 \to \mathbb{R}^2$ und $\widetilde{P}_{r\mathbf{k}} : S_2 \to \mathbb{R}^2$, wobei S_1 die nördliche und S_2 die südliche Halbsphäre ist. Für diese Projektionen aus den Polen vereinfacht sich die Darstellung (4.20) wie folgt:

$$\mathbf{x} = (x, y, z), \qquad \widetilde{P}_{-r\mathbf{k}}(\mathbf{x}) = (X, Y) \text{ mit } X = \sigma x,\ Y = \sigma y,\ \sigma = \frac{2r}{r+z} \tag{4.21}$$

und

$$\mathbf{x} = (x, y, z), \qquad \widetilde{P}_{r\mathbf{k}}(\mathbf{x}) = (X, Y) \text{ mit } X = \sigma x,\ Y = \sigma y,\ \sigma = \frac{2r}{r-z}.$$

Im folgenden untersuchen wir, in welcher Weise ein kleines Dreieck $\widetilde{T} \subset S_h$ durch die Projektion $\widetilde{P}_{\mathbf{a}}$ seine Gestalt ändert. Dazu fragen wir zunächst, wie zwei nahe beieinanderliegende Punkte auf S unter der Projektion $\widetilde{P}_{\mathbf{a}}$ ihren Abstand ändern.

Satz 4.9. *Es seien $\mathbf{x} = (x, y, z)$ und $\mathbf{x} + d\mathbf{x} = (x + dx, y + dy, z + dz)$ zwei Punkte auf der durch $x^2 + y^2 + z^2 = r^2$ gegebenen Kugel S mit dem infinitesimal kleinen Abstand $|d\mathbf{x}|$. Dann haben die Bildpunkte $\widetilde{P}_{\mathbf{a}}(\mathbf{x})$ und $\widetilde{P}_{\mathbf{a}}(\mathbf{x}+d\mathbf{x})$ der stereographischen Projektion $\widetilde{P}_{\mathbf{a}}$ den Abstand $\sigma|d\mathbf{x}|$, wobei $\sigma = 2r^2/(r^2 - \mathbf{a}\cdot\mathbf{x})$ der in (4.20) definierte Streckungsfaktor ist.*

Beweis. Aus Symmetriegründen reicht es aus, den Satz für einen speziellen Vektor $\mathbf{a}$ zu zeigen. Wir wählen $\mathbf{a} = -r\mathbf{k}$, d.h. wir betrachten die stereographische Projektion aus dem Südpol auf die im Nordpol angeheftete Tangentialebene. Nach der Formel (4.21) gilt

$$X = \frac{2rx}{r+z}, \qquad X + dX = \frac{2r(x+dx)}{r+z+dz}.$$

Es folgt

$$dX = 2r\frac{(x+dx)(r+z) - x(r+z+dz)}{(r+z+dz)(r+z)},$$

woraus durch Vernachlässigung des Terms dz im Nenner die Aussage

$$dX = 2r\frac{(r+z)dx - xdz}{(r+z)^2}$$

folgt. Analog erhält man

$$dY = 2r\frac{(r+z)dy - ydz}{(r+z)^2}.$$

Demnach gilt

$$\begin{aligned} dX^2 + dY^2 &= \frac{4r^2}{(r+z)^4}\left(((r+z)dx - xdz)^2 + ((r+z)dy - ydz)^2\right) \qquad (4.22)\\ &= \frac{4r^2}{(r+z)^4}\left((r+z)^2(dx^2+dy^2) + (x^2+y^2)dz^2 - 2(r+z)dz(xdx+ydy)\right). \end{aligned}$$

Da der Vektor $\mathbf{x} + d\mathbf{x}$ den Betrag r hat, folgt

$$(x+dx)^2 + (y+dy)^2 + (z+dz)^2 = x^2+y^2+z^2+2xdx+2ydy+2zdz+dx^2+dy^2+dz^2 = r^2.$$

Wegen $x^2+y^2+z^2 = r^2$ ergibt sich, bei Vernachlässigung der Terme dx^2, dy^2 und dz^2, die Gleichung $xdx + ydy + zdz = 0$. Setzt man das in (4.22) ein, so folgt

$$\begin{aligned} dX^2 + dY^2 &= \frac{4r^2}{(r+z)^4}\left((r+z)^2(dx^2+dy^2) + (x^2+y^2+2z(r+z))dz^2\right)\\ &= \frac{4r^2}{(r+z)^4}\left((r+z)^2(dx^2+dy^2) + (r^2-z^2+2z(r+z))dz^2\right)\\ &= \frac{4r^2}{(r+z)^2}(dx^2+dy^2+dz^2). \end{aligned}$$

Zieht man daraus die Wurzel, so ergibt sich die Behauptung. □

Man beachte, daß die Streckung um den Faktor σ, die der Vektor $d\mathbf{x}$ durch die Projektion $P_{\mathbf{a}}$ erfährt, nur von dem Skalarprodukt $\mathbf{a}\cdot\mathbf{x}$, also insbesondere *nicht von der Orientierung* des Vektors $d\mathbf{x}$ abhängt. Damit ist die stereographische Projektion für infinitesimal kleine Dreiecke eine Ähnlichkeits-Transformation, bei der die Innenwinkel erhalten bleiben. Die Winkeltreue der Projektion $P_{\mathbf{a}}$ wurde somit ebenfalls gezeigt.

Zur Gittererzeugung auf der nördlichen Halbsphäre S_1 verwenden wir, wie oben bereits erwähnt, die stereographische Projektion $P_{-r\mathbf{k}}$. Ein verbleibendes Problem ist die Erzeugung eines geeigneten Grundgitters im Projektionsgebiet $P_{-r\mathbf{k}}(S_1)$. Unser Ziel ist es, ein uniformes Gitter *auf der Sphäre* zu erzeugen. Demnach sollte das Gitter auf dem Kreis $P_{-r\mathbf{k}}(S_1)$ am Rand größere Dreiecke enthalten als in der Mitte. Die Größenverhältnisse sind durch den Streckungsfaktor σ vorgegeben. Wir geben auf $P_{-r\mathbf{k}}(S_1)$ gewisse Gitterpunkte G_i vor, deren Urbilder $P^{-1}_{-r\mathbf{k}}(G_i)$ auf S nahezu den gleichen Abstand voneinander haben. Das wird durch die folgende Vorgehensweise erreicht.

Es seien φ und ϑ die sphärischen Koordinaten auf S, gegeben durch

$$\begin{aligned} x &= r\cos\varphi\cos\vartheta,\\ y &= r\sin\varphi\cos\vartheta,\\ z &= r\sin\vartheta \end{aligned}$$

Zunächst projizieren wir die $n+1$ Breitenkreise[3] $B_k := \{(\varphi, \vartheta) : \vartheta = k\pi/2n\},\ k = 0, \ldots, n$ mittels $P_{-r\mathbf{k}}$ auf das Bildgebiet $P_{-r\mathbf{k}}(S_1)$. In der Bildebene ergibt das konzentrische Kreise $P_{-r\mathbf{k}}(B_k)$ um den Ursprung mit den Radien $R_k = 2r\tan((n-k)\pi/4n)$. Die Gitterpunkte werden nun auf den Kreisen $P_{-r\mathbf{k}}(B_k)$ plaziert. Wir bezeichnen mit N_k die Anzahl der Punkte, die auf dem Kreis $P_{-r\mathbf{k}}(B_k)$ liegen sollen. Der zu einem Punkt degenerierte Kreis $P_{-r\mathbf{k}}(B_n)$ enthält den Gitterpunkt $G_{n,0} = (0,0)$. Der nächste Kreis $P_{-r\mathbf{k}}(B_{n-1})$ sollte sechs Gitterpunkte $G_{n-1,0}, \ldots, G_{n-1,5}$ enthalten, die äquidistant angeordnet werden. Nun soll für $m = 0, \ldots, n-1$ die Anzahl N_m der Punkte auf dem Kreis $P_{-r\mathbf{k}}(B_k)$ näherungsweise proportional zur Länge des Breitenkreises B_k sein. Die Länge der Breitenkreise beträgt

$$|B_k| = 2\pi r \sin\frac{(n-k)\pi}{2n}, \qquad k = 0, \ldots, n.$$

Mit $N_{n-1} = 6$ kann man daraus die Zahlen $N_0, \ldots, N_{n-2}$ berechnen:

$$N_k = \text{round}\left(\frac{6\sin\frac{(n-k)\pi}{2n}}{\sin\frac{\pi}{2n}}\right).$$

Dabei liefert die Funktion round() den gerundeten Wert auf die nächstliegende ganze Zahl. Die Gitterpunkte sollen auf den Kreisen $P_{-r\mathbf{k}}(B_k)$ äquidistant verteilt werden. Wir wählen die folgenden Positionen

$$G_{m,k} = \left(R_m \cos\frac{2k\pi}{N_m}, R_m \sin\frac{2k\pi}{N_m}\right), \qquad k = 0, \ldots, N_m - 1,\ m = 0, \ldots, n-1.$$

Sind die Punkte $G_{m,k}$ bestimmt, so kann das Gitter durch eine Delaunay-Triangulierung erzeugt werden. Abbildung 4.13 zeigt die Kreise $P_{-r\mathbf{k}}(B_k)$ im Bildgebiet $P_{-r\mathbf{k}}(S_1)$ sowie das Dreiecksgitter. Im Anschluß an die Gittererzeugung kann noch eine Gitterglättung vorgenommen werden, die einzelne Gitterknoten geringfügig verschiebt, um sehr spitze bzw. stumpfe Innenwinkel zu vermeiden. Für unsere Berechnungen benutzen wir die Delaunay-Triangulierung und die Gitterglättung aus MATLAB [114]. Die Punkte $P^{-1}_{-r\mathbf{k}}(G_{m,k})$ sind dann die Eckpunkte der Polyederfläche S_h.

4.5.2 Gittererzeugung auf dem Ellipsoid

Es sei S das durch die Gleichung

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$$

gegebene achsenparallele Ellipsoid. Um auf S ein Gitter zu erzeugen, transformieren wir das Ellipsoid zunächst auf die Einheitskugel B. Dazu verwenden wir die Abbildung $K : S \to B$, die durch $K(x,y,z) = (x/a, y/b, z/c)$ gegeben ist. Wir erzeugen nun ein Gitter auf B durch die stereographischen Projektionen $P_{\mathbf{a}}$ und $P_{-\mathbf{a}}$ in der in Abschnitt 4.5.1 beschriebenen Weise. Anschließend bilden wir die Gitterpunkte auf B mittels der Abbildung K^{-1} auf S ab und verbinden die entsprechenden Gitterpunkte geradlinig. Im Ergebnis erhalten wir eine das Ellipsoid S approximierende Polyederfläche S_h.

[3] Die Zahl $n \in \mathbb{N}$ sollte in der Größenordnung von 10 liegen.

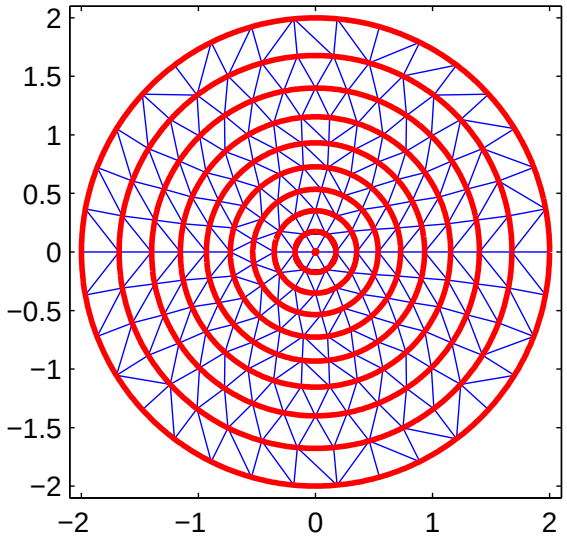


Abbildung 4.13: Bildgebiet $P_{-r\mathbf{k}}(S_1)$. **rot:** Kreise $P_{-r\mathbf{k}}(B_k)$, **blau:** Gitter

Die Abbildung K ist jedoch im allgemeinen *nicht winkeltreu*. Je stärker sich die Halbachsenlängen a, b und c unterscheiden, umso stärker verzerrte Dreiecke können in S_h auftreten. Die genannte Methode der Gittererzeugung ist daher für Ellipsoide mit extrem verschiedenen Halbachsenlängen nicht geeignet.

Kapitel 5

Zeitintegration durch Runge-Kutta-Verfahren

Im Ergebnis der Ortsdiskretisierung eines Reaktions-Diffusions-Problems von erhält man ein System gewöhnlicher Differentialgleichungen, welches entweder in der Form

$$\mathbf{u}_t = \mathbf{f}(t, \mathbf{u}), \tag{5.1}$$

siehe (3.15), (3.16), (3.33), oder in der Form

$$\mathbf{M}\mathbf{u}_t = \mathbf{f}(t, \mathbf{u}), \tag{5.2}$$

siehe (3.9), (3.13), vorliegt. Dieses System muß mit einem geeigneten numerischen Verfahren gelöst werden. Das folgende Kapitel befaßt sich daher mit numerischen Lösungsverfahren für ein System der Form (5.1). Auch für Systeme der Form (5.2) existieren numerische Verfahren, jedoch werden wir diese hier nicht betrachten.

Zur numerischen Behandlung gewöhnlicher Differentialgleichungen existiert eine umfangreiche Theorie. Als wichtige Standardwerke auf diesem Gebiet seien die Bücher von Hairer/Nørsett/Wanner [77, 78], Deuflhard/Bornemann [51] und Strehmel/Weiner [155] empfohlen. Wir werden im Rahmen dieser Arbeit nur auf die **Runge-Kutta-Verfahren** eingehen, die eine wichtige Verfahrensklasse zur Lösung gewöhnlicher Differentialgleichungen bilden.

5.1 Runge-Kutta-Verfahren

Wir betrachten das System gewöhnlicher Differentialgleichungen

$$\mathbf{u}_t = \mathbf{f}(t, \mathbf{u}) \tag{5.3}$$

mit einer Anfangsbedingung $\mathbf{u}(t_0) = \mathbf{u}_0$. Es sei $t \in [t_0, t_e]$ die Zeitvariable und $\mathbf{u}$ eine vektorwertige differenzierbare Funktion $\mathbf{u} : [t_0, t_e] \to \mathbb{R}^n$. Das Zeitintervall werde gemäß $t_0 < t_1 < \cdots < t_N = t_e$ diskretisiert und die Länge der Teilintervalle mit

$$\tau_i := t_{i+1} - t_i \tag{5.4}$$

bezeichnet. Das numerische Verfahren liefert eine nur in den Zeitpunkten t_i gegebene Lösung mit Werten $\mathbf{u}_i$, die den exakten Lösungswert $\mathbf{u}(t_i)$ approximieren.

Das einfachste Verfahren zur numerischen Lösung des Problems (5.3) wurde bereits im Jahre 1768 von EULER [60] angegeben. Hierbei ersetzt man die Zeitableitung $\mathbf{u}_t$ zur Zeit $t = t_i$ durch den Differenzenquotienten $(\mathbf{u}_{i+1} - \mathbf{u}_i)/\tau_i$ und erhält so das **Eulersche Polygonzug-Verfahren** (auch: explizites Euler-Verfahren)

$$\frac{\mathbf{u}_{i+1} - \mathbf{u}_i}{\tau_i} = \mathbf{f}(\mathbf{u}_i, t_i), \qquad i = 0, \ldots, N.$$

Der Vektor $\mathbf{u}_0$ ist durch die Anfangsbedingung gegeben.

Die Genauigkeit der numerischen Lösung kann durch kompliziertere Verfahren erhöht werden. RUNGE [137] und HEUN [82] entwickelten gegen Ende des 19. Jahrhunderts verschiedene derartige Verfahren. KUTTA [100] gab 1901 eine allgemeine Form dieser Verfahren an, die später als **Runge-Kutta-Verfahren** bezeichnet wurden:

Definition 5.1. Gegeben sei eine Matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{s\times s}$ und ein Vektor $\mathbf{b} = (b_i) \in \mathbb{R}^s$. Mit $c_i = \sum_{j=1}^{s} a_{ij}$ ist ein s-stufiges Runge-Kutta-Verfahren durch die Vorschrift

$$\begin{aligned} \mathbf{k}_j &= \mathbf{f}\left(t_i + c_j\tau_i, \mathbf{u}_i + \tau_i \sum_{l=1}^{s} a_{jl}\mathbf{k}_l\right), \qquad j = 1, \ldots, s, \\ \mathbf{u}_{i+1} &= \mathbf{u}_i + \tau_i \sum_{l=1}^{s} b_l \mathbf{k}_l \end{aligned} \tag{5.5}$$

definiert. □

Die Koeffizienten a_{ij}, b_i und c_i werden üblicherweise in einem sogenannten **Butcher-Tableau**

$$\begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array}$$

angeordnet, eine Schreibweise, die 1964 von BUTCHER [31] eingeführt wurde. Das eingangs erwähnte Eulersche Polygonzugverfahren wird durch das Butcher-Tableau

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

dargestellt.

5.2 Konsistenzordnung

Die Konsistenzordnung eines numerischen Verfahrens gibt an, wie schnell sich die Näherungslösung der exakten Lösung annähert, wenn man die Zeitschritte τ_i gegen 0 gehen läßt.

Definition 5.2. Ein Runge-Kutta-Verfahren ist von der Konsistenzordnung p (kurz: von p-ter Ordnung), wenn für hinreichend glatte Probleme (5.3) die Abschätzung

$$\|\mathbf{u}(t_1) - \mathbf{u}_1\| \leq C\tau_0^{p+1}$$

gilt. □

Man berechnet die Konsistenzordnung von Runge-Kutta-Verfahren durch Taylor-Entwicklung der exakten Lösung $\mathbf{u}(t_1)$ an der Entwicklungsstelle t_0.

Beispiel 5.3. Die Taylor-Entwicklung von $\mathbf{u}(t_1)$ an der Stelle t_0 ergibt

$$\mathbf{u}(t_1) = \mathbf{u}(t_0) + \tau_0\mathbf{u}_t(t_0) + O(\tau_0^2) = \mathbf{u}_0 + \tau_0\mathbf{f}(t_0, \mathbf{u}_0) + O(\tau_0^2).$$

Für das Eulersche Polygonzugverfahren

$$\mathbf{u}_1 = \mathbf{u}_0 + \tau_0\mathbf{f}(t_0, \mathbf{u}_0)$$

folgt $\mathbf{u}(t_1) - \mathbf{u}_1 = O(\tau_0^2)$, also auch $\|\mathbf{u}(t_1) - \mathbf{u}_1\| \leq C\tau_0^2$. Das Eulersche Polygonzugverfahren ist demnach von erster Ordnung, da die Größe $\mathbf{u}(t_1) - \mathbf{u}_1$ zweiter Ordnung in τ_0 ist. □

5.3 Explizite und implizite Verfahren

Ist die Matrix $\mathbf{A}$ in Definition 5.1 eine untere Dreiecksmatrix mit Nulldiagonale, so liegt ein **explizites** Runge-Kutta-Verfahren vor. In diesem Falle kann die rechte Seite in der j-ten Gleichung von (5.5) direkt aus den linken Seiten der Gleichungen $1, \ldots, j-1$ berechnet werden. Genügt $\mathbf{A}$ jedoch nicht dieser Bedingung, so handelt es sich um ein **implizites** Verfahren. Die Gleichungen (5.5) bilden dann ein Gleichungssystem, dessen Berechnung aufwendiger ist. Klassische explizite Runge-Kutta-Verfahren sind das bereits erwähnte Eulersche Polygonzugverfahren, das von erster Ordnung ist, sowie die durch die folgenden Butcher-Tableaus gegebenen Verfahren (von links: Runge 2. Ordnung, Runge 3. Ordnung [137], Kutta 4. Ordnung [100]).

0	0	0
1/2	1/2	0
	0	1

0	0	0	0	0
1/2	1/2	0	0	0
1	0	1	0	0
1	0	0	1	0
	1/6	2/3	0	1/6

0	0	0	0	0
1/2	1/2	0	0	0
1/2	0	1/2	0	0
1	0	0	1	0
	1/6	1/3	1/3	1/6

Implizite Runge-Kutta-Verfahren wurden erstmalig im Jahre 1824 von Cauchy [38] benutzt. Er entwickelte das sogenannte **ϑ-Verfahren**

ϑ	ϑ
	1

, $\qquad 0 \leq \vartheta \leq 1.$

Für $\vartheta = 0$ liegt das explizite Euler-Verfahren, für $\vartheta = 1$ das sogenannte **implizite Euler-Verfahren** vor. Falls $\vartheta \neq 1/2$ ist, hat das ϑ-Verfahren die Ordnung 1. Für $\vartheta = 1/2$ erhält

man ein Verfahren zweiter Ordnung, die sogenannte **implizite Mittelpunktsregel**. Von Hammer und Hollingsworth [79] (1955) stammt das Verfahren dritter Ordnung

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 2/3 & 1/3 & 1/3 \\ \hline & 1/4 & 3/4 \end{array}.$$

Neben den angegebenen existiert noch eine Vielzahl weiterer Verfahren. Durch Verwendung hinreichend vieler Stufen läßt sich eine beliebig hohe Ordnung erzielen. Allerdings nimmt die Anzahl der Ordnungsbedingungen an die Koeffizienten stark zu, so daß deren Bestimmung für Verfahren höherer Ordnung komplizierter wird.

5.4 Schrittweitensteuerung

Bei einem konvergenten Verfahren verringert sich der Fehler der numerischen Lösung, wenn die Zeitschrittweiten τ_i verkleinert werden. Die Idee der Zeitschrittsteuerung ist es, τ_i so groß zu wählen, daß der lokale Fehler des Verfahrens unter einer vorgegebenen Toleranz TOL_t bleibt. Wir definieren zunächst den **lokalen Fehler** des Verfahrens:

Definition 5.4. Gegeben sei das System gewöhnlicher Differentialgleichungen

$$\mathbf{u}_t = \mathbf{f}(t, \mathbf{u}) \tag{5.6}$$

auf dem Intervall $t \in [t_0, t_e]$ mit der Anfangsbedingung $\mathbf{u}(t_0) = \mathbf{u}_0$. Die diskreten Zeitpunkte, Zeitschrittweiten und numerischen Lösungswerte seien wie oben mit t_i, τ_i und $\mathbf{u}_i$ bezeichnet. Weiterhin sei $\widetilde{\mathbf{u}}_i$ eine Lösung von (5.6), die der Anfangsbedingung $\widetilde{\mathbf{u}}_i(t_i) = \mathbf{u}_i$ genügt. Dann erhält man mit

$$\mathbf{e}_{\mathrm{loc},i} := \mathbf{u}_i - \widetilde{\mathbf{u}}_{i-1}(t_i)$$

den lokalen Fehler des Verfahrens im i-ten Zeitschritt. □

Bemerkung 5.5. Für ein Verfahren p-ter Ordnung ist der lokale Fehler von der Ordnung τ_i^{p+1}, d.h. es gilt $\|\mathbf{e}_{\mathrm{loc},i+1}\| \le C\tau_i^{p+1}$ mit einer Konstante C. □

Der lokale Fehler ist bei der praktischen Rechnung jedoch nicht bekannt. Man verwendet daher eine geeignete Schätzung dieses Fehlers. Eine Möglichkeit der Abschätzung erhält man durch Einbettung eines weiteren Verfahrens, ein Prinzip, das wir im folgenden am Beispiel der Runge-Kutta-Verfahren erläutern wollen. Wir folgen dabei im wesentlichen der Darstellung von Deuflhard/Bornemann [51, Abschnitt 5.3].

Runge-Kutta-Verfahren mit Einbettung sind Verfahren der Form (5.5), bei denen mit

$$\widehat{\mathbf{u}}_{i+1} = \mathbf{u}_i + \tau_i \sum_{l=1}^{s} \widehat{b}_l \mathbf{k}_l$$

eine weitere konsistente Lösung $\widehat{\mathbf{u}}_{i+1}$ vorliegt. Wir nehmen an, daß diese Lösung $\widehat{\mathbf{u}}_{i+1}$ bezüglich einer Norm $\|\cdot\|_*$ von geringerer Genauigkeit als $\mathbf{u}_{i+1}$ ist. Die Ordnungen von $\mathbf{u}$ und $\widehat{\mathbf{u}}$ seien p und $\widehat{p}$; es gelte $p \ge \widehat{p}$. Für die lokalen Fehler

$$\mathbf{e}_{\mathrm{loc},i+1} = \widetilde{\mathbf{u}}_i(t_{i+1}) - \mathbf{u}_{i+1} \qquad \text{und} \qquad \widehat{\mathbf{e}}_{\mathrm{loc},i+1} = \widetilde{\mathbf{u}}_i(t_{i+1}) - \widehat{\mathbf{u}}_{i+1}$$

gilt demnach

$$\vartheta := \frac{\|\mathbf{e}_{\mathrm{loc},i+1}\|_*}{\|\widehat{\mathbf{e}}_{\mathrm{loc},i+1}\|_*} < 1.$$

In Analogie zu Abschnitt 4.1 definieren wir einen Fehlerschätzer auf die folgende Weise.

Definition 5.6. Es sei $t_0 < t_1 < \cdots < t_N = t_e$ die Diskretisierung des Zeitintervalls und $\mathbf{u}_0, \ldots, \mathbf{u}_N$ die diskrete Lösung des Systems $\mathbf{u}_t = \mathbf{f}(t, \mathbf{u})$. Eine Funktion $\varepsilon : \{0, \ldots, N\} \to \mathbb{R}_+$ wird als **Fehlerschätzer** für den lokalen Fehler $\mathbf{e}_{\mathrm{loc},i}$ der diskreten Lösung bezeichnet, wenn ε bezüglich einer Norm $\|\cdot\|_*$ die folgenden Eigenschaften erfüllt:

- **Zuverlässigkeit**: Es gibt eine Konstante $C_1 > 0$, so daß $\|\mathbf{e}_{\mathrm{loc},i}\|_* \leq C_1 \varepsilon(i)$ für alle $i = 0, \ldots, N$ erfüllt ist.
- **Effizienz**: Es gibt eine Konstante $C_2 > 0$, so daß $\varepsilon(i) \leq C_2 \|\mathbf{e}_{\mathrm{loc},i}\|_*$ für alle $i = 0, \ldots, N$ erfüllt ist.

Gilt zusätzlich für $\tau := \max\{\tau_i\}$

$$\lim_{\tau \to 0} C_1, C_2 = 1,$$

so heißt der Fehlerschätzer ε **asymptotisch exakt**. □

Satz 5.7. *Die Größe*

$$\widehat{\varepsilon} := \|\widehat{\mathbf{e}}_{\mathrm{loc},i+1} - \mathbf{e}_{\mathrm{loc},i+1}\|_* \tag{5.7}$$

ist ein Fehlerschätzer für den Fehler $\widehat{\mathbf{e}}_{\mathrm{loc},i+1}$ des ungenaueren Verfahrens. Falls $\mathbf{u}$ von echt höherer Ordnung als $\widehat{\mathbf{u}}$ ist, d.h. falls $p > \widehat{p}$ gilt, so ist dieser Fehlerschätzer asymptotisch exakt. Für den Fehler $\mathbf{e}_{\mathrm{loc},i+1}$ des genaueren Verfahrens gilt

$$\|\mathbf{e}_{\mathrm{loc},i+1}\|_* \leq \frac{\vartheta}{1-\vartheta} \|\widehat{\varepsilon}_{\mathrm{loc},i+1}\|_*.$$

Beweis. Siehe Deuflhard/Bornemann [51, Abschnitt 5.3]. □

Wenn sogar $\|\mathbf{e}_{\mathrm{loc},i+1}\|_* \leq \|\widehat{\mathbf{e}}_{\mathrm{loc},i+1}\|_*/2$ gilt, so folgt aus Satz 5.7, daß der Fehler $\|\mathbf{e}_{\mathrm{loc},i+1}\|_*$ durch die Größe $\widehat{\varepsilon}$ nach oben beschränkt ist. In diesem Falle ist

$$\tau_{\mathrm{opt}} := \beta\ \tau_i \left(\frac{TOL_t}{\widehat{\varepsilon}} \right)^{1/(\widehat{p}+1)}, \qquad \beta \lessapprox 1 \tag{5.8}$$

ein Zeitschritt, der garantiert, daß der lokale Fehler $\|\mathbf{e}_{\mathrm{loc},i+1}\|_*$ unter einer vorgegebenen Toleranz TOL_t bleibt. Die Herleitung von τ_{opt} bedarf einiger regelungstechnischer Überlegungen. Wir verweisen erneut auf Deuflhard/Bornemann [51, Abschnitt 5.2.2].

Um allzu große Sprünge des Zeitschritts zu vermeiden, schlagen wir den folgenden neuen Zeitschritt vor:

$$\tau_{\mathrm{neu}} = \begin{cases} \beta_{\max}\ \tau_i, & \tau_{\mathrm{opt}} > \beta_{\max}\ \tau_i, \\ \beta_{\min}\ \tau_i, & \tau_{\mathrm{opt}} < \beta_{\min}\ \tau_i, \\ \tau_{\mathrm{opt}}, & \text{sonst.} \end{cases} \tag{5.9}$$

Die Größen $\beta_{\max} > 1$ und $\beta_{\min} < 1$ sind die Sprungbegrenzungen. Eine mögliche Wahl der Parameter ist beispielsweise $\beta = 0{,}8$, $\beta_{\max} = 2$, $\beta_{\min} = 0{,}5$.

Ist $\widehat{\varepsilon} \leq TOL_t$, so wird die Lösung $\mathbf{u}_{i+1}$ akzeptiert und für den nächsten Zeitschritt $\tau_{i+1} = \tau_{\text{neu}}$ gesetzt. Wenn aber $\widehat{\varepsilon} > TOL_t$ ist, so wird der Schritt mit $\tau_i = \tau_{\text{neu}}$ wiederholt. Der Vorgang der Zeitschrittsteuerung wird in dem folgenden Algorithmus dargestellt.

Algorithmus 5.8 (Zeitschrittsteuerung).

Vorgabe von t_0, $t_{\max}$, τ_0, TOL_t, $\mathbf{u}_0$

$i = 0$

while $t_i < t_{\max}$

Lösungsverfahren: $\mathbf{u}_i,\ \tau_i \Longrightarrow \mathbf{u}_{i+1},\ \widehat{\mathbf{u}}_{i+1}$

bestimme $\widehat{\varepsilon}$ nach (5.7), (5.9)

bestimme τ_{neu} nach (5.8), (5.9)

if $\widehat{\varepsilon} \leq TOL_t$

$t_{i+1} = t_i + \tau_i$

$i := i + 1$

end

$\tau_i = \tau_{\text{neu}}$

end □

Dieser Algorithmus garantiert, daß $\widehat{\varepsilon} \leq TOL_t$ gilt. Falls $\|\mathbf{e}_{\text{loc},i+1}\|_* \leq \|\widehat{\mathbf{e}}_{\text{loc},i+1}\|_*/2$ erfüllt ist, so folgt auch für den lokalen Fehler $\|\mathbf{e}_{\text{loc},i+1}\|_* \leq TOL_t$. Wenn jedoch

$$\vartheta = \frac{\|\mathbf{e}_{\text{loc},i+1}\|_*}{\|\widehat{\mathbf{e}}_{\text{loc},i+1}\|_*} \ll \frac{1}{2}$$

ist, dann ist $\|\mathbf{e}_{\text{loc},i+1}\|_* \ll TOL_t$, d.h. die gewünschte Kontrolle des lokalen Fehlers wird übererfüllt.

In Algorithmus 5.8 wird mit der *genaueren* Lösung $\mathbf{u}_{i+1}$ im nächsten Zeitschritt weitergerechnet. Vor allem bei historisch älteren Verfahren wurde im Gegensatz dazu die *ungenauere* Lösung $\widehat{\mathbf{u}}_{i+1}$ zur weiteren Rechnung verwendet. Ein Grund dafür war, daß die Größe $\widehat{\varepsilon}$ für $\widehat{\mathbf{u}}_{i+1}$ ein Fehlerschätzer ist, während sie für $\mathbf{u}_{i+1}$ nur eine obere Schranke des Fehlers darstellt. Verfahren, die $\mathbf{u}_{i+1}$ weiterverwenden, haben sich jedoch weitestgehend durchgesetzt, da hierbei die höhere Genauigkeit dieser Lösung ausgenutzt wird. Für die Ordnung von Verfahren mit Einbettung verwenden wir die folgende Schreibweise: Verfahren, die mit der genaueren Lösung weiterrechnen, sind von der Ordnung $p(\widehat{p})$, Verfahren, die mit der ungenaueren Lösung weiterrechnen, von der Ordnung $\widehat{p}(p)$.

Bemerkung 5.9. Bei der Definition von $\widehat{\varepsilon}$ in (5.7) können verschiedene Vektornormen $\|\cdot\|_*$ verwendet werden. Häufig benutzt man entweder eine skalierte Euklidische Vektornorm

$$\|\mathbf{v}\|_* := \sqrt{\frac{1}{n}\sum_{i=1}^{n} v_i^2} \tag{5.10}$$

oder die Maximumnorm

$$\|\mathbf{v}\|_* := \max_{i=1,\ldots,n} |v_i|.$$

Bei Verwendung der Maximumnorm kommt es nach einer Gitterverfeinerung häufig zu einem starken Einbruch des Zeitschritts, verbunden mit etlichen Zeitschrittverwerfungen, weil die dabei auftretenden lokal begrenzten Fehler die Norm deutlich beeinflussen. Dieser Effekt erhöht die Rechenzeit, und man wird deshalb oft die skalierte Euklidische Norm verwenden. Die Zeitschrittsteuerung mit der skalierten Euklidischen Norm ist jedoch mitunter anfälliger für das Auftreten räumlich lokaler Instabilitäten des Verfahrens, da räumlich lokal begrenzte Fehler in diesem Falle nicht so stark bestraft werden. Da kann zu irreparabel schlechten Lösungen und zum Abbruch des Verfahrens führen. Bei Problemen, die zu instabilem Verhalten neigen, kann deshalb die Verwendung der Maximumnorm eine Alternative sein. In den in dieser Arbeit durchgeführten numerischen Beispielen wird, wenn es nicht anders vermerkt ist, die skalierte Euklidische Norm verwendet. □

Ein explizites Runge-Kutta-Verfahren mit Einbettung wurde erstmalig 1957 von Merson [117] konstruiert. Es ist durch das Butcher-Tableau

0	0	0	0	0	0
1/3	1/3	0	0	0	0
1/3	1/6	1/6	0	0	0
1/2	1/8	0	3/8	0	0
1	1/2	0	−3/2	2	0
u_{i+1}	1/6	0	0	2/3	1/6
$\widehat{u}_{i+1}$	1/10	0	3/10	2/5	1/5

gegeben und von vierter Ordnung. Das eingebettete Verfahren ist i.a. dritter Ordnung, im Falle linearer Gleichungen mit konstanten Koeffizienten sogar fünfter Ordnung. Häufig benutzt werden die in den Jahren 1968 und 1969 vorgestellten Verfahren von Fehlberg [62, 63] (Ordnung 7(8), 13-stufig und Ordnung 4(5), 6-stufig) sowie die Verfahren von Dormand und Prince [52, 53] aus den Jahren 1980 und 1981 (Ordnung 5(4), 7-stufig und Ordnung 8(7), 13-stufig). Die Koeffizienten dieser Verfahren können den Büchern von Hairer/Nørsett/Wanner [77] und Deuflhard/Bornemann [51] entnommen werden.

Auch in impliziten Runge-Kutta-Verfahren wird Einbettung zur Schrittweitensteuerung verwendet. Ein Beispiel hierfür ist das Verfahren RADAU5 von Hairer und Wanner [78], das auf einem impliziten Verfahren von Ehle [55] aus dem Jahre 1969 basiert. Dieses Verfahren ist dreistufig und hat die Ordnung 5(4).

5.5 Stabilität

Viele explizite und einige implizite Verfahren zeigen nur dann ein stabiles Verhalten, wenn die Zeitschrittweite hinreichend klein gewählt wurde. Stabilität ist jedoch eine notwendige Voraussetzung dafür, überhaupt sinnvolle numerische Näherungslösungen zu erhalten. Instabile Verfahren führen auf oszillierende numerische Lösungen, häufig mit schnell anwachsender Amplitude, die keine Aussage über die exakte Lösung mehr gestatten. Eine Analyse instabiler Probleme wurde erstmals – für hyperbolische Differentialgleichungen – von Courant, Friedrichs und Lewy [43] im Jahre 1928 vorgenommen.

Der Begriff der Stabilität ist sowohl auf die Differentialgleichung selbst, als auch auf das numerische Verfahren anwendbar. Unter Stabilität verstehen wir die Eigenschaft, daß zwei benachbarte Lösungskurven in beschränktem Abstand voneinander bleiben. Betrachten wir zunächst ein lineares autonomes System $\mathbf{u}_t = \mathbf{A}\mathbf{u}$ und zwei Lösungen $\mathbf{u}$ und $\mathbf{v}$ zu den Anfangswerten $\mathbf{u}(t_0) = \mathbf{u}_0$ und $\mathbf{v}(t_0) = \mathbf{v}_0$. Die Differenz der beiden Lösungen $\mathbf{w} := \mathbf{u} - \mathbf{v}$ erfüllt dann die Gleichung $\mathbf{w}_t(t) = \mathbf{A}\mathbf{w}(t)$. Das Verhalten dieses linearen Systems wird wesentlich durch die Eigenwerte von $\mathbf{A}$ bestimmt. Wir nennen das Problem **stabil**, wenn alle Eigenwerte λ_i der Matrix $\mathbf{A}$ die Bedingung $\text{Re}\ \lambda_i \leq 0$ und die mehrfachen Eigenwerte λ_j sogar $\text{Re}\ \lambda_i < 0$ erfüllen.

Bei nichtlinearen Systemen reicht eine derartige Bedingung an die Eigenwerte der Jacobi-Matrix *nicht* aus, um Stabilität zu erhalten. Die Stabilitätstheorie nichtlinearer Probleme wurde 1877 von Routh [136] und Poincaré begründet etwas später in einer berühmten Arbeit von Ljapunov [108] weiterentwickelt. Mit Hilfe sogenannter Ljapunov-Funktionen erhält man hier eine hinreichende Bedingung für Stabilität. Für einen kurzen Überblick über diese Thematik verweisen wir auf das Buch von Hairer, Nørsett und Wanner [77].

Zur Untersuchung der Stabilität eines numerischen Verfahrens führte Dahlquist die Testgleichung $u_t = \lambda u,\ u(0) = 1$ ein. Die Lösung eines Runge-Kutta-Verfahrens nach einem Zeitschritt kann in der Form

$$u_1 = R(\lambda\tau_0)$$

geschrieben werden.

Definition 5.10. Die rationale Funktion R wird als **Stabilitätsfunktion** des Verfahrens bezeichnet. Das Gebiet $S = \{z \in \mathbb{C} : |R(z)| \leq 1\}$ heißt **Stabilitätsgebiet** des Verfahrens. □

Falls $\lambda\tau_0 \in S$ ist, so folgt $|u_1| \leq |u_0| = 1$, d.h. der Betrag der numerischen Lösung der Testgleichung nimmt ab.

Im Falle eines linearen autonomen Systems $\mathbf{u}_t = \mathbf{A}\mathbf{u}$ ist es sinnvoll, die folgende Stabilitätsforderung an ein numerisches Verfahren zu stellen:

Stabilitätsforderung

Es gelte

$$\tau_i \lambda_j(\mathbf{A}) \in S \tag{5.11}$$

für alle Eigenwerte $\lambda_j(\mathbf{A})$, die $\text{Re}\ \lambda_j(\mathbf{A}) \leq 0$ erfüllen.

Die Stabilitätsforderung kann wie folgt interpretiert werden: In Eigenrichtungen, in denen die exakte Lösung stabil ist, soll auch die numerische Lösung stabiles Verhalten zeigen.

Die Stabilitätsforderung kann zu einer notwendigen Beschränkung des Zeitschrittes führen. Wir bezeichnen mit $\tau_{\text{stab},i}$ den nach der Stabilitätsforderung maximal zulässigen Zeitschritt τ_i. Wird die Stabilitätsbedingung nur sehr knapp erfüllt, was der Fall ist, wenn Eigenwerte nahe am Rand von S liegen, so kann es durchaus noch zu störenden Oszillationen der numerischen Lösung kommen. Durch die Stabilitätsbedingung werden nur Oszillationen mit wachsender Amplitude verhindert. Das unten angegebene Beispiel 5.16 illustriert diesen Sachverhalt. Man muß also gegebenenfalls den Zeitschritt stärker beschränken, als es die Bedingung vorschreibt.

Verfahren, bei denen die Stabilitätsforderung (5.11) niemals auf eine Zeitschrittbeschränkung führt, werden A-stabile Verfahren genannt. Die folgende Definition geht auf DAHLQUIST [45] (1963) zurück:

Definition 5.11. Ein numerisches Verfahren heiße **A-stabil**, wenn die linke Halbebene $\{z \in \mathbb{C} : \operatorname{Re} z \leq 0\}$ ganz im Stabilitätsgebiet S des Verfahrens enthalten ist. □

Eine etwas schwächere Bedingung wird für die sogenannte A(α)-Stabilität gestellt:

Definition 5.12 (Widlund [171] (1967)). Sei $\alpha \in]0, \pi/2[$. Ein numerisches Verfahren heißt **A(α)-stabil**, falls der Ausschnitt $\{z \in \mathbb{C} : |\arg(-z)| < \alpha\}$ ganz in S enthalten ist. □

Eine weitere wünschenswerte Eigenschaft eines numerischen Verfahrens ist, daß Komponenten des linearen autonomen Systems, die sehr schnell gegen 0 gehen, die also den Eigenwerten mit stark negativem Realteil entsprechen, auch in der numerischen Lösung schnell gedämpft werden. Ein solches Verhalten zeigen L-stabile Verfahren, die wie folgt definiert sind:

Definition 5.13 (Ehle [55] (1969)). Ein numerisches Verfahren heißt **L-stabil**, falls es A-stabil ist und zusätzlich der Bedingung $\lim_{z\to\infty} R(z) = 0$ genügt. □

Bemerkung 5.14. Für eine rationale Funktion R gilt $\lim_{z\to\infty} R(z) = \lim_{z\to-\infty} R(z)$. □

Beispiel 5.15. Die folgende Darstellung zeigt Stabilitätsfunktion und Stabilitätsgebiet des expliziten und impliziten Euler-Verfahrens sowie der impliziten Trapezregel.

Euler explizit:	$\begin{array}{c\|c} 0 & 0 \\ \hline & 1 \end{array}$,	$R(z) = 1 + z$,	$S = \{z \in \mathbb{C} : \lvert z + 1\rvert \leq 1\}$,	nicht A-stabil
Euler implizit:	$\begin{array}{c\|c} 1 & 1 \\ \hline & 1 \end{array}$,	$R(z) = \dfrac{1}{1-z}$,	$S = \{z \in \mathbb{C} : \lvert z - 1\rvert \geq 1\}$,	L-stabil
implizite Trapezregel:	$\begin{array}{c\|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}$,	$R(z) = \dfrac{2+z}{2-z}$,	$S = \{z \in \mathbb{C} : \operatorname{Re} z \leq 0\}$,	A-stabil, nicht L-stabil

□

Beispiel 5.16. Zur Untersuchung der Stabilität verschiedener Verfahren konstruierten CURTISS und HIRSCHFELDER [44] 1952 die Differentialgleichung

$$u_t = -50(u - \cos t), \qquad u(0) = 0,$$

deren Lösung durch

$$u(t) = -\frac{2500}{2501}e^{-50t} + \frac{2500}{2501}\cos t + \frac{50}{2501}\sin t$$

gegeben ist. Nach einer kurzen transienten Phase, die durch den Exponentialterm bestimmt wird, nähert sich die Lösung der Funktion

$$g(t) = \frac{2500}{2501}\cos t + \frac{50}{2501}\sin t$$

an. Abbildung 5.1 zeigt die Lösung der drei Verfahren aus Beispiel 5.15 mit konstanter Zeitschrittweite.

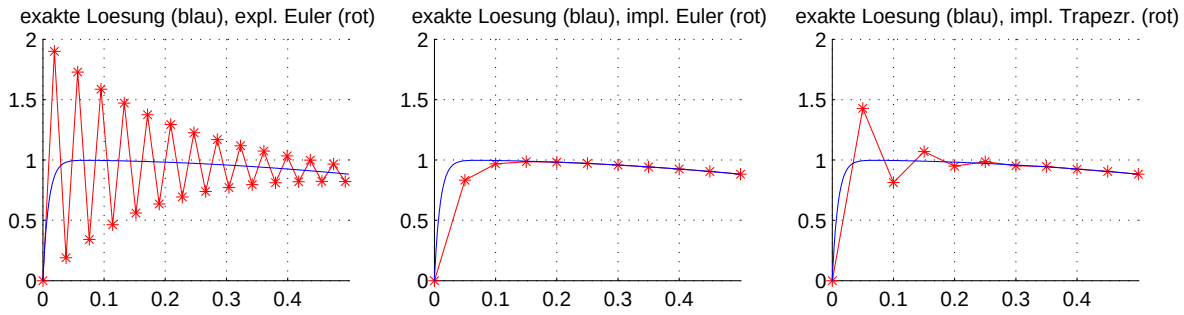


Abbildung 5.1: Lösung des Problems von Curtiss & Hirschfelder, **von links:** expl. Euler ($\tau = 0.019$), impl. Euler ($\tau = 0.05$), impl. Trapezregel ($\tau = 0.05$)

Das explizite Eulerverfahren erfüllt gerade noch die Stabilitätsbedingung, trotzdem ist die Lösung stark oszillierend. Die Lösung der Trapezregel zeigt ebenfalls Oszillationen, da diese nicht L-stabil ist. Das L-stabile implizite Eulerverfahren ist zwar anfangs etwas ungenau, oszillert aber nicht. □

Im allgemeinen zeigen implizite Verfahren bessere Stabilitätseigenschaften als explizite. Viele implizite Verfahren sind A-stabil. Für Probleme mit betraglich großen Eigenwerten, die negativen Realteil haben, sind deshalb oftmals explizite Verfahren nicht mehr effizient, da deren Stabilitätsforderung den Zeitschritt zu stark einschränkt. Hier müssen stabile implizite Verfahren verwendet werden. Probleme dieser Art werden als steif bezeichnet. Wir werden im folgenden Abschnitt näher auf das Phänomen der Steifheit gewöhnlicher Differentialgleichungen eingehen.

5.6 Steifheit

Bei einigen Differentialgleichungen führen die Stabilitätsbedingungen expliziter numerischer Verfahren zu einer starken Einschränkung des Zeitschrittes, obwohl aus der Sicht der benötigten Genauigkeit der Zeitschritt größer gewählt werden könnte. Eine solche Problematik wird als **Steifheit** der Differentialgleichung bezeichnet. Auf mathematisch exakte Weise ist der Begriff der Steifheit schwer zu fassen; es gibt viele Größen, von denen letztendlich abhängt, ob man ein Problem als steif bezeichnet oder nicht. Insbesondere sind dies

- die Stabilitätseigenschaften des Systems,
- die gewünschte Genauigkeit, repräsentiert durch die Fehlertoleranz TOL_t und
- die verwendeten expliziten und impliziten Verfahren, deren Zeitschrittweiten man miteinander vergleicht.

Aus diesem Grunde gibt es in der Literatur auch keine einheitliche Definition der Steifheit sondern eher eine Vielzahl verbaler Beschreibungen des Phänomens. Die historisch älteste geht auf Curtiss und Hirschfelder [44] (1952) zurück und lautet: „Stiff equations are equations

where certain implicit methods, in particular BDF, perform better, usually tremendously better, than explicit ones."[1]

Differentialgleichungen, die aus der Ortsdiskretisierung parabolischer Gleichungen hervorgehen, sind häufig derartige steife Probleme. Wir wollen daher im folgenden Abschnitt speziell auf die Steifheit linearer autonomer Reaktion-Diffusions-Gleichungen eingehen.

5.6.1 Steifheit linearer autonomer Reaktions-Diffusions-Gleichungen

Die Steifheit von Reaktions-Diffusions-Systemen kann Ursachen im Diffusionsanteil und in der Reaktionsfunktion haben. Wir wollen das Auftreten von Steifheit exemplarisch an einer skalaren semilinearen Gleichung der Form

$$\begin{aligned} \frac{\partial u}{\partial t}(\mathbf{x},t) &= \nabla \cdot (d(\mathbf{x})\nabla u(\mathbf{x},t)) + r(\mathbf{x})(u(\mathbf{x},t)), \\ u(\mathbf{x},0) &= u_0(\mathbf{x}). \end{aligned} \tag{5.12}$$

studieren. Wir nehmen Neumannsche Randbedingungen an:

$$d(\mathbf{x})\mathbf{n}_{\partial\Omega} \cdot \nabla u(\mathbf{x},t) = g_{\text{Neu}}(\mathbf{x},t).$$

Nach der Ortsdiskretisierung durch lineare finite Elemente mit reduzierter Massenmatrix liegt ein System gewöhnlicher Differentialgleichungen der Form

$$\mathbf{u}_t = -\mathbf{L}^{-1}\mathbf{S}\mathbf{u} + \text{diag}(\mathbf{r})\mathbf{u} + \mathbf{L}^{-1}\mathbf{B}\mathbf{g}_{\text{Neu}} =: \mathbf{f}(\mathbf{u})$$

vor, vgl. (3.15). Dabei ist $\mathbf{r}$ der Vektor der Funktionswerte von r in den Gitterpunkten $\mathbf{x}_i$. Die übrigen Vektoren und Matrizen sind wie in Abschnitt 3.1.4 definiert.

Um das Stabilitätsverhalten der Differentialgleichung zu untersuchen, benötigen wir eine Aussage über die Eigenwerte der Jacobi-Matrix $\mathbf{J} := \partial\mathbf{f}/\partial\mathbf{u} = -\mathbf{L}^{-1}\mathbf{S} + \text{diag}(\mathbf{r})$. Wir zeigen zunächst, daß die Eigenwerte dieser Matrix sämtlich reell sind und schätzen dann die Größenordnung der Eigenwerte ab.

Zunächst benötigen wir das folgende Lemma:

Lemma 5.17. *Sind* $\mathbf{A}$ *und* $\mathbf{B}$ *symmetrische Matrizen und ist* $\mathbf{B}$ *zusätzlich positiv oder negativ definit, so sind die durch die Gleichung* $\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}$ *gegebenen Eigenwerte des Matrizenpaares* $\mathbf{A};\mathbf{B}$ *sämtlich reell.*

Beweis. Der Beweis wird hier nicht angegeben. Wir verweisen auf das Buch von Zurmühl und Falk [178, Abschnitt 15.2]. □

Ausgerüstet mit diesem Lemma, können wir jetzt den folgenden Satz beweisen:

Satz 5.18. *Für skalare Reaktions-Diffusions-Gleichungen sind die Eigenwerte der Jacobi-Matrix* $\mathbf{J} = \partial\mathbf{f}/\partial\mathbf{u} = -\mathbf{L}^{-1}\mathbf{S} + \text{diag}(\mathbf{r})$ *sämtlich reell.*

[1] „Steife Gleichungen sind Gleichungen, bei denen gewisse implizite Verfahren, insbesondere BDF-Verfahren, besser – oft wesentlich besser – abschneiden als explizite."

Beweis. Nach Konstruktion ist $\mathbf{S}$ eine symmetrische Matrix. $\mathbf{L}$ ist eine Diagonalmatrix mit sämtlich positiven Diagonalelementen, also positiv definit. Die Matrix $\operatorname{diag}(\mathbf{r})$ ist ebenfalls Diagonalmatrix. Die Eigenwerte von $\mathbf{J}$ genügen der Gleichung

$$(-\mathbf{S} + \mathbf{L}\operatorname{diag}(\mathbf{r}))\,\mathbf{x} = \lambda \mathbf{L}\mathbf{x}.$$

Die Matrix $(-\mathbf{S} + \mathbf{L}\operatorname{diag}(\mathbf{r}))$ ist symmetrisch. Aus Lemma 5.17 folgt nun, daß die Eigenwerte sämtlich reell sind. □

Bemerkung 5.19. Für nichtskalare Systeme von Reaktions-Diffusions-Gleichungen geht die Symmetrie der Jacobi-Matrix verloren. Deshalb können dann auch konjugiert komplexe Eigenwerte auftreten. □

Eine grobe Lokalisierung der Eigenwerte gewinnt man mit dem Satz von Gerschgorin [72] (1931), siehe etwa Zurmühl/Falk [179, Abschnitt 36.2]:

Satz 5.20 (Gerschgorin [72] (1931)). *Die n Eigenwerte der Matrix $\mathbf{A} = (a_{ij})_{i,j=1,\dots,n}$, die durch die Gleichung $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ gegeben sind, liegen in der Vereinigungsmenge der n Kreise*

$$|\lambda - a_{ii}| \le \min\left\{ \sum_{j=1,\dots,n,\ j\neq i} |a_{ij}|, \sum_{j=1,\dots,n,\ j\neq i} |a_{ji}| \right\}, \qquad i = 1,\dots,n.$$

Wir diskutieren zunächst das Auftreten von Steifheit im Spezialfall $r \equiv 0$, also für die Wärmeleitungsgleichung.

Steifheit der Wärmeleitungsgleichung

Für $d = \text{const.} > 0$ und $r \equiv 0$ in (5.12) erhält man die Wärmeleitungsgleichung mit konstantem Diffusionskoeffizienten. Der kleinste Eigenwert $\lambda_{\min}(\mathbf{J})$ der Jacobi-Matrix bestimmt das Stabilitätsverhalten. Der Satz von Gerschgorin liefert die Eingrenzung

$$-2\max\{\left(\mathbf{L}^{-1}\mathbf{S}\right)_{ii}\} \le \lambda_{\min}(\mathbf{J}) \le 0.$$

Miller [119] gibt die schärfere Einschränkung

$$\lambda_{\min}(\mathbf{J}) \le -\max\{\left(\mathbf{L}^{-1}\mathbf{S}\right)_{ii}\}$$

an. Der Ausdruck $\max\{\left(\mathbf{L}^{-1}\mathbf{S}\right)_{ii}\}$ wird im wesentlichen durch die Größe des kleinsten Elements der Triangulierung bestimmt. Für ein Gitter aus gleichseitigen Dreiecken mit Seitenlänge h gilt etwa

$$\left(\mathbf{L}^{-1}\mathbf{S}\right)_{ii} = \frac{4}{h^2}.$$

Folglich ist $-\lambda_{\min}(\mathbf{J})$ proportional zu $1/h^2$, d.h. für nicht A-stabile Verfahren ist der maximal zulässige Zeitschritt τ_{stab} proportional zu h^2. Bei starker Gitterverfeinerung kann τ_{stab} demnach sehr klein werden; es liegt dann ein steifes Problem vor und man sollte ein A-stabiles Verfahren verwenden.

Steifheit linearer autonomer Reaktions-Diffusions-Gleichungen

Der Satz von Gerschgorin liefert die Abschätzungen

$$\min\{-2\left(\mathbf{L}^{-1}\mathbf{S}\right)_{ii} + r(\mathbf{x}_i)\} \le \lambda_{\min}(\mathbf{J}) \le 0.$$

Zusätzlich zur Steifheit durch Diffusion können demnach negative Werte von $p'(u_i)$ zu einer weiteren Verkleinerung der Eigenwerte der Jacobi-Matrix $\mathbf{J}$ führen. Die Steifheit von Reaktions-Diffusions-Gleichungen kann somit eine Folge der Gitterverfeinerung und negativer Funktionswerte der Reaktionsfunktion r sein.

5.7 W-Methoden

Wir betrachten das System gewöhnlicher Differentialgleichungen

$$\mathbf{u}_t = \mathbf{f}(t, \mathbf{u}), \qquad \mathbf{u}(t_0) = \mathbf{u}_0.$$

In dem in (5.1) angegebenen s-stufigen Runge-Kutta-Verfahren müssen in jedem Zeitschritt s nichtlineare Gleichungssysteme gelöst werden, falls die zugrundeliegende Differentialgleichung nichtlinear ist. Häufig wird hierfür das Newton-Verfahren verwendet. Eine Alternative dazu sind **linear-implizite Runge-Kutta-Verfahren**, bei denen die Nichtlinearität nur explizit auftritt, also nicht auf ein nichtlineares Gleichungssystem führt. Linear-implizite Verfahren wurden erstmals 1963 von Rosenbrock [135] formuliert. Bei diesen Verfahren geht die Jacobi-Matrix $\partial\mathbf{f}/\partial\mathbf{u}$ in das lineare Gleichungssystem ein, sie muß also in jedem Zeitschritt neu berechnet werden. Später wurden von Steihaug und Wolfbrandt [152] die sogenannten **W-Methoden** entwickelt.

Definition 5.21. Eine s-stufige W-Methode mit Einbettung ist durch die Vorschrift

$$\begin{aligned}
\mathbf{k}_j &= \mathbf{f}\left(t_i + c_j\tau_i, \mathbf{u}_i + \tau_i \sum_{l=1}^{j-1} \alpha_{jl}\mathbf{k}_l\right) + \tau_i \mathbf{T} \sum_{l=1}^{j} \gamma_{jl}\mathbf{k}_l, \qquad && j = 1, \dots, s, \qquad (5.13)\\
\gamma_{jj} &= \gamma, && j = 1, \dots, s,\\
\mathbf{u}_{i+1} &= \mathbf{u}_i + \tau_i \sum_{l=1}^{s} b_l \mathbf{k}_l,\\
\widehat{\mathbf{u}}_{i+1} &= \mathbf{u}_i + \tau_i \sum_{l=1}^{s} \widehat{b}_l \mathbf{k}_l
\end{aligned}$$

gegeben. Die Koeffizienten werden in den Matrizen $\mathbf{A}$, $\mathbf{\Gamma}$ und den Vektoren $\mathbf{b}$, $\widehat{\mathbf{b}}$ und $\mathbf{c}$ zusammengefaßt. □

Die Matrix $\mathbf{T}$ kann beliebig gewählt werden, ohne daß sich die Ordnung des Verfahrens ändert, allerdings ergibt sich die größte Genauigkeit und die besten Stabilitätseigenschaften, wenn $\mathbf{T}$ eine Approximation an die Jacobi-Matrix $\mathbf{J} = \partial\mathbf{f}/\partial\mathbf{u}$ ist. Das Verfahren mit $\mathbf{T} = \mathbf{J}$ bezeichnen wir als **W-Methode mit Jacobi-Matrix**. Insbesondere entspricht die W-Methode (5.21)

für ein *lineares* System $\mathbf{u}_t = \mathbf{B}\mathbf{u}$ mit konstanten Koeffizienten gerade dem durch das Butcher-Tableau

$$\begin{array}{c|cccc} \gamma & 0 & \dots & 0 \\ a_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ a_{s1} & \dots & a_{s,s-1} & \gamma \\ \hline b_1 & \dots & \dots & b_s \end{array}, \qquad a_{ij} = \alpha_{ij} + \gamma_{ij}, \quad j < i \le s$$

gegebenen impliziten Runge-Kutta-Verfahren, falls $\mathbf{T} = \mathbf{J}$ ist. Ist dieses implizite Verfahren A-stabil, so ist die W-Methode für das lineare Problem sogar unbedingt stabil.

Die ursprüngliche Idee bei der Entwicklung der W-Methoden war, die Jacobi-Matrix zu verwenden, sie jedoch über einige Zeitschritte konstant zu lassen. W-Methoden bieten jedoch auch einen geeigneten Ausgangspunkt zur Konstruktion von Partitionierungs-Verfahren. Da für $\mathbf{T} = \mathbf{0}$ ein explizites Runge-Kutta-Verfahren vorliegt, so kann durch Nullsetzen gewisser Untermatrizen von $\mathbf{T}$ ein Verfahren gebildet werden, das *nur bezüglich einiger Komponenten* implizit ist. Dieser Ansatz führt auf lokale Partitionierungs-Verfahren, die in Kapitel 8 vorgestellt werden sollen. Auch die in Kapitel 7 beschriebenen Krylov-W-Verfahren lassen sich als W-Verfahren mit einer speziellen Matrix $\mathbf{T}$ interpretieren.

Bei der praktischen Implementierung wird in der Regel nicht die Form (5.13) verwendet. Durch geeignete lineare Transformation der $\mathbf{k}_i$ kann man nämlich ein äquivalentes System erzeugen, das sich mit deutlich geringerem Aufwand lösen läßt. Wir bilden dafür die Matrizen $\mathbf{K} = (\mathbf{k}_1, \dots, \mathbf{k}_s)$, $\mathbf{U}_i = (\mathbf{u}_i, \dots, \mathbf{u}_i)$, $\widehat{\mathbf{U}}_i = (\widehat{\mathbf{u}}_i, \dots, \widehat{\mathbf{u}}_i)$ [2] und den Vektor $\mathbf{e} = (1, \dots, 1)^T$. Das System (5.13) kann damit in Matrixform als

$$\begin{aligned} \mathbf{K} &= \mathbf{f}(t_i\mathbf{e}^T + \tau_i\mathbf{c}^T, \mathbf{U}_i + \tau_i\mathbf{K}\mathbf{A}^T) + \tau_i\mathbf{T}\mathbf{K}\mathbf{\Gamma}^T, \\ \mathbf{u}_{i+1} &= \mathbf{u}_i + \tau_i\mathbf{K}\mathbf{b}, \\ \widehat{\mathbf{u}}_{i+1} &= \mathbf{u}_i + \tau_i\mathbf{K}\widehat{\mathbf{b}} \end{aligned} \tag{5.14}$$

geschrieben werden. Die lineare Transformation

$$\widetilde{\mathbf{K}} = \frac{1}{\gamma}\mathbf{K}\mathbf{\Gamma}^T \tag{5.15}$$

führt auf das äquivalente System

$$\begin{aligned} (\mathbf{I} - \tau_i\gamma\mathbf{T})\widetilde{\mathbf{K}} &= \mathbf{f}(t_i\mathbf{e}^T + \tau_i\mathbf{c}^T, \mathbf{U}_i + \tau_i\gamma\widetilde{\mathbf{K}}\mathbf{\Gamma}^{-T}\mathbf{A}^T) + \widetilde{\mathbf{K}}(\mathbf{I} - \gamma\mathbf{\Gamma}^{-T}), \\ \mathbf{u}_{i+1} &= \mathbf{u}_i + \tau_i\gamma\widetilde{\mathbf{K}}\mathbf{\Gamma}^{-T}\mathbf{b}, \\ \widehat{\mathbf{u}}_{i+1} &= \mathbf{u}_i + \tau_i\gamma\widetilde{\mathbf{K}}\mathbf{\Gamma}^{-T}\widehat{\mathbf{b}}. \end{aligned} \tag{5.16}$$

Wegen der Dreiecksgestalt von $\mathbf{A}$ und $\mathbf{\Gamma}$ sind die Matrizen $\mathbf{\Gamma}^{-T}\mathbf{A}^T$ und $\mathbf{I} - \gamma\mathbf{\Gamma}^{-T}$ *obere Dreiecksmatrizen mit Null-Diagonale*. Wir führen die Bezeichnungen

$$\mathbf{\Phi} = \begin{pmatrix} 0 & \varphi_{12} & \dots & \varphi_{1s} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \varphi_{s-1,s} \\ 0 & \dots & \dots & 0 \end{pmatrix} := \mathbf{\Gamma}^{-T}\mathbf{A}^T,$$

[2] Diese Schreibweise bedeutet, daß die Matrizen aus den entsprechenden Spaltenvektoren zusammengesetzt sind.

$$\mathbf{\Theta} = \begin{pmatrix} 0 & \vartheta_{12} & \dots & \vartheta_{1s} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \vartheta_{s-1,s} \\ 0 & \dots & \dots & 0 \end{pmatrix} := \mathbf{I} - \gamma \mathbf{\Gamma}^{-T},$$

$$\mathbf{g} = (g_1, \dots, g_s)^T := \gamma \mathbf{\Gamma}^{-T} \mathbf{b}$$

und

$$\widehat{\mathbf{g}} = (\widehat{g}_1, \dots, \widehat{g}_s)^T := \gamma \mathbf{\Gamma}^{-T} \widehat{\mathbf{b}}$$

ein. Das System (5.16) hat nun die Form

$$\begin{aligned} (\mathbf{I} - \tau_i \gamma \mathbf{T}) \widetilde{\mathbf{k}}_j &= \mathbf{f}\left(t_i + \tau_i c_j, \mathbf{u}_i + \tau_i \gamma \sum_{l=1}^{j-1} \varphi_{lj} \widetilde{\mathbf{k}}_l \right) + \sum_{l=1}^{j-1} \vartheta_{lj} \widetilde{\mathbf{k}}_l, \qquad j = 1, \dots, s, \quad (5.17) \\ \mathbf{u}_{i+1} &= \mathbf{u}_i + \tau_i \sum_{l=1}^{s} g_l \widetilde{\mathbf{k}}_l, \\ \widehat{\mathbf{u}}_{i+1} &= \mathbf{u}_i + \tau_i \sum_{l=1}^{s} \widehat{g}_l \widetilde{\mathbf{k}}_l. \end{aligned}$$

Dabei ist $\widetilde{\mathbf{k}}_j$ die j-te Spalte der Matrix $\widetilde{\mathbf{K}}$. Das System zerfällt also in s lineare Gleichungssysteme, die *nacheinander gelöst* werden können. Ein weiterer Vorteil gegenüber dem System (5.14) ist die Tatsache, daß die Multiplikation der i.a. sehr großen Matrizen $\mathbf{T}$ und $\mathbf{K}$ auf der rechten Seite vermieden wird.

Beispiel 5.22. Die dreistufige W-Methode mit den Koeffizienten

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad \mathbf{\Gamma} = \begin{pmatrix} 1 - \frac{1}{2}\sqrt{2} & 0 & 0 \\ -1 & 1 - \frac{1}{2}\sqrt{2} & 0 \\ -\frac{1}{2}\sqrt{2} & -2 + \frac{3}{2}\sqrt{2} & 1 - \frac{1}{2}\sqrt{2} \end{pmatrix},$$

$$\mathbf{c} = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 1/2 \\ 0 \\ 1/2 \end{pmatrix}, \qquad \widehat{\mathbf{b}} = \begin{pmatrix} 9/20 \\ -\sqrt{2}/20 \\ (11 + \sqrt{2})/20 \end{pmatrix}$$

wurde von Schmitt und Weiner [143] zur Konstruktion eines Krylov-W-Verfahrens (siehe (7.2)) verwendet. Sie ist L-stabil [154] und von zweiter Ordnung mit Einbettung erster Ordnung. Mit der Transformation (5.15) ergibt sich das Verfahren

$$\begin{aligned} \left(\mathbf{I} - \tau_i \left(1 - \sqrt{2}/2\right) \mathbf{T}\right) \widetilde{\mathbf{k}}_1 &= \mathbf{f}(t_i, \mathbf{u}_i), \qquad (5.18) \\ \left(\mathbf{I} - \tau_i \left(1 - \sqrt{2}/2\right) \mathbf{T}\right) \widetilde{\mathbf{k}}_2 &= \mathbf{f}\left(t_{i+1}, \mathbf{u}_i + \tau_i \widetilde{\mathbf{k}}_1\right) - \left(2 + \sqrt{2}\right) \widetilde{\mathbf{k}}_1, \\ \left(\mathbf{I} - \tau_i \left(1 - \sqrt{2}/2\right) \mathbf{T}\right) \widetilde{\mathbf{k}}_3 &= \mathbf{f}\left(t_{i+1}, \mathbf{u}_i + \tau_i \widetilde{\mathbf{k}}_1\right) - \widetilde{\mathbf{k}}_1 + \left(-1 + \sqrt{2}\right) \widetilde{\mathbf{k}}_2, \\ \mathbf{u}_{i+1} &= \mathbf{u}_i + \frac{\tau_i}{2}\left(2\widetilde{\mathbf{k}}_1 + \left(1 - \sqrt{2}\right) \widetilde{\mathbf{k}}_2 + \widetilde{\mathbf{k}}_3\right), \\ \widehat{\mathbf{u}}_{i+1} &= \mathbf{u}_i + \frac{\tau_i}{20}\left(\left(18 - \sqrt{2}\right) \widetilde{\mathbf{k}}_1 + \left(9 - 11\sqrt{2}\right) \widetilde{\mathbf{k}}_2 + \left(11 + \sqrt{2}\right) \widetilde{\mathbf{k}}_3\right). \end{aligned}$$

Setzt man $\mathbf{T} = \mathbf{0}$, so erhält man aus (5.14) das zugehörige explizite Runge-Kutta-Verfahren. In unserem Fall ergibt sich $\mathbf{k}_2 = \mathbf{k}_3$, so daß das explizite Verfahren nur zweistufig ist:

$$\begin{aligned} \mathbf{k}_1 &= \mathbf{f}(t_i, \mathbf{u}_i), \\ \mathbf{k}_2 &= \mathbf{f}(t_{i+1}, \mathbf{u}_i + \tau_i \mathbf{k}_1), \\ \mathbf{u}_{i+1} &= \mathbf{u}_i + \frac{\tau_i}{2}(\mathbf{k}_1 + \mathbf{k}_2), \\ \widehat{\mathbf{u}}_{i+1} &= \mathbf{u}_i + \frac{\tau_i}{20}(9\mathbf{k}_1 + 11\mathbf{k}_2). \end{aligned} \tag{5.19}$$

□

Die Zeitschrittsteuerung von W-Methoden mit Einbettung erfolgt in der gleichen Weise wie bei Runge-Kutta-Verfahren. Wenn das zu lösende System gewöhnlicher Differentialgleichungen aus der Semidiskretisierung einer partiellen Differentialgleichung hervorgeht, so haben die auftretenden linearen Gleichungssysteme dünnbesetzte Matrizen. Liegt eine räumlich eindimensionale partielle Differentialgleichung vor, so ist, wegen der Bandstruktur der Systemmatrix, oft das Gaußsche Eliminationsverfahren zur Lösung das geeignetste. Bei räumlich mehrdimensionalen Problemen sind in der Regel iterative Gleichungslöser effizienter. Im nächsten Kapitel gehen wir auf verschiedene iterative Verfahren zur Lösung linearer Gleichungssystemen ein.

Kapitel 6

Iterative Lösung dünnbesetzter linearer Gleichungssysteme

6.1 Überblick über verschiedene Verfahren

Bei der Diskretisierung parabolischer Differentialgleichungen treten lineare Gleichungssysteme $\mathbf{Ax} = \mathbf{b}$ mit einer dünnbesetzten Matrix $\mathbf{A}$ auf, d.h. nur ein geringer Teil der Elemente von $\mathbf{A}$ ist von 0 verschieden. Zur Lösung derartiger Systeme können **direkte** oder **iterative Verfahren** eingesetzt werden. Bei den iterativen Verfahren unterscheidet man wiederum zwischen **stationären** und **instationären Verfahren**. Direkte und stationäre iterative Verfahren sind beispielsweise in Hämmerlin/Hoffmann [80] beschrieben. Eine Übersichtsdarstellung iterativer Verfahren findet sich in Barrett et al. [19], eine ausführlichere Beschreibung in Hackbusch [76].

Direkte Verfahren liefern, abgesehen von eventuellen Rundungsfehlern, die exakte Lösung des Gleichungssystems. Das klassische direkte Verfahren ist das Gaußsche Eliminationsverfahren, welches erstmalig in einer Arbeit von Gauss aus dem Jahre 1810 erschien, in der die Störungen des Planetoiden Pallas untersucht wurden [69, Bd. VI, S. 3-24, Bd. VII, S. 307-308]. Außerdem gehören Cholesky- und Householder-Verfahren zu dieser Gruppe.

Iterative Verfahren werden vor allem zur Lösung großer Systeme mit dünnbesetzter Systemmatrix eingesetzt. Sie sind daher für die hier betrachteten Probleme von besonderem Interesse. Diese Verfahren berechnen eine Folge von Näherungslösungen $\mathbf{x}_i$ für das Gleichungssystem, die gegen die exakte Lösung konvergiert. Stationäre iterative Verfahren sind stets von der Form

$$\mathbf{x}_i = \mathbf{B}\mathbf{x}_{i-1} + \mathbf{c}, \tag{6.1}$$

wobei weder die Matrix $\mathbf{B}$ noch der Vektor $\mathbf{c}$ von dem Iterationsschritt i abhängen.

Das historisch älteste stationäre Iterationsverfahren geht ebenfalls auf Gauss zurück; es wurde 1822 zur Berechnung von Gleichungssystemen, die bei der Methode der kleinsten Quadrate auftreten, entwickelt und in der Arbeit „Supplementum theoriae combinationis observationum erroribus minime obnoxiae“ veröffentlicht. Das Verfahren wurde 1874 von Seidel [147] weiterentwickelt und wird heute als Gauß-Seidel-Verfahren bezeichnet. Andere stationäre Verfahren

sind das 1845 von JACOBI [88] angegebene Jacobi-Verfahren und das SOR-Verfahren von YOUNG [175] (1950).

Instationäre Iterationsverfahren sind jene, die sich nicht auf die Form (6.1) bringen lassen. Die ersten derartigen Verfahren wurden Anfang der fünfziger Jahre des vorigen Jahrhunderts entwickelt. Zu nennen sind hier die Verfahren von ARNOLDI [8] (1951), LANCZOS [101] (1952) sowie das Verfahren der konjugierten Gradienten („CG-Verfahren“) von HESTENES und STIEFEL [81] (1952). Das CG-Verfahren ist, im Gegensatz zu den beiden anderen instationären Verfahren, nur für Systeme mit symmetrischer und positiv definiter Matrix anwendbar.

Viele der später vorgestellten instationären Gleichungslöser sind Weiterentwicklungen der genannten klassischen Verfahren. Wichtige Beispiele hierfür sind das BiCG-Verfahren („bi-conjugate gradient“) von FLETCHER [67] (1975), das GMRES-Verfahren („generalized minimal residual“) von SAAD und SCHULTZ [141] (1986), das CGS-Verfahren („conjugate gradient squared“) von SONNEVELD [151] (1989), das QMR-Verfahren („quasi-minimal residual“) von FREUND und NACHTIGAL [68] (1991) und das BiCGstab-Verfahren („bi-conjugate gradient stabilized“) von VAN DER VORST [166] (1992). Alle diese Verfahren erlauben die Lösung von Gleichungssystemen mit beliebiger regulärer Matrix. Sie unterscheiden sich jedoch in ihrem Konvergenzverhalten. Welcher Methode man den Vorzug geben sollte, hängt von dem zu lösenden Gleichungssystem ab und ist eine i.a. schwer zu beantwortende Frage. Ein Vergleich einiger instationärer Iterationsverfahren, darunter CG, BiCG, GMRES und CGS, der von NACHTIGAL, REDDY und TREFETHEN [123] durchgeführt wurde, zeigt, daß jede dieser Methoden für eine bestimmte Klasse von Problemen die geeignetste ist. Zur Lösung mehrerer Systeme mit gleicher Systemmatrix wurde von SCHMITT und WEINER [143, 168] (1995) der multiple Arnoldi-Prozeß, ein auf der Arnoldi-Iteration basierendes Verfahren, entwickelt.

Im folgenden werden wir das CG-, das BiCGstab- und das Arnoldi-Verfahren sowie den multiplen Arnoldi-Prozeß näher beschreiben, da diese Verfahren im Rahmen dieser Arbeit bei der numerischen Lösung partieller Differentialgleichungen verwendet werden.

6.2 Das Verfahren der konjugierten Gradienten (CG-Verfahren)

6.2.1 Algorithmus des CG-Verfahrens

Für Systeme $\mathbf{A}\mathbf{x} = \mathbf{b}$ mit symmetrischer und positiv definiter Koeffizientenmatrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ wurde 1952 von HESTENES und STIEFEL [81] das **Verfahren der konjugierten Gradienten** entwickelt. Grundlage des Verfahrens ist eine Umwandlung des Gleichungssystems in die äquivalente Minimierungsaufgabe

$$F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x} \to \min\,.$$

Ausgehend von einem Startvektor $\mathbf{x}_0$ werden geeignete Suchrichtungen bestimmt, um das Minimum von F zu finden. Eine ausführliche Beschreibung dieses Vorgehens ist z.B. bei HACKBUSCH [76] zu finden. Hier soll nur der Algorithmus angegeben werden.

Algorithmus 6.1 (CG-Verfahren).

gegeben: $\mathbf{A}$, $\mathbf{b}$, $\max_{\mathrm{It}}$, TOL_{LSS}

wähle Startvektor $\mathbf{x}_0$

$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$

for $i = 1, 2, \ldots, \max_{\mathrm{It}}$

$\varrho_{i-1} = \|\mathbf{r}_{i-1}\|^2$
if $i = 1$
$\mathbf{p}_1 = \mathbf{r}_0$
else
$\beta_{i-1} = \varrho_{i-1}/\varrho_{i-2}$
$\mathbf{p}_i = \mathbf{r}_{i-1} + \beta_{i-1}\mathbf{p}_{i-1}$
end
$\mathbf{q}_i = \mathbf{A}\mathbf{p}_i$
$\alpha_i = \varrho_{i-1}/\mathbf{p}_i^T\mathbf{q}_i$
$\mathbf{x}_i = \mathbf{x}_{i-1} + \alpha_i\mathbf{p}_i$
$\mathbf{r}_i = \mathbf{r}_{i-1} - \alpha_i\mathbf{q}_i$
if $\|\mathbf{r}_i\| < TOL_{\mathrm{LSS}}$
break
end

end

□

Bemerkung 6.2. Die Lösung des CG-Verfahrens $\mathbf{x}_i$ liegt in dem affinen Unterraum $\mathbf{x}_0 + \mathcal{K}_i$, wobei der Raum $\mathcal{K}_i = \mathrm{span}\{\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, \ldots, \mathbf{A}^{i-1}\mathbf{r}_0\}$ als **Krylov-Raum** bezeichnet wird. Da der Krylov-Raum $\mathcal{K}_n$ mit dem $\mathbb{R}^n$ übereinstimmt, liefert das CG-Verfahren spätestens nach n Schritten die exakte Lösung des Systems $\mathbf{A}\mathbf{x} = \mathbf{b}$. Für große Systeme wird man das Verfahren jedoch in der Regel eher abbrechen. Daher ist die Konvergenzrate der Iterierten von Interesse. □

6.2.2 Vorkonditionierung

Die Konvergenz des CG-Verfahrens hängt wesentlich von der Kondition der Matrix $\mathbf{A}$ ab. Diese wird durch die Konditionszahl $\varkappa(\mathbf{A}) = \varrho(\mathbf{A})\varrho(\mathbf{A}^{-1})$ beschrieben, wobei ϱ der Spektralradius ist. Für symmetrische und positiv definite Matrizen läßt sich die Konditionszahl durch die extremalen Eigenwerte beschreiben; es gilt $\varkappa(\mathbf{A}) = \lambda_{\max}(\mathbf{A})/\lambda_{\min}(\mathbf{A})$. Ist $\mathbf{x}$ die exakte Lösung des Systems, d.h. $\mathbf{A}\mathbf{x} = \mathbf{b}$, so gilt die Konvergenzaussage

$$\|\mathbf{x}_i - \mathbf{x}\|_{\mathbf{A}} \leq 2\left(\frac{\sqrt{\varkappa(\mathbf{A})} - 1}{\sqrt{\varkappa(\mathbf{A})} + 1}\right)^i \|\mathbf{x}_0 - \mathbf{x}\|_{\mathbf{A}} \tag{6.2}$$

in der Norm $\|\mathbf{y}\|_{\mathbf{A}}^2 = \mathbf{y}^T\mathbf{A}\mathbf{y}$, siehe GOLUB/VAN LOAN [73]. Bei schlecht konditionierten Problemen, d.h. solchen mit großer Konditionszahl, sollte das **vorkonditionierte CG-Verfahren** verwendet weden.

Algorithmus 6.3 (Vorkonditioniertes CG-Verfahren).

gegeben: $\mathbf{A}$, $\mathbf{P}$, $\mathbf{b}$, $\max_{\mathrm{It}}$, TOL_{LSS}

wähle Startvektor $\mathbf{x}_0$

$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$

for $i = 1, 2, \ldots, \max_{\mathrm{It}}$

löse $\mathbf{P}\mathbf{z}_{i-1} = \mathbf{r}_{i-1}$ (6.3)
$\varrho_{i-1} = \mathbf{r}_{i-1}^T\mathbf{z}_{i-1}$
if $i = 1$
$\mathbf{p}_1 = \mathbf{z}_0$
else
$\beta_{i-1} = \varrho_{i-1}/\varrho_{i-2}$
$\mathbf{p}_i = \mathbf{z}_{i-1} + \beta_{i-1}\mathbf{p}_{i-1}$
end
$\mathbf{q}_i = \mathbf{A}\mathbf{p}_i$
$\alpha_i = \varrho_{i-1}/\mathbf{p}_i^T\mathbf{q}_i$
$\mathbf{x}_i = \mathbf{x}_{i-1} + \alpha_i\mathbf{p}_i$
$\mathbf{r}_i = \mathbf{r}_{i-1} - \alpha_i\mathbf{q}_i$
if $\|\mathbf{r}_i\| < TOL_{\mathrm{LSS}}$
break
end

end

□

Wie aus dem Algorithmus ersichtlich ist, erfolgt der Abbruch der Iteration, wenn die Norm des Residuums $\mathbf{r}_i$ kleiner als eine vorgegebene Toleranz TOL_{LSS} wird. Die Abkürzung LSS steht dabei für „linear system solver“.

In Algorithmus 6.3 ist $\mathbf{P}$ die Vorkonditionierungs-Matrix. Diese sollte so gewählt werden, daß die Konditionszahl $\varkappa(\mathbf{P}^{-1}\mathbf{A})$ möglichst klein ist, denn es gilt jetzt die zu (6.2) analoge Konvergenzaussage

$$\|\mathbf{x}_i - \mathbf{x}\|_{\mathbf{A}} \leq 2\left(\frac{\sqrt{\varkappa(\mathbf{P}^{-1}\mathbf{A})} - 1}{\sqrt{\varkappa(\mathbf{P}^{-1}\mathbf{A})} + 1}\right)^i \|\mathbf{x}_0 - \mathbf{x}\|_{\mathbf{A}}. \tag{6.4}$$

Es gibt verschiedene Methoden, die Vorkonditionierungs-Matrix $\mathbf{P}$ zu wählen. Wir wollen hier die **SSOR-Vorkonditionierung** beschreiben. SSOR steht für „symmetric succesive overrelaxation“ und ist ein stationäres Iterationsverfahren zur Lösung des Systems $\mathbf{Ax} = \mathbf{b}$. Zerlegt man $\mathbf{A} = \mathbf{L}+\mathbf{D}+\mathbf{U}$ in unteren Dreiecksanteil $\mathbf{L}$, Diagonalanteil $\mathbf{D}$ und oberen Dreiecksanteil $\mathbf{U}$, so lautet die Iterationsvorschrift des SSOR-Verfahrens

$$\mathbf{x}_k = \mathbf{x}_{k-1} - \mathbf{P}^{-1}(\mathbf{A}\mathbf{x}_{k-1} - \mathbf{b}),$$

wobei

$$\mathbf{P} = \frac{1}{\omega(2-\omega)}(\mathbf{D} + \omega\mathbf{L})\mathbf{D}^{-1}(\mathbf{D} + \omega\mathbf{U})$$

ist. Der Parameter ω muß geeignet gewählt werden. Für optimales $\omega = \omega_{\text{opt}}$ kann man zeigen, daß für die Konditionszahl

$$\varkappa(\mathbf{P}_{\text{opt}}^{-1}\mathbf{A}) = O(\sqrt{\varkappa(\mathbf{A})})$$

gilt, siehe Axelsson/Barker [10]. Die Matrix $\mathbf{P}$ ist daher eine gute Wahl als Vorkonditionierungs-Matrix des CG-Verfahrens. Das in Zeile (6.3) zu lösende System $\mathbf{P}\mathbf{z}_{i-1} = \mathbf{r}_{i-1}$ kann wegen der speziellen Struktur von $\mathbf{P}$ in Systeme mit Dreiecks- und Diagonalmatrizen zerlegt werden, die mit geringem Aufwand lösbar sind.

Bemerkung 6.4. Der optimale Parameter ω_{opt} ist von der Matrix $\mathbf{A}$ abhängig und seine Bestimmung i.a. sehr aufwendig. Häufig erreicht man jedoch mit einem pauschal gewählten Parameter $\omega > 1,\ \omega \approx 1$ bereits eine Verbesserung der Konvergenz. In den numerischen Berechnungen in dieser Arbeit verwenden wir $\omega = 1{,}3$. □

Die Idee, Matrizen aus der Iterationsvorschrift stationärer Verfahren zur Vorkonditionierung des CG-Verfahrens zu verwenden, ist nicht auf das SSOR-Verfahren beschränkt. Bei der Jacobi-Vorkonditionierung wird beispielsweise die Matrix des Jacobi-Verfahrens als Vorkonditionierung verwendet.

Untersuchung 6.5 (Einfluß der Vorkonditionierung auf die Konvergenz des CG-Verfahrens). Das folgende Beispiel illustriert die Verbesserung der Konvergenz des CG-Verfahrens durch SSOR-Vorkonditionierung. Gegeben sei die eindimensionale Wärmeleitungsgleichung

$$u_t = a\ u_{xx}, \qquad a > 0 \tag{6.5}$$

auf dem Intervall $x \in [0, 10]$. Die Anfangsbedingung sei $u(x, 0) = u_0(x) = \sin(\pi x/5)$, die Randbedingung $u(0, t) = u(1, t) = 0$. Wir legen auf dem Intervall $[0, 1]$ ein uniformes Gitter der Maschenweite $h = 1/(N + 1),\ N \in \mathbb{N}$ fest; die inneren Gitterpunkte sind dann mit $x_i = ih,\ i = 1, \ldots, N$ gegeben. Diskretisiert man die Gleichung (6.5) auf diesem Gitter mit einem Differenzenverfahren, so geht sie in das System gewöhnlicher Differentialgleichungen

$$\mathbf{u}_t = \mathbf{B}\mathbf{u} \tag{6.6}$$

über, wobei $\mathbf{u}$ die Werte der Näherungslösung an den Stellen $x_i,\ i = 1, \ldots, N$ enthält und die Matrix $\mathbf{B}$ von der Form

$$\mathbf{B} = \frac{a}{h^2}\begin{pmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{N\times N}$$

ist. Wir bezeichnen mit $\mathbf{u}_0 = (u_0(x_1), \ldots, u_0(x_N))^T$ den Vektor der Anfangswerte und lösen das System (6.6) mit dem impliziten Euler-Verfahren. Dieses liefert im ersten Zeitschritt der Länge τ_0 das Gleichungssystem

$$(\mathbf{I} - \tau_0 \mathbf{B})\mathbf{u}_1 = \mathbf{u}_0.$$

Der Vektor $\mathbf{u}_1$ enthält die Werte der Näherungslösung zum Zeitpunkt $t = \tau_0$.

Die Systemmatrix $(\mathbf{I} - \tau_0 \mathbf{B})$ ist symmetrisch und positiv definit. Das Gleichungssystem wird nun mit dem CG-Verfahren gelöst. Für $a = 10^{-4}$, $h = 0{,}01$, $\tau_0 = 0{,}1$ ist in Abbildung 6.1 die Konvergenz des CG-Verfahrens mit und ohne SSOR-Vorkonditionierung dargestellt. Man erkennt, daß in diesem Falle die Vorkonditionierung die Konvergenz wesentlich verbessert. □

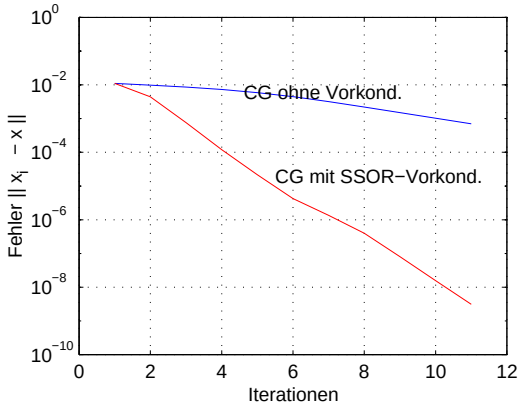


Abbildung 6.1: Konvergenz des CG-Verfahrens mit und ohne SSOR-Vorkonditionierung

6.3 Das BiCGstab-Verfahren

6.3.1 Algorithmus des BiCGstab-Verfahrens mit Vorkonditionierung

Die Konvergenz des CG-Verfahrens ist nur für symmetrische und positiv definite Systemmatrizen gewährleistet. Zur Lösung von Systemen, die diese Bedingung nicht erfüllen, existiert jedoch eine Reihe von Modifikationen dieses Verfahrens. Wir wollen hier das 1992 von VAN DER VORST [166] vorgestellte **BiCGstab-Verfahren** beschreiben. Dieses Verfahren liefert, wie das CG-Verfahren, die exakte Lösung des Gleichungssystems $\mathbf{Ax} = \mathbf{b}$ nach endlich vielen Iterationsschritten, jedoch bricht man bei großen Systemen in der Regel eher ab. Durch Vorkonditionierung läßt sich die Konvergenz oft verbessern.

Algorithmus 6.6 (Vorkonditioniertes BiCGstab-Verfahren).

gegeben: $\mathbf{A}$, $\mathbf{P}$, $\mathbf{b}$, $\max_{\mathrm{It}}$, TOL_{LSS}

wähle Startvektor $\mathbf{x}_0$

$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$

for $i = 1, 2, \ldots, \max_{\mathrm{It}}$

$\varrho_{i-1} = \mathbf{r}_0^T \mathbf{r}_{i-1}$
if $\varrho_{i-1} = 0$
    Verfahren versagt
    break
end
if $i = 1$
    $\mathbf{p}_1 = \mathbf{r}_0$
else
    $\beta_{i-1} = (\varrho_{i-1}/\varrho_{i-2})(\alpha_{i-1}/\omega_{i-1})$
    $\mathbf{p}_i = \mathbf{r}_{i-1} + \beta_{i-1}(\mathbf{p}_{i-1} - \omega_{i-1}\mathbf{v}_{i-1})$
end
löse $\mathbf{P}\mathbf{y}_i = \mathbf{p}_i$
$\mathbf{v}_i = \mathbf{A}\mathbf{y}_i$
$\alpha_i = \varrho_{i-1}/\mathbf{r}_0^T \mathbf{v}_i$
$\mathbf{s}_i = \mathbf{r}_{i-1} - \alpha_i \mathbf{v}_i$
löse $\mathbf{P}\mathbf{z}_i = \mathbf{s}_i$
if $\|\mathbf{z}_i\| < TOL_{\mathrm{LSS}}$
    $\mathbf{x}_i = \mathbf{x}_{i-1} + \alpha_i \mathbf{y}_i$
    break
end
$\mathbf{t}_i = \mathbf{A}\mathbf{z}_i$
$\omega_i = \mathbf{t}_i{}^T \mathbf{s}_i / \mathbf{t}_i^T \mathbf{t}_i$
if $\omega_i = 0$
    Verfahren versagt
    break
end
$\mathbf{x}_i = \mathbf{x}_{i-1} + \alpha_i \mathbf{y}_i + \omega_i \mathbf{z}_i$
$\mathbf{r}_i = \mathbf{s}_i - \omega_i \mathbf{t}_i$
end

□

Für $\mathbf{P} = \mathbf{I}$ liegt das Verfahren ohne Vorkonditionierung vor. Wie beim CG-Verfahren kann SSOR-Vorkonditionierung verwendet werden: Zerlegt man $\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}$ in unteren Dreiecksanteil $\mathbf{L}$, Diagonalanteil $\mathbf{D}$ und oberen Dreiecksanteil $\mathbf{U}$, so ist mit

$$\mathbf{P} = \frac{1}{\omega(2-\omega)}(\mathbf{D} + \omega\mathbf{L})\mathbf{D}^{-1}(\mathbf{D} + \omega\mathbf{U}) \tag{6.7}$$

die SSOR-Vorkonditionierungs-Matrix gegeben. Wir wählen im folgenden stets $\omega = 1{,}3$, vgl. Bemerkung 6.4.

6.3.2 Abbruch der Iteration

Von entscheidender Bedeutung für die Effizienz des Verfahrens ist eine geeignete Abbruchbedingung. Häufig fordert man, daß die Norm des Residuums $\mathbf{b} - \mathbf{A}\mathbf{x}_i$ kleiner als eine vorgegebene Toleranz TOL_{LSS} wird, siehe etwa Barrett et al. [19]. Wir verwenden das BiCGstab-Verfahren zur Lösung der in einer W-Methode, siehe (5.17), auftretenden Gleichungssysteme. In diesem Falle ist es günstig, anstelle einer Beschränkung des Residuums eine Abbruchbedingung der Form

$$\|\mathbf{P}^{-1}(\mathbf{b} - \mathbf{A}\mathbf{x}_i)\|_* \leq TOL_{\mathrm{LSS}} \tag{6.8}$$

zu stellen, wobei die Norm $\|\cdot\|_*$ die gleiche ist, die auch zur Schätzung des zeitlichen Fehlers in Abschnitt 5.4 verwendet wird. Außerdem ist es aus Effizienzgründen oft sinnvoll, die maximal zulässige Anzahl der Iterationen durch eine vorgegebene Zahl $\max_{\mathrm{It}}$ zu beschränken. Der Verlust an Genauigkeit, den die Lösung des Gleichungssystems dadurch erleidet, wirkt sich nämlich im Kontext einer W-Methode nicht unbedingt negativ aus, da diese lediglich eine gewisse Näherung der Lösung benötigt. Numerische Untersuchungen in Abschnitt 9.5 zeigen, daß Iterationsbeschränkung oftmals einen Gewinn an Effizienz erbringt.

Welchen Wert man für $\max_{\mathrm{It}}$ wählen sollte, hängt vom konkreten Problem ab. Allgemeingültige Aussagen sind hier schwer zu machen. Die Größe TOL_{LSS} in der Abbruchbedingung (6.8) sollte geeignet an die Toleranz für den lokalen Fehler TOL_t gekoppelt werden. Eine Möglichkeit der Kopplung wurde von Blom, Verwer und Trompert [25] für iterative Gleichungslöser, die in impliziten BDF-Verfahren eingesetzt werden, vorgestellt. Wir übertragen diese Überlegungen auf das in einer W-Methode eingesetzte BiCGstab-Verfahren für den Fall, daß $\|\cdot\|_*$ die in (5.10) definierte skalierte Euklidische Norm ist.

Wir betrachten das System gewöhnlicher Differentialgleichungen

$$\mathbf{u}_t = \mathbf{f}(t, \mathbf{u}) \tag{6.9}$$

und verwenden zu dessen Lösung die in (5.17) dargestellte s-stufige W-Methode

$$\begin{aligned}
(\mathbf{I} - \tau_i\gamma\mathbf{T})\widetilde{\mathbf{k}}_j &= \mathbf{f}\left(t_i + \tau_i c_j, \mathbf{u}_i + \tau_i\gamma\sum_{l=1}^{j-1}\varphi_{lj}\widetilde{\mathbf{k}}_l\right) + \sum_{l=1}^{j-1}\vartheta_{lj}\widetilde{\mathbf{k}}_l, \qquad j = 1,\ldots,s, \quad (6.10)\\
\mathbf{u}_{i+1} &= \mathbf{u}_i + \tau_i\sum_{l=1}^{s} g_l\widetilde{\mathbf{k}}_l,\\
\widehat{\mathbf{u}}_{i+1} &= \mathbf{u}_i + \tau_i\sum_{l=1}^{s}\widehat{g}_l\widetilde{\mathbf{k}}_l.
\end{aligned}$$

Der Kürze wegen führen wir die Bezeichnungen

$$\mathbf{M} = \mathbf{I} - \tau_i\gamma\mathbf{T}, \qquad \mathbf{z}_j = \mathbf{f}\left(t_i + \tau_i c_j, \mathbf{u}_i + \tau_i\gamma\sum_{l=1}^{j-1}\varphi_{lj}\widetilde{\mathbf{k}}_l\right) + \sum_{l=1}^{j-1}\vartheta_{lj}\widetilde{\mathbf{k}}_l, \quad j = 1,\ldots,s$$

ein, so daß die in (6.10) zu lösenden Gleichungssysteme in der Form

$$\mathbf{M}\widetilde{\mathbf{k}}_j = \mathbf{z}_j, \qquad j = 1,\ldots,s \tag{6.11}$$

geschrieben werden können. Die rechte Seite $\mathbf{z}_j$ hängt dabei nur von $\widetilde{\mathbf{k}}_1, \ldots, \widetilde{\mathbf{k}}_{j-1}$ ab. Die Lösung dieser Systeme erfolge mit dem BiCGstab-Verfahren mit einer SSOR-Vorkonditionierung, die durch eine Matrix $\mathbf{P}$ beschrieben wird. Wir verwenden in diesem Abschnitt die folgenden Bezeichnungen:

- $\mathbf{u}(t)$ – exakte Lösung der Differentialgleichung (6.9) mit dem Anfangswert $\mathbf{u}(t_0)$
- $\widetilde{\mathbf{u}}_i(t)$ – exakte Lösung der Differentialgleichung (6.9) mit dem Anfangswert $\widetilde{\mathbf{u}}_i(t_i) = \mathbf{u}(t_i)$
- $\mathbf{u}_{i+1}$ – Lösung der Differentialgleichung (6.9) mit der W-Methode (6.10) und einem exakten linearen Löser
- $\overline{\mathbf{u}}_{i+1}$ – Lösung der Differentialgleichung (6.9) mit der W-Methode (6.10) und einem iterativen Löser (BiCGstab-Verfahren)
- $\widetilde{\mathbf{k}}_j$ – exakte Lösung des Gleichungssystems (6.11)
- $\overline{\mathbf{k}}_j$ – Lösung des Gleichungssystems (6.11) mit iterativem Löser (BiCGstab-Verfahren)

Das Residuum $\mathbf{r}_j := \mathbf{M}(\widetilde{\mathbf{k}}_j - \overline{\mathbf{k}}_j) = \mathbf{z}_j - \mathbf{M}\overline{\mathbf{k}}_j$ ist, im Gegensatz zu dem Fehler des Gleichungslösers $\widetilde{\mathbf{k}}_j - \overline{\mathbf{k}}_j$, eine berechenbare Größe. Das Ziel der in Abschnitt 5.8 vorgestellten Zeitschrittsteuerung ist es, den lokalen Fehler $\|\widetilde{\mathbf{u}}_i(t_{i+1}) - \overline{\mathbf{u}}_{i+1}\|_*$ unter einer Toleranz TOL_t zu halten, wobei $\|\cdot\|_*$ die skalierte Euklidische Vektornorm ist. Eine sinnvolle Forderung wäre daher

$$\|\widetilde{\mathbf{u}}_i(t_{i+1}) - \overline{\mathbf{u}}_{i+1}\|_* \leq \|\widetilde{\mathbf{u}}_i(t_{i+1}) - \mathbf{u}_{i+1}\|_* + \|\mathbf{u}_{i+1} - \overline{\mathbf{u}}_{i+1}\|_* \leq TOL_t.$$

Wir geben nun vor, daß der Fehler $\|\mathbf{u}_{i+1} - \overline{\mathbf{u}}_{i+1}\|_*$ des Gleichungslösers von ähnlicher Größenordnung wie der lokale Fehler der W-Methode $\|\widetilde{\mathbf{u}}_i(t_{i+1}) - \mathbf{u}_{i+1}\|_*$ sein sollte, also

$$\|\mathbf{u}_{i+1} - \overline{\mathbf{u}}_{i+1}\|_* \approx \|\widetilde{\mathbf{u}}_i(t_{i+1}) - \mathbf{u}_{i+1}\|_* \tag{6.12}$$

gelte. Deshalb wäre eine Ungleichung der Form

$$\|\mathbf{u}_{i+1} - \overline{\mathbf{u}}_{i+1}\|_* \leq \alpha TOL_t, \qquad \alpha \approx 1/2 \tag{6.13}$$

wünschenswert. Die Bildungsvorschrift

$$\mathbf{u}_{i+1} = \mathbf{u}_i + \tau_i \sum_{l=1}^{s} g_l \widetilde{\mathbf{k}}_l$$

in (6.10) gilt in der gleichen Form auch für $\overline{\mathbf{u}}_{i+1}$, d.h. es gilt ebenfalls

$$\overline{\mathbf{u}}_{i+1} = \mathbf{u}_i + \tau_i \sum_{l=1}^{s} g_l \overline{\mathbf{k}}_l.$$

Folglich ist $\mathbf{u}_{i+1} - \overline{\mathbf{u}}_{i+1} = \tau_i \sum_{l=1}^{s} g_l(\widetilde{\mathbf{k}}_l - \overline{\mathbf{k}}_l)$. Man erhält die Abschätzung

$$\begin{aligned} \|\mathbf{u}_{i+1} - \overline{\mathbf{u}}_{i+1}\|_* &\leq \tau_i \sum_{l=1}^{s} |g_l| \|\widetilde{\mathbf{k}}_l - \overline{\mathbf{k}}_l\|_* = \tau_i \sum_{l=1}^{s} |g_l| \|\mathbf{M}^{-1}\mathbf{r}_l\|_* \qquad (6.14) \\ &= \tau_i \sum_{l=1}^{s} |g_l| \|(\mathbf{M}^{-1}\mathbf{P})\mathbf{P}^{-1}\mathbf{r}_l\|_* \leq \tau_i \sum_{l=1}^{s} |g_l| \|\mathbf{M}^{-1}\mathbf{P}\|_2 \|\mathbf{P}^{-1}\mathbf{r}_l\|_* \qquad (6.15) \\ &\leq \tau_i \sum_{l=1}^{s} |g_l| \sigma_{\max}(\mathbf{M}^{-1}\mathbf{P}) \|\mathbf{P}^{-1}\mathbf{r}_l\|_*, \end{aligned}$$

wobei $\sigma_{\max}$ der maximale Singulärwert ist. Man geht davon aus, daß die Vorkonditionierung so gut ist, daß $\sigma_{\max}(\mathbf{M}^{-1}\mathbf{P}) = 1/\sigma_{\min}(\mathbf{P}^{-1}\mathbf{M}) \approx 1$ ist. Um (6.13) zu gewährleisten, sollte daher eine Bedingung der Form

$$\|\mathbf{P}^{-1}\mathbf{r}_l\|_* \leq \frac{\alpha_{\mathrm{LSS}}}{\tau_i} TOL_t =: TOL_{\mathrm{LSS}} \qquad (6.16)$$

erfüllt werden, wobei α_{LSS} eine Konstante von moderater Größe, etwa $\alpha_{\mathrm{LSS}} = 1/2$, ist. Die Größe $\mathbf{P}^{-1}\mathbf{r}_l$ entspricht dem $\mathbf{z}_i$ in Algorithmus 6.6.

Bemerkung 6.7. Der genaue Wert eines bezüglich der Effizienz des Verfahrens optimalen α_{LSS} kann auf diese Weise in der Regel nicht bestimmt werden, wofür besonders zwei Gründe verantwortlich sind. Zum einen ist es nicht klar, wie die Fehler $\|\mathbf{u}_{i+1} - \overline{\mathbf{u}}_{i+1}\|_*$ und $\|\widetilde{\mathbf{u}}_i(t_{i+1}) - \mathbf{u}_{i+1}\|_*$ sinnvollerweise gewichtet werden sollten und ob (6.12) wirklich die effizienteste Variante darstellt. Zum anderen ist die Größe des Singulärwerts $\sigma_{\max}(\mathbf{A}^{-1}\mathbf{P})$ in (6.14) schwer einzuschätzen. Die Suche nach einem α_{LSS}, das größtmögliche Effizienz bietet, bleibt daher numerischen Testrechnungen vorbehalten. Im Grunde ist die wesentliche Erkenntnis aus der Beziehung (6.16) die Tatsache, daß α_{LSS} *nicht vom Zeitschritt* τ_i abhängen sollte. □

6.4 Das Arnoldi-Verfahren

6.4.1 Definition und algorithmische Umsetzung des Verfahrens

Das **Verfahren von** Arnoldi [8] stammt aus dem Jahre 1951 und ist damit eines der ältesten instationären Iterationsverfahren zur Lösung linearer Gleichungssysteme. Für Systeme der Form $\mathbf{Ax} = \mathbf{b}$ ist der Algorithmus des Arnoldi-Verfahrens in Anhang B angegeben. Im Rahmen dieser Arbeit benötigen wir das Arnoldi-Verfahren jedoch nur für spezielle Gleichungssysteme der Form

$$(\mathbf{I} - \delta\mathbf{A})\mathbf{x} = \mathbf{b}, \qquad \delta \in \mathbb{R}, \qquad (6.17)$$

die bei der impliziten Diskretisierung von Differentialgleichungen auftreten. Für derartige Systeme existiert eine spezielle Variante des Arnoldi-Verfahrens, die etwa in Büttner et al. [33] dargestellt wird. Wir werden uns daher im folgenden auf diese Variante beschränken.

In dem Gleichungssystem (6.17) sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ und $\mathbf{b} \in \mathbb{R}^n$. Die Matrix $\mathbf{I}$ ist hier die $n \times n$-Einheitsmatrix[1]. Wir definieren eine Folge von Unterräumen des $\mathbb{R}^n$, die sogenannten **Krylov-**

[1] In diesem Abschnitt wird das Symbol $\mathbf{I}$ für Einheitsmatrizen verschiedener Dimension verwendet. Die Dimension ist jedoch stets aus dem Kontext erkennbar.

Räume $\mathcal{K}_i$, durch

$$\mathcal{K}_i = \operatorname{span}\{\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^{i-1}\mathbf{b}\}. \tag{6.18}$$

Die Arnoldi-Iteration

$$\mathbf{w}_j := \left(\mathbf{I} - \sum_{k=1}^{j} \mathbf{q}_k \mathbf{q}_k^T\right) \mathbf{A}\mathbf{q}_j, \qquad \mathbf{q}_{j+1} := \mathbf{w}_j/|\mathbf{w}_j|, \qquad j = 0, \ldots, i-1. \tag{6.19}$$

erzeugt eine Orthonormalbasis $\{\mathbf{q}_1, \ldots, \mathbf{q}_i\}$ des Krylov-Raumes $\mathcal{K}_i$, siehe etwa Saad [140]. In jedem Iterationsschritt wird die Menge der Basisvektoren um einen neuen Vektor erweitert. Wir fassen die Basisvektoren in den Matrizen $\mathbf{Q}_i = \begin{pmatrix}\mathbf{q}_1 & \ldots & \mathbf{q}_i\end{pmatrix}$ zusammen.

Als Näherungslösung $\mathbf{x}_i$ des Systems $(\mathbf{I} - \delta\mathbf{A})\mathbf{x} = \mathbf{b}$ wird jeweils die exakte Lösung des Systems

$$(\mathbf{I} - \delta\mathbf{T}_i)\mathbf{x}_i = \mathbf{b} \tag{6.20}$$

gewählt, wobei $\mathbf{T}_i = \mathbf{Q}_i\mathbf{Q}_i^T\mathbf{A}$ ist. Dieses System ist, wie das Ausgangssystem (6.17), von der Dimension n, jedoch hat die Matrix $\mathbf{T}_i$ nur den Rang i. Man kann leicht zeigen, daß (6.20) äquivalent zu einem i-dimensionalen Gleichungssystem ist.

Satz 6.8. *Für beliebiges $i = 1, 2, \ldots$ sei $\mathbf{H}_i = \mathbf{Q}_i^T\mathbf{A}\mathbf{Q}_i$. Die Matrizen $\mathbf{I} - \delta\mathbf{T}_i$ und $\mathbf{I} - \delta\mathbf{H}_i$ seien regulär. Dann ist $\mathbf{x}_i \in \mathcal{K}_i$, und das System (6.20) ist äquivalent zu dem i-dimensionalen System*

$$(\mathbf{I} - \delta\mathbf{H}_i)\mathbf{y}_i = \mathbf{Q}_i^T\mathbf{b}, \tag{6.21}$$

wobei $\mathbf{x}_i = \mathbf{Q}_i\mathbf{y}_i$ ist.

Beweis. Es gilt

$$\mathbf{Q}_i\mathbf{H}_i = \mathbf{Q}_i\mathbf{Q}_i^T\mathbf{A}\mathbf{Q}_i = \mathbf{T}_i\mathbf{Q}_i$$

und damit

$$\mathbf{Q}_i(\mathbf{I} - \delta\mathbf{H}_i) = (\mathbf{I} - \delta\mathbf{T}_i)\mathbf{Q}_i$$

sowie

$$(\mathbf{I} - \delta\mathbf{T}_i)^{-1}\mathbf{Q}_i = \mathbf{Q}_i(\mathbf{I} - \delta\mathbf{H}_i)^{-1}.$$

Nach der Definition von $\mathbf{x}_i$ in (6.20) folgt

$$\mathbf{x}_i = (\mathbf{I} - \delta\mathbf{T}_i)^{-1}\mathbf{b}.$$

Weil wegen (6.18) $\mathbf{b} \in \mathcal{K}_i$ ist, existiert ein Vektor $\mathbf{z}_i \in \mathbb{R}^i$, so daß $\mathbf{b} = \mathbf{Q}_i\mathbf{z}_i$ ist. Da die Spalten von $\mathbf{Q}_i$ orthonormale Vektoren sind, ist $\mathbf{Q}_i$ eine orthogonale Matrix, d.h. $\mathbf{Q}_i^T\mathbf{Q}_i = \mathbf{I}$. Es folgt $\mathbf{Q}_i^T\mathbf{b} = \mathbf{z}_i$ und

$$\mathbf{x}_i = (\mathbf{I} - \delta\mathbf{T}_i)^{-1}\mathbf{Q}_i\mathbf{z}_i = \mathbf{Q}_i(\mathbf{I} - \delta\mathbf{H}_i)^{-1}\mathbf{z}_i.$$

Setzt man $\mathbf{y}_i = (\mathbf{I} - \delta\mathbf{H}_i)^{-1}\mathbf{z}_i$, so erhält man das äquivalente System

$$(\mathbf{I} - \delta\mathbf{H}_i)\mathbf{y}_i = \mathbf{Q}_i^T\mathbf{b}$$

mit $\mathbf{x}_i = \mathbf{Q}_i\mathbf{y}_i$. Insbesondere gilt also $\mathbf{x}_i \in \mathcal{K}_i$. □

Bemerkung 6.9. Wenn die Matrix $\mathbf{A}$ symmetrisch und positiv definit ist, so entspricht der durch das Arnoldi-Verfahren aufgebaute Krylov-Raum dem Krylov-Raum des CG-Verfahrens zur Lösung des Systems $(\mathbf{I} - \delta\mathbf{A})\mathbf{x} = \mathbf{b}$, falls das CG-Verfahren mit dem Startvektor $\mathbf{x}_0 = \mathbf{0}$ begonnen wurde, siehe Bemerkung 6.2. □

Lemma 6.10. *Die Matrix $\mathbf{H}_i$ hat Hessenberg-Form, d.h. sie ist eine obere Dreiecksmatrix mit einer zusätzlichen von 0 verschiedenen Subdiagonale.*

Beweis. Der Vektor $\mathbf{q}_{k+1}$ entsteht nach der Definition in (6.19) durch Orthogonalisieren von $\mathbf{A}\mathbf{q}_k$ bezüglich des Krylov-Raums $\mathcal{K}_k = \text{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_k\}$. Daraus folgt

$$\mathbf{A}\mathbf{q}_k \in \text{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_{k+1}\} = \mathcal{K}_{k+1}.$$

Wegen $\mathbf{q}_j \perp \mathcal{K}_{k+1}$ für alle $j > k+1$ folgt

$$(\mathbf{H}_i)_{jk} = \mathbf{q}_j^T \mathbf{A}\mathbf{q}_i = 0, \qquad k+1 < j \leq i,$$

wobei $(\mathbf{H}_i)_{jk}$ das Element der Matrix $\mathbf{H}_i$ in Zeile j, Spalte k bezeichnet[2]. Das bedeutet aber gerade, daß $\mathbf{H}_i$ eine Hessenberg-Matrix ist. □

Aus Satz 6.8 folgt, daß im i-ten Iterationsschritt nur noch das i-dimensionale Gleichungssystem (6.21) gelöst werden muß. Es ergibt sich die folgende Kurzform des Arnoldi-Verfahrens für ein System $(\mathbf{I} - \delta\mathbf{A})\mathbf{x} = \mathbf{b}$.

Algorithmus 6.11 (Arnoldi-Verfahren für $(\mathbf{I} - \delta\mathbf{A})\mathbf{x} = \mathbf{b}$ − Kurzform).

gegeben: $\mathbf{A}$, $\mathbf{b}$, $\max_{\text{It}}$

$\mathbf{q}_1 = \mathbf{b}/|\mathbf{b}|$

$\mathbf{Q}_1 = (\mathbf{q}_1)$

for $i = 2, \ldots, \max_{\text{It}}$

$\mathbf{v}_{i-1} = \mathbf{A}\mathbf{q}_{i-1}$ (Krylov-Schritt)

$\mathbf{w}_{i-1} = (\mathbf{I} - \mathbf{Q}_{i-1}\mathbf{Q}_{i-1}^T)\mathbf{v}_{i-1}$ (Gram-Schmidt-Orthogonalisierung)

$\mathbf{q}_i = \mathbf{w}_{i-1}/|\mathbf{w}_{i-1}|$

$\mathbf{Q}_i = \begin{pmatrix}\mathbf{Q}_{i-1} & \mathbf{q}_i\end{pmatrix}$

$\mathbf{H}_i = \mathbf{Q}_i^T \mathbf{A}\mathbf{Q}_i$

$\mathbf{z}_i = \mathbf{Q}_i^T \mathbf{b}$

löse $(\mathbf{I} - \delta\mathbf{H}_i)\mathbf{y}_i = \mathbf{z}_i$

end

$\mathbf{x}_i = \mathbf{Q}_i \mathbf{y}_i$

□

[2]Wir werden in diesem Abschnitt des öfteren eine derartige Bezeichnungsweise verwenden.

Algorithmus 6.11 gibt jedoch noch nicht die effizienteste Variante des Arnoldi-Verfahrens wieder. Diese geben wir in dem folgenden Algorithmus an:

Algorithmus 6.12 (Arnoldi-Verfahren für $(\mathbf{I}-\delta\mathbf{A})\mathbf{x}=\mathbf{b}$ – effiziente Form).

gegeben: $\mathbf{A}$, $\mathbf{b}$, $\max_{\mathrm{It}}$, TOL_{LSS}

$\widetilde{\mathbf{q}}_1 = \mathbf{b}/|\mathbf{b}|$

for $i = 1, \ldots, \max_{\mathrm{It}}$

$\mathbf{u}_{1i} = \mathbf{A}\widetilde{\mathbf{q}}_i$

for $j = 1, \ldots, i$

$\widetilde{h}_{ji} = \widetilde{\mathbf{q}}_j^T \mathbf{u}_{ji}$

$\mathbf{u}_{j+1,i} = \mathbf{u}_{ji} - \widetilde{h}_{ji}\widetilde{\mathbf{q}}_j$ (6.22)

end

$\widetilde{h}_{i+1,i} = |\mathbf{u}_{i+1,i}|$

$$\widetilde{\mathbf{H}}_i = \begin{pmatrix} \widetilde{h}_{11} & \cdots & \widetilde{h}_{1i} \\ \vdots & & \vdots \\ \widetilde{h}_{i1} & \cdots & \widetilde{h}_{ii} \end{pmatrix}$$

$\widetilde{\mathbf{z}}_i = (|\mathbf{b}|\ \mathbf{0})^T \in \mathbb{R}^i$

löse $(\mathbf{I} - \delta\widetilde{\mathbf{H}}_i)\widetilde{\mathbf{y}}_i = \widetilde{\mathbf{z}}_i$ (6.23)

$\mathbf{r}_i = -\delta(\widetilde{\mathbf{y}}_i)_i \mathbf{u}_{i+1,i}$ (Residuum)

if $\|\mathbf{r}_i\|_* \leq TOL_{\mathrm{LSS}}$

break

end

$\widetilde{\mathbf{q}}_{i+1} = \mathbf{u}_{i+1,i}/|\mathbf{u}_{i+1,i}|$

end

$\widetilde{\mathbf{Q}}_i = \begin{pmatrix} \widetilde{\mathbf{q}}_1 & \ldots & \widetilde{\mathbf{q}}_i \end{pmatrix}$

$\widetilde{\mathbf{x}}_i = \widetilde{\mathbf{Q}}_i \widetilde{\mathbf{y}}_i$

□

Bemerkung 6.13. Für die Abbruchbedingung können verschiedene Normen $\|\cdot\|_*$ des Residuums verwendet werden. Häufig wird die Maximum-Norm $\|(a_1, \ldots, a_n)\|_{\max} = \max_{i=1,\ldots,n} |a_i|$, die Euklidische Vektornorm $\|(a_1, \ldots, a_n)\|_2 = \sqrt{\sum_{i=1}^n a_i^2}$ oder die skalierte Euklidische Norm $\|(a_1, \ldots, a_n)\| = \|(a_1, \ldots, a_n)\|_2/\sqrt{n}$ benutzt. □

Satz 6.14. *Es gilt* $\mathbf{q}_i = \widetilde{\mathbf{q}}_i$, $\mathbf{Q}_i = \widetilde{\mathbf{Q}}_i$, $\mathbf{H}_i = \widetilde{\mathbf{H}}_i$, $\mathbf{z}_i = \widetilde{\mathbf{z}}_i$, $\mathbf{y}_i = \widetilde{\mathbf{y}}_i$ *und* $\mathbf{x}_i = \widetilde{\mathbf{x}}_i$. *Damit erzeugt Algorithmus 6.11 die gleiche Näherungslösung wie Algorithmus 6.12. Die Größe* $\mathbf{r}_i = -\delta(\widetilde{\mathbf{y}}_i)_i \mathbf{u}_{i+1,i}$ *aus Algorithmus 6.12 ist gleich dem Residuum* $(\mathbf{I} - \delta\mathbf{A})\mathbf{x}_i - \mathbf{b}$.

Wir beweisen den Satz schrittweise durch einige Lemmata.

Lemma 6.15. *Es gilt* $\mathbf{q}_i = \widetilde{\mathbf{q}}_i$ *und* $\mathbf{Q}_i = \widetilde{\mathbf{Q}}_i$.

Beweis. Der Beweis erfolgt durch vollständige Induktion über i. Offensichtlich ist $\mathbf{q}_1 = \widetilde{\mathbf{q}}_1 = \mathbf{b}/|\mathbf{b}|$. Wir nehmen an, daß $\mathbf{q}_j = \widetilde{\mathbf{q}}_j$ für $j = 1, \ldots, i-1$ gelte. Aus (6.19) folgt, daß die Vektoren $\mathbf{q}_j$ für $j = 1, \ldots, i-1$ ein Orthonormalsystem bilden, da sie mit dem Gram-Schmidt-Verfahren erzeugt wurden. Nach der Definition von $\mathbf{q}_i$ in (6.19) gilt

$$\mathbf{q}_i = \mathbf{w}_{i-1}/|\mathbf{w}_{i-1}|, \qquad \mathbf{w}_{i-1} = (\mathbf{I} - \mathbf{Q}_{i-1}\mathbf{Q}_{i-1}^T)\mathbf{A}\mathbf{q}_{i-1}.$$

In Algorithmus 6.12 ist $\widetilde{\mathbf{q}}_i$ gemäß

$$\widetilde{\mathbf{q}}_i = \mathbf{u}_{i,i-1}/|\mathbf{u}_{i,i-1}|$$

definiert. Nach der Definition von $\mathbf{u}_{j+1,i}$ in (6.22) gilt

$$\mathbf{u}_{j+1,i} = \mathbf{u}_{ji} - \widetilde{h}_{ji}\widetilde{\mathbf{q}}_j = \mathbf{u}_{ji} - \widetilde{\mathbf{q}}_j^T\mathbf{u}_{ji}\widetilde{\mathbf{q}}_j = (\mathbf{I} - \widetilde{\mathbf{q}}_j\widetilde{\mathbf{q}}_j^T)\mathbf{u}_{ji}$$

für $j = 1, \ldots, i$, also auch

$$\mathbf{u}_{i,i-1} = (\mathbf{I} - \widetilde{\mathbf{q}}_{i-1}\widetilde{\mathbf{q}}_{i-1}^T)\mathbf{u}_{i-1,i-1}.$$

In analoger Weise erhält man

$$\mathbf{u}_{i-1,i-1} = (\mathbf{I} - \widetilde{\mathbf{q}}_{i-2}\widetilde{\mathbf{q}}_{i-2}^T)\mathbf{u}_{i-2,i-1}$$

usw.. Es folgt

$$\mathbf{u}_{i,i-1} = (\mathbf{I} - \widetilde{\mathbf{q}}_{i-1}\widetilde{\mathbf{q}}_{i-1}^T)\cdots(\mathbf{I} - \widetilde{\mathbf{q}}_1\widetilde{\mathbf{q}}_1^T)\mathbf{u}_{1,i-1} = (\mathbf{I} - \widetilde{\mathbf{q}}_{i-1}\widetilde{\mathbf{q}}_{i-1}^T)\cdots(\mathbf{I} - \widetilde{\mathbf{q}}_1\widetilde{\mathbf{q}}_1^T)\mathbf{A}\widetilde{\mathbf{q}}_{i-1}.$$

Wegen der Orthogonalität der $\widetilde{\mathbf{q}}_j$ für $j = 1, \ldots, i-1$ folgt daraus

$$\mathbf{u}_{i,i-1} = (\mathbf{I} - \widetilde{\mathbf{q}}_1\widetilde{\mathbf{q}}_1^T - \cdots - \widetilde{\mathbf{q}}_{i-1}\widetilde{\mathbf{q}}_{i-1}^T)\mathbf{A}\widetilde{\mathbf{q}}_{i-1}.$$

Ebenfalls aus der genannten Orthogonalität folgt $\widetilde{\mathbf{Q}}_{i-1}\widetilde{\mathbf{Q}}_{i-1}^T = \widetilde{\mathbf{q}}_1\widetilde{\mathbf{q}}_1^T + \cdots + \widetilde{\mathbf{q}}_{i-1}\widetilde{\mathbf{q}}_{i-1}^T$. Also ist

$$\mathbf{u}_{i,i-1} = (\mathbf{I} - \widetilde{\mathbf{Q}}_{i-1}\widetilde{\mathbf{Q}}_{i-1}^T)\mathbf{A}\widetilde{\mathbf{q}}_{i-1}.$$

Da nach Induktionsvoraussetzung $\mathbf{Q}_{i-1} = \widetilde{\mathbf{Q}}_{i-1}$ ist, so folgt $\mathbf{u}_{i,i-1} = \mathbf{w}_{i-1}$ und damit auch $\mathbf{q}_i = \widetilde{\mathbf{q}}_i$. □

Lemma 6.16. *Es gilt*

$$\mathbf{u}_{j+1,i} = (\mathbf{I} - \mathbf{q}_1\mathbf{q}_1^T - \cdots - \mathbf{q}_j\mathbf{q}_j^T)\mathbf{A}\mathbf{q}_i = (\mathbf{I} - \mathbf{Q}_j\mathbf{Q}_j^T)\mathbf{A}\mathbf{q}_i = \begin{cases} \mathbf{u}_{j+1,i}, & j \le i, \\ \mathbf{0}, & j > i. \end{cases}$$

Beweis. Die Gleichheit $\mathbf{I} - \mathbf{q}_1\mathbf{q}_1^T - \cdots - \mathbf{q}_j\mathbf{q}_j^T = \mathbf{I} - \mathbf{Q}_j\mathbf{Q}_j^T$ folgt aus der Orthogonalität der Vektoren $\mathbf{q}_k$ für $k = 1, \ldots, j$. Im Beweis des Lemmas 6.15 wurde die Beziehung

$$\mathbf{u}_{i,i-1} = (\mathbf{I} - \widetilde{\mathbf{Q}}_{i-1}\widetilde{\mathbf{Q}}_{i-1}^T)\mathbf{A}\widetilde{\mathbf{q}}_{i-1}$$

gezeigt. Auf analoge Weise kann man

$$\mathbf{u}_{j+1,i} = (\mathbf{I} - \widetilde{\mathbf{Q}}_j\widetilde{\mathbf{Q}}_j^T)\mathbf{A}\widetilde{\mathbf{q}}_i = (\mathbf{I} - \mathbf{Q}_j\mathbf{Q}_j^T)\mathbf{A}\mathbf{q}_i$$

für $j \le i$ beweisen. Für $i < j$ gilt $\mathbf{A}\mathbf{q}_i \in \mathcal{K}_{i+1} \subseteq \mathcal{K}_j$. Die Multiplikation mit der Matrix $(\mathbf{I} - \widetilde{\mathbf{Q}}_j\widetilde{\mathbf{Q}}_j^T)$ beschreibt die Orthogonalisierung bezüglich $\mathcal{K}_j$. Da $\mathbf{A}\mathbf{q}_i$ bereits in $\mathcal{K}_j$ enthalten ist, folgt $(\mathbf{I} - \widetilde{\mathbf{Q}}_j\widetilde{\mathbf{Q}}_j^T)\mathbf{A}\mathbf{q}_i = \mathbf{0}$. □

Lemma 6.17. *Die Matrix $\mathbf{H}_{i-1}$ ist eine Untermatrix von $\mathbf{H}_i$, d.h. $\mathbf{H}_i$ ist von der Form*

$$\mathbf{H}_i = \left(\begin{array}{ccccc|c} & & \mathbf{H}_{i-1} & & & h_{1i} \\ & & & & & \vdots \\ & & & & & h_{i-1,i} \\ \hline 0 & \dots & 0 & h_{i,i-1} & & h_{ii} \end{array}\right). \tag{6.24}$$

Beweis. Da $\mathbf{Q}_i$ die Darstellung $\mathbf{Q}_i = \begin{pmatrix}\mathbf{Q}_{i-1} & \mathbf{q}_i\end{pmatrix}$ besitzt, ist $\mathbf{H}_{i-1}$ die entsprechende Untermatrix von $\mathbf{H}_i$. Die Darstellung (6.24) ist korrekt, weil $\mathbf{H}_i$ nach Lemma 6.10 eine Hessenberg-Matrix ist. □

Lemma 6.18. *Es gilt $\mathbf{H}_i = \widetilde{\mathbf{H}}_i$.*

Beweis. Der Beweis erfolgt durch vollständige Induktion über i. Es gilt $\mathbf{H}_1 = \mathbf{q}_1^T\mathbf{A}\mathbf{q}_1$ und $\widetilde{\mathbf{H}}_1 = (\widetilde{h}_{11}) = \mathbf{q}_1^T\mathbf{u}_{11} = \mathbf{q}_1^T\mathbf{A}\mathbf{q}_1$, also $\mathbf{H}_1 = \widetilde{\mathbf{H}}_1$. Wir nehmen an, daß $\mathbf{H}_j = \widetilde{\mathbf{H}}_j$ für $j = 1, \dots, i-1$ gelte. Wegen der Aussage von Lemma 6.17 ist nur noch zu zeigen, daß $h_{ji} = \widetilde{h}_{ji}$ für $j = 1, \dots, i$ und $h_{i,i-1} = \widetilde{h}_{i,i-1}$ gelten. Die erste Beziehung folgt aus

$$h_{ji} = (\mathbf{H}_i)_{ji} = (\mathbf{Q}_i^T\mathbf{A}\mathbf{Q}_i)_{ji} = \mathbf{q}_j^T\mathbf{A}\mathbf{q}_i = \mathbf{q}_j^T\mathbf{u}_{ji} = \widetilde{h}_{ji}, \qquad j = 1, \dots, i$$

die verbleibende Beziehung $h_{i,i-1} = \widetilde{h}_{i,i-1}$ zeigen wir folgendermaßen:

Aus Algorithmus 6.11 ergibt sich

$$h_{i,i-1} = \mathbf{q}_i^T\mathbf{A}\mathbf{q}_{i-1}$$

Wegen der Darstellung $\mathbf{q}_i^T = \widetilde{\mathbf{q}}_i^T = \mathbf{u}_{i,i-1}/|\mathbf{u}_{i,i-1}|$ in Algorithmus 6.12 folgt weiter

$$h_{i,i-1} = \frac{1}{|\mathbf{u}_{i,i-1}|}\mathbf{u}_{i,i-1}^T\mathbf{A}\mathbf{q}_{i-1}.$$

Lemma 6.16 liefert

$$h_{i,i-1} = \frac{1}{|\mathbf{u}_{i,i-1}|}\mathbf{q}_{i-1}^T\mathbf{A}^T(\mathbf{I} - \mathbf{Q}_{i-1}\mathbf{Q}_{i-1}^T)\mathbf{A}\mathbf{q}_{i-1}.$$

Aus der Orthogonalität von $\mathbf{Q}_{i-1}$ folgt $\mathbf{I} - \mathbf{Q}_{i-1}\mathbf{Q}_{i-1}^T = (\mathbf{I} - \mathbf{Q}_{i-1}\mathbf{Q}_{i-1}^T)^T(\mathbf{I} - \mathbf{Q}_{i-1}\mathbf{Q}_{i-1}^T)$. Folglich ist

$$h_{i,i-1} = \frac{1}{|\mathbf{u}_{i,i-1}|}\left(\mathbf{q}_{i-1}\mathbf{A}(\mathbf{I} - \mathbf{Q}_{i-1}\mathbf{Q}_{i-1}^T)\right)^T\left((\mathbf{I} - \mathbf{Q}_{i-1}\mathbf{Q}_{i-1}^T)\mathbf{A}\mathbf{q}_{i-1}\right) = |\mathbf{u}_{i,i-1}| = \widetilde{h}_{i,i-1}.$$

□

Lemma 6.19. *Es gilt $\mathbf{z}_i = (|\mathbf{b}|, 0, \dots, 0)^T = \widetilde{\mathbf{z}}_i$.*

Beweis. Zu zeigen ist nur die linke Gleichheit. Wegen $\mathbf{q}_1 = \mathbf{b}/|\mathbf{b}|$ und der Orthogonalität der Spalten von $\mathbf{Q}_i$ gilt $\mathbf{z}_i = \mathbf{Q}_i^T\mathbf{b} = (\mathbf{q}_1^T\mathbf{b}, 0, \dots, 0)^T = (|\mathbf{b}|, 0, \dots, 0)^T$. □

Beweis des Satzes 6.14. Die Gleichheitsaussagen $\mathbf{q}_i = \widetilde{\mathbf{q}}_i$, $\mathbf{Q}_i = \widetilde{\mathbf{Q}}_i$, $\mathbf{H}_i = \widetilde{\mathbf{H}}_i$, $\mathbf{z}_i = \widetilde{\mathbf{z}}_i$, $\mathbf{y}_i = \widetilde{\mathbf{y}}_i$ und $\mathbf{x}_i = \widetilde{\mathbf{x}}_i$ folgen direkt aus den Lemmata 6.15, 6.18 und 6.19. Zu zeigen ist noch, daß $\mathbf{r}_i$ das Residuum ist. Aus $(\mathbf{I} - \delta\mathbf{T}_i)\mathbf{x}_i = \mathbf{b}$ folgt mit $\mathbf{T}_i = \mathbf{Q}_i\mathbf{Q}_i^T\mathbf{A}$ und $\mathbf{x}_i = \mathbf{Q}_i\mathbf{y}_i$

$$\mathbf{Q}_i\mathbf{y}_i - \delta\mathbf{Q}_i\mathbf{Q}_i^T\mathbf{A}\mathbf{Q}_i\mathbf{y}_i = \mathbf{b}.$$

Damit hat das Residuum $(\mathbf{I} - \delta\mathbf{A})\mathbf{x}_i - \mathbf{b}$ die Darstellung

$$(\mathbf{I} - \delta\mathbf{A})\mathbf{x}_i - \mathbf{b} = \mathbf{Q}_i\mathbf{y}_i - \delta\mathbf{A}\mathbf{Q}_i\mathbf{y}_i - \mathbf{b} = -\delta(\mathbf{I} - \mathbf{Q}_i\mathbf{Q}_i^T)\mathbf{A}\mathbf{Q}_i\mathbf{y}_i.$$

Die Matrix $(\mathbf{I} - \mathbf{Q}_i\mathbf{Q}_i^T)\mathbf{A}\mathbf{Q}_i$ hat aber nach Lemma 6.16 gerade die Form

$$(\mathbf{I} - \mathbf{Q}_i\mathbf{Q}_i^T)\mathbf{A}\mathbf{Q}_i = (\mathbf{0}, \ldots, \mathbf{0}, \mathbf{u}_{i+1,i}),$$

woraus

$$(\mathbf{I} - \delta\mathbf{A})\mathbf{x}_i - \mathbf{b} = -\delta(\mathbf{y}_i)_i\mathbf{v}_{i+1,i} = \mathbf{r}_i$$

folgt. □

Bemerkung 6.20. Im Gegensatz zu den anderen hier erwähnten Iterationsverfahren steigt der Rechenaufwand pro Iterationsschritt beim Arnoldi-Verfahren mit zunehmender Iteration an. Man kann zeigen, daß der Aufwand des Verfahrens von der Ordnung $O(\varkappa^2 n)$ ist, wobei $\varkappa$ die Anzahl der verwendeten Iterationsschritte ist. Das Arnoldi-Verfahren ist daher in der Regel nur für solche Gleichungssysteme effizient, die mit einer geringen Anzahl von Iterationen auskommen. □

6.4.2 Effiziente Lösung der linearen Gleichungssysteme aus Algorithmus 6.12

Die spezielle Struktur der Matrizen $\mathbf{H}_i$ und der rechten Seite $\mathbf{z}_i$ in dem linearen Gleichungssystem (6.23) kann ausgenutzt werden, um den Aufwand zur Lösung dieses Systems deutlich zu senken. Die Matrizen $\mathbf{H}_i,\ i = 1, \ldots, \max_{\mathrm{It}}$ sind nach Lemma 6.17 ineinander eingebettet und haben wegen Lemma 6.10 Hessenberg-Form. Nach Lemma 6.19 haben auch die Vektoren $\mathbf{z}_i$ die Einbettungs-Eigenschaft

$$\mathbf{z}_i = \begin{pmatrix} \mathbf{z}_{i-1} \\ 0 \end{pmatrix}, \qquad i = 2, \ldots, \max_{\mathrm{It}}.$$

Wir setzen $\mathbf{M}_i = \mathbf{I} - \delta\mathbf{H}_i$ und lösen das System $\mathbf{M}_i\mathbf{y}_i = \mathbf{z}_i$ aus (6.23) durch LU-Zerlegung ohne Pivotisierung, siehe etwa Hämmerlin/Hoffmann [80, Abschnitt 2.1.3], d.h. es sei $\mathbf{M}_i = \mathbf{L}_i\mathbf{U}_i$, wobei $\mathbf{L}_i$ eine untere Dreiecksmatrix mit Diagonalelementen gleich 1 und $\mathbf{U}_i$ eine obere Dreiecksmatrix ist. Das Gleichungssystem $\mathbf{M}_i\mathbf{y}_i = \mathbf{z}_i$ zerfällt in die beiden Dreiecksysteme $\mathbf{L}_i\mathbf{a}_i = \mathbf{z}_i$ und $\mathbf{U}_i\mathbf{y}_i = \mathbf{a}_i$, die nacheinander gelöst werden. Die Matrizen $\mathbf{M}_i$ und $\mathbf{L}_i$ sind wie $\mathbf{H}_i$ Hessenberg-Matrizen. Die Einbettungs-Eigenschaft der $\mathbf{H}_i$ und $\mathbf{z}_i$ überträgt sich auf die Matrizen $\mathbf{M}_i,\ \mathbf{L}_i,\ \mathbf{U}_i$ und den Vektor $\mathbf{a}_i$. Daher müssen in jedem Iterationsschritt nur die neu hinzugekommenen Elemente dieser Matrizen und Vektoren berechnet werden. Das leistet der folgende Algorithmus.

Algorithmus 6.21 (Effiziente Lösung des Systems $(\mathbf{I} - \delta\mathbf{H}_i)\mathbf{y}_i = \mathbf{z}_i$).

Schritt 1: neue Elemente der Matrix $\mathbf{M}_i$ hinzufügen

for $j = 1, \ldots, i-1$

$m_{ji} = -\delta h_{ji}$

end

if $i > 1$

$m_{i,i-1} = -\delta h_{i,i-1}$

end

$m_{ii} = 1 - \delta h_{ii}$

Schritt 2: LU-Zerlegung

for $j = 2, \ldots, i-1$

$m_{ji} = m_{ji} - m_{j,j-1} m_{j-1,i}$

end

if $i > 1$

$m_{i,i-1} = m_{i,i-1} / m_{i-1,i-1}$
$m_{ii} = m_{ii} - m_{i,i-1} m_{i-1,i}$

end

Schritt 3: löse $\mathbf{L}_i \mathbf{a}_i = \mathbf{z}_i$

if $i = 1$

$a_1 = |\mathbf{b}|$

else

$a_i = -m_{i,i-1} a_{i-1}$

end

Schritt 4: löse $\mathbf{U}_i \mathbf{y}_i = \mathbf{a}_i$

for $j = i, i-1, \ldots, 1$

$(\mathbf{y}_i)_j = \left(a_j - \sum_{k=j+1}^{i} m_{jk} (\mathbf{y}_i)_k \right) / m_{jj}$

end

□

Der Ausdruck $(\mathbf{y}_i)_j$ in der vorletzten Zeile des Algorithmus steht für das j-te Element des Vektors $\mathbf{y}_i$. Wir werden diese Schreibweise auch im folgenden gelegentlich verwenden.

Für eine effiziente Lösung muß die Zeile (6.23) in Algorithmus 6.12 durch den Algorithmus 6.21 ersetzt werden. Die Zahlen m_{jk} in Algorithmus 6.21 entsprechen in Schritt 1 den Elementen der Matrix $\mathbf{M}_i$ und werden in Schritt 2 zu Elementen der LU-Matrix $\widetilde{\mathbf{M}}_i = \mathbf{L}_i + \mathbf{U}_i - \mathbf{I}$ umgewandelt. Anstelle von $\mathbf{L}_i$ und $\mathbf{U}_i$ wird nur $\widetilde{\mathbf{M}}_i$ gespeichert.

6.4.3 Approximation äußerer Eigenvektoren

Wird ein lineares Gleichungssystem $(\mathbf{I} - \delta\mathbf{A})\mathbf{x} = \mathbf{b}$ mit dem Arnoldi-Verfahren gelöst, so ist die Näherungslösung $\mathbf{x}_i$ entlang bestimmter Eigenrichtungen von $\mathbf{A}$ besonders genau. Wir bezeichnen mit $\mathbf{v}_k$ die Eigenvektoren und mit λ_k die zugehörigen Eigenwerte von $\mathbf{A}$ sowie mit $G \subset \mathbb{C}$ ein Gebiet, das das Spektrum von $\mathbf{A}$ enthält und hinreichend eng umschließt. Ferner sei $\vartheta(\mathcal{K}_i, \mathbf{v}_k)$ der Winkel zwischen dem nach i Iterationsschritten erzeugten Krylovraum $\mathcal{K}_i$ und dem Eigenvektor $\mathbf{v}_k$. In numerischen Untersuchungen stellt man fest, daß $\vartheta(\mathcal{K}_i, \mathbf{v}_k)$ für $i \to \infty$ besonders schnell gegen 0 geht, wenn der zugehörige Eigenwert λ_k in der Nähe des Randes von G liegt. Wir bezeichnen im folgenden derartige λ_k und $\mathbf{v}_k$ als **äußere Eigenwerte** bzw. **-vektoren** von $\mathbf{A}$, siehe Abbildung 6.2.

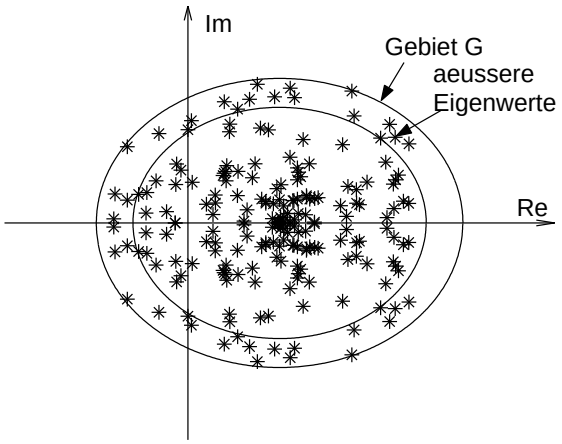


Abbildung 6.2: Gebiet G, äußere Eigenwerte

Die Approximation der äußeren Eigenvektoren durch den Krylovraum macht das Arnoldi-Verfahren besonders für die Lösung der Gleichungssysteme attraktiv, die in impliziten Diskretisierungen gewöhnlicher Differentialgleichungen auftreten, da diese Eigenschaft der Arnoldi-Iteration sich positiv auf die Stabilität der impliziten Verfahren auswirkt. Auf diesen Sachverhalt werden wir im Zusammenhang mit den Krylov-W-Verfahren in Kapitel 7 noch näher eingehen.

Für den Spezialfall, daß $\mathbf{A}$ ein rein reelles Spektrum besitzt, kann der oben eingeführte Winkel $\vartheta(\mathcal{K}_i, \mathbf{v}_k)$ nach oben abgeschätzt werden.

Satz 6.22. *Das System* $(\mathbf{I} - \delta\mathbf{A})\mathbf{x} = \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{N\times N}$ *werde mit dem in Abschnitt 6.4.1 angegebenen Arnoldi-Verfahren gelöst. Die Eigenwerte* $\lambda_1 > \cdots > \lambda_N$ *von* $\mathbf{A}$ *seien sämtlich reell und*

einfach. Sei $\mathcal{K}_i = \operatorname{span}\{\mathbf{b}, \mathbf{Ab}, \ldots, \mathbf{A}^{i-1}\mathbf{b}\}$ *der Krylov-Raum, in dem die Näherungslösung* $\mathbf{x}_i$ *liegt. Ferner seien die folgenden Größen gegeben.*

$$\varkappa_1 = 1, \qquad \varkappa_i = \prod_{j=1}^{i-1} \frac{\lambda_j - \lambda_N}{\lambda_j - \lambda_i}, \quad \textit{falls } i > 1, \qquad \gamma_i = 1 + 2\frac{\lambda_i - \lambda_{i+1}}{\lambda_{i+1} - \lambda_N}$$

$$\widetilde{\varkappa}_i = \prod_{j=i+1}^{N} \frac{\lambda_1 - \lambda_j}{\lambda_i - \lambda_j}, \quad \textit{falls } i < N, \qquad \widetilde{\varkappa}_N = 1, \qquad \widetilde{\gamma}_i = 1 + 2\frac{\lambda_{i-1} - \lambda_i}{\lambda_1 - \lambda_{i-1}}.$$

Es sei $\vartheta(\mathcal{K}_i, \mathbf{v}_k)$ *der Winkel zwischen dem Krylovraum* $\mathcal{K}_i$ *und dem Eigenvektor* $\mathbf{v}_k$ *von* $\mathbf{A}$*. Dann gelten die Abschätzungen*

$$\tan\vartheta(\mathcal{K}_i, \mathbf{v}_k) \le \frac{\varkappa_k}{T_{i-k}(\gamma_k)} \tan\vartheta(\mathcal{K}_1, \mathbf{v}_k), \qquad \textit{falls } k \le i \tag{6.25}$$

und

$$\tan\vartheta(\mathcal{K}_i, \mathbf{v}_k) \le \frac{\widetilde{\varkappa}_k}{T_{i+k-N-1}(\widetilde{\gamma}_k)} \tan\vartheta(\mathcal{K}_1, \mathbf{v}_k), \qquad \textit{falls } k \ge N+1-i \textit{ ist,} \tag{6.26}$$

wobei $T_i(x) = \frac{1}{2}\left(\left(x + \sqrt{x^2-1}\right)^i + \left(x - \sqrt{x^2-1}\right)^i\right)$ *das Tschebyschev-Polynom erster Art vom Grade* i *ist.*

Beweis. Die Abschätzung (6.25) wurde von Saad in [139] und [140] gezeigt. Aus diesem Beweis läßt sich auch die Abschätzung 6.26 durch einfache Symmetrieargumente herleiten. □

Die Abschätzungen in Satz 6.22 sind für $k = 1$ und $k = N$ optimal, wie für (6.25) zeigt. In vielen anderen Fällen sind sie jedoch sehr grob. Für äquidistant verteilte Eigenwerte und $N \gg 1$ etwa wächst $\varkappa_k$ bereits für $k = 2, 3, \ldots$ sehr stark an, was durch den Faktor $T_{i-k}(\gamma_k)$ nicht kompensiert werden kann, so daß die angegebenen Schranken nicht mehr relevant sind. Falls $k \approx 1$ ist und γ_k nicht zu nahe bei 1 liegt, so geht der Ausdruck $1/T_{i-k}(\gamma_k)$ schnell gegen 0. Analoges gilt für $k \approx N$. Im diesem Fall läßt sich demnach für die extremalen Eigenvektoren zeigen, daß sie relativ schnell durch den Krylovraum $\mathcal{K}_i$ approximiert werden.

Numerische Untersuchungen zeigen jedoch, daß vielfach auch in Fällen, in denen die angegebenen Schranken sehr grob sind, die äußeren Eigenvektoren schneller als die inneren durch den Krylovraum approximiert werden. Wir verdeutlichen das an dem folgenden numerischen Beispiel.

Untersuchung 6.23. Wir betrachten die Konvektions-Diffusions-Gleichung

$$u_t = \Delta u + \mathbf{b} \cdot \nabla u, \qquad \mathbf{b} = \begin{pmatrix} b_1(t) \\ b_2(t) \end{pmatrix} = c \begin{pmatrix} -\sin\ t \\ \cos\ t \end{pmatrix}$$

im Gebiet $\Omega =]-5, 5[^2$ mit Anfangsbedingung $u(x, y, 0) = e^{-(x^2+y^2)}$ und homogenen Dirichlet-Randbedingungen $u(x, y, t) = 0$ für $(x, y) \in \partial\Omega$. Die Diskretisierung erfolge auf einem äquidistanten quadratischen Gitter der Maschenweite $h = 1$. Die Ortsdiskretisierung werde mit zentralen Differenzen vorgenommen. Es sei $N = 10/h - 1$. Bei lexikographischer Numerierung der inneren Gitterpunkte, d.h. $\mathbf{x}_{(i-1)N+j} = (-5+jh, -5+ih)$, $i, j = 1, \ldots, N$, siehe Abbildung 6.3, ergibt sich nach der Ortsdiskretisierung das System gewöhnlicher Differentialgleichungen

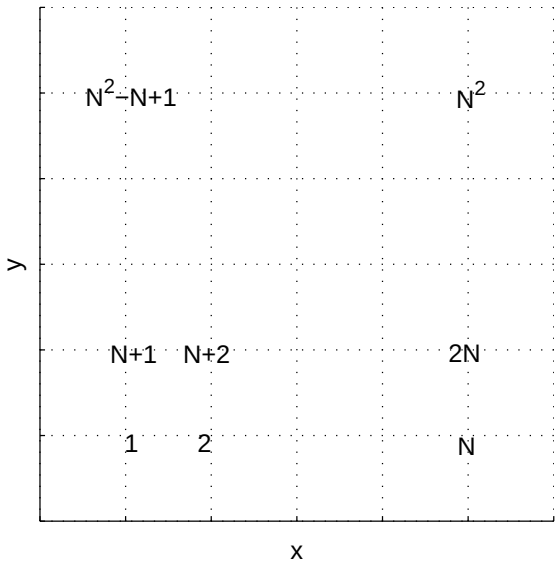


Abbildung 6.3: Lexikographische Anordnung der inneren Gitterpunkte

$$\mathbf{u}_t = \mathbf{A}\mathbf{u}, \tag{6.27}$$

wobei $\mathbf{u}$ die Näherung für $u(\mathbf{x}, \cdot)$ ist und $\mathbf{A} \in \mathbb{R}^{N^2 \times N^2}$ sich wie folgt aus $N \times N$-Blockmatrizen zusammensetzt.

$$\mathbf{D}_0 = \frac{1}{h^2}\begin{pmatrix} -4 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -4 \end{pmatrix}, \qquad \mathbf{D}_1 = \frac{b_1(t)}{2h}\begin{pmatrix} 0 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 0 \end{pmatrix}, \qquad \mathbf{D} = \mathbf{D}_0 + \mathbf{D}_1,$$

$$\mathbf{R} = \left(\frac{1}{h^2} + \frac{b_2(t)}{2h}\right)\mathbf{I}, \qquad \mathbf{L} = \left(\frac{1}{h^2} - \frac{b_2(t)}{2h}\right)\mathbf{I}, \qquad \mathbf{A} = \begin{pmatrix} \mathbf{D} & \mathbf{R} & & \\ \mathbf{L} & \ddots & \ddots & \\ & \ddots & \ddots & \mathbf{R} \\ & & \mathbf{L} & \mathbf{D} \end{pmatrix}$$

Wird das System (6.27) etwa mit dem impliziten Euler-Verfahren gelöst, so erhält man im ersten Zeitschritt der Länge $\tau_0 = 0{,}01$ das Gleichungssystem

$$(\mathbf{I} - \tau_0\mathbf{A})\mathbf{u}_1 = \mathbf{u}_0. \tag{6.28}$$

Dabei sind $\mathbf{u}_0$ und $\mathbf{u}_1$ die diskreten Lösungen zu den Zeiten 0 und τ_0. Wir wählen im folgenden den Konvektionskoeffizienten $c \in \{0, 1, 2\}$. Die Eigenwerte von $\mathbf{A}$ sind in diesem Falle sämtlich reell; wir bezeichnen sie mit λ_k, die zugehörigen Eigenvektoren mit $\mathbf{v}_k$.

Das System 6.28 werde nun mit dem Arnoldi-Verfahren gelöst. Wie oben sei $\vartheta(\mathcal{K}_i, \mathbf{v}_k)$ der Winkel zwischen dem Krylovraum $\mathcal{K}_i$ und dem Eigenvektor $\mathbf{v}_k$. In der Graphik 6.4 stellen wir für $i = 10, 20, 30$ und $c = 2$ den Winkel $\vartheta(\mathcal{K}_i, \mathbf{v}_k)$ über dem zugehörigen Eigenwert λ_k dar. Wählen wir stattdessen $c = 1$, so erhalten wir das in Abbildung 6.5 dargestellte Verhalten.

Ergebnisse der Untersuchung. Man erkennt, daß für die äußeren Eigenwerte der Winkel $\vartheta(\mathcal{K}_i, \mathbf{v}_k)$ besonders schnell gegen 0 geht, daß also der Krylov-Raum $\mathcal{K}_i$ sich den äußeren Eigenvektoren am schnellsten annähert. Interessant ist, daß die Approximation der Eigenvektoren wesentlich von dem Konvektionsterm abhängt. Für $c = 1$ fällt die Approximation schwächer als für $c = 2$ aus, siehe Abbildung 6.5. Noch geringer ist die – hier nicht graphisch dargestellte – Approximation für $c = 0$. □

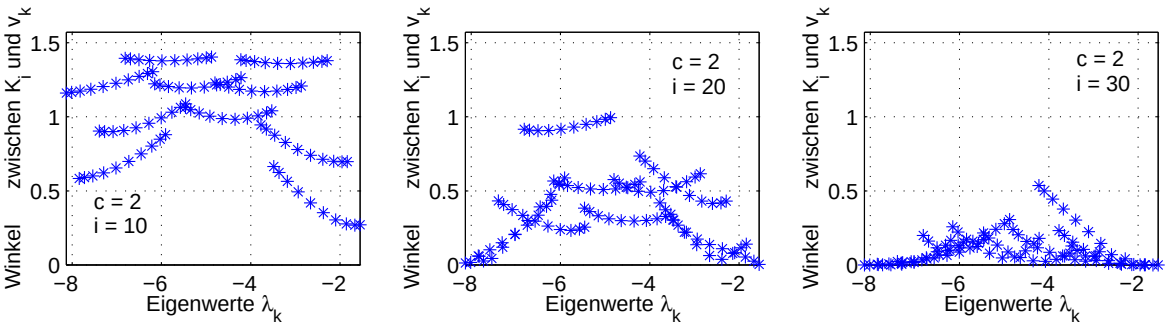


Abbildung 6.4: Winkel $\vartheta(\mathcal{K}_i, \mathbf{v}_k)$ für $c = 2$

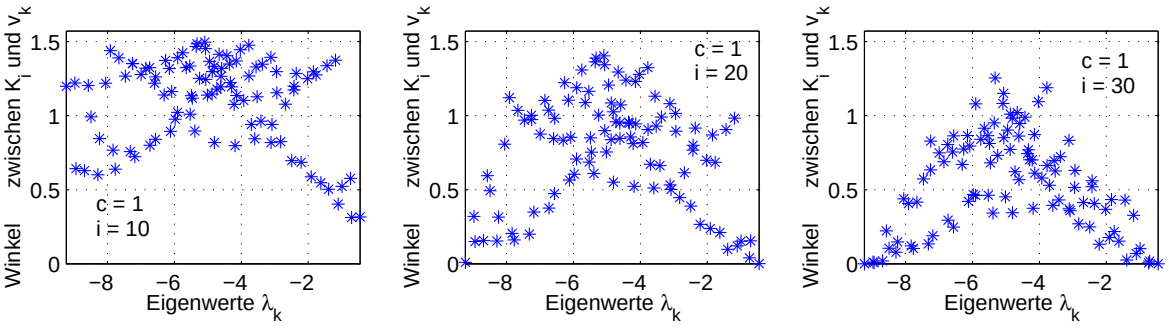


Abbildung 6.5: Winkel $\vartheta(\mathcal{K}_i, \mathbf{v}_k)$ für $c = 1$

6.5 Der multiple Arnoldi-Prozeß

In einigen Anwendungen, insbesondere bei den in Abschnitt 5.7 beschriebenen W-Methoden, treten mehrere Gleichungssysteme der Form

$$(\mathbf{I} - \delta\mathbf{A})\mathbf{x}^i = \mathbf{b}^i, \qquad i = 1, \ldots, s$$

mit gleicher Systemmatrix auf. Dabei kann die rechte Seite $\mathbf{b}^i$ von der Lösung $\mathbf{x}^j$ abhängen, falls $j < i$ ist. Für diesen Fall konstruierten Schmitt und Weiner [143, 168] (1995) den **multiplen Arnoldi-Prozeß**, ein Verfahren, bei dem jedes der s Gleichungssysteme mit dem Arnoldi-Verfahren gelöst wird. Es sei $\mathcal{K}^i,\ i = 1, \ldots, s$ der durch das Arnoldi-Verfahren im i-ten System aufgebaute Krylov-Raum. Wichtige Eigenschaften des multiplen Arnoldi-Prozesses sind

- die Einbettung der Krylov-Räume, d.h. $\mathcal{K}^1 \subset \cdots \subset \mathcal{K}^s$ und

- das Einfügen der rechten Seite $\mathbf{b}^i$ des i-ten Systems in den Krylov-Raum $\mathcal{K}^i$.

Während der erste Punkt die Effizienz des Prozesses erhöht, wirkt sich der zweite Punkt positiv auf das Konvergenzverhalten des Arnoldi-Verfahrens aus.

Wir geben im folgenden einen Algorithmus für den multiplen Arnoldi-Prozeß an; dieser ist in etwas verkürzter Form in Weiner und Schmitt [168] zu finden. Um die Darstellung wesentlich zu vereinfachen, lassen wir die Iterationsindizes von Matrizen und Vektoren weg. Gemeint ist jeweils die aktuelle Iterierte. In Übereinstimmung mit Schmitt und Weiner [143, 168] bezeichnen wir die maximal zulässigen Dimensionen des Krylov-Raums mit $\varkappa_1, \ldots, \varkappa_s$.

Zunächst definieren wir die folgenden Funktionen auf Matrizen.

Definition 6.24. Es sei $\mathbf{A}$ eine Matrix. Dann bezeichne

- $\mathrm{col}_i(\mathbf{A})$ die i-te Spalte von $\mathbf{A}$, (col – „column“),
- $\mathrm{nc}(\mathbf{A})$ die Anzahl der Spalten von $\mathbf{A}$, (nc – „number of columns“),
- $\mathrm{nzc}_i(\mathbf{A})$ die von links gesehen i-te vom Nullvektor verschiedene Spalte von $\mathbf{A}$, (nzc – „non-zero column“),
- $\mathrm{nnzc}(\mathbf{A})$ die Anzahl der Spaltenvektoren von $\mathbf{A}$, die nicht gleich dem Nullvektor sind, (nnzc – „number of non-zero columns“). □

Algorithmus 6.25 (Multipler Arnoldi-Prozeß – Kurzform).

gegeben: $\mathbf{A},\ \delta,\ \varkappa_1,\ \varkappa_2,\ \varkappa_3,\ \mathbf{b}^i,\ i = 1, \ldots, s$

System 1:

$\mathbf{q} = \mathbf{b}^1 / \|\mathbf{b}^1\|$

$\mathbf{Q} = (\mathbf{q})$

for $j = 2, 3, \ldots, \varkappa_1$

$\mathbf{v} = (\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)\mathbf{A}\mathbf{q}$ (Gram-Schmidt-Orthogonalisierung)

$\mathbf{q} = \mathbf{v}/|\mathbf{v}|$

$\mathbf{Q} = \begin{pmatrix} \mathbf{Q} & \mathbf{q} \end{pmatrix}$

$\mathbf{H} = \mathbf{Q}^T\mathbf{A}\mathbf{Q}$

löse $(\mathbf{I} - \delta\mathbf{H})\mathbf{y}^1 = \mathbf{Q}^T\mathbf{b}^1$

$\mathbf{r} = -\delta(\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)\mathbf{A}\mathbf{Q}\mathbf{y}^1$ (Residuum)

if $\|\mathbf{r}\|_* < TOL_{\mathrm{LSS}}$

break

end

end

Systeme 2 *bis s:*

for $i = 2, \ldots, s$

$\mathbf{v} = (\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)\mathbf{b}^i$ (Gram-Schmidt-Orthogonalisierung)

if $\mathbf{v} \neq \mathbf{0}$

$\mathbf{q} = \mathbf{v}/|\mathbf{v}|$

$\mathbf{Q} = \begin{pmatrix}\mathbf{Q} & \mathbf{q}\end{pmatrix}$

end

$\mathbf{H} = \mathbf{Q}^T\mathbf{A}\mathbf{Q}$ (6.29)

löse $(\mathbf{I} - \delta\mathbf{H})\mathbf{y}^i = \mathbf{Q}^T\mathbf{b}^i$

$\mathbf{G} = (\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)\mathbf{A}\mathbf{Q}$

$\mathbf{r} = -\delta\mathbf{G}\mathbf{y}^i$ (Residuum)

if $\|\mathbf{r}\|_* < TOL_{\text{LSS}}$

continue

end

$c = \text{nnzc}(\mathbf{G})$

for $j = 1, \ldots, c$

$\mathbf{v} = \text{nzc}_j(\mathbf{G})$

$\mathbf{q} = \mathbf{v}/|\mathbf{v}|$

$\mathbf{Q} = \begin{pmatrix}\mathbf{Q} & \mathbf{q}\end{pmatrix}$

end

for $k = \varkappa_{i-1}, \ldots, \varkappa_i$

$\mathbf{H} = \mathbf{Q}^T\mathbf{A}\mathbf{Q}$ (6.30)

löse $(\mathbf{I} - \delta\mathbf{H})\mathbf{y}^i = \mathbf{Q}^T\mathbf{b}^i$

$\mathbf{r} = -\delta(\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)\mathbf{A}\mathbf{Q}\mathbf{y}^i$ (Residuum)

if $\|\mathbf{r}\|_* < TOL_{\text{LSS}}$

break

end

for $j = 1, \ldots, c$

$\mathbf{q}_{\text{alt}} = \text{col}_{\text{nc}(\mathbf{Q})-c+1}(\mathbf{Q})$

$\mathbf{v} = (\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)\mathbf{A}\mathbf{q}_{\text{alt}}$ (Gram-Schmidt-Orthogonalisierung)

$\mathbf{q} = \mathbf{v}/|\mathbf{v}|$

$\mathbf{Q} = \begin{pmatrix}\mathbf{Q} & \mathbf{q}\end{pmatrix}$

end

end

end

$\varkappa = \text{nc}(Q)$

for $i = 1, \ldots, s$

$$\mathbf{y}^i = \begin{pmatrix} \mathbf{y}^i \\ \mathbf{0} \end{pmatrix} \in \mathbb{R}^{\varkappa}$$

$$\mathbf{x}^i = \mathbf{Q}\mathbf{y}^i$$

end □

Bemerkung 6.26. Der erste Teil des Algorithmus („System 1“) entspricht weitestgehend dem Algorithmus 6.11. Es wurde lediglich die Abbruchbedingung aufgenommen. Die Matrizen $\mathbf{H}$ haben eine verallgemeinerte Hessenberg-Struktur: Die für das i-te System in (6.29) und (6.30) aufgestellte Matrix $\mathbf{H}$ ist eine obere Dreiecksmatrix mit maximal i zusätzlichen Subdiagonalen. □

Dieser Algorithmus kann in einer effizienten Form programmiert werden, die wir hier jedoch nicht angeben werden. Für den Fall, daß der multiple Arnoldi-Prozeß in einer speziellen W-Methode eingesetzt wird, ist die effiziente Form in Abschnitt 7.2 aufgeführt.

Kapitel 7

Ein Krylov-W-Verfahren

7.1 Eine dreistufige W-Methode

Zur Lösung des Systems gewöhnlicher Differentialgleichungen $\mathbf{u}_t = \mathbf{f}(t, \mathbf{u})$ betrachten wir die in (5.17) angegebene W-Methode

$$\begin{aligned}
(\mathbf{I} - \tau_i\gamma\mathbf{T})\widetilde{\mathbf{k}}^j &= \mathbf{f}\left(t_i + \tau_i c_j, \mathbf{u}_i + \tau_i\gamma\sum_{l=1}^{j-1}\varphi_{lj}\widetilde{\mathbf{k}}^l\right) + \sum_{l=1}^{j-1}\vartheta_{lj}\widetilde{\mathbf{k}}^l, \qquad j = 1, \ldots, s \quad (7.1)\\
\mathbf{u}_{i+1} &= \mathbf{u}_i + \tau_i\sum_{l=1}^{s} g_l\widetilde{\mathbf{k}}^l\\
\widehat{\mathbf{u}}_{i+1} &= \mathbf{u}_i + \tau_i\sum_{l=1}^{s}\widehat{g}_l\widetilde{\mathbf{k}}^l.
\end{aligned}$$

In dieser Methode treten mehrere lineare Gleichungssysteme mit der gleichen Systemmatrix $\mathbf{I} - \tau_i\gamma\mathbf{T}$ auf, die nacheinander gelöst werden können. Verwendet man zur Lösung dieser Systeme den in Abschnitt 6.5 beschriebenen multiplen Arnoldi-Prozeß, so spricht man auch von einem **Krylov-W-Verfahren**[1]. Derartige Verfahren wurden von Schmitt und Weiner [143] (1995) entwickelt. Das Programm ROWMAP von Weiner, Schmitt und Podhaisky [169] (1997) basiert auf einem Krylov-W-Verfahren vierter Ordnung. Wir wollen in den in Kapitel 9 dargestellten numerischen Berechnungen u.a. ein Krylov-W-Verfahren verwenden, dem die W-Methode zweiter Ordnung (5.18)

$$\begin{aligned}
\left(\mathbf{I} - \tau_m\left(1 - \sqrt{2}/2\right)\mathbf{T}\right)\widetilde{\mathbf{k}}^1 &= \mathbf{f}(t_m, \mathbf{u}_m) =: \mathbf{b}^1 \quad (7.2)\\
\left(\mathbf{I} - \tau_m\left(1 - \sqrt{2}/2\right)\mathbf{T}\right)\widetilde{\mathbf{k}}^2 &= \mathbf{f}\left(t_{m+1}, \mathbf{u}_m + \tau_m\widetilde{\mathbf{k}}^1\right) - \left(2 + \sqrt{2}\right)\widetilde{\mathbf{k}}^1 =: \mathbf{b}^2\\
\left(\mathbf{I} - \tau_m\left(1 - \sqrt{2}/2\right)\mathbf{T}\right)\widetilde{\mathbf{k}}^3 &= \mathbf{f}\left(t_{m+1}, \mathbf{u}_m + \tau_m\widetilde{\mathbf{k}}^1\right) - \widetilde{\mathbf{k}}^1 + \left(-1 + \sqrt{2}\right)\widetilde{\mathbf{k}}^2 =: \mathbf{b}^3\\
\mathbf{u}_{m+1} &= \mathbf{u}_m + \frac{\tau}{2}\left(2\widetilde{\mathbf{k}}^1 + \left(1 - \sqrt{2}\right)\widetilde{\mathbf{k}}^2 + \widetilde{\mathbf{k}}^3\right)\\
\widehat{\mathbf{u}}_{m+1} &= \mathbf{u}_m + \frac{\tau}{20}\left(\left(18 - \sqrt{2}\right)\widetilde{\mathbf{k}}^1 + \left(9 - 11\sqrt{2}\right)\widetilde{\mathbf{k}}^2 + \left(11 + \sqrt{2}\right)\widetilde{\mathbf{k}}^3\right)
\end{aligned}$$

[1] Der Name Krylov bezieht sich auf den durch das Arnoldi-Verfahren aufgebauten Krylov-Raum, in dem die numerische Lösung des Systems liegt, siehe Abschnitt 6.4.1

zugrundeliegt.[2] Diese W-Methode wurde ebenfalls von SCHMITT und WEINER [143] zur Konstruktion eines Krylov-W-Verfahrens verwendet.

7.2 Effizienter Algorithmus für den multiplen Arnoldi-Prozeß

In Abschnitt 6.5 wurde der multiple Arnoli-Prozeß in Kurzform angegeben. Durch Ausnutzung der speziellen in der W-Methode (7.2) auftretenden rechten Seiten erhält man die folgende effiziente Form des multiplen Arnoldi-Prozesses. Die Stufen 1 bis 3 in diesem Algorithmus entsprechen den Stufen der W-Methode. Die Größen $\varkappa_1$, $\varkappa_{12}$ und $\varkappa_{23}$ legen die maximal erlaubte Anzahl der Iterationen in den Stufen 1, 2 und 3 fest.

Algorithmus 7.1 (Multipler Arnoldi-Prozeß für die W-Methode (7.2) – effiziente Form).

gegeben: W-Mehtode (7.2), $\varkappa_1$, $\varkappa_{12}$, $\varkappa_{23}$

setze $\delta = \tau_m \left(1 - \sqrt{2}/2\right)$, $\mathbf{A} = \mathbf{T}$

Stufe 1:

$\mathbf{b}^1 = \mathbf{f}(t_m, u_m)$

$\mathbf{q}_1 = \mathbf{b}^1/|\mathbf{b}^1|$

for $i = 1, \ldots, \varkappa_1$

$\mathbf{u}_{1i} = \mathbf{A}\mathbf{q}_i$

for $j = 1, \ldots, i$

$h_{ji} = \mathbf{q}_j^T \mathbf{u}_{ji}$

$\mathbf{u}_{j+1,i} = \mathbf{u}_{ji} - h_{ji}\mathbf{q}_j$

end

$h_{i+1,i} = |\mathbf{u}_{i+1,i}|$

$$\mathbf{H}_i = \begin{pmatrix} h_{11} & \cdots & h_{1i} \\ \vdots & & \vdots \\ h_{i1} & \cdots & h_{ii} \end{pmatrix}$$

$\mathbf{z}_i^1 = (|\mathbf{b}^1|, 0, \ldots, 0)^T \in \mathbb{R}^i$

löse $(\mathbf{I} - \delta\mathbf{H}_i)\mathbf{y}_i^1 = \mathbf{z}_i^1$ (7.3)

$\mathbf{r}_i^1 = -\delta(\mathbf{y}_i^1)_i \mathbf{u}_{i+1,i}$ (Residuum)

if $\|\mathbf{r}_i^1\|_* \leq TOL_{\text{LSS}}$

$\varkappa_1 = i$

break

[2] Um eine Überschneidung mit Indizes in dem später angegebenen multiplen Arnoldi-Prozeß zu vermeiden, wurde der Index i hier durch m ersetzt.

end

$\mathbf{q}_{i+1} = \mathbf{u}_{i+1,i}/|\mathbf{u}_{i+1,i}|$

end

$\mathbf{Q}_{\varkappa_1} = (\mathbf{q}_1, \ldots, \mathbf{q}_{\varkappa_1})$

$\mathbf{x}^1 = \mathbf{Q}_{\varkappa_1}\mathbf{y}^1_{\varkappa_1}$

$\widetilde{\mathbf{k}}_1 = \mathbf{x}^1$

Stufe 2:

$\varkappa_2 = \varkappa_1 + \varkappa_{12}$

$\mathbf{b}^2 = \mathbf{f}\left(t_{m+1}, \mathbf{u}_m + \tau_m\widetilde{\mathbf{k}}_1\right) - \left(2+\sqrt{2}\right)\widetilde{\mathbf{k}}_1$

$\mathbf{w}_1 = \mathbf{b}^2$

for $i = 1, \ldots, \varkappa_1$

$z_i = \mathbf{q}_i^T\mathbf{b}^2$

$\mathbf{w}_{i+1} = \mathbf{w}_i - z_i\mathbf{q}_i$

end

$z_{\varkappa_1+1} = |\mathbf{w}_{\varkappa_1+1}|$

$\mathbf{z}^2_{\varkappa_2} = (z_1, \ldots, z_{\varkappa_1+1}, 0, \ldots, 0)^T \in \mathbb{R}^{\varkappa_2}$

$\mathbf{q}_{\varkappa_1+1} = \mathbf{w}_{\varkappa_1+1}/z_{\varkappa_1+1}$

$h_{\varkappa_1+1,\varkappa_1} = \mathbf{q}^T_{\varkappa_1+1}\mathbf{u}_{\varkappa_1+1,\varkappa_1}$ ($\mathbf{u}_{\varkappa_1+1,\varkappa_1}$ aus Stufe 1)

$\mathbf{u}_{\varkappa_1+2,\varkappa_1} = \mathbf{u}_{\varkappa_1+1,\varkappa_1} - h_{\varkappa_1+1,\varkappa_1}\mathbf{q}_{\varkappa_1+1}$

$h_{\varkappa_1+2,\varkappa_1} = |\mathbf{u}_{\varkappa_1+2,\varkappa_1}|$

$\mathbf{q}_{\varkappa_1+2} = \mathbf{u}_{\varkappa_1+2,\varkappa_1}/h_{\varkappa_1+2,\varkappa_1}$

for $i = \varkappa_1 + 1, \ldots, \varkappa_2$

$\mathbf{u}_{1i} = \mathbf{A}\mathbf{q}_i$

for $j = 1, \ldots, i+1$

$h_{ji} = \mathbf{q}_j^T\mathbf{u}_{ji}$

$\mathbf{u}_{j+1,i} = \mathbf{u}_{ji} - h_{ji}\mathbf{q}_j$

end

$h_{i+2,i} = |\mathbf{u}_{i+2,i}|$

$$\mathbf{H}_i = \begin{pmatrix} h_{11} & \cdots & h_{1,i} \\ \vdots & & \vdots \\ h_{i,1} & \cdots & h_{ii} \end{pmatrix}$$

$\mathbf{z}_i^2 = (z_1, \ldots, z_{\varkappa_1+1}, 0, \ldots, 0)^T \in \mathbb{R}^i$

löse $(\mathbf{I} - \delta \mathbf{H}_i)\mathbf{y}_i^2 = \mathbf{z}_i^2$ (7.4)

$\mathbf{q}_{i+2} = \mathbf{u}_{i+2,i}/h_{i+2,i}$

$\mathbf{r}_i^2 = -\delta \left(\left(h_{i+1,i-1}(\mathbf{y}_i^2)_{i-1} + h_{i+1,i}(\mathbf{y}_i^2)_i \right) \mathbf{q}_{i+1} + (\mathbf{y}_i^2)_i \mathbf{u}_{i+2,i} \right)$ (Residuum)

if $\|\mathbf{r}_i^2\|_* \leq TOL_{\text{LSS}}$

$\varkappa_2 = i$

break

end

end

Stufe 3:

$\varkappa_3 = \varkappa_2 + \varkappa_{23}$

$\mathbf{y}_{\varkappa_2}^1 = (\mathbf{y}_{\varkappa_1}^1 \ \mathbf{0})^T \in \mathbb{R}^{\varkappa_2}$

$\mathbf{z}_{\varkappa_2}^3 = \mathbf{z}_{\varkappa_2}^2 + (1+\sqrt{2})\mathbf{y}_{\varkappa_2}^1 + (-1+\sqrt{2})\mathbf{y}_{\varkappa_2}^2$

$$\mathbf{H}_{\varkappa_2} = \begin{pmatrix} h_{11} & \cdots & h_{1,\varkappa_2} \\ \vdots & & \vdots \\ h_{\varkappa_2,1} & \cdots & h_{\varkappa_2,\varkappa_2} \end{pmatrix}$$

löse $(\mathbf{I} - \delta \mathbf{H}_{\varkappa_2})\mathbf{y}_{\varkappa_2}^3 = \mathbf{z}_{\varkappa_2}^3$ (7.5)

$\mathbf{r}_{\varkappa_2}^3 = -\delta \left(\left(h_{\varkappa_2+1,\varkappa_2-1}(\mathbf{y}_{\varkappa_2}^3)_{\varkappa_2-1} + h_{\varkappa_2+1,\varkappa_2}(\mathbf{y}_{\varkappa_2}^3)_{\varkappa_2} \right) \mathbf{q}_{\varkappa_2+1} + (\mathbf{y}_{\varkappa_2}^3)_{\varkappa_2} \mathbf{u}_{\varkappa_2+2,\varkappa_2} \right)$

(Residuum)

if $\|\mathbf{r}_{\varkappa_2}^3\|_* \leq TOL_{\text{LSS}}$

$\varkappa_3 = \varkappa_2$

break

end

for $i = \varkappa_2 + 1, \ldots, \varkappa_3$

$\mathbf{u}_{1i} = \mathbf{A}\mathbf{q}_i$

for $j = 1, \ldots, i+1$

$h_{ji} = \mathbf{q}_j^T \mathbf{u}_{ji}$

$\mathbf{u}_{j+1,i} = \mathbf{u}_{ji} - h_{ji}\mathbf{q}_j$

end

$h_{i+2,i} = \|\mathbf{u}_{i+2,i}\|$

$\mathbf{H}_i = \begin{pmatrix} h_{11} & \cdots & h_{1i} \\ \vdots & & \vdots \\ h_{i1} & \cdots & h_{ii} \end{pmatrix}$

$\mathbf{z}_i^3 = (\mathbf{z}_{\varkappa_2}^3\ \mathbf{0})^T \in \mathbb{R}^i$

löse $(\mathbf{I} - \delta\mathbf{H}_i)\mathbf{y}_i^3 = \mathbf{z}_i^3$ (7.6)

$\mathbf{q}_{i+2} = \mathbf{u}_{i+2,i}/h_{i+2,i}$

$\mathbf{r}_i^3 = -\delta\left(\left(h_{i+1,i-1}(\mathbf{y}_i^3)_{i-1} + h_{i+1,i}(\mathbf{y}_i^3)_i\right)\mathbf{q}_{i+1} + (\mathbf{y}_i^3)_i\mathbf{u}_{i+2,i}\right)$ (Residuum)

if $\|\mathbf{r}_i^3\|_* \le TOL_{\mathrm{LSS}}$

$\varkappa_3 = i$

break

end

end

Lösung zum neuen Zeitschritt:

$\mathbf{y}_{\varkappa_3}^1 = (\mathbf{y}_{\varkappa_1}^1\ \mathbf{0})^T \in \mathbb{R}^{\varkappa_3}$

$\mathbf{y}_{\varkappa_3}^2 = (\mathbf{y}_{\varkappa_2}^2\ \mathbf{0})^T \in \mathbb{R}^{\varkappa_3}$

$\mathbf{Q}_{\varkappa_3} = (\mathbf{q}_1, \ldots, \mathbf{q}_{\varkappa_3})$

$\mathbf{u}(t+\tau) = \mathbf{u}(t) + \frac{\tau}{2}\mathbf{Q}_{\varkappa_3}\left(2\mathbf{y}_{\varkappa_3}^1 + (1-\sqrt{2})\mathbf{y}_{\varkappa_3}^2 + \mathbf{y}_{\varkappa_3}^3\right)$

$\widehat{\mathbf{u}}(t+\tau) = \mathbf{u}(t) + \frac{\tau}{20}\mathbf{Q}_{\varkappa_3}\left((18-\sqrt{2})\mathbf{y}_{\varkappa_3}^1 + (9-11\sqrt{2})\mathbf{y}_{\varkappa_3}^2 + (11+\sqrt{2})\mathbf{y}_{\varkappa_3}^3\right)$

□

Bemerkung 7.2. Wie in Abschnitt 7.5 noch ausgeführt werden wird, sollte die zur Residuenabschätzung verwendete Norm $\|\cdot\|_*$ die gleiche Norm sein, die auch in der Zeitschrittsteuerung verwendet wird. Häufig benutzt man die skalierte Euklidische Norm

$$\|(v_1, \ldots, v_N)\|_* := |(v_1, \ldots, v_N)|/\sqrt{N}.$$

In diesem Falle gilt für die Norm der Residuen $\|\mathbf{r}_i^k\|_*,\ k = 2, 3$ die Abschätzung

$$\begin{aligned} \|\mathbf{r}_i^k\|_* &\le |\delta|\left(\left|h_{i+1,i-1}(\mathbf{y}_i^k)_{i-1} + h_{i+1,i}(\mathbf{y}_i^k)_i\right|\|\mathbf{q}_{i+1}\|_* + |(\mathbf{y}_i^k)_i|\|\mathbf{u}_{i+2,i}\|_*\right) \\ &= |\delta|\left(\left|h_{i+1,i-1}(\mathbf{y}_i^k)_{i-1} + h_{i+1,i}(\mathbf{y}_i^k)_i\right| + h_{i+2,i}|(\mathbf{y}_i^k)_i|\right)/\sqrt{N} =: r_{i,\mathrm{est}}^k. \end{aligned} \quad (7.7)$$

Die Berechnung von $\|\mathbf{r}_i^k\|_*$ ist aufwendiger als die von $r_{i,\mathrm{est}}^k$. Deshalb kann die Bedingung $\|\mathbf{r}_i^k\|_* \le TOL_{\mathrm{LSS}}$ durch die oft nur geringfügig schärfere Bedingung $r_{i,\mathrm{est}}^k \le TOL_{\mathrm{LSS}}$ ersetzt werden. □

7.3 Effiziente Lösung der linearen Gleichungssysteme in Algorithmus 7.1

Wie bei dem in Abschnitt 6.4.1 beschriebenen einfachen Arnoldi-Verfahren, so kann auch im multiplen Arnoldi-Prozeß die Lösung der in (7.3), (7.4), (7.5) und (7.6) auftretenden linearen Gleichungssysteme

$$(\mathbf{I} - \delta\mathbf{H}_i)\mathbf{y}_i^k = \mathbf{z}_i^k \tag{7.8}$$

auf effiziente Weise erfolgen, indem die LU-Zerlegung vorangegangener Iterationsschritte genutzt wird. Wir bezeichnet die Systemmatrizen $\mathbf{I} - \delta\mathbf{H}_i$ mit $\mathbf{M}_i$. Die Matrix $\mathbf{M}_{\varkappa_3}$ hat jedoch im Unterschied zu $\mathbf{M}_{\varkappa_1}$ nicht mehr Hessenberg-Form, sondern eine bei Spalte $\varkappa_1$ beginnende weitere Subdiagonale, d.h. $\mathbf{M}_{\varkappa_3}$ ist von der Form

$$\mathbf{M}_{\varkappa_3} = \begin{pmatrix} * & \dots & \dots & \dots & \dots & \dots & * \\ * & \ddots & & & & & \vdots \\ & \ddots & \ddots & & & & \vdots \\ & & \ddots & \ddots & & & \vdots \\ & & * & \ddots & \ddots & & \vdots \\ & & & \ddots & \ddots & \ddots & \vdots \\ 0 & & & & * & * & * \end{pmatrix}$$

Wir lösen die Systeme (7.8) durch LU-Zerlegung ohne Pivotisierung und erhalten $\mathbf{M}_i = \mathbf{L}_i\mathbf{U}_i$, $\mathbf{L}_i\mathbf{a}_i^k = \mathbf{z}_i^k$, $\mathbf{U}_i\mathbf{y}_i^k = \mathbf{a}_i^k$. Die Matrizen $\mathbf{L}_i$ und $\mathbf{U}_i$ werden lediglich durch die Matrix $\widetilde{\mathbf{M}}_i = \mathbf{L}_i + \mathbf{U}_i - \mathbf{I}$ gespeichert. Wie bei dem einfachen Arnoldi-Verfahren, siehe Abschnitt 6.4.2, überträgt sich die modifizierte Hessenberg-Gestalt der Matrix $\mathbf{M}_i$ auch auf $\widetilde{\mathbf{M}}_i$, und es gelten ebenfalls die Einbettungs-Eigenschaften

$$\mathbf{M}_i = \left(\begin{array}{ccc|c} & \mathbf{M}_{i-1} & & * \\ & & & \vdots \\ & & * & * \\ \hline * & \dots & \dots & * \end{array}\right), \qquad \widetilde{\mathbf{M}}_i = \left(\begin{array}{ccc|c} & \widetilde{\mathbf{M}}_{i-1} & & * \\ & & & \vdots \\ & & * & * \\ \hline * & \dots & \dots & * \end{array}\right),$$

$$\mathbf{z}_i^k = \begin{pmatrix} \mathbf{z}_{i-1}^k \\ * \end{pmatrix}, \qquad \mathbf{a}_i^k = \begin{pmatrix} \mathbf{a}_{i-1}^k \\ * \end{pmatrix}.$$

Daher müssen im i-ten Iterationsschritt nur die von 0 verschiedenen Elemente in der i-ten Zeile und Spalte von $\widetilde{\mathbf{M}}_i$ sowie die i-te Komponente von $\mathbf{a}_i^k$ neu berechnet werden. Das geschieht durch die Algorithmen 7.3 und 7.4. Algorithmus 7.3 dient zur Lösung der Systeme (7.3), (7.4) und (7.6), Algorithmus 7.4 zur Lösung des Systems (7.5).

Algorithmus 7.3 (Effiziente Lösung der Systeme (7.3), (7.4) und (7.6)).

gegeben: System $(\mathbf{I} - \delta\mathbf{H}_i)\mathbf{y}_i^k = \mathbf{z}_i^k$, $\varkappa_1$, $m_{11}, \dots, m_{i-1,i-1}$, $a_1, \dots, a_{i-1}$

Schritt 1: neue Elemente der Matrix $\mathbf{M}_i$ hinzufügen

for $j = 1, \dots, i-1$

$m_{ji} = -\delta h_{ji}$

end

if $i > 1$

$m_{i,i-1} = -\delta h_{i,i-1}$

end

$m_{ii} = 1 - \delta h_{ii}$

if $i > \varkappa_1 + 1$

$m_{i,i-2} = -\delta h_{i,i-2}$

end

Schritt 2: LU-Zerlegung

for $j = 2, \ldots, \min\{i-1, \varkappa_1 + 1\}$

$m_{ji} = m_{ji} - m_{j,j-1} m_{j-1,i}$

end

for $j = \varkappa_1 + 2, \ldots, i-1$

$m_{ji} = m_{ji} - m_{j,j-1} m_{j-1,i} - m_{j,j-2} m_{j-2,i}$

end

if $1 < i \leq \varkappa_1 + 1$

$m_{i,i-1} = m_{i,i-1} / m_{i-1,i-1}$

$m_{ii} = m_{ii} - m_{i,i-1} m_{i-1,i}$

end

Schritt 3: löse $\mathbf{L}_i \mathbf{a}_i^k = \mathbf{z}_i^k$

if $i = 1$

$a_1 = (\mathbf{z}_1^1)_1$

end

if $1 < i \leq \varkappa_1$

$a_i = (\mathbf{z}_i^1)_i - m_{i,i-1} a_{i-1}$

end

if $i = \varkappa_1 + 1$

$a_1 = (\mathbf{z}^2_{\varkappa_1+1})_1$

for $j = 2, \ldots, \varkappa_1 + 1$

$a_j = (\mathbf{z}^2_{\varkappa_1+1})_j - m_{j,j-1}a_{j-1}$

end

end

if $i > \varkappa_1 + 1$

$a_i = (\mathbf{z}^k_i)_i - m_{i,i-2}a_{i-2} - m_{i,i-1}a_{i-1}$

end

Schritt 4: löse $\mathbf{U}_i\mathbf{y}^k_i = \mathbf{a}^k_i$

for $j = i, i-1, \ldots, 1$

$(\mathbf{y}^k_i)_j = \left(a_j - \sum_{l=j+1}^{i} m_{jl}(\mathbf{y}^k_i)_l\right)/m_{jj}$

end

□

Algorithmus 7.4 (Effiziente Lösung des Systems (7.5)).

gegeben: System $(\mathbf{I} - \delta\mathbf{H}_{\varkappa_2})\mathbf{y}^3_{\varkappa_2} = \mathbf{z}^3_{\varkappa_2}$, $\varkappa_1$, $\varkappa_2$, $m_{11}, \ldots, m_{\varkappa_2-1,\varkappa_2-1}$

Schritt 1 und *Schritt 2* entfallen

Schritt 3: löse $\mathbf{L}_{\varkappa_2}\mathbf{a}^3_{\varkappa_2} = \mathbf{z}^3_{\varkappa_2}$

$a_1 = (\mathbf{z}^3_i)_1$

for $j = 2, \ldots, \varkappa_1 + 1$

$a_j = (\mathbf{z}^3_i)_j - m_{j,j-1}a_{j-1}$

end

for $j = \varkappa_1 + 2, \ldots, \varkappa_2$

$a_j = (\mathbf{z}^3_i)_j - m_{j,j-1}a_{j-1} - m_{j,j-2}a_{j-2}$

end

Schritt 4: löse $\mathbf{U}_{\varkappa_2}\mathbf{y}^3_{\varkappa_2} = \mathbf{a}^3_{\varkappa_2}$

for $j = \varkappa_2, \varkappa_2 - 1, \ldots, 1$

$(\mathbf{y}_i^3)_j = \left(a_j - \sum_{l=j+1}^{\varkappa_2} m_{jl}(\mathbf{y}_i^3)_l\right)/m_{jj}$

end

□

Bemerkung 7.5. Wie in Algorithmus 6.21 entsprechen die Zahlen m_{jk} in Schritt 1 den Elementen der Matrix $\mathbf{M}_i$ und werden in Schritt 2 zu Elementen der LU-Matrix $\widetilde{\mathbf{M}}_i = \mathbf{L}_i + \mathbf{U}_i - \mathbf{I}$ umgewandelt. □

7.4 Zur Stabilität von Krylov-W-Verfahren

Die in Abschnitt 6.4.3 angesprochene Eigenschaft des Arnoldi-Verfahrens, die *äußeren* Eigenvektoren besonders schnell zu approximieren, wirkt sich günstig auf die Stabilität eines Krylov-W-Verfahrens aus. Wir betrachten das System $\mathbf{u}_t = \mathbf{f}(t, \mathbf{u})$ und die W-Methode (7.1) mit $\mathbf{T} = \partial\mathbf{f}/\partial\mathbf{u} = \mathbf{J}$, die dann die Form

$$\begin{aligned}(\mathbf{I} - \tau_i\gamma\mathbf{J})\widetilde{\mathbf{k}}^j &= \mathbf{f}\left(t_i + \tau_i c_j, \mathbf{u}_i + \tau_i\gamma\sum_{l=1}^{j-1}\varphi_{lj}\widetilde{\mathbf{k}}^l\right) + \sum_{l=1}^{j-1}\vartheta_{lj}\widetilde{\mathbf{k}}^l, \qquad j = 1, \ldots, s \qquad (7.9)\\ \mathbf{u}_{i+1} &= \mathbf{u}_i + \tau_i\sum_{l=1}^{s} g_l\widetilde{\mathbf{k}}^l\\ \widehat{\mathbf{u}}_{i+1} &= \mathbf{u}_i + \tau_i\sum_{l=1}^{s}\widehat{g}_l\widetilde{\mathbf{k}}^l\end{aligned}$$

hat. Die Gleichungssysteme in dieser Methode werden durch den multiplen Arnoldi-Prozeß näherungsweise gelöst. Wir bezeichnen die Näherungslösungen mit $\overline{\mathbf{k}}^j$. Setzen wir

$$\overline{\mathbf{T}}^j := \mathbf{Q}_{\varkappa_j}\mathbf{Q}_{\varkappa_j}^T\mathbf{J} \qquad \text{und} \qquad \overline{\mathbf{b}}^j := \mathbf{f}\left(t_i + \tau_i c_j, \mathbf{u}_i + \tau_i\gamma\sum_{l=1}^{j-1}\varphi_{lj}\overline{\mathbf{k}}^l\right) + \sum_{l=1}^{j-1}\vartheta_{lj}\overline{\mathbf{k}}^l,$$

dann sind die Näherungslösungen $\overline{\mathbf{k}}^j$ von (7.9) exakte Lösungen der modifizierten W-Methode

$$\begin{aligned}(\mathbf{I} - \tau_i\gamma\overline{\mathbf{T}}^j)\overline{\mathbf{k}}^j &= \overline{\mathbf{b}}^j, \qquad j = 1, \ldots, s \qquad (7.10)\\ \mathbf{u}_{i+1} &= \mathbf{u}_i + \tau_i\sum_{l=1}^{s} g_l\overline{\mathbf{k}}^l\\ \widehat{\mathbf{u}}_{i+1} &= \mathbf{u}_i + \tau_i\sum_{l=1}^{s}\widehat{g}_l\overline{\mathbf{k}}^l,\end{aligned}$$

siehe (6.20). Die Stabilität des Krylov-W-Verfahrens hängt von den Matrizen $\overline{\mathbf{T}}^j$ in der exakt gelösten modifizierten W-Methode ab. Wäre $\overline{\mathbf{T}}^j = \mathbf{0}$ für $j = 1, \ldots, s$, so würde das explizite

Runge-Kutta Verfahren

$$\begin{aligned}
\overline{\mathbf{k}}^j &= \mathbf{f}\left(t_i + \tau_i c_j, \mathbf{u}_i + \tau_i \gamma \sum_{l=1}^{j-1} \varphi_{lj} \overline{\mathbf{k}}^l\right) + \sum_{l=1}^{j-1} \vartheta_{lj} \overline{\mathbf{k}}^l, \qquad j = 1, \ldots, s \qquad (7.11)\\
\mathbf{u}_{i+1} &= \mathbf{u}_i + \tau_i \sum_{l=1}^{s} g_l \overline{\mathbf{k}}^l\\
\widehat{\mathbf{u}}_{i+1} &= \mathbf{u}_i + \tau_i \sum_{l=1}^{s} \widehat{g}_l \overline{\mathbf{k}}^l
\end{aligned}$$

gelöst, welches ein beschränktes Stabilitätsgebiet hat und damit eine Einschränkung des Zeitschritts erfordert. Wäre hingegen $\overline{\mathbf{T}}^j = \mathbf{J},\ j = 1, \ldots, s$, dann wären die Stabilitätseigenschaften des Krylov-W-Verfahrens optimal.

Aus der Stabilitätsbedingung für das explizite Verfahren (7.11) geht hervor, daß betragsgroße Eigenwerte mit negativem Realteil für die Steifheit eines Problems verantwortlich sind, da diese den maximal erlaubten Zeitschritt einschränken. Wir bezeichnen diese Eigenwerte und den zugehörigen Eigenraum als **dominant**. Aus Stabilitätsgründen wäre es daher günstig, wenn die Eigenräume der Matrizen $\overline{\mathbf{T}}^j$ den dominanten Eigenraum der Jacobi-Matrix $\mathbf{J}$ besonders gut approximieren würden.

Wie aus Abschnitt 6.4.3 hervorgeht, approximiert der durch das Arnoldi-Verfahren aufgebaute Krylov-Raum $\mathcal{K}_i$ gerade die äußeren Eigenvektoren von $\mathbf{J}$, zu denen auch die dominanten Eigenvektoren gehören, besonders schnell. Der folgende Satz zeigt, daß dann auch die Eigenräume der Matrizen $\overline{\mathbf{T}}^j$ den dominanten Eigenraum von $\mathbf{J}$ schnell approximieren.

Satz 7.6. *Es sei* $\mathbf{v}$ *ein Eigenvektor der Matrix* $\mathbf{J}$ *zum Eigenwert* λ*. Das System* $(\mathbf{I} - \delta\mathbf{J})\mathbf{k} = \mathbf{b}$ *werde mit dem in Abschnitt 6.4.1 beschriebene Arnoldi-Verfahren gelöst. Es sei* $\mathbf{Q}$ *die durch dieses Verfahren erzeugte Matrix, die die Orthonormalbasis des Krylov-Raumes* $\mathcal{K}$ *enthält. Dann gilt die Ungleichung*

$$|\overline{\mathbf{T}}\mathbf{v} - \lambda\mathbf{v}| \leq |\lambda| \ \mathrm{dist}(\mathcal{K}, \mathbf{v}),$$

wobei $\overline{\mathbf{T}} = \mathbf{Q}\mathbf{Q}^T\mathbf{J}$ *und* $\mathrm{dist}(\mathcal{K}, \mathbf{v})$ *der Euklidische Abstand zwischen* $\mathcal{K}$ *und* $\mathbf{v}$ *ist.*

Beweis. Das Minimum

$$\mathrm{dist}(\mathcal{K}, \mathbf{v}) = \min\{|\mathbf{w} - \mathbf{v}| : \mathbf{w} \in \mathcal{K}\}$$

wird für einen Vektor $\mathbf{w}^* \in \mathcal{K}$ angenommen. Es gilt

$$\begin{aligned}
\overline{\mathbf{T}}\mathbf{v} &= \mathbf{Q}\mathbf{Q}^T\mathbf{J}\mathbf{v} = \lambda\mathbf{Q}\mathbf{Q}^T\mathbf{v} = \lambda\mathbf{Q}\mathbf{Q}^T\mathbf{w}^* - \lambda\mathbf{Q}\mathbf{Q}^T(\mathbf{w}^* - \mathbf{v}) = \lambda\mathbf{w}^* - \lambda\mathbf{Q}\mathbf{Q}^T(\mathbf{w}^* - \mathbf{v})\\
&= \lambda\mathbf{v} + \lambda(\mathbf{w}^* - \mathbf{v}) - \lambda\mathbf{Q}\mathbf{Q}^T(\mathbf{w}^* - \mathbf{v}) = \lambda\mathbf{v} + \lambda(\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)(\mathbf{w}^* - \mathbf{v}).
\end{aligned}$$

Die Multiplikation mit der Matrix $\mathbf{I} - \mathbf{Q}\mathbf{Q}^T$ beschreibt einen Schritt der Gram-Schmidt-Orthogonalisierung bezüglich der in $\mathbf{Q}$ enthaltenen Orthonormalbasis. Insbesondere folgt daraus, daß für einen beliebigen Vektor $\mathbf{a}$ die Vektoren $\mathbf{a}$, $\mathbf{Q}\mathbf{Q}^T\mathbf{a}$ und $(\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)\mathbf{a}$ die Seitenvektoren eines rechtwinkligen Dreiecks sind, wobei $\mathbf{a}$ die Hypothenuse bildet, ein Resultat, von dem man sich auch durch direkte Rechnung leicht überzeugt. Demnach ist für einen beliebigen Vektor $\mathbf{a}$ stets

$$|(\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)\mathbf{a}| \leq |\mathbf{a}|.$$

Damit folgt

$$|\overline{\mathbf{T}}\mathbf{v} - \lambda\mathbf{v}| = |\lambda||(\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)(\mathbf{w}^* - \mathbf{v})| \leq |\lambda||(\mathbf{w}^* - \mathbf{v})| = |\lambda|\text{dist}(\mathcal{K}, \mathbf{v}).$$

□

Ist $\mathbf{v}$ ein dominanter Eigenvektor von $\mathbf{J}$, so geht $\text{dist}(\mathcal{K}, \mathbf{v})$ für wachsende Krylov-Dimension schnell gegen 0. Satz 7.6 zeigt, daß dann ein Eigenvektor von $\mathbf{T}$ schnell gegen $\mathbf{v}$ konvergiert. Daraus folgt, daß insbesondere der dominante Eigenraum von $\mathbf{J}$ schnell durch Eigenvektoren der Matrix $\mathbf{T}$ approximiert wird. Damit ist der Einsatz des Arnoldi-Verfahrens in einer W-Methode besonders stabilitätserhaltend.

7.5 Abbruch der Iteration

Die Abbruchbedingung für die Iteration entspricht der, die wir für das BiCGstab-Verfahren ohne Vorkonditionierung verwenden. Diese Abbruchbedingung wird auch von Schmitt und Weiner [143] vorgeschlagen. Es gelten die gleichen Überlegungen, wie in Abschnitt 6.3.2 dargestellt, nur daß jetzt $\mathbf{P} = \mathbf{I}$ zu setzen ist. Wir erhalten also wieder mit

$$TOL_{\text{LSS}} = \frac{\alpha_{\text{LSS}} TOL_t}{\tau_i} \tag{7.12}$$

eine an die Toleranz für den lokalen Fehler TOL_t gekoppelte Toleranz des linearen Lösers. Wie beim BiCGstab-Verfahren ist der optimale Wert für α_{LSS} problemabhängig. Man ist hier auf numerische Testrechnungen angewiesen, vgl. Bemerkung 6.7. Wir werden uns in Abschnitt 9.6 unter anderem mit der Ermittlung eines optimalen α_{LSS} befassen.

Ein Verfahren, das den Abbruch seiner Iteration allein durch die Toleranz TOL_{LSS} steuert, ist jedoch in vielen Fällen nicht effizient. Es müssen zusätzlich die durch $\varkappa_1$, $\varkappa_{12}$ und $\varkappa_{23}$ vorgegebenen maximalen Krylov-Dimensionen geeignet und bei einigen Problemen recht niedrig gewählt werden. Das Problem der Ineffizienz bei zu vielen Iterationen ist beim Krylov-W-Verfahren, wo der Aufwand mit der Iteration quadratisch ansteigt, gravierender als bei BiCGstab. Auch wenn man durch Beschränkung der Krylov-Dimension das Gleichungssystem selbst mit nur geringer Genauigkeit löst, so sollte der lokale Fehler der W-Methode dadurch nicht wesentlich zunehmen, da dieser über die Zeitschrittsteuerung reguliert wird. Möglicherweise verliert die Methode aber an Stabilität, so daß der Zeitschritt reduziert wird.

Geeignete Werte für die maximalen Krylov-Dimensionen sind ebenfalls problemabhängig und müssen durch numerische Testrechnungen gefunden werden. Für einige Beispielprobleme wird die Effizienz in Abhängigkeit der maximalen Krylov-Dimension in Abschnitt 9.5 untersucht.

In einigen Implementationen von Krylov-W-Verfahren wird der residuumsabhängige Abbruch nur in Stufe 1 vorgenommen. In den höheren Stufen wird dann immer bis zur vorgegebenen maximalen Krylov-Dimension iteriert. Eine solche Vorgehensweise wird beispielsweise in dem Programm ROWMAP von Weiner, Schmitt und Podhaisky [169] verfolgt. Sie hat den Vorteil, daß die Bestimmung der Residuen und einiger zu deren Berechnung benötigter Größen in den höheren Stufen entfällt. Insbesondere müssen in den höheren Stufen die Gleichungssysteme $(\mathbf{I} - \delta\mathbf{H}_i)\mathbf{y}_i^k = \mathbf{z}_i^k$ *nur noch einmal am Ende* gelöst werden. Ein Nachteil dieser Vorgehensweise ist natürlich die fehlende Fehlerkontrolle in den Stufen 2 bis s, jedoch kann

auf diese eventuell verzichtet werden, wenn $\varkappa_{12}$ und $\varkappa_{23}$ hinreichend klein sind. In unseren numerischen Rechnungen schlagen wir diesen Weg nicht ein, sondern stellen die Abbruchbedingung in jeder Iteration.

Erfolgt der Abbruch in einer Stufe j aufgrund des Erreichens der maximalen Krylov-Dimension $\varkappa_j$, so ist das Residuum $\|\mathbf{r}^j_{\varkappa_j}\|_*$ größer als die Toleranz TOL_{LSS}. In diesem Falle ist es möglicherweise sinnvoll, TOL_{LSS} für die nachfolgenden Stufen *auf den Wert* $\|\mathbf{r}^j_{\varkappa_j}\|_*$ *zu erhöhen*, da die gewünschte Genauigkeit bereits in Stufe j verletzt wurde. Diese Strategie wenden wir auch in dem in Kapitel 9 zu numerischen Untersuchungen verwendeten Krylov-W-Verfahren an.

7.6 Numerische Untersuchungen zur Konvergenz

Untersuchung 7.7. Wir wollen die Konvergenz von Fehler und Residuum des betrachteten Krylov-W-Verfahrens anhand des folgenden numerischen Beispiels untersuchen und mit der Konvergenz des in Abschnitt 6.3 beschriebenen BiCGstab-Verfahrens vergleichen. Wir betrachten die Wärmeleitungsgleichung mit Quellterm

$$u_t = \frac{1}{4\pi^2}\Delta u + (1+t)\sin(2\pi x) - t\sin(2\pi y), \qquad (x,y) \in \Omega =]0,1[^2$$

mit der Anfangsbedingung $u(x,y,0) = \sin(2\pi y)$ und der Dirichlet-Randbedingung

$$u(x,y,t) = \begin{cases} t\sin(2\pi x), & y \in \{0,1\}, \\ (1-t)\sin(2\pi y), & x \in \{0,1\}. \end{cases}$$

Die exakte Lösung dieser Differentialgleichung ist mit

$$u(x,y,t) = t\sin(2\pi x) + (1-t)\sin(2\pi y)$$

gegeben. Die Ortsdiskretisierung erfolgt mit linearen finiten Elementen auf einem uniformen Dreiecksgitter. Die mittlere Seitenlänge der Dreiecke ist $h = 0{,}0125$. Im Ergebnis erhalten wir ein System gewöhnlicher Differentialgleichungen der Form $\mathbf{u}_t = \mathbf{f}(t,\mathbf{u})$. Die Zeitdiskretisierung wird mit der W-Methode zweiter Ordnung (7.2) durchgeführt, wobei $\mathbf{T} = \mathbf{J} = \partial\mathbf{f}/\partial\mathbf{u}$ ist.

Wir betrachten die Lösung der W-Methode im ersten Zeitschritt, $\tau_0 = 0{,}01$. Der Vektor $\widetilde{\mathbf{k}}^j$ sei die exakte Lösung des in (7.2) auftretenden Gleichungssystems

$$\left(\mathbf{I} - \tau_m\left(1 - \sqrt{2}/2\right)\mathbf{J}\right)\widetilde{\mathbf{k}}^j = \mathbf{b}^j$$

und $\overline{\mathbf{k}}^j$ die Näherungslösung, die mit einem iterativen Löser gewonnen wurde. Als Fehler-Toleranz benutzen wir zunächst $TOL_{\mathrm{LSS}} = 10^{-6}$. In Abbildung 7.1 (links) ist der Fehler des Gleichungslösers $\|\widetilde{\mathbf{k}}^j - \overline{\mathbf{k}}^j\|_*$ über der Anzahl der Iterationen grün dargestellt. Die Norm $\|\cdot\|_*$ ist dabei die skalierte Euklidische Vektornorm, siehe Bermerkung 7.2. Die blauen Kurven zeigen das Residuum $\|\mathbf{r}^j_i\|_*$ und die roten Kurven das in (7.7) definierte geschätzte Residuum $r^j_{i,\mathrm{est}}$. In diesem Falle ist das Residuum etwas größer als der Fehler, die Schätzung $r^j_{i,\mathrm{est}}$ des Residuums ist sehr gut. Wir bezeichnen mit $\widetilde{\mathbf{k}}^j_i$ und $\overline{\mathbf{k}}^j_i$ den Wert von $\widetilde{\mathbf{k}}^j$ bzw. $\overline{\mathbf{k}}^j$ im i-ten Iterationsschritt der Stufe j. Die Konvergenzgeschwindigkeit beträgt in der ersten Stufe etwa

$$\|\widetilde{\mathbf{k}}^1_i - \overline{\mathbf{k}}^1_i\|_* \approx C \cdot 0{,}5^i.$$

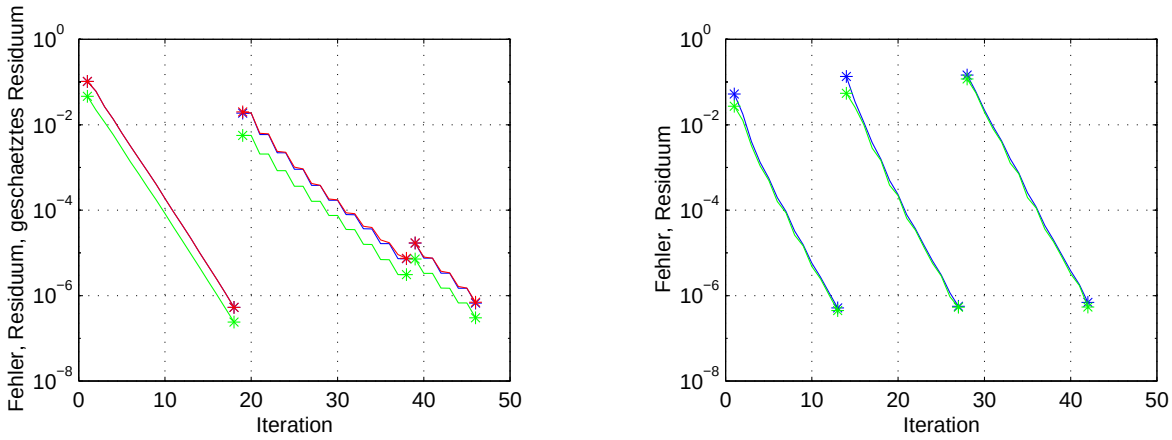


Abbildung 7.1: Konvergenz bei $TOL_{\mathrm{LSS}} = 10^{-6}$: Fehler (grün), Residuum (blau) und geschätztes Residuum (rot) in Abhängigkeit von der Anzahl der Iterationen; **links:** Krylov-W-Verfahren, **rechts:** BiCGstab-Verfahren ohne Vorkonditionierung

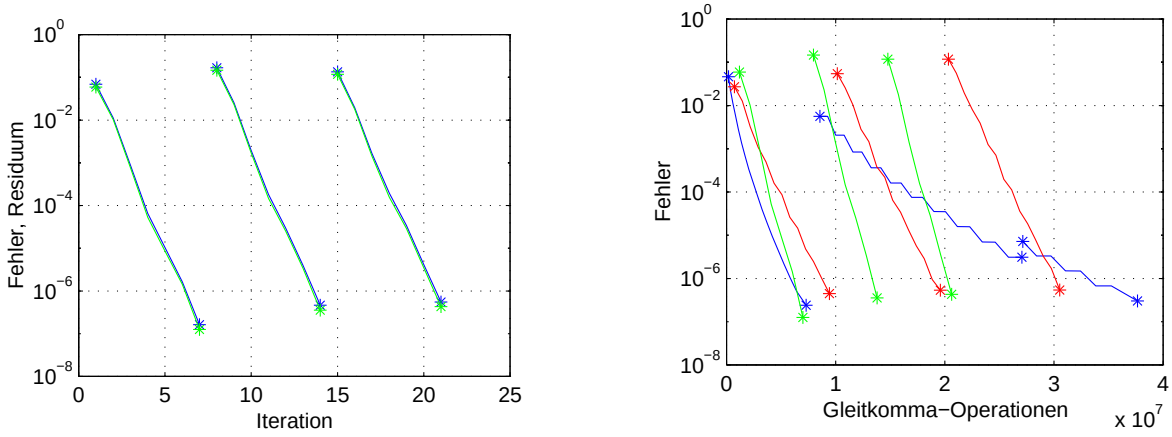


Abbildung 7.2: Konvergenz bei $TOL_{\mathrm{LSS}} = 10^{-6}$: **links:** BiCGstab-Verfahren mit SSOR-Vorkonditionierung, Fehler (grün) und Residuum (blau) in Abhängigkeit von der Anzahl der Iterationen, **rechts:** Fehler in Abhängigkeit der Gleitkomma-Operationen: Krylov-W-Verfahren (blau), BiCGstab-Verfahren ohne (rot) und mit SSOR-Vorkonditionierung (grün)

In den folgenden Stufen ist die Konvergenz deutlich langsamer; sie beträgt etwa

$$\|\widetilde{\mathbf{k}}_i^j - \overline{\mathbf{k}}_i^j\|_* \approx C \cdot 0{,}65^i.$$

für $i = 2, 3$. Der Fehler zu Beginn der dritten Stufe ist nur geringfügig höher als der zum Ende der zweiten Stufe. Das ist darauf zurückzuführen, daß die rechte Seite $\mathbf{z}^3$ bereits in dem in der zweiten Stufe aufgebauten Krylov-Raum $\mathcal{K}_{\varkappa_2}$ enthalten ist.

Zum Vergleich betrachten wir die Lösung der W-Methode mit dem BiCGstab-Verfahren. Abbildung 7.1 (rechts) zeigt Fehler und Residuum, wenn keine Vorkonditionierung eingesetzt wird. In Abbildung 7.2 (links) sind Fehler und Residuum für das BiCGstab-Verfahren mit SSOR-Vorkonditionierung dargestellt. In beiden Fällen sind Residuum und Fehler beinahe

identisch. Die Konvergenzraten der einzelnen Stufen unterscheiden sich nicht so stark wie beim multiplen Arnoldi-Prozeß. Die Konvergenzrate beträgt im Falle ohne Vorkonditionierung näherungsweise

$$\|\widetilde{\mathbf{k}}_i^j - \overline{\mathbf{k}}_i^1\|_* \approx C \cdot 0{,}35^i, \qquad j = 1, 2, 3.$$

mit der Vorkonditionierung kann eine Steigerung auf etwa

$$\|\widetilde{\mathbf{k}}_i^j - \overline{\mathbf{k}}_i^1\|_* \approx C \cdot 0{,}1^i, \qquad j = 1, 2, 3.$$

erreicht werden.[3]

Um den Rechenaufwand der genannten Verfahren zu vergleichen, stellen wir in Abbildung 7.2 (rechts) jeweils den Fehler über der Anzahl der Gleitkomma-Operationen dar. Das vorkonditionierte BiCGstab-Verfahren schneidet am besten ab. Der Aufwand des multiplen Arnoldi-Prozesses ist wegen seines quadratisch steigenden Aufwandes und der hier benötigten relativ hohen Anzahl von Iterationen den anderen beiden Verfahren unterlegen.

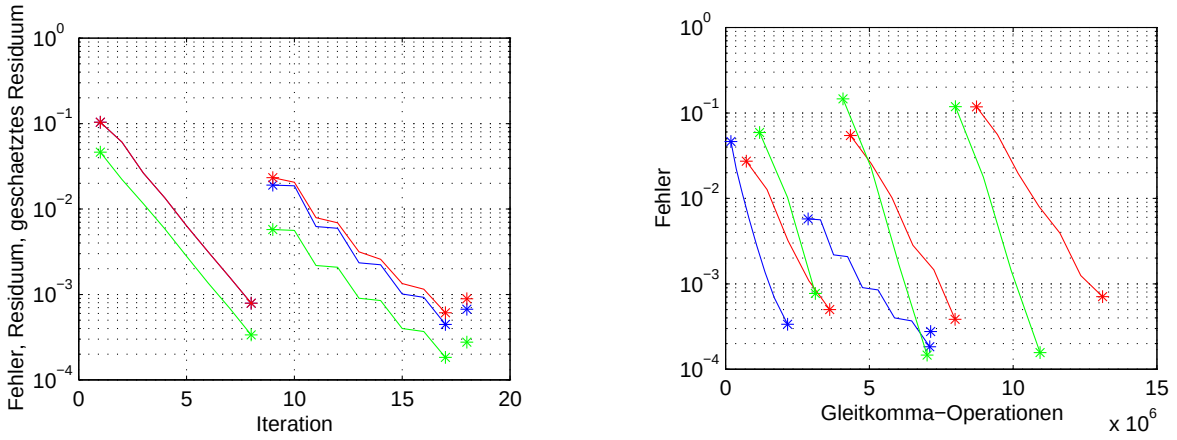


Abbildung 7.3: Konvergenz bei $TOL_{\text{LSS}} = 10^{-3}$: **links:** Krylov-W-Verfahren: Fehler (grün), Residuum (blau) und geschätztes Residuum (rot) in Abhängigkeit von der Anzahl der Iterationen, **rechts:** Fehler in Abhängigkeit der Gleitkomma-Operationen: Krylov-W-Verfahren (blau), BiCGstab-Verfahren ohne (rot) und mit SSOR-Vorkonditionierung (grün)

Ein anderes Bild ergibt sich bei einer geringeren Genauigkeitsforderung. Wir wiederholen die Untersuchung für die Toleranz $TOL_{\text{LSS}} = 10^{-3}$ und stellen in Abbildung 7.3 die Konvergenz des Krylov-W-Verfahrens und den Vergleich der drei Verfahren bezüglich der Gleitkomma-Operationen dar. Das Krylov-W-Verfahren ist in diesem Falle das effizienteste Verfahren.

Ergebnisse der Untersuchung

- Beim Krylov-W-Verfahren ist die Schätzung des Residuums sehr gut.
- Die Übereinstimmung zwischen Residuum und Fehler ist bei dem BiCGstab-Verfahren deutlich besser als bei dem Krylov-W-Verfahren.

[3] Die Konstanten C sind selbstverständlich in allen Fällen unterschiedlich. Der Wert von C hat jedoch keinen Einfluß auf die Konvergenz*geschwindigkeit*, deshalb wird er hier nicht angegeben.

- Die Konvergenzgeschwindigkeit des Krylov-W-Verfahrens ist in der ersten Stufe höher als in den darauf folgenden Stufen 2 und 3.
- Beim BiCGstab-Verfahren beschleunigt die SSOR-Vorkonditionierung die Konvergenz.
- Für kleine Toleranzen TOL_{LSS} erweist sich das vorkonditionierte BiCGstab-Verfahren als effizient, für große Toleranzen hingegen das Krylov-W-Verfahren.

□

Wie das Beispiel illustriert, ist der multiple Arnoldi-Prozeß oft im Vorteil, wenn keine sehr hohe Genauigkeit der Lösung erforderlich ist. Diese Eigenschaft macht das Verfahren gerade als Löser in einer W-Methode interessant, da hier oft moderate Genauigkeiten zur Sicherung der Stabilität ausreichen und die Ordnung der W-Methode auch bei ungenauer Lösung der Gleichungssysteme erhalten bleibt.

Kapitel 8

Partitionierung

Das in Abschnitt 5.6 beschriebene Phänomen der Steifheit tritt bei der Diskretisierung von Reaktions-Diffusions-Systemen mitunter nur *lokal* auf, d.h. das Problem ist *nur in einem Teilgebiet* $\Omega_{\text{steif}}(t)$ von Ω steif, welches oftmals auch zeitabhängig ist. Wie wir in 5.6.1 gesehen haben, sind bei Reaktions-Diffusions-Gleichungen insbesondere die Feinheit des Gitters im Zusammenhang mit Diffusion sowie steile Gradienten der Reaktionsfunktion Ursachen für Steifheit. Treten diese Ursachen nur lokal auf, so ist auch die Steifheit nur lokaler Natur. In diesem Falle ist es oft nicht effizient, eine Zeitdiskretisierung für steife Probleme – etwa ein implizites Runge-Kutta-Verfahren – auf dem gesamten Gebiet Ω einzusetzen, da derartige Verfahren i.a. mehr Rechenzeit pro Zeitschritt benötigen als die für nichtsteife Probleme geeigneten expliziten Verfahren. Wir werden uns daher in diesem Kapitel mit Verfahren befassen, die auf dem steifen Teilgebiet Ω_{steif} eine implizite und auf dem nichtsteifen Teilgebiet $\Omega \backslash \Omega_{\text{steif}}$ eine explizite Zeitdiskretisierung verwenden. Zusätzlich soll die Möglichkeit bestehen, nur den Diffusionsteil oder nur den Reaktionsteil implizit zu lösen. Verfahren dieser Art wollen wir als **lokale Partitionierungs-Verfahren** bezeichnen.

Die in Abschnitt 5.7 beschriebenen W-Methoden sind zur lokalen Partitionierung besonders gut geeignet, da man allein durch Änderung der Matrix $\mathbf{T}$ in (5.13) zwischen einem für steife Probleme geeigneten linear impliziten Verfahren und einem expliziten Verfahren „schalten“ kann. Bei lokal steifen Problemen kann hier auch eine lokale, d.h. zeilen- und spaltenweise Anpassung der Matrix $\mathbf{T}$ vorgenommen werden.

Bei einigen Problemen kann man bereits a-priori eine Einteilung in steife und nichtsteife Komponenten vornehmen. In diesem Falle erübrigt sich die im nächsten Abschnitt beschriebene Steifigkeitserkennung. Derartige Verfahren werden als **feste Partitionierungs-Verfahren** bezeichnet. Demgegenüber wird bei einem **automatischen Partitionierungs-Verfahren** in regelmäßigen Abständen die lokale Steifigkeit untersucht und die Matrix $\mathbf{T}$ entsprechend angepaßt. Ein automatisches Partitionierungs-Verfahren wurde erstmals 1979 von Enright und Kamel [56] vorgestellt. Weiner, Arnold, Rentrop und Strehmel [167] entwickelten 1993 ein automatisches Partitionierungs-Verfahren auf Grundlage einer W-Methode. Hundsdorfer [85] wandte automatische Partitionierung in einem BDF-Verfahren an.

In dem von uns verwendeten Verfahren benutzen wir eine etwas andere Methode zur Steifigkeitserkennung als in [167]. Außerdem führen wir die Möglichkeit ein, Diffusions- und Reak-

tionsteil getrennt auf Steifheit zu untersuchen und zu partitionieren.

8.1 Vollständig automatische Partitionierung

Im Falle der automatischen Partitionierung müssen zunächst die steifen Komponenten ermittelt werden. Wir betrachten das in (2.2) angegebene Reaktions-Diffusions-System

$$\begin{aligned}\frac{\partial u_k}{\partial t}(\mathbf{x},t) &= \text{div }(d_k(\mathbf{x})\nabla u_k(\mathbf{x},t)) + r_k(\mathbf{x})u_k(\mathbf{x},t) + p_k(u_1(\mathbf{x},t),\ldots,u_m(\mathbf{x},t)) + q_k(\mathbf{x},t),\\ &\qquad k=1,\ldots,m, \quad \mathbf{x}\in\Omega\subset\mathbb{R}^n, \quad t\in[t_0,t_e],\\ u(\mathbf{x},t_0) &= u_0(\mathbf{x}), \qquad \mathbf{x}\in\Omega\subset\mathbb{R}^n.\end{aligned}$$

Im Falle $k=1$ liegt nur eine einzelne Reaktions-Diffusions-Gleichung vor. Die Ortsdiskretisierung erfolge mit linearen finiten Elementen, siehe Abschnitt 3.1. Wir verwenden die in Abschnitt 3.1.5 beschriebene Reduktion der Massen-Matrx. Im Ergebnis erhält man je nach Randbedingung eines der in (3.15), (3.16) und (3.33) angegebenen Systeme gewöhnlicher Differentialgleichungen, welches wir in der Form

$$\mathbf{u}_t = \mathbf{f}(t,\mathbf{u}) \tag{8.1}$$

schreiben. Die Zeitdiskretisierung werde mit der in (5.13) dargestellten W-Methode

$$\begin{aligned}\mathbf{k}_j &= \mathbf{f}\left(t_i + c_j\tau_i, \mathbf{u}_i + \tau_i\sum_{l=1}^{j-1}\alpha_{jl}\mathbf{k}_l\right) + \tau_i\mathbf{T}\sum_{l=1}^{j}\gamma_{jl}\mathbf{k}_l, \qquad j=1,\ldots,s \qquad (8.2)\\ \gamma_{jj} &= \gamma, \qquad j=1,\ldots,s\\ \mathbf{u}_{i+1} &= \mathbf{u}_i + \tau_i\sum_{l=1}^{s} b_l\mathbf{k}_l\\ \widehat{\mathbf{u}}_{i+1} &= \mathbf{u}_i + \tau_i\sum_{l=1}^{s}\widehat{b}_l\mathbf{k}_l\end{aligned}$$

durchgeführt. Diese Methode kann mit der in Abschnitt 5.7 dargestellten Transformation auf die Form (5.17)

$$\begin{aligned}(\mathbf{I}-\tau_i\gamma\mathbf{T})\widetilde{\mathbf{k}}_j &= \mathbf{f}\left(t_i+\tau_i c_j, \mathbf{u}_i + \tau_i\gamma\sum_{l=1}^{j-1}\varphi_{lj}\widetilde{\mathbf{k}}_l\right) + \sum_{l=1}^{j-1}\vartheta_{lj}\widetilde{\mathbf{k}}_l, \qquad j=1,\ldots,s \qquad (8.3)\\ \mathbf{u}_{i+1} &= \mathbf{u}_i + \tau_i\sum_{l=1}^{s} g_l\widetilde{\mathbf{k}}_l\\ \widehat{\mathbf{u}}_{i+1} &= \mathbf{u}_i + \tau_i\sum_{l=1}^{s}\widehat{g}_l\widetilde{\mathbf{k}}_l.\end{aligned}$$

gebracht werden. Zum Zeitpunkt t_i liegt eine Triangulierung $\mathcal{T}_h$ und ein Vektor $\mathbf{u}_i$ vor, der die Knotenwerte der Näherungslösung enthält. Mit Hilfe der oben angegebenen W-Methode wird der zur Zeit t_{i+1} gehörige Lösungsvektor $\mathbf{u}_{i+1}$ berechnet. Die Größe der Zeitschritte wird durch eine Genauigkeitsforderung an den lokalen Fehler eingeschränkt.

Es sei $\mathbf{J} := \partial\mathbf{f}/\partial\mathbf{u}$ die Jacobi-Matrix der rechten Seite der Differentialgleichung. Die W-Methode erreicht ihre höchste Stabilität, wenn man

$$\mathbf{T} = \mathbf{J} \tag{8.4}$$

wählt. Ziel der Partitionierung ist es nun, in der Matrix $\mathbf{T}$ möglichst viele Zeilen und Spalten durch Nullzeilen/-spalten zu ersetzen, dabei aber die Stabilität des Verfahrens zu erhalten.

Wird in (8.2) $\mathbf{T} = \mathbf{0}$ gesetzt, so erhält man das explizite Runge-Kutta-Verfahren

$$\begin{aligned}
\mathbf{k}_j &= \mathbf{f}\left(t_i + c_j\tau_i, \mathbf{u}_i + \tau_i \sum_{l=1}^{j-1} \alpha_{jl}\mathbf{k}_l\right), && j = 1, \ldots, s, \\
\gamma_{jj} &= \gamma, && j = 1, \ldots, s, \\
\mathbf{u}_{i+1} &= \mathbf{u}_i + \tau_i \sum_{l=1}^{s} b_l\mathbf{k}_l, \\
\widehat{\mathbf{u}}_{i+1} &= \mathbf{u}_i + \tau_i \sum_{l=1}^{s} \widehat{b}_l\mathbf{k}_l
\end{aligned} \tag{8.5}$$

mit dem in Definition 5.10 definierten Stabilitätsgebiet S. Für dieses Verfahren betrachten wir die Zeitschrittbeschränkung

$$\tau_i\lambda_j(\mathbf{J}) \in S \tag{8.6}$$

für alle Eigenwerte $\lambda_j(\mathbf{J})$ mit $\mathrm{Re}\ \lambda_j(\mathbf{J}) \leq 0$. Für lineare autonome Probleme bedeutet diese Bedingung, daß das Verfahren in den Eigenrichtungen stabil sein soll, in denen auch das Differentialgleichungs-System stabil ist, vgl (5.11). Für nichtlineare oder nichtautonome Probleme ist (8.6) *keine* hinreichende Bedingung für stabiles Verhalten. Trotzdem gibt es eine Reihe nichtlinearer Probleme, die ein ähnliches Stabilitätsverhalten wie ihre Linearisierung besitzen. Wir legen daher die Bedingung (8.6) – mangels eines besseren Kriteriums – auch bei nichtlinearen Problemen der im folgenden erläuterten Steifigkeitserkennung zugrunde. Falls die Bedingung (8.6) für den durch die Genauigkeitsforderung bestimmten Zeitschritt τ_i *des impliziten Verfahrens* (8.3), (8.4) nicht erfüllt ist, so nennen wir das Problem **S-steif**.

Der Sachverhalt vereinfacht sich, wenn wir anstelle der Eigenwerte den Spektralradius $\varrho(\mathbf{J})$ heranziehen. Wir approximieren das Stabilitätsgebiet S durch einen Halbkreis $H := \{z \in \mathbb{C} : |z| \leq r_{\mathrm{stab}},\ \mathrm{Re}\ z \leq 0\}$ mit dem Radius r_{stab}, den wir als **Stabilitäts-Radius** bezeichnen. Die Stabilitätsbedingung (8.6) ersetzen wir durch die Bedingung

$$\tau_i\varrho(\mathbf{J}) \leq r_{\mathrm{stab}}. \tag{8.7}$$

Wird diese Bedingung durch den Zeitschritt τ_i aus (8.3), (8.4) verletzt, so bezeichnen wir das Problem als **H-steif**.

Bei einem Partitionierungs-Verfahren versucht man, die Komponenten, d.h. die Zeilen und Spalten von $\mathbf{J}$ zu finden, die für S-Steifheit verantwortlich sind, denn diese sollen durch ein implizites Verfahren abgedeckt werden. In der praktischen Realisierung sucht man, wie wir im folgenden zeigen werden, jedoch nur nach Komponenten, die H-Steifheit hervorrufen. Die beiden Steifheitsbegriffe stimmen aber nicht ganz überein. Wir wollen das am Beispiel eines

beliebigen dreistufigen expliziten Runge-Kutta-Verfahrens dritter Ordnung erläutern. Alle derartigen Verfahren besitzen die Stabilitätsfunktion

$$R(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6}$$

und das Stabilitätsgebiet $S = \{z \in \mathbb{C} : |R(z)| \leq 1\}$.

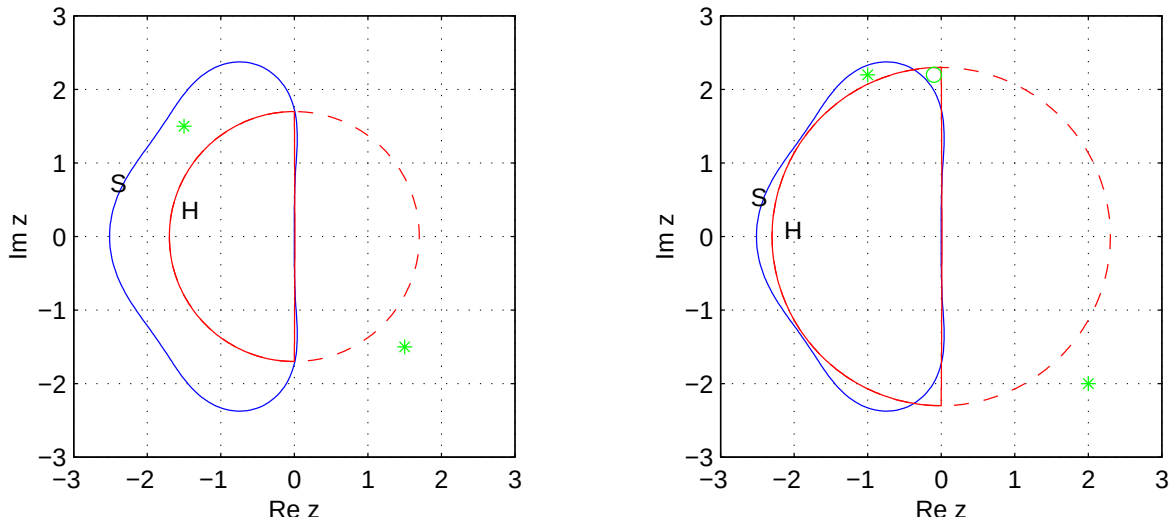


Abbildung 8.1: Stabilitätsgebiet S und Halbkreis H, **links:** Fall 1, **rechts:** Fall 2

Fall 1: Wählt man $r_{\text{stab}} \leq \sqrt{3}$, so ist $H \subset S$, siehe Abbildung 8.1, links. Liegt $\tau_i \lambda_j(\mathbf{J})$ für einen Eigenwert $\lambda_j(\mathbf{J})$ an einer der durch $(*)$ markierten Positionen, so ist das Problem H-steif, aber nicht S-steif. Es gibt in diesem Fall kein Problem, das S-steif, aber nicht H steif ist.

Fall 2: Wählt man $r_{\text{stab}} > \sqrt{3}$, so ist $H \not\subset S$, siehe Abbildung 8.1, rechts. Liegt $\tau_i \lambda_j(\mathbf{J})$ an einer der durch $(*)$ markierten Positionen, so ist das Problem H-steif, aber nicht S-steif. Liegt $\tau_i \lambda_j(\mathbf{J})$ an der durch (o) markierten Position, so ist das Problem S-steif, aber nicht H-steif.

Im Fall 1 ist man auf der sicheren Seite, da mit den H-steifen Komponenten auch alle S-steifen Komponenten erfaßt werden. Hingegen kann es im Fall 2 vorkommen, daß gewisse S-steife Komponenten nicht erkannt werden. Dafür ist im Fall 2 jedoch eine bessere Approximation von S durch H möglich. Zudem ist es bei einigen expliziten Runge-Kutta-Verfahren gar nicht möglich, einen Halbkreis H in der oben angegebenen Form zu finden, der gänzlich in S liegt. In der Praxis wird man oft einen Kompromiß suchen zwischen dem Risiko, das Fall 2 mit sich bringt und der besseren Approximation bei größerem r_{stab}. Mitunter kann man auch gewisse Informationen über die Eigenwerte $\lambda_j(\mathbf{J})$ ausnutzen. Ist beispielsweise bekannt, daß zumindest die betragsgroßen Eigenwerte reell sind, so kann $r_{\text{stab}} := \min(S \cap \mathbb{R})$ gewählt werden.

Da der Spektralradius $\varrho(\mathbf{J})$ keinen Aufschluß über einzelne Zeilen oder Spalten von $\mathbf{J}$ gibt, approximieren wir ihn durch eine geeignete Matrixnorm. Hier bieten sich insbesondere die Zeilen- und die Spaltensummennorm an, die folgendermaßen definiert sind.

Definition 8.1. Es sei $\mathbf{A} = (a_{ij})_{i,j=1,\ldots,n}$ eine $n \times n$-Matrix. Die Matrixnormen

$$\|\mathbf{A}\|_{M,1} := \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ji}| \qquad \text{und} \qquad \|\mathbf{A}\|_{M,\infty} := \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|$$

werden als **Spalten-** bzw. **Zeilensummennorm** bezeichnet. □

Wir definieren noch die folgenden Vektornormen:

Definition 8.2. Für einen Vektor $\mathbf{v} = (v_1, \ldots, v_n)$ sind mit

$$\|\mathbf{v}\|_{V,1} := \sum_{i=1}^{n} |v_i| \qquad \text{und} \qquad \|\mathbf{v}\|_{V,\infty} := \max_{i=1,\ldots,n} |v_i|$$

die l^1- und die l^∞-Norm, die auch als Maximumnorm bezeichnet wird, definiert. □

Es zeigt sich, daß Spalten- und Zeilensummennorm gerade die durch l^1- und l^∞-Norm erzeugten Matrixnormen sind.

Satz 8.3. *Für eine $n \times n$-Matrix $\mathbf{A}$ ist*

$$\|\mathbf{A}\|_{M,1} = \max_{\mathbf{v}\in\mathbb{R}^n, \|v\|_{V,1}=1} \|\mathbf{A}\mathbf{v}\|_{V,1} \qquad \textit{und} \qquad \|\mathbf{A}\|_{M,\infty} = \max_{\mathbf{v}\in\mathbb{R}^n, \|v\|_{V,\infty}=1} \|\mathbf{A}\mathbf{v}\|_{V,\infty}.$$

Beweis. Siehe Zurmühl/Falk [179, Abschnitt 25.2]. □

Jede Matrixnorm, die in der angegebenen Weise aus einer Vektornorm erzeugt wurde, ist eine obere Schranke des Spektralradius:

Satz 8.4. *Es sei $\|\cdot\|_M$ die einer Vektornorm $\|\cdot\|_V$ zugeordnete Matrixnorm, d.h. für alle $\mathbf{A} \in \mathbb{R}^{n\times n}$ gelte*

$$\|\mathbf{A}\|_M := \max_{\mathbf{v}\in\mathbb{R}^n, \|v\|_V=1} \|\mathbf{A}\mathbf{v}\|_V.$$

Dann gilt stets $\varrho(\mathbf{A}) \leq \|\mathbf{A}\|_M$.

Beweis. Siehe Zurmühl/Falk [179, Abschnitt 25.3]. □

Zeilen- und Spaltensummennorm liefern i.a. eine eher grobe Abschätzung des Spektralradius. Im Hinblick auf die Partitionierung bieten diese beiden Normen jedoch den Vorteil, daß man sofort erkennt, welche Zeile bzw. Spalte für die Höhe der Norm einer Matrix verantwortlich ist. Wir betrachten im folgenden die Approximation des Spektralradius durch die Zeilensummennorm. Die Elemente von $\mathbf{J}$ bezeichnen wir mit j_{kl}. Ersetzen wir die Bedingung (8.7) durch die i.a. etwas schärfere Bedingung

$$\tau_i \|\mathbf{J}\|_{M,\infty} \leq r_{\text{stab}}$$

so können wir unschwer feststellen, welche Zeilen von $\mathbf{J}$ für eine eventuelle Verletzung dieser Bedingung verantwortlich sind. Die Indizes dieser Zeilen fassen wir in der Menge

$$I := \left\{ k \in \mathbb{N} : \tau_i \sum_{l=1}^{n} |j_{lk}| > r_{\text{stab}} \right\}$$

zusammen.

Wie wir in Abschnitt (5.6.1) gesehen hatten, kann die Steifheit einer Reaktions-Diffusions-Gleichung sowohl vom Diffusions- als auch vom Reaktionsterm ausgehen. So kann auch die Jacobi-Matrix $\mathbf{J}$ in einen Diffusionsteil $\mathbf{J}^{\text{diff}}$ und einen Reaktionsteil $\mathbf{J}^{\text{reac}}$ aufgespalten werden. Im Ergebnis der Ortsdiskretisierung erhalten wir das System (8.1), dessen rechte Seite wir in einen Diffusions- und einen Reaktionsteil aufspalten können:

$$\mathbf{u}_t = \mathbf{f}(t, \mathbf{u}) = \mathbf{f}^{\text{diff}}(t, \mathbf{u}) + \mathbf{f}^{\text{reac}}(t, \mathbf{u})$$

schreiben. In naheliegender Weise definieren wir $\mathbf{J}^{\text{diff}} := \partial \mathbf{f}^{\text{diff}} / \partial \mathbf{u}$ und $\mathbf{J}^{\text{reac}} := \partial \mathbf{f}^{\text{reac}} / \partial \mathbf{u}$. Analog zu I bilden wir die Indexmengen I^{diff} und I^{reac} der bezüglich Diffusion bzw. Reaktion steifen Komponenten nach dem folgenden Algorithmus:

Algorithmus 8.5 (Steifheitserkennung nach der Zeilensumme).

gegeben: $\mathbf{J} = (j_{kl})$, $\mathbf{J}^{\text{diff}} = (j_{kl}^{\text{diff}})$, $\mathbf{J}^{\text{reac}} = (j_{kl}^{\text{reac}}) \in \mathbb{R}^{N \times N}$, Zeitschritt τ_i, Sicherheitsfaktor $\alpha = 0{,}8$, Stabilitätsradius r_{stab}

$I := I^{\text{diff}} := I^{\text{reac}} := \emptyset$

for $k = 1, \ldots, N$

if $\tau_i \sum_{l=1}^{N} |j_{kl}| > \alpha r_{\text{stab}}$ (8.8)
$I := I \cup \{k\}$
end

if $\tau_i \sum_{l=1}^{N} |j_{kl}^{\text{diff}}| > \alpha r_{\text{stab}}$ (8.9)
$I^{\text{diff}} := I^{\text{diff}} \cup \{k\}$
end

if $\tau_i \sum_{l=1}^{N} |j_{kl}^{\text{reac}}| > \alpha r_{\text{stab}}$ (8.10)
$I^{\text{reac}} := I^{\text{reac}} \cup \{k\}$
end

if $k \in I$ and $k \notin I^{\text{diff}}$ and $k \notin I^{\text{reac}}$
$I^{\text{diff}} := I^{\text{diff}} \cup \{k\}$
$I^{\text{reac}} := I^{\text{reac}} \cup \{k\}$
end

end

□

Dieser Algorithmus garantiert, daß zu jeder H-steifen Zeile von $\mathbf{J}$ mindestens ein Verursacher – Diffusion oder Reaktion – gefunden wird.

Definition 8.6 (Reaktions- und Diffusions-Steifheit). Es sei N die Anzahl der Gitterknoten zum Zeitpunkt t_i. Wir bezeichnen die Größe

$$\sigma_{k,i}^{\text{reac}} := \frac{\tau_i}{r_{\text{stab}}} \sum_{l=1}^{N} |j_{kl}^{\text{reac}}|$$

als **Reaktions-Steifheit** und die Größe

$$\sigma_{k,i}^{\text{diff}} := \frac{\tau_i}{r_{\text{stab}}} \sum_{l=1}^{N} |j_{kl}^{\text{diff}}|$$

als **Diffusions-Steifheit** des Gitterknotens $\mathbf{x}_k$ zum Zeitpunkt t_i. □

In Algorithmus 8.5 wird eine Komponente als steif bezüglich der Diffusion betrachtet, wenn die Diffusions-Steifheit des zugehörigen Gitterknotens größer als α ist. Das entsprechende gilt für die Reaktion.

In analoger Weise kann man auch den Spektralradius $\varrho(\mathbf{J})$ durch die Spaltensummennorm $\|\mathbf{J}\|_{M,1}$ approximieren, was letztendlich auf einen analogen Algorithmus zur Steifheitserkennung führt, den man erhält, wenn man in Algorithmus 8.5

- Zeile (8.8) durch: if $\tau_i \sum_{l=1}^{N} |j_{lk}| > \alpha r_{\text{stab}}$,
- Zeile (8.9) durch: if $\tau_i \sum_{l=1}^{N} |j_{lk}^{\text{diff}}| > \alpha r_{\text{stab}}$ und
- Zeile (8.10) durch: if $\tau_i \sum_{l=1}^{N} |j_{lk}^{\text{reac}}| > \alpha r_{\text{stab}}$

ersetzt. In den in dieser Arbeit vorgenommenen numerischen Berechnungen verwenden wir jedoch ausschließlich die Steifheitserkennung nach Algorithmus 8.5.

Wir stellen im folgenden weitere Varianten der Partitionierung vor, bei denen die lokale Steifheitserkennung nach Algorithmus 8.5

- nur für einen Teil der Gitterknoten oder
- nur für Reaktions- bzw. Diffusionsteil

vorgenommen wird. Im Unterschied zu diesen Varianten bezeichnen wir die Bildung der Indexmengen I^{diff} und I^{reac} nach Algorithmus 8.5 als **vollständig automatische Partitionierung**.

8.2 Weitere Varianten der Partitionierung

Diffusions-Partitionierung

Bei einigen Problemen ist es sinnvoll, nur den Diffusionsteil zu partitionieren, den gesamten Reaktionsteil jedoch durch ein implizites Verfahren zu lösen. Wir bezeichnen ein derartiges Vorgehen hier als **Diffusions-Partitionierung**. Bei einer *skalaren* Reaktions-Diffusions-Gleichung etwa ist die implizite Lösung des Reaktionsteils wegen der Diagonalgestalt der entsprechenden Jacobi-Matrix kaum aufwendiger als die Lösung durch das explizite Verfahren.

Mitunter ist das implizite Verfahren aber genauer als das explizite. Ein derartiges Verhalten werden wir bei den in Kapitel 9 untersuchten Problemen antreffen.

Wir bezeichnen mit $A = \{1, \ldots, N\}$ die Menge aller Knotenindizes. Bei Diffusions-Partitionierung in der hier beschriebenen Form wird nur die Indexmenge I^{diff} entsprechend Algorithmus 8.5 berechnet; hingegen setzt man $I^{\text{reac}} := A$.

Feste Partitionierung

Bei der bereits eingangs erwähnten festen Partitionierung legt man a-priori fest, welche Komponenten des Problems (8.1) in das implizite und welche in das explizite Verfahren eingehen. In diesem Fall muß keine lokale Steifigkeitserkennung durchgeführt werden.

Komponenten-Partitionierung von Reaktions-Diffusions-Systemen

Bei einem Reaktions-Diffusions-System

$$\begin{aligned}\frac{\partial u_k}{\partial t}(\mathbf{x},t) &= \text{div } (d_k(\mathbf{x})\nabla u_k(\mathbf{x},t)) + r_k(\mathbf{x})u_k(\mathbf{x},t) + p_k(u_1(\mathbf{x},t),\ldots,u_m(\mathbf{x},t)) + q_k(\mathbf{x},t),\\ &\quad k = 1,\ldots,m\end{aligned}$$

können bezüglich der einzelnen Komponenten u_k unterschiedliche Partitionierungs-Verfahren zum Einsatz kommen. Wir betrachten ein derartiges Beispiel in Abschnitt 9.2.

8.3 Die Bildung der Partitionierungsmatrix

Nachdem durch Algorithmus 8.5 oder eine entsprechende Variante aus Abschnitt 8.2 die bezüglich Diffusion und Reaktion steifen Komponenten gefunden wurden, wird für die W-Methode (8.3) eine geeignete Matrix $\mathbf{T}$ bestimmt. Hierbei gehen wir nach dem folgenden Algorithmus vor:

Algorithmus 8.7 (Bildung der Partitionierungsmatrix T).

gegeben: $\mathbf{J} = (j_{kl}) \in \mathbb{R}^{N\times N}$, $\mathbf{J}^{\text{diff}} = (j_{kl}^{\text{diff}}) \in \mathbb{R}^{N\times N}$, $\mathbf{J}^{\text{reac}} = (j_{kl}^{\text{reac}}) \in \mathbb{R}^{N\times N}$, I^{diff}, I^{reac}

for $k = 1, \ldots, N$

  for $l = 1, \ldots, N$

    if $k \in I^{\text{diff}}$ and $l \in I^{\text{diff}}$ (8.11)

      $t_{kl}^{\text{diff}} := j_{kl}^{\text{diff}}$

    else

      $t_{kl}^{\text{diff}} := 0$

    end

  end

end

for $k = 1, \ldots, N$

  for $l = 1, \ldots, N$

```
            if k ∈ I^reac and l ∈ I^reac                    (8.12)
                t_kl^reac := j_kl^reac
            else
                t_kl^reac := 0
            end
        end
    end
    T^diff := (t_kl^diff),   T^reac := (t_kl^reac)
    T := T^diff + T^reac
```

□

In der W-Methode (8.3) sind lineare Gleichungssysteme mit der Systemmatrix $(\mathbf{I} - \tau_i\gamma\mathbf{T})$ zu lösen. Bei entsprechender Umordnung von Zeilen und Spalten hat die Systemmatrix die Form

$$\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{M} \end{pmatrix},$$

wobei $\mathbf{I}$ die Einheitsmatrix, $\mathbf{0}$ die Nullmatrix und $\mathbf{M}$ eine den steifen Komponenten entsprechende Untermatrix ist. Die oberen Zeilen entsprechen den bezüglich Diffusion und Reaktion nichtsteifen Komponenten. Das System zerfällt also in einen oberen Teil, dessen Lösung explizit vorliegt, und eine unteren Teil, der die Lösung eines linearen Gleichungssystems erfordert. Hierzu verwenden wir eines der in Kapitel 6 beschriebenen iterativen Lösungsverfahren.

Bemerkung 8.8. Wird die Partitionierungsmatrix entsprechend Algorithmus 8.7 gebildet, so spricht man von einer **schwachen Kopplung** steifer und nichtsteifer Komponenten. Ersetzt man

- Zeile (8.11) durch: if $k \in I^{\text{diff}}$ und
- Zeile (8.12) durch: if $k \in I^{\text{reac}}$,

so erhält man die **starke Kopplung** steifer und nichtsteifer Komponenten. Welche Form der Kopplung sinnvoller ist, hängt von dem zugrundeliegenden Problem ab. Wir verwenden bei den numerischen Berechnungen in dieser Arbeit ausschließlich die schwache Kopplung.

8.4 Das Krylov-W-Verfahren als spezielles Partitionierungs-Verfahren

In gewisser Weise kann man auch das Krylov-W-Verfahren aus Kapitel 7 als Partitionierungs-Verfahren ansehen, denn auch bei dem Krylov-W-Verfahren wird eine W-Methode (8.3) mit einer Matrix $\mathbf{T}$ gelöst, die niedrigen Ranges ist und eine Approximation an die Jacobi-Matrix $\mathbf{J}$ bezüglich des dominanten Eigenraumes darstellt. Es ist daher interessant festzustellen, wie

das dem Krylov-W-Verfahren zugrundeliegende Arnoldi-Verfahren auf ein Problem reagiert, das eigentlich ein klassischer Kandidat für lokale Partitionierung ist. Gelingt es dem Arnoldi-Verfahren hier, die steifen Komponenten gut zu approximieren? Wir wollen das an einem Beispiel untersuchen.

Untersuchung 8.9. Wir betrachten eine Matrix $\mathbf{J} = (j_{kl}) \in \mathbb{R}^{N\times N}$, die gemäß

$$\mathbf{J} = \begin{pmatrix} \mathbf{J}_1 & \mathbf{J}_2 \\ \mathbf{J}_3 & \mathbf{J}_4 \end{pmatrix}$$

aus den vier Teilmatrizen $\mathbf{J}_1 \in \mathbb{R}^{n\times n}$, $\mathbf{J}_2 \in \mathbb{R}^{n\times(N-n)}$, $\mathbf{J}_3 \in \mathbb{R}^{(N-n)\times n}$ und $\mathbf{J}_4 \in \mathbb{R}^{(N-n)\times(N-n)}$ zusammengesetzt ist. Alle betragsgroßen Elemente von $\mathbf{J}$ sollen sich in $\mathbf{J}_1$ befinden. Wir erreichen das durch die folgende Wahl der Elemente j_{kl}:

$$j_{kl} = \begin{cases} -e^{-5\ \mathrm{rand}(0,1)}, & k,l = 1,\ldots,n, \\ -0{,}01e^{-5\ \mathrm{rand}(0,1)}, & \text{sonst.} \end{cases}$$

Die Funktion rand(0,1) liefert gleichmäßig verteilte Zufallszahlen im Intervall $[0,1]$. Wir wählen $n = 10$ und $N = 100$.

Die Elemente einer solchen Matrix $\mathbf{J}$ sind in Abbildung 8.2, links logarithmisch dargestellt. Wir lösen nun das lineare Gleichungssystem $(\mathbf{I} - 0{,}5 \cdot \mathbf{J})\mathbf{x} = \mathbf{b}$ für $\mathbf{b} = (1,\ldots,1)^T$ mit dem in Abschnitt 6.4.1 definierten Arnoldi-Verfahren. Abbildung 8.2, rechts zeigt die Elemente der Matrix $\mathbf{T} := \mathbf{Q}_\varkappa \mathbf{Q}_\varkappa^T \mathbf{J}$ für $\varkappa = 15$. In Abbildung 8.3 ist der Approximationsfehler $|\mathbf{J} - \mathbf{T}|$ logarithmisch dargestellt[1].

Ergebnis der Untersuchung. Es wird deutlich, daß die den steifen Komponenten entsprechende Teilmatrix $\mathbf{J}_1$ am besten approximiert wird. Das Krylov-W-Verfahren reagiert in dieser Hinsicht ähnlich einem automatischen Partitionierungs-Verfahren, welches die Untermatrix $\mathbf{J}_1$ exakt in die Matrix $\mathbf{T}$ einfügen würde. □

[1] Der Betrag $|\cdot|$ wird hierbei elementweise verstanden.

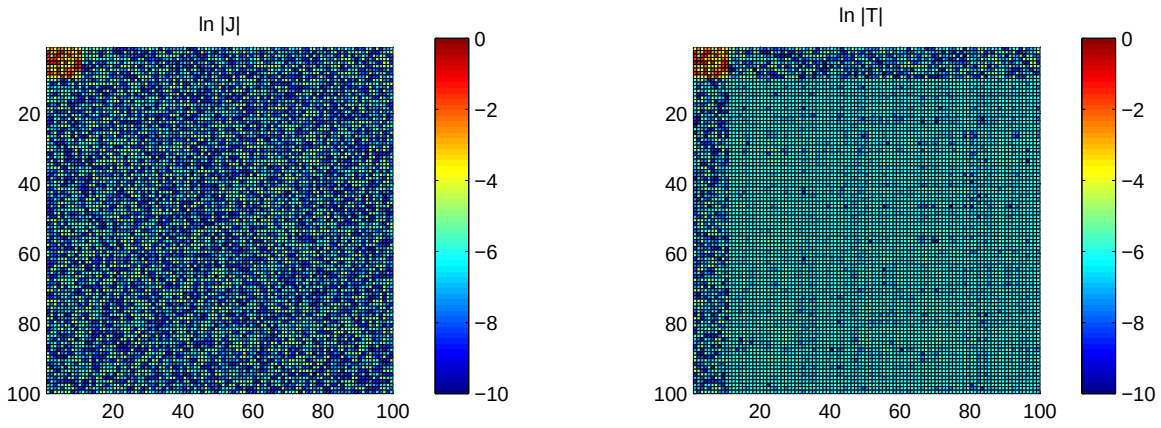


Abbildung 8.2: Betrag der Elemente von $\mathbf{J}$ und $\mathbf{T}$, logarithmische Darstellung

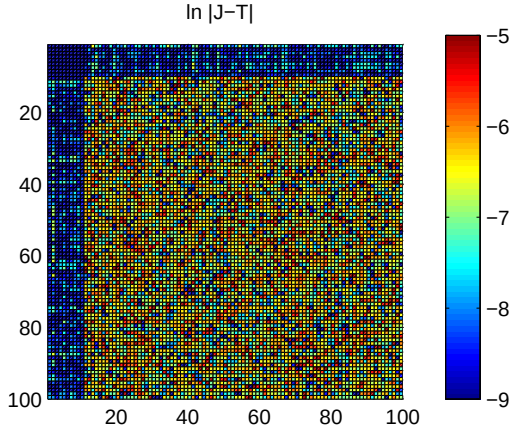


Abbildung 8.3: Betrag der Elemente von $\mathbf{J} - \mathbf{T}$, logarithmische Darstellung

Kapitel 9

Vergleich numerischer Verfahren zur Zeitdiskretisierung

In diesem Kapitel untersuchen wir einige Verfahren zur Zeitdiskretisierung und vergleichen sie bezüglich ihrer Effizienz[1] bei der Lösung dreier Reaktions-Diffusions-Probleme, die wir in den Abschnitten 9.1.1, 9.1.2 und 9.1.3 zunächst vorstellen werden.

9.1 Drei Reaktions-Diffusions-Probleme

9.1.1 TANH – ein Frontproblem mit bekannter Lösung

Es sei $\Omega =]-5,5[^2$. Wir betrachten die Reaktions-Diffusions-Gleichung

$$\frac{\partial u}{\partial t} = \Delta u + r(1-u^2) + 2q^2(u-u^3) \tag{9.1}$$

für eine Funktion $u : \Omega \times [0,1] \to \mathbb{R}$ mit der exakten Lösung

$$u(x,y,t) = \tanh(q(x\cos\varphi + y\sin\varphi - p) + rt), \qquad (x,y) \in \Omega,\ t \in [0,1]. \tag{9.2}$$

Anfangsbedingung und Dirichlet-Randbedingung seien entsprechend der exakten Lösung vorgegeben. Wir verwenden die Parameter $p = 1, q = r = 3, \varphi = \pi/4$.

Die Lösung der Differentialgleichung bildet eine gerade Front, die sich bei $t = 0$ auf der Linie

$$y = -(\cot\varphi)x + \frac{p}{\sin\varphi}$$

befindet und sich mit gleichmäßiger Geschwindigkeit r/q in Richtung des Vektors $(-\cos\varphi, -\sin\varphi)$ bewegt. In Abbildung 9.1 ist die Lösung zur Zeit $t = 0$ dargestellt.

Das Problem ist im Grunde ein eindimensionales, denn die Lösung ist entlang der Geraden $y = -(\cot\varphi)x + C$ konstant. Die Differentialgleichung läßt sich unschwer aus der räumlich

[1] Der Begriff der Effizienz wird in Abschnitt 9.4 näher erläutert.

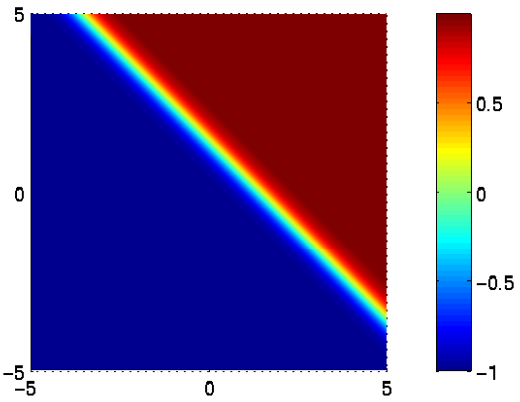


Abbildung 9.1: Lösung des TANH-Problems zur Zeit $t = 0$. Die Front bewegt sich nach links unten.

eindimensionalen Gleichung

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} + r(1-u^2) + 2q^2(u-u^3), \\ u(x,t) &= \tanh(q(x-p)+rt) \end{aligned} \tag{9.3}$$

herleiten, die von Nowak [125] und Lang [103] zu numerischen Untersuchungen herangezogen wurde. Offenbar erfüllt die auf zwei Ortsvariablen definierte Funktion

$$u(x,y,t) = \tanh(q(x-p)+rt) \tag{9.4}$$

die Differentialgleichung

$$\frac{\partial u}{\partial t} = \Delta u + r(1-u^2) + 2q^2(u-u^3), \tag{9.5}$$

da für die Lösung (9.4) $\partial^2 u/\partial y^2 = 0$ und damit $\partial^2 u/\partial x^2 = \Delta u$ ist. Das Problem (9.1), (9.2) entsteht aus (9.5), (9.4) durch eine Drehung des Koordinatensystems um den Winkel φ, da diese Drehung keinen Einfluß auf Δu hat.

9.1.2 BSVD – eine bistabile Diffusionsgleichung mit ortsabhängiger Diffusion

Es sei $\Omega =]0,1[^2$. Wir betrachten die bistabile Diffusionsgleichung

$$\begin{aligned} \frac{\partial u}{\partial t} &= \nabla \cdot (d(x,y)\nabla u) + \beta_0(1-u^2)(u-\beta_1), \\ d(x,y) &= \alpha_0 \left(\sum_{i=1}^{3} e^{-\alpha_i((x-x_i)^2+(y-y_i)^2)} \right), && (x,y) \in \Omega, \\ u(x,y,0) &= 2e^{-(\gamma_0(x-x_0)^2+(y-y_0)^2)} - 1, && (x,y) \in \Omega \end{aligned}$$

für eine Funktion $u : \Omega \times [t_0, t_e] \to \mathbb{R}$ mit der Neumannschen Randbedingung

$$d(x,y)\mathbf{n}_{\partial\Omega}(x,y) \cdot \nabla u(x,y,t) = 0, \qquad (x,y) \in \partial\Omega,\ t \in [0,t_e].$$

Die auftretenden Parameter seien wie folgt gewählt:

$$\alpha_0 = 0{,}1, \qquad \alpha_1 = \alpha_2 = \alpha_3 = 100,$$

$$\beta_0 = 10, \qquad \beta_1 = -0{,}6, \qquad \gamma_0 = 10,$$

$$x_0 = x_1 = x_2 = x_3 = 0{,}5,$$

$$y_0 = -0{,}1, \qquad y_1 = 0{,}6, \qquad y_2 = 0{,}75, \qquad y_3 = 0{,}9.$$

In dieser Differentialgleichung ist der Diffusionskoeffizient d stark vom Ort abhängig, siehe Abbildung 9.2.

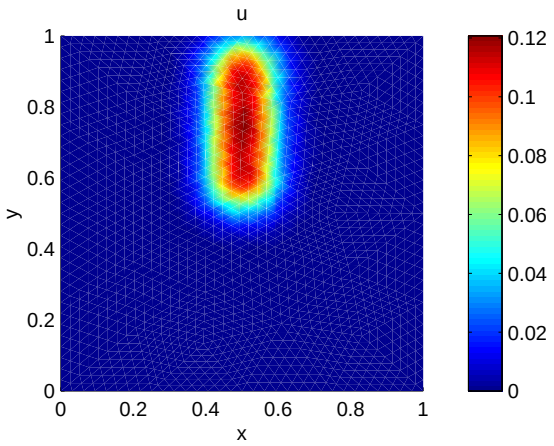


Abbildung 9.2: Diffusionskoeffizient d

Die Lösung der bistabilen Diffusionsgleichung bildet eine bewegte Front, deren Geschwindigkeit stark von der Diffusion abhängt – eine hohe Diffusion beschleunigt die Frontbewegung. Anfangs befindet sich die Front größtenteils in einem Bereich geringer Diffusion. Lediglich in einem kleinen mittleren Frontabschnitt ist die Diffusion etwas höher. Dort wird die Front stark beschleunigt und dadurch in den Bereich hoher Diffusion hineingezogen. Später verläßt die Front das Gebiet hoher Diffusion wieder. Die Lösung des Problems ist in Abbildung 9.3 dargestellt.

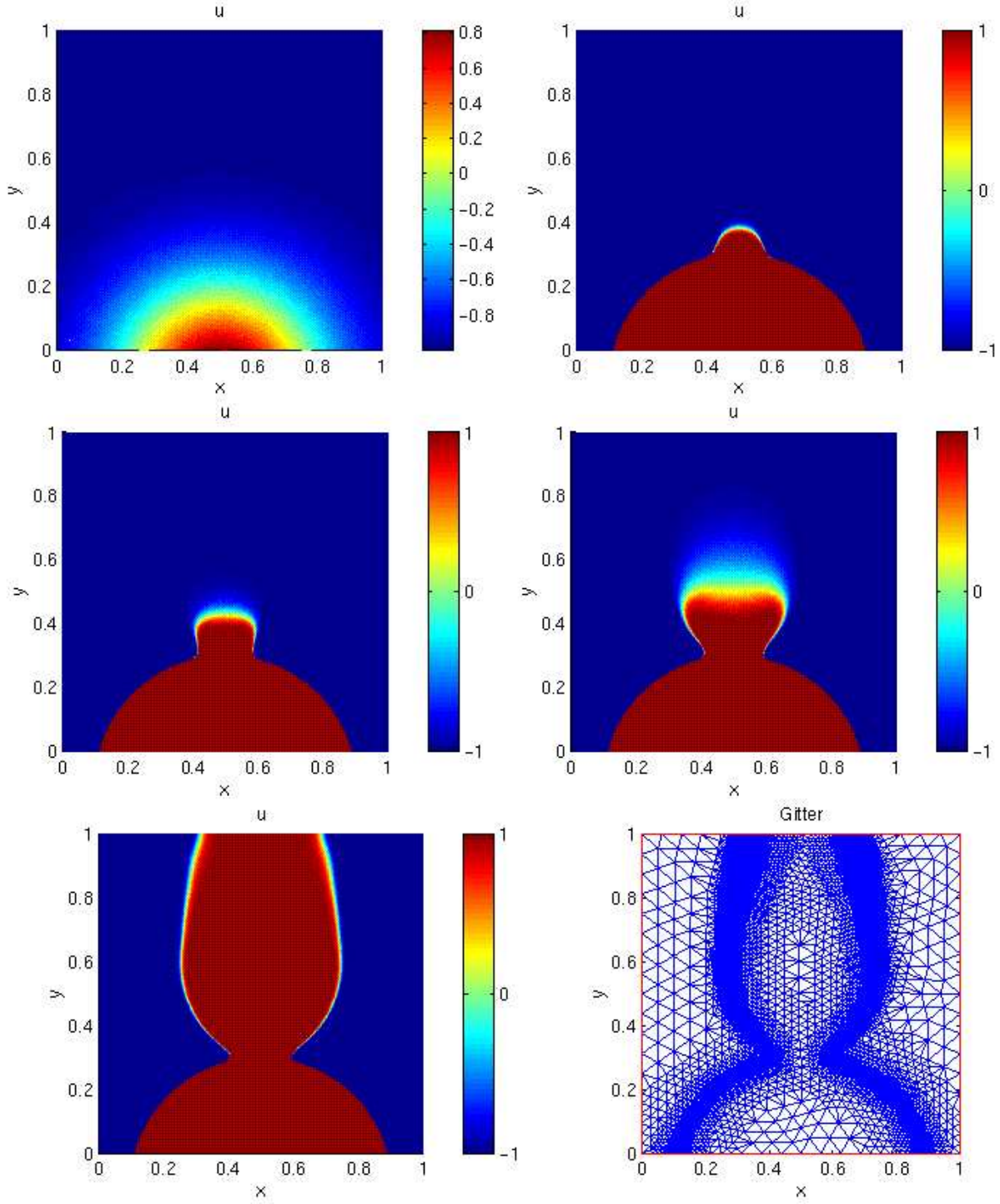


Abbildung 9.3: Lösung des BSVD-Problems zu den Zeiten $t = 0, 5, 6, 7, 8$, Gitter zur Zeit $t = 8$

9.1.3 KRINSKY – das System von Krinsky et al.

Das dritte Problem, das wir untersuchen wollen, ist das System von KRINSKY et al. zur Modellierung eines erregbaren Mediums. In Kapitel 10 werden wir erregbare Medien ausführlich beschreiben und eine Reihe von Untersuchungen darstellen, die mit diesem Reaktions-Diffusions-System durchgeführt wurden. Es sei $\Omega =]-20,20[^2$. Wir betrachten das System

$$\begin{aligned}\frac{\partial u}{\partial t} &= d_1 \Delta u + f(u,v),\\ \frac{\partial v}{\partial t} &= d_2 \Delta v + g(u,v)\end{aligned}$$

für Funktionen $u, v : \Omega \times [0,10] \to \mathbb{R}$. Die rechten Seiten seien

$$f(u,v) = \begin{cases} -k_1 u - v, & u < \sigma, \\ k_f(u-a) - v, & \sigma < u < 1-\sigma, \\ k_2(1-u) - v, & 1-\sigma < u, \end{cases} \qquad g(u,v) = \begin{cases} \varepsilon(k_g u - v), & k_g u \geq v, \\ \varepsilon k_\varepsilon (k_g u - v), & k_g u < v. \end{cases}$$

Wir wählen die Parameter

$$d_1 = 1, \qquad d_2 = 0, \qquad k_1 = 15{,}3, \qquad k_2 = 151{,}3, \qquad k_f = 1{,}7,$$

$$k_g = 2, \qquad k_\varepsilon = 6, \qquad a = 0{,}1, \qquad \sigma = 0{,}01, \qquad \varepsilon = 0{,}2.$$

Auf $\partial\Omega$ werden homogene Neumannsche Randbedingungen $\partial u/\partial \mathbf{n}_{\partial\Omega} = \partial v/\partial \mathbf{n}_{\partial\Omega} = 0$ vorgegeben. Als Anfangsbedingung wählen wir

$$\begin{aligned} u(x,y,0) &= u_0(x,y) = \begin{cases} 1, & y \in [-17,-14], \\ 0, & \text{sonst}, \end{cases} \\ v(x,y,0) &= v_0(x,y) = \begin{cases} 1{,}5, & y \leq -17, \\ -y/2 + 7, & -17 \leq y \leq -14, \\ 0, & y \geq -14. \end{cases} \end{aligned}$$

Das Reaktions-Diffusions-System modelliert eine waagerechte Erregungswelle, die sich mit gleichmäßiger Geschwindigkeit nach oben bewegt. Da die Anfangsverteilung (u_0, v_0) nicht von x abhängt, trifft dies auch auf die Lösung des Problems zu einem beliebigen Zeitpunkt t zu. Die Lösung des Problems KRINSKY kann demnach auch durch Berechnung eines räumlich eindimensionalen Systems gewonnen werden.

9.2 Die ausgewählten numerischen Verfahren

Zur Ortsdiskretisierung der Reaktions-Diffusions-Probleme verwenden wir stets lineare finite Elemente auf einem adaptiv verfeinerten Dreiecksgitter. Die in diesem Kapitel untersuchten

numerischen Verfahren zur Zeitdiskretisierung basieren alle auf der dreistufigen W-Methode

$$\begin{aligned}
\left(\mathbf{I}-\tau_i\left(1-\sqrt{2}/2\right)\mathbf{T}\right)\widetilde{\mathbf{k}}_1 &= \mathbf{f}(t_i,\mathbf{u}_i), \qquad (9.6)\\
\left(\mathbf{I}-\tau_i\left(1-\sqrt{2}/2\right)\mathbf{T}\right)\widetilde{\mathbf{k}}_2 &= \mathbf{f}\left(t_{i+1},\mathbf{u}_i+\tau_i\widetilde{\mathbf{k}}_1\right)-\left(2+\sqrt{2}\right)\widetilde{\mathbf{k}}_1,\\
\left(\mathbf{I}-\tau_i\left(1-\sqrt{2}/2\right)\mathbf{T}\right)\widetilde{\mathbf{k}}_3 &= \mathbf{f}\left(t_{i+1},\mathbf{u}_i+\tau_i\widetilde{\mathbf{k}}_1\right)-\widetilde{\mathbf{k}}_1+\left(-1+\sqrt{2}\right)\widetilde{\mathbf{k}}_2,\\
\mathbf{u}_{i+1} &= \mathbf{u}_i+\frac{\tau_i}{2}\left(2\widetilde{\mathbf{k}}_1+\left(1-\sqrt{2}\right)\widetilde{\mathbf{k}}_2+\widetilde{\mathbf{k}}_3\right),\\
\widehat{\mathbf{u}}_{i+1} &= \mathbf{u}_i+\frac{\tau_i}{20}\left(\left(18-\sqrt{2}\right)\widetilde{\mathbf{k}}_1+\left(9-11\sqrt{2}\right)\widetilde{\mathbf{k}}_2+\left(11+\sqrt{2}\right)\widetilde{\mathbf{k}}_3\right),
\end{aligned}$$

siehe (5.18). Die W-Methode ist unabhängig von der Wahl der Matrix $\mathbf{T}$ von zweiter Ordnung. In den Testrechnungen verwenden wir die folgenden Verfahren, die wir abkürzend durch einen Buchstaben in Klammern kennzeichnen:

(E): das explizite Verfahren

$$\begin{aligned}
\mathbf{k}_1 &= \mathbf{f}(t_i,\mathbf{u}_i), \qquad (9.7)\\
\mathbf{k}_2 &= \mathbf{f}(t_{i+1},\mathbf{u}_i+\tau_i\mathbf{k}_1),\\
\mathbf{u}_{i+1} &= \mathbf{u}_i+\frac{\tau_i}{2}(\mathbf{k}_1+\mathbf{k}_2),\\
\widehat{\mathbf{u}}_{i+1} &= \mathbf{u}_i+\frac{\tau_i}{20}(9\mathbf{k}_1+11\mathbf{k}_2),
\end{aligned}$$

das aus (9.6) hervorgeht, wenn $\mathbf{T}=\mathbf{0}$ gesetzt wird,

(I): die entsprechende, in Abschnitt 5.7 definierte W-Methode mit Jacobi-Matrix[2],

(A): das in Abschnitt 8.1 angegebene Verfahren zur vollständig automatischen Partitionierung sowie

(D): das in Abschnitt 8.2 angegebene Verfahren zur Diffusions-Partitionierung.

Bei den Partitionierungs-Verfahren benutzen wir stets die Steifheitserkennung nach der Zeilensumme, siehe Algorithmus 8.5. Zur Lösung der in den Verfahren (I), (A) und (D) auftretenden linearen Gleichungssysteme verwenden wir wahlweise das BiCGstab-Verfahren (B), siehe Abschnitt 6.3, oder den multiplen Arnoldi-Prozeß (K)[3], siehe Abschnitt 6.5 und Kapitel 7. Wir verwenden das BiCGstab-Verfahren

- ohne Vorkonditionierung oder
- mit SSOR-Vorkonditionierung (S).

Zur Abkürzung bezeichnen wir die Verfahren mit den entsprechenden Buchstabenkombinationen – die Bezeichnung (IBS) etwa steht für implizites Verfahren mit BiCGstab-Löser und SSOR-Vorkonditionierung. Für das Problem KRINSKY können unterschiedliche Verfahren für die Komponenten u und v verwendet werden, wir benutzen jedoch stets *den gleichen linearen Löser für beide Komponenten*. Die Bezeichnung (IEB) bedeutet hier beispielsweise: implizit bezüglich u-, explizit bezüglich v-Komponente, BiCGstab-Löser ohne Vorkonditionierung.

[2] Die Bezeichnung (I) steht für „implizit“.
[3] für „Krylov-W-Verfahren“

9.3 Referenzlösungen

Zur Auswertung der bei den einzelnen Verfahren auftretenden Fehler benötigen wir die exakte Lösung der Testprobleme. Da die exakte Lösung der Probleme BSVD und KRINSKY nicht bekannt ist, müssen wir sie durch eine hinreichend genaue Referenzlösung simulieren. Für das Problem Krinsky kann die Referenzlösung auf einem räumlich eindimensionalen Gebiet gewonnen werden. In der folgenden Tabelle geben wir die Verfahren und Parameter[4] an, die wir zur Berechnung der Referenzlösung verwenden:

	BSVD	KRINSKY	siehe
Gebiet	$\Omega =]0,1[^2$	$\Omega =]-20,20[^2$	
Zeitintervall	$[t_0, t_e] = [0,7{,}5]$	$[t_0, t_e] = [0,10]$	
Ortsdiskretisierung	lin. FE auf adaptivem Dreiecksgitter	Differenzenverfahren auf uniformem Gitter	Kap. 3
Maschenweite des uniformen (Grund-)Gitters	$h_0 = 0{,}2$	$h = 0{,}01$	
Fehlerindikator	Z^2	–	4.1.1
max. Gitterfeinheit	$M = 6$	–	4.3
Sprungstellen von zfh	$\mu(0) = \mu(1) = 0$, $\mu(2) = \mu(3) = 1/24$, $\mu(4) = 1/12$, $\mu(5) = 5/12$	–	(4.9)
Strategie der Verfeinerung	Strategie 3, $\alpha = 0{,}05$, $k_{\max} = 10$	–	4.3.2
Zeitdiskretisierungs-Verfahren	(IBS)	(EE)	9.2
Zeitschritt	$\tau = 5 \cdot 10^{-4}$, falls $t \in [0,1[$; $\tau = 10^{-3}$, falls $t \in [1,5[$; $\tau = 10^{-4}$, falls $t \in [5,7{,}5]$	$\tau = 10^{-6}$, falls $t \in [0,0{,}02[$; $\tau = 10^{-5}$, falls $t \in [0{,}02, 10]$	(5.4)
Toleranz des zeitlich lokalen Fehlers	10^{-6}	–	(6.16)
Parameter des linearen Lösers	$\alpha_{\mathrm{LSS}} = 0{,}2$	–	(6.16)
Norm des Residuums des linearen Lösers	skalierte Euklidische Norm	–	Bem. 5.9
max. Anzahl der Iterationen	$\max_{\mathrm{It}} = 10$	–	Alg. 6.6
SSOR-Parameter	$\omega = 1{,}3$	–	(6.7)

Tabelle 9.1: Parameter der Referenzrechnungen

[4] Die in der Tabelle angegebene Toleranz des zeitlich lokalen Fehlers wird hier *nicht* zur Zeitschrittsteuerung sondern nur zur Berechnung von TOL_{LSS} verwendet!

9.4 Effizienz

Die **Effizienz** eines numerischen Verfahrens wird durch zwei Größen bestimmt: den Fehler der numerischen Lösung und den Rechenaufwand. Ein Verfahren ist effizienter als ein anderes, wenn

- es bei gleichem Fehler einen geringeren Aufwand verlangt oder
- bei gleichem Aufwand der Fehler geringer ausfällt.

Folglich läßt sich die Effizienz von Verfahren besonders gut in einem Koordinatensystem darstellen, in dem der Aufwand über dem Fehler aufgetragen wird.

Sowohl Fehler als auch Rechenaufwand können in verschiedener Form gemessen werden. Bei dem Problem TANH benutzen wir den zeitlich gemittelten L^2-Fehler:

Definition 9.1. Gegeben sei das TANH-Problem aus Abschnitt 9.1.1. Es sei $u(\mathbf{x},t) \in L^2(]t_0,t_e[,H^1(\Omega))$ die exakte Lösung und $u_h^\tau(\mathbf{x},t)$ die bezüglich der Zeit stückweise konstante und bezüglich des Ortes stückweise lineare Näherungslösung des Problems. Dann bezeichnen wir den Wert

$$ERR_{L^2}^{\mathrm{T}} := \frac{1}{t_e - t_0} \int_{t_0}^{t_e} \|u(\mathbf{x},t) - u_h^\tau(\mathbf{x},t)\|_{L^2(\Omega)}\, dt$$

als den **mittleren L^2-Fehler der Näherungslösung des Problems TANH** im Zeitintervall $[t_0,t_e]$. □

Bei dem Problem BSVD werten wir den mittleren relativen Fehler der Frontposition aus:

Definition 9.2. Gegeben sei das BSVD-Problem aus Abschnitt 9.1.2. Mit Hilfe der in Abschnitt 9.3 beschriebenen Referenzlösung werde die exakte Lösung des Problems simuliert. Wir bezeichnen die Referenzlösung mit $u(x,y,t)$ und die Näherungslösung mit $u_h^\tau(x,y,t)$. Die Frontposition $(0,y_{\mathrm{fr}}(t))$ der Referenzlösung sei der Schnittpunkt der Kurve $u(x,y,t)=0$ mit der Geraden $x=0$ zur Zeit t. Analog sei die numerische Frontposition $(0,y_{\mathrm{fr,num}}(t))$ der Schnittpunkt der Kurve $u_h^\tau(x,y,t)=0$ mit der Geraden $x=0$ zur Zeit t. Dann definieren wir den **mittleren relativen Fehler der Frontposition für das Problem BSVD** als den Wert

$$ERR_{\mathrm{fr}}^{\mathrm{B}} := \frac{1}{t_e - t_0} \int_{t_0}^{t_e} \left| \frac{y_{\mathrm{fr}}(t) - y_{\mathrm{fr,num}}(t)}{y_{\mathrm{fr}}(t_e) - y_{\mathrm{fr}}(t_0)} \right| dt.$$

□

Bei dem Problem Krinsky verwenden wir hingegen den relativen Fehler der Frontposition zum Zeitpunkt t_e:

Definition 9.3. Gegeben sei das KRINSKY-Problem aus Abschnitt 9.1.3. Es seien $u(x,y,t)$ die Referenzlösung und $u_h^\tau(x,y,t)$ die Näherungslösung der ersten Komponente dieses Reaktions-Diffusions-Systems. Die Frontposition $(0,y_{\mathrm{fr}}(t))$ der Referenzlösung sei der Schnittpunkt

der Kurve $u(x, y, t) = 0{,}5$ mit der Geraden $x = 0$ zur Zeit t. Analog sei die numerische Frontposition $(0, y_{\mathrm{fr,num}}(t))$ der Schnittpunkt der Kurve $u_h^\tau(x, y, t) = 0{,}5$ mit der Geraden $x = 0$ zur Zeit t. Wir definieren den **relativen Fehler der Frontposition zur Zeit $t = t_e$ für das Problem KRINSKY** als den Wert

$$ERR_{\mathrm{fr}}^{\mathrm{K}} := \left| \frac{y_{\mathrm{fr}}(t_e) - y_{\mathrm{fr,num}}(t_e)}{y_{\mathrm{fr}}(t_e) - y_{\mathrm{fr}}(t_0)} \right|.$$

□

Zur Bewertung des Rechenaufwandes verwenden wir

- die Messung der Rechenzeit oder
- eine Zählung der ausgeführten Multiplikationen.

Die Anzahl der Multiplikationen erweist sich als ein zuverlässiges Maß für die benötigte Rechenzeit. Eine direkte Messung der Rechenzeit hat den Nachteil, daß diese Größe maschinenabhängig ist und eventuelle Schwankungen in der Rechengeschwindigkeit die Messung beeinflussen. Diese Methode vermeiden wir deshalb bei Langzeitrechnungen.

9.5 Beschränkung der Iteration – eine numerische Studie

Die Frage, wann die Iterationen des Gleichungslösers – also des BiCGstab- oder des Arnoldi-Verfahrens – abgebrochen werden sollten, ist für eine effiziente Lösung von entscheidender Bedeutung. In den Abschnitten 6.3.2 und 7.5 wurden hierfür zwei Möglichkeiten erörtert:

- Ein Abbruch der Iteration erfolgt, wenn das Residuum unter eine Toleranz TOL_{LSS} fällt. Die Größe dieser Toleranz wird durch den Parameter α_{LSS} bestimmt, siehe (6.16), (7.12).
- Eine fest gewählte Obergrenze beschränkt die Anzahl der Iterationen. Diese ist beim BiCGstab-Verfahren durch $\max_{\mathrm{It}}$, beim multiplen Arnoldi-Prozeß durch $\varkappa_1$, $\varkappa_{12}$ und $\varkappa_{23}$ gegeben.

Numerische Experimente zeigen, daß oft eine Kombination aus beiden Formen der Iterationsbeschränkung am erfolgreichsten ist; jedoch halten wir es für angebracht, der Steuerung durch den Parameter α_{LSS} eine gewisse Priorität einzuräumen, da eine solche Steuerung des Iterationsabbruchs theoretisch begründbar ist[5] und flexibel auf Größen wie TOL_t und τ reagieren kann. Der Iterationsprozeß sollte also in der Regel durch das Toleranz-Kriterium abgebrochen werden.

Untersuchung 9.4 (Beschränkung der Iteration). In einer umfangreichen Parameterstudie anhand der in den Abschnitten 9.1.1 und 9.1.2 dargestellten Probleme TANH und BSVD wurden die Parameter α_{LSS}, $\max_{\mathrm{It}}$, $\varkappa_1$, $\varkappa_{12}$, $\varkappa_{23}$ und TOL_t variiert und die Effizienz der in Abschnitt 9.2 genannten Verfahren verglichen.

[5] siehe Abschnitt 6.3.2

Durch Auswertung der Ergebnisse dieser Studie erscheinen uns die folgenden Werte für die genannten Parameter als eine sinnvolle Wahl:

- Verfahren mit BiCGstab-Löser ohne Vorkonditionierung: $\max_{\mathrm{It}} = 8$,
- Verfahren mit BiCGstab-Löser mit SSOR-Vorkonditionierung: $\max_{\mathrm{It}} = 5$,
- Verfahren mit dem multiplen Arnoldi-Prozeß: $\varkappa_1 = 10$, $\varkappa_{12} = \varkappa_{23} = 2$

Wir verdeutlichen den erheblichen Nutzen einer Iterationsbeschränkung am Beispiel des BSVD-Problems, das mit dem Verfahren (IK) gelöst wird. Tabelle 9.2 gibt die verwendeten Parameter an.

		siehe
Gebiet	$\Omega =]0,1[^2$	
Zeitintervall	$[t_0, t_e] = [0,7{,}5]$	
Maschenweite des uniformen Grundgitters	$h_0 = 0{,}2$	
Fehlerindikator	Z^2	4.1.1
max. Gitterfeinheit	$M = 6$	4.3
Sprungstellen von zfh	$\mu(0) = \mu(1) = 0$, $\mu(2) = \mu(3) = 1/24$, $\mu(4) = 1/12$, $\mu(5) = 5/12$	(4.9)
Strategie der Verfeinerung	Strategie 3, $\alpha = 0{,}05$, $k_{\max} = 10$	4.3.2
Zeitdiskretisierungs-Verfahren	(IK)	9.2
Anfangszeitschritt	$\tau_0 = 0{,}01$	(5.4)
Parameter der Zeitschrittsteuerung	$\beta = 0{,}8$, $\beta_{\min} = 0{,}5$, $\beta_{\max} = 1{,}1$	(5.8), (5.9)
Toleranz des zeitlich lokalen Fehlers	10^{-4}	(5.8)
Norm des zeitlichen Fehlerschätzers	skalierte Euklidische Norm	Bem. 5.9
Stabilitäts-Radius	$r_{\mathrm{stab}} = 2$	(8.6), Alg. 8.5
Steifigkeits-Sicherheitsparameter	$\alpha = 0{,}8$	Alg. 8.5
Parameter des linearen Lösers	$\alpha_{\mathrm{LSS}} = 0{,}5,\ 1,\ 2,\ 5,\ 10,\ 20$	(6.16), (7.12)
max. Anzahl der Iterationen	3 Fälle: 1. $\varkappa_1 = \varkappa_{12} = \varkappa_{23} = 20$, 2. $\varkappa_1 = \varkappa_{12} = \varkappa_{23} = 10$, 3. $\varkappa_1 = 10$, $\varkappa_{12} = \varkappa_{23} = 2$	Alg. 6.6, 7.1
SSOR-Parameter	$\omega = 1{,}3$	(6.7)

Anzahl der Testrechnungen	3 Fälle zur max. Anzahl der Iterationen $\times$ 6 verschiedene $\alpha_{\mathrm{LSS}} = 18$ Testrechnungen	

Tabelle 9.2: BSVD-Problem, Verfahren (IK): Untersuchungen zur Beschränkung der Iteration

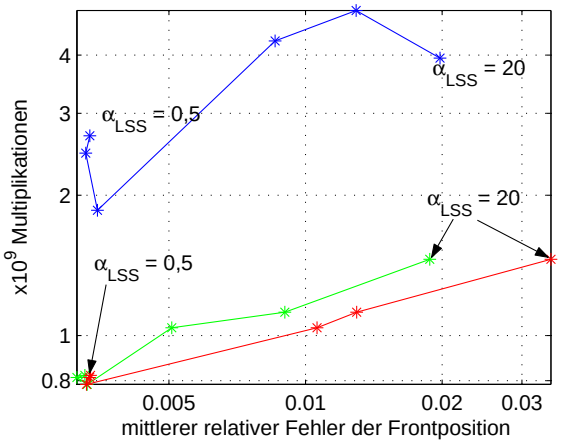


Abbildung 9.4: BSVD-Problem, Verfahren (IK): **blau:** $\varkappa_1 = \varkappa_{12} = \varkappa_{23} = 20$, **rot:** $\varkappa_1 = \varkappa_{12} = \varkappa_{23} = 10$, **grün:** $\varkappa_1 = 10$, $\varkappa_{12} = \varkappa_{23} = 2$, jeweils $\alpha_{\mathrm{LSS}} = 0{,}5, 1, 2, 5, 10, 20$

In Abbildung 9.4 stellen wir den Rechenaufwand über dem mittleren relativen Fehler der Frontposition dar. Aus der Graphik geht hervor, daß die beiden Varianten mit $\varkappa_1 = 10$ (rot, grün) wesentlich effizienter sind als die Variante mit $\varkappa_1 = 20$ (blau). Die Steuerung des Iterationsabbruchs durch α_{LSS} reicht deshalb nicht aus, sondern $\varkappa_1$ muß hinreichend niedrig gewählt werden, um den Rechenaufwand zu senken. Die Beschränkung von $\varkappa_{12}$ und $\varkappa_{23}$ auf zwei (grüne Kurve) ist möglich, aber nicht zwingend notwendig. □

9.6 Weitere numerische Untersuchungen

9.6.1 Problem TANH

In diesem Abschnitt stellen wir die Ergebnisse dreier numerischer Untersuchungen mit dem Problem TANH dar. Die folgende Tabelle veranschaulicht die in den einzelnen Untersuchungen verwendeten Verfahren und Parameter.

Name	Untersuchung 9.5	Untersuchung 9.6	Untersuchung 9.7	siehe
Dgl.-Parameter	$p = 1,\ q = r = 3,\ \varphi = \pi/4$			(9.1)
Gebiet	$\Omega =]-5,5[^2$			
Zeitintervall	$[t_0, t_e] = [0, 1]$		$[t_0, t_e] = [0, 5]$	
Maschenweite des uniformen Grundgitters	$h_0 = 2$			
Fehlerindikator	Z^2			4.1.1
max. Gitterfeinheit	$M = 4, 5, 6$	$M = 5$		4.3
Sprungstellen von zfh	$\mu(0) = \mu(1) = 0,\ \mu(2) = \cdots = \mu(5) = 1/60$			(4.9)
Strategie der Verfeinerung	Strategie 1, $\tau_{\text{adapt}} = 0{,}05$			4.3.2
Zeit-diskretisierungs-Verfahren	(IBS)	(IB), (IBS), (IK), (AB), (ABS), (AK), (DB), (DBS), (DK)	(E), (IB), (IBS), (IK), (AB), (ABS), (AK), (DB), (DBS), (DK)	9.2
Anfangs-zeitschritt	$\tau_0 = 0{,}01$			(5.4)
Parameter der Zeitschritt-steuerung	$\beta = 0{,}8,\ \beta_{\min} = 0{,}5,\ \beta_{\max} = 1{,}1$			(5.8), (5.9)
Toleranz des zeitlich lokalen Fehlers	$TOL_t = 10^{-5}$, 10^{-4}, 10^{-3}, 10^{-2}	$TOL_t = 10^{-5}$, 10^{-4}, 10^{-3}	$TOL_t = 2{\cdot}10^{-5}$, $5 \cdot 10^{-5}$, 10^{-4}, $2 \cdot 10^{-4}$, $5 \cdot 10^{-4}$, 10^{-3}, $2 \cdot 10^{-3}$, $5 \cdot 10^{-3}$, 10^{-2}	(5.8)
Norm des zeitlichen Fehlerschätzers	skalierte Euklidische Norm			Bem. 5.9
Stabilitäts-Radius	$r_{\text{stab}} = 2$			(8.6), Alg. 8.5
Steifigkeits-Sicherheits-parameter	$\alpha = 0{,}8$			Alg. 8.5
Parameter des linearen Lösers	$\alpha_{\text{LSS}} = 0{,}1$	$\alpha_{\text{LSS}} = 0{,}5$, 1, 2, 5, 10, 20, 30	$\alpha_{\text{LSS}} = 2$ (ohne SSOR) bzw. 0,5 (mit SSOR)	(6.16), (7.12)
max. Anzahl der Iterationen	$\max_{\text{It}} = \infty$	$\max_{\text{It}} = 8$ (ohne SSOR) bzw. 5 (mit SSOR), $\varkappa_1 = 10$, $\varkappa_{12} = \varkappa_{23} = 2$		Alg. 6.6, 7.1

SSOR-Parameter	$\omega = 1{,}3$			(6.7)
Anzahl der Testrechnungen	4 verschiedene TOL_t × 3 verschiedene $M = 12$ Testrechnungen	9 Verfahren × 3 verschiedene TOL_t × 7 verschiedene $\alpha_{\text{LSS}} = 189$ Testrechnungen	10 Verfahren × 9 verschiedene $TOL_t = 90$ Testrechnungen	

Tabelle 9.3: Untersuchungen zum Problem TANH

Untersuchung 9.5 (TANH: Einfluß von Gitterfeinheit und Toleranz TOL_t). Wir vergleichen die Rechnungen auf drei Gittern unterschiedlicher Feinheit, die wir durch Variation des Parameters M erhalten, der die maximale Gitterfeinheit angibt. Die minimalen Maschenweiten der Gitter betragen näherungsweise $h_{\min,1} = 2^{-3}$, $h_{\min,2} = 2^{-4}$ und $h_{\min,3} = 2^{-5}$. Abbildung 9.5 zeigt links die Abhängigkeit des L^2-Fehlers der Lösung zum Zeitpunkt $t = 1$ vom mittleren Zeitschritt τ. In der rechten Graphik ist die Rechenzeit über dem L^2-Fehler aufgetragen.

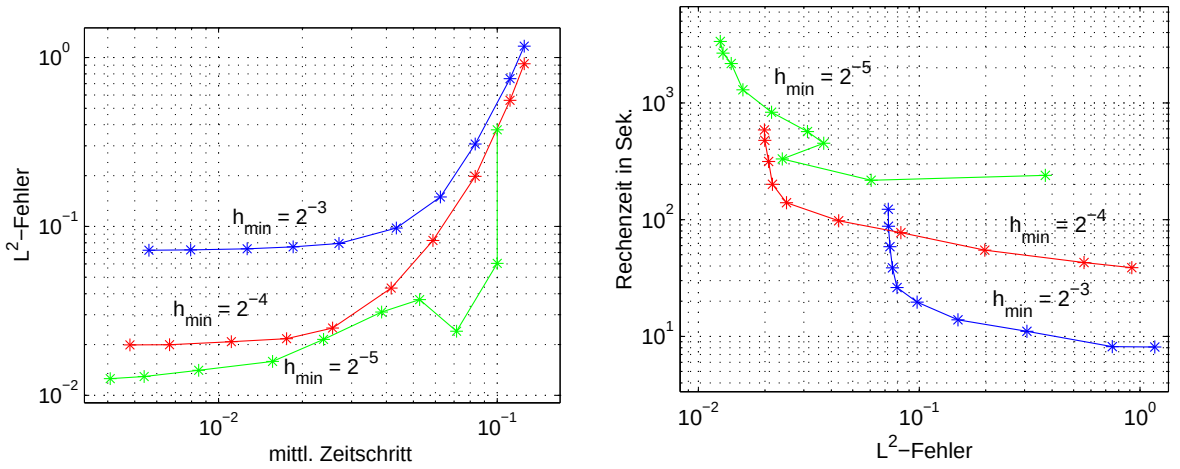


Abbildung 9.5: **links:** L^2-Fehler in Abhängigkeit vom mittleren Zeitschritt τ, **rechts:** L^2-Fehler-Rechenzeit-Diagramm; verschiedene Gitter (blau, rot, grün), verschiedene Toleranzen zwischen $TOL_t = 10^{-5}$ und $TOL_t = 10^{-2}$

Die Bestimmung von Konvergenzraten ist schwierig, da sich mehrere Fehler überlagern: Fehler der Orts- und der Zeitdiskretisierung, des linearen Lösers, der Interpolation nach Verfeinerungen. Man erkennt jedoch, daß sowohl die Verwendung kleiner Zeitschritte auf einem groben Gitter als auch umgekehrt große Zeitschritte auf einem feinen Gitter nicht effizient sind, da hier entweder der Fehler der Ortsdiskretisierung oder der der Zeitdiskretisierung dominiert. Die beiden gröberen Gitter zeigen ein regelmäßiges Fehlerverhalten. □

Bei den folgenden numerischen Untersuchungen verwenden wir das mittlere Gitter, d.h. $M = 5$.

Untersuchung 9.6 (TANH: Der Parameter α_{LSS}). Wenn wir die Effizienz der verschiedenen Verfahren miteinander vergleichen wollen, so stehen wir zunächst vor der Frage, welchen Wert der Parameter α_{LSS} annehmen sollte, damit ein Verfahren seine größtmögliche Effizienz erreicht. Daher beobachten wir in einer zweiten Untersuchungsreihe den Einfluß von α_{LSS} auf Fehler und Aufwand jedes nicht-expliziten Vefahrens. Sämtliche in dieser Untersuchung verwendeten Verfahren und Parameter gehen aus Tabelle 9.3 hervor. In allen Testrechnungen wurde der Rechenaufwand und der mittlere L^2-Fehler $ERR^{\mathrm{T}}_{L^2}$, siehe Definition 9.1, ausgewertet.

Die umfangreichen Ergebnisse dieser Studie können hier nicht vollständig präsentiert werden. Wir beschränken uns auf die Darstellung der Effizienz für ein Verfahren, an dem einige typische Phänomene sichtbar werden. Abbildung 9.6 zeigt das Fehler-Aufwand-Diagramm für das Verfahren (IB) mit den zwei Toleranzen $TOL_t = 10^{-4}$ und 10^{-3}.

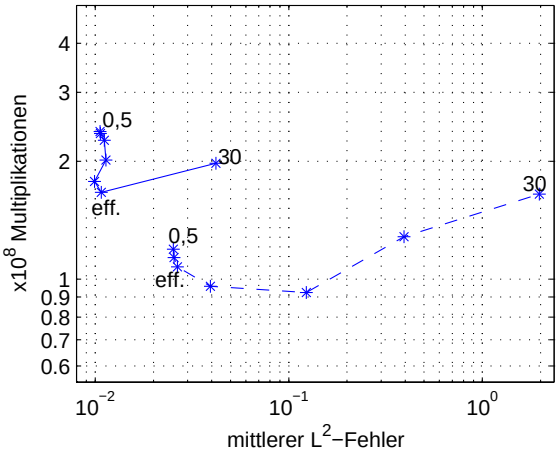


Abbildung 9.6: Effizienz des Verfahrens (IB) für die Toleranzen $TOL_t = 10^{-4}$ (–) und 10^{-3} (- -). Die Zahlen geben die Werte von α_{LSS} an.

Man erkennt, daß der Wert von α_{LSS} die Effizienz des Verfahrens entscheidend beeinflußt. Der günstigste Wert beträgt $\alpha_{\mathrm{LSS}} = 20$ für $TOL_t = 10^{-4}$ und $\alpha_{\mathrm{LSS}} = 2$ für $TOL_t = 10^{-3}$. Der bezüglich der Effizienz optimale Wert für α_{LSS} ist demnach von der Toleranz TOL_t abhängig. Diese Abhängigkeit ist jedoch nicht erwünscht, denn die Größe α_{LSS} soll in der Beziehung

$$TOL_{\mathrm{LSS}} = \frac{\alpha_{\mathrm{LSS}}}{\tau_i} TOL_t$$

siehe (6.16), ja gerade eine von TOL_t *unabhängige* Konstante darstellen. Auch bei den anderen untersuchten Verfahren hängt das optimale α_{LSS} von TOL_t ab, wie die folgende Übersicht zeigt:

Wir wollen uns für jedes Verfahren auf ein *für alle untersuchten Toleranzen günstiges* α_{LSS} festlegen. Dieser Wert ist in der letzten Spalte der Tabelle 9.4 angegeben. Wir werden diesen Wert in der folgenden Untersuchung 9.7 benutzen. □

Untersuchung 9.7 (TANH: Effizienz der Verfahren). In dieser Untersuchung vergleichen wir die Effizienz der zehn betrachteten Verfahren in Abhängigkeit der Toleranz TOL_t. In Abbildung 9.7 ist das Fehler-Aufwand-Diagramm dargestellt.

Verfahren	$TOL_t = 10^{-4}$	$TOL_t = 10^{-3}$	in Untersuchung 9.7 gewählter Wert
(IB), (IK)	$\alpha_{\mathrm{LSS}} = 10 \ldots 20$	$\alpha_{\mathrm{LSS}} = 0{,}5 \ldots 2$	$\alpha_{\mathrm{LSS}} = 2$
(AB), (AK), (DB), (DK)	$\alpha_{\mathrm{LSS}} = 5 \ldots 10$	$\alpha_{\mathrm{LSS}} = 0{,}5 \ldots 2$	$\alpha_{\mathrm{LSS}} = 2$
(IBS), (ABS), (DBS)	$\alpha_{\mathrm{LSS}} = 0{,}5 \ldots 1$	$\alpha_{\mathrm{LSS}} = 0{,}5$	$\alpha_{\mathrm{LSS}} = 0{,}5$

Tabelle 9.4: Problem TANH, Untersuchung 9.6: Bezüglich der Effizienz optimale Werte von α_{LSS} für zwei verschiedene Toleranzen; in Untersuchung 9.7 gewählter Wert von α_{LSS}

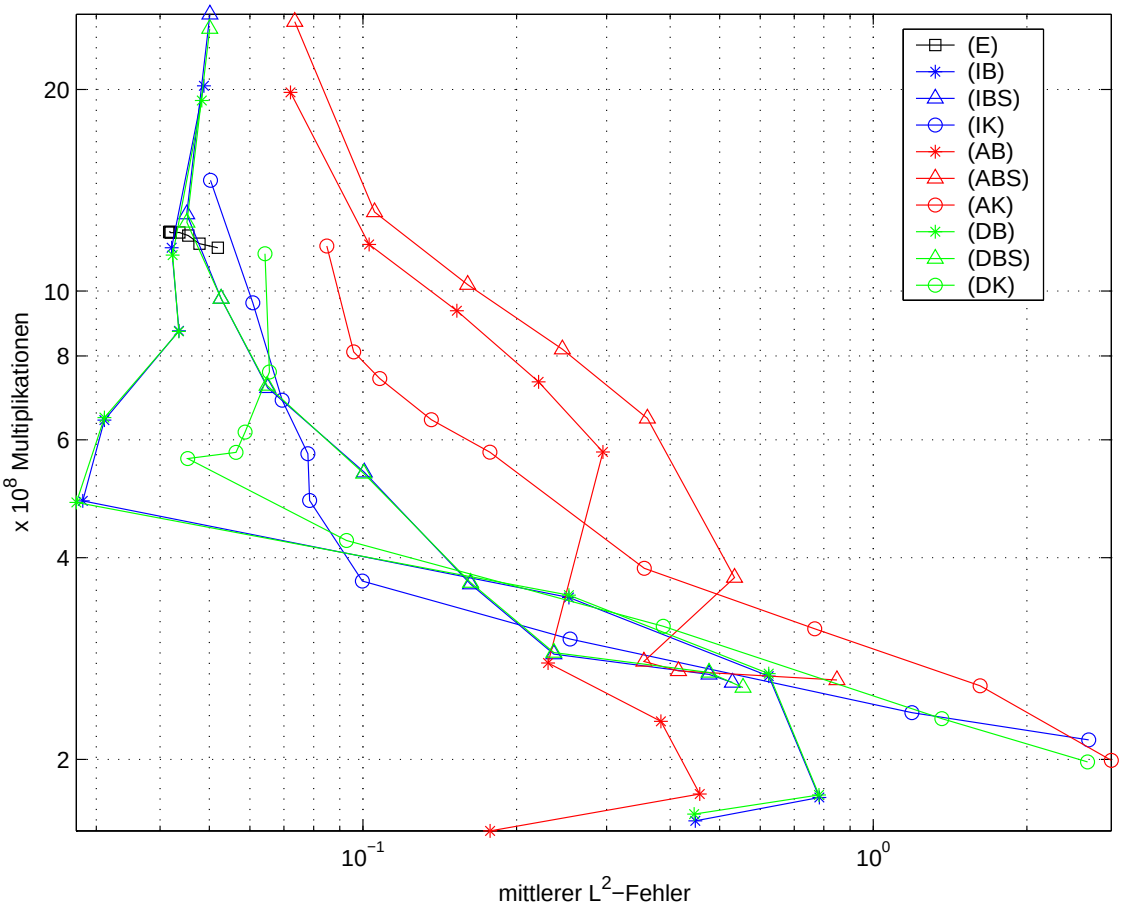


Abbildung 9.7: Fehler-Aufwand-Diagramm. Die Toleranz TOL_t variiert zwischen $2{\cdot}10^{-5}$ (links oben) und 10^{-2} (rechts unten).

Die Abhängigkeit des Fehlers von der Toleranz TOL_t weist hier einige Besonderheiten auf. Während bei den Verfahren (IBS), (DBS), (IK) und (AK) der Fehler wie erwartet mit steigender Toleranz zunimmt, zeigen die Verfahren (IB), (AB), (ABS), (DB) und (DK) ein unregelmäßiges Fehlerverhalten. Diese Verfahren erreichen für relativ grobe Toleranzen mitunter

eine ungewöhnlich hohe Genauigkeit. In Abbildung 9.8 zeigen wir die zeitliche Entwicklung des L^2-Fehlers für die Verfahren (DB) und (AK). Bei dem Verfahren (DB) erkennt man starke Schwankungen des Fehlers bei einigen Toleranzen.

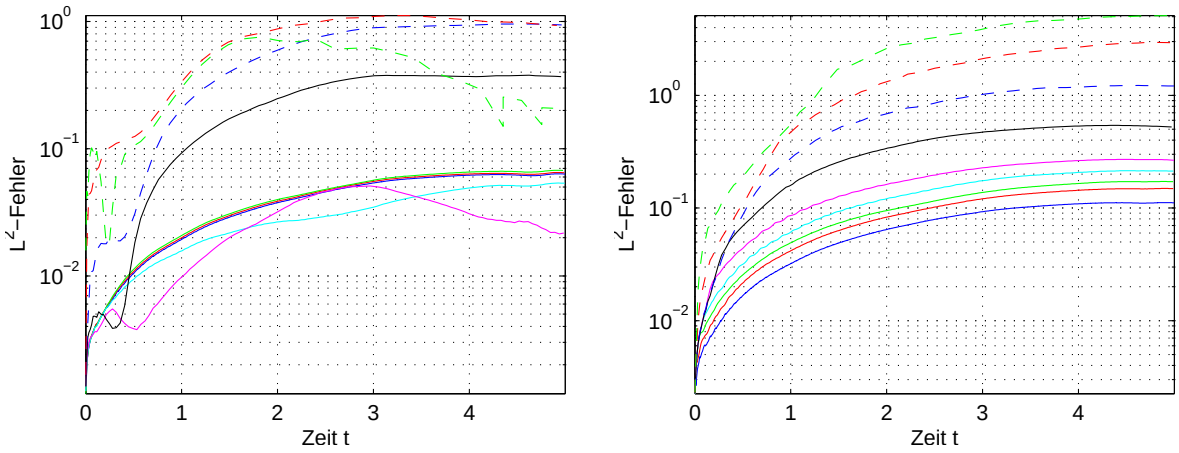


Abbildung 9.8: L^2-Fehler, **links:** Verfahren (DB), **rechts:** Verfahren (AK). Legende in Abbildung 9.9 rechts

Für das Verfahren (DB) stellen wir in Abbildung 9.9 die Zeitschritte und die Anzahl der Gitterknoten über der Zeit t dar. Die Zeitschritte nehmen mit steigender Toleranz TOL_t zu. Die Anzahl der Gitterknoten variiert kaum bei einer Änderung der Toleranz. Diese Untersuchungen zeigen, daß das ungewöhnliche Fehlerverhalten des Verfahrens (DB) trotz einer erwartungsgemäß reagierenden Zeitschrittsteuerung und Gitteradaption zustande kommt. Die Ursachen für die Schwankungen des Fehlers bei diesem und einigen anderen Verfahren konnten nicht geklärt werden. Bei den Problemen BSVD und KRINSKY tritt eine solche Anomalie nicht auf.

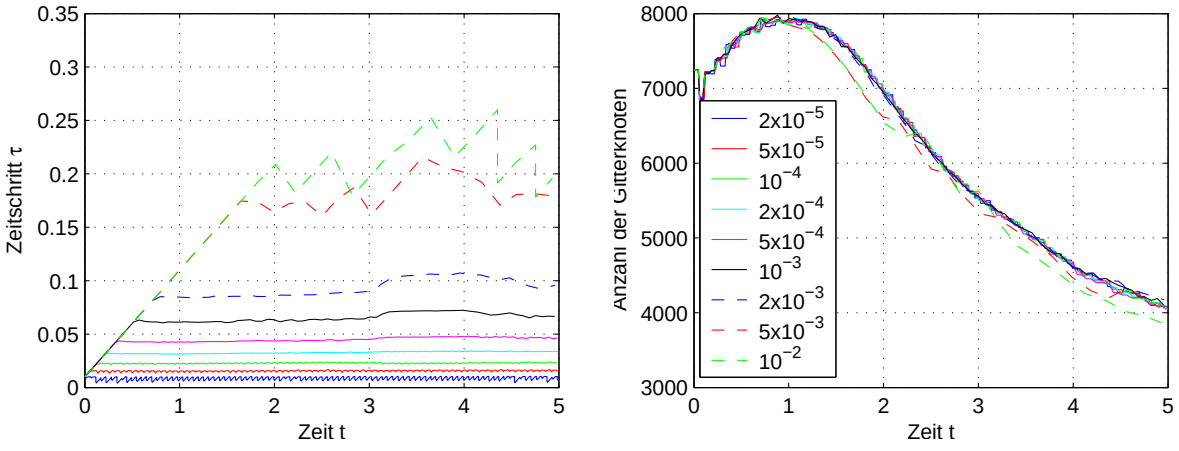


Abbildung 9.9: Verfahren (DB): Zeitschritte und Anzahl der Gitterknoten. Legende gilt für beide Graphiken

Die Varianten (I) und (D) ähneln sich bezüglich ihrer Effizienz, was darauf zurückzuführen ist, daß fast alle Gitterknoten diffusions-steif sind. Die vollständig automatische Partitionierung

(A) ist für geringe Toleranzen oft deutlich ungenauer als die Verfahren (I) und (D). Für uns stellt sich daher die Frage, ob der relativ hohe Fehler bei Verfahren (A) möglicherweise von den Übergangsbereichen zwischen steifen und nichtsteifen Teilgebieten ausgeht. Um das zu untersuchen, stellen wir in Abbildung 9.10 den Fehler der Verfahren (IB), (DB) und (AB) sowie die Partitionierung in Teilgebiete für das Verfahren (AB) graphisch dar.

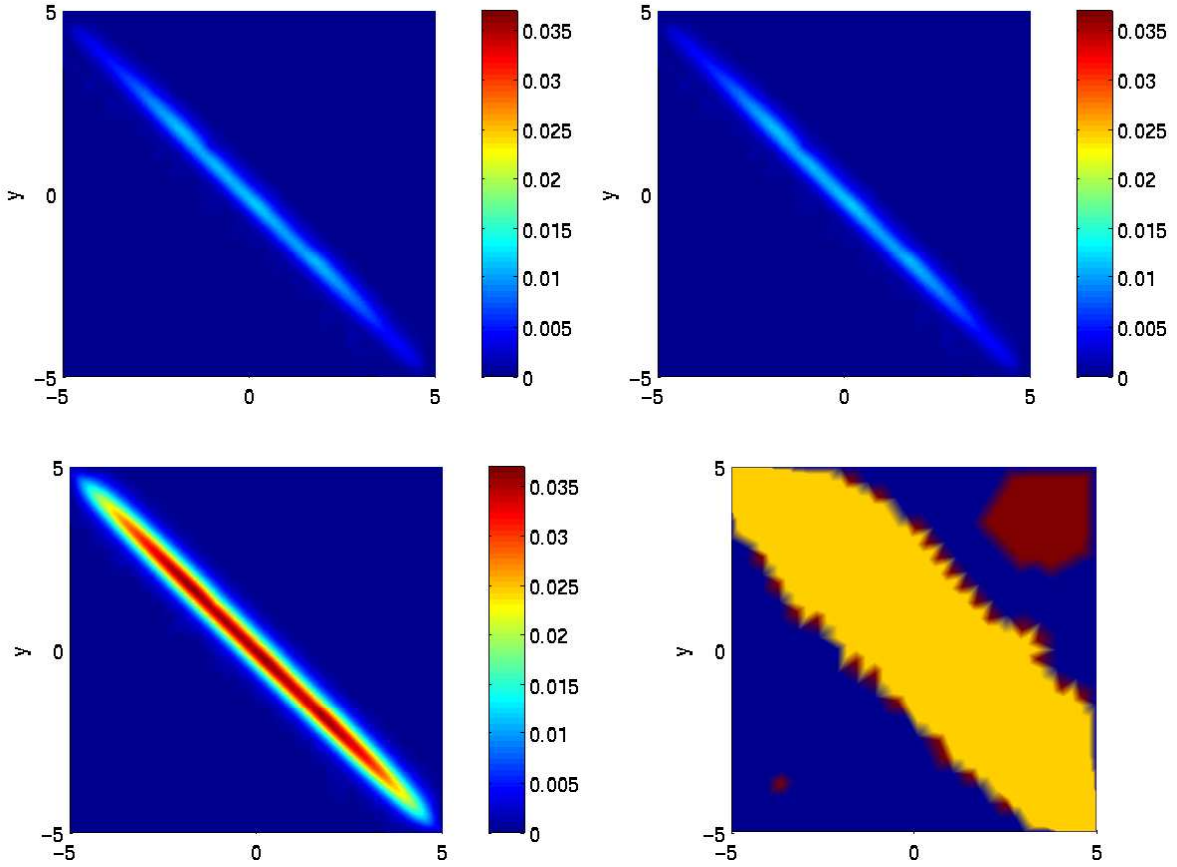


Abbildung 9.10: Fehler zur Zeit $t = 1$: **links oben:** (IB), **rechts oben:** (DB), **links unten:** (AB); **rechts unten:** Partitionierung bei Verfahren (AB): *blau* = Diffusion und Reaktion explizit, *gelb* = Diffusion implizit, Reaktion explizit, *braun* = Diffusion und Reaktion implizit

Der oben erwähnte Verdacht wird durch Abbildung 9.10 jedoch nicht bestätigt. Der Fehler von (AB) tritt im wesentlichen an der Front auf und *nicht* an den Partitionierungsgrenzen. Offenbar ist bei geringer Toleranz die implizite Berechnung des Reaktionsteils genauer als die explizite.

In Abbildung 9.11 ist die durchschnittliche Anzahl der Gitterknoten über der Toleranz TOL_t aufgetragen. Interessant ist, daß die Krylov-W-Verfahren bei groben Toleranzen ein wesentlich stärker verfeinertes Gitter benötigen als die übrigen Verfahren. Vermutlich kommt es hier zu leichten Oszillationen der numerischen Lösung, die zu einer stärkeren Gitterverfeinerung führen.

Die Verfahren mit SSOR-Vorkonditionierung sind in der Regel etwas aufwendiger und unge-

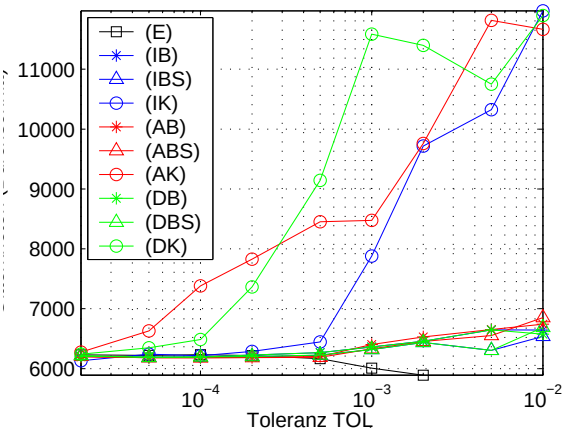


Abbildung 9.11: Durchschnittliche Anzahl der Gitterknoten

nauer als die Verfahren mit BiCGstab-Löser ohne Vorkonditionierung. Die Vorkonditionierung lohnt sich daher bei diesem Problem nicht. Das explizite Verfahren (E) ist wegen der Steifheit des Problems recht aufwendig und kaum durch die Toleranz TOL_t zu beeinflussen, da die Wahl des Zeitschritts hier der Stabilitätsbedingung (5.11) folgt. □

9.6.2 Problem BSVD

Wir dokumentieren in diesem Abschnitt erneut drei numerische Untersuchungen, die mit dem Problem BSVD durchgeführt wurden. Die folgende Tabelle gibt einen Überblick über die verwendeten Verfahren und Parameter.

	Untersuchung 9.8	Untersuchung 9.9	Untersuchung 9.11	siehe
Gebiet	$\Omega =]0,1[^2$			
Zeitintervall	$[t_0, t_e] = [0,8]$	$[t_0, t_e] = [0,7{,}5]$		
Maschenweite des uniformen Grundgitters	$h_0 = 0{,}2$			
Fehlerindikator	Z^2			4.1.1
max. Gitterfeinheit	$M = 5,6$	$M = 6$		4.3
Sprungstellen von zfh	$\mu(0) = \mu(1) = 0$, $\mu(2) = \mu(3) = 1/24$, $\mu(4) = 1/12$, $\mu(5) = 5/12$			(4.9)
Strategie der Verfeinerung	Strategie 3, $\alpha = 0{,}05$, $k_{\max} = 10$			4.3.2
Zeit-diskretisierungs-Verfahren	(IBS)	(IB), (IBS), (IK), (AB), (ABS), (AK), (DB), (DBS), (DK)		9.2

Anfangs-zeitschritt	$\tau_0 = 0{,}01$			(5.4)
Parameter der Zeitschritt-steuerung	$\beta = 0{,}8,\ \beta_{\min} = 0{,}5,\ \beta_{\max} = 1{,}1$			(5.8), (5.9)
Toleranz des zeitlich lokalen Fehlers	$TOL_t = 10^{-5}$	$TOL_t = 10^{-5}$, 10^{-4}, 10^{-3}	$TOL_t = 2{\cdot}10^{-5}$, $5 \cdot 10^{-5}$, 10^{-4}, $2 \cdot 10^{-4}$, $5 \cdot 10^{-4}$, 10^{-3}, $2 \cdot 10^{-3}$, $5 \cdot 10^{-3}$, 10^{-2}	(5.8)
Norm des zeitlichen Fehlerschätzers	skalierte Euklidische Norm			Bem. 5.9
Stabilitäts-Radius	$r_{\text{stab}} = 2$			(8.6), Alg. 8.5
Steifigkeits-Sicherheits-parameter	$\alpha = 0{,}8$			Alg. 8.5
Parameter des linearen Lösers	$\alpha_{\text{LSS}} = 0{,}1$	$\alpha_{\text{LSS}} = 0{,}5$, 1, 2, 5, 10, 20, 30	$\alpha_{\text{LSS}} = 0{,}5$	(6.16), (7.12)
max. Anzahl der Iterationen	$\max_{\text{It}} = \infty$	$\max_{\text{It}} = 8$ (ohne SSOR) bzw. 5 (mit SSOR), $\varkappa_1 = 10$, $\varkappa_{12} = \varkappa_{23} = 2$		Alg. 6.6, 7.1
SSOR-Parameter	$\omega = 1{,}3$			(6.7)

Tabelle 9.5: Untersuchungen zum Problem BSVD

Untersuchung 9.8 (BSVD: Feinheit des Gitters an der Front). Bei dem Problem BSVD hat die Feinheit des Gitters einen sehr starken Einfluß auf die Genauigkeit der numerischen Lösung. Abbildung 9.12 zeigt die numerische Lösung zur Zeit $t = 8$ links auf einem groben Gitter der maximalen Feinheit $M = 5$ und rechts auf einem feineren Gitter der maximalen Feinheit $M = 6$. Das entspricht einer minimalen Maschenweite von etwa 0,006 bzw. 0,003. Ein Vergleich mit einer Referenzlösung auf einem noch feineren Gitter zeigt, daß die Lösung mit $M = 6$ eine recht gute Approximation der exakten Lösung darstellt. Die Lösung auf dem groben Gitter, $M = 5$, ist extrem ungenau und praktisch nicht verwendbar. Bei den folgenden Berechnungen wird daher stets $M = 6$ verwendet. □

Untersuchung 9.9 (BSVD: Der Parameter α_{LSS}). Wie beim Problem TANH untersuchen wir auch für das BSVD-Problem, wie sich die Wahl des Parameters α_{LSS} auf die Effizienz der Verfahren auswirkt. Sämtliche untersuchten Verfahren und Paramater sind in Tabelle 9.5 dargestellt. Der Übersichtlichkeit halber zeigen wir in Abbildung 9.13 nur die Ergebnisse für die Verfahren (IBS), (ABS) und (DBS) bei Verwendung der Toleranzen $TOL_t = 10^{-4}$ und 10^{-3}.

Große Werte von α_{LSS} sind in allen Fällen sehr ineffizient. Für jedes Verfahren soll ein be-

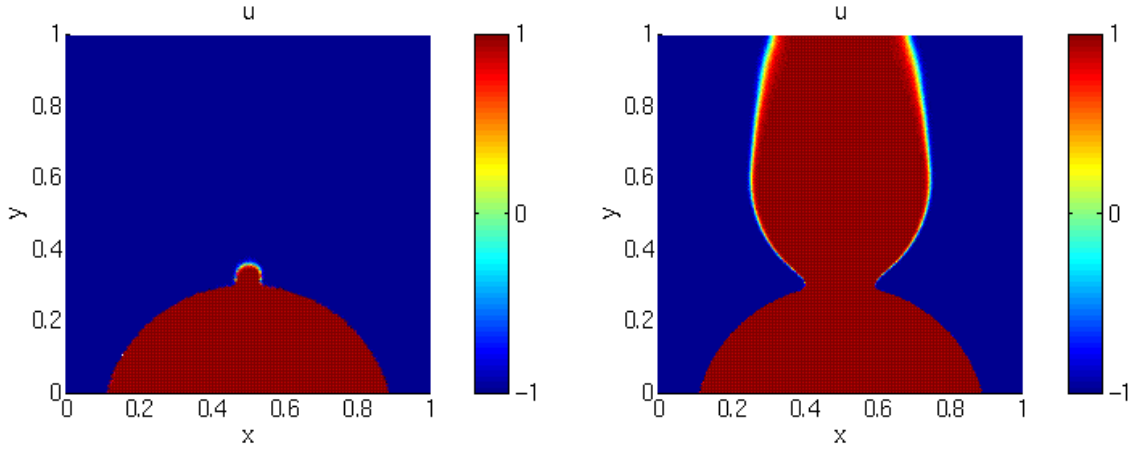


Abbildung 9.12: Numerische Lösung des BSVD-Problems zur Zeit $t = 8$, maximale Gitterfeinheit: **links:** $M = 5$, **rechts:** $M = 6$

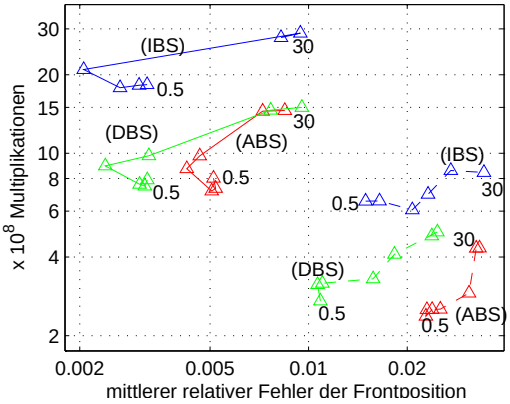


Abbildung 9.13: Einfluß von α_{LSS} auf die Effizienz der Verfahren (IBS), (ABS) und (DBS) mit den Toleranzen $TOL_t = 10^{-4}$ (–) und 10^{-3} (- -). Die Zahlen in der Graphik sind die Werte von α_{LSS}.

züglich der Effizienz optimales α_{LSS} bestimmt werden. Für die Wahl dieses optimalen Wertes gelten ähnliche Überlegungen wie im Falle des Problems TANH, siehe Untersuchung 9.6. Nach Auswertung der numerischen Untersuchungen halten wir *für alle Verfahren* den Wert $\alpha_{\mathrm{LSS}} = 0{,}5$ für eine günstige Wahl. □

Zur Steifheit des Problems BSVD

Für den Rechenaufwand der Partitionierungs-Verfahren (A) und (D) ist die Anzahl der *bezüglich der Diffusion* steifen Gitterknoten von entscheidender Bedeutung. Die Steifheit des Reaktionsterms wirkt sich hingegen nicht so stark auf den Rechenaufwand aus, vgl. dazu die Bemerkung zur Diffusionspartitionierung in Abschnitt 8.2. Das BSVD-Problem ist im wesentlichen dort diffusions-steif, wo der Diffusions-Koeffizient hinreichend groß ist. Die Front bewegt sich im Laufe der Zeit in dieses Gebiet hinein. Dadurch nimmt auch die Anzahl der

diffusions-steifen Gitterknoten und damit die Dimension des linearen Gleichungssystems im Laufe der Rechnung stark zu. Aus diesem Grunde ist BSVD ein Problem, bei dem man eine deutliche Effizienzsteigerung durch lokale Partitionierung erwarten kann.

Untersuchung 9.10 (BSVD: Steifheit). Wir berechnen das Problem BSVD mit dem Verfahren (DB) und der Toleranz $TOL_t = 5 \cdot 10^{-5}$. Die übrigen Parameter entsprechen den in Tabelle 9.5 für Untersuchung 9.11 angegebenen. In Abbildung 9.14 sind Anzahl der Gitterknoten und Dimension des linearen Gleichungssystems dargestellt.

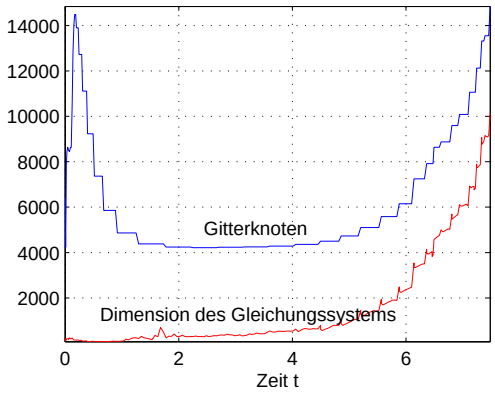


Abbildung 9.14: Verfahren (DB), Toleranz $TOL_t = 5 \cdot 10^{-5}$: Anzahl der Gitterknoten und Dimension des linearen Gleichungssystems.

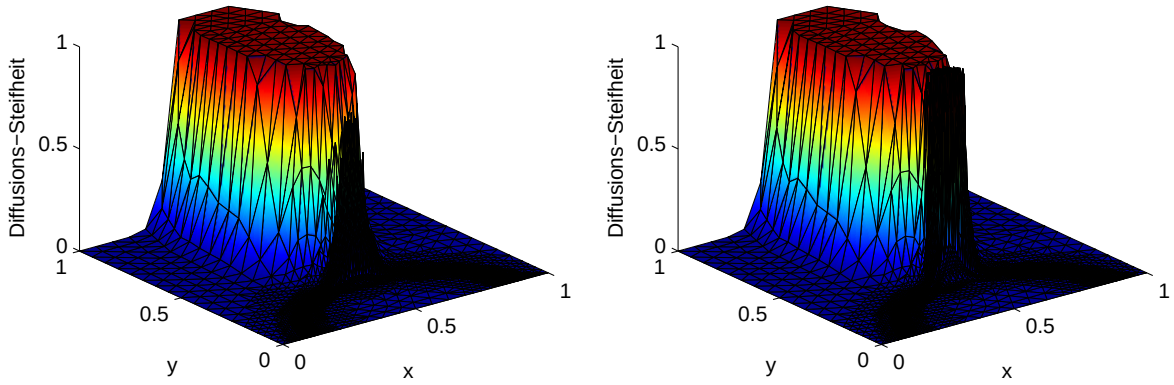


Abbildung 9.15: Verfahren (DB), $TOL_t = 5 \cdot 10^{-5}$. Diffusions-Steifheit zu den Zeitpunkten $t = 2$ und 4

Das Problem wird somit anfangs nahezu vollständig explizit gelöst, während gegen Ende des Zeitintervalls $[0,7{,}5]$ ein großer Teil implizit gelöst wird. In den Abbildungen 9.15 und 9.16 ist die Diffusions-Steifheit $\sigma_{k,i}^{\text{diff}}$ graphisch dargestellt[6]. Man erkennt, wie die Anzahl der Gitterknoten in dem diffusions-steifen Gebiet zunimmt. □

[6]Die Diffusions-Steifheit wurde in Definition 8.6 eingeführt. Zur besseren Übersichtlichkeit wurde die Diffusions-Steifheit „nach oben abgeschnitten“, d.h. in der Graphik wird die Größe $\min\{\sigma_{k,i}^{\text{diff}}, 1\}$ dargestellt. Ein Gitterknoten ist diffusions-steif, wenn $\sigma_{k,i}^{\text{diff}} > \alpha = 0{,}8$ ist.

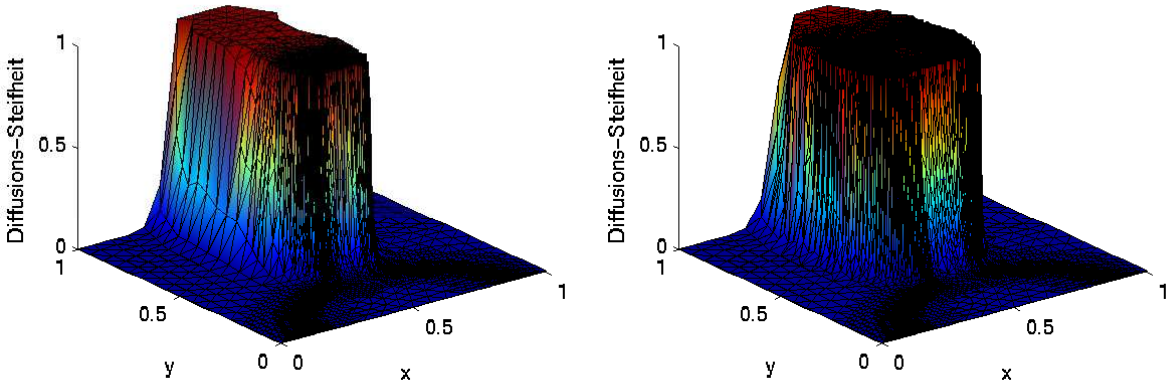


Abbildung 9.16: Verfahren (DB), $TOL_t = 5 \cdot 10^{-5}$. Diffusions-Steifheit zu den Zeitpunkten $t = 6$ und 7

Untersuchung 9.11 (BSVD: Effizienz der Verfahren). Wir vergleichen die neun in Tabelle 9.5 angegebenen Verfahren bezüglich ihrer Effizienz. Dabei setzen wir $\alpha_{\mathrm{LSS}} = 0{,}5$ und variieren die Toleranz TOL_t zwischen $2 \cdot 10^{-5}$ und 10^{-2}. Das Fehler-Aufwand-Diagramm ist in Abbildung 9.17 dargestellt.

Da die Anzahl der diffusions-steifen Gitterknoten hier deutlich geringer als die Anzahl aller Gitterknoten ist[7], sind Partitionierungs-Verfahren deutlich im Vorteil. Für $TOL_t \leq 10^{-3}$ ist die Diffusions-Partitionierung in allen Fällen der vollständig automatischen Partitionierung überlegen, da sie genauer ist – ein Effekt, der bereits bei dem Problem TANH beobachtet wurde. Das Verfahren (DB) ist für einen weiten Bereich mittlerer Genauigkeitsanforderung der Favorit. Für sehr hohe Genauigkeiten ist (IK) geringfügig besser, während für geringe Genauigkeitsforderung die Verfahren (AB) und (ABS) etwas effizienter sind. Die SSOR-Vorkonditionierung bringt nur für grobe Toleranzen eine Einsparung an Rechenaufwand.

Es sei N die Anzahl der in der Rechnung benötigten Zeitschritte. In Abbildung 9.18 stellen wir für die untersuchten Verfahren jeweils

- den mittleren Zeitschritt $\tau_{\mathrm{av}} = t_e/N$,
- den mittleren relativen Fehler der Frontposition $ERR^{\mathrm{B}}_{\mathrm{fr}}$,
- die durchschnittliche Anzahl der Iterationen und
- die durchschnittliche Anzahl der Gitterknoten

über der Toleranz TOL_t dar. Besonders auffällig ist hier die sehr große Anzahl von Gitterknoten, die das Verfahren (IK) für grobe Toleranzen TOL_t benötigt. Ein ähnlicher Effekt wurde bei dem Problem TANH bereits beobachtet. Räumliche Oszillationen in der numerischen Lösung sind hier wahrscheinlich die Ursache für eine starke Gitterverfeinerung. □

[7] Siehe etwa Abbildung 9.14.

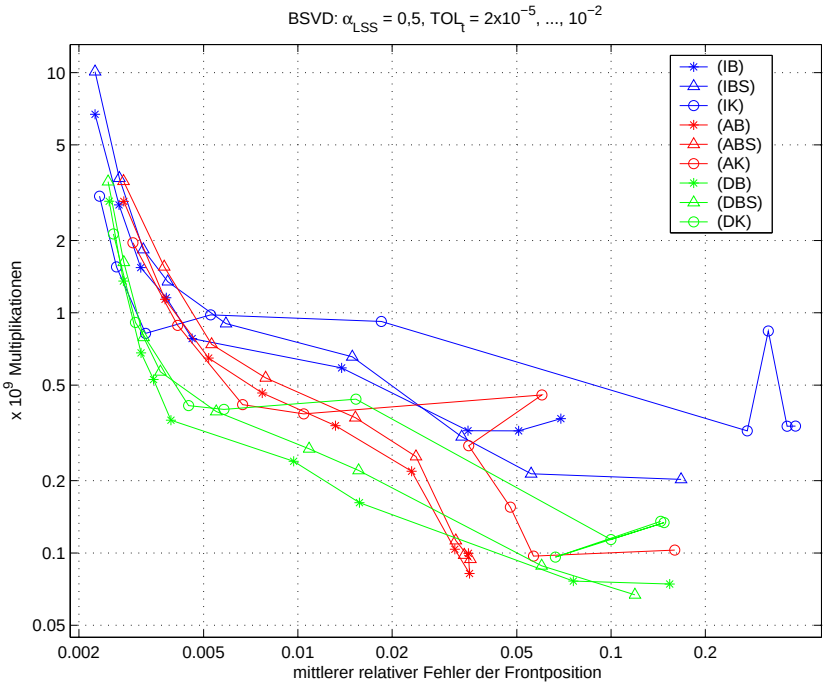


Abbildung 9.17: Fehler-Aufwand-Diagramm. Die Toleranz TOL_t variiert auf jeder Kurve zwischen $2 \cdot 10^{-5}$ (links oben) und 10^{-2} (rechts unten).

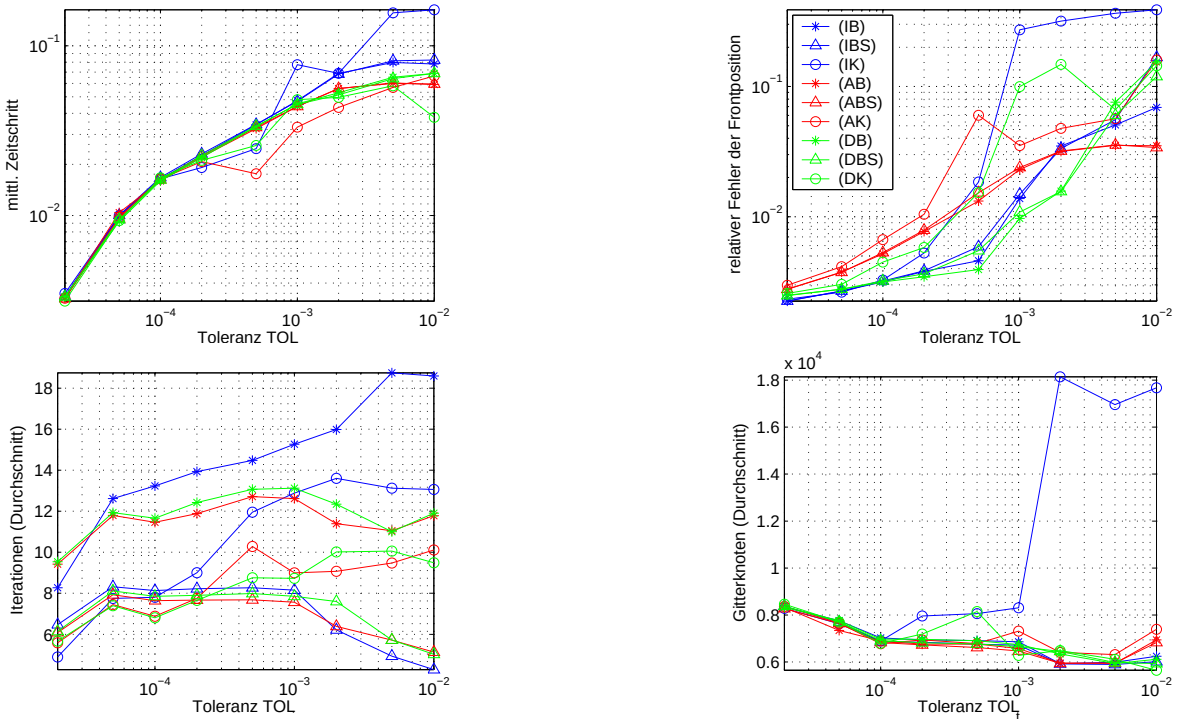


Abbildung 9.18: Mittlerer Zeitschritt, mittlerer relativer Fehler der Frontposition, durchschnittliche Anzahl der Iterationen und Gitterknoten

9.6.3 Problem KRINSKY

In diesem Abschnitt stellen wir die Ergebnisse dreier numerischer Untersuchungen dar, die für das Problem Krinsky durchgeführt wurden. Die folgende Tabelle gibt einen Überblick über die verwendeten Verfahren und Parameter.

	Untersuchung 9.12	Untersuchung 9.13	Untersuchung 9.15	siehe
Gebiet	$\Omega =]-20,20[^2$			
Zeitintervall	$[t_0, t_e] = [0, 10]$			
Maschenweite des uniformen Grundgitters	$h_0 = 2$			
Fehlerindikator	Z^2			4.1.1
max. Gitterfeinheit	$M = 2, 3, 4$	$M = 4$		4.3
Sprungstellen von zfh	$\mu(0) = \mu(1) = \mu(2) = \mu(3) = 1/240$			(4.9)
Strategie der Verfeinerung	Strategie 2, $k_{\text{adapt}} = 10$			4.3.2
Zeitdiskretisierungs-Verfahren	(IEBS)	(EE), (IEB), (IEBS), (IEK), (AEB), (AEBS), (AEK), (DEB), (DEBS), (DEK), (IIB), (IIBS), (IIK)		9.2
Anfangszeitschritt	$\tau_0 = 0{,}015$			(5.4)
Parameter der Zeitschrittsteuerung	$\beta = 0{,}8,\ \beta_{\min} = 0{,}5,\ \beta_{\max} = 1{,}1$			(5.8), (5.9)
Toleranz des zeitlich lokalen Fehlers	$TOL_t = 10^{-5}$, $2\cdot10^{-5}$, $5\cdot10^{-5}$, 10^{-4}, $2\cdot10^{-4}$, $5\cdot10^{-4}$, 10^{-3}, $2\cdot10^{-3}$, $5\cdot10^{-3}$, 10^{-2}	$TOL_t = 10^{-5}$, 10^{-4}, 10^{-3}	$TOL_t = 2\cdot10^{-5}$, $5\cdot10^{-5}$, 10^{-4}, $2\cdot10^{-4}$, $5\cdot10^{-4}$, 10^{-3}, $2\cdot10^{-3}$, $5\cdot10^{-3}$, 10^{-2}	(5.8)
Norm des zeitlichen Fehlerschätzers	skalierte Euklidische Norm			Bem. 5.9
Stabilitäts-Radius	$r_{\text{stab}} = 2$			(8.6), Alg. 8.5
Steifigkeits-Sicherheitsparameter	$\alpha = 0{,}8$			Alg. 8.5

Parameter des linearen Lösers	$\alpha_{\mathrm{LSS}} = 1$	$\alpha_{\mathrm{LSS}} = 0{,}5,\ 1,\ 2,\ 5,\ 10,\ 20,\ 30$	$\alpha_{\mathrm{LSS}} = 0{,}5$ (bei (IEK), (AEK), (DEK)) bzw. 1 (sonst)	(6.16), (7.12)
max. Anzahl der Iterationen	$\max_{\mathrm{It}} = 10$	$\max_{\mathrm{It}} = 8$ (ohne SSOR) bzw. 5 (mit SSOR), $\varkappa_1 = 10$, $\varkappa_{12} = \varkappa_{23} = 2$		Alg. 6.6, 7.1
SSOR-Parameter	$\omega = 1{,}3$			(6.7)

Tabelle 9.6: Untersuchungen zum Problem KRINSKY

Untersuchung 9.12 (KRINSKY: Einfluß von Gitterfeinheit und Toleranz TOL_t). Wir untersuchen zunächst am Beispiel des Verfahrens (IEBS), wie sich Gitterfeinheit und Toleranz TOL_t auf die Genauigkeit der Lösung auswirken. Die verwendeten Parameter können Tabelle 9.6 entnommen werden. Das Ergebnis ist in Abbildung 9.20 dargestellt. Gitterverfeinerung und Senkung der Toleranz TOL_t führen erwartungsgemäß zu einer Erhöhung der Genauigkeit. Eine Verringerung der Toleranz unter den Wert $2 \cdot 10^{-4}$ bringt jedoch kaum noch eine Verbesserung, da in diesem Fall der Fehler der Ortsdiskretisierung überwiegt. □

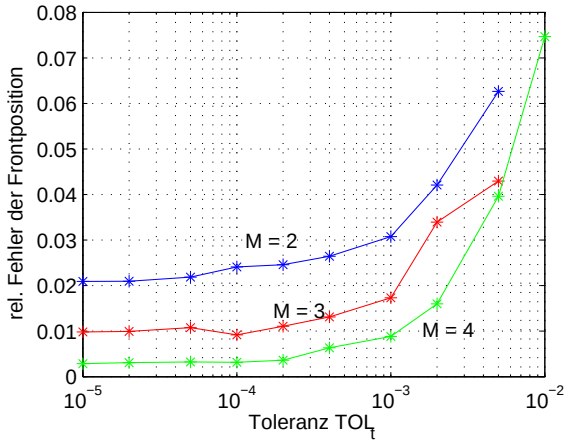


Abbildung 9.19: Verfahren (IEBS), Einfluß von Gitterfeinheit und Toleranz TOL_t auf den Fehler $ERR_{\mathrm{fr}}^{\mathrm{K}}$

Untersuchung 9.13 (KRINSKY: Der Parameter α_{LSS}). Zur Bestimmung eines bezüglich der Effizienz optimalen Parameters α_{LSS} werden die zehn in Tabelle 9.6 angegebenen Verfahren mit den drei Toleranzen $TOL_t = 10^{-5},\ 10^{-4}$ und 10^{-3} getestet. Der besseren Übersicht wegen stellen wir hier nur die Ergebnisse für das Verfahren (IEK) dar, siehe Abbildung 9.20.

Die Auswahl eines optimalen α_{LSS} ist bei diesem Problem komplizierter als bei den Problemen TANH und BSVD, da ein Verfahren oft für eine weite Bandbreite von Parametern α_{LSS} effizient ist. Beispielsweise zeigt Abbildung 9.20, daß sich das Verfahren (IEK) bei $TOL_t = 10^{-4}$ für $\alpha_{\mathrm{LSS}} \in [0{,}5,\ 20]$ im effizienten Bereich befindet. Erst für $\alpha_{\mathrm{LSS}} = 30$ ist das Verfahren

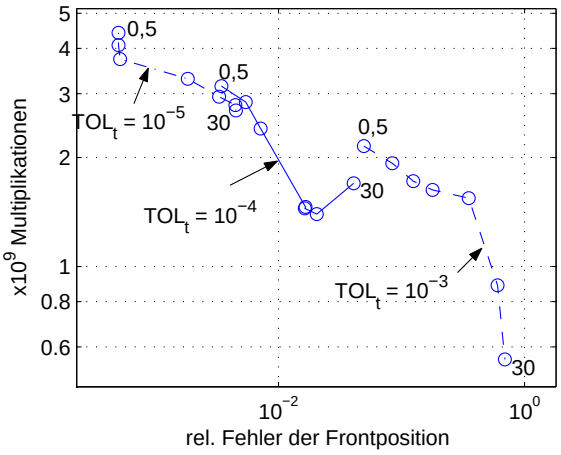


Abbildung 9.20: Effizienz in Abhängigkeit von α_{LSS} beim Verfahren (IEK). Der Parameter α_{LSS} variiert zwischen 0,5 und 30, siehe Tabelle 9.6. Die Zahlen in der Graphik geben den Wert von α_{LSS} an.

eindeutig weniger effizient als für $\alpha_{\text{LSS}} = 20$. Wir haben uns hier dafür entschieden, die Genauigkeit des Verfahrens nur durch die Toleranz TOL_t und nicht durch den Parameter α_{LSS}, d.h. durch die Qualität des linearen Lösers, steuern zu lassen. Daher bevorzugen wir von den Werten von α_{LSS} im effizienten Bereich jeweils den mit dem kleinsten Fehler. Für das Verfahren (IEK) bedeutet das beispielsweise das folgende, vgl. Abbildung 9.20:

Bei $TOL_t = 10^{-5}$ sind $\alpha_{\text{LSS}} \in [2,\ 30]$ effizient; wir wählen $\alpha_{\text{LSS}} = 2$. Bei $TOL_t = 10^{-4}$ sind $\alpha_{\text{LSS}} \in [0{,}5,\ 20]$ effizient; wir wählen hier $\alpha_{\text{LSS}} = 0{,}5$. Bei $TOL_t = 10^{-3}$ sind $\alpha_{\text{LSS}} \in [0{,}5,\ 30]$ effizient; daher wählen wir hier $\alpha_{\text{LSS}} = 0{,}5$. Wir suchen jedoch einen *gemeinsamen* Wert für alle Toleranzen. Bei Nichtübereinstimmung nehmen wir das Minimum, also hier den Wert $\alpha_{\text{LSS}} = 0{,}5$. Auch für $TOL_t = 10^{-5}$ ist dieser Wert durchaus noch akzeptabel, wie Abbildung 9.20 zeigt.

Nach Auswertung der numerischen Ergebnisse für alle untersuchten Verfahren erscheinen uns die folgenden Werte für α_{LSS} optimal:

$$\alpha_{\text{LSS}} = \begin{cases} 0{,}5 & \text{bei (IEK), (AEK), (DEK),} \\ 1 & \text{bei den übrigen Verfahren.} \end{cases}$$

Diese Werte werden in den später dargestellten Untersuchungen 9.14 und 9.15 verwendet. □

Untersuchung 9.14 (KRINSKY: Steifheit). Verwenden wir in dem Problem KRINSKY die maximale Gitterfeinheit $M = 4$ und eine moderate Toleranz, so ist die u-Komponente des Systems stark diffusions-steif und leicht reaktions-steif. Bei der v-Komponente liegt keine Steifheit vor. Wir verdeutlichen die lokal auftretende Steifheit der u-Komponente am Beispiel einer numerischen Berechnung mit dem Verfahren (AEBS) und der Toleranz $TOL_t = 10^{-4}$. Alle übrigen Parameter werden wie in Untersuchung 9.15 gewählt, siehe Tabelle 9.6. Abbildung 9.21 zeigt rechts die Verteilung der Diffusions- und Reaktions-Steifheit[8] über das Gebiet

[8] Die Größen Diffusions-Steifheit $\sigma_{k,i}^{\text{diff}}$ und Reaktions-Steifheit $\sigma_{k,i}^{\text{reac}}$ wurden in Definition 8.6 eingeführt. Zur besseren Übersichtlichkeit wurde die Diffusions-Steifheit „nach oben abgeschnitten“, d.h. in der Graphik wird die Größe $\min\{\sigma_{k,i}^{\text{diff}}, 2\}$ dargestellt. Ein Gitterknoten gilt als diffusions- bzw. reaktions-steif, wenn $\sigma_{k,i}^{\text{diff}}$ bzw. $\sigma_{k,i}^{\text{reac}}$ größer als 0,8 ist.

Ω zur Zeit $t = 9$.

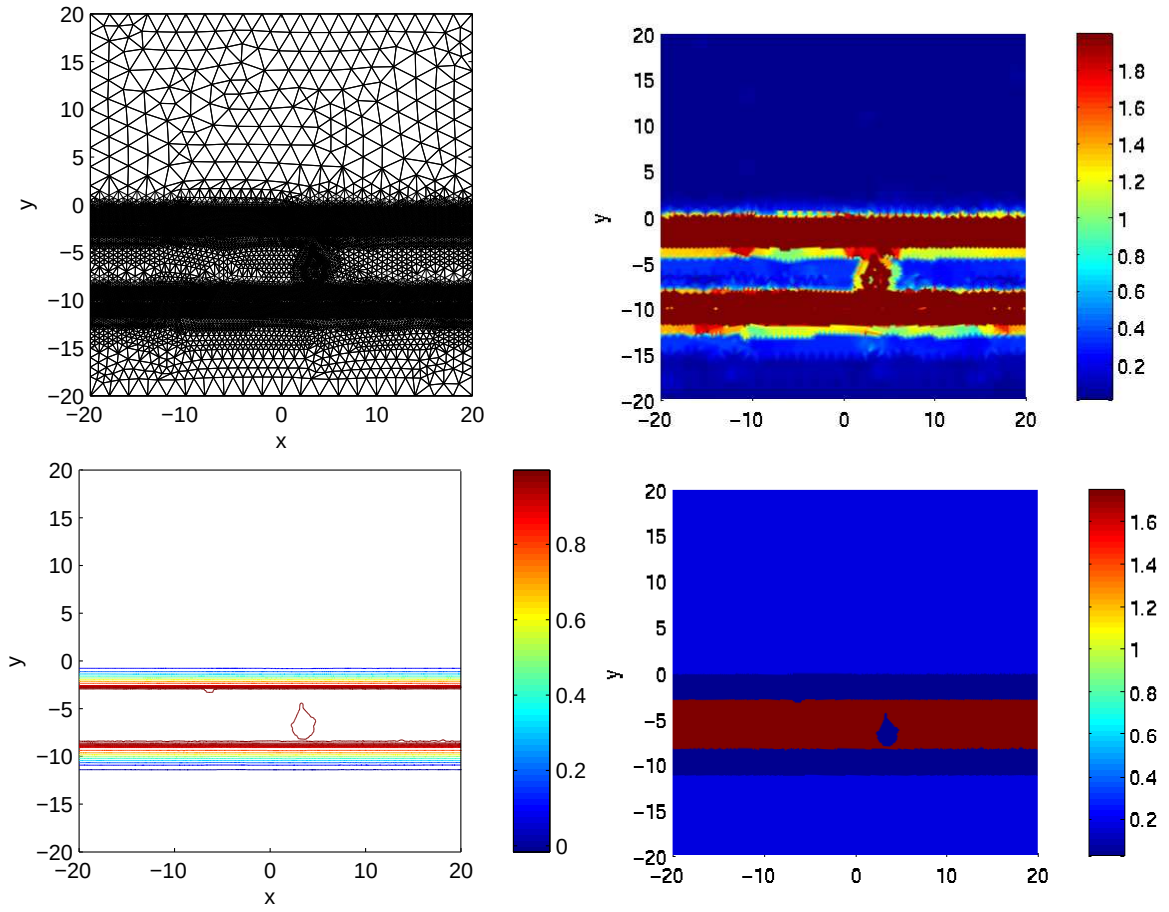


Abbildung 9.21: Verfahren (AEBS), $t = 9$; **links oben:** Gitter, **rechts oben:** Diffusions-Steifheit der u-Komponente, **links unten:** u-Komponente (Höhenlinien-Darstellung), **rechts unten:** Reaktions-Steifheit der u-Komponente

Die Diffusions-Steifheit der u-Komponente wird durch starke Gitterverfeinerung hervorgerufen. Reaktions-Steifheit liegt dort vor, wo $u \approx 1$ ist. Die Unregelmäßigkeit in der Steifheit rechts der Mitte geht auf numerische Fehler zurück. Wegen des unstrukturierten – und damit auch unsymmetrischen – Gitters treten diese Fehler nicht symmetrisch zur y-Achse auf. Bereits ein geringfügiges Absinken der u-Komponente unter den Maximalwert 1 läßt den Reaktions-Term hier seine Steifheit verlieren!

Der weitaus größte Teil der Gitterknoten – hier sind es 94% – liegt im diffusions-steifen Gebiet. Man erwartet daher, daß die Partitionierungs-Verfahren (D) und (A) nur geringe Einsparungen gegenüber dem impliziten Verfahren (I) erzielen.

In Abbildung 9.22 zeigen wir am Beispiel des Verfahrens (DEB)[9] zur Zeit $t = 10$, wie der Anteil der bezüglich der u-Komponente diffusions-steifen Gitterknoten von der Toleranz TOL_t abhängt. Aus der Graphik geht hervor, daß der Anteil steifer Gitterknoten bei sinkender Toleranz leicht abnimmt. In der Tat zeigen die im folgenden dargestellten Ergebnisse der

[9]Es wurden die gleichen Parameter wie in Untersuchung 9.15 verwendet, siehe Tabelle 9.6.

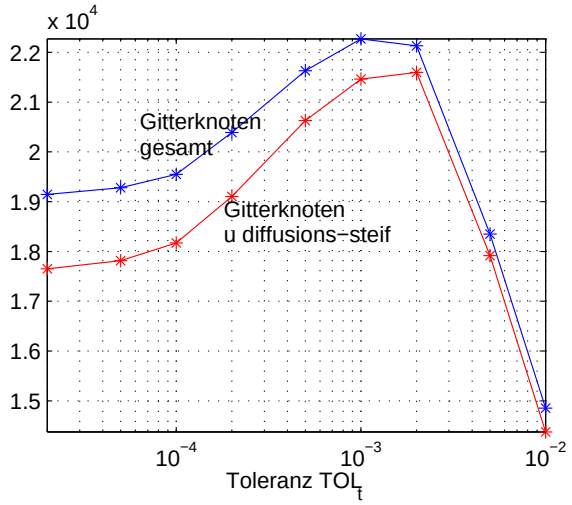


Abbildung 9.22: Verfahren (DEB): Anzahl der Gitterpunkte und Anzahl der bezüglich u diffusions-steifen Gitterpunkte in Abhängigkeit von der Toleranz TOL_t

Untersuchung 9.15, daß für kleine Toleranzen TOL_t die Diffusions-Partitionierung – also das Verfahren (D) – gegenüber dem impliziten Verfahren (I) im Vorteil ist. □

Untersuchung 9.15 (KRINSKY: Effizienz der Verfahren). In Abbildung 9.23 ist das Fehler-Aufwand-Diagramm für die untersuchten Verfahren dargestellt. Das Diagramm läßt die folgenden Schlüsse zu.

- Die Verfahren (IIB), (IIBS) und (IIK) sind nicht effizient, da ihr Aufwand zu hoch ist. Der Grund dafür ist, daß die v-Komponente des Systems nicht steif ist, bei den genannten Verfahren aber implizit gerechnet wird.
- Das explizite Verfahren (EE) ist nicht effizient, da die u-Komponente des Systems steif ist. Die Zeitschritte des Verfahrens sind somit sehr klein.
- Für grobe Toleranzen verbessert sich durch die SSOR-Vorkonditionierung die Effizienz
- Die Krylov-W-Verfahren (IEK), (AEK) und (DEK) sind für eine hohe Genauigkeitsforderung effizient, für eine niedrige jedoch nicht.
- Die Favoriten zur Lösung des Problems sind
  - bei niedriger Toleranz die Verfahren (IEK), (AEK) und (DEK),
  - bei großer Toleranz das Verfahren (IEBS).

Bemerkung 9.16. Wie in Abschnitt 9.4 erläutert, wird der Fehler $ERR_{\text{fr}}^{\text{K}}$ durch Vergleich mit einer Referenzlösung ermittelt. In Untersuchung 9.15 stellt sich heraus, daß dieser Fehler für einige Verfahren *sehr kleine Werte* annimmt. So ergibt sich beispielsweise für das Verfahren (IIB) mit Toleranz $TOL_t = 5 \cdot 10^{-5}$ der Fehler $ERR_{\text{fr}}^{\text{K}} = 3{,}04 \cdot 10^{-5}$. Ein solcher Wert erscheint uns jedoch nicht als seriös, da wir für die Referenzlösung selbst keine derartig hohe Genauigkeit voraussetzen können. Daher wurde in Abbildung 9.23 nicht der Wert $ERR_{\text{fr}}^{\text{K}}$ sondern der Wert

$$\max\{ERR_{\text{fr}}^{\text{K}}, 5 \cdot 10^{-4}\}$$

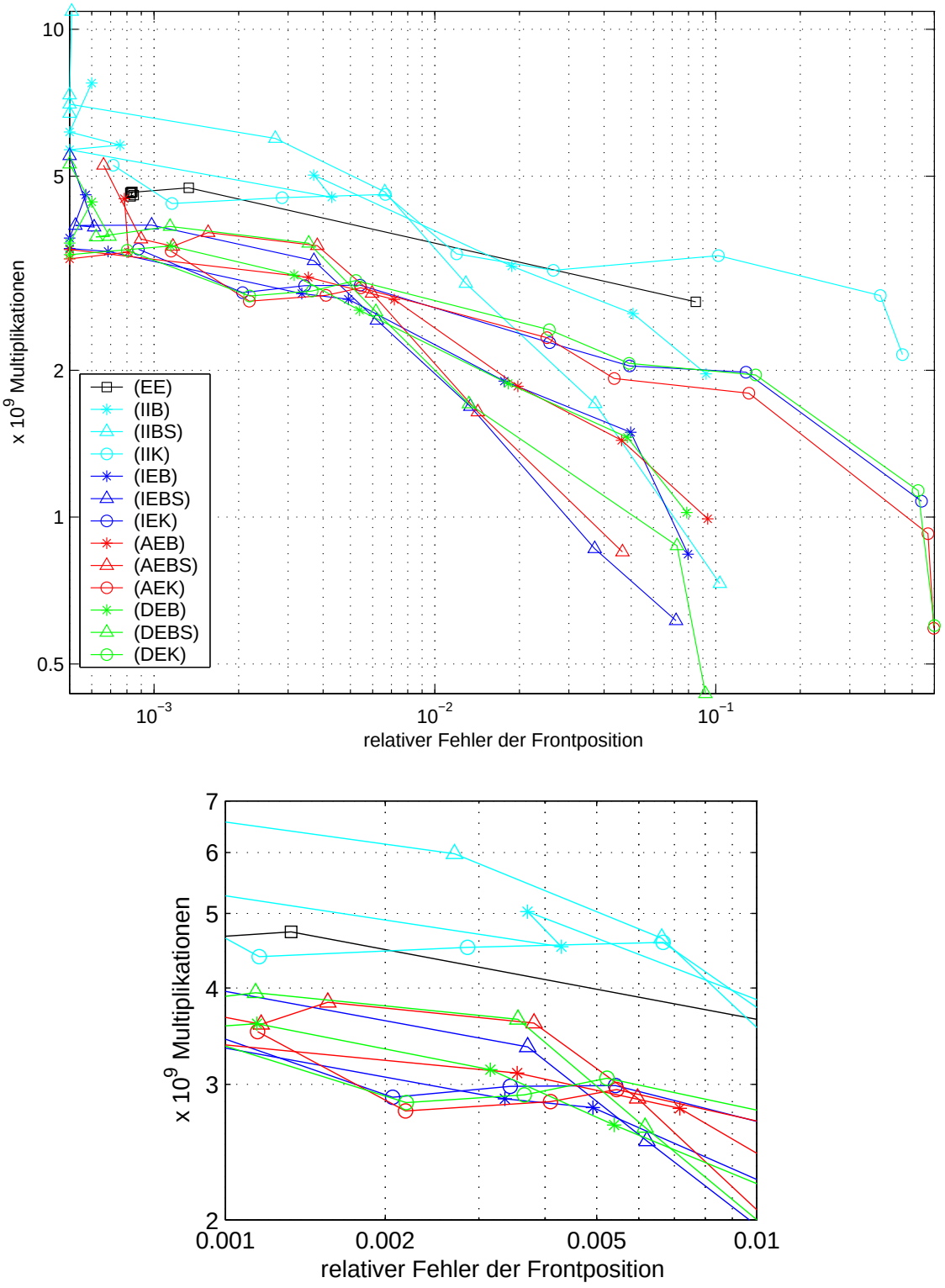


Abbildung 9.23: **oben:** Fehler-Aufwand-Diagramm. Die Toleranz TOL_t variiert auf jeder Kurve zwischen $2 \cdot 10^{-5}$ (links oben) und 10^{-2} (rechts unten). **unten:** Ausschnitt vergrößert

als „relativer Fehler der Frontposition“ dargestellt.

In Abbildung 9.24 links stellen wir für alle untersuchten Verfahren den mittleren Zeitschritt über der Toleranz TOL_t dar. Das explizite Verfahren (EE) erlaubt wegen der Steifheit der u-Komponente nur sehr kleine Zeitschritte. Die mittleren Zeitschritte der übrigen Verfahren stimmen bemerkenswert gut überein.

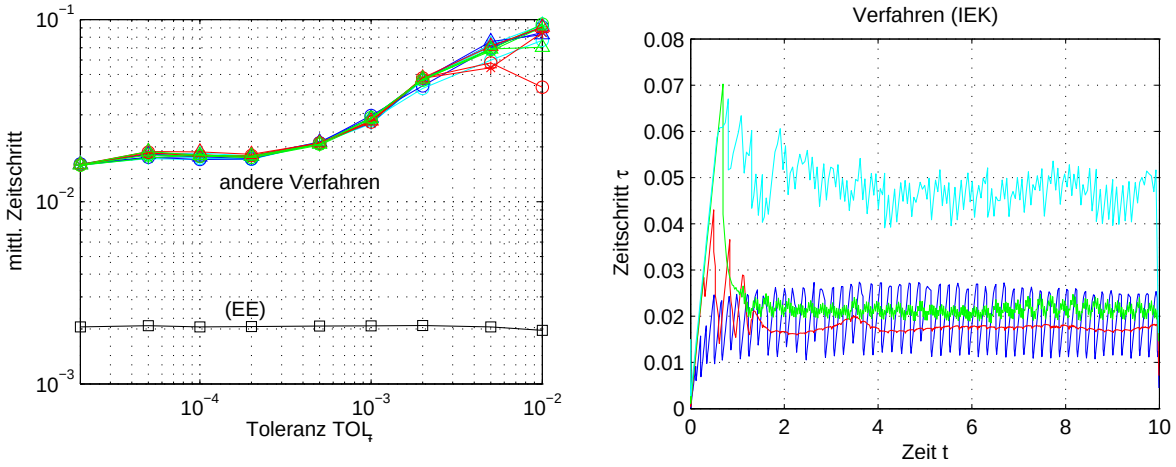


Abbildung 9.24: **links:** Einfluß der Toleranz TOL_t auf den mittleren Zeitschritt; **rechts:** Zeitschritte bei Verfahren (IEK) für die Toleranzen $TOL_t = 2 \cdot 10^{-5}$ (dunkelblau), 10^{-4} (rot), $5 \cdot 10^{-4}$ (grün), $2 \cdot 10^{-3}$ (hellblau)

Auffällig ist, daß die Variation der Toleranz TOL_t zwischen $2 \cdot 10^{-5}$ und $5 \cdot 10^{-4}$ kaum Einfluß auf den mittleren Zeitschritt hat. In Abbildung 9.24 rechts stellen wir für das Verfahren (IEK) die Zeitschritte τ_i über der Zeitachse dar. Man erkennt hier ebenfalls, daß sich der Zeitschritt im Mittel bei den kleinen Toleranzen kaum unterscheidet. Für $TOL_t = 2 \cdot 10^{-5}$ treten jedoch *starke Zeitschrittschwankungen* auf. Auch für die übrigen impliziten Verfahren[10] trifft diese Beobachtung zu. □

9.7 Zusammenfassung

Die drei untersuchten Probleme stellen unterschiedliche Anforderungen an eine Partitionierung. Bei den Problemen BSVD und KRINSKY führen geeignete Partitionierungs-Verfahren zu einer effizienteren Lösung als Verfahren ohne Partitionierung. Beim BSVD-Problem führt lokale Partitionierung teilweise zu Einsparungen von etwa 50% des Rechenaufwandes, der für die Lösung der semidiskreten Probleme benötigt wird. Bei der Wahl des Partitionierungs-Verfahrens sollte die gewünschte Genauigkeit berücksichtigt werden. Die folgende Tabelle gibt einen groben Überblick über die Eignung von Partitionierungs-Verfahren bei den untersuchten Problemen. Sie stellt eine Zusammenfassung der Ergebnisse dar, die in den Untersuchungen zur Effizienz erbracht wurden.

[10] d.h. alle außer (EE)

<table>
<tr><td></td><td>hohe Genauigkeit</td><td>mittlere Genauigkeit</td><td>geringe Genauigkeit</td></tr>
<tr><td rowspan="2">TANH

effizient:</td><td colspan="2">keine Partitionierung benötigt</td><td rowspan="2">Unregelmäßiges Fehlerverhalten erschwert eine Aussage.</td></tr>
<tr><td>(IB)</td><td>(IK)</td></tr>
<tr><td rowspan="2">BSVD

effizient:</td><td colspan="3">lokale Partitionierung sehr sinnvoll</td></tr>
<tr><td>(IK), (DB), (DBS), (DK)</td><td>(DB)</td><td>(AB), (ABS)</td></tr>
<tr><td rowspan="3">KRINSKY

effizient:</td><td colspan="3">Komponenten-Partitionierung nötig</td></tr>
<tr><td colspan="3">lokale Partitionierung nicht benötigt</td></tr>
<tr><td>(IEK), (AEK), (DEK), (IEB)</td><td colspan="2">(IEBS), (AEBS)</td></tr>
</table>

Tabelle 9.7: Eignung von Partitionierungs-Verfahren bei den Problemen TANH, BSVD und KRINSKY

Kapitel 10

Erregbare Medien

„Reiz ist Schönheit in Bewegung."

Gotthold Ephraim Lessing ⟨1729 – 1781⟩

10.1 Beispiele erregbarer Systeme

Ein wichtiger Anwendungsfall für Systeme von Reaktions-Diffusions-Gleichungen ist die Simulation sogenannter erregbarer Medien. Diese treten in den unterschiedlichsten Bereichen in Physik, Chemie und Biologie auf. Eine Gemeinsamkeit aller dieser Prozesse ist das Phänomen, daß ein Medium durch eine oft geringfügige Störung aus einem Gleichgewichtszustand in einen erregten Zustand versetzt wird, der sich dann in Form einer Welle räumlich ausbreitet. Besonders faszinierend sind rotierende Spiralwellen, die in vielen erregbaren Medien auftreten. Den Mechanismus, der zur Entstehung dieser Spiralen führt, werden wir in Abschnitt 10.5 näher erläutern. Ein Überblick über erregbare Medien in den verschiedensten biologischen und chemischen Systemen findet sich etwa bei Manz [111]. Einige Beispiele seien im folgenden aufgeführt.

Auf dem Gebiet der Biologie wurden bereits 1944 von Leão [105] Depolarisationswellen im neuronalen Gewebe wie der Hirnrinde beschrieben. Im Jahre 1946 untersuchten Wiener und Rosenblueth [172] Wellenbewegungen im Herzmuskel und entwickelten dabei die Grundlagen der kinematischen Wellenbeschreibung, die in Abschnitt 10.8 betrachtet werden soll. Spiralförmige Wellen im Herzgewebe sind medizinisch von besonderem Interesse, da sie zu Herzrhytmusstörungen führen [46]. In den fünfziger und sechziger Jahren des 20. Jahrhun-

derts entdeckten HODGKIN und HUXLEY [84] (1952) sowie FITZHUGH [66] (1961) und NAGUMO [124] (1962) Erregungswellen bei der Reizübertragung in Nerven und modellierten diese durch inzwischen klassische Reaktions-Diffusions-Systeme, die in Abschnitt 2.2 bereits aufgeführt wurden. Weitere Beispiele aus der Biologie sind Wellen des zyklischen Adenosinmonophosphats während der Aggregation des Schleimpilzes *Dictyostelium discoideum* [71, 134, 149], Calcium-Wellen im Zytoplasma nach der Befruchtung der Oozyten des Krallenfrosches *Xenopus laevis* [106] und menschlicher Eizellen [157], aber auch die Ausbreitung von Krankheitswellen [36, 122] und wellenartige Populationsentwicklung in Ökosystemen [133, 24].

Erregungswellen bei chemischen Reaktionen wurden schon im Jahre 1906 von LUTHER [109] bei der Oxydation von Oxalsäure durch Permanganat festgestellt. Erregungsfronten bilden sich auch bei der seit 1921 von BRAY und LIEBHAFSKY [27, 28] untersuchten katalytischen Zersetzung von Wasserstoffperoxid durch Jodsäure. Mit der „Bray-Liebhafsky-Reaktion“ wurde erstmalig eine oszillierende chemische Reaktion in homogener Lösung beschrieben.

Eines der bekanntesten Beispiele für eine chemische Reaktion, in der Erregungswellen auftreten, ist die Belousov-Zhabotinsky-Reaktion. BELOUSOV [21, 22] untersuchte 1950/51[1] die Oxydation von Zitronensäure durch Bromat-Ionen bei Verwendung von Cer als Katalysator. ZHABOTINSKY [176] ersetzte 1961, zur Verstärkung des Farbkontrastes, die Zitronensäure durch Malonsäure. Später wurde dann der Katalysator Cer durch Ferroin ersetzt, was einen deutlich sichtbaren Farbumschlag zwischen rot und blau hervorruft. Im Jahre 1972 wurden von WINFREE [173] die für erregbare Medien charakteristischen Spiralwellen in der BELOUSOV-ZHABOTINSKY-Reaktion entdeckt. FIELD, KÖRÖS und NOYES [126, 64] (1972) unterteilten die BELOUSOV-ZHABOTINSKY-Reaktion in elf Elementarreaktionen, an denen zwölf Spezies beteiligt sind und legten eine quantitative Beschreibung dieser Reaktionen vor, den sogenannten „FKN-Mechanismus“. Eine mathematische Modellierung durch Reaktions-Diffusions-Gleichungen, das sogenannte „Oregonator-Modell“, wurde 1974 von FIELD und NOYES [65] angegeben und 1979/80 von TYSON und FIFE [161, 163] weiter vereinfacht. Wir werden in Abschnitt 10.2 auf dieses Modell noch näher eingehen.

Eine Kombination von Bray-Liebhafsky- und Belousov-Zhabotinsky-Reaktion wurde 1973 von BRIGGS und RAUSCHER [30] untersucht. Dabei läuft zwischen Wasserstoffperoxid, Jodat, Malonsäure und Perchlor- oder Schwefelsäure eine oszillierende Reaktion ab, bei der auch eine räumliche Ausbreitung von Wellen stattfindet. Diese Reaktion wird auch als „Jod-Uhr“ bezeichnet. Weitere Beispiele für erregbare chemische Systeme sind die katalytische Oxydation von Kohlenmonoxid auf Platinoberflächen im Vakuum [89, 59] oder auch die Korrosion von Metallen in elektrochemischen Systemen [99, 5, 4]. Schließlich können selbst Galaxien mit ihren rotierenden Spiralarmen als Wellen in einem erregbaren Medium aus Gas und Staub angesehen werden [145, 110].

10.2 Die Modellierung erregbarer Systeme durch Reaktions-Diffusions-Gleichungen

Auch wenn in der Natur vorkommende Systeme i.a. aus einer Vielzahl einzelner Mechanismen bestehen, so lassen sich jedoch die einfachsten erregbaren Medien bereits durch zwei

[1] Die Arbeiten wurden erst einige Jahre später veröffentlicht.

Reaktions-Diffusions-Gleichungen modellieren. Auch die Dynamik von Erregungswellen kann am anschaulichsten an einem Zwei-Komponenten-System verdeutlicht werden. Ein solches System läßt sich stets in der Form

$$\begin{aligned} \frac{\partial u}{\partial t}(\mathbf{x},t) &= d_1 \Delta u(\mathbf{x},t) + f(u(\mathbf{x},t), v(\mathbf{x},t)), \qquad \mathbf{x} \in \Omega,\ t \in [t_0, t_e] \\ \frac{\partial v}{\partial t}(\mathbf{x},t) &= d_2 \Delta v(\mathbf{x},t) + g(u(\mathbf{x},t), v(\mathbf{x},t)) \end{aligned} \tag{10.1}$$

angeben. Im Rahmen dieser Arbeit betrachten wir lediglich zweidimensionale erregbare Medien, bei denen Ω entweder ein Gebiet in $\mathbb{R}^2$ oder ein Gebiet auf einer zweidimensionalen Mannigfaltigkeit $S \subset \mathbb{R}^3$ ist. Die Größe u sei die erregbare Komponente; sie kann von einem Gleichgewichtszustand in einen erregten Zustand springen. Die Komponente v dient der Dämpfung von u. Sie führt auf der Wellenrückseite zum Abklingen der Erregung von u. Die Größe u wird oft als **Aktivator**, v als **Inhibitor** bezeichnet.

Damit bewegte Wellen entstehen, muß – wie wir später sehen werden – die Diffusionskonstante $d_1 > 0$ sein. Für d_2 muß lediglich $d_2 \geq 0$ gelten. In der Praxis gibt es durchaus Fälle, in denen $d_2 = 0$ ist, so etwa bei Erregungswellen in neuromuskulärem Medium oder auch bei der Belousov-Zhabotinsky-Reaktion mit immobilisiertem Katalysator. Bei der flüssigen Belousov-Zhabotinsky-Reaktion ist $d_1 \approx d_2$. Der Fall $d_2 \gg d_1$ ist charakteristisch für stehende Wellen, sogenannte Turing-Muster[2], die in der Natur beispielsweise auf Schneckenhäusern und Muschelschalen vorkommen, wie etwa Untersuchungen von Meinhardt und Klinger [116] zeigen.

Für die Wahl der Funktionen f und g gibt es verschiedene Möglichkeiten, die jedoch alle gewisse qualitative Gemeinsamkeiten besitzen. Wir geben im folgenden drei typische Modelle an, die in der Literatur oft verwendet werden. Das klassische Modell, das speziell zur Beschreibung der Belousov-Zhabotinsky-Reaktion entwickelt wurde, ist das bereits erwähnte zweikomponentige Oregonator-Modell, gegeben durch das System (10.1) mit den Quelltermen

$$f(u,v) = \frac{1}{\varepsilon}\left(u - u^2 - pv\frac{u-q}{u+q}\right) \qquad \text{und} \qquad g(u,v) = u - v. \tag{10.2}$$

Die Bestimmung der auftretenden Parameter aus der chemischen Reaktion ist kompliziert. Von Jahnke und Winfree [87] existieren numerische Untersuchungen zur Abhängigkeit der Wellendynamik von den Parametern ε und p. Rotierende Spiralwellen treten beispielsweise für die Parameter

$$d_1 = 1, \qquad d_2 = 0{,}6, \qquad \varepsilon = 0{,}01, \qquad p = 1{,}4, \qquad q = 0{,}002$$

auf. Eine durch numerische Simulation gewonnene Lösung mit diesen Parametern ist in Abschnitt 10.6.3 dargestellt.

Ein weiteres Modell wurde erstmals 1972 von Krinsky, Pertsov und Reshetilov [98] angegeben. Zykov [181] verwendete dieses Modell für umfangreiche Untersuchungen erregbarer

[2] nach Turing [159] (1952)

Medien. Hierbei sind f und g von der Form

$$f(u,v) = \begin{cases} -k_1 u - v, & u < \sigma, \\ k_f(u-a) - v, & \sigma < u < 1-\sigma, \\ k_2(1-u) - v, & 1-\sigma < u, \end{cases} \qquad g(u,v) = \begin{cases} \varepsilon(k_g u - v), & k_g u \geq v, \\ \varepsilon k_\varepsilon (k_g u - v), & k_g u < v. \end{cases} \tag{10.3}$$

k_1 und k_2 werden dabei so gewählt, daß f stetig ist. Mögliche Parameter sind etwa

$$d_1 = 1, \qquad d_2 = 0, \qquad k_1 = 15{,}3, \qquad k_2 = 151{,}3, \qquad k_f = 1{,}7, \tag{10.4}$$
$$k_g = 2, \qquad k_\varepsilon = 6, \qquad a = 0{,}1, \qquad \sigma = 0{,}01, \qquad \varepsilon = 0{,}25.$$

In Abschnitt 10.6 zeigen wir eine numerische Simulation mit diesen Parametern.

Ein drittes Modell, gegeben durch

$$f(u,v) = \frac{1}{\varepsilon} u(1-u)\left(u - \frac{v+b}{a}\right), \qquad g(u,v) = u - v$$

mit den Parametern

$$d_1 = 1, \qquad d_2 = 0, \qquad a = 0{,}3, \qquad b = 0{,}01, \qquad \varepsilon = 0{,}005,$$

geht auf Barkley, Kness und Tuckerman [18] zurück.

10.3 Dynamik erregbarer Medien

Die Dynamik erregbarer Medien läßt sich am besten mit Hilfe eines der genannten Modelle verdeutlichen. Wir betrachten hier das Modell von Krinsky, Pertsov und Reshetilov. Ohne Berücksichtigung der Diffusion erhalten wir aus (10.1) die zwei gewöhnlichen Differentialgleichungen

$$\frac{du}{dt} = f(u,v), \qquad \frac{dv}{dt} = g(u,v)$$

mit den im vorigen Abschnitt angegebenen Funktionen f und g. Abbildung 10.1 zeigt das Phasenportrait dieser Differentialgleichungen.

Bei $u = v = 0$ liegt eine stabile Ruhelage vor, jedoch führt bereits eine geringe Auslenkung nach rechts dazu, daß u gegen 1 strebt und erst nach längerer Zeit wieder in die Nähe der Ruhelage gelangt. Die Linie $u_S(v)$ gibt einen Schwellenwert für u an. Falls $u < u_S$ ist, so strebt der Orbit ohne großen Umweg der Ruhelage zu. Für $u > u_S$ wird u zunächst erregt, d.h. u strebt gegen 1. Dann nimmt v zu, was die Erregung von u wiederum abklingen läßt. Die Trajektorie bewegt sich somit anfangs von der Ruhelage weg und dann in einem großen Bogen wieder auf die Ruhelage zu.

Bewegte Wellen benötigen eine positive Diffusion $d_1 > 0$. Sie treten bereits im räumlich eindimensionalen Fall auf, wenn eine entsprechende Anfangsbedingung gewählt wird. Wir betrachten das o.g. Modell von Krinsky et al. in einer Raumdimension. Eine Anfangsbedingung, die eine bewegte Welle erzeugt, ist beispielsweise

$$u(x,0) = \begin{cases} 0, & x \in \mathbb{R}\backslash[-5,5], \\ 1, & x \in [-5,5], \end{cases} \qquad v(x,0) = \begin{cases} 0, & x \in]-\infty,-5], \\ x/5+1, & x \in [-5,5], \\ 2, & x \in [5,\infty[. \end{cases}$$

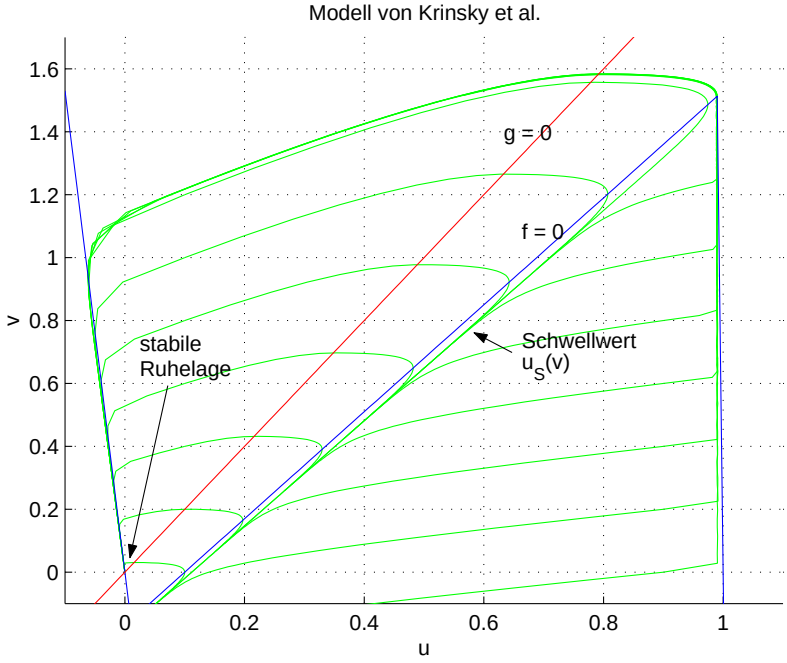


Abbildung 10.1: Dynamik des Modells von KRINSKY et al.

Starten wir die Rechnung mit dieser Anfangsbedingung, so stellt sich nach einiger Zeit eine sich mit konstanter Geschwindigkeit nach links bewegende Welle ein, deren Form sich nicht mehr ändert. Eine derartige Wellenbewegung bezeichnen wir als **stationär**. Die Gestalt der Welle ist in Abbildung 10.2 angegeben.

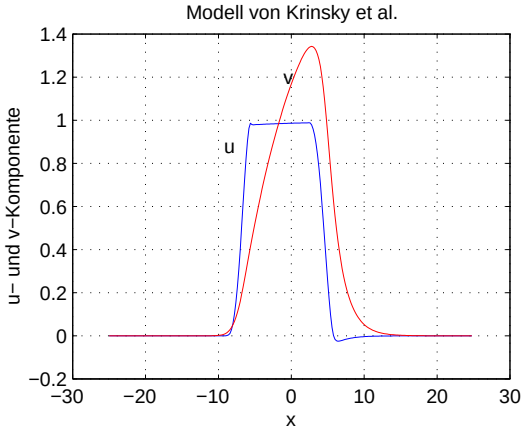


Abbildung 10.2: Form der stationären Welle, räumlich eindimensional mit dem Modell von KRINSKY et al. berechnet

Für die Wellenausbreitung sorgt der folgende Mechanismus: An der Wellenfront (anfangs bei $x = -5$) sorgt die Diffusion von u für eine Überschreitung des Schwellenwertes $u_S(v)$. Damit steigt u links der Front an, was zu einer Bewegung der Front nach links führt. An der Wellenrückseite (anfangs bei $x = 5$) ist dagegen $u_S(v)$ durch den hohen v-Wert so groß, daß die Schwelle nicht überschritten wird. Es kommt hier zu einem Abklingen der u-Werte

entsprechend der oben angegebenen Trajektorien, so daß sich auch die Rückseite der Welle nach links bewegt. Die v-Komponente ist so an die u-Komponente gekoppelt, daß die v-Welle der u-Welle immer etwas nacheilt. Damit bleibt auch für spätere Zeitpunkte der oben beschriebene Zustand erhalten.

Gestalt und Geschwindigkeit der stationären Welle, die sich nach hinreichend langer Zeit herausbildet, sind charakteristische Größen des Reaktions-Diffusions-Systems und hängen *nicht* von der Anfangsbedingung ab. Bewegen sich zwei Wellen aufeinander zu, so kommt es zur Auslöschung beider Wellen.

Um bewegte Wellen zu erzeugen, müssen die folgenden Punkte bei der Wahl der Anfangsbedingungen berücksichtigt werden:

- u und v müssen vor der Wellenfront nahe am Äquilibrium sein.
- An der Wellenrückseite muß v hinreichend groß sein, um eine Rückwärtsbewegung zu verhindern.
- Eine Mindestbreite der Welle wird benötigt, damit sie nicht von der u-Diffusion zerstört wird.
- Der Vektor (u, v) darf sich nicht in instabilen Bereichen des Phasenportraits befinden.

Erfüllt die Anfangsbedingung diese Kriterien, so stellt sich eine bewegte Welle ein, deren Profil und Geschwindigkeit nach einer gewissen Übergangszeit konstant bleiben.

10.4 Numerische Untersuchungen räumlich eindimensionaler Erregungswellen

In diesem Abschnitt werden einige numerische Untersuchungen von Erregungswellen präsentiert. Die Berechnungen wurden mit dem Modell von Krinsky et al. in einer Raumdimension durchgeführt. In einer ersten Meßreihe studieren wir den Einfluß der Gitterfeinheit und der Toleranz für den lokalen Fehler TOL_t auf die Genauigkeit der Lösung. Eine zweite Untersuchung befaßt sich mit der Abhängigkeit der Geschwindigkeit der Welle von dem Erregungsparameter ε und der Diffusionskonstante d_2. Beide Untersuchungen dienen im wesentlichen einem Vergleich mit entsprechenden Messungen bei räumlich zweidimensionalen Wellen, die in Abschnitt 10.6 vorgestellt werden.

Untersuchung 10.1 (Einfluß von Gitterfeinheit und Fehlertoleranz). Wir untersuchen die Abhängigkeit der Genauigkeit der Lösung von der Maschenweite h eines uniformen Gitters und der Toleranz für den lokalen Fehler TOL_t. Grundlage der Berechnungen ist das in (10.1), (10.3) angegebene Modell von Krinsky et al.. Das zugrundeliegende Gebiet ist das Intervall $\Omega =]-20, 20[$. Die Modellparameter sind

$$d_1 = 1, \qquad d_2 = 0, \qquad k_1 = 15{,}3, \qquad k_2 = 151{,}3, \qquad k_f = 1{,}7, \tag{10.5}$$

$$k_g = 2, \qquad k_\varepsilon = 6, \qquad a = 0{,}1, \qquad \sigma = 0{,}01, \qquad \varepsilon = 0{,}2.$$

Es werden homogene Neumannsche Randbedingungen vorgegeben. Die Anfangsbedingungen sind durch

$$u(x,0) = \begin{cases} 1, & x \in [14,17], \\ 0, & \text{sonst}, \end{cases} \qquad v(x,0) = \begin{cases} 0, & x \le 14, \\ x/2 - 7, & x \in [14,17], \\ 1{,}5, & x \ge 17 \end{cases}$$

gegeben. Zur Ortsdiskretisierung werden zentrale Differenzen verwendet. Die Zeitdiskretisierung erfolgt mit dem in Kapitel 7 beschriebenen Krylov-W-Verfahren zweiter Ordnung.

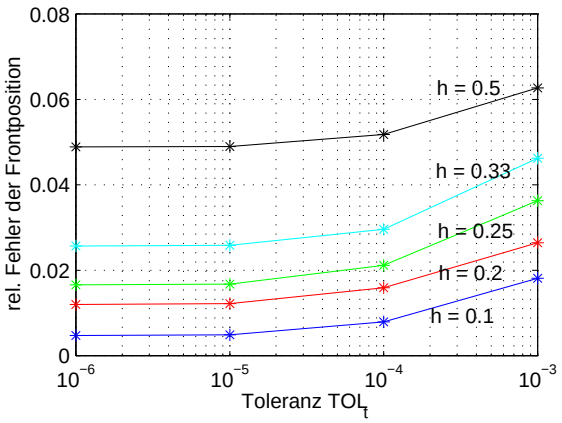


Abbildung 10.3: Abhängigkeit des Fehlers der Frontposition von Maschenweite h und Toleranz TOL_t

Als ein Maß für die Genauigkeit der Lösung wird der Fehler der Frontposition $\|x_{\text{fr}} - x_{\text{fr,ex}}\|$ zum Zeitpunkt $t_e = 10$ betrachtet. Dabei sei $x_{\text{fr}} \in \Omega$ die Frontposition, definiert durch die Beziehungen $u(x_{\text{fr}}) = 0{,}5$ und $du/dt(x_{\text{fr}}) > 0$. Eine „näherungsweise exakte“ Frontposition $x_{\text{f,ex}}$ erhält man durch eine Referenzrechnung mit dem in (5.19) angegebenen expliziten Verfahren zweiter Ordnung bei Verwendung eines sehr feinen Gitters und sehr kleiner Zeitschritte. Wir wählen hierfür die Gitterweite $h = 0{,}01$ und den Zeitschritt

$$\tau = \begin{cases} 10^{-6}, & t < 0{,}02, \\ 10^{-5}, & t \ge 0{,}02. \end{cases}$$

Ergebnis. Die Resultate sind in Abbildung 10.3 dargestellt. Es zeigt sich eine deutliche Abhängigkeit der Genauigkeit von h und TOL_t, wobei für die hier untersuchten Werte von h und TOL_t die maximale Genauigkeit bereits bei $TOL_t = 10^{-5}$ erreicht wird. □

Untersuchung 10.2 (Einfluß von Erregbarkeit und Diffusion). In dem Modell von Krinsky et al. kann mit dem Parameter ε die Erregbarkeit des zugrundeliegenden Mediums gesteuert werden, was sich u.a. in unterschiedlichen Geschwindigkeiten einer Welle äußert. Dabei nimmt die Erregbarkeit ab, wenn ε vergrößert wird. Außerdem haben auch die Diffusionskoeffizienten Einfluß auf die Wellengeschwindigkeit. Wir untersuchen diesen Zusammenhang bei Verwendung der gleichen Parameter, Rand- und Anfangsbedingungen wie in Untersuchung 10.1 mit der Ausnahme, daß nun d_2 die Werte $0,\ 0{,}2,\ 0{,}4,\ 0{,}6,\ 0{,}8,\ 1$ und ε die Werte $0{,}1,\ 0{,}12,\ 0{,}14, \ldots, 0{,}5$ durchläuft. Es sei $h = 0{,}1$ und $TOL_t = 10^{-5}$. Zur Ortsdiskretisierung des Diffusionsteils werden zentrale Differenzen, zur Zeitdiskretisierung das in Kapitel

7 beschriebene Krylov-W-Verfahren verwendet. Abbildung 10.4 zeigt die Abhängigkeit der Frontgeschwindigkeit zum Zeitpunkt $t_e = 20$ von ε und d_2. Wie aus der Graphik hervorgeht, sinkt die Frontgeschwindigkeit bei einer Vergrößerung von ε und d_2. □

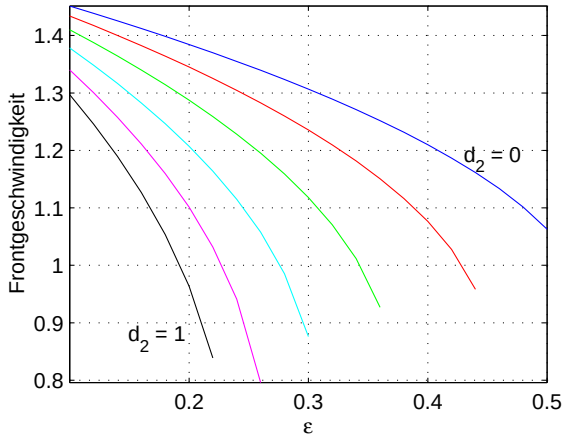


Abbildung 10.4: Abhängigkeit der Frontgeschwindigkeit von ε für $d_2 = 0,\ 0{,}2,\ 0{,}4,\ 0{,}6,\ 0{,}8,\ 1$

10.5 Spiralwellen in erregbaren Medien

Betrachtet man Erregungswellen in räumlich mehrdimensionalen Gebieten, so können verschiedene Wellenformen auftreten. Man beobachtet geschlossene Wellenformen, wie etwa expandierende Kreiswellen, oder auch stehende Wellen, die bereits erwähnten Turing-Muster, die dann auftreten, wenn $d_2 \gg d_1$ ist. Eine weitere Wellenform, die in biologischen, chemischen und numerischen Experimenten beobachtet werden kann, ist die rotierende Spirale.

Eine Spiralwelle kann entstehen, wenn eine Welle ein offenes Ende besitzt, also weder unendlich ausgedehnt noch, wie etwa eine ringförmige Welle, geschlossen ist. Werden beispielsweise im $\mathbb{R}^2$ die Modellgleichungen von Krinsky et al. mit den Parametern (10.5) betrachtet, so kann bei Verwendung der Anfangsbedingung

$$u(x,y,0) = \begin{cases} 1, & (x,y) \in [-10,10] \times [-2,2], \\ 0, & \text{sonst}, \end{cases}$$

$$v(x,y,0) = \begin{cases} 1{,}5, & y \le -2, \\ 1{,}5 - 0{,}375\ y, & -2 < y < 2, \\ 0, & y \ge 2 \end{cases}$$

eine Doppelspirale simuliert werden. Die Entwicklung einer derartigen Spiralwelle ist in Abbildung 10.5 schematisch dargestellt; sie wird durch den folgenden Mechanismus hervorgerufen:

Es sei E der erregte Bereich, d.h. der Bereich hoher Aktivatorkonzentration (hoher u-Werte) und R der refraktäre Bereich (hohe Inhibitorkonzentration, hohe v-Werte). Da die Ausbreitung von E durch R gehemmt wird, biegt sich E an den Enden um R herum. R wird jedoch von E angezogen, was dazu führt, daß R der Biegung von E folgt. Das führt schließlich zur Ausbildung einer rotierenden Spirale.

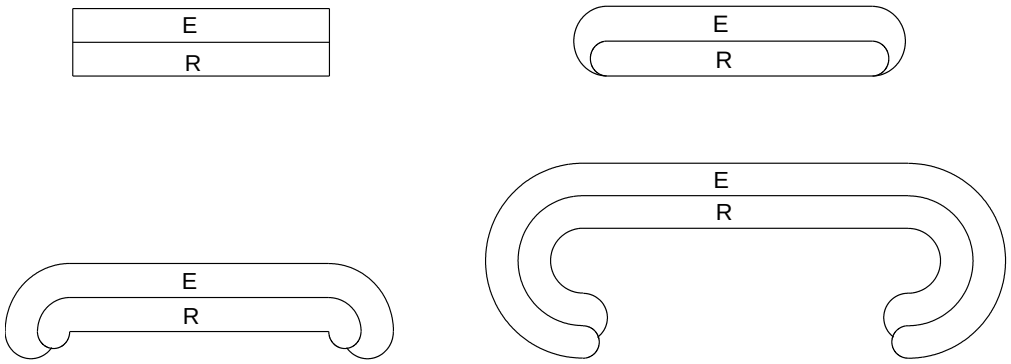


Abbildung 10.5: Entwicklung von Spiralen aus den offenen Enden einer Welle

Wir wollen im folgenden die Rotation der Spiralwellen näher untersuchen. Ein wichtiges Kriterium zur Klassifizierung unterschiedlicher Rotationsformen ist die Bahnkurve, auf der sich die Wellenspitze bewegt. Zunächst müssen wir die Position der Wellenspitze definieren. Wir folgen hierin der Vorgehensweise, die beispielsweise von Zykov und Müller [183] verwendet wird. Im Inneren der Welle befindet sich die Aktivator-Komponente u im erregten Zustand $u \approx u_1$, außerhalb der Welle in der Nähe des Gleichgewichtszustandes u_0. Folglich umrandet die Höhenlinie $u = u_2$ für $u_2 \in (u_0, u_1)$ die Welle. Auf der Vorderseite der bewegten Welle gilt $du/dt > 0$, auf der Rückseite $du/dt < 0$. Man kann daher die Wellenspitze als den Schnittpunkt der Kurven

$$u = u_2 \qquad \text{und} \qquad du/dt = 0 \tag{10.6}$$

definieren. Für das Modell von Krinsky et al. gilt, wenn dessen Parameter in der Größenordnung von (10.4) liegen, stets $u_0 = 0,\ u_1 \approx 1$. Wir wählen in diesem Modell immer

$$u_2 = 1/2 \tag{10.7}$$

zur Bestimmung der Wellenspitze.

Wir unterscheiden die folgenden Formen der Rotation von Spiralwellen:

- die **stationäre Rotation**: Hierbei rotiert die Welle mit konstanter Geschwindigkeit um einen festen Punkt. Die Wellenspitze beschreibt eine Kreisbahn. Die Gestalt der Welle bleibt unverändert, nur ihre Lage ändert sich. In den Abschnitten 10.6.1 und 10.7 zeigen wir numerische Simulationen stationär rotierender Spiralwellen.

Andere – also **instationäre** – Formen der Rotation können wiederum nach der Bahnkurve der Wellenspitze klassifiziert werden.

- Die Wellenspitze bewegt sich auf einer Kreisbahn, aber die Gestalt der Welle ändert sich, beispielsweise durch Zusammenstoß mit anderen Wellen. Beispiele hierfür werden in den Untersuchungen 10.4 und 10.7 dargestellt.
- Die Wellenspitze rotiert um einen Punkt, der wiederum eine Bewegung ausführt. Im Ergebnis beschreibt die Wellenspitze eine Art Zykloide. Wir bezeichnen die Bewegung des Rotationszentrums als **Drift** der Spiralwelle. Eine derartige Drift tritt bei einigen

mit dem Oregonator-Modell erzeugten Wellen auf, siehe Abschnitt 10.6.3, ebenso bei kurzen Wellen, siehe Abschnitt 10.6.2, und bei Wellen auf nichtgleichmäßig gekrümmten Flächen, siehe Abschnitt 10.9.

- Die Wellenspitze beschreibt eine kompliziertere, mäandrierende Bahn. Ein derartiges Verhalten wurde beim Oregonator-Modell beobachtet, siehe Abschnitt 10.6.3.

Bewegt sich die Wellenspitze auf einer Kreisbahn, so kann der Radius dieses Kreises angegeben werden. Wir bezeichnen ihn als den **numerischen Kernradius** ϱ_{num} der Welle.[3]

Die Untersuchung von Spiralwellen in erregbaren Medien geht auf die wegweisende Arbeit von Wiener und Rosenblueth [172] zurück. In der Belousov-Zhabotinsky-Reaktion wurden von Winfree [173] erstmals Spiralwellen beobachtet. Müller, Plesser und Hess [121] beschrieben die Form einer solchen Welle als Archimedische Spirale. Numerische Simulationen von Spiralwellen in ebenen Medien wurden beispielsweise von Barkley, Kness und Tuckerman [18], Michailov und Zykov [118], Zykov, Steinbock und Müller [184] sowie Pardhanani und Carey [129] vorgenommen.

Im Gegensatz zum Verhalten in einem ebenen Medium zeigt die Bewegung von Erregungswellen einige qualitative Unterschiede, wenn sich das zugrundeliegende Medium auf einer gekrümmten Fläche befindet. Das ist zum einen auf topologische Faktoren zurückzuführen. Die Abmessungen des Raumes, der einer Welle zur Verfügung steht, ändern sich naturgemäß, wenn die Fläche gekrümmt wird. Außerdem beeinflußt die lokale Geometrie, insbesondere die Krümmung der Fläche, direkt die Bewegung einer Welle. In [183] untersuchten Zykov und Müller den Einfluß der Krümmung einer Kugel auf die Rotationsfrequenz rotierender Spiralen. In Abschnitt 10.9 werden wir uns besonders mit der Drift von Spiralwellen auf Flächen nichtkonstanter Krümmung beschäftigen. Untersuchungen zu diesem Sachverhalt wurden von Davidov, Zykov, Michailov und Yamaguchi [49, 48, 50] mit Hilfe der kinematischen Theorie und direkter numerischer Simulation durchgeführt.

Erregbare Medien auf gekrümmten Flächen sind nicht nur von theoretischem Interesse. Insbesondere bei einigen der in Abschnitt 10.1 genannten biologischen Systeme treten gekrümmte Oberflächen auf. Beispiele sind die von Davidenko et al. [46] beobachteten Wellen auf der Herzoberfläche, Wellen auf der Hühnerretina [74] und die Calcium-Wellen auf der Oberfläche von Froscheiern [35]. Bereits in der frühen Arbeit von Wiener und Rosenblueth [172] (1946) wurde der Einfluß der Geometrie auf bestimmte Aspekte der Wellendynamik angesprochen.

Die meisten chemischen Experimente, etwa mit der BZ-Reaktion, wurden bisher nur in der Ebene ausgeführt. Zu Erregungswellen auf gekrümmten Medien finden sich nur einige wenige Ergebnisse; hier seien etwa die Untersuchungen von Maselko, Showalter, Steinbock, Davydov, Manz, Zykov und Müller [113, 153, 47, 111] genannt. Chemische Experimente zur Spiraldrift auf gekrümmten Flächen wurden von Davydov, Zykov, Yamaguchi, Manz, Müller und Bär [50, 112] ausgeführt.

Numerische Simulationen gekrümmter erregbarer Medien wurden beispielsweise von Zykov, Müller, Davydov, Michailov, Yamaguchi, Yagisita, Mimura, Yamada und Steinbock durchgeführt [183, 49, 48, 50, 174, 153]. Chávez und Kapral [39] berechneten drei-

[3] Im Unterschied dazu wird in der kinematischen Theorie der kinematische Kernradius eingeführt, siehe Definition 10.8.

dimensionale Erregungswellen in einer dünnen Kugelschale.

10.6 Untersuchungen zu Spiralwellen in der Ebene

10.6.1 Eine stationär rotierende Spiralwelle im Modell von Krinsky et al.

Eine stationär rotierende Spiralwelle kann mit dem in (10.1), (10.3) angegebenen Modell von Krinsky et al. numerisch simuliert werden. In den Abbildungen 10.6, 10.7 und 10.8 ist eine mit diesem Modell berechnete Spiralwelle dargestellt. Dabei wurden die folgenden Modellparameter zugrundegelegt.

$$d_1 = 1, \qquad d_2 = 0, \qquad k_1 = 15{,}3, \qquad k_2 = 151{,}3, \qquad k_f = 1{,}7, \tag{10.8}$$

$$k_g = 2, \qquad k_\varepsilon = 6, \qquad a = 0{,}1, \qquad \sigma = 0{,}01, \qquad \varepsilon = 0{,}25.$$

Als Rechengebiet wurde das Rechteck $\Omega =]-50, 50[^2$ gewählt. Die Anfangsbedingungen zum Zeitpunkt $t = 0$ waren

$$u(x,y,0) = \begin{cases} 1, & (x,y) \in [-10,10] \times [-8,-5], \\ 0, & \text{sonst}, \end{cases}$$

$$v(x,y,0) = \begin{cases} 1, & (x,y) \in ([-50,-10] \times [-50,50]) \cup ([10,50] \times [-50,50]), \\ 0, & (x,y) \in [-10,10] \times [-5,50], \\ 1{,}5, & (x,y) \in [-10,10] \times [-50,-8], \\ -(y+5)/2, & (x,y) \in [-10,10] \times [-8,-5]. \end{cases}$$

Auf $\partial\Omega$ wurden homogene Neumannsche Randbedingungen $\partial u/\partial \mathbf{n}_{\partial\Omega} = \partial v/\partial \mathbf{n}_{\partial\Omega} = 0$ vorgegeben. Die Ortsdiskretisierung erfolgte mit linearen finiten Elementen, die Zeitdiskretisierung mit dem in (5.19) beschriebenen expliziten Verfahren zweiter Ordnung.

Die in den Abbildungen 10.6 und 10.7 dargestellte Lösung veranschaulicht einige typische Mechanismen, die bei Erregungswellen auftreten können. Aus den Enden einer kurzen ebenen Welle entwickeln sich rotierende Spiralen. Wenn diese genügend angewachsen sind, kommt es zum Zusammenstoß zweier Wellenfronten und infolgedessen zur Auslöschung eines Teils der Welle. Das ehemalige Mittelstück der Welle wird zu einer expandierenden Ringwelle, während die beiden verbliebenen Endstücken zu einer neuen kurzen Wellenfront fusionieren, die erneut Spiralen ausbildet. Auf diese Weise wird der Zyklus der Spiralbildung fortlaufend wiederholt. In regelmäßigen Abständen wird eine Ringwelle nach außen abgestoßen.

Untersuchung 10.3 (Einfluß von Erregungsparameter und Diffusion auf Kernradius und Winkelgeschwindigkeit). Im Modell von Krinsky et al. haben der Erregungsparameter ε und die Inhibitor-Diffusion d_2 wesentlichen Einfluß auf numerischen Kernradius und Winkelgeschwindigkeit stationär rotierender Wellen. Um diesen Zusammenhang zu untersuchen, führen wir Berechnungen mit dem genannten Modell durch. Das Gebiet Ω, die Parameter $d_1,\ k_1,\ k_2,\ k_f,\ k_g,\ k_\varepsilon,\ a,\ \sigma$ und die Randbedingungen entsprechen den in (10.8)

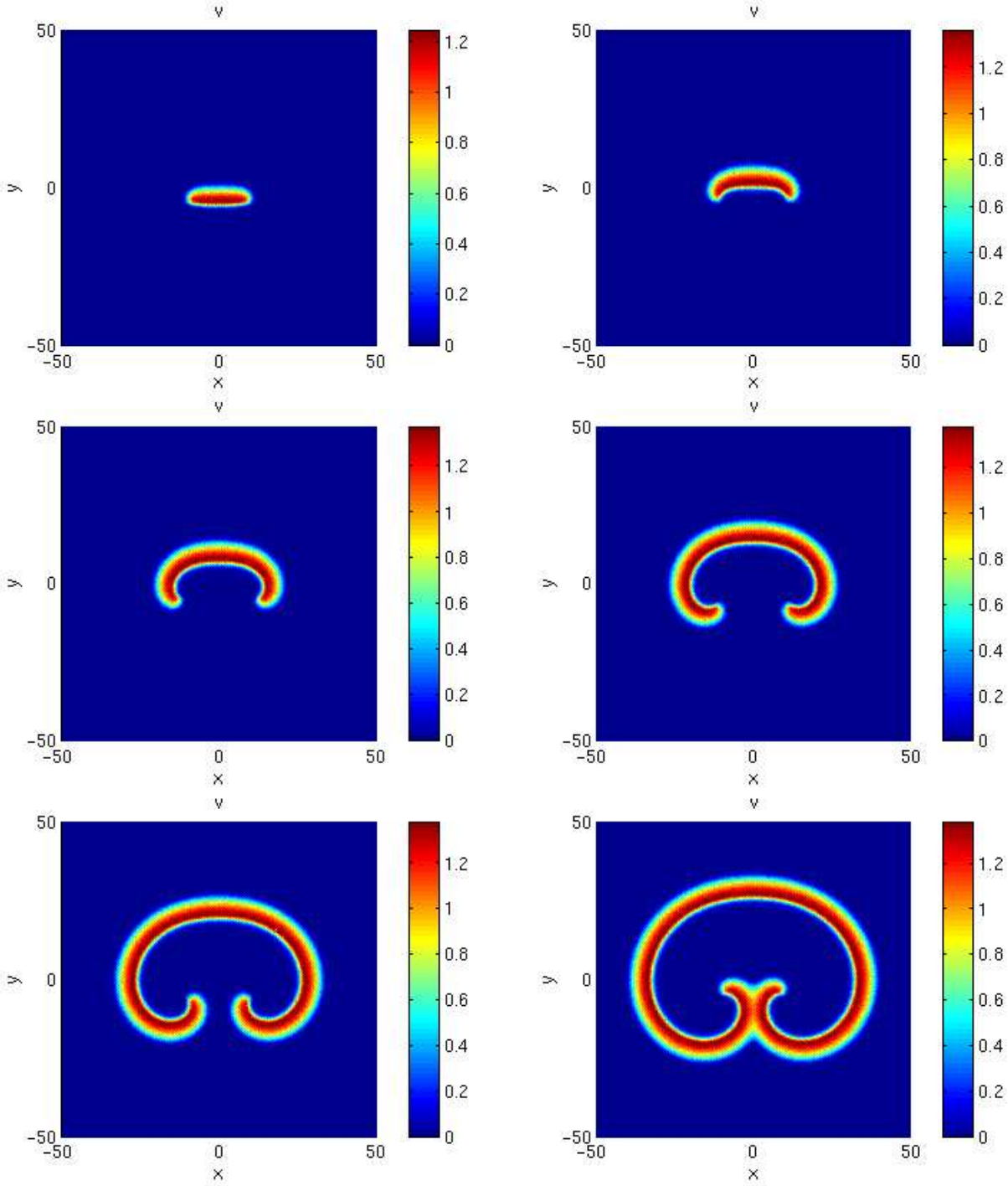


Abbildung 10.6: Lösung des Systems von KRINSKY et al, v-Komponente zu den Zeiten $t = 5, 10, 15, 20, 25, 30$

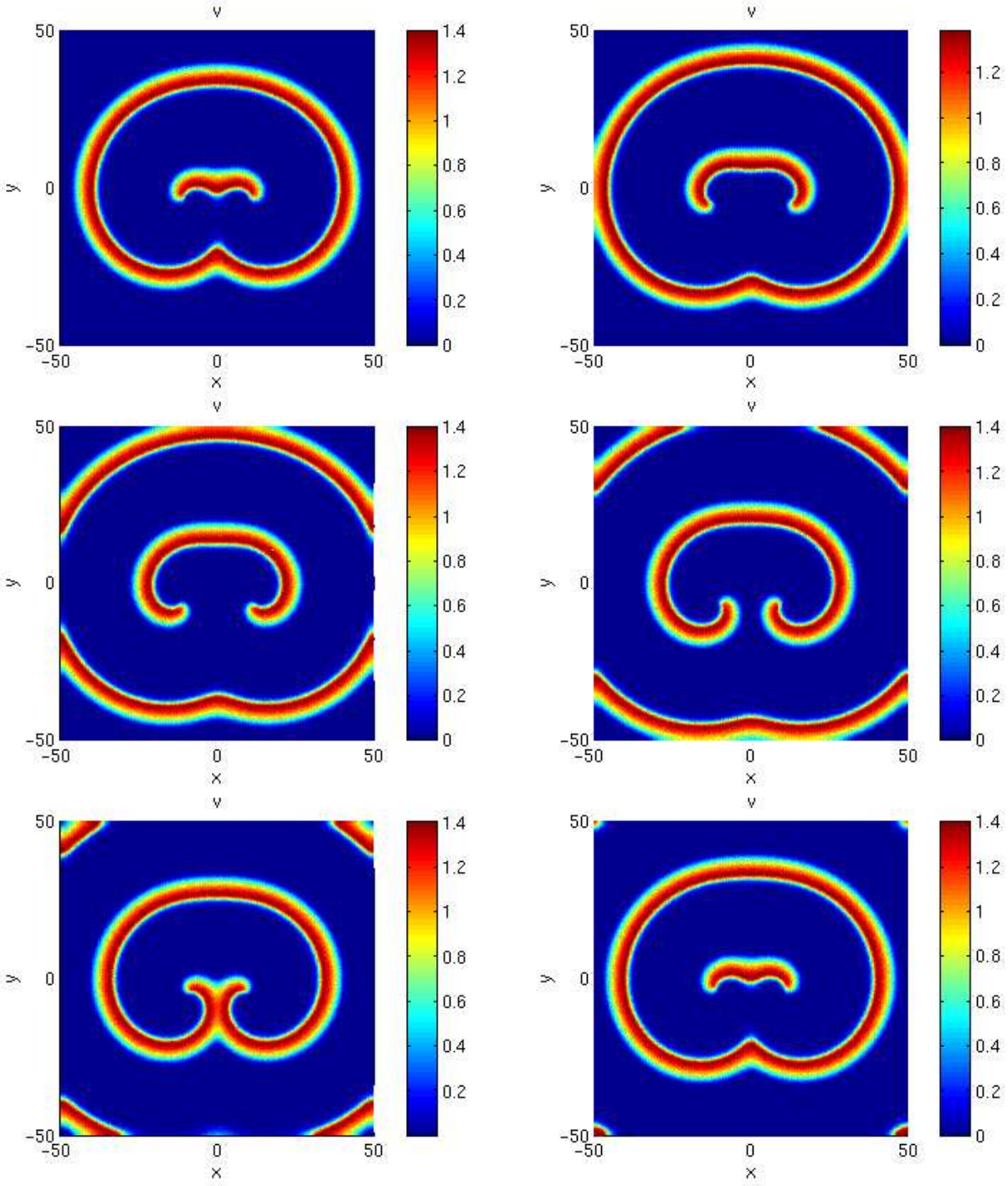


Abbildung 10.7: Lösung des Systems von KRINSKY et al, v-Komponente zu den Zeiten $t = 35, 40, 45, 50, 55, 60$

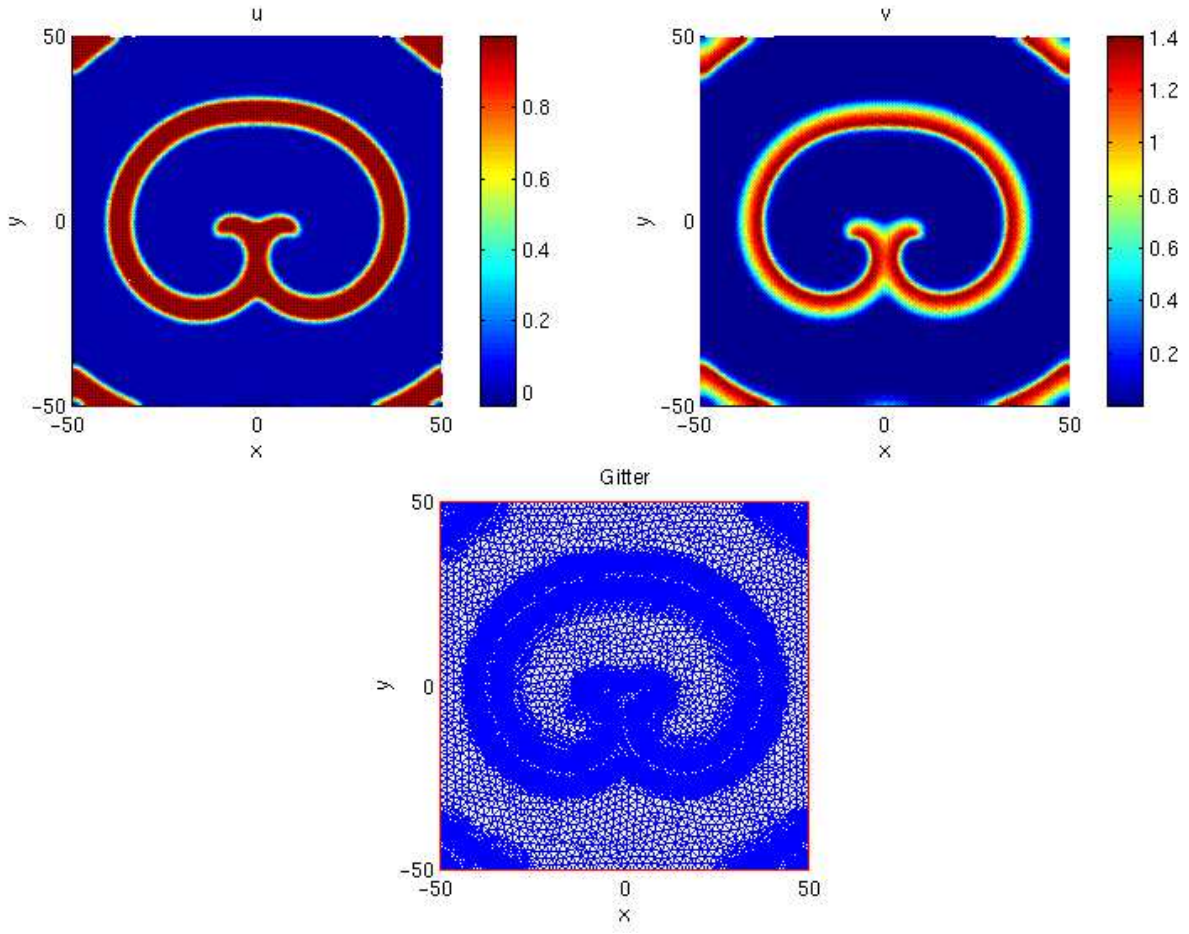


Abbildung 10.8: Lösung des Systems von KRINSKY et al, u-Komponente, v-Komponente und Gitter zur Zeit $t = 55$

angegebenen. Als Anfangsbedingungen wählen wir jetzt

$$u(x,y,t) = \begin{cases} 1, & (x,y) \in [0,50] \times [-10,0], \\ 0, & \text{sonst}, \end{cases}$$

$$v(x,y,t) = \begin{cases} 1, & (x,y) \in [-50,0] \times [-50,50], \\ 0, & (x,y) \in [0,50] \times [0,50], \\ 1{,}5, & (x,y) \in [0,50] \times [-50,-10], \\ -0{,}15y, & (x,y) \in [0,50] \times [-10,0]. \end{cases}$$

Für $d_2 = 0$ und $d_2 = 0{,}4$ sowie verschiedene Werte für ε bestimmen wir den numerischen Kernradius ϱ_{num} und die Winkelgeschwindigkeit ω der nach einer gewissen Zeit stationär rotierenden Welle. Die Ergebnisse sind in Abbildung 10.9 dargestellt. Abbildung 10.10 zeigt die u-Komponente für $d_2 = 0{,}4$, $\varepsilon = 0{,}2$ und $d_2 = 0$, $\varepsilon = 0{,}3$ jeweils zur Zeit $t = 70$. Der weiße Kreis stellt die Bahnkurve der Wellenspitze dar. □

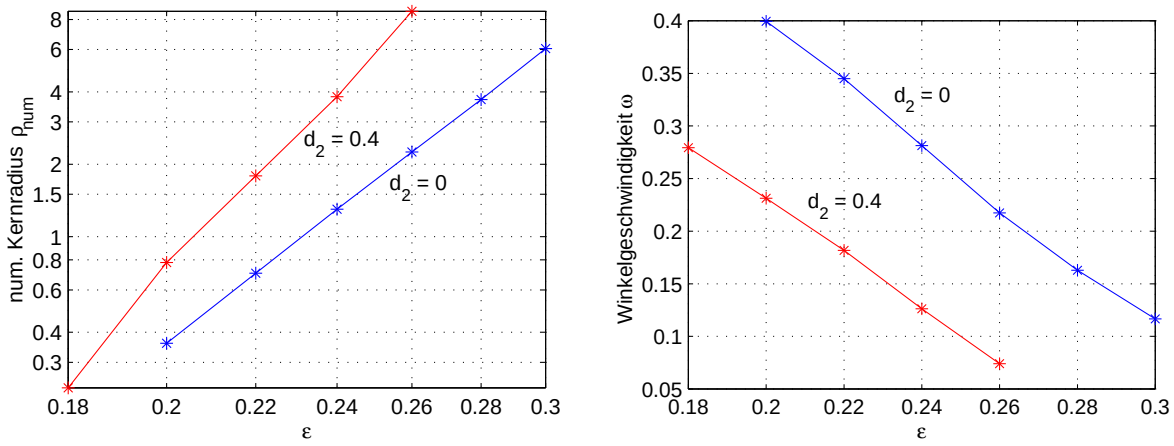


Abbildung 10.9: Numerischer Kernradius ϱ_{num} und Winkelgeschwindigkeit ω in Abhängigkeit von d_2 und ε

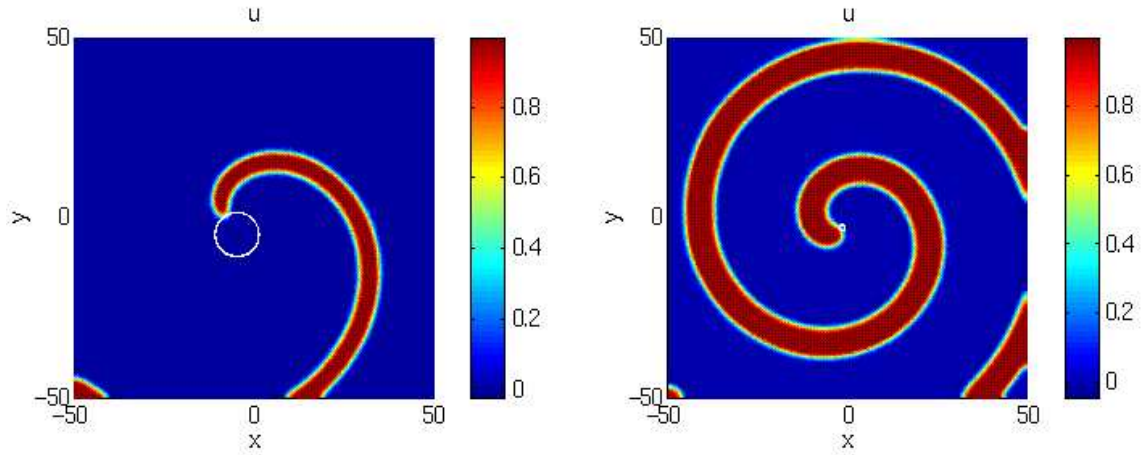


Abbildung 10.10: u-Komponente zur Zeit $t = 70$ und Bahn der Wellenspitze. links: $d_2 = 0$, $\varepsilon = 0{,}3$, rechts: $d_2 = 0{,}4$, $\varepsilon = 0{,}2$

10.6.2 Drift kurzer Spiralwellen im Modell von Krinsky et al.

Wie in Abschnitt 10.6.1 gezeigt wurde, stellen die Lösungen der Modellgleichungen von Krinsky et al. mit den dort angegebenen Parametern und Anfangswerten stationär rotierende Spiralwellen dar. Die Rotationsachse ist dabei ein fester Punkt in der Ebene. Das gilt jedoch nur, wenn die Spiralwellen hinreichend lang sind. Startet man hingegen mit einer kurzen Welle, so ist die Rotationsachse der rotierenden Wellenspitze selbst ein bewegter Punkt, d.h. die Bahnkurve der Wellenspitze ist eine Zykloide. Die Bewegung der Rotationsachse bezeichnen wir als Drift der Welle. Offenbar bestehen bei kurzen Wellen gewisse Wechselwirkungen zwischen den beiden nahe beieinanderliegenden Wellenenden.

Untersuchung 10.4 (Drift kurzer Spiralwellen). Wir wollen im folgenden die Drift kurzer Spiralwellen experimentell untersuchen. Dazu führen wir Berechnungen mit dem in Abschnitt 10.2 beschriebenen Modell von Krinsky et al. durch. Als Rechengebiet wird $\Omega =$

$]-50,50[^2$ gewählt. Die auftretenden Parameter sind $t_0 = 0$, $t_e = 150$, $d_1 = 1$, $d_2 = 0$, $k_f = 1{,}7$, $k_g = 2$, $k_\varepsilon = 6$, $a = 0{,}1$, $\sigma = 0{,}01$, $\varepsilon = 0{,}3$. Wir geben homogene Neumann-Randbedingungen vor. Die einzelnen Testrechnungen unterscheiden sich in den Anfangsbedingungen. Diese sind

$$u(x,y,0) = \begin{cases} 1, & (x,y) \in [-l/2, l/2] \times [-3, 0], \\ 0, & \text{sonst}, \end{cases}$$

$$v(x,y,0) = \begin{cases} 1, & (x,y) \in ([-50,-l/2] \times [-50,50]) \cup ([l/2,50] \times [-50,50]), \\ 0, & (x,y) \in [-l/2, l/2] \times [0,50], \\ 1{,}5, & (x,y) \in [-l/2, l/2] \times [-50,-3], \\ -(y+5)/2, & (x,y) \in [-l/2, l/2] \times [-3,0], \end{cases}$$

wobei l die Ausgangslänge der Welle ist. Die Berechnungen werden für die Werte $l = 8$, 10, 12, 14, 16, 18, 20 durchgeführt.

In Abbildung 10.11 links ist für die zwei Wellen mit den Ausgangslängen $l = 8$ und $l = 20$ jeweils die Bahn der Wellenspitze dargestellt. Man erkennt für die kurze Welle eine starke Drift, die im wesentlichen nach unten verläuft. Bei der langen Welle ist kaum noch Drift erkennbar.

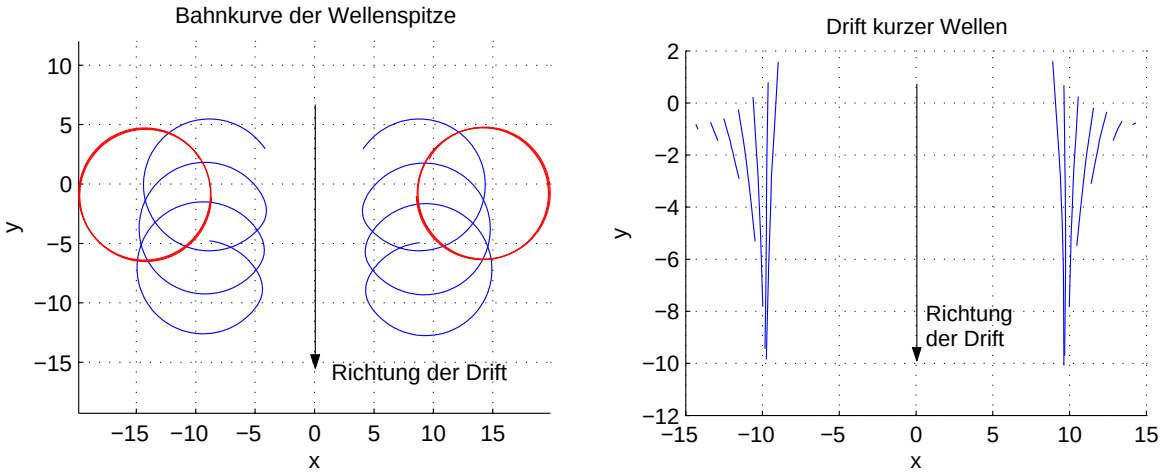


Abbildung 10.11: **links:** Bahnkurven der Wellenspitzen von Wellen der Ausgangslänge $l = 8$ (blau) und $l = 20$ (rot); **rechts:** Bahn der Rotationsachse für Wellen der Ausgangslängen $l = 8, 10, 12, 14, 16, 18, 20$ in der Zeit $t \in [0, 150]$

Wir bezeichnen im folgenden mit $(x_{0l}, y_{0l})(t)$ und $(x_{0r}, y_{0r})(t)$ die Bahn der driftenden Rotationsachse des linken bzw. rechten Wellenendes. Die Ergebnisse einer quantitativen Untersuchung der Driftgeschwindigkeit $d(x_{0l}, y_{0l})/dt$ werden in Abbildung 10.12 links gezeigt. Demnach nimmt die Geschwindigkeitskomponente dy_{0l}/dt und auch der Betrag der Driftgeschwindigkeit zu, je näher sich die Wellenspitze an der Symmetrieachse $x = 0$ befindet. Die Komponente dx_{0l}/dt ist für $|x| > 9{,}8$ positiv und ansonsten negativ. Die rechte Wellenspitze verhält sich aus Symmetriegründen analog. Folglich bewegen sich die Rotationszentren (x_{0l}, y_{0l}) und (x_{0r}, y_{0r}) auf die Geraden $x = \pm 9{,}8$ zu. Abbildung 10.11 rechts zeigt die Bahn dieser Rotationszentren jeweils in der Zeitspanne $t \in [0, 150]$.

Die Darstellung der Bahnkurven der Wellenspitze in Abbildung 10.11 links verdeutlicht bereits, daß die Krümmung dieser Bahnkurven in der Nähe der Symmetrieachse $x = 0$ offenbar

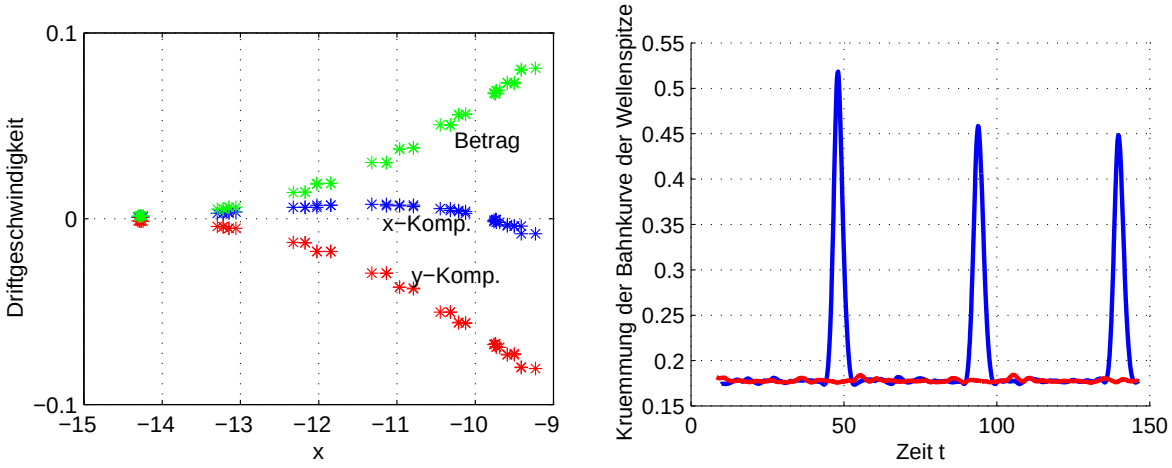


Abbildung 10.12: **links:** Driftgeschwindigkeit $d(x_{0l}, y_{0l})/dt$ in Abhängigkeit der x-Koordinate, *blau:* x-Komponente, *rot:* y-Komponente, *grün:* Betrag der Geschwindigkeit; **rechts:** Krümmung der Bahn der Wellenspitze für $l = 8$ (blau) und $l = 20$ (rot)

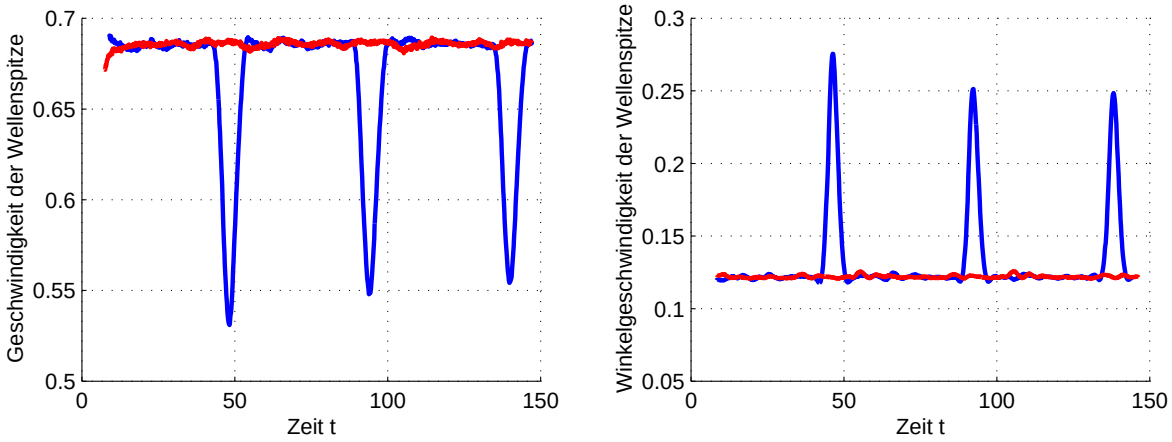


Abbildung 10.13: **links:** Bahngeschwindigkeit der Wellenspitze für $l = 8$ (blau) und $l = 20$ (rot); **rechts:** Winkelgeschwindigkeit der Wellenspitze für $l = 8$ (blau) und $l = 20$ (rot)

zunimmt. Eine Berechnung der Krümmung für die Wellen mit den Ausgangslängen $l = 8$ und $l = 20$ untermauert diese Beobachtung. In Abbildung 10.12 rechts wird die Krümmung über der Zeit aufgetragen. Man erkennt, daß in beiden Fällen die Krümmung die meiste Zeit konstant und gleich ist. Wenn die Wellenspitze sich nahe an der Symmetrieachse $x = 0$ befindet, kommt es jedoch bei der kurzen Welle zu einer Vergrößerung der Krümmung. Dieses Verhalten führt zu der starken Drift der kurzen Welle in Richtung der negativen y-Achse. Abbildung 10.13 links zeigt zusätzlich, daß die Bahngeschwindigkeit der Wellenspitze in den Punkten hoher Krümmung abnimmt.

Das Produkt aus Bahngeschwindigkeit $d\mathbf{x}/dt$ und Krümmung k ist die Winkelgeschwindigkeit

$$\omega = k \left| \frac{d\mathbf{x}}{dt} \right|,$$

die in Abbildung 10.13 rechts über der Zeit t dargestellt ist. Abgesehen von den Zeiten, in denen sich die Spitze der kurzen Welle in der Nähe der Symmetrieachse $x = 0$ befindet, liegt eine konstante Winkelgeschwindigkeit von etwa $\omega = 0{,}12$ vor. Von ZYKOV und MÜLLER [183] wurde mit Hilfe der kinematischen Theorie bei gleichen Parametern der Wert $\omega = 0{,}13$ ermittelt, der mit unserem experimentellen Wert beinahe übereinstimmt. □

Ein ähnliches Driftverhalten wie das hier beschriebene wurde ebenfalls von ZYKOV und MÜLLER [183] beobachtet. In dieser Arbeit wird die Bahnkurve einer Wellenspitze betrachtet, die sich in der Nähe des Randes eines kreisförmigen Gebietes befindet. Es werden homogene Neumann-Randbedingungen angenommen. Hier kommt es ebenfalls zu einer Drift der Spiralwelle entlang des Randes. Die von uns betrachtete kurze Welle verhält sich, aus Symmetriegründen, äquivalent zu einer randnahen Welle mit homogenen Neumann-Randbedingungen bei $x = 0$. Damit liegt der Drift kurzer Spiralwellen der gleiche Mechanismus zugrunde wie der von ZYKOV und MÜLLER beschriebenen Drift randnaher Wellen.

10.6.3 Mit dem Oregonator-Modell erzeugte Spiralwellen

Das Oregonator-Modell ist das klassische System zur Beschreibung der Belousov-Zhabotinsky-Reaktion. Es wird daher häufig zur Modellierung erregbarer Medien herangezogen. Numerische Untersuchungen mit diesem Modell wurden beispielsweise von PARDHANANI und CAREY [129] dokumentiert.

Wir stellen in diesem Abschnitt zwei Simulationen mit dem Oregonator-Modell vor, in denen rotierende Spiralwellen erzeugt werden. Im ersten Fall führt die Wellenspitze eine Driftbewegung aus, die vermutlich unabhängig von der Länge der Welle vorliegt, also nicht mit dem in Abschnitt 10.6.2 beschriebenen Phänomen der Drift kurzer Wellen zu erklären ist. Im zweiten Fall beschreibt die Wellenspitze eine kompliziertere, mäanderförmige Bahn. PARDHANANI und CAREY [129] erhalten mit den gleichen Parametern – allerdings auf einem gröberen Gitter ($h_{\min} = 0{,}23$) – im erstem Fall eine Kreisbahn und im zweiten Fall eine ähnliche mäandrierende Bahn der Wellenspitze. Auch von ZYKOV, STEINBOCK und MÜLLER [184] wurden derartige mäanderförmige Bahnkurven der Wellenspitze beobachtet und untersucht. Trotz umfangreicher Modifikation der Modellparameter gelang es uns jedoch nicht, stationär rotierende Wellen mit dem Oregonator-Modell zu erzeugen.

Untersuchung 10.5 (Spiralwellen mit dem Oregonator Modell). Wir gehen von dem in (10.1), (10.2) dargestellten System auf dem Gebiet $\Omega =]-10, 10[^2$ aus und verwenden die Parameter

$$d_1 = 1, \qquad d_2 = 0{,}6, \qquad \varepsilon = 0{,}01, \qquad p = 1{,}4, \qquad q = 0{,}002$$

Die Anfangsbedingung geben wir in Polarkoordinaten an. Sie lautet

$$\begin{aligned} u(x,y,0) &= \begin{cases} 0{,}8, & 0 < \varphi < 0{,}5, \\ q(p+1)/(p-1), & \text{sonst}, \end{cases} \\ v(x,y,0) &= q(p+1)/(p-1) + \varphi/(8\pi p), \end{aligned}$$

wobei

$$\varphi = \begin{cases} \arctan(y/x), & x > 0,\ y \geq 0, \\ \pi/2, & x = 0,\ y > 0, \\ \pi + \arctan(y/x), & x < 0, \\ 3\pi/2, & x = 0,\ y < 0, \\ 2\pi + \arctan(y/x), & x > 0,\ y < 0, \\ \text{nicht definiert}, & x = y = 0 \end{cases}$$

der Polarwinkel ist. Wir verwenden homogene Neumann-Randbedingungen. Die Wellenspitze sei gemäß (10.6) definiert, wobei $u_2 = 0{,}4$ gewählt werde. Das Problem verlangt eine starke Gitterverfeinerung an der Wellenfront. Numerische Untersuchungen zeigen, daß für ein grobes Gitter mit der Maschenweite $h_{\min} > 0{,}15$ die berechnete Bahn der Wellenspitze stark von der exakten Bahn abweicht. Wir verfeinern das Gitter bis auf $h_{\min} = 0{,}125$. Abbildung 10.14 zeigt links die Lösung dieses Problems zur Zeit $t = 5$ und die Bahn der Wellenspitze, die eine leichte Drift nach oben ausführt.

Die gleiche Rechnung wird nun mit dem Parameter $p = 3$ wiederholt. Die Lösung zur Zeit $t = 5$ und die Bahn der Wellenspitze ist in Abbildung 10.14 rechts zu sehen. □

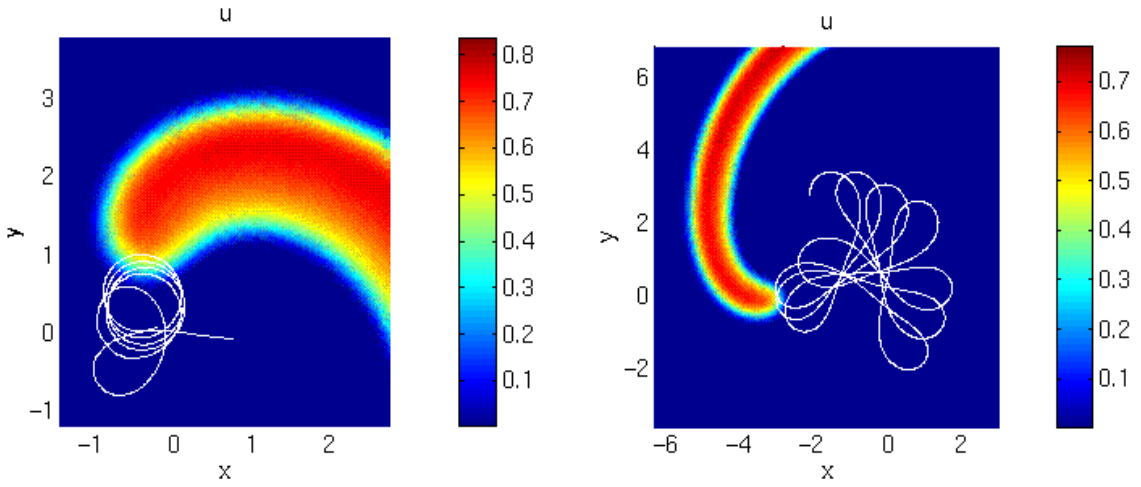


Abbildung 10.14: Mit dem Oregonator-Modell erzeugte Spiralwelle und Bahnkurve der Wellenspitze. **links:** $p = 1{,}4$, **rechts:** $p = 3$

10.7 Spiralwellen auf der Kugeloberfläche

Betrachtet man ein erregbares Medium in einem Gebiet Ω, das auf einer gekrümmten Fläche $S \subset \mathbb{R}^3$ liegt, so muß in dem zugrundeliegenden Reaktions-Diffusions-System (10.1) der Laplace-Operator Δ durch den Laplace-Beltrami-Operator Δ_S ersetzt werden, der die Diffusion einer Größe auf der Fläche S beschreibt. Der Laplace-Beltrami-Operator ist im Anhang in (A.2) definiert.

In diesem Abschnitt sei $S = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = R^2\}$ eine Sphäre vom Radius R

und es sei $\Omega = S$. Wir führen numerische Simulationen mit dem Oregonatormodell

$$\begin{aligned} \frac{du}{dt} &= d_1 \Delta_S u + \frac{1}{\varepsilon_1}\left((1-u)u + fv\frac{q-u}{q+u}\right), \\ \frac{dv}{dt} &= d_2 \Delta_S v + u - v \end{aligned}$$

und mit dem Modell von KRINSKY et al.

$$\begin{aligned} \frac{du}{dt} &= d_1 \Delta_S u + f(u,v), \qquad (10.9) \\ \frac{dv}{dt} &= d_2 \Delta_S v + g(u,v), \\ f(u,v) &= \begin{cases} -k_1 u - v, & u < \sigma, \\ k_f(u-a) - v, & \sigma < u < 1-\sigma, \\ k_2(1-u) - v, & 1-\sigma < u, \end{cases} \\ g(u,v) &= \begin{cases} k_g u - v, & k_g u \geq v, \\ k_\varepsilon(k_g u - v), & k_g u < v, \end{cases} \end{aligned}$$

durch. Die im einzelnen gewählten Parameter werden in den Untersuchungen 10.6 und 10.7 angegeben. Die Anfangsbedingungen werden jeweils symmetrisch zum Äquator $\gamma = \{(x,y,z) \in S : z = 0\}$ vorgegeben. Wegen der die Symmetrie erhaltenden Struktur der Reaktions-Diffusions-Systeme bleibt die Lösung für alle Zeiten t symmetrisch. Daher reicht es aus, die numerische Berechnung auf der nördlichen Halbsphäre $\Omega_1 = \{(x,y,z) \in S : z > 0\}$ durchzuführen. Homogene Neumannsche Randbedingungen $\partial u / \partial \mathbf{n}_{\Omega_1} = 0$ auf dem Äquator γ simulieren eine symmetrische Fortsetzung der Lösung, siehe Abschnitt 3.3.

Zur Ortsdiskretisierung verwenden wir die in Abschnitt 3.2.2 beschriebenen finiten Elemente auf Mannigfaltigkeiten. In den hier dargestellten Simulationen wird die Zeitdiskretisierung mit dem in Kapitel 7 dargestellten Krylov-W-Verfahren durchgeführt. Die Gittererzeugung und -adaption erfolgt wie in Abschnitt 4.5 beschrieben.

Untersuchung 10.6 (Stationäre Rotation). Zur Erzeugung einer stationär rotierenden Welle benutzen wir das oben dargestellte Modell von KRINSKY et al. mit den Parametern $R = 20$, $d_1 = 0{,}1$, $d_2 = 0$, $k_1 = 15{,}3$, $k_2 = 151{,}3$, $k_f = 1{,}7$, $k_g = 2$, $k_\varepsilon = 6$, $a = 0{,}1$, $\sigma = 0{,}01$ und $\varepsilon = 0{,}2$. Es sei S die oben definierte Sphäre, $\Omega_1 = \{(x,y,z) \in S : z > 0\}$ die „nördliche“ Halbsphäre und $\mathbf{k} = (0,0,1)$ der Einheitsvektor in z-Richtung. Durch die Abbildung

$$\Phi : \Omega_1 \to \{(X,Y) \in \mathbb{R}^2 : X^2 + Y^2 < 4\}, \qquad \Phi(x,y,z) = \widetilde{P}_{-\mathbf{k}}(x/R, y/R, z/R) \qquad (10.10)$$

wird das Gebiet Ω_1 auf einen Kreis vom Radius 2 abgebildet. Die Abbildung $\widetilde{P}_{-\mathbf{k}}$ ist die in Abschnitt 4.5.1 definierte stereographische Projektion aus dem Südpol. Wir geben die Anfangsbedingungen in den Bildkoordinaten $(X,Y) = \Phi(x,y,z)$ an. Sie lauten

$$u(\Phi(x,y,z)) = u(X,Y) = \begin{cases} 1, & X < 0{,}025, Y \in [-0{,}2, 0{,}1], \\ 0, & \text{sonst}, \end{cases}$$

$$v(\Phi(x,y,z)) = v(X,Y) = \begin{cases} -5Y + 0{,}5, & Y \in [-0{,}2, 0{,}1], \\ 0, & Y > 0{,}1, \\ 1{,}5, & Y < -0{,}2. \end{cases}$$

Nach einiger Zeit stellt sich bei dieser Wahl der Parameter eine (nahezu) stationäre Spiralwelle ein, die um die z-Achse rotiert, siehe Abbildung 10.15. Neben der Krümmung der beiden Spiralenden tritt auch in der Mitte der Welle – am Äquator – eine Krümmung in entgegengesetzter Richtung auf. Dieses Phänomen geht auf die konvexe Geometrie der Sphäre zurück; es wird auch durch die in Abschnitt 10.8 dargestellte kinematische Theorie gestützt. □

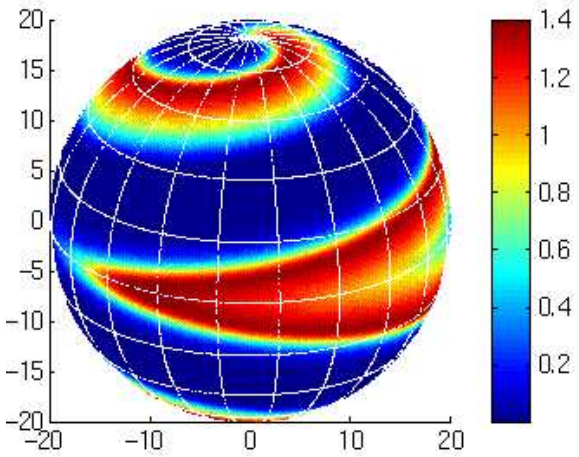


Abbildung 10.15: Lösung des Modells von KRINSKY et al. auf der Sphäre: stationär rotierende Welle, v-Komponente

Um eine stationär rotierende Welle zu erhalten, wurde zur Zeit $t = 0$ ein erregter Streifen vorgegeben, der fast vom Südpol bis zum Nordpol reicht. Instationär rotierende Wellen erhält man, wenn dieser Streifen länger oder kürzer gewählt wird.

Untersuchung 10.7 (Instationär rotierende Wellen). Eine instationäre rotierende Spiralwelle erzeugen wir mit dem eingangs angegebenen Oregonator-Modell bei Verwendung der Parameter $R = 20$, $d_1 = 1$, $d_2 = 0$, $\varepsilon = 0{,}07$, $f = 3$, $q = 0{,}002$. Die Sphäre S werde gemäß

$$\begin{aligned} x &= R\cos\varphi\cos\vartheta, \\ y &= R\sin\varphi\cos\vartheta, \\ z &= R\sin\vartheta, \\ \varphi &\in [-\pi,\pi], \quad \vartheta \in [-\pi/2,\pi/2] \end{aligned}$$

parametrisiert. Als Anfangsbedingungen wählen wir

$$u(\varphi,\vartheta,0) = \begin{cases} 0{,}8, & \varphi \in [0,0{,}05\pi],\ |\vartheta| \le 0{,}2\pi, \\ u_{\mathrm{rel}}, & \text{sonst}, \end{cases}$$

$$v(\varphi,\vartheta,0) = \begin{cases} u_{\mathrm{rel}} + \frac{\varphi}{8\pi f}, & \varphi \ge 0, |\vartheta| \le 0{,}2\pi, \\ u_{\mathrm{rel}} + \frac{\varphi+2\pi}{8\pi f}, & \varphi < 0, |\vartheta| \le 0{,}2\pi, \\ 0{,}05, & |\vartheta| > 0{,}2\pi, \end{cases}$$

wobei $u_{\mathrm{rel}} = q(f+1)/(f-1)$ ist. Die Position der Wellenspitze wird gemäß (10.6), $u_2 = 0{,}3$ festgelegt. In Abbildung 10.16 ist die v-Komponente der Lösung zu den Zeiten $t = 2, 4, 6, 8, 10$

dargestellt. Das Bild rechts unten zeigt die u-Komponente der Lösung zur Zeit $t = 6$ zusammen mit der Bahnkurve der Wellenspitze. Die anfangs kurze Welle dehnt sich aus und stößt mit sich selbst zusammen. Dabei wird ein Teil der Welle ausgelöscht. Die beiden Wellenenden fügen sich zu einer neuen Welle zusammen, während das ehemalige Mittelstück nun eine Ringwelle bildet, die sich bis zur Selbstauslöschung zusammenzieht. Dieser Prozeß kann sich mehrere Male wiederholen. Man kann aus der Bahnkurve der Wellenspitzen erkennen, daß die Rotationszentren der Wellenspitzen ostwärts, aber auch zum Äquator hin driften. Daher werden die nach dem Zusammenstoß neugebildeten Wellen immer kürzer. Schließlich kommt es zur Auslöschung der Welle. □

10.8 Approximation stationärer Wellenfronten durch die kinematische Theorie

10.8.1 Stationäre Wellen in der Ebene

Um die Bewegung von Wellen in einem vereinfachten Modell näherungsweise zu beschreiben, wurde die kinematische Theorie entwickelt, die bereits auf die 1946 erschienene Arbeit von Wiener und Rosenblueth [172] zurückgeht und später vor allem durch Zykov [180] sowie Brazhnik, Davydov und Michailov [29] erweitert wurde. Einen guten Überblick liefert der Artikel von Davydov, Zykov und Michailov [49]. Wir betrachten zunächst ein begrenztes ebenes zweidimensionales Medium, in dem sich das eine Ende einer Erregungswelle befindet. In der kinematischen Theorie wird die Welle lediglich durch eine Kurve γ dargestellt. Diese werde durch ihre Bogenlänge s parametrisiert. Der Endpunkt der Kurve γ, der der Wellenspitze entspricht, erhält den Parameter $s = 0$. Die Kurve γ ist von der Zeit t abhängig. Jedem Punkt der Kurve wird seine Geschwindigkeit in Normalenrichtung $v(s,t)$ zugeordnet, dem Endpunkt noch zusätzlich die Geschwindigkeitskomponente in tangentialer Richtung $c(t)$, siehe Abbildung 10.17.

Die Krümmung der Kurve γ bezeichnen wir mit $k(s,t)$. Sie erfüllt die rein geometrisch begründbare **Grundgleichung der Kinematik** in der Ebene

$$\frac{\partial k}{\partial s}(s,t)\left(\int_0^s k(\sigma,t)\; v(\sigma,t)\; d\sigma + c(s,t)\right) + \frac{\partial k}{\partial t}(s,t) + k(s,t)^2 v(s,t) + \frac{\partial^2 v}{\partial^2 s}(s,t) = 0, \quad (10.11)$$

die auf Zykov [180] zurückgeht. Eine Herleitung ist in [49] angegeben.

Die Bewegung einer Welle in einem errgebaren Medium wird durch die beiden Bewegungsgleichungen

$$v(s,t) = v_0 - Dk(s,t), \qquad c(t) = \beta(k^* - k(0,t)) \quad (10.12)$$

eindeutig festgelegt. Dabei sind die nichtnegativen Parameter v_0, D, β und k^* charakteristische Größen des Reaktions-Diffusions-Systems. Der Parameter v_0 ist die Ausbreitungsgeschwindigkeit einer geraden Wellenfront in der Ebene. Die Größe k^* wird als **kritische Krümmung** bezeichnet. Es handelt sich dabei um die maximale Krümmung einer stationär rotierenden Spiralwelle; diese tritt an der Wellenspitze auf. Die Parameter v_0, D, β und k^* können experimentell durch Auswertung von $v(s,t)$, $k(s,t)$ und $c(t)$ bestimmt werden.

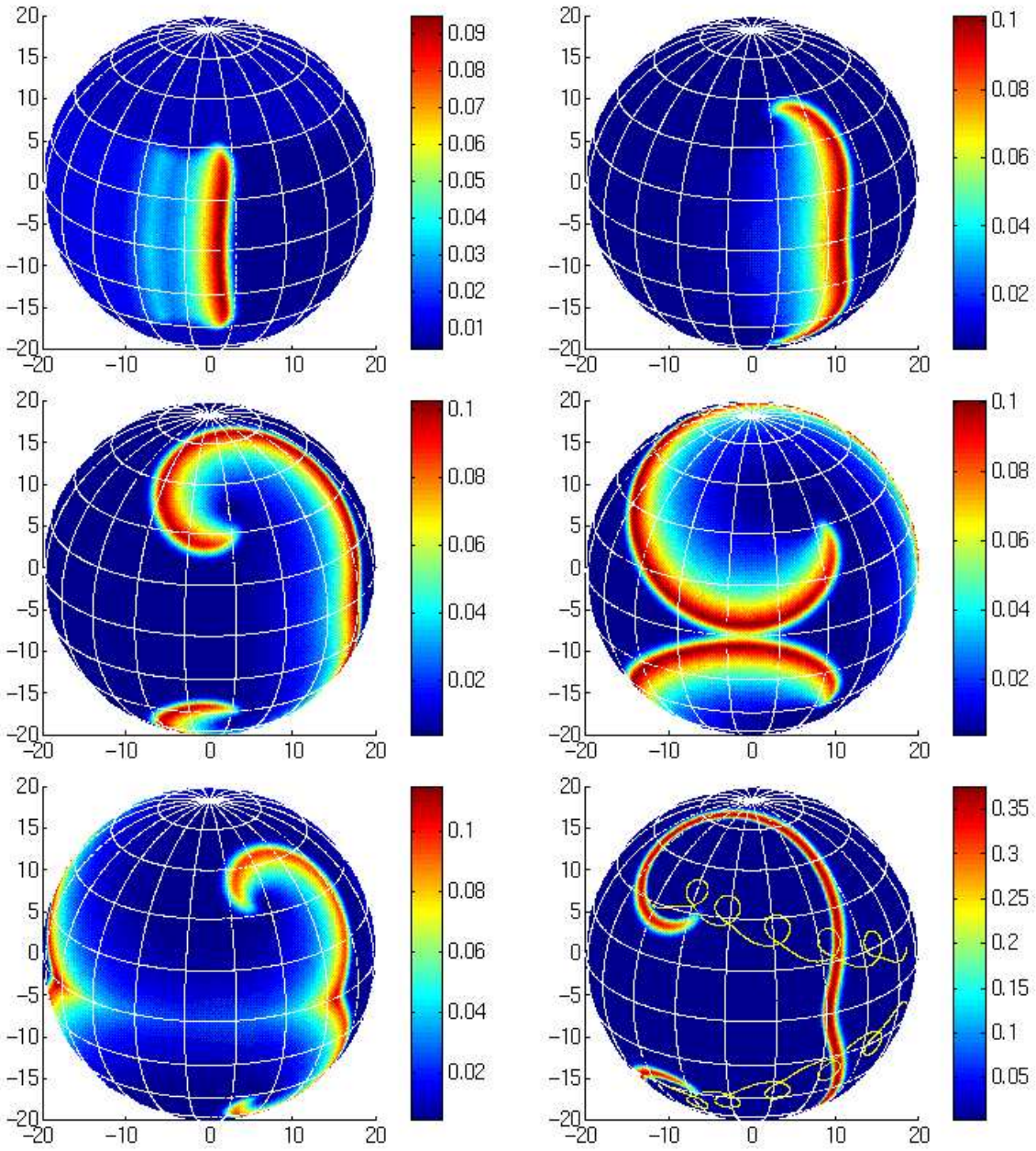


Abbildung 10.16: Lösung des Oregonator-Modells auf der Sphäre: v-Komponente zu den Zeitpunkten $t = 2, 4, 6, 8, 10$, **rechts unten:** u-Komponente zur Zeit $t = 6$ und Bahnkurve der Wellenspitze

Die triviale Lösung $k \equiv 0$ der Gleichung (10.11) erweist sich als instabil. Ein wichtiger Spezialfall, den wir im folgenden näher betrachten werden, sind stationär rotierende Wellen. Für diese Wellen ändert sich die Gestalt der Wellenfront nicht, d.h. v und k hängen nicht von t ab. Die Wellenspitze bewegt sich in diesem Falle auf einer Kreisbahn.

Definition 10.8. Der Radius der Kreisbahn, auf dem sich die durch $\gamma(0, t)$ definierte Spitze

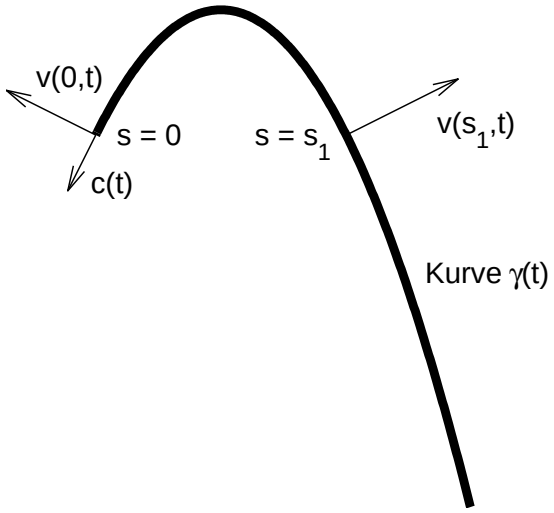


Abbildung 10.17: Kurve γ zum Zeitpunkt t mit Geschwindigkeitskomponenten v und c

einer in der Ebene stationär rotierenden Spiralwelle bewegt, wird als **kinematischer Kernradius** $\varrho_{\mathrm{kin},0}$ bezeichnet.

Bemerkung 10.9. Die Position der Wellenspitze wurde in Abschnitt 10.5 als der Punkt festgelegt, in dem die Aktivatorkonzentration u einen mittleren Wert annimmt und deren Zeitableitung $u_t = 0$ ist. Diese Definition der Wellenspitze weicht mitunter geringfügig von der Wellenspitze im kinematischen Modell ab. Das äußert sich darin, daß der kinematische Kernradius $\varrho_{\mathrm{kin},0}$ *nicht* mit dem in Abschnitt 10.5 definierten numerischen Kernradius $\varrho_{\mathrm{num},0}$ übereinstimmen muß.

Die stationäre Front mit konstanter Normalgeschwindigkeit

Wir betrachten im folgenden den Spezialfall einer stationären und mit konstanter Normalgeschwindigkeit bewegten Front, d.h. es gelte $k(0,t) \equiv k^*$ und $D = 0$, woraus wegen der Bewegungsgleichungen (10.12) $c(t) \equiv 0$ und $v(s,t) \equiv v_0 = \text{const.}$ für alle s und t folgt. Der Einfachheit halber lassen wir daher die Variable t weg, schreiben also $k(s),\ v(s)$ etc. Aus der Grundgleichung (10.11) erhält man in diesem Falle

$$\frac{\partial k}{\partial s}(s) v_0 \int_0^s k(\sigma)\ d\sigma + k(s)^2 v_0 = 0, \tag{10.13}$$

nach Division durch v_0 ergibt sich

$$\frac{\partial k}{\partial s}(s) \int_0^s k(\sigma)\ d\sigma + k(\sigma)^2 = 0.$$

Durch Integration erhält man daraus

$$k(s) \int_0^s k(\sigma)\ d\sigma = \text{const.} =: C. \tag{10.14}$$

Division durch $k(s)$ und anschließende Differentiation nach s liefert die gewöhnliche Differentialgleichung

$$\frac{\partial k}{\partial s}(s) = -k(s)^3/C$$

mit der Lösung

$$k(s) = \sqrt{\frac{C}{2s + CC_2}}.$$

Für eine nichttriviale Lösung gilt $C \neq 0$. In diesem Falle muß jedoch $C_2 = 0$ sein, da aus (10.14) die Aussage $\lim_{s\to 0} k(s) = \infty$ folgt. Demnach ist

$$k(s) = \sqrt{\frac{C}{2s}}, \qquad C > 0$$

die Lösung des Problems (10.13).

Damit entspricht die stationäre Front mit konstanter Normalgeschwindigkeit gerade der **Evolvente eines Kreises** mit dem Radius

$$R_{\mathrm{Ev}} = 1/C, \tag{10.15}$$

ein Sachverhalt, der bereits von Wiener und Rosenblueth [172] gezeigt wurde. Es gilt nämlich das folgende Lemma:

Lemma 10.10 (Evolvente des Kreises). *Die Gleichungen*

$$x = R(\cos\varphi + \varphi\sin\varphi), \qquad y = R(\sin\varphi - \varphi\cos\varphi)$$

beschreiben eine Evolvente des Kreises $x^2 + y^2 = R^2$. Wird diese durch ihre Bogenlänge $s \geq 0$ parametrisiert, wobei $s = 0$ dem Anfang der Kurve bei $x = R,\ y = 0$ entspricht, so ist ihre Krümmung durch $k(s) = 1/\sqrt{2Rs}$ gegeben.

Beweis. Die Evolvente des Kreises ist dessen „Abwickelkurve“, d.h. die Kurve, die das Ende eines straffen Fadens beschreibt, der von dem Kreis abgewickelt wird. Aus dieser geometrischen Deutung erhält man mühelos die angegebene Gleichung einer Evolvente. Die Bogenlänge der Evolvente ist definiert durch

$$s = \int_0^{\varphi} \sqrt{\left(\frac{dx}{d\varphi}(t)\right)^2 + \left(\frac{dy}{d\varphi}(t)\right)^2}\, dt = \frac{R}{2}\varphi^2.$$

Die Krümmung der Evolvente berechnet sich nach der Formel

$$k = \frac{\frac{dx}{d\varphi}\frac{d^2y}{d\varphi^2} - \frac{dy}{d\varphi}\frac{d^2x}{d\varphi^2}}{\left(\left(\frac{dx}{d\varphi}\right)^2 + \left(\frac{dy}{d\varphi}\right)^2\right)^{3/2}} = \frac{1}{R\varphi}.$$

Aus beidem folgt $k(s) = 1/\sqrt{2Rs}$. □

Aus der Beziehung (10.15) folgt insbesondere $C = \omega_0/v_0$, wobei ω_0 die **Winkelgeschwindigkeit** der rotierenden Spirale ist[4]. Die Krümmung genügt daher der Gleichung

$$k(s) = \sqrt{\frac{\omega_0}{2v_0 s}}.$$

Der Radius R_{Ev} ist der kinematische Kernradius der Welle, also der Radius des Kreises, den die bei $s = 0$ angenommene Wellenspitze bei der Rotation beschreibt.

Die stationäre Front mit krümmungsabhängiger Normalgeschwindigkeit

Eine genauere Approximation einer stationären Wellenfront erhält man unter der Annahme, daß die Normalgeschwindigkeit *linear* von der Krümmung abhängt, also in der durch die Bewegungsgleichung (10.12) gegebenen Form

$$v(s) = v_0 - Dk(s), \qquad D > 0$$

vorliegt. Da wir eine stationäre Front betrachten, ist wieder $k(0) = k^*$ und damit $c \equiv 0$. Durch Integration der Grundgleichung (10.11) erhält man die Beziehung

$$k(s) \int_0^s k(\sigma)(v_0 - Dk(\sigma))\, d\sigma - D\frac{\partial k}{\partial s}(s) = \text{const.} = C. \tag{10.16}$$

Die Winkelgeschwindigkeit ω_0 erfüllt, wenn $D \neq 0$ ist, gerade die Beziehung $dv/ds(0) = \omega_0$, siehe etwa [49]. Somit ist $C = \omega_0$, und man kann Gleichung (10.16) auf die Form

$$\int_0^s k(\sigma)(v_0 - Dk(\sigma))\, d\sigma = \frac{\omega_0 + D\frac{\partial k}{\partial s}(s)}{k(s)}$$

bringen. Differenziert man nach s, so ergibt sich die gewöhnliche Differentialgleichung

$$-Dk(s)\frac{\partial^2 k}{\partial s^2}(s) + \left(\omega_0 + D\frac{\partial k}{\partial s}(s)\right)\frac{\partial k}{\partial s}(s) + k(s)^3(v_0 - Dk(s)) = 0. \tag{10.17}$$

Aus (10.16) gewinnen wir die Randbedingung

$$\frac{\partial k}{\partial s}(0) = -\frac{\omega_0}{D},$$

außerdem setzen wir $\lim\limits_{s\to\infty} k(s) = 0$ als weitere Randbedingung.

Falls $Dk(0) \ll v_0$ ist, so spricht man von einem **schwach erregbaren Medium**. In diesem Falle ist (10.17) ein singulär gestörtes Problem. Die Lösung dieses Problems entspricht außerhalb einer Grenzschicht bei $s = 0$ näherungsweise der des Grenzproblems, das sich für $D \to 0$ ergibt:

$$\omega_0\frac{\partial k}{\partial s}(s) + v_0 k(s)^3 = 0, \qquad \lim_{s\to 0} k(s) = \infty.$$

[4]Der Index 0 bei ω_0 dient hier der Unterscheidung von der Winkelgeschwindigkeit einer Welle *auf der Sphäre*, die im nächsten Abschnitt betrachtet und dort mit ω bezeichnet wird. Gleiches gilt für die im folgenden auftretenden Größen $\varrho_{\mathrm{num},0}$, $\varrho_{\mathrm{kin},0}$ und $k_{C,0}$.

Die Lösung des Grenzproblems ist

$$k(s) = \sqrt{\frac{\omega_0}{2v_0 s}}, \tag{10.18}$$

also nach Lemma 10.10 die Gleichung einer Evolvente eines Kreises mit dem Radius $R_{\mathrm{Ev}} = v_0/\omega_0$. Am linken Rand, bei $s = 0$, befindet sich eine Grenzschicht, die durch die dortige Randbedingung erzwungen wird. In Abbildung 10.18 ist für die Parameter $D = 0{,}01$, $v_0 = 1$, $\omega_0 = 1$ die numerisch berechnete Lösung der Differentialgleichung (10.17) zusammen mit der Grenzlösung (10.18) dargestellt.

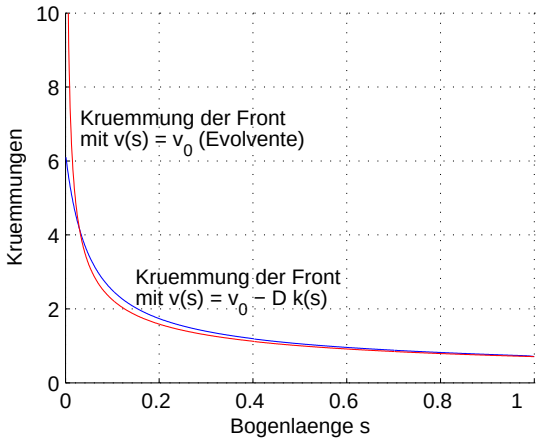


Abbildung 10.18: Vergleich der Krümmung der Grenzlösung ($v(s) \equiv v_0$, Evolvente) mit der Krümmung der Welle für $v(s) = v_0 - Dk(s)$

Der kinematische Kernradius $\varrho_{\mathrm{kin},0}$ der Welle ist definiert durch

$$\varrho_{\mathrm{kin},0} = \frac{v(0)}{\omega_0} = \frac{v_0 - Dk(0)}{\omega_0}.$$

Für schwach erregbare Medien gilt

$$\varrho_{\mathrm{kin},0} \approx R_{\mathrm{Ev}} = v_0/\omega_0. \tag{10.19}$$

Beispiel 10.11. In Abbildung 10.19 ist die u-Komponente der Lösung des in (10.1), (10.3) angegebenen Reaktions-Diffusions-Systems mit den Parametern

$$d_1 = 1, \qquad d_2 = 0, \qquad k_1 = 15{,}3, \qquad k_2 = 151{,}3, \qquad k_f = 1{,}7,$$

$$k_g = 2, \qquad k_\varepsilon = 6, \qquad a = 0{,}1, \qquad \sigma = 0{,}01, \qquad \varepsilon = 0{,}25$$

zur Zeit $t = 60$ dargestellt. Der kleine gelbe Kreis markiert die Bahnkurve der gemäß (10.6), (10.7) definierten Wellenspitze, sein Radius ist $\varrho_{\mathrm{num},0} = 1{,}73$. Die weiße Kurve ist die Evolvente des ebenfalls weiß gezeichneten Kreises. Die Evolvente stimmt sehr gut mit der Welle überein; geringfügige Abweichungen ergeben sich lediglich nahe der Wellenspitze. In diesem Falle ist $R_{\mathrm{Ev}} = 5{,}15$, $v_0 = 1{,}36$ und $\omega_0 = 0{,}26$. Der numerische Kernradius $\varrho_{\mathrm{num},0}$ unterscheidet sich hier deutlich von dem kinematischen Kernradius $\varrho_{\mathrm{kin},0} \approx R_{\mathrm{Ev}}$, siehe dazu Bemerkung 10.9.

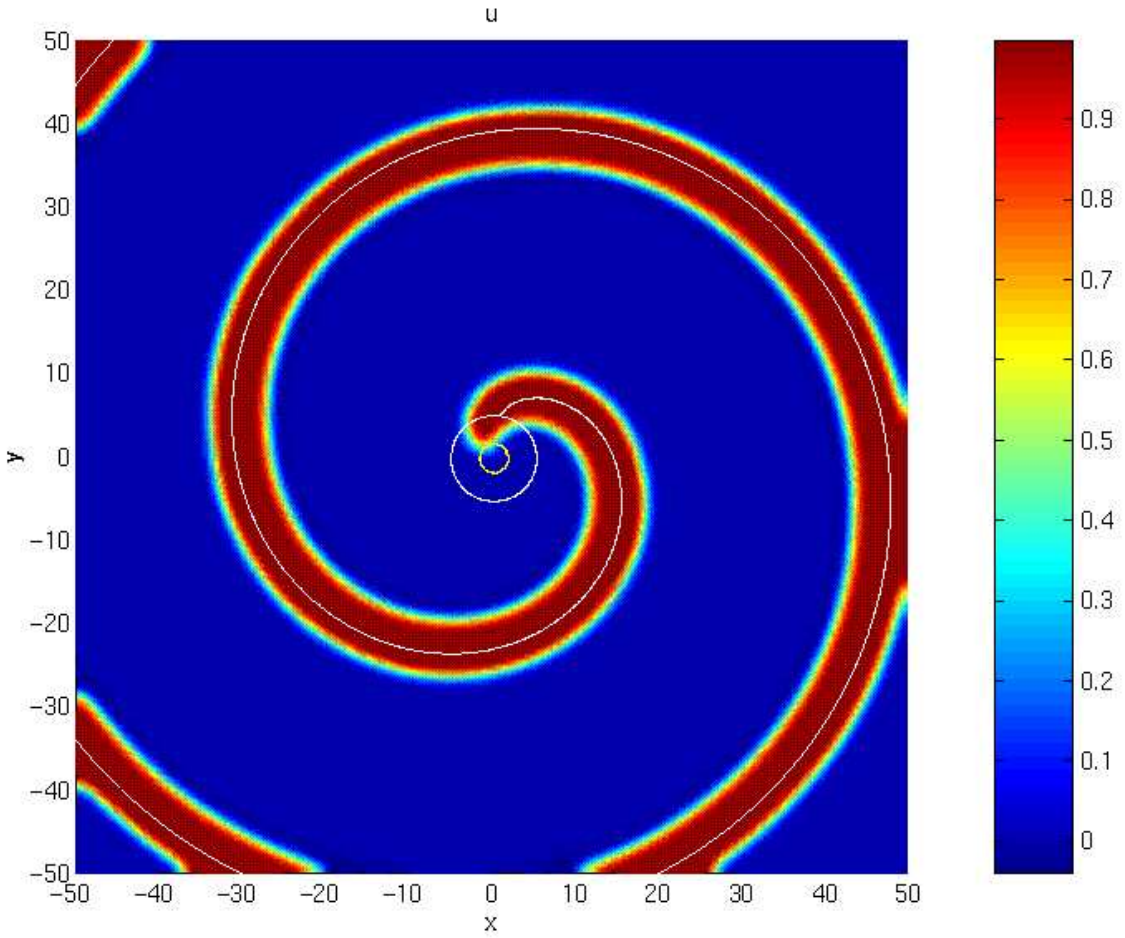


Abbildung 10.19: Spiralwelle, Bahn der Wellenspitze und Evolvente

10.8.2 Stationäre Wellen auf der Sphäre

Auf einer gekrümmten Fläche S geht deren Gaußsche Krümmung Γ in die Grundgleichung der Kinematik ein:

$$\frac{\partial k_g}{\partial s}(s,t)\left(\int_0^s k_g(\sigma,t)v(\sigma,t)\,d\sigma + c(s,t)\right) + \frac{\partial k_g}{\partial t}(s,t) \tag{10.20}$$

$$+ (k_g(s,t))^2 v(s,t) + \frac{\partial^2 v}{\partial^2 s}(s,t) \quad = \quad -\Gamma(s,t)v(s,t), \tag{10.21}$$

eine Beziehung, die von Brazhnik, Davydov und Michailov [29] hergeleitet wurde. Die Größe k_g ist hierbei die *geodätische* Krümmung, die im Anhang, Abschnitt A.3 definiert ist. In den Bewegungsgleichungen (10.12) wird lediglich k durch k_g ersetzt. Man erhält

$$v(s,t) = v_0 - Dk_g(s,t), \qquad c(t) = \beta(k^* - k_g(0,t)) \tag{10.22}$$

Auf einer Sphäre S vom Radius R gilt $\Gamma \equiv 1/R^2$. Wir betrachten nun eine Welle endlicher Ausdehnung mit zwei Wellenspitzen. Für eine stationäre Welle ist wieder $k_g(0,t) \equiv k^*$ und

daher $c \equiv 0$. Außerdem gilt in diesem Falle $\partial k/\partial t \equiv 0$. Wir lassen im stationären Fall wieder die Variable t in $v(s,t)$, $k_g(s,t)$ etc. weg. Durch Integration der Grundgleichung (10.20) erhält man

$$k_g(s) \int_0^s k_g(\sigma)(v_0 - Dk_g(\sigma))\, d\sigma - D\frac{\partial k_g}{\partial s}(s) + \Gamma \int_0^s (v_0 - Dk_g(\sigma))\, d\sigma = \text{const.} =: C. \quad (10.23)$$

Für den Fall eines schwach erregbaren Mediums, $Dk_g(0) \ll v_0$, ergibt sich wie im ebenen Fall ein singulär gestörtes Problem mit Grenzschichten nahe der Wellenspitzen. Außerhalb dieser Grenzschichten entspricht die Lösung von (10.23) näherungsweise der Lösung des Grenzproblems für $D \to 0$. Dieses Grenzproblem ist von der Form

$$k_g(s)v_0 \int_0^s k_g(\sigma)\, d\sigma + \Gamma v_0 s = \text{const.} =: C$$

und hat die beiden Lösungen

$$k_{g1}(s) = \left(\frac{C}{v_0} - \Gamma s\right)\left(\frac{2Cs}{v_0} - \Gamma s^2\right)^{-1/2}, \qquad k_{g2}(s) = |k_{g1}(s)| - \frac{2C}{v_0}\Gamma^{-1/2}\delta\left(s - \frac{C}{v_0\Gamma}\right),$$

siehe [29]. Das Symbol δ steht hierbei für die Delta-Distribution. Die Konstante C steht mit der Rotationsfrequenz ω in der Beziehung

$$C^2 = \omega^2 - v_0^2\Gamma.$$

Die Rotationsfrequenz ω unterscheidet sich von der Rotationsfrequenz ω_0 der entsprechenden ebenen Welle. Von Abramychev, Davydov und Zykov [1] wird die Beziehung

$$\omega = \omega_0 + \frac{v_0^2\Gamma}{2\omega_0} \quad (10.24)$$

angegeben. Die Lösung k_{g1} beschreibt eine antisymmetrische, die Lösung k_{g2} eine symmetrische Wellenfront.

Aus der Darstellung von k_{g1} und k_{g2} lassen sich insbesondere die folgenden Sachverhalte ableiten: Im ebenen Falle $\Gamma = 0$ gehen die Lösungen k_{g1} und k_{g2} in die ebene Lösung (10.18) über. Für kleines s verhalten sich k_{g1} und k_{g2} so, als wäre $\Gamma = 0$. Es gilt daher:

Satz 10.12. *Die geodätische Krümmung einer stationären Wellenfront auf der Kugeloberfläche entspricht in der Nähe der Wellenspitze näherungsweise der Krümmung einer ebenen Welle.*

In Abbildung 10.20 sind die geodätischen Krümmungen k_{g1} und k_{g2} gemeinsam mit der ebenen Krümmung k aus (10.18) graphisch dargestellt.

Die Bahnkurve der Wellenspitze

Die Wellenspitze einer stationär rotierenden Welle γ beschreibt einen Kreis C auf der Sphäre S. Wir betrachten diesen Kreis als eingebettet in den $\mathbb{R}^3$ und bezeichnen seinen Radius, den

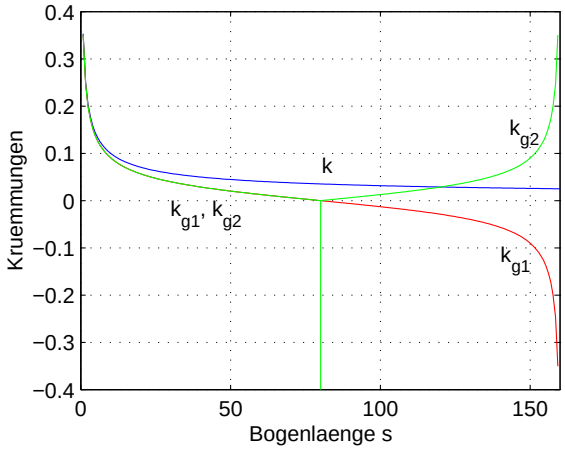


Abbildung 10.20: Vergleich der Krümmung k einer ebenen Welle mit der geodätischen Krümmung einer antisymmetrischen (k_{g1}) und einer symmetrischen (k_{g2}) Welle für die Werte $v_0 = 0{,}1,\ \omega_0 = 0{,}02,\ R = 20$

sogenannten Kernradius der Welle, mit ϱ_{kin}. Wie oben sei $\varrho_{\text{kin},0}$ der kinematische Kernradius der ebenen Welle. Die Winkelgeschwindigkeit der ebenen Welle ist für ein schwach erregbares Medium in guter Näherung durch $\omega_0 = v_0/\varrho_{\text{kin},0}$ gegeben, siehe (10.19). Analog gilt auch für die Winkelgeschwindigkeit ω der Welle auf der Sphäre $\omega = v_0/\varrho_{\text{kin}}$. Zusammen mit (10.24) ergibt sich der kinematische Kernradius zu

$$\varrho_{\text{kin}} = \frac{2R^2\varrho_{\text{kin},0}}{2R^2 + \varrho_{\text{kin},0}^2}. \tag{10.25}$$

Beispiel 10.13. Der numerische Kernradius wurde im ebenen Fall in Beispiel 10.11 berechnet; er betrug $\varrho_{\text{num},0} = 1{,}73$. Simuliert man das gleiche Problem auf einer Sphäre vom Radius $R = 20$, so ergibt sich der numerische Kernradius zu $\varrho_{\text{num}} = 1{,}54$, ist also *kleiner* als $\varrho_{\text{num},0}$. Das korrespondiert zu der Tatsache, daß auch der kinematische Kernradius auf der Sphäre kleiner als in der Ebene ist. □

Als *lokale* Größe der Bahn C der Wellenspitze ist for allem die geodätische Krümmung von C von Interesse. In Abschnitt A.4 im Anhang wird die geodätische Krümmung des Breitenkreises $\vartheta = \text{const.}$ einer Sphäre vom Radius R berechnet, sie beträgt

$$k_g = \frac{\tan\vartheta}{R}.$$

Die Bahnkurve C hat die gleiche Gestalt wie der Breitenkreis $\vartheta = \text{const.}$, wenn $\varrho_{\text{kin}} = R\cos\vartheta$ ist. Daraus folgt die geodätische Krümmung von C:

$$k_{gC} = \frac{1}{R\varrho_{\text{kin}}}\sqrt{R^2 - \varrho_{\text{kin}}^2}.$$

Setzt man ϱ_{kin} aus (10.25) ein, so ergibt sich

$$k_{gC} = \sqrt{\frac{1}{\varrho_{\text{kin},0}^2} + \frac{\varrho_{\text{kin},0}^2}{4R^4}}.$$

Es sei nun $k_{C,0} := 1/\varrho_{\text{kin},0}$ die Krümmung der Bahn der Wellenspitze im ebenen Fall. Mit $\Gamma = 1/R^2$ folgt

$$k_{gC} = k_{C,0}\sqrt{1 + \frac{\Gamma^2}{4k_{C,0}^4}}. \tag{10.26}$$

Aufgrund der Resultate (10.25) und (10.26) gilt der folgende Satz:

Satz 10.14. *Es sei $\alpha := \varrho_{\text{kin},0}/R < 1$. Dann gilt für den kinematischen Kernradius ϱ_{kin} einer Erregungswelle in einem schwach erregbaren Medium auf der Sphäre S mit dem Radius R*

$$\varrho_{\text{kin}} = \varrho_{\text{kin},0}\left(1 - \frac{\alpha^2}{2} + O(\alpha^4)\right).$$

Für die geodätische Krümmung k_{gC} der Bahn der Wellenspitze gilt

$$k_{gC} = k_{C,0}\left(1 + O(\alpha^4)\right).$$

Beweis. Aus der Beziehung (10.25) folgt

$$\varrho_{\text{kin}} = \frac{2R^2\varrho_{\text{kin},0}}{2R^2 + \varrho_{\text{kin},0}^2} = \varrho_{\text{kin},0}\frac{2}{2+\alpha}.$$

Durch Taylor-Entwicklung an der Stelle $\alpha = 0$ ergibt sich

$$\frac{2}{2+\alpha^2} = 1 - \frac{\alpha^2}{2} + O(\alpha^4).$$

Damit erhält man die erste Aussage des Satzes

$$\varrho_{\text{kin}} = \varrho_{\text{kin},0}\left(1 - \frac{\alpha^2}{2} + O(\alpha^4)\right).$$

Wegen $k_{C,0} = 1/\varrho_{\text{kin},0}$ und $\Gamma = 1/R^2$ gilt

$$\alpha = \frac{\varrho_{\text{kin},0}}{R} = \frac{\sqrt{\Gamma}}{k_{C,0}}.$$

Zusammen mit (10.26) folgt

$$k_{gC} = k_{C,0}\sqrt{1 + \frac{\alpha^4}{4}}.$$

Durch Taylor-Entwicklung an der Stelle $\alpha = 0$ erhält man daraus

$$k_{gC} = k_{C,0}\left(1 + O(\alpha^4)\right),$$

die zweite Aussage des Satzes. □

Ist α hinreichend klein, so stellt $k_{gC} \approx k_{C,0}$ eine gute Näherung dar. Die geodätische Krümmung der Bahn der Wellenspitze unterscheidet sich dann also kaum von der entsprechenden Krümmung im ebenen Fall.

10.9 Spiralwellen auf dem Ellipsoid

Spiralwellen, die auf Flächen konstanter Gaußscher Krümmung, etwa der Sphäre, um einen festen Punkt rotieren, zeigen auf nichtgleichmäßig gekrümmten Flächen ein anderes Verhalten; sie driften zusätzlich zu ihrer Rotation in eine bestimmte, von der Krümmung der Fläche abhängige Richtung, so daß die Wellenspitze eine Zykloide beschreibt. Eine Erklärung dieser Wellendrift wurde von Davydov, Zykov, Michailov und Yamaguchi [49, 50] mit Hilfe der kinematischen Theorie formuliert. Die Drift wird in diesen beiden Arbeiten jedoch nur für rotationssymmetrische Flächen angegeben. Wir wollen die in [50] präsentierte Gleichung für die Drift im folgenden kurz wiedergeben.

Wir betrachten ein Reaktions-Diffusions-System der Form (10.1), welches in der Ebene eine nach einer gewissen Zeit stationär rotierende Spiralwelle erzeugt. Die in Abschnitt 10.8 beschriebene kinematische Theorie reduziert diese Welle auf eine Kurve γ, für die die Bewegungsgleichungen

$$v(s,t) = v_0 - Dk(s,t), \qquad c(t) = \beta(k^* - k(0,t)) \tag{10.27}$$

gelten, siehe (10.12). Die Parameter v_0, k^* und β seien, ebenso wie der in Definition 10.8 eingeführte kinematische Kernradius $\varrho_{\mathrm{kin},0}$ der Spiralwelle, experimentell bestimmt worden.

Es sei S eine bezüglich der z-Achse eines kartesischen Koordinatensystems rotationssymmetrische Fläche in $\mathbb{R}^3$. Die Fläche S wird durch eine glatte positive Funktion $r(z)$ bestimmt. Der Punkt $\mathbf{x} = (x, y, z)$ gehört genau dann zu S, wenn die Beziehungen

$$x = r(z)\cos\varphi, \qquad y = r(z)\sin\varphi$$

für beliebiges $\varphi \in]-\pi, \pi]$ erfüllt sind. Wir bezeichnen mit

$$\zeta = \begin{cases} \arctan\frac{r(z)}{z}, & z > 0, \\ \pi/2, & z = 0, \\ \pi + \arctan\frac{r(z)}{z}, & z < 0 \end{cases}$$

den **Polarwinkel** des Punktes $\mathbf{x}$. Auf der Fläche S gelten dann die zu (10.27) analogen Bewegungsgleichungen

$$v(s,t) = v_0 - Dk_g(s,t), \qquad c(t) = \beta(k^* - k_g(0,t))$$

einer Wellenfront. Mit (φ_0, ζ_0) bezeichnen wir die (φ, ζ)-Koordinaten des Rotationszentrums einer rotierenden Spiralwelle auf S. In [50] wird die folgende mit Hilfe der kinematischen Theorie gewonnene Gleichung für die Drift der Welle angegeben.

$$\frac{d\varphi_0}{dt} = \frac{v_0\varrho_{\mathrm{kin},0}}{4\sin\zeta_0}\frac{d\Gamma}{d\zeta}, \qquad \frac{d\zeta_0}{dt} = \frac{\beta k^*\varrho_{\mathrm{kin},0}}{4}\frac{d\Gamma}{d\zeta}. \tag{10.28}$$

Dabei ist Γ die Gaußsche Krümmung der Fläche S im Punkt (φ_0, ζ_0) und $\varrho_{\mathrm{kin},0}$ der in Definition 10.8 eingeführte kinematische Kernradius der rotierenden Welle.

Wir stellen in Abschnitt 10.9.4 die Ergebnisse zur Wellendrift bei einer Vielzahl numerischer Simulationen vor, die mit dem Modell von Krinsky et al. durchgeführt wurden. Für dieses Modell und die verwendeten Parameter wurde auf der Sphäre eine stationäre Rotation

ohne Drift beobachtet. Auf dem Ellipsoid driftet die Wellenspitze. Allerdings steht bei der Verwendung gewisser Parameter die Richtung der Drift im Widerspruch zu der in (10.28) angegebenen Beziehung. In den Abschnitten 10.9.3 und 10.9.4 wird näher auf diese Problematik eingegangen. Versuche mit dem Oregonator-Modell schlugen hier fehl, da für eine Vielzahl verwendeter Parameter auf der Kugel niemals eine *stationär* rotierende Welle erzeugt werden konnte. Der angesprochene Widerspruch numerisch gewonnener Resultate zu den Aussagen der kinematischen Theorie wurde bereits von Davydov, Zykov und Yamaguchi [50] beobachtet und konnte bisher nicht geklärt werden.

10.9.1 Halbellipsoide als Rechengebiet

Es sei $\mathcal{E}$ das Ellipsoid

$$\frac{x^2}{A^2} + \frac{y^2}{B^2} + \frac{z^2}{C^2} = 1$$

mit den Halbachsenlängen $A, B, C > 0$ und $\mathcal{S}$ die durch $x^2 + y^2 + z^2 = 1$ gegebene Einheitssphäre. Das Ellipsoid $\mathcal{E}$ kann gemäß

$$x = A\cos\varphi\cos\vartheta, \qquad y = B\sin\varphi\cos\vartheta, \qquad z = C\sin\vartheta$$

parametrisiert werden. Die Funktion

$$K(x, y, z) = (x/A, y/B, z/C)$$

bildet $\mathcal{E}$ auf $\mathcal{S}$ ab. Für $\mathbf{a} \in \mathcal{S}$ ist die in (4.20) definierte stereographische Projektion $\widetilde{P}_{\mathbf{a}}$ eine Abbildung von $\mathcal{S}$ nach $\mathbb{R}^2$. Wir definieren damit eine Abbildung $\Phi_{\mathbf{a}} : \mathcal{E} \to \mathbb{R}^2$ gemäß

$$\Phi_{\mathbf{a}}(x, y, z) = \widetilde{P}_{\mathbf{a}}(K(x, y, z)).$$

Es sei $\mathcal{K} = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \le 4\}$. Die Menge $\widetilde{P}_{\mathbf{a}}^{-1}(\mathcal{K})$ ist dann eine Halbsphäre auf $\mathcal{S}$. Folglich ist $\Phi_{\mathbf{a}}^{-1}(\mathcal{K})$ ein Gebiet auf dem Ellipsoid $\mathcal{E}$, das dessen halbe Oberfläche einnimmt. Wir bezeichnen dieses Gebiet als **Halbellipsoid** $\mathcal{E}_{\mathbf{a}}$.

Für die folgenden numerischen Berechnungen benutzen wir stets ein derartiges Halbellipsoid $\mathcal{E}_{\mathbf{a}}$ als Rechengebiet. Durch verschiedene Wahl des Vektors $\mathbf{a}$ kann das Rechengebiet unterschiedlich auf $\mathcal{E}$ plaziert werden. Wir wählen im folgenden den Vektor $\mathbf{a}$ stets so, daß er in der (x, z)-Ebene liegt; er ist dann von der Form

$$\mathbf{a} = (\sin\alpha, 0, -\cos\alpha). \tag{10.29}$$

Der Winkel α ist der Winkel zwischen $\mathbf{a}$ und der negativen z-Achse.

10.9.2 Die Lösung des Systems von Krinsky et al. auf einem Ellipsoid

Untersuchung 10.15. Wir betrachten das System von Krinsky et al. (10.9) mit den Parametern

$$d_1 = 1, \qquad d_2 = 0, \qquad k_1 = 15{,}3, \qquad k_2 = 151{,}3, \qquad k_f = 1{,}7, \qquad k_g = 2, \qquad k_\varepsilon = 6,$$

$$a = 0{,}1, \qquad \sigma = 0{,}01, \qquad \varepsilon = 0{,}3.$$

Es sei $\Omega = S = \mathcal{E} = \{(x,y,z) \in \mathbb{R}^3 : x^2/20^2 + y^2/15^2 + z^2/30^2 = 1\}$. Die Anfangsbedingungen seien symmetrisch zur (x,y)-Ebene gewählt. Die Lösung behält dann für alle Zeiten t diese Symmetrie. Deshalb reicht es aus, die numerische Simulation auf dem Halbellipsoid $\Omega_1 = \mathcal{E}_{-\mathbf{k}}$ durchzuführen[5] und die Lösung entsprechend symmetrisch fortzusetzen. Wir geben die Anfangsbedingungen in den Bildkoordinaten $(X,Y) = \Phi_{-\mathbf{k}}(x,y,z)$ an. Sie lauten

$$u(\Phi_{-\mathbf{k}}(x,y,z)) = u(X,Y) = \begin{cases} 1, & X > 0{,}1, Y \in [-0{,}3, 0] \\ 0, & \text{sonst} \end{cases},$$

$$v(\Phi_{-\mathbf{k}}(x,y,z)) = v(X,Y) = \begin{cases} -5Y, & Y \in [-0{,}3, 0] \\ 0, & Y > 0 \\ 1{,}5, & Y < -0{,}3 \end{cases}.$$

Wie in Abschnitt 3.3 nachgewiesen wurde, entsprechen die benötigten symmetrischen Randbedingungen gerade den homogenen Neumannschen Randbedingungen

$$\frac{\partial u}{\partial \mathbf{n}_{\Omega_1}} = \frac{\partial v}{\partial \mathbf{n}_{\Omega_1}} = 0.$$

Die Ortsdiskretisierung erfolgt mit linearen finiten Elementen, siehe Abschnitt 3.2.2. Das Gitter wird wie in Abschnitt 4.5.2 beschrieben erzeugt. Zur Zeitdiskretisierung verwenden wir hier das in Kapitel 7 dargestellte Krylov-W-Verfahren. Abbildung 10.21 zeigt die v-Komponente der Lösung zu den Zeitpunkten $t = 5,\ 15,\ 20,\ 45,\ 55,\ 60$. □

10.9.3 Abhängigkeit der Wellendrift von den Parametern d_2 und ε

Untersuchung 10.16. Wie sich in numerischen Untersuchungen zeigt, haben beim Modell von Krinsky et al. offenbar die Parameter d_2 und ε einen starken Einfluß auf Richtung und Geschwindigkeit der Wellendrift. Um das zu verdeutlichen, stellen wir an dieser Stelle drei Simulationen gegenüber. Wir verwenden ein Ellipsoid $\mathcal{E}$ mit den Halbachsen $A = B = 20$, $C = 30$. Für alle drei Berechnungen wird das Gebiet $\Omega = \mathcal{E}_{\mathbf{a}}$ mit $\mathbf{a} = (\sin\alpha, 0, -\cos\alpha)$, $\alpha = -0{,}55$ zugrundegelegt. Es werden die Parameter

$$d_1 = 1, \qquad k_1 = 15{,}3, \qquad k_2 = 151{,}3, \qquad k_f = 1{,}7, \qquad k_g = 2, \qquad k_\varepsilon = 6,$$

$$a = 0{,}1, \qquad \sigma = 0{,}01$$

verwendet. In den Bildkoordinaten $(X,Y) = \Phi_{\mathbf{a}}$ lauten die Anfangsbedingungen

$$u(\Phi(x,y,z)) = u(X,Y) = \begin{cases} 1, & X > 0{,}1, Y \in [-0{,}8, 0], \\ 0, & \text{sonst}, \end{cases}$$

$$v(\Phi(x,y,z)) = v(X,Y) = \begin{cases} -1{,}875Y, & Y \in [-0{,}8, 0], \\ 0, & Y > 0, \\ 1{,}5, & Y < -0{,}8. \end{cases}$$

[5] Der Vektor $\mathbf{k} = (0,0,1)$ ist hierbei der Einheitsvektor in z-Richtung.

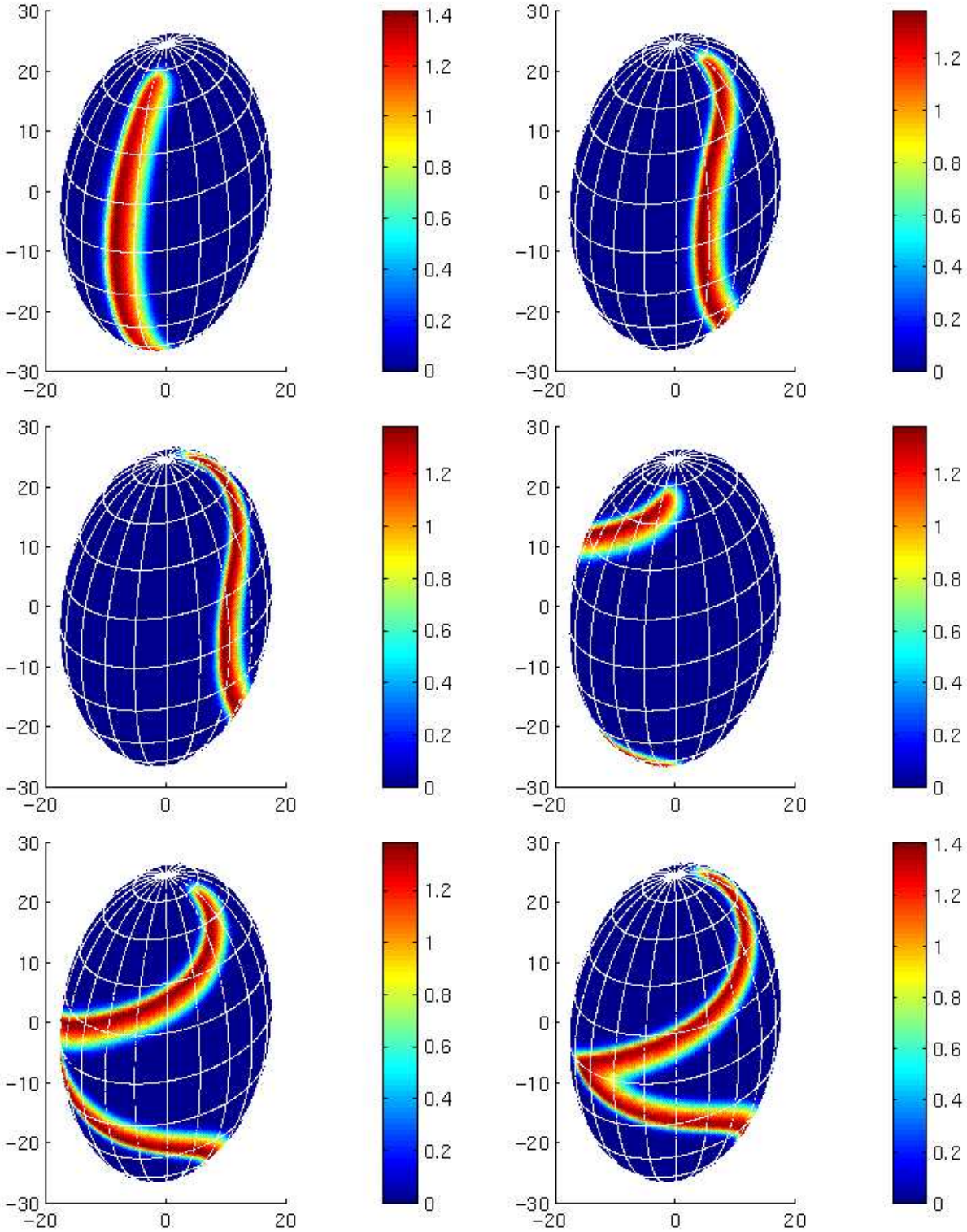


Abbildung 10.21: Lösung des Systems von KRINSKY et al. auf einem Ellipsoid zu den Zeitpunkten $t = 5, 15, 20, 45, 55, 60$

Auf dem Rand von Ω verwenden wir homogene Neumannsche Randbedingungen

$$\frac{\partial u}{\partial \mathbf{n}_\Omega} = \frac{\partial v}{\partial \mathbf{n}_\Omega} = 0.$$

Die drei Rechnungen unterscheiden sich in der Wahl der Parameter d_2 und ε. Wir setzen

- im Fall (1): $d_2 = 1$, $\varepsilon = 0{,}15$,
- im Fall (2): $d_2 = 0{,}5$, $\varepsilon = 0{,}2$ und
- im Fall (3): $d_2 = 0$, $\varepsilon = 0{,}25$.

Ergebnis. In Abbildung 10.22 sind die Bahnkurven der Wellenspitzen in diesen drei Fällen dargestellt. Im Fall (1) (blau in der Abbildung) ist die Drift nach „Westen“, im Fall (2) (rot in der Abbildung) nach „Südosten“ und im Fall (3) (grün in der Abbildung) nach „Osten“ gerichtet. Wir wollen die beobachtete Wellendrift mit der theoretischen Voraussage (10.28)

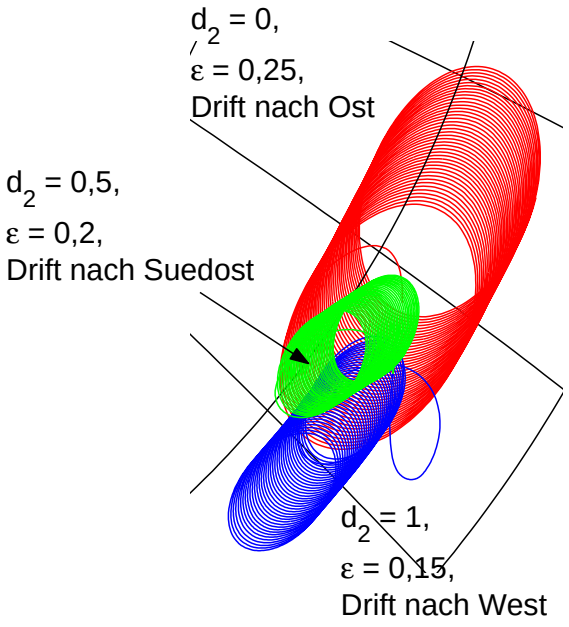


Abbildung 10.22: Bahnkurve der Wellenspitze für unterschiedliche Werte von d_2 und ε

vergleichen. In dieser Gleichung ist $v_0 > 0$, $\varrho_{\mathrm{kin},0} > 0$ und $k^* > 0$. Da sich die Wellenspitze auf der nördlichen Hälfte des Ellipsoids befindet, ist $d\Gamma/d\zeta < 0$. Im Fall (1) ist $\beta = 0$, siehe [49], in den anderen beiden Fällen $\beta > 0$. Demnach sagt die Theorie im Fall (1) eine Drift nach „Westen“ und in den Fällen (2) und (3) nach „Nordwesten“ voraus.

Eine – zumindest qualitative – Übereinstimmung herrscht also nur im Fall (1)! In den Fällen (2) und (3) ist die von uns beobachtete Richtung der Drift der theoretischen Voraussage geradezu entgegengesetzt. □

Dieser Widerspruch konnte bislang nicht geklärt werden. Interessant ist, daß die erwähnte Diskrepanz zwischen numerisch beobachteter und theoretisch berechneter Drift bereits in der

Arbeit von Davydov, Zykov und Yamaguchi [50] zutage tritt. Auch dort wurde für ein ähnliches Beispiel eine „westliche“ Wellendrift berechnet, während die numerische Simulation eine „östliche“ Drift ergab.

10.9.4 Abhängigkeit der Wellendrift von der Gaußschen Krümmung

Für den Fall, daß $d_2 = 0$ ist, wollen wir in diesem Abschnitt die Abhängigkeit der Wellendrift von der Gaußschen Krümmung der Fläche untersuchen. Die in diesem Abschnitt vorgestellten numerischen Berechnungen beziehen sich sämtlich auf das Modell von Krinsky et al.. Das Gebiet Ω ist in allen Fällen ein Halbellipsoid $\mathcal{E}_{\mathbf{a}}$ mit $\mathbf{a} = (\sin\alpha, 0, -\cos\alpha)$, siehe (10.29). Es werden – abgesehen von den im Anschluß erwähnten Ausnahmen – die in Abschnitt 10.9.2 angegebenen Parameter, Anfangs- und Randbedingungen verwendet.

Die folgenden Größen werden variiert und daher stets angegeben:

- die Halbachsenlängen A, B und C,
- der in (10.29) definierte Winkel α, der die Lage von Ω bestimmt,
- der Erregungsparameter ε.

Bemerkung 10.17. Falls $A \neq C$ und $\alpha \neq k\pi/2$, $k \in \mathbb{Z}$ ist, so ist das genannte Problem *nicht* symmetrisch auf ganz $\mathcal{E}$ fortsetzbar, da in diesem Falle $\partial\Omega$ keine Kurve ist, die $\mathcal{E}$ in zwei spiegelsymmetrische Hälften zerlegt[6]. □

Die Drift einer Spiralwelle kann aus der Bahnkurve $C(t)$, die die Wellenspitze beschreibt, gewonnen werden, wobei die Wellenspitze wie in (10.6), (10.7) angegeben definiert wird. Eine driftende Welle beschreibt eine Bahnkurve in Form einer Zykloide. Bei jeder Umdrehung der Wellenspitze wird das aktuelle Rotationszentrum $\mathbf{x}_0(t)$ bestimmt. Die Bahn dieser Mittelpunkte beschreibt die Drift der Welle.

Wir vergleichen die Bahnkurve der Wellenspitze auf einer Sphäre mit der auf einem Ellipsoid. Abbildung 10.23 zeigt den Unterschied. Während die Wellenspitze auf der Sphäre eine Kreisbahn beschreibt, liegt auf dem Ellipsoid zusätzlich eine Driftbewegung vor. Es wurde eine Sphäre mit dem Radius $A = B = C = 20$ und ein Ellipsoid mit den Halbachsen $A = 15$, $B = 20$ und $C = 30$ verwendet, $\varepsilon = 0{,}25$ und $\alpha = -0{,}4$ gesetzt.

Untersuchung 10.18. Die von Davydov, Zykov und Yamaguchi [50] angegebene Beziehung (10.28) beschreibt auf rotationssymmetrischen Flächen eine Abhängigkeit der Drift von der Ableitung der Gaußschen Krümmung $\partial\Gamma/\partial\zeta$. Wir wollen zunächst dieses Resultat der kinematischen Theorie mit den Ergebnissen vergleichen, die wir aus der numerischen Berechnung des Modells von Krinsky et al. auf einem Rotationsellipsoid mit den Halbachsen $A = B = 20,\ C = 30$ erhalten. Wir setzen $\varepsilon = 0{,}2$. Durch Variation des in (10.29) definierten Winkels α wird die Wellenspitze an verschiedenen Stellen des Ellipsoids plaziert. Wir führen Berechnungen mit $\alpha = 0, -0{,}1, -0{,}2, \ldots, -1, -1{,}5$ durch. In Abbildung 10.24 stellen wir die Größen $\sin\zeta_0\ d\varphi_0/dt$ und $d\zeta_0/dt$ über $d\Gamma/d\zeta$ dar.

[6] vgl. Abschnitt 3.3.

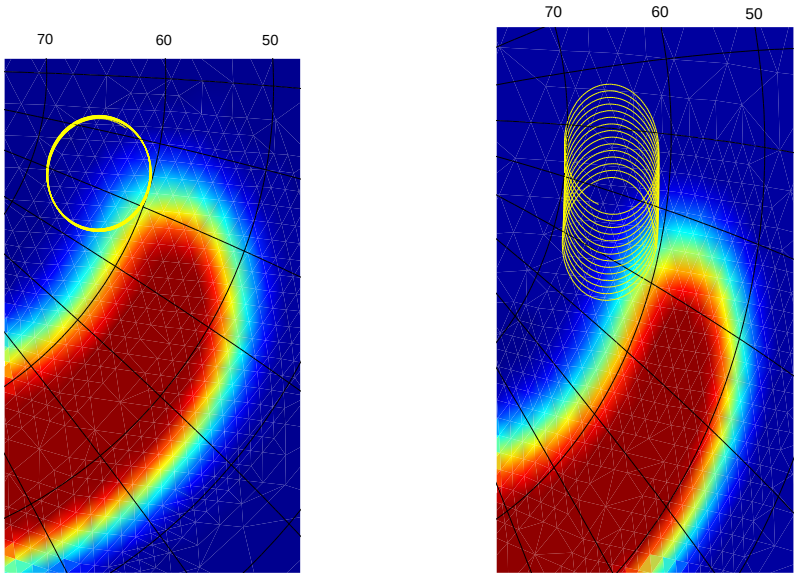


Abbildung 10.23: Bahnkurve der Wellenspitze, **links:** auf einer Sphäre vom Radius $R = 20$, **rechts:** auf einem Ellipsoid mit Halbachsen $A = 15$, $B = 20$, $C = 30$. Die Drift bei dem Ellipsoid geht in „östliche“ Richtung, d.h. auf dem Bild nach oben.

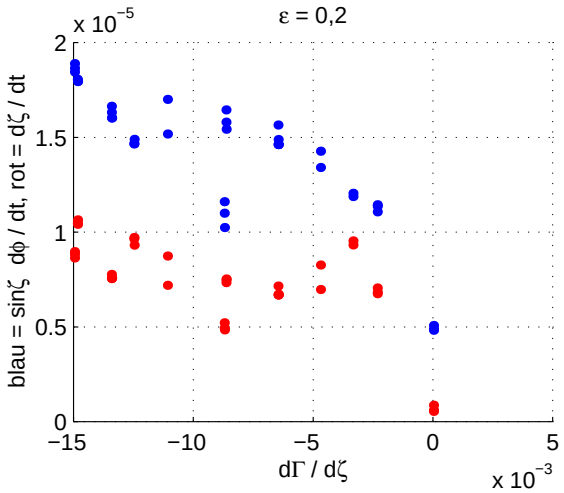


Abbildung 10.24: Die Größen $\sin\zeta_0\ d\varphi_0/dt$ und $d\zeta_0/dt$ in Abhängigkeit von $d\Gamma/d\zeta$

Ergebnis. Die kinematische Theorie sagt für beide Größen eine *lineare* Abhängigkeit von $d\Gamma/d\zeta$ voraus, siehe Gleichung (10.28). Eine solche lineare Beziehung wird durch die in Abbildung 10.24 dargestellten numerische Ergebnisse jedoch nicht bestätigt. Außerdem erwartet man nach der kinematischen Theorie, daß $\sin\zeta_0\ d\varphi_0/dt < 0$ und $d\zeta_0/dt < 0$ gelten, da in allen Fällen die Wellenspitze auf der nördlichen Hälfte des Ellipsoids liegt, also $d\Gamma/d\zeta < 0$ ist. Damit stimmt auch das Vorzeichen der gemessenen Drift nicht mit der kinematischen Theorie überein, ein Umstand, der schon bei den Berechnungen in Abschnitt 10.9.3 festgestellt wurde. □

Trotzdem zeigt sich, daß die Gaußsche Krümmung und deren Ableitung sehr wohl einen Einfluß auf die Driftgeschwindigkeit ausüben. Wir wollen diesen Zusammenhang im folgenden in einer Reihe von Beispielrechnungen auf verschiedenen Ellipsoiden $\mathcal{E}$ näher beleuchten. Um von einer Parametrisierung unabhängig zu sein, betrachten wir die Driftgeschwindigkeit $d\mathbf{x}_0/dt$ in den Koordinaten des $\mathbb{R}^3$ und beschreiben deren Abhängigkeit von der Gaußschen Krümmung Γ und deren tangentialem Gradienten $\nabla_S\Gamma$ auf der Fläche $S = \mathcal{E}$.

Untersuchung 10.19. Wir führen numerische Berechnungen auf Ellipsoiden mit verschiedenen Halbachsenlängen A, B, C durch und variieren ferner den Erregungsparameter ε und den Winkel α, der die Anfangslage der Welle auf dem Ellipsoid bestimmt. In der folgende Tabelle sind die Parameter, für die numerische Experimente durchgeführt werden, mit • gekennzeichnet.

A	15				15				20	20	
B	20				20				15	20	
C	30				25				30	30	
ε	0,2	0,21	0,22	0,25	0,2	0,21	0,22	0,25	0,25	0,2	0,25
$\alpha =$											
0	•	•	•	•	•	•	•	–	•	•	•
−0,1	•	•	•	•	•	•	•	–	•	•	–
−0,2	•	•	•	•	•	•	•	•	•	•	–
−0,3	•	•	•	•	•	•	•	–	•	•	–
−0,4	•	•	•	•	•	•	•	•	•	•	–
−0,5	•	•	•	•	•	•	•	–	•	•	•
−0,6	•	•	•	•	•	•	•	•	–	•	–
−0,7	•	•	•	•	•	•	•	–	–	•	–
−0,8	•	•	•	•	•	•	•	•	–	•	–
−0,9	•	•	•	•	•	•	•	–	–	•	–
−1	•	•	•	•	•	•	•	–	•	•	•
−1,5	•	•	•	•	•	•	•	–	•	•	•

Wir stellen die Driftgeschwindigkeit $d\mathbf{x}_0/dt$ als Linearkombination der Einheits-Tangentialvektoren $\mathbf{e}_1 = \nabla_S\Gamma/|\nabla_S\Gamma|$ und $\mathbf{e}_2 = \nabla_S\Gamma \times \mathbf{n}_S/|\nabla_S\Gamma|$ dar, siehe Abbildung 10.25.

Ergebnis. Die numerischen Untersuchungen ergeben näherungsweise eine Beziehung der Form

$$\frac{d\mathbf{x}_0}{dt} = K|\nabla_S\Gamma|\mathbf{e}_1 + L\frac{|\nabla_S\Gamma|}{\sqrt{\Gamma}}\mathbf{e}_2, \tag{10.30}$$

wobei die Größen K und L Konstanten sind, die von dem Erregungsparameter ε, aber nicht von der Geometrie der Fläche S abhängen, also nicht von A, B, C oder α. Um die genannte Beziehung zu verdeutlichen, sind in den Abbildungen 10.26 und 10.27 die beiden Geschwindigkeitskomponenten $v_1 = (d\mathbf{x}_0/dt)^T\mathbf{e}_1$ und $v_2 = (d\mathbf{x}_0/dt)^T\mathbf{e}_2$ über $|\nabla_S\Gamma|$ bzw. $|\nabla_S\Gamma|/\sqrt{\Gamma}$ aufgetragen.

Die Beziehung (10.30), die auch in der Form

$$\frac{d\mathbf{x}_0}{dt} = K\nabla_S\Gamma + L\frac{\nabla_S\Gamma \times \mathbf{n}_S}{\sqrt{\Gamma}} \tag{10.31}$$

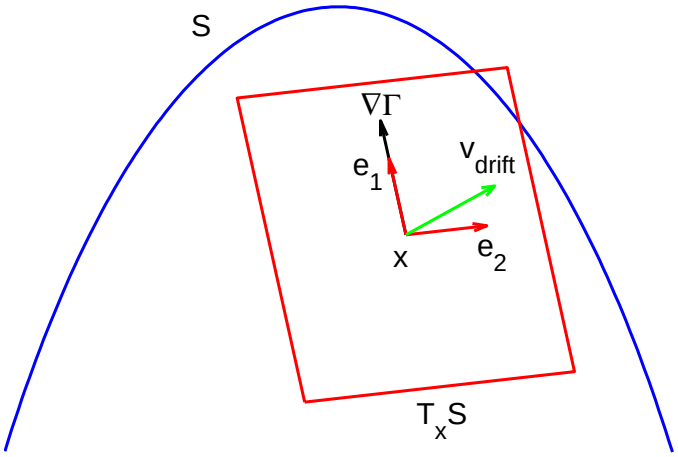


Abbildung 10.25: Tangentialebene, Einheitsvektoren $\mathbf{e}_1$ und $\mathbf{e}_2$, Driftgeschwindigkeit $v_{\text{drift}} = d\mathbf{x}_0/dt$

geschrieben werden kann, ist jedoch nur eine Vermutung, die sich auf die hier untersuchten numerischen Testbeispiele stützt. Theoretische Aussagen, etwa eine Präzisierung der kinematischen Theorie, wären wünschenswert, liegen aber zur Zeit nicht vor.

Ausgehend von den in den Abbildungen 10.26 und 10.27 dargestellten quantitativen Untersuchungen der Wellendrift können die Konstanten K und L in Abhängigkeit von ε näherungsweise bestimmt werden. Sie sind in der folgenden Tabelle aufgeführt.

ε	K	L
0,2	$-1{,}2$	0,12
0,21	$-0{,}98$	0,15
0,22	$-0{,}74$	0,20
0,25	0,0	0,59

Aus der Differentialgleichung (10.31) lassen sich die Bahnkurven bestimmen, die der driftende Mittelpunkt des Wellenkernes auf dem Ellipsoid beschreibt. Für $\varepsilon = 0{,}2, A = 20, B = 15, C = 30$ sind in Abbildung 10.28 diese Bahnkurven dargestellt. Abbildung 10.29 zeigt den Betrag der Driftgeschwindigkeit $|d\mathbf{x}_0/dt|$.

Die Differentialgleichung (10.31) hat in diesem Falle zwei stabile Ruhelagen in den Punkten $\varphi = \pm\pi/2, \vartheta = 0$, zwei instabile Ruhelagen in Nord- und Südpol, d.h. für $\vartheta = \pm\pi/2$ und zwei Sattelpunkte bei $\varphi = 0, \vartheta = 0$ und $\varphi = \pi, \vartheta = 0$. Anders als in dem von Davydov, Zykov und Yamaguchi in [50] beschriebenen Beispiel, dem das Oregonator-Modell zugrunde liegt, bewegt sich in unserem Falle die driftende Welle von den Punkten maximaler Gaußscher Krümmung weg und hin zu den Punkten minimaler Gaußscher Krümmung. □

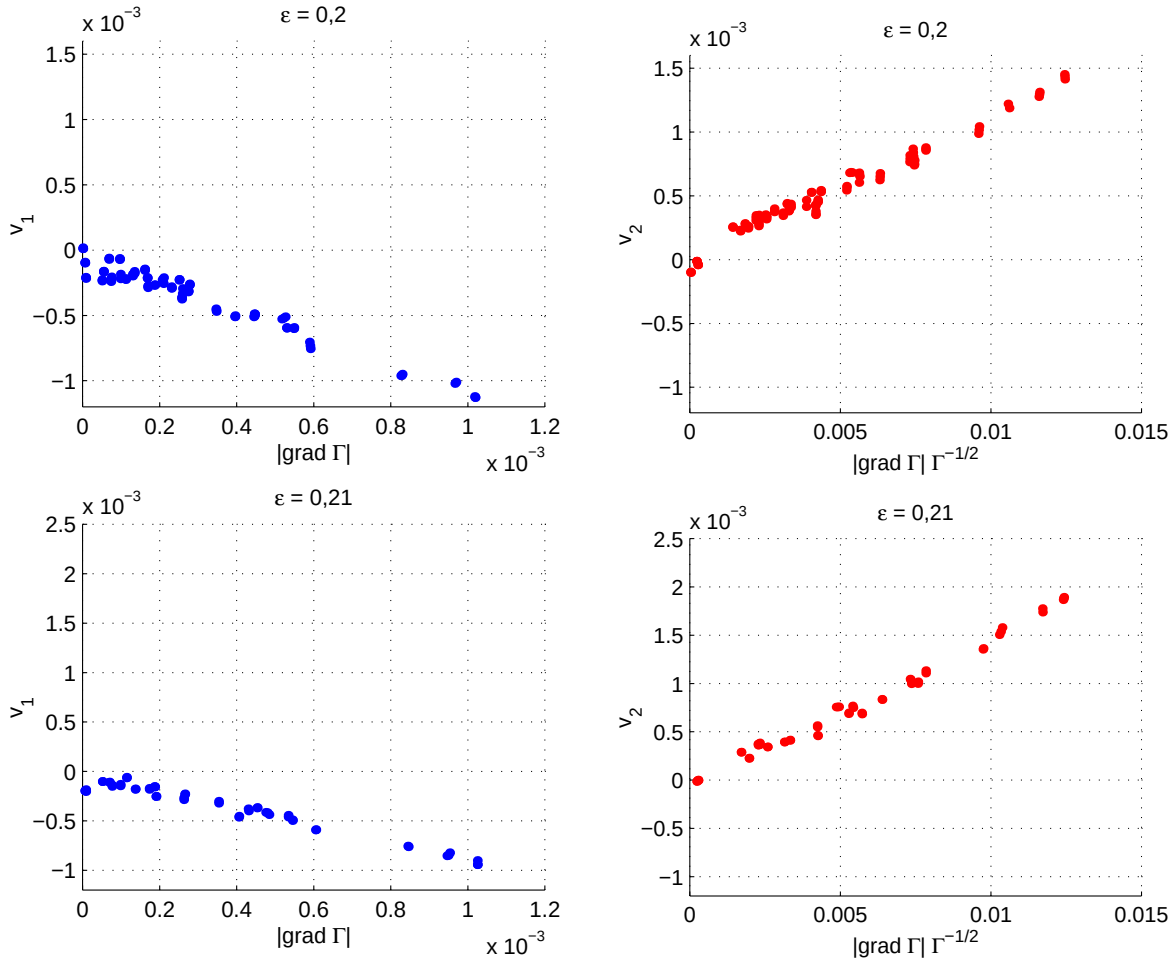


Abbildung 10.26: Lineare Abhängigkeit der Driftgeschwindigkeits-Komponenten v_1 und v_2 von $|\nabla_S \Gamma|$ bzw. $|\nabla_S \Gamma|/\sqrt{\Gamma}$ für $\varepsilon = 0{,}2$ und $\varepsilon = 0{,}21$

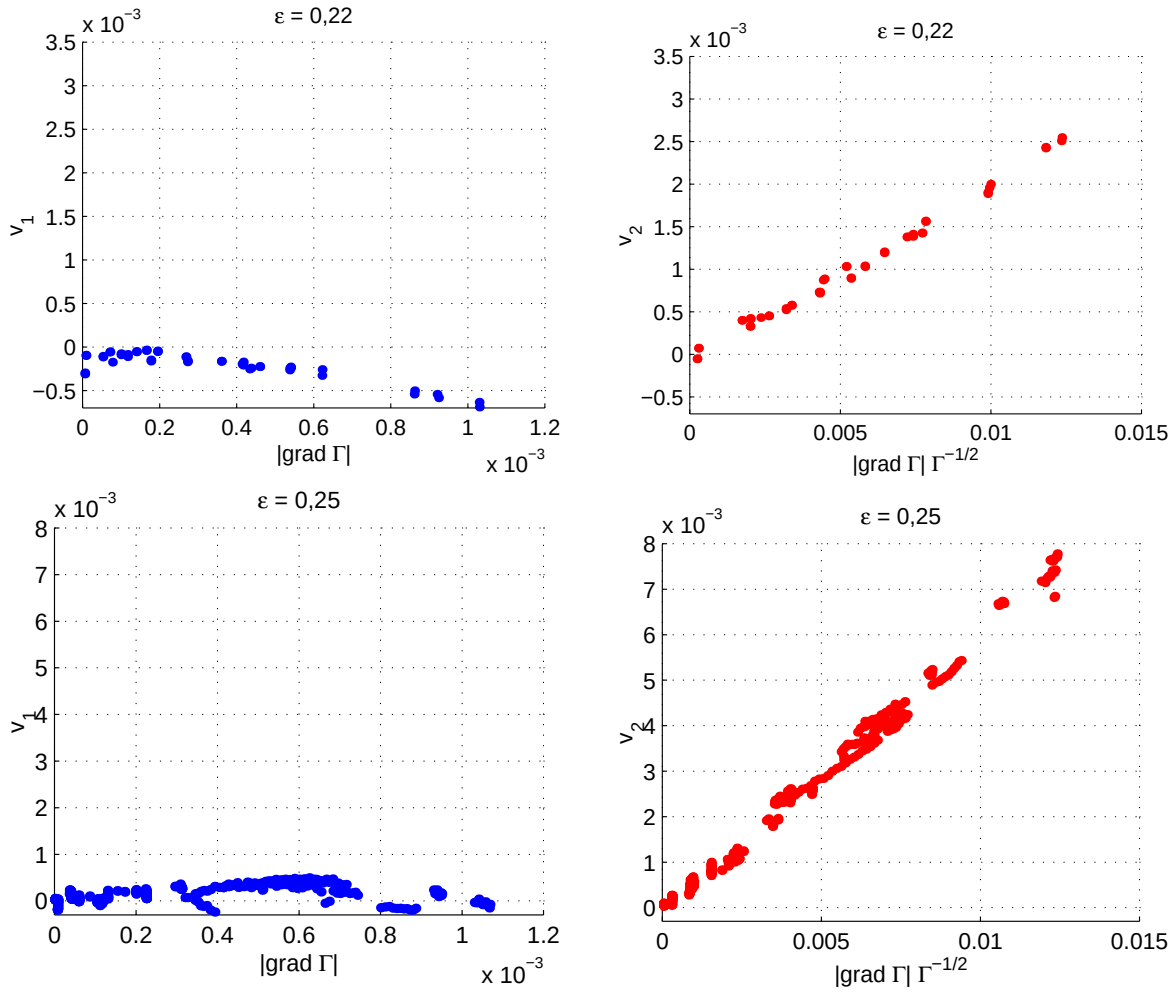


Abbildung 10.27: Lineare Abhängigkeit der Driftgeschwindigkeits-Komponenten v_1 und v_2 von $|\nabla_S \Gamma|$ bzw. $|\nabla_S \Gamma|/\sqrt{\Gamma}$ für $\varepsilon = 0{,}22$ und $\varepsilon = 0{,}25$

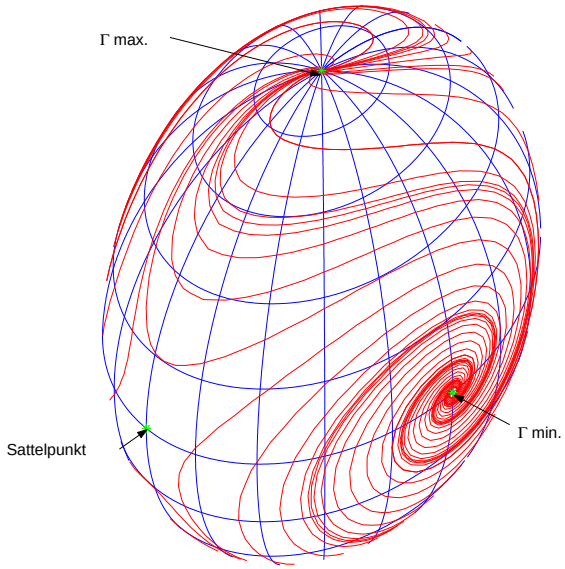


Abbildung 10.28: Bahnkurven der Wellendrift auf dem Ellipsoid $A = 20, B = 15, C = 30$ bei $\varepsilon = 0{,}2$. Die Drift ist von den Polen ($\vartheta = \pm\pi/2$) zu den Punkten $\varphi = \pm\pi/2, \vartheta = 0$ gerichtet.

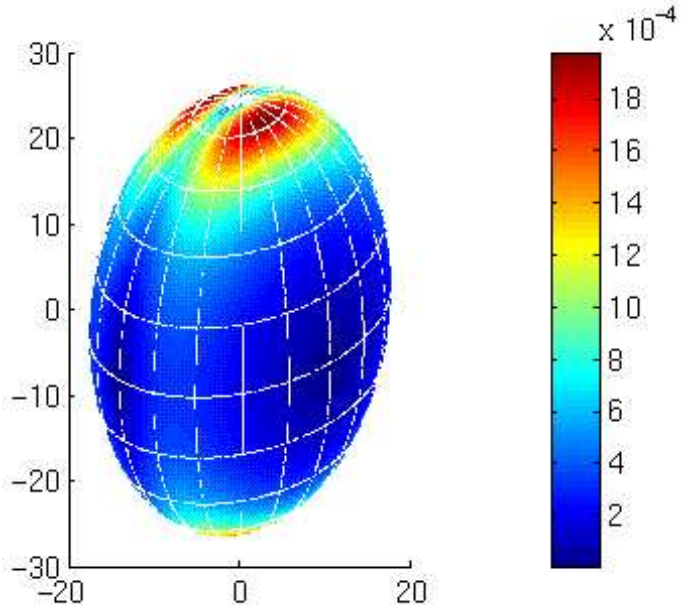


Abbildung 10.29: Geschwindigkeit der Wellendrift auf dem Ellipsoid $A = 20, B = 15, C = 30$ bei $\varepsilon = 0{,}2$

Zusammenfassung und Ausblick

Bei der Lösung vieler Reaktions-Diffusions-Probleme sind Diskretisierungsfehler und Steifheit lokal von unterschiedlicher Größenordnung. Adaptive Verfahren passen sich dieser Situation an und erreichen damit oftmals einen Zuwachs an Effizienz. In Kapitel 4 haben wir die Adaption des Gitters in Abhängigkeit der Verteilung des räumlichen Diskretisierungsfehlers beschrieben. In Abschnitt 4.3 wurde ein flexibler Algorithmus zur Steuerung der Gitteradaption angegeben, der über das oftmals verfolgte Ziel einer Gleichverteilung des Fehlers über das Rechengebiet hinausgeht. Mit diesem Algorithmus ist es möglich, jedem Wert des Fehlers eine gewünschte Gitterfeinheit zuzuordnen. Wie an einem numerischen Beispiel nachgewiesen wurde, läßt sich auf diese Weise die Effizienz mitunter deutlich erhöhen.

Ein wesentliches Ziel der vorliegenden Arbeit war die in Kapitel 8 beschriebene Konstruktion geeigneter Partitionierungs-Verfahren für lokal steife semidiskrete Probleme. Die zur Zeitdiskretisierung eingesetzten W-Methoden sind in einfacher Weise durch eine Manipulation der darin auftretenden Matrix zur Partitionierung geeignet. Mit dem Blick auf Reaktions-Diffusions-Probleme wurden Verfahren entwickelt, die Reaktions- und Diffusions-Term getrennt partitionieren. Somit wurde auf Probleme eingegangen, bei denen die Steifheit nur von einem dieser beiden Terme ausgeht.

Zur Lösung der im Ergebnis der W-Methode vorliegenden linearen Gleichungssysteme wurden zwei iterative Verfahren betrachtet: das BiCGstab-Verfahren von van der Vorst [166] und der multiple Arnoldi-Prozeß von Schmitt und Weiner [143], der aufgrund der schnellen Approximation des dominanten Eigenraumes für lokal steife Probleme besonders geeignet erscheint.

Im Ergebnis erhielten wir eine Reihe numerischer Lösungsverfahren, die in einem umfangreichen Test an drei ausgewählten Reaktions-Diffusions-Problemen bezüglich ihrer Effizienz miteinander verglichen wurden. Die Resultate wurden in Kapitel 9 dokumentiert. Die verwendeten Probleme waren in unterschiedlichem Maße zur Partitionierung geeignet. Besonders bei dem BSVD-Problem, siehe 9.1.2, konnten erhebliche Einsparungen durch lokale Partitionierung erzielt werden. Die numerische Untersuchung in Kapitel 9 beschränkte sich allerdings auf zwei Reaktions-Diffusions-Gleichungen und ein zweikomponentiges Reaktions-Diffusions-System. Bei diesen Problemen ist eine vollständig implizite Lösung des Reaktions-Anteils kaum aufwendiger – aber oftmals genauer – als eine Partitionierung. Es wäre daher interessant, die konstruierten Partitionierungs-Verfahren auch an mehrkomponentigen Systemen zu untersuchen, bei denen eine Partitionierung steifer gekoppelter Reaktionsterme stärkere Einsparungen erwarten läßt.

Die in Kapitel 10 dargestellten numerischen Untersuchungen zur Drift von Erregungswellen

auf gekrümmten Flächen liefern interessante Ergebnisse, die im Widerspruch zu gewissen Aussagen der kinematischen Theorie in [50] stehen, ein Umstand, der noch endgültiger Klärung bedarf. Es wäre wünschenswert, die Anwendbarkeit der vorliegenden kinematischen Theorie auf das von uns betrachtete Modellproblem zu prüfen und die Theorie so zu erweitern, daß der in den numerischen Berechnungen beobachtete Zusammenhang (10.31) zwischen der Gaußschen Krümmung der Fläche und der Driftgeschwindigkeit gestützt wird. Wie sich in der numerischen Untersuchung zeigt, ist die Richtung der Drift stark von der Diffusion des Inhibitors abhängig, siehe Abschnitt 10.9.3. Es wäre interessant, diesen Zusammenhang weiter zu untersuchen. Ein Vergleich der in der Simulation beobachteten Wellendrift auf Ellipsoiden mit der Drift auf anderen Flächen – auch solchen negativer Gaußscher Krümmung – wäre ebenfalls eine gewinnbringende Fortsetzung der in dieser Arbeit durchgeführten Untersuchungen.

Anhang A

Einige Grundbegriffe aus der Differentialgeometrie

Im folgenden geben wir einige differentialgeometrische Grundlagen an, die in dieser Arbeit benötigt werden. Die gewählte Bezeichnung orientiert sich weitestgehend an dem Buch von doCarmo [37], welches eine sehr anschauliche Darstellung der Grundlagen der Differentialgeometrie enthält. Die hier angegebenen Definitionen und Aussagen finden sich beispielsweise in den Lehrbüchern von Jost [90] und Bär [13].

Im folgenden sei S eine differenzierbare zweidimensionale Mannigfaltigkeit, eingebettet im $\mathbb{R}^3$. Auf S liege eine Parametrisierung

$$\mathbf{x} : \mathbb{R}^2 \to S \tag{A.1}$$

vor. Die Vektoren

$$\frac{\partial \mathbf{x}}{\partial \varphi} = \begin{pmatrix} \partial x/\partial\varphi(\mathbf{p}) \\ \partial y/\partial\varphi(\mathbf{p}) \\ \partial z/\partial\varphi(\mathbf{p}) \end{pmatrix} \qquad \text{und} \qquad \frac{\partial \mathbf{x}}{\partial \vartheta} = \begin{pmatrix} \partial x/\partial\vartheta(\mathbf{p}) \\ \partial y/\partial\vartheta(\mathbf{p}) \\ \partial z/\partial\vartheta(\mathbf{p}) \end{pmatrix}$$

bilden eine lokale Basis des im Punkt $\mathbf{p} \in S$ angehängten Tangentialraumes $T_{\mathbf{p}}S$. Auf $T_{\mathbf{p}}S$ wird durch das Euklidische Skalarprodukt im $\mathbb{R}^3$ ein lokales Skalarprodukt $\langle \cdot, \cdot \rangle_{\mathbf{p}} : S \times S \to \mathbb{R}$ induziert, man definiert einfach $\langle \mathbf{v}, \mathbf{w} \rangle_{\mathbf{p}} = \mathbf{v} \cdot \mathbf{w}$. Die zugehörige quadratische Form $I_{\mathbf{p}}(\mathbf{v}) = \langle \mathbf{v}, \mathbf{v} \rangle_{\mathbf{p}}$ bezeichnet man als die **erste Fundamentalform** oder **Riemannsche Metrik** der Fläche S. Die Skalarprodukte

$$E = \frac{\partial \mathbf{x}}{\partial \varphi} \cdot \frac{\partial \mathbf{x}}{\partial \varphi}, \qquad F = \frac{\partial \mathbf{x}}{\partial \varphi} \cdot \frac{\partial \mathbf{x}}{\partial \vartheta} \qquad \text{und} \qquad G = \frac{\partial \mathbf{x}}{\partial \vartheta} \cdot \frac{\partial \mathbf{x}}{\partial \vartheta}$$

werden Koeffizienten der ersten Fundamentalform genannt.

Bemerkung A.1. Die Größen E, F und G sind die Komponenten des metrischen Tensors, der oft mit g bezeichnet wird; und zwar ist $g_{11} = E$, $g_{12} = g_{21} = F$, $g_{22} = G$. Der Einfachheit halber verzichten wir in dieser Arbeit auf die Tensornotation, die in der Literatur oft verwendet wird. □

Der **Normalvektor** der Fläche ist durch

$$\mathbf{n}_S = \frac{\partial \mathbf{x}/\partial\varphi \times \partial \mathbf{x}/\partial\vartheta}{|\partial \mathbf{x}/\partial\varphi \times \partial \mathbf{x}/\partial\vartheta|}$$

gegeben. Er ist – unabhängig von der Parametrisierung – bis auf sein Vorzeichen eindeutig bestimmt.

Ist S eine Fläche mit Rand, so existiert in jedem Randpunkt $\mathbf{p}$ der **äußere Normalvektor an den Rand** $\mathbf{n}_{\partial S}$. Dieser Vektor liegt im Tangentialraum $T_{\mathbf{p}}$, steht senkrecht auf dem Rande ∂S und ist von der Fläche nach außen gerichtet.

A.1 Gradient, Divergenz und Laplace-Beltrami-Operator

Die aus dem $\mathbb{R}^n$ bekannten Operatoren Gradient, Divergenz und Laplace-Operator können auch auf gekrümmte Flächen übertragen werden. Man definiert nämlich

- für eine Funktion $u \in C^1(S, \mathbb{R})$ den **tangentialen Gradienten**

$$\nabla_S\, u = \frac{1}{EG - F^2} \left(\left(G \frac{\partial u}{\partial \varphi} - F \frac{\partial u}{\partial \vartheta} \right) \frac{\partial \mathbf{x}}{\partial \varphi} + \left(-F \frac{\partial u}{\partial \varphi} + E \frac{\partial u}{\partial \vartheta} \right) \frac{\partial \mathbf{x}}{\partial \vartheta} \right),$$

- für ein tangentiales Vektorfeld $\mathbf{w} = u\partial\mathbf{x}/\partial\varphi + v\partial\mathbf{x}/\partial\vartheta$ mit $u, v \in C^1(S, \mathbb{R})$ die **tangentiale Divergenz**

$$\operatorname{div}_S \left(u \frac{\partial \mathbf{x}}{\partial \varphi} + v \frac{\partial \mathbf{x}}{\partial \vartheta} \right) = \frac{1}{\sqrt{EG - F^2}} \left(\frac{\partial}{\partial \varphi} \left(\sqrt{EG - F^2}\; u \right) + \frac{\partial}{\partial \vartheta} \left(\sqrt{EG - F^2}\; v \right) \right),$$

- für eine Funktion $u \in C^2(S, \mathbb{R})$ den **Laplace-Beltrami-Operator**

$$\begin{aligned} \Delta_S\, u = \operatorname{div}_S\, (\nabla_S\, u) \;\; = \;\; & \frac{1}{\sqrt{EG - F^2}} \left(\frac{\partial}{\partial \varphi} \left(\frac{1}{\sqrt{EG - F^2}} \left(G \frac{\partial u}{\partial \varphi} - F \frac{\partial u}{\partial \vartheta} \right) \right) \right. \\ & \left. + \frac{\partial}{\partial \vartheta} \left(\frac{1}{\sqrt{EG - F^2}} \left(-F \frac{\partial u}{\partial \varphi} + E \frac{\partial u}{\partial \vartheta} \right) \right) \right). \end{aligned} \tag{A.2}$$

Lemma A.2. *Für glatte Funktionen u und v gilt auf S die* **Greensche Formel**

$$\int_S (\Delta_S\, u)\; v\; dA = -\int_S \nabla_S\, u \cdot \nabla_S\, v\; dA + \int_{\partial S} \frac{\partial u}{\partial \mathbf{n}_{\partial S}} v\; ds.$$

Für eine Funktion, die nicht nur auf S, sondern in einer offenen Umgebung von S im $\mathbb{R}^3$ gegeben ist, kann der tangentiale Gradient mit Hilfe des $\mathbb{R}^3$-Gradienten ∇u ausgedrückt werden.

Lemma A.3. *Sei*

$$\nabla u = \begin{pmatrix} \partial u/\partial x \\ \partial u/\partial y \\ \partial u/\partial z \end{pmatrix}$$

der $\mathbb{R}^3$-Gradient von u. Es gilt die Beziehung

$$\nabla_S\, u = \nabla u - (\nabla u \cdot \mathbf{n}_S)\mathbf{n}_S, \tag{A.3}$$

d.h. $\nabla_S\, u$ ist gerade der tangentiale Anteil des $\mathbb{R}^3$-Gradienten von u.

Beweis. Mit $\partial u/\partial\varphi = \nabla u \cdot \partial\mathbf{x}/\partial\varphi$ und $\partial u/\partial\vartheta = \nabla u \cdot \partial\mathbf{x}/\partial\vartheta$ folgt

$$\nabla_S\, u = \begin{pmatrix} \frac{\partial\mathbf{x}}{\partial\varphi} & \frac{\partial\mathbf{x}}{\partial\vartheta} \end{pmatrix} \begin{pmatrix} E & F \\ F & G \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial\mathbf{x}}{\partial\varphi} & \frac{\partial\mathbf{x}}{\partial\vartheta} \end{pmatrix}^T \nabla u.$$

Nach Definition ist

$$\begin{pmatrix} E & F \\ F & G \end{pmatrix} = \begin{pmatrix} \frac{\partial\mathbf{x}}{\partial\varphi} & \frac{\partial\mathbf{x}}{\partial\vartheta} \end{pmatrix}^T \begin{pmatrix} \frac{\partial\mathbf{x}}{\partial\varphi} & \frac{\partial\mathbf{x}}{\partial\vartheta} \end{pmatrix}.$$

Setzt man $D = \begin{pmatrix} \frac{\partial\mathbf{x}}{\partial\varphi} & \frac{\partial\mathbf{x}}{\partial\vartheta} \end{pmatrix}$, so folgt $\nabla_S\, u = D(D^TD)^{-1}D^T\nabla u$. Durch Ausrechnen erhält man jedoch $D(D^TD)^{-1}D^T = I - \mathbf{n}_S\mathbf{n}_S^T$. Es folgt

$$\nabla_S\, u = \nabla u - (\nabla u \cdot \mathbf{n}_S)\mathbf{n}_S.$$

□

Bemerkung A.4. Lemma A.3 kann benutzt werden, um den Gradienten $\nabla_S\, u$ einer Funktion zu berechnen, ohne den Weg über die Parametrisierung zu gehen. Es reicht aus, die Funktion *u auf beliebige Weise* zu einer glatten Funktion im $\mathbb{R}^3$ fortzusetzen. □

Bemerkung A.5. Der Normalvektor $\mathbf{n}_S$ ist nur bis auf das Vorzeichen eindeutig bestimmt, der tangentiale Gradient hängt jedoch nicht vom Vorzeichen von $\mathbf{n}_S$ ab. □

Eine ähnliche Beziehung wie für den Gradienten liegt für die Divergenz for. Sie wird hier ohne Beweis angegeben.

Lemma A.6. *Es seien n_1, n_2 und n_3 die Komponenten des Flächennormalvektors $\mathbf{n}_S$. Die tangentiale Divergenz eines tangentialen Vektorfeldes $\mathbf{w} = u\ \partial\mathbf{x}/\partial\varphi + v\ \partial\mathbf{x}/\partial\vartheta$ erfüllt die Gleichung*

$$\operatorname{div}_S\, \mathbf{w} = \operatorname{div}\ \mathbf{w} - \sum_{i=1}^{3} (\nabla(w_i) \cdot \mathbf{n}_S)\, n_i. \tag{A.4}$$

A.2 Die Gaußsche Krümmung

Um die Krümmung einer Fläche S zu definieren führen wir zunächst die **Koeffizienten der zweiten Fundamentalform**

$$\begin{aligned} e &= -\frac{\partial\mathbf{n}_S}{\partial\varphi} \cdot \frac{\partial\mathbf{x}}{\partial\varphi} && = \mathbf{n}_S \cdot \frac{\partial^2\mathbf{x}}{\partial\varphi^2} \\ f &= -\frac{\partial\mathbf{n}_S}{\partial\vartheta} \cdot \frac{\partial\mathbf{x}}{\partial\varphi} = -\frac{\partial\mathbf{n}_S}{\partial\varphi} \cdot \frac{\partial\mathbf{x}}{\partial\vartheta} && = \mathbf{n}_S \cdot \frac{\partial^2\mathbf{x}}{\partial\varphi\partial\vartheta} \\ g &= -\frac{\partial\mathbf{n}_S}{\partial\vartheta} \cdot \frac{\partial\mathbf{x}}{\partial\vartheta} && = \mathbf{n}_S \cdot \frac{\partial^2\mathbf{x}}{\partial\vartheta\partial\vartheta} \end{aligned}$$

ein.

Bemerkung A.7. Die Koeffizienten e, f, g werden in der Literatur mitunter auch mit L, M, N bezeichnet. □

Die **Gaußsche Krümmung** einer Fläche S ist dann durch

$$\Gamma = \frac{eg - f^2}{EG - F^2}$$

gegeben.

Betrachtet man eine in den $\mathbb{R}^3$ eingebettete Fläche S, so besitzt die Gaußsche Krümmung die folgende geometrische Bedeutung:

In einem Punkte $\mathbf{p} \in S$ sei $\mathbf{n}_S$ der Flächennormalvektor. $\mathcal{E}$ sei das Büschel aller Ebenen, in denen $\mathbf{n}_S$ liegt. Die Schnittkurven der Ebenen aus $\mathcal{E}$ mit S werden Normalschnitte genannt; unter diesen Schnittkurven existiert eine mit maximaler Krümmung $k_{\max}$ und eine mit minimaler Krümmung $k_{\min}$ bei $\mathbf{p}$. Die extremalen Krümmungen $k_{\max}$ und $k_{\min}$ werden als Hauptkrümmungen bezeichnet, ihr Produkt bildet gerade die Gaußsche Krümmung der Fläche S bei $\mathbf{p}$:

$$\Gamma = k_{\max} k_{\min}.$$

Insbesondere hat eine Sphäre vom Radius R die Gaußsche Krümmung $\Gamma \equiv 1/R^2 = \text{const.}$. Zylinder und Kegel sind Beispiele für Flächen mit verschwindender Gaußscher Krümmung.

A.3 Die geodätische Krümmung einer auf einer Fläche gelegenen Kurve

Wir betrachten wieder die durch die Parametrisierung (A.1) gegebene Fläche S. Auf dieser Fläche befinde sich der Graph einer glatten Kurve $\mathbf{c}(t)$, die in der Form $\mathbf{c}(t) = \mathbf{x}(\varphi(t), \vartheta(t))$, $t \in [t_0, t_e]$ durch die beiden Funktionen φ und ϑ eindeutig bestimmt ist. Wenn der Parameter t gerade die Bogenlänge der Kurve $\mathbf{c}$ ist, so bezeichnen wir ihn mit s. In diesem Falle gilt $\|d\mathbf{c}/ds\| \equiv 1$.

Wird die Kurve $\mathbf{c}(t)$ in einem Punkte $\mathbf{x} = \mathbf{c}(t^*)$ orthogonal auf die Tangentialebene $T_{\mathbf{x}}S$ projiziert, so erhält man eine Bildkurve $\mathbf{c}^*(t) \subset T_{\mathbf{x}}S$. Die Krümmung dieser ebenen Kurve $\mathbf{c}^*$ in Punkte $\mathbf{x}$ wird als **geodätische Krümmung** k_g von $\mathbf{c}$ im Punkte $\mathbf{x}$ bezeichnet.

Wir bezeichnen mit $\mathbf{n}_S$ die Flächennormale an S in einem Punkte $\mathbf{x}$ und mit

$$\mathbf{t} := d\mathbf{c}/ds$$

den **Tangentenvektor** von $\mathbf{c}$ in $\mathbf{x}$. Dann ist die geodätische Krümmung gerade das Spatprodukt der Vektoren $\mathbf{t}$, $d\mathbf{t}/ds$ und $\mathbf{n}_S$, d.h. es gilt

$$k_g = (\mathbf{t} \times d\mathbf{t}/ds) \cdot \mathbf{n}_S.$$

Die geodätische Krümmung kann auch direkt aus den Parametrisierungen von Fläche S und Kurve $\mathbf{c}$ gewonnen werden. Zunächst definieren wir die **Christoffel-Symbole erster Art** Γ_{ijk}:

$$\Gamma_{111} = \frac{\partial^2 \mathbf{x}}{\partial \varphi^2} \cdot \frac{\partial \mathbf{x}}{\partial \varphi}, \qquad \Gamma_{112} = \frac{\partial^2 \mathbf{x}}{\partial \varphi^2} \cdot \frac{\partial \mathbf{x}}{\partial \vartheta}, \qquad \Gamma_{121} = \frac{\partial^2 \mathbf{x}}{\partial \varphi \partial \vartheta} \cdot \frac{\partial \mathbf{x}}{\partial \varphi}, \qquad \Gamma_{122} = \frac{\partial^2 \mathbf{x}}{\partial \varphi \partial \vartheta} \cdot \frac{\partial \mathbf{x}}{\partial \vartheta},$$

$$\Gamma_{211} = \frac{\partial^2 \mathbf{x}}{\partial\vartheta\partial\varphi} \cdot \frac{\partial \mathbf{x}}{\partial\varphi}, \qquad \Gamma_{212} = \frac{\partial^2 \mathbf{x}}{\partial\vartheta\partial\varphi} \cdot \frac{\partial \mathbf{x}}{\partial\vartheta}, \qquad \Gamma_{221} = \frac{\partial^2 \mathbf{x}}{\partial\vartheta^2} \cdot \frac{\partial \mathbf{x}}{\partial\varphi}, \qquad \Gamma_{222} = \frac{\partial^2 \mathbf{x}}{\partial\vartheta^2} \cdot \frac{\partial \mathbf{x}}{\partial\vartheta}.$$

Durch die Beziehung

$$\begin{pmatrix} \Gamma^1_{11} & \Gamma^2_{11} \\ \Gamma^1_{12} & \Gamma^2_{12} \\ \Gamma^1_{21} & \Gamma^2_{21} \\ \Gamma^1_{22} & \Gamma^2_{22} \end{pmatrix} = \frac{1}{EG - F^2} \begin{pmatrix} \Gamma_{111} & \Gamma_{112} \\ \Gamma_{121} & \Gamma_{122} \\ \Gamma_{211} & \Gamma_{212} \\ \Gamma_{221} & \Gamma_{222} \end{pmatrix} \begin{pmatrix} G & -F \\ -F & E \end{pmatrix}$$

berechnen sich aus diesen die **Christoffel-Symbole zweiter Art** Γ^k_{ij}. Die geodätische Krümmung ergibt sich dann zu

$$\begin{aligned} k_g = \sqrt{EG - F^2} \Bigg(& \Gamma^2_{11} \left(\frac{d\varphi}{ds}\right)^3 + (2\Gamma^2_{12} - \Gamma^1_{11}) \left(\frac{d\varphi}{ds}\right)^2 \frac{d\vartheta}{ds} \\ & -(2\Gamma^1_{12} - \Gamma^2_{22}) \frac{d\varphi}{ds} \left(\frac{d\vartheta}{ds}\right)^2 - \Gamma^1_{22} (\frac{d\vartheta}{ds})^3 + \frac{d\varphi}{ds} \frac{d^2\vartheta}{ds^2} - \frac{d^2\varphi}{ds^2} \frac{d\vartheta}{ds} \Bigg). \end{aligned}$$

Dabei sind $d\varphi/ds$, $d\vartheta/ds$, $d^2\varphi/ds^2$ und $d^2\vartheta/ds^2$ die ersten und zweiten Ableitungen der die Kurve $\mathbf{c}$ definierenden Funktionen φ und ϑ nach der Bogenlänge s. Ein Beweis dieser Beziehung ist beispielsweise bei KREYSZIG [97] angegeben.

A.4 Ein Beispiel

Wir wollen die erwähnten Begriffe an dem einfachen Beispiel einer Sphäre mit Radius R verdeutlichen. Als Parametrisierung wählen wir die geographische Länge φ und die geographische Breite ϑ:

$$\begin{aligned} x &= R\cos\varphi\cos\vartheta, \\ y &= R\sin\varphi\cos\vartheta, \\ z &= R\sin\vartheta \end{aligned} \tag{A.5}$$

Bemerkung A.8. In der Literatur wird häufig anstelle der geographischen Breite ϑ der Polarwinkel $\pi/2 - \vartheta$ zur Parametrisierung verwendet. Wir benutzen in dieser Arbeit jedoch immer die genannten geographischen Koordinaten. □

Wir erhalten

$$\frac{\partial \mathbf{x}}{\partial\varphi} = \begin{pmatrix} -R\sin\varphi\cos\vartheta \\ R\cos\varphi\cos\vartheta \\ 0 \end{pmatrix}, \qquad \frac{\partial \mathbf{x}}{\partial\vartheta} = \begin{pmatrix} -R\cos\varphi\sin\vartheta \\ -R\sin\varphi\sin\vartheta \\ R\cos\vartheta \end{pmatrix}$$

und damit die Koeffizienten der ersten Fundamentalform

$$E = R^2\cos^2\vartheta, \qquad F = 0, \qquad G = R^2.$$

Der in (A.2) definierte Laplace-Beltrami-Operator hat dann die Form

$$\Delta_S\, u = \frac{1}{R^2\cos^2\vartheta} \frac{\partial^2 u}{\partial\varphi^2} - \frac{\tan\vartheta}{R^2} \frac{\partial u}{\partial\vartheta} + \frac{1}{R^2} \frac{\partial^2 u}{\partial\vartheta^2}.$$

Wir betrachten als Beispiel die Funktion

$$u(x,y,z) = x + y + z = R(\cos\varphi + \sin\varphi)\cos\vartheta + R\sin\vartheta$$

und berechnen ihren Laplace-Beltrami-Operator:

$$\Delta_S\ u = -\frac{2}{R}((\cos\varphi + \sin\varphi)\cos\vartheta + \sin\vartheta) = -\frac{2}{R^2}(x+y+z)$$

Im folgenden soll der Laplace-Beltrami-Operator auf einem anderen Wege ermittelt werden, nämlich ohne die Parametrisierung der Fläche S zu verwenden. Der Normalvektor von S ergibt sich zu $\mathbf{n}_S = (x,y,z)^T/R$. Nach der Formel (A.3) erhalten wir erhalten den tangentialen Gradienten

$$\nabla_S\ u = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \frac{x+y+z}{R^2} \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

Dieser ist zunächst nur auf S definiert, wir setzen ihn jedoch in natürlicher Weise in den gesamten $\mathbb{R}^3$ fort. Nach Gleichung (A.4) berechnen wir von diesem Ausdruck die tangentiale Divergenz und erhalten ebenfalls

$$\Delta_S\ u = \mathrm{div}_S(\nabla_S\ u) = -\frac{2}{R^2}(x+y+z).$$

Zur Berechnung der Gaußschen Krümmung der Sphäre S benötigen wir zunächst die Koeffizienten der zweiten Fundamentalform

$$\begin{aligned}
e = \mathbf{n_S} \cdot \frac{\partial^2 \mathbf{x}}{\partial\varphi^2} &= \frac{1}{R}\begin{pmatrix} x \\ y \\ z \end{pmatrix} \cdot R \begin{pmatrix} -\cos\varphi\cos\vartheta \\ -\sin\varphi\cos\vartheta \\ 0 \end{pmatrix} = -R + \frac{z^2}{R} \\
f = \mathbf{n_S} \cdot \frac{\partial^2 \mathbf{x}}{\partial\varphi\partial\vartheta} &= \frac{1}{R}\begin{pmatrix} x \\ y \\ z \end{pmatrix} \cdot R \begin{pmatrix} -\sin\varphi\sin\vartheta \\ -\cos\varphi\sin\vartheta \\ 0 \end{pmatrix} = 0 \\
g = \mathbf{n_S} \cdot \frac{\partial^2 \mathbf{x}}{\partial\vartheta^2} &= \frac{1}{R}\begin{pmatrix} x \\ y \\ z \end{pmatrix} \cdot R \begin{pmatrix} -\cos\varphi\cos\vartheta \\ -\sin\varphi\cos\vartheta \\ -\sin\vartheta \end{pmatrix} = -R
\end{aligned}$$

Damit ergibt sich die Gaußsche Krümmung zu

$$\Gamma = \frac{eg - f^2}{EG - F^2} = \frac{R^2 - z^2}{R^4\cos^2\vartheta} = \frac{1}{R^2}.$$

Als Beispiel zur Berechnung der geodätischen Krümmung betrachten wir einen Breitenkreis der Kugel, gegeben durch die Beziehung $\vartheta = \text{const.}$. Ein Punkt $\mathbf{x}$ auf diesem Breitenkreis hat die in (A.5) angegebenen Koordinaten. Der Normalvektor der Sphäre in diesem Punkte ist $\mathbf{n}_S = \mathbf{x}/R$. Der normierte Tangentenvektor des Breitenkreises ist mit

$$\mathbf{t} = \frac{\partial\mathbf{x}/\partial\varphi}{|\partial\mathbf{x}/\partial\varphi|} = \begin{pmatrix} -\sin\varphi \\ \cos\varphi \\ 0 \end{pmatrix}$$

gegeben. Für das Bogenelement gilt $ds = R\cos\vartheta\ d\varphi$. Wir erhalten damit

$$\frac{d\mathbf{t}}{ds} = \frac{1}{R\cos\vartheta}\frac{d\mathbf{t}}{d\varphi} = \frac{1}{R\cos\vartheta}\begin{pmatrix}-\cos\varphi\\-\sin\varphi\\0\end{pmatrix}.$$

Wegen

$$\mathbf{n}_S \times \mathbf{t} = \frac{\partial\mathbf{x}/\partial\vartheta}{|\partial\mathbf{x}/\partial\vartheta|} = \begin{pmatrix}-\cos\varphi\sin\vartheta\\-\sin\varphi\sin\vartheta\\\cos\vartheta\end{pmatrix}$$

ergibt sich schließlich die geodätische Krümmung

$$k_g = (\mathbf{n}_S \times \mathbf{t}) \cdot \frac{d\mathbf{t}}{ds} = \frac{\tan\vartheta}{R}.$$

Erwartungsgemäß hat etwa der Äquator als Großkreis die geodätische Krümmung 0, während für $\vartheta \to \pi/2$ die Krümmung gegen unendlich geht.

Anhang B

Das Arnoldi-Verfahren für Systeme der Form Ax = b

Das Arnoldi-Verfahren wurde in Abschnitt 6.4.1 nur in einer Variante für Systeme der speziellen Form $(\mathbf{I} - \delta\mathbf{A})\mathbf{x} = \mathbf{b}$ behandelt. Der Algorithmus zur Lösung des Systems $\mathbf{Ax} = \mathbf{b}$ findet sich bei ARNOLDI [8], siehe auch SAAD [140, 139, 138].

Algorithmus B.1 (Arnoldi-Verfahren für $\mathbf{Ax} = \mathbf{b}$).

$\mathbf{q}_1 = \mathbf{b}/\|\mathbf{b}\|$

$\mathbf{Q}_1 = (\mathbf{q}_1)$

for $i = 2, 3, \ldots, m$

$\mathbf{v}_{i-1} = A\mathbf{q}_{i-1}$ (Krylov-Schritt)

$\mathbf{w}_{i-1} = (\mathbf{I} - \mathbf{Q}_{i-1}\mathbf{Q}_{i-1}^T)\mathbf{v}_{i-1}$ (Gram-Schmidt-Orthogonalisierung)

$\mathbf{q}_i = \mathbf{w}_{i-1}/\|\mathbf{w}_{i-1}\|$

$\mathbf{Q}_i = (\mathbf{Q}_{i-1}\ \mathbf{q}_i)$

end

$\mathbf{H}_m = \mathbf{Q}_m^T\mathbf{A}\mathbf{Q}_m$

löse $\mathbf{H}_m\mathbf{y}_m = \mathbf{Q}_m^T\mathbf{b}$

$\mathbf{x}_m = \mathbf{Q}_m\mathbf{y}_m$

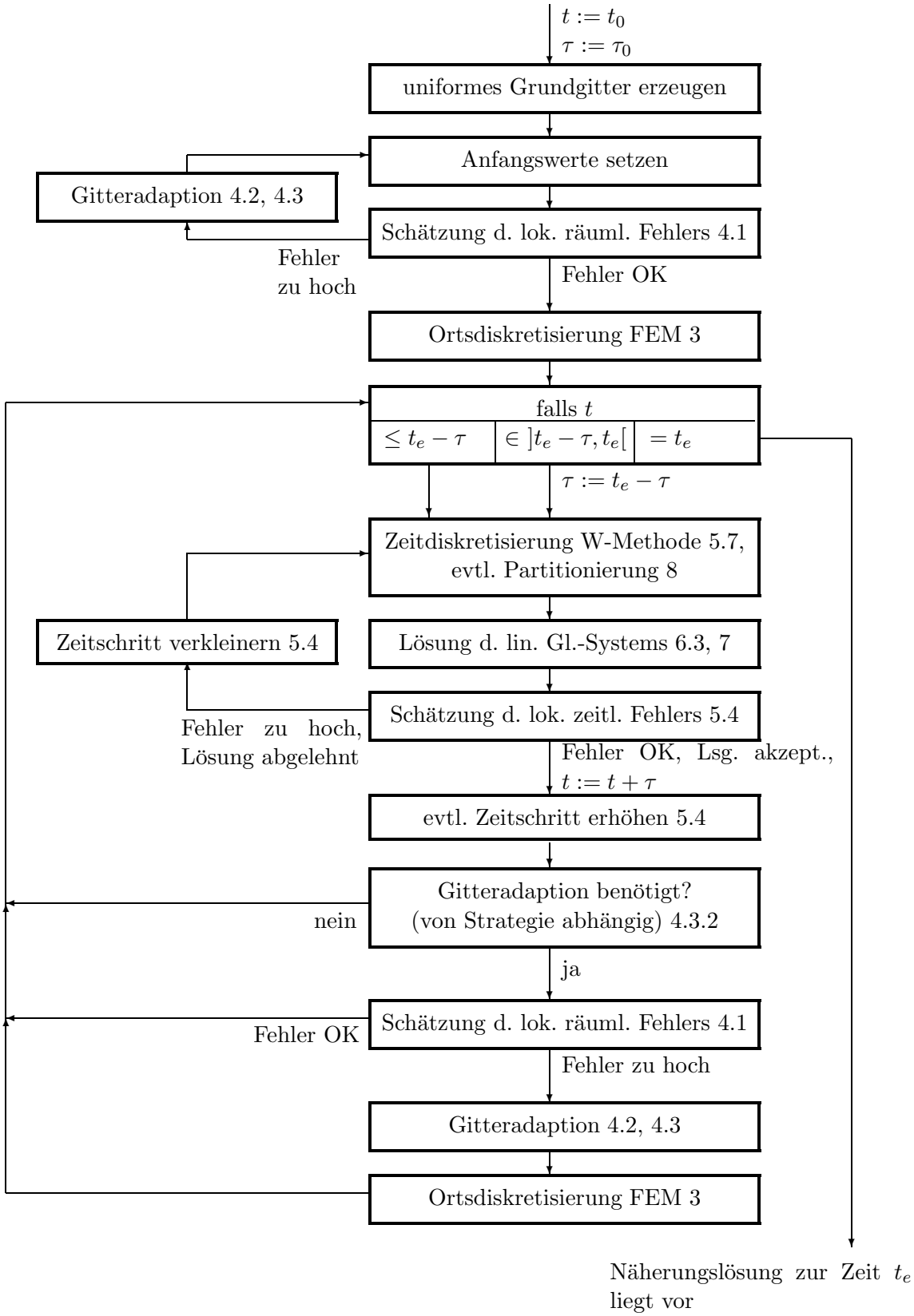
□

Der Vektor $\mathbf{x}_m$ ist die Näherung der gesuchten Lösung $\mathbf{x}$.

Anhang C

Ablaufplan zur Diskretisierung von Reaktions-Diffusions-Systemen

Den in dieser Arbeit verwendeten Diskretisierungsverfahren für Reaktions-Diffusions-Systeme liegt der umseitig dargestellte Ablaufplan zugrunde. Die angegebenen Zahlen bezeichnen jeweils die Abschnitte, in denen auf den entsprechenden Schritt Bezug genommen wird.


$t := t_0$
$\tau := \tau_0$
uniformes Grundgitter erzeugen
Anfangswerte setzen
Gitteradaption 4.2, 4.3
Schätzung d. lok. räuml. Fehlers 4.1
Fehler zu hoch
Fehler OK
Ortsdiskretisierung FEM 3
falls t
$\leq t_e - \tau$
$\in]t_e - \tau, t_e[$
$= t_e$
$\tau := t_e - \tau$
Zeitdiskretisierung W-Methode 5.7, evtl. Partitionierung 8
Zeitschritt verkleinern 5.4
Lösung d. lin. Gl.-Systems 6.3, 7
Schätzung d. lok. zeitl. Fehlers 5.4
Fehler zu hoch, Lösung abgelehnt
Fehler OK, Lsg. akzept., $t := t + \tau$
evtl. Zeitschritt erhöhen 5.4
Gitteradaption benötigt? (von Strategie abhängig) 4.3.2
nein
ja
Schätzung d. lok. räuml. Fehlers 4.1
Fehler OK
Fehler zu hoch
Gitteradaption 4.2, 4.3
Ortsdiskretisierung FEM 3
Näherungslösung zur Zeit t_e liegt vor

Literaturverzeichnis

[1] A.Y. Abramychev, V.A. Davydov, V.S. Zykov. Zh. Eksp. Teor. Fiz. 97 (1990) 1188 (russisch), Sov. Phys. JETP 70 (1990) 666 (englisch).

[2] R.A. Adams. *Sobolev Spaces.* Academic Press 1975.

[3] S. Adjerid, J.E. Flaherty. *Second-order finite element approximations and a posteriori error estimation for two-dimensional parabolic systems.* Num. Math. 53 (1988) 183-198.

[4] K. Agladze, R.R. Aliev, T. Yamaguchi, K. Yoshikawa. *Chemical Diode.* J. Phys. Chem. 100 (1996) 13895-13897.

[5] K. Agladze, O. Steinbock. *Waves and Vortices of Rust on the Surface of Corroding Steel.* J. Phys. Chem. A 104 (2000) 9816-9819.

[6] H.W. Alt. *Lineare Funktionalanalysis.* Springer-Verlag Berlin, Heidelberg, New York, 1992.

[7] H. Amann. *Nonhomogeneous linear and quasilinear elliptic and parabolic boundary value problems.* in H.-J. Schmeisser, H. Triebel. (Hrsg.) Function spaces, differential operators and nonlinear analysis. Teubner, Leipzig, 1993.

[8] W.E. Arnoldi. *The principle of minimized iterations in the solution of the matrix eigenvalue problem.* Quart. Appl. Math. 9 (1951) 17-29.

[9] D.G. Aronson, H.F. Weinberger. *Multidimensional nonlinear diffusion arising in population genetics.* Adv. in Math. 30 (1978)33-76.

[10] O. Axelsson, A. Barker. *Finite element solution of boundary value problems. Theory and computation.* Academic Press, Orlando/Florida 1984.

[11] I. Babuška, M. Feistauer, P. Šolín. *On one approach to a posteriori estimates for evolution problems solved by the method of lines.* Num. Math. 89 (2001) 225-256.

[12] I. Babuška, W.C. Rheinboldt. *Error estimates for adaptive finite element computations.* SIAM J. Num. Anal. 15 (1978) 736-754.

[13] C. Bär. *Elementare Differentialgeometrie.* de Gruyter 2001.

[14] R.E. Bank. *PLTMG: A software package for solving elliptic partial differential equations. User's Guide 6.0.* SIAM, Philadelphia 1990.

[15] R.E. Bank, A.H. Sherman, A. Weiser. *Refinement algorithms and data structures for regular local mesh refinement.* in Scientific Computing, IMACS, North-Holland, Amsterdam 1983.

[16] R.E. Bank, A. Weiser. *Some a posteriori error estimators for elliptic partial differential equations.* Math. Comp. 44 (1985) 283-301.

[17] D. Barkley. *A model for fast computer simulation of waves in excitable media.* Physica D 49 (1991) 61-70.

[18] D. Barkley, M. Kness, L.S. Tuckerman. *Spiral-wave dynamics in a simple model of excitable media: The transition from simple to profound rotation.* Phys. Rev. A 42 (1990) 2489-2492.

[19] R. Barrett, M. Berry, T.F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, H. van der Vorst. *Templates for the solution of linear systems: building blocks for iterative methods.* SIAM, Philadelphia 1994.

[20] P. Bastian, K. Birken, K. Johannsen, S. Lang, N. Neuss, H. Rentz-Reichert, C. Wieners. *UG - a flexible software toolbox for solving partial differential equations.* Computing and Visualization in Science 1 (1997) 27-40, erhältlich unter `http://cox.iwr.uni-heidelberg.de/~ug/`

[21] B.P. Belousov. *Eine periodische Reaktion und ihr Mechanismus.* Sbornik referatov po radiacionnoi medicine za 1958 (Sammlung von Arbeiten über Strahlungsmedizin des Jahres 1958) 1 (1959) 145-147 (russisch), deutsche Übersetzung in: L. Kuhnert, U. Niedersen. Ostwalds Klassiker der exakten Wissenschaften, Band 272, Selbstorganisation chemischer Strukturen. Verlag Harri Deutsch, Thun, Frankfurt am Main, 1999.

[22] B.P. Belousov. *A periodic reaction and its mechanism.* in: M.T. Grekhova. (Hrsg.) Avtovolnovye processy v sistemakh s diffuziejj. (Autowave Processes in Systems with Diffusion) Gorki, 1981 (russisch), deutsche Übersetzung in: L. Kuhnert, U. Niedersen. Ostwalds Klassiker der exakten Wissenschaften, Band 272, Selbstorganisation chemischer Strukturen. Verlag Harri Deutsch, Thun, Frankfurt am Main, 1999, englische Übersetzung in: R.J. Field, M. Burger. Oscillations and Traveling Waves in Chemical Systems. John Wiley & Sons, New York 1985.

[23] M. Bieterman, I. Babuška. *An adaptive method of lines with error control for parabolic equations of the reaction-diffusion type.* J. Comp. Phys. 63 (1986) 33-66.

[24] B. Blaius, A. Huppert, L. Stone. *Complex dynamics and phase synchronization in spatially extended ecological systems.* Nature 399 (1999) 354-359.

[25] J.G. Blom, J.G. Verwer, R.A. Trompert. *A comparison between direct and iterative methods to solve the linear systems arising from a time-dependent 2D groundwater flow model.* Comp. Fluid Dyn. 1 (1993) 95-113.

[26] F.A. Bornemann. *An adaptive multilevel approach to parabolic equations. III. 2D error estimation and multilevel preconditioning.* IMPACT Comp. Sci. Eng. 4 (1992) 1-45.

[27] W.C. Bray. *A periodic reaction in homogeneous solution and its relation to catalysis.* J. Am. Chem. Soc. 43 (1921) 1262-1267.

[28] W.C. Bray, H.A. Liebhafsky. *Reactions involving hydrogen peroxide, iodine and iodate ion. I. Introduction.* J. Am. Chem. Soc. 53 (1931) 38-44.

[29] P.K. Brazhnik, V.A. Davydov, A.S. Mikhailov. Teor. Mat. Fiz. 74 (1988) 444 (russisch), Theor. Math. Phys. (USSR) 74 (1988) 300 (englische Übersetzung).

[30] T.S. Briggs, W.C. Rauscher. *An Oscillating Iodine Clock.* J. Chem. Educ. 50 (1973) 496.

[31] J.C. Butcher. *On Runge-Kutta processes of high order.* J. Austral. Math. Soc. IV/2 (1964) 179-194.

[32] M. Büttner. *W-Methoden für steife Systeme – B-Konvergenz und Implementierung mittels Krylovtechniken.* Dissertation, Martin-Luther-Universität Halle-Wittenberg, 1992.

[33] M. Büttner, B.A. Schmitt, R. Weiner. *Automatic partitioning in linearly-implicit Runge-Kutta methods.* Appl. Num. Math. 13 (1993) 41-55.

[34] M. Büttner, B.A. Schmitt, R. Weiner. *W-methods with automatic partitioning by Krylov techniques for large stiff systems.* SIAM J. Num. Anal. 32/1 (1995) 260-284.

[35] P. Camacho, J.D. Lechleiter. *Increased Frequency of Calcium Waves in Xenopus laevis Oocytes that Express a Calcium-ATPase.* Science 260 (1993) 226-229.

[36] A.B. Carey, R.H. Giles, R.G. McLean. *The landscape epidemiology of rabies in Virginia.* Am J. Trop. Med. Hyg. 27 (1978) 573-580.

[37] M.P. do Carmo. *Differential geometry of curves and surfaces.* Prentice Hall 1976.

[38] A.L. Cauchy. *Résumé des Leçons données à l'Ecole Royale Polytechnique. Suite du Calcul Infinitésimal.* 1824. veröffentlicht: Equations différentielles ordinaires. Hrsg. C. Gilain, Johnson 1981.

[39] F. Chávez, R. Kapral. *Scroll waves in spherical shell geometries.* Chaos 11/4 (2001) 757-765.

[40] K.N. Chueh, C.C. Conley, J.A. Smoller. *Positively Invariant Regions for Systems of Nonlinear Diffusion Equations.* Indiana University Math. J. 26 (1977) 373-392.

[41] R.W. Clough. *The finite element method in plane stress analysis.* in Proceedings 2nd ASCE Conference on Electronic Computation, Pittsburgh 1960.

[42] R. Courant. *Variational methods for the solution of problems of equilibrium and vibration.* Bull. Amer. Math. Soc. 49 (1943) 1-23.

[43] R. Courant, K. Friedrichs, H. Lewy. *Ueber die partiellen Differenzengleichungen der mathematischen Physik.* Math. Ann. 100 (1928) 32-74.

[44] C.F. Curtiss, J.O. Hirschfelder. *Integration of stiff equations.* Proc. Nat. Acad. Sci. 38 (1952) 235-243.

[45] G. Dahlquist. *A special stability problem for linear multistep methods.* BIT 3 (1963) 27-43.

[46] J.M. Davidenko, A.V. Pertsov, R. Salomosz, W. Baxter, J. Jalife. *Stationary and drifting spiral waves of excitation in isolated cardiac muscle.* Nature 355 (1992) 349-351.

[47] V.A. Davydov, N. Manz, O. Steinbock, V.S. Zykov, S.C. Müller. *Excitation fronts on a periodically modulated curved surface.* Phys. Rev. Lett. 85/4 (2000) 868-871.

[48] V.A. Davydov, V.S. Zykov. *Kinematics of spiral waves on nonuniformly curved surfaces.* Physica D 49 (1991) 71-74.

[49] V.A. Davydov, V.S. Zykov, A.S. Mikhailov. *Kinematics of autowave structures in excitable media.* Sov. Phys. Usp. 34/8 (1991) 665-684.

[50] V.A. Davydov, V.S. Zykov, T. Yamaguchi. *Drift of spiral waves on nonuniformly curved surfaces.* in: A.R. Khokhlov (Hrsg.): International Conference on Nonlinear Dynamics in Polymer Science and Related Fields, Conference Center Desna, Moscow Region, Russia, October 11 - 15, 1999. Macromol. Symp. 160, 99-106, Wiley-VCH, Weinheim 2000.

[51] P. Deuflhard, F. Bornemann. *Numerische Mathematik II. Integration gewöhnlicher Differentialgleichungen.* de Gruyter Berlin 1994.

[52] J.R. Dormand, P.J. Prince. *A family of embedded Runge-Kutta formulae.* J. Comp. Appl. Math. 6 (1980) 19-26.

[53] J.R. Dormand, P.J. Prince. *Higher order embedded Runge-Kutta formulae.* J. Comp. Appl. Math. 7 (1981) 67-75.

[54] G. Dziuk. *Finite elements for the Beltrami operator on arbitrary surfaces.* in S. Hildebrand, R. Leis. (Hrsg.) Partial differential equations and calculus of variations. Lecture notes in mathematics 1357, Springer, 1988.

[55] B.L. Ehle. *On Padé approximations to the exponential function and A-stable methods for the numerical solution of initial value problems.* Research Report CSRR 2010, Dept. AACS, Univ. of Waterloo, Ontario, Canada 1969.

[56] W.H. Enright, M. Kamel. *Automatic partitioning of stiff systems and exploiting the resulting structure.* ACM-TOMS 5 (1979) 374-385.

[57] K. Eriksson, C. Johnson. *Adaptive finite element methods for linear elliptic problems.* Math. Comp. 50 (1988) 361-383.

[58] K. Eriksson, C. Johnson. *Adaptive finite element methods for parabolic problems IV: Nonlinear problems.* SIAM J. Num. Math. 32 (1995) 1729-1749.

[59] G. Ertl. *Oscillatory Kinetics and Spatio-Temporal Self-Organisation in Reactions at Solid Surfaces.* Science 254 (1991) 1750-1755.

[60] L. Euler. *Institutionum Calculi Integralis.* Volumen Primum, Opera Omnia, Vol. XI, 1768.

[61] E. Fehlberg. *New high-order Runge-Kutta formulas with step size control for systems of first and second order differential equations.* ZAMM 44 (1964), Sonderheft T17-T19.

[62] E. Fehlberg. *Classical fifth-, sixth-, seventh-, and eighth order Runge-Kutta formulas with step size control.* NASA Technical Report 287 (1968), auszugsweise veröffentlicht in Computing 4 (1969) 93-106.

[63] E. Fehlberg. *Low-order classical Runge-Kutta formulas with step size control and their application to some heat transfer problems.* NASA Technical Report 315 (1969), auszugsweise veröffentlicht in Computing 6 (1970) 61-71.

[64] J.R. Field, E. Körös, R.M. Noyes. *Oscillations in Chemical Systems. II. Thorough Analysis of Temporal Oscillation in the Bromate-Cerium-Malonic Acid System.* J. Am. Chem. Soc. 94 (1972) 8649-8664.

[65] J.R. Field, R.M. Noyes. *Oscillations in chemical systems. IV. Limit cycle behavior in a model of a real chemical reaction.* J. Chem. Physics 60 (1974) 1877-1884.

[66] R. FitzHugh. *Impulses and physiological states in theoretical models of nerve membrane.* Biophys. J. 1 (1961) 445-466.

[67] R. Fletcher. *Conjugate gradient methods for indefinite systems.* in: G. Watson (Hrsg.): Numerical Analysis Dundee 1975, Springer-Verlag, Berlin, New York, 1976.

[68] R. Freund, N. Nachtigal. *QMR: A quasi-minimal residual method for non-Hermitian linear systems.* Num. Math. 60 (1991) 315-339.

[69] C.F. Gauß, Werke, Göttingen 1903.

[70] C.W. Gear, Y. Saad. *Iterative solution of linear equations in ODE codes.* SIAM J. Sci. Stat. Comput. 4/4 (1983) 583-601.

[71] G. Gerisch. *Periodische Signale steuern die Musterbildung in Zellverbänden.* Naturwissenschaften 58 (1971) 430-438.

[72] S. Gerschgorin. *Über die Abgrenzung der Eigenwerte einer Matrix.* Bull. Acad. Sci. Leningrad (1931) 749-754.

[73] G. Golub, C. Van Loan. *Matrix computations.* 2. Auflage. The Johns Hopkins University Press, Baltimore 1989.

[74] N.A. Gorelova, J. Bureš. *Spiral Waves of Spreading Depression in the Isolated Chicken Retina.* J. Neurobiol. 14 (1983) 353-363.

[75] C. Großmann, H.-G. Roos. *Numerik partieller Differentialgleichungen.* Teubner, Stuttgart, 2. Aufl. 1994.

[76] W. Hackbusch. *Iterative Lösung großer schwachbesetzter Gleichungssysteme.* Teubner, Stuttgart, 2. Aufl. 1993.

[77] E. Hairer, S.P. Nørsett, G. Wanner. *Solving ordinary differential equations I.* Springer-Verlag, Berlin, Heidelberg, 2. Aufl. 1993.

[78] E. Hairer, G. Wanner. *Solving ordinary differential equations II.* Springer-Verlag, Berlin, Heidelberg, 2. Aufl. 1996.

[79] P.C. Hammer, J.W. Hollingsworth. *Trapezoidal methods of approximating solutions of differential equations.* MTAC 9 (1955) 92-96.

[80] G. Hämmerlin, K.-H. Hoffmann. *Numerische Mathematik.* Springer-Verlag, Berlin, Heidelberg, New York, 4. Aufl. 1994.

[81] M. Hestenes, E. Stiefel. *Methods of conjugate gradients for solving linear systems.* J. Res. Nat. Bur. Stand. 49 (1952) 409-436.

[82] K. Heun. Neue Methode zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen. Zeitschr. f. Math. u. Phys. 45 (1900) 23-38.

[83] H. Heuser. *Lehrbuch der Analysis, Teil 1.* Teubner Stuttgart, 11. Aufl. 1990.

[84] A.L. Hodgkin, A.F. Huxley. *A quantitative description of membrane current and its application to conduction and excitation in nerves.* J. Physiol. 117 (1952) 500-544.

[85] W.H. Hundsdorfer. *Partially implicit BDF2 blends for convection dominated flows.* CWI-Report MAS-R9831. 1998.

[86] K.R. Jackson, W.L. Seward. *Adaptive linear equation solvers in codes for large stiff systems of ODEs.* SIAM J. Sci. Comput. 14 (1993) 800-823.

[87] W. Jahnke, A.T. Winfree. *A survey of spiral wave behaviors in the Oregonator model.* Int. J. Bifurc. Chaos 1 (1991) 445-466.

[88] C.G. Jacobi. *Über eine neue Auflösungsart der bei der Methode der kleinsten Quadrate vorkommenden linearen Gleichungen.* Astronom. Nachr. (1845).

[89] S. Jakubith, H.H. Rothermund, W. Engel, A. von Oertzen, G. Ertl. *Spatio-Temporal Concentration Patterns in a Surface Reaction: Propagating and Standing Waves, Rotating Spirals, and Turbulence.* Phys. Rev. Lett. 65 (1990) 3013-3016.

[90] J. Jost. *Riemannian geometry and geometric analysis.* Springer 1998.

[91] J.I. Kanel. *Über die Stabilität der Lösungen des Cauchyschen Problems für Gleichungen, die in der Theorie der Verbrennung auftreten.* Mat. Sb. 59/101 Ergänzungsband (1962) 245-288 (russisch).

[92] R. Kapral, K. Showalter. (Hrsg.) *Chemical waves and patterns.* Kluwer Ac. Publ., Dordrecht 1995.

[93] J.P. Keener, J.J. Tyson. *Spiral waves in the Belousov-Zhabotinskii reaction.* Physica D 21D (1986) 307-324.

[94] P. Knabner, L. Angermann. *Numerik partieller Differentialgleichungen.* Springer-Verlag, Berlin, Heidelberg 2000.

[95] M. Koecher. *Lineare Algebra und analytische Geometrie.* Springer 1997.

[96] A. Kolmogorov, I. Petrovsky, N. Piskunov. *Étude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un probleme biologique.* Bull. Univ. État Moscou, Sér. Int., Sect. A: Math. et Mécan. 1, Fasc. 6 (1937) 1-25 (französisch), Byull. Moskov. Gos. Univ. 17 (1937) 1-72 (russisch).

[97] E. Kreyszig. *Differentialgeometrie.* 2. Aufl., Akademische Verlagsgesellschaft Geest & Portig, Leipzig 1968.

[98] V.I. Krinsky, A.M. Pertsov, A.N. Reshetilov. *Study of the mechanism of initiation of ectopic excitation center on modified Hodgkin-Huxley equations.* Biofizika 17/2 (1972) 271-277 (russisch), Biophys. (USSR) 17 (1972) 282 (englische Übersetzung).

[99] K. Krischer. *Principles of temporal and spatial pattern formation in electrochemical systems.* in: B.E. Conway, R.E. White. (Hrsg.) Modern Aspects of Electrochemistry, Number 32, Kluwer Academic / Plenum Publishers, New York 1999.

[100] W. Kutta. *Beitrag zur näherungsweisen Integration totaler Differentialgleichungen.* Zeitschr. f. Math. u. Phys. 46 (1901) 435-453.

[101] C. Lanczos. *Solution of systems of linear equations by minimized iterations.* J. Res. Nat. Bur. Stand. 49 (1952) 33-53.

[102] J. Lang. *Adaptive FEM for reaction-diffusion equations.* Appl. Num. Math. 26 (1998) 105-116.

[103] J. Lang. *Adaptive multilevel solution of nonlinear parabolic PDE systems: theory, algorithm and applications.* Springer-Verlag Berlin, Heidelberg 2001.

[104] J. Lang, A. Walter. *A finite element method adaptive in space and time for nonlinear reaction-diffusion systems.* IMPACT Comp. Sci. Eng. 4 (1992) 269-314.

[105] A.A.P. Leão. *Spreading depression of activity in the cerebral cortex.* J. Neurophysiol. 7 (1944) 359-390.

[106] J.D. Lechleiter, S. Girard, E. Peralta, D.E. Clapham. *Spiral Calcium Wave Propagation and Annihilation in Xenopus laevis Oocytes.* Science 252 (1991) 123-126.

[107] R. Lefever, G. Nicolis. *Chemical instabilities and sustained oscillations.* J. theor. Biol. 30 (1971) 267-284.

[108] A.M. Liapunov. *Problème général de la stabilité du mouvement.* russisch 1892, französische Übersetzung in: Annales de la Faculté des Sciences de Toulouse, 1907, Nachdruck: Princeton University Press, 1947.

[109] R. Luther. *Räumliche Fortpflanzung chemischer Reaktionen.* Z. Elektrochem. 12 (1906) 596-600 (deutsch), englische Übersetzung in: R. Arnold, K. Showalter, J.J. Tyson. Translation of Luther's "Propagation of chemical reactions in space". J. Chem. Educ. 64 (1987) 740-742.

[110] B.F. Madore, W.L. Freedman. *Self-organizing structures.* Am. Sci. 75 (1987) 252-259.

[111] N. Manz. *Untersuchung chemischer Wellen in der* Belousov-Zhabotinsky-*Reaktion: räumlich modulierte Systeme und anomale Dispersion.* Cuvillier-Verlag Göttingen 2002.

[112] N. Manz, V.A. Davydov, S.C. Müller, M. Bär. *Dependence of the spiral rotation frequency on the surface curvature of reaction-diffusion systems.* Phys. Lett. A 316/5 (2003) 311-316.

[113] J. Maselko, K. Showalter. *Chemical waves on spherical surfaces.* Nature 339 (1989) 609-611.

[114] The MathWorks, Inc. *MATLAB.* Version 6.5.0.180913a Release 13, 2002.

[115] P. McQuillan, J. Gomatam. *Rotating chemical waves on the sphere.* J. Phys. Chem. 100/13 (1996) 5157-5159.

[116] H. Meinhardt, M. Klinger. *A Model for Pattern Formation on the Shells of Molluscs.* J. Theor. Biol. 126 (1987) 63-89.

[117] R.H. Merson. *An operational method for the study of integration processes.* Proc. Symp. Data Processing, Weapons Research Establishment, Salisbury, Australia (1957) 110-1 bis 110-25.

[118] A.S. Mikhailov, V.S. Zykov. *Kinematical theory of spiral waves in excitable media: Comparison with numerical simulations.* Physica D 52 (1991) 379-397.

[119] A.D. Miller. *Distorted elements in finite element analysis.* in R.L. May, A.K. Easton. (Hrsg.) Computational Techniques and Applications: CTAC95, World Scientific, 1996.

[120] P.K. Moore. *A posteriori error estimation with finite element semi- and fully discrete methods for nonlinear parabolic equations in one space dimension.* SIAM J. Num. Anal. 31 (1994) 149-169.

[121] S.C. Müller, T. Plesser, B. Hess. *Two-Dimensional Spectrophotometry of Spiral Wave Propagation in the Belousov-Zhabotinskii Reaction, 2. Geometric and Kinematic Parameters.* Physica D 24 (1987) 87-96.

[122] J.D. Murray, E.A. Stanley, D.L. Brown. *On the spatial spread of rabies among foxes.* Proc. Roy. Soc. London B 229 (1986) 111-150.

[123] N. Nachtigal, S. Reddy, L. Trefethen. *How fast are nonsymmetric matrix iterations?* SIAM J. Matrix Anal. Appl. 13 (1992) 778-795.

[124] J. Nagumo, S. Arimoto, S. Yoshizawa. *An active pulse transmission line simulating nerve axon.* Proc. IEEE 50 (1962) 2061-2070.

[125] U. Nowak. *Adaptive Linienmethoden für nichtlineare parabolische Systeme in einer Raumdimension.* Dissertation, Freie Universität Berlin, 1993.

[126] R.M. Noyes, R.J. Field, E. Körös. *Oscillations in Chemical Systems. I. Detailed Mechanism in a System Showing Temporal Oscillations.* J. Am. Chem. Soc. 94 (1972) 1394-1395.

[127] J.T. Oden. *Finite Elements: An Introduction.* in P.G. Ciarlet, J.L. Lions. (Hrsg.) Handbook of Numerical Analysis. Bd. II, North Holland, Amsterdam, 1991.

[128] A. Papastavrou. *Adaptive Finite Element Methoden für Konvektions-Diffusionsprobleme.* Dissertation, Ruhr-Universität Bochum, 1998.

[129] A.L. Pardhanani, G.F. Carey. *Efficient simulation of complex patterns in reaction-diffusion systems.* J. Comp. Appl. Math. 74 (1996) 295-311.

[130] T. Plesser, S.C. Müller, B. Hess. J. Phys. Chem. 94 (1990) 7501.

[131] B. van der Pol. *On "Relaxation Oscillations".* Phil. Mag., Vol. 2, 978-992, neu aufgelegt in B. van der Pol. Selected Scientific Papers. Vol. I, North Holland, Amsterdam, 1960.

[132] A. Quarteroni, A. Valli. *Numerical approximation of partial differential equations.* Springer-Verlag Berlin, Heidelberg 1994.

[133] E. Ranta, V. Kaitala. *Travelling waves in vole population dynamics.* Nature 390 (1997) 456.

[134] A. Robertson, M.H. Cohen. *Control of Developing Fields.* Ann. Rev. Biophys. Bioeng. 1 (1972) 409-464.

[135] H.H. Rosenbrock. *Some general implicit processes for the numerical solution of differential equations.* Comp. J. 5 (1963) 329-331.

[136] E.J. Routh. *A Treatise on the stability of a given state of motions.* Being the essay to which the Adams prize was adjudged in 1877, in the University of Cambridge. London 1877.

[137] C. Runge. *Ueber die numerische Auflösung von Differentialgleichungen.* Math. Ann. 46 (1895) 167-178.

[138] Y. Saad. *Krylov subspace methods for solving large unsymmetric linear systems.* Math. Comp. 37 (1981) 105-126.

[139] Y. Saad. *On the Rates of Convergence of the Lanczos and the Block-Lanczos Methods.* SIAM J. Num. Anal. 17/5 (1980) 687-706.

[140] Y. Saad. *Variations on Arnoldi's Method for Computing Eigenelements of Large Unsymmetric Matrices.* Lin. Alg. Appl. 34 (1980) 269-295.

[141] Y. Saad, M. Schultz. *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems.* SIAM J. Sci. Statist. Comput. 7 (1986) 856-869.

[142] K. Schellbach. *Probleme der Variationsrechnung.* J. Reine Angew. Math. 41 (1851) 293-363.

[143] B.A. Schmitt, R. Weiner. *Matrix-free W-methods using a multiple Arnoldi iteration.* Appl. Num. Math. 18 (1995) 307-320.

[144] B.A. Schmitt, R. Weiner. *Polynomial preconditioning in Krylov-ROW-methods.* Appl. Num. Math. 28 (1998) 427-437.

[145] L.S. Schulman, P.E. Seiden. *Percolation and Galaxies.* Science 233 (1986) 425-431.

[146] H.R. Schwarz. *Methode der finiten Elemente.* Teubner, Stuttgart, 3. Aufl. 1991.

[147] P.L. Seidel. *Über ein Verfahren, die Gleichungen, auf welche die Methode der kleinsten Quadrate führt, sowie lineare Gleichungen überhaupt, durch successive Annäherung aufzulösen.* Münch. Abh. (1847).

[148] L.F. Shampine. *Numerical Solution of Ordinary Differential Equations.* Chapman & Hall 1994.

[149] F. Siegert, C.J. Weijer. *Digital image processing of optical density wave propagation in Dictyostelium discoideum and analysis of the effects of caffeine and ammonia.* J. Cell Sci. 93 (1989) 325-335.

[150] J. Smoller. *Shock waves and reaction diffusion equations.* Springer-Verlag, New York, 2. Aufl. 1994.

[151] P. Sonneveld. *CGS, a fast Lanczos-type solver for nonsymmetric linear systems.* SIAM J. Sci. Statist. Comput. 10 (1989) 36-52.

[152] T. Steihaug, A. Wolfbrandt. *An attempt to avoid exact Jacobian and nonlinear equations in the numerical solution of stiff differential equations.* Math. Comp. 33 (1979) 521-534.

[153] O. Steinbock. *Excitation Waves on Cylindrical Surfaces: Rotor Competition and Vortex Drift.* Phys. Rev. Lett. 78 (1997) 745-748.

[154] K. Strehmel, R. Weiner. *Linear-implizite Runge-Kutta-Methoden und ihre Anwendung.* Teubner, Leipzig, 1992.

[155] K. Strehmel, R. Weiner. *Numerik gewöhnlicher Differentialgleichungen.* Teubner, Stuttgart, 1995.

[156] H.L. Swinney. *Observations of complex dynamicsand chaos.* in: E.G.D. Cohen. (Hrsg.) *Fundamental problems in statistical mechanics VI.* North Holland, Amsterdam, 1985.

[157] J. Tesarik, M. Sousa, J. Testart. *Human oocyte activation after intracytoplasmic sperm injection.* Hum. Reprod. 9 (1994) 511-518.

[158] H. Triebel. *Höhere Analysis.* VEB Verlag der Wissenschaften, Berlin, 1972.

[159] A.M. Turing. *The Chemical Basis of Morphogenesis.* Philos. Trans. Roy. Soc. London Ser. B 237 (1952) 37-72.

[160] M.J. Turner, R.W. Clough, H.C. Martin, L.J. Topp. *Stiffness and deflection analysis of complex structures.* J. Aero. Sci. 23 (1956) 805-823.

[161] J.J. Tyson. *Oscillations, Bistability, and Echo Waves in Models of the* Belousov-Zhabotinsky *Reaction.* Ann. NY Acad. Sci. 316 (1979) 279-295.

[162] J.J. Tyson. *A Quantitative Account of Oscillations, Bistability, and Traveling Waves in the Belousov-Zhabotinskii Reaction.* in R.J. Field, M. Burger. (Hrsg.) Oscillations and Traveling Waves in Chemical Systems. John Wiley & Sons, 1985, S. 93-144.

[163] J.J. Tyson, P.C. Fife. *Target patterns in a realistic model of the Belousov-Zhabotinskii reaction.* J. Phys. Chem. 73 (1980) 2224-2237.

[164] R. Verfürth. *A posteriori error estimates for non-linear problems. $L^r(0,T;L^r)$-error estimates for finite element discretizations of parabolic equations.* Math. Comp. 67 (1998) 1335-1360.

[165] R. Verfürth. *A review of a posteriori error estimation and adaptive mesh-refinement techniques.* Wiley & Teubner, Chichester, Stuttgart 1996.

[166] H. van der Vorst. *Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems.* SIAM J. Sci. Statist. Comput. 13 (1992) 631-644.

[167] R. Weiner, M. Arnold, P. Rentrop, K. Strehmel. *Partitioning strategies in Runge-Kutta type methods.* IMA J. Num. Anal. 13 (1993) 303-319.

[168] R. Weiner, B.A. Schmitt. *Consistency of Krylov-W-methods in initial value problems.* Techn. Report 14, Martin-Luther-Universität Halle-Wittenberg, FB Mathematik und Informatik, 1996.

[169] R. Weiner, B.A. Schmitt, H. Podhaisky. *ROWMAP – a ROW-code with Krylov techniques for large stiff ODEs.* Appl. Num. Math. 25 (1997) 303-319.

[170] D. Werner. *Funktionalanalysis.* Springer-Verlag Berlin, Heidelberg, 2. Auflage 1997.

[171] O.B. Widlund. *A note on unconditionally stable linear multistep methods.* BIT 7 (1967) 65-70.

[172] N. Wiener, A. Rosenblueth. *The mathematical formulation of the problem of conduction of impulses in a network of connected excitable elements, specifically in cardiac muscle.* Arch. Inst. Card. Mex. 16 (1946) 205-265.

[173] A.T. Winfree. *Spiral waves of chemical activity.* Science 175 (1972) 634-636.

[174] H. Yagisita, M. Mimura, M. Yamada. *Spiral wave behaviors in an excitable reaction-diffusion system on a sphere.* Physica D 124 (1998) 126-136.

[175] D.M. Young. *Iterative methods for solving partial differential equations of elliptic type.* Doctoral thesis, Harvard University, 1950.

[176] A.M. Zhabotinsky. *Periodic Processes of the Oxidation of Malonic Acid in Solution (Investigation of the Kinetics of the Reaction of Belousov.* Biofizika 9/3 (1964) 306-311 (russisch), deutsche Übersetzung in: L. Kuhnert, U. Niedersen. Ostwalds Klassiker der exakten Wissenschaften, Band 272, Selbstorganisation chemischer Strukturen. Verlag Harri Deutsch, Thun, Frankfurt am Main, 1999, englische Übersetzung in: Biophysics 9 (1964) 329-335.

[177] O.C. Zienkiewicz, J.Z. Zhu. *A simple error estimator and adaptive procedure for practical engineering analysis.* Int. J. Num. Meth. Eng. 24 (1987) 337-357.

[178] R. Zurmühl, S. Falk. *Matrizen und ihre Anwendungen 1. Grundlagen. Für Ingenieure, Physiker und Angewandte Mathematiker.* Springer-Verlag, Berlin, Heidelberg, 7. Aufl. 1997.

[179] R. Zurmühl, S. Falk. *Matrizen and ihre Anwendungen für Angewandte Mathematiker, Physiker und Ingenieure. Teil 2: Numerische Methoden.* Springer-Verlag, Berlin, Heidelberg, 5. Aufl. 1986.

[180] V.S. Zykov. *Control of Complex Systems.* (russisch) Nauka, Moskau 1975.

[181] V.S. Zykov. *Simulation of wave processes in excitable media.* Manchester University Press, Manchester 1987.

[182] V.S. Zykov, A.S. Mikhailov, S.C. Müller. *Controlling spiral waves in confined geometries by global feedback.* Phys. Rev. Lett. 78/17 (1997) 3398-3401.

[183] V.S. Zykov, S.C. Müller. *Spiral waves on circular and spherical domains of excitable medium.* Physica D 97 (1996) 322-332.

[184] V.S. Zykov, O. Steinbock, S.C. Müller. *External forcing of spiral waves.* Chaos 4/3 (1994) 509-518.