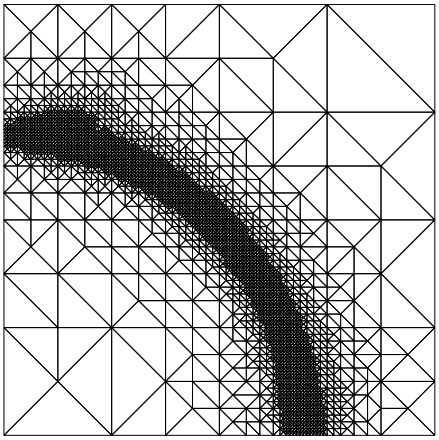
Adaptive Finite Element Methoden für Konvektions-Diffusionsprobleme

Dissertation
zur Erlangung des Grades eines Doktor-Ingenieurs
der Fakultät für Bauingenieurwesen
an der Ruhr-Universität Bochum



von
Areti Papastavrou
aus Hanau

Graduiertenkolleg
Computational Structural Dynamics

Tag der mündlichen Prüfung: 19.10.1998

Prüfungsvorsitzender: Prof. Dr.-Ing. H. Niemann
Referent: Prof. Dr.rer.nat. R. Verfürth
Korreferent: Prof. G. Schmid, Ph.D.
Prüfungsbeisitzer: Prof. Dr.-Ing. K. Krass

Zusammenfassung

Der Schwerpunkt der vorliegenden Arbeit liegt auf der Entwicklung effektiver adaptiver Finiter Element Methoden für Konvektions-Diffusionsprobleme mit dominanter Konvektion. Zunächst wird von einem stationären Modellproblem ausgegangen. Für dieses werden verschiedene Finite Element Methoden der Bubnov-Galerkin und der Petrov-Galerkin Art, sowie Modifikationen davon, unter Verwendung linearer Ansatzfunktionen auf Dreieckselementen dargestellt. Für das Modellproblem mit dominater Konvektion schneidet im Vergleich das SU/PG-Verfahren als sehr praktikabel ab.

Zur Lokalisierung des Diskretisierungsfehler, z.B. im Bereich scharfer Schockfronten, werden verschiedene Möglichkeiten zur *a posteriori* Abschätzung bereitgestellt. Neben der Darstellung eines Fehlerindikators aus einer Mittelungstechnik werden ausgehend von einer klassischen Formulierung eines residuellen Fehlerschätzers modifizierte Vorfaktoren der Residuenterme entwickelt. Desweiteren wird ein Fehlerschätzer basierend auf einem lokalen Konvektions-Diffusionsproblem mit Neumann Randbedingungen eingeführt. Numerische Beispiele dienen dem Vergleich der verschiedenen Fehlerschätzer und zeigen deren Wirksamkeit.
Zur Durchführung einer anschließenden Verfeinerung des zugrundeliegenden Diskretisierungsnetzes werden verschiedene Markierungsmethoden und Verfeinerungsstrategien einander gegenübergestellt. Anhand von Beispielen wird ihre Wirkungsweise veranschaulicht und ein Vergleich geführt.

Eine weitere Anpassung des Netzes an die berechnete numerische Lösung kann auch durch die Optimierung der Plazierung der Netzknoten erfolgen. Zu diesem Zweck wird ein Netzglättungsalgorithmus basierend auf Informationen eines *a posteriori* Fehlerschätzers vorgestellt und dieser insbesondere an den eingeführten Neumann-Fehlerschätzer angepaßt. Beispiele zeigen, daß diese Netzanpassungsmethode zur Minimierung des Fehlers beiträgt, dies jedoch nicht in vergleichbar starkem Maß wie eine weitere Netzverfeinerung. Dennoch stellt diese neue Methode eine sinnvolle Ergänzung zur Netzverfeinerung dar.

Den Abschluß der Arbeit bildet die Erweiterung der Betrachtungen auf ein instationäres Modellproblem. Zur numerischen Lösung wird dabei auf die räumliche Diskretisierung des stationären Problems zurückgegriffen. Die Zeitdiskretisierung erfolgt mittels eines θ-Schemas. Zur Steuerung der Größe der Zeitschritte werden Standardtechniken verwendet.
Informationen über den Diskretisierungsfehler können in jedem einzelnen Zeitschritt die vorgestellten Fehlerschätzer liefern. Zur Anpassung des Netzes an die zeitlich veränderliche Lösung dienen unterschiedliche Vorgehensweisen, die u.a. auf Vergröberungstechniken und Delaunay-Triangulierungen basieren. Beispiele belegen die Effektivität des Verfahrens.

Vorwort

Die vorliegende Arbeit entstand während meiner Förderungszeit durch das Graduiertenkolleg "Computational Structural Dynamics" an der Ruhr-Universität Bochum.
Mein besonderer Dank gilt dem Hauptreferenten Herrn Prof. Dr.rer.nat. Rüdiger Verfürth, der die Arbeit angeregt und deren Fortgang stets mit Engagement und konstruktiver Kritik verfolgt hat.
Herrn Prof. Dr. Günther Schmid PhD danke ich herzlich für die Übernahme des Korreferats und die kritische Durchsicht dieser Arbeit.
Mein Dank gilt in besonderer Weise den Mitgliedern des Graduiertenkollegs für die Schaffung einer fruchtbaren wissenschaftlichen Umgebung und der DFG für die finanzielle Unterstützung.
Schließlich danke ich allen derzeitigen und ehemaligen Kollegen am Lehrstuhl für Numerik, deren Hilfsbereitschaft und Freundschaft die Grundlage einer angenehmen, den wissenschaftlichen Austausch fördernden Atmosphäre darstellt. Die vielen wertvollen Diskussionen mit Daniel Flensberg, Matthias Wiedmer, Regina Sarazin und Hermann Weber haben mir sehr geholfen.

Bochum, im Oktober 1998 Areti Papastavrou

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

Mathematische Modelle, die eine Kombination von Konvektions-Diffusionsprozessen beschreiben, gehören zu den weitverbreitesten in den Natur- und Ingenieurwissenschaften. Oft ist der dimensionslose Parameter, der die relative Größe des Diffusionsanteils mißt, sehr klein. Die Konvektions-Diffusionsgleichung nimmt dann einen vornehmlich hyperbolischen Charakter an und es können dünne Grenzschichten am Rand und im Inneren des Gebietes, sowie scharfe Schockfronten auftreten. Im strengsten Sinne, d.h. im rein hyperbolischen Fall, stellen Grenzschichten bzw. Schockfronten Diskontinuitäten in der Lösung dar, die mit zunehmender Diffusion ausgeschmiert werden.
Unter all diesen Umständen wird man, anders als für elliptische und parabolische Probleme, mit Standardverfahren der Methode der Finiten Elemente auf numerische Schemata kommen, die physikalisch unsinnige Ergebnisse liefern. Insbesondere wurde dieses schlechte Verhalten für die oben genannten Fälle beobachtet, in denen die exakte Lösung eines Problems von hyperbolischem Charakter nicht-glatt ist. Wenn die exakte Lösung z.B. eine Diskontinuität in Form eines Sprungs aufweist, zeigt die Lösung einer Standard Methode der Finiten Elemente i.a. große Oszillationen, auch in weiter Entfernung vom Sprung. Damit ist die approximierte Lösung nirgends nah an der exakten Lösung und damit praktisch unbrauchbar. Da viele hyperbolische Probleme nicht-glatte exakte Lösungen besitzen, werden Approximationsverfahren mit höherer Stabilität benötigt. In der jüngeren Vergangenheit gelang es, diese Schwierigkeiten durch die Konstruktion neuer spezieller Verfahren für diese Problemklasse mit befriedigenden Konvergenzeigenschaften zu bewältigen. In diesem Zusammenhang seien die *SUPG-* oder auch *Stromlinien-Diffusions-Methode* und deren *Shock-Capturing Modifikationen* zu nennen, die im Rahmen dieser Arbeit vorgestellt werden.
Bei der numerischen Behandlung partieller Differentialgleichungen haben sich adaptive Methoden als zusätzliche wertvolle Hilfsmittel erwiesen. In Fällen, in denen die Genauigkeit der numerischen Approximation durch lokale Singularitäten, wie einspringende Ecken, Grenzschichten oder Schockfronten gestört wird, ist es sinnvoll, die Diskretisierung dort zu verfeinern. Um eine optimale Genauigkeit zu erhalten, müssen die Regionen, in denen die Lösung weniger regulär ist, identifiziert und entsprechend verfeinert werden. Dabei muß eine gute Balance zwischen verfeinerten und unverfeinerten Teilgebieten erreicht werden. Die Netzadaption umfaßt somit die Behandlung von zwei Problemstellungen: Zum einen muß ein Verdichtungsindikator, der auf einer *a posteriori* Abschätzung des Fehlers der numerischen Berechnung in einer geeigneten Norm basiert, bestimmt werden. Zum anderen muß ein Netzgenerator ausgehend von der Verteilung

der Indikatoren die Diskretisierung in geschickter Weise so umstrukturieren, daß der lokale Fehler in den einzelnen Elementen reduziert wird.
Hinsichtlich der Abschätzung des Diskretisierungsfehlers stehen verschiedene Kategorien von *a posteriori* Fehlerschätzer, z.B. *residuelle* Fehlerschätzer, Indikatoren aus Mittelungstechniken oder Fehlerschätzer basierend auf *lokalen Hilfsproblemen* zur Verfügung.
Neben der h-Adaption, in der Verfeinerungstrategien und Markierungsmethoden eine große Rolle spielen, können auch sogenannte Netzglättungsmethoden einen positiven Beitrag zur Reduzierung des Fehlers der numerische Lösung leisten. Im Verfeinerungsprozeß der h-Adaption entsteht eine Folge von hierarchischen Netzen, in denen die Positionierung der einzelnen Knotenpunkte und damit die Ausrichtung der Elemente vom Ursprungsnetz abhängt. Die Verschiebung von Knoten unter Beibehaltung der Netztopologie zum Abschluß einer solchen Netzverfeinerung kann schon mit relativ kleinen Positionsänderungen große Effekte erzielen.

1.1 Zielsetzung

Aus dem zuvor Gesagten ergibt sich zunächst die Forderung nach der Bereitstellung stabiler Methoden der Finiten Elemente mit zufriedenstellenden Konvergenzeigenschaften zur Lösung eines Konvektions-Diffusionsproblems auch mit dominanter Konvektion. Für ein stationäres Modellproblem sollen ausgehend von der Bubnov-Galerkin Methode verschiedene spezielle Methoden diskutiert werden.
Das mögliche Auftreten von Singularitäten, in Form von Schockfronten oder Grenzschichten, läßt den Wunsch nach effizienten adaptiven Methoden, welche basierend auf Informationen über den Fehler eine Approximation gewünschter Genauigkeit mit nahezu minimalem Aufwand liefern, aufkommen. Die benötigten Informationen über die Verteilung und Größe der Fehlers sollen dabei mittels *a posteriori* Fehlerschätzer ermittelt werden. Insbesondere soll ein passender residueller Fehlerschätzer und ein Fehlerschätzer basierend auf einem lokalen Konvektions-Diffusionsproblem entwickelt werden.
Zur eigentlichen Verfeinerung der Triangulierungen sollen verschiedene Markierungs- und Verfeinerungsstrategien zur Auswahl gestellt werden.
Desweiteren soll ein Netzglättungsalgorithmus basierend auf Informationen eines *a posteriori* Fehlerschätzers vorgestellt werden. Aufgabe dieses Netzglättungsalgorithmus ist die Optimierung der Netzknoten einer aus dem adaptiven Prozeß hervorgegangenen Triangulierung hinsichtlich des Fehlers. Sein Einfluß auf die Minimierung des Fehlers soll diskutiert werden.
Abschließend soll die Erweiterung dieser Betrachtungen auf den instationären Fall geführt werden. Zur Zeitdiskretisierung soll ein θ-Schema verwendet werden. Die zeitlich veränderliche Lösung erfordert eine entsprechene Anpassung der Netze. Zu diesem Zweck sollen verschiedene Vorgehensweisen vorgestellt werden.

1.2 Gliederung

Der Inhalt dieser Arbeit gliedert sich folgendermaßen:
In **Kapitel 2** werden zunächst einige repräsentative Beispiele für Konvektions-Diffusionsprobleme vorgestellt und anhand des Stofftransports in einem Grundwasserleiter die

einzelnen Prozesse erläutert. Anschließend wird das stationäre Modellproblem formuliert.

Die numerische Behandlung der betrachteten Differentialgleichung mittels Methoden Finiter Elemente wird in **Kapitel 3** vorgeführt. Insbesondere werden die Bubnov-Galerkin Methode, die SUPG bzw. SDFEM, sowie Shock-Capturing Modifikationen dargestellt. Anhand von Beispielen werden die Besonderheiten der einzelnen Methoden beleuchtet.

Das **Kapitel 4** ist iterativen Algorithmen zur Lösung der aus den Diskretisierungen durch Finite Elemente entstehenden unsymmetrischen Gleichungssysteme gewidmet. Besonderen Platz nimmt dabei das BiCG-Stab Verfahren ein.

In **Kapitel 5** werden verschiedene Möglichkeiten zur *a posteriori* Abschätzung des Approximationsfehlers bereitgestellt. Zentrale Punkte sind hierbei ein Indikator aus einer Mittelungstechnik, ein klassischer residueller Fehlerschätzer und ein Fehlerschätzer basierend auf einem lokalen Neumann Problem.

Für die verwendeten hierachischen strukturierten Netze werden in **Kapitel 6** verschiedene Markierungsmethoden und Verfeinerungsstrategien zur Durchführung einer Netzverfeinerung basierend auf den vorliegenden Indikatoren vorgestellt und anhand von Beispielen verglichen.

In **Kapitel 7** wird ein Netzglättungsalgorithmus zur Optimierung der Plazierung der Knoten einer Triangulierung, die aus einem adaptiven Verfeinerungsprozeß gewonnen wurde, präsentiert.

Schließlich erfolgt in **Kapitel 8** die Erweiterung der Betrachtung auf ein instationäres Modellproblem. Die Diskretisierung der Zeit erfolgt mittels eines θ-Schemas. Die räumliche Diskretisierung wird aus der Behandlung des stationären Problems übernommen. Die Anpassung der Netze an die zeitlich veränderliche Lösung erfolgt mittels der vorgestellten Fehlerindikatoren und verschiedener Verfahren, die auf Verfeinerungs- bzw. Vergröberungsmethoden oder auf Delaunay-Triangulierungen basieren.

Kapitel 2

Die Konvektions-Diffusionsgleichung

Zunächst soll die Vielfältigkeit der Probleme aus Natur- und Ingenieurwissenschaften, in denen Konvektions-Diffusions-Phänomene auftreten, anhand einiger Beispiele verdeutlicht werden. In vielen Fällen treten Konvektions-Diffusions-Phänomene eingebettet in einem größeren Zusammenhang auf und werden deshalb häufig nicht unbedingt als solche wahrgenommen. In anderen Fällen umfassen und beschreiben sie das Gesamtproblem. Eine umfangreiche Zusammenstellung von verschiedenen Problemen sind in [48] zu finden.
Anschließend an die Vorstellung einiger beispielhafter Konvektions-Diffusionsprobleme sollen speziell anhand des Problems des Stofftransports in einem Grundwasserleiter die Bedeutung und Problematik der Konvektions-Diffusionsgleichung sowie die auftretenden Prozesse veranschaulicht werden. Vor diesem Hintergrund wird am Ende dieses Kapitels ein stationäres Modellproblem vorgestellt.

2.1 Beispielhafte Probleme

Ein sehr anschauliches Beispiel für ein Konvektions-Diffusionsproblem ist die Verschmutzungsausbreitung in einer Flußmündung. Man stelle sich eine Flußmündung als zweidimensionales Gebiet, wie in der folgenden, stark vereinfachten Abbildung 2.1 angedeutet, vor.

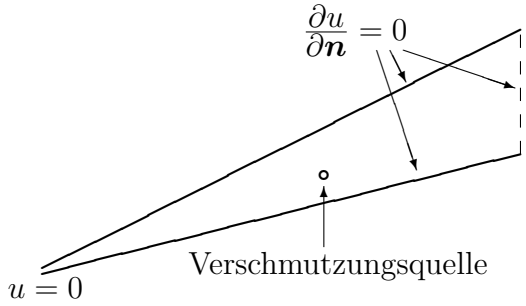


Abbildung 2.1: Flußmündung

Mit u sei die über die Tiefe gemittelte Konzentration der Verschmutzung, z.B. Ab-

wasser, chemischer oder radioaktiver Abfall oder auch eine Temperaturverteilung (z.B. durch Kühlwasser), bezeichnet. Dann hat die typische Gleichung zur Beschreibung ihrer Ausbreitung die Form

$$\frac{\partial u}{\partial t} + \boldsymbol{b} \cdot \nabla u = \nabla \cdot (\epsilon \nabla u) + f \ .$$

Dabei ist f die Quelle, die räumlich sehr begrenzt, aber mit der Zeit veränderlich sein kann. Der Vektor $\boldsymbol{b}$ beschreibt ein gegebenes ebenes Geschwindigkeitsfeld, welches ebenfalls von der Zeit abhängen kann, z.B. aufgrund von Tidebewegungen. In den meisten Fällen wird $\boldsymbol{b}$ als divergenzfrei ($\nabla \cdot \boldsymbol{b} = 0$), d.h. das Fluid als inkompressibel angenommen, so daß man $\boldsymbol{b} \cdot \nabla u$ durch $\nabla \cdot (\boldsymbol{b}u)$ ersetzen kann. Der Koeffizient ϵ beschreibt die Diffusivität, welche die Mischung durch vertikale Bewegungen in gemittelter Form, sowie turbulente und molekulare Diffusion berücksichtigt.
Die dimensionslose *Peclet-Zahl* $\mathrm{Pe} = |\boldsymbol{b}|L/\epsilon$ gibt das Verhältnis von Konvektion zu Diffusion wieder. Für typische Dimensionen des Problems, Geschwindigkeiten von 0.5 bis 3m/sec, eine Diffusivität von 0.05 bis $5\mathrm{m}^2$/sec und einer charakteristischen Längenabmessung L von einigen 100m, nimmt die *Peclet-Zahl* Werte von 10 bis $6{\cdot}10^3$ an.
Eine repräsentative lineare Konvektions-Diffusionsgleichung erhält man, wenn ϵ als unabhängig von der Konzentration u angenommen wird.

Die Modellierung einer atmosphärischen Verschmutzung, wie z.B. nach der Tschernobyl-Katastrophe durch radioaktive Partikel oder nach dem Ausbruch des Vesuv in der Antike durch Asche, stellt ein komplexes dreidimensionales Problem dar. Äußere Parameter sind dabei die gemittelte horizontale Windgeschwindigkeit, als Quellen und Senken fungieren z.B. der Abfall der Radioaktivität oder die Transformation der Isotope im System. Desweiteren geht ein vertikaler Diffusionskoeffizient ein, der eine Funktion der Höhe ist. Bemerkenswert an diesem Problem ist, daß die Skalen in horizontaler und vertikaler Richtung weit auseinanderliegen können. Insbesondere bei einer Verschmutzung wie nach Tschernobyl können in horizontaler Richtung die Längenskalen Tausende von Kilometern betragen und Geschwindigkeiten von 10m/sec auftreten, während in vertikaler Richtung die Skalen wesentlich kleiner anzusetzen sind. So kann die Bandbreite von auftretenden *Peclet-Zahlen* von 10^{-2} in atmosphärischen Grenzschichten bis zu 10 in freier Atmosphäre reichen.

Anhand des Problems des Stofftransports durch einen Grundwasserleiter soll die Bedeutung und Problematik der Konvektions-Diffusion-Gleichung im folgenden Abschnitt etwas veranschaulicht werden.

2.2 Stofftransport im Grundwasser

Ein typisches makroskopisches phänomenologisches Modell für den Grundwassertransport einer Substanz in Lösung in einem Grundwasserleiter mit der Konzentration u hat die Form

$$\frac{\partial (n_e u)}{\partial t} + \nabla \cdot (n_e \boldsymbol{b} u) - \nabla \cdot (n_e (D^m \mathbf{1} + \boldsymbol{D}) \cdot \nabla u) - f = 0$$

Zur Erläuterung der Bedeutung und des Hintergrunds der einzelnen Terme soll etwas weiter ausgeholt werden. Details zu den folgenden Ausführungen können in [43] nachgelesen werden.
Die Ausbreitung von Stoffen im Untergrund kann in Abhängigkeit von der Art des Stoffes, der Art seiner Eintragung, dem Aquifertyp, dem Typ der Strömung etc. sehr unterschiedlich ablaufen. Die Prozesse lassen sich nach folgenden Kriterien klassifizieren:

1. Art des Grundwasserleiters

   - Porengrundwasserleiter oder Kluftgrundwasserleiter
     Ersterer ist insbesondere für die Trinkwasserversorgung wichtig.

2. Art der Strömung

   - ungesättigte und gesättigte Verhältnisse
     In der gesättigten Zone vollzieht sich im wesentlichen die horizontale Ausbreitung von Stoffen, während durch die ungesättigten Zonen mit dem Sickerwasser Stoffe i.a. in vertikaler Richtung bewegt werden. Durch das zugrundeliegende Strömungsfeld ergeben sich i.a. Unterschiede zwischen dem Transport in beide Richtungen und diese werden meist getrennt betrachtet. Innerhalb dieser Arbeit sollen nur Strömungs- und Transportprozesse in gesättigten Zonen betrachtet werden.

   - mit Wasser nicht mischbare und mischbare/gelöste Stoffe
     Die nicht mischbaren Stoffe (z.B. Mineralölprodukte) stellen eine dem Wasser gleichwertige Phase dar, da der Transport nach den Gesetzmäßigkeiten der Mehrphasenströmung erfolgt. I.a. ist ein Transport über größere Distanzen nur für gelöste Stoffe möglich und dieser soll im weiteren betrachtet werden.

   - hydrodynamische aktive und inaktive Konzentrationen
     Abhängig von seiner Konzentration kann der Inhaltsstoff die Strömung, z.B. durch Gravitationseffekte, Veränderung der Dichte und kinematischen Zähigkeit des Fluids, beeinflussen. Bei hohen Konzentrationen kann also eine Rückkopplung des Transportprozesses auf die Stömung erfolgen. Bei niedrigen Konzentrationen können diese Einflüsse auf die Stömung vernachlässigt werden.

3. Art des transportierten Stoffes

   - ideale Tracer und Stoffe mit Adsorption und chemischer Reaktion
     Unter einem idealen Tracer versteht man einen Wasserinhaltsstoff, welcher die Wasserbewegung mitvollzieht. Er erfährt weder Adsorption am umgebenden Gestein noch unterliegt er chemischem Abbau oder Zerfall und er liegt in hydrodynamisch inaktiver Konzentration vor.

4. Art des Transportes

   - Advektion und Diffusion, Dispersion
     Für einen idealen Tracer gibt es in mikroskopischer Betrachtungsweise, d.h. das Kontrollvolumen ist klein gegenüber einer Einzelpore, nur zwei grundlegende Transportprozesse: Die durch das Geschwindigkeitsfeld der Grundströmung verursachte

Konvektion, auch Advektion genannt, und die durch die Brownsche Molekularbewegung verursachte Diffusion, die eine Ausbreitung in Richtung abnehmender Konzentration bewirkt.
In makroskopischer Betrachtungsweise (d.h. in größeren Gebieten müssen zur Beschreibung des Transportes gemittelte Geschwindigkeiten berücksichtigt werden) werden alle Transporteffekte, welche in den Inhomogenitäten des Strömungsfeldes begründet liegen und deren Größe unterhalb der des Mittelungsvolumens liegen, in dem Begriff der Dispersion zusammengefaßt.

2.2.1 Konvektion, Advektion

Auf mikroskopischer Ebene stellt man neben dem Prozeß der molekularen Diffusion die konvektive Bewegung von Wasserinhaltsstoffen in Richtung des Wassers mit dessen lokaler Geschwindigkeit (Advektion) fest. An jedem Punkt des Wassers innerhalb einer Pore läßt sich die Massenstromdichte des transportieren Stoffes infolge Advektion wie folgt darstellen.

$$\boldsymbol{N}_{advek} = u\boldsymbol{b}$$

Dabei bezeichnet u die Konzentration und $\boldsymbol{b}$ die lokale Geschwindigkeit.

2.2.2 Diffusion

Die molekulare Diffusion bewirkt einen Ausgleich von Konzentrationsunterschieden in Richtung abfallender Konzentration und damit eine Vermischung von Wasser und Wasserinhaltsstoff. Dem diffusiven Anteil der Massenstromdichte liegt das Ficksche Gesetz zugrunde:

$$\boldsymbol{N}_{diff} = -D^m \nabla u$$

mit der Diffusionskonstanten D^m. Aufgrund der geringen Größe der Diffusionskonstanten ist die Vermischung zunächst nur auf der Ebene einer Einzelpore von Bedeutung. Um den Transport über makroskopische Bereiche beschreiben zu können, müssen Mittelwerte des Massenstromdichtevektors betrachtet werden. Diese können entweder über den effektiven Porenraum oder über das Gesamtvolumen einschließlich der Gesteinskörner ermittelt werden. Durch die effektive Porosität n_e sind die beiden Mittelwerte miteinander gekoppelt.

$$\boldsymbol{N}^{Gesamtporenvolumen} = n_e \boldsymbol{N}^{Gesamtvolumen}$$

Die beteiligten Größen werden in ihren Mittelwert über den effektiven Porenraum und die lokale Abweichung bzw. Schwankung zerlegt.

$$\begin{aligned} \boldsymbol{b} &= \bar{\boldsymbol{b}} + \boldsymbol{b}' \\ u &= \bar{u} + u' \end{aligned}$$

Man erhält folgende Massenstromdichte nach der Mittelung

$$\boldsymbol{N} = n_e(\bar{u}\bar{\boldsymbol{b}} - D^m \nabla \bar{u} + \overline{u'\boldsymbol{b}'})$$

Dabei sind die Terme, in denen die Schwankungsgrößen einfach enthalten sind, verschwunden, während Produkte aus diesen Größen nicht im Mittelungsprozeß herausfallen. Der erste Term beschreibt den Massenfluß aufgrund der mittleren Geschwindigkeit und wird Konvektionsterm genannt. Der zweite Term gibt den Massenfluß infolge Diffusion an und der dritte Term wird als Dispersionsterm bezeichnet.

2.2.3 Dispersion

Der Dispersionsterm beschreibt den Massenfluß, der über den mittleren Wert $\bar{u}\bar{\boldsymbol{b}}$ hinaus aufgrund der Geschwindigkeitsfluktuationen innerhalb des Mittelungsvolumens auftritt. Die Größe des Mittelungsvolumens legt einen Schnitt bei der Skalengröße, bis zu welcher Inhomogenitäten des Strömungsfeldes noch explizit bekannt sind. Alle Transporteffekte, die durch kleinerskalige Inhomogenitäten des Strömungsfeldes verursacht werden, werden in der Dispersion zusammengefaßt.
In Analogie zur Diffusion wird der Dispersionterm i.a. durch einen Fickschen Ansatz beschrieben:

$$\overline{u'\boldsymbol{b}'} = -\mathbf{D} \cdot \nabla \bar{u}.$$

Der Dispersionstensor $\mathbf{D}$ ist ein Tensor zweiter Stufe. I.a. ist die Dispersion in Strömungsrichtung eine Größenordnung größer als quer zur Richtung der Strömung. In die Koeffizienten gehen sowohl Aquifer- als auch Strömungseigenschaften ein.

Aus der Massenbilanz im Kontrollvolumen erhält man die Transportgleichung, in welcher der über die Berandung ein- bzw. austretende Massenstrom und die Quellen und Senken innerhalb des Kontrollvolumens mit den Änderungen der Masse im Kontrollvolumen im Gleichgewicht stehen müssen. Unter Berücksichtigung der bisher betrachteten Prozesse erhält die Transportgleichung folgende Form:

$$\frac{\partial(n_e u)}{\partial t} + \nabla \cdot (n_e \boldsymbol{b} u) - \nabla \cdot (n_e (D^m \mathbf{1} + \boldsymbol{D}) \cdot \nabla u) - f = 0.$$

Die Zugabe bzw. Entnahme des Inhaltsstoffes, die nicht mit einer Zugabe/Entnahme von Wasser verbunden ist, ist in der Größe f gegeben, wobei mit $f > 0$ eine Quelle und mit $f < 0$ eine Senke vorliegt.
Der spezifische Durchfluß $\boldsymbol{q} = n_e \boldsymbol{b}$ des Grundwassers, der in der Transportgleichung den konvektiven Term ausmacht, kann aus einem Grundwasserströmungsmodell der allgemeinen Form

$$\frac{\partial(\rho n_e)}{\partial t} + \nabla \cdot (\rho \boldsymbol{q}) = 0$$

berechnet werden, wobei ρ die Dichte ist. Eine Verallgemeinerung des Darcy-Gesetzes stellt den Zusammenhang zwischen $\boldsymbol{q}$ und der äquivalenten Piezometerhöhe her

$$\boldsymbol{q} = -\mathbf{K} \cdot \nabla \phi \ .$$

Hierin ist $\mathbf{K}$ die Matrix der effektiven hydraulischen Durchlässigkeiten und ϕ die äquivalente Piezometerhöhe, die als Funktion der Dichte ρ und der vertikalen Höhe z betrachtet werden kann. In diesem Zusammenhang soll auf die umfassenden Arbeiten von [5]

hingewiesen werden. Die Werte für die Größen in diesen Gleichungen können sehr stark schwanken und sind häufig sehr schlecht zu bestimmen.
Zur Lösung der Differentialgleichung des Stofftransports müssen sowohl Anfangswerte als auch Randbedingungen gegeben sein. Sind Werte für die Konzentration auf dem Rand vorgegeben, so spricht man von *Dirichlet*- Randbedingungen. Ist der Konzentrationsgradient (und damit der dispersive Fluß) senkrecht zum Gebietsrand gegeben, handelt es sich um sogenannte *Neumann*-Randbedingungen. Darüberhinaus können auch gemischte Randbedingungen vorliegen, die den gesamten Stoffffluß über den Rand spezifizieren, in diesem Fall spricht man von *Cauchy*-Randbedingungen.

2.2.4 Adsorption

Unter Adsorption versteht man die physikalische oder chemische Bindung des gelösten Stoffes an der Oberfläche fester Stoffe oder Gesteinskörner. Die physikalische Adsorption ist i.a. reversibel und läuft relativ schnell ab, die chemische hingegen ist i.a. irreversibel und hat einen langsamen Verlauf. Häufig wird auch der Effekt, den 'dead-end-pores' auf das Transportverhalten ausüben, als Adsorption bezeichnet, obwohl es sich dabei um einen anderen Prozeß, z.B. die molekulare Diffusion in diese Poren, handelt.
In der Massenbilanz muß nicht nur die adsorbierte sondern auch die gelöste Substanz berücksichtigt werden. Man kann von einem lokalen Gleichgewicht zwischen adsorbierter und gelöster Substanz ausgehen, wenn der Adsorptionsprozeß im Vergleich zur Zeitskala der Strömung schnell verläuft. Die Konzentration u_a der adsorbierten Phase stellt sich dann als algebraische Funktion der Konzentration u der gelösten Phase dar:

$$u_a = f(u).$$

Die Funktion $f(u)$ wird Isotherme genannt, sie beschreibt das Gleichgewicht bei konstanter Temperatur. Im einfachsten Fall liegt sie als lineare Funktion der Konzentration vor, man spricht von dem K_D-Konzept, welches für kleine Konzentrationen immer verwendet werden kann:

$$u_a = K_D u.$$

Hieraus läßt sich ein Retardierungs- oder Verzögerungsfaktor ableiten, der das Verhältnis von gesamter zu gelöster Stoffmenge pro Volumeneinheit angibt. Bei einer Gleichgewichtsadsorption werden die Abstandsgeschwindigkeit, die Koeffizienten des Dispersionstensors und die äußeren Quellterme um den Retardierungsfaktor R vermindert. Dies wird als Verzögerung der Ausbreitung beobachtet.
Weitere häufig benutzte Funktionen sind die *Freundlich*-Isotherme und die *Langmuir*-Isotherme

$$u_a^F = k_1 u^{k_2} \ , \qquad u_a^L = \frac{k_1 u}{k_2 + u}$$

mit den Konstanten k_1, k_2. Nichtlineare Isothermen verzögern und deformieren die Schadstoffverteilung, außerdem führen sie Nichtlinearitäten in die Transportgleichung ein, für deren Lösung dann iterative Methoden herangezogen werden müssen.
Langsame Adsorptionsprozesse lassen sich nur durch Differentialgleichungen bezüglich der Zeit erfassen. Der Einfachheit halber sollen diese hier nicht weiter betrachtet werden.

2.2.5 Chemische Reaktionen

Der Vollständigkeit halber soll hier noch auf chemische Reaktionen hingewiesen sein, diese werden im weiteren jedoch nicht berücksichtigt.
Chemische Reaktionen eines Wasserinhaltsstoffes werden in der Transportgleichung als Quellen- bzw. Senkenterm eingebaut. Schnelle Reaktionen (Gleichgewichtsreaktionen) lassen sich durch ein System von miteinander gekoppelter Differentialgleichungen für die Konzentrationsverteilung und algebraischer Gleichungen zur Beschreibung der ablaufenden Reaktionen darstellen.

2.3 Modellproblem

Für die weiteren Betrachungen wird folgendes Modellproblem formuliert, ein stationäres, lineares Konvektions-Diffusionsproblem mit Adsorption und zugehörigen Randbedingungen:

$$\begin{aligned} -\nabla \cdot (\mathbf{D} \cdot \nabla u) + \boldsymbol{b} \cdot \nabla u + \alpha u &= f && \text{in } \Omega \\ u &= u_D && \text{auf } \Gamma_D \\ \boldsymbol{n} \cdot \mathbf{D} \cdot \nabla u &= t_N && \text{auf } \Gamma_N \end{aligned}$$

Der Notation des vorhergegangenen Kapitels folgend beschreibt u die unbekannte Skalargröße der Schadstoffkonzentration, $\boldsymbol{b} = [b_1, b_2]$ das gegebene Geschwindigkeitsfeld der zugrundeliegenden Strömung, $\alpha \geq 0$ den Koeffizient einer linearen Adsorptionsisotherme, f die Funktion der Quellen und Senken und $\boldsymbol{n}$ den Randnormalenvektor. $\mathbf{D} = \mathbf{D}(\boldsymbol{x})$ ist der zweistufige Tensor, in dem Diffusion und Dispersion zusammengefaßt wurden. Diese mathematische Vereinfachung ist möglich, da beide Prozesse durch das Ficksche Gesetz beschrieben werden und bei makroskopischer Betrachtung der Natur beide Phänomene schwer auseinanderzuhalten sind. Der entsprechende Term kann als hydrodynamische Dispersion bezeichnet werden, aber der Einfachheit halber soll er im folgenden als Diffusion bezeichnet werden. Unter Vernachlässigung von Dispersion und für den isotropen und homogenen Fall reduziert sich die Größe $\mathbf{D}$ auf den skalaren Diffusionskoeffizient $\epsilon \geq 0$. Das Modellproblem hat dann die Form

$$\begin{aligned} -\epsilon \Delta u + \boldsymbol{b} \cdot \nabla u + \alpha u &= f && \text{in } \Omega \\ u &= u_D && \text{auf } \Gamma_D \\ \epsilon \nabla u \cdot \boldsymbol{n} &= t_N && \text{auf } \Gamma_N \ . \end{aligned}$$

Es wird von einem divergenzfreien Strömungsfeld ausgegangen, d.h. statt $\nabla \cdot (\boldsymbol{b}u)$ läßt sich $\boldsymbol{b} \cdot \nabla u$ schreiben. Auf dem Teil Γ_N der Gebietsberandung seien Neumann-Bedingungen und auf dem Teil Γ_D Dirichlet-Bedingungen gegeben.

Die Differentialgleichung ist elliptisch mit mehr oder weniger hyperbolischem Charakter, abhängig von der Größe der Parameter ϵ und $\boldsymbol{b}$. Im Grenzfall $\epsilon = 0$ liegt eine rein hyperbolische Gleichung vor, wird ϵ hingegen groß, so nimmt sie mehr elliptischen Charakter an.

Insbesondere soll der Fall "ϵ klein" betrachtet werden, d.h. die Gleichung ist von stark hyperbolischer Natur und die zuvor beschriebenen Effekte sind von zunehmender Bedeutung.

Die Schwache Form der Differentialgleichung erhält man durch die Multiplikation mit einer geeigneten glatten Testfunktion w, die Integration über das Gebiet Ω und die partielle Integration des ersten Terms zu:

$$-(t_N, w)_{\Gamma_N} + (\mathbf{D} \cdot \nabla u, \nabla w) + (\boldsymbol{b} \cdot \nabla u + \alpha u, w) = (f, w) \qquad \forall w.$$

Mit $(\cdot, \cdot)$ wird dabei das innere L_2-Produkt über Ω von skalaren oder vektoriellen Größen bezeichnet

$$\begin{aligned} (\boldsymbol{g}, \boldsymbol{h}) &:= \int_\Omega \boldsymbol{g} \cdot \boldsymbol{h} \, \mathrm{d}\Omega \\ (\boldsymbol{g}, \boldsymbol{h})_{\Gamma_N} &:= \int_{\Gamma_N} \boldsymbol{g} \cdot \boldsymbol{h} \, \mathrm{d}\Gamma. \end{aligned}$$

Zur Abkürzung wird eine Bilinearform $a(\cdot, \cdot)$ und ein lineares Funktional $l(\cdot)$ für die rechte Seite, welches die Daten enthält, definiert

$$\begin{aligned} a(u, w) &:= (\mathbf{D} \cdot \nabla u, \nabla w) + (\boldsymbol{b} \cdot \nabla u + \alpha u, w) \\ l(w) &:= (f, w) + (t_N, w)_{\Gamma_N}. \end{aligned}$$

Die Bilinearform ist für $u, w \in H^1(\Omega)$, dem Sobolev-Raum der Funktionen, die wie auch ihre ersten Ableitungen über Ω quadratisch integrierbar sind, wohldefiniert. Die Lösung u muß auf dem Dirichlet-Rand Γ_D gleich dem vorgegebenen Wert u_D sein und die Testfunktionen dort identisch Null. Wir definieren

$$\begin{aligned} H^1_D(\Omega) &= \{v \in H^1(\Omega) : v = u_D \text{ auf } \Gamma_D\} \\ H^1_{D_0}(\Omega) &= \{v \in H^1(\Omega) : v = 0 \text{ auf } \Gamma_D\}. \end{aligned}$$

Somit schreibt sich die Schwache Formulierung des Modellproblems wie folgt:

Finde $u \in H^1_D(\Omega)$, so daß

$$a(u, w) = l(w) \qquad \forall w \in H^1_{D_0}(\Omega),$$

wobei $a(\cdot, \cdot)$ und $l(\cdot)$ wie oben definiert sind.

Kapitel 3

Finite Element Methoden

Die grundlegende Idee aller numerischen Methoden zur Lösung von Differentialgleichungen ist die *Diskretisierung* und damit Algebraisierung des gegebenen kontinuierlichen Problems, welches unendlich viele Freiheitsgrade besitzt. Das entstehende diskrete Problem bzw. das Gleichungssystem enthält hingegen eine endliche Anzahl von Unbekannten und kann mit Hilfe eines Computers gelöst werden.
In der *Finiten Differenzen Methode* (FDM), als der klassischen numerischen Methode für partielle Differentialgleichungen, werden die Ableitungen durch Differenzenquotienten ersetzt. Die Werte der Unbekannten werden an einer endlichen Zahl von festgelegten Punkten eingeführt, diese stellen einen Mittelwert über die entsprechende Gitterzelle dar.
Der Diskretisierungsprozeß innerhalb der *Methode der Finiten Elemente* (FEM) verläuft anders. Es wird von einer Reformulierung der gegebenen Differentialgleichung als äquivalentes Variationsproblem ausgegangen. Die unbekannte Größe, in diesem Fall die Konzentration, wird an jedem Punkt des modellierten Gebietes durch eine Interpolationsfunktion beschrieben. Die Stützstellen der Interpolationsfunktion bilden als Knoten ein beliebiges Elementnetz, welches begrenzt durch den Rand $\Gamma = \partial\Omega$ das Lösungsgebiet Ω überdeckt.
Vorteile der FEM gegenüber der FDM liegen in der relativ einfachen Handhabbarkeit von komplizierten Geometrien, allgemeinen Randbedingungen und veränderlichen oder nichtlinearen Materialeigenschaften. Außerdem ermöglicht die klare Struktur und Vielseitigkeit der FEM die Entwicklung von allgemeinen Softwarecodes für verschiedene Anwendungen. Darüberhinaus besitzt diese Methode ein starkes theoretisches Fundament, welches weitere Zuverlässigkeit gibt und die Möglichkeit eröffnet, den Fehler in der approximierten FE-Lösung mathematisch zu analysieren und abzuschätzen.
Zunächst wird das Bubnov-Galerkin-Verfahren als Standard Finite Element Methode angewandt. Für die vorliegende Konvektions-Diffusionsgleichung zeigen sich allerdings bei dominierender Konvektion unphysikalische Oszillationen in der numerischen Lösung. Aus diesem Grunde wurden stabilere Methoden, die streamline upwind/Petrov-Galerkin Methode [30] oder auch streamline diffusion Methode [36] u.a. entwickelt. Diese werden im Anschluß vorgestellt.

3.1 Standard-Galerkin, Bubnov-Galerkin

Es sei $\mathcal{T}_h$ eine Aufteilung des Gebietes Ω in beispielsweise dreieckige Finite Elemente, Triangulierung genannt. Mit T oder Ω_e sei ein Element, mit h_e sein Durchmesser und mit ∂T oder Γ_e sein Rand bezeichnet. Sei $H^1(\Omega)$ der Hilbert-Raum der quadratisch integrierbaren Funktionen mit quadratisch integrierbaren Ableitungen erster Ordnung. Der endlich dimensionale Unterraum $H^{1h} \subset H^1(\Omega)$ ist definiert als der Raum, welcher durch stückweise polynominale C^0-kontinuierliche Basisfunktionen über die Triangulierung $\mathcal{T}_h$ aufgespannt wird

$$H^{1h} = \{\phi^h : \phi^h \in C^0(\bar{\Omega}),\ \phi^h|_T \in P_k,\ \forall T \in \mathcal{T}_h\}.$$

Mit $\bar{\Omega}$ ist der Gebietsabschluß bezeichnet und P_k steht für die Polynome k-ter Ordnung. Innerhalb dieser Arbeit wird im wesentlichen $k = 1$ gesetzt, sofern nichts anderes vereinbart wird. Damit werden im Zweidimensionalen lineare Dreieckselemente, die sogenannten P1-Elemente, verwendet.
Eine in H^{1h} enthaltene Funktion kann als Linearkombination der Knotenbasisfunktionen beschrieben werden

$$\phi^h(x, y) = \sum_j \phi_j N_j(x, y),$$

wobei die Koeffizienten ϕ_j als Knotenfreiheitsgrade bezeichnet werden. Die lineare Interpolationsfunktion $N_j(x, y)$ nimmt an dem assoziierten Knoten j den Wert 1 an und verschwindet an den anderen Knoten, d.h. es gilt $N_j(x_i, y_i) = \delta_{ij}$ für jede Paarung von zwei Knoten i und j des Netzes, wobei δ_{ij} das Kroneckerdelta ist. Der Freiheitsgrad ϕ_j ist also der Wert der Funktion ϕ^h im Knoten j, $\phi_j = \phi^h(x_j, y_j)$.
In der Finiten-Element-Formulierung wird die Approximationslösung u^h in entsprechender Weise an den Knoten beschrieben, d.h. die Unbekannten des diskreten Problems liegen als Knotengrößen vor.
Die Integral-Darstellung zur Ermittlung dieser Unbekannten wird als *schwache Form* bezeichnet. Man erhält sie aus der Wichtung der Differentialgleichung mit der Test- oder Wichtungsfunktion w und anschließendem Integrieren. Bezeichend für die Bubnov-Galerkin Formulierung ist, daß der Raum der Testfunktion mit dem der Approximationslösung bis auf die Randbedingungen übereinstimmt. Dies bedeutet, daß die Wichtungsfunktion w eine kontinuierliche, stückweise lineare Funktion ist. Auf dem Dirichlet-Rand nimmt die Testfunktion den Wert Null an, $w \in H^1$ mit $w = 0$ auf Γ_D.

Die gesamte Gleichung stellt sich wie folgt in der schwachen Form für das Standard-Galerkin bzw. Bubnov-Galerkin Verfahren dar:

$$\int_\Omega \overset{(1)}{(\mathbf{D} \cdot \nabla u^h)} \nabla w^h \,\mathrm{d}\Omega + \int_\Omega \overset{(2)}{\boldsymbol{b} \cdot \nabla u^h} w^h \,\mathrm{d}\Omega + \alpha \int_\Omega \overset{(3)}{u^h} w^h \,\mathrm{d}\Omega = \int_\Omega \overset{(4)}{f} w^h \,\mathrm{d}\Omega + \int_{\Gamma_N} \overset{(5)}{t_N} w^h \,\mathrm{d}\Gamma$$

Für die Testfunktion w^h und die Konzentration u^h bzw. deren Gradienten werden die oben beschriebenen linearen Ansätze nach dem isoparametrischen Konzept gewählt:

$$w^h = \sum_{k=1}^{n_{nod}} N^k w_k = \boldsymbol{N}\boldsymbol{w} \quad , \quad u^h = \sum_{k=1}^{n_{nod}} N^k u_k = \boldsymbol{N}\boldsymbol{u} \qquad \text{in } T$$

$$\nabla w^h = \sum_{k=1}^{n_{nod}} \nabla N^k w_k = \mathbf{B}\boldsymbol{w} \quad , \quad \nabla u^h = \sum_{k=1}^{n_{nod}} \nabla N^k u_k = \mathbf{B}\boldsymbol{u}$$

Wobei n_{nod} die Anzahl der Knoten pro Element sind. Insbesondere für lineare Dreieckselemente ($n_{nod} = 3$) gilt dementsprechend

$$\mathbf{N} = [N^1, N^2, N^3] \ , \quad \mathbf{B} = \begin{bmatrix} N^1_{,x}, N^2_{,x}, N^3_{,x} \\ N^1_{,y}, N^2_{,y}, N^3_{,y} \end{bmatrix}, \quad \boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \ , \quad \boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}.$$

Die Darstellung der Anteile der einzelnen Terme der schwachen Form erfolgt in Matrizenschreibweise:

$$(1) \ \int_T (\mathbf{D} \cdot \nabla u^h) \nabla w^h \, \mathrm{d}\Omega_e \quad \rightsquigarrow \quad \boldsymbol{w}^t \underbrace{\int_T \mathbf{B}^t \mathbf{D} \mathbf{B} \, \mathrm{d}\Omega_e}_{\mathbf{k} \ : \ \text{symm}} \boldsymbol{u} = \boldsymbol{w}^t \mathbf{k} \boldsymbol{u}$$

$$(2) \ \int_T \boldsymbol{b} \cdot \nabla u^h \ w^h \, \mathrm{d}\Omega_e \quad \rightsquigarrow \quad \boldsymbol{w}^t \underbrace{\int_T \mathbf{N}^t \boldsymbol{b} \mathbf{B} \, \mathrm{d}\Omega_e}_{\mathbf{g} \ : \ \text{unsymm}} \boldsymbol{u} = \boldsymbol{w}^t \mathbf{g} \boldsymbol{u}$$

$$(3) \ \alpha \int_T u^h w^h \, \mathrm{d}\Omega_e \quad \rightsquigarrow \quad \boldsymbol{w}^t \alpha \underbrace{\int_T \mathbf{N}^t \mathbf{N} \, \mathrm{d}\Omega_e}_{\mathbf{m} \ : \ \text{symm}} \boldsymbol{u} = \boldsymbol{w}^t \alpha \mathbf{m} \boldsymbol{u}$$

$$(4),(5) \ \int_T f w^h \, \mathrm{d}\Omega_e + \int_{\Gamma_N} t_N w^h \, \mathrm{d}\Gamma_e \quad \rightsquigarrow \quad \boldsymbol{w}^t \underbrace{\left(\int_T f \mathbf{N}^t \, \mathrm{d}\Omega_e + \int_{\Gamma_N \cap T} t_N \mathbf{N}^t \, \mathrm{d}\Gamma_e \right)}_{\boldsymbol{f}} = \boldsymbol{w}^t \boldsymbol{f}$$

In Anhang 1 ist eine kurze Darstellung sogenannter linearer Dreieckselemente und eine Übersicht über die zur Implementierung nötigen Elementmatrizen zu finden.
Der Zusammenbau der lokalen Größen, in n_{el} Elementen, zu den globalen Größen erfolgt formal mittels der Boolschen Matrizen und führt auf die Form:

$$\begin{aligned} \bigcup_{e=1}^{n_{el}} \boldsymbol{w}^t \ \{ \underbrace{\mathbf{k} + \mathbf{g} + \alpha \mathbf{m}}_{\text{lokale Größen}} \} \ \boldsymbol{u} \ &= \ \bigcup_{e=1}^{n_{el}} \boldsymbol{w}^t \boldsymbol{f} \\ \rightarrow \ \boldsymbol{W}^t \ \{ \underbrace{\mathbf{K} + \mathbf{G} + \alpha \mathbf{M}}_{\text{globale Größen}} \} \ \boldsymbol{U} \ &= \ \boldsymbol{W}^t \boldsymbol{F} \qquad \forall \ \boldsymbol{W} \end{aligned}$$

Durch Zusammenfassen der Matrizen zu einer Systemmatrix $\mathbf{A}$, die durch den Anteil $\mathbf{G}$ unsymmetrisch ist,

$$\mathbf{A} \ = \ \mathbf{K} + \alpha \mathbf{M} + \mathbf{G} = \mathbf{E} + \mathbf{G}$$

$$\begin{aligned} \text{mit} \qquad & \mathbf{E} = \mathbf{K} + \alpha \mathbf{M} : \text{symmetrisch}, \\ & \mathbf{G} : \text{unsymmetrisch} \end{aligned}$$

gelangt man auf das globale Gleichungssystem $\mathbf{A}\boldsymbol{U} = \mathbf{F}$.

Der Einfluß der Konvektion gegenüber der Diffusion wird durch die dimensionslose *Peclet-Zahl* wiedergegeben, $\mathrm{Pe} = |\boldsymbol{b}|L/\epsilon$, wobei L eine charakteristische Längenabmessung bezeichnet, auf die später noch eingegangen werden soll.
Die Standard-Galerkin-Methode arbeitet für relativ große Werte von ϵ ($\epsilon > h$) zufriedenstellend. Aber für kleine ϵ bzw. große Peclet-Zahlen, d.h. wenn $\epsilon \to 0$ und die Konvektion dominant ist, treten in der berechneten Lösung Oszillationen auf, sofern die exakte Lösung nicht glatt ist, z.B. Grenzschichten und Diskontinuitäten enthält. Diese Oszillationen werden zudem über das gesamte Gebiet mit der Advektion transportiert und machen die numerische Lösung wegen ihres unphysikalischen Verhaltens unbrauchbar. Dieser Effekt begründet sich aus der schwachen Stabilität von Methoden vom Zentralen-Differenzen-Typ für kleine Werte von ϵ. Die Methode arbeitet nur dann gut, sofern die lokale Netzweite derart verfeinert wird, daß die Konvektion auf Elementebene nicht über die Diffusion dominiert.
Eine Möglichkeit, die Stabilität von Festgitter-Methoden wie der FEM zu erhöhen, besteht darin, künstliche Diffusion in das numerische Schema zu bringen. Die einfachste Art stellt das Addieren eines Diffusionsterms $-\delta\Delta u$, mit $\delta = h - \epsilon$, auf der linken Seite der Differentialgleichung dar. Danach wird auf die modifizierte Differentialgleichung die Galerkin-Methode angewendet. Dies führt auf stabile, aber überdiffusive Lösungen, die insbesondere starke "Crosswind Diffusion", d.h. Diffusion senkrecht zur Stromrichtung, aufweisen.
Eine weitere Art, die Stabilität zu sichern, besteht darin, einen Term $-\dfrac{\partial^2 u}{\partial b^2} =: \partial_{bb} u$, wobei $\dfrac{\partial u}{\partial b} = \boldsymbol{b} \cdot \nabla u$ die Ableitung in Stromrichtung ist, hinzuzufügen. Da das Originalproblem gestört wurde, fehlt es dieser Methode an Konsistenz und die beste zu erwartende Konvergenzrate liegt bei $\mathcal{O}(\delta)$ [22].
Ein anderer Ansatz läßt die Differentialgleichung unberührt und führt numerische Diffusion in die schwache Form durch Wichten des Konvektionsterms mit einer um einen Term erweiterten Testfunktion ein. Diese Idee entstammt wiederum den Finiten Differenzen Verfahren, wo eine nicht-zentrale Differenzen-Approximation unter Berücksichtigung der Flußrichtung für die erste Ableitung gewählt wird. Für die FEM entspricht dies einem *Petrov-Galerkin-Verfahren*, d.h. die Testfunktionen unterscheiden sich von den Ansatzfunktionen. Allerdings sei betont, daß bei diesem Vorgehen nur der Konvektionsterm in dieser Art gewichtet wurde, was auch dort zu einem Mangel an Konsistenz führt.
Um zu einer stabilen und konsistenten Methode zu gelangen, stellten Hughes & Brooks 1982 die *streamline upwind/Petrov-Galerkin method (SUPG)* vor [30], deren mathematischer Hintergrund durch Johnson und seine Mitarbeiter beleuchtet wird [36]. In der mathematischen Literatur wird der Name *streamline diffusion method (SDFEM)* für diese im folgenden beschriebene Methode bevorzugt.
Der Übersicht halber soll die *Galerkin/least-square method* nicht unerwähnt bleiben. Diese Methode wurde durch Hughes und Mitarbeiter im wesentlichen für das Stokes-Problem begründet, für welches sich beweisen läßt, daß die Babuška-Brezzi-Bedingung mittels dieser Methode umgangen werden kann. Da aber diese Methode für das Konvektions-Diffusionsproblem keine Vorteile gegenüber der SUPG-Methode aufzuweisen scheint [11], soll in diesem Rahmen nicht weiter auf sie eingegangen werden.

3.2 Streamline Upwind/Petrov-Galerkin, SDFEM

Die zugrundeliegende Idee der SUPG Methode entstammt dem *upwinding* der Finiten Differenzen Methode. Den Knoten, die stromaufwärts (upwind) liegen, wird mehr Gewicht verliehen als denen, die stromabwärts liegen. Für die FEM wurde diese Idee zuerst durch Christie, Griffiths, Mitchell & Zienkiewicz (1976) [8] im eindimensionalen Kontext angewendet. Die Testfunktion wurde aus der Ansatzfunktion und der Addition einer Modifikation höherer Ordnung gewonnen, siehe Abbildung 3.1(2) . Die Abbildung 3.1 zeigt die Ansatzfunktion N_i und die Testfunktion W_i des Bubnov-Galerkin Verfahrens (1), sowie die modifizierten Testfunktionen W_i nach Christie et al. (2) und nach Hughes & Brooks (3) für den eindimensionalen Fall.

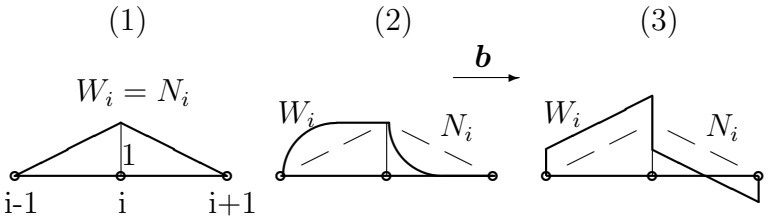


Abbildung 3.1: Ansatzfunktionen und modifizierte Testfunktionen

Einen konzeptionell anderen Ansatz wählten Hughes & Brooks [30]. Im Falle der SDFEM bzw. SUPG ergibt sich die Testfunktion aus der Addition der Ansatzfunktion und der Ableitung der Ansatzfunktion, siehe Abbildung 3.1(3). Diese Methode hat sich als generell sehr geeignet für hyperbolische Probleme herausgestellt. In dem hier betrachteten Fall ergibt sich die Testfunktion zu: $w = v + \delta \boldsymbol{b} \cdot \nabla v =: v + v_b$ mit v der üblichen Testfunktion wie zuvor. Durch Wichtung der gesamten Differentialgleichung mit der Testfunktion w ergibt sich folgende schwache Form:

$$\begin{aligned} & -\int_\Omega \nabla \cdot (\mathbf{D} \cdot \nabla u^h) v_b^h \,\mathrm{d}\Omega + \int_\Omega (\mathbf{D} \cdot \nabla u^h) \nabla v^h \,\mathrm{d}\Omega + \int_\Omega \boldsymbol{b} \cdot \nabla u^h (v^h + v_b^h) \,\mathrm{d}\Omega \\ + \quad & \alpha \int_\Omega u^h (v^h + v_b^h) \,\mathrm{d}\Omega = \int_\Omega f (v^h + v_b^h) \,\mathrm{d}\Omega + \int_{\Gamma_N} t_N v^h \,\mathrm{d}\Gamma \end{aligned}$$

Der erste Term auf der linken Seite ist so für FE-Funktionen u^h unsinnig. Er muß als Summe der entsprechenden elementweisen Terme interpretiert werden

$$\int_\Omega \Delta u^h v_b^h \,\mathrm{d}\Omega = \sum_{T \in \mathcal{T}_h} \int_T \Delta u^h v_b^h \,\mathrm{d}\Omega.$$

Zur Vereinfachung wird er aber wie zunächst oben dargestellt. Allerdings muß u in dem soeben beschriebenen Sinne interpretiert werden.
Die gesamte Gleichung stellt sich, unter Fortlassung der Kopfzeiger h, wie folgt in der schwachen Form für das SUPG-Verfahren dar:

$$\int_\Omega \overset{(1)}{(\mathbf{D} \cdot \nabla u)} \nabla v \,\mathrm{d}\Omega + \int_\Omega \overset{(2)}{\boldsymbol{b} \cdot \nabla u}\ v \,\mathrm{d}\Omega + \alpha \int_\Omega \overset{(3)}{uv} \,\mathrm{d}\Omega$$

$$
\begin{aligned}
- & \overset{(1a)}{\int_\Omega \nabla\cdot(\mathbf{D}\cdot\nabla u)\delta\boldsymbol{b}\cdot\nabla v\,\mathrm{d}\Omega} + \overset{(2a)}{\int_\Omega \boldsymbol{b}\cdot\nabla u\ \delta\boldsymbol{b}\cdot\nabla v\,\mathrm{d}\Omega} + \alpha\overset{(3a)}{\int_\Omega u\ \delta\boldsymbol{b}\cdot\nabla v\,\mathrm{d}\Omega} \\
= & \overset{(4)}{\int_\Omega fv\,\mathrm{d}\Omega} + \overset{(4a)}{\int_\Omega f\ \delta\boldsymbol{b}\cdot\nabla v\,\mathrm{d}\Omega} + \overset{(5)}{\int_{\Gamma_N} t_N v\,\mathrm{d}\Gamma}
\end{aligned}
$$

Der Term $(2a)$ beschreibt zusätzliche Diffusion in Stromrichtung:

$$
\delta\int_\Omega \boldsymbol{b}\cdot\nabla u\ \boldsymbol{b}\cdot\nabla v\,\mathrm{d}\Omega \qquad \rightsquigarrow \qquad -\delta\underbrace{\partial_{bb}u}_{\text{2.Ableitung von u in Richtung } \boldsymbol{b}}
$$

Es werden wie beim Vorgehen mit dem Standard-Galerkin Verfahren lineare Ansätze für die Konzentration u und die Testfunktion v bzw. deren Gradienten gewählt:

$$
\begin{aligned}
v^h &= \sum_{k=1}^{n_{nod}} N^k v_k = \boldsymbol{N}\boldsymbol{v} \qquad \text{in } T \\
\nabla v^h &= \sum_{k=1}^{n_{nod}} \nabla N^k v_k = \mathbf{B}\boldsymbol{v} \\
v_b^h &= \delta\boldsymbol{b}\sum_{k=1}^{n_{nod}} \nabla N^k v_k = \delta\boldsymbol{b}\mathbf{B}\boldsymbol{v}
\end{aligned}
$$

Für die gewählten linearen Ansatzfunktionen verschwindet der Term $(1a)$ ($\Delta u = 0$). Die Terme $(1), (2), (3), (4), (5)$ entsprechen denen beim Standard-Galerkin Verfahren dargestellten. Die zusätzlichen Terme lassen sich analog wie folgt in Matrizenschreibweise darstellen:

$$
(1a)\ \int_T \nabla\cdot(\mathbf{D}\cdot\nabla u)\delta\boldsymbol{b}\cdot\nabla v\,\mathrm{d}\Omega_e \qquad \rightsquigarrow \qquad \text{verschwindet für lineare Ansätze}
$$

$$
(2a)\ \int_T \boldsymbol{b}\cdot\nabla u\ \delta\boldsymbol{b}\cdot\nabla v\,\mathrm{d}\Omega_e \qquad \rightsquigarrow \qquad \boldsymbol{v}^t\underbrace{\int_T \delta\mathbf{B}^t\boldsymbol{b}^t\boldsymbol{b}\mathbf{B}\,\mathrm{d}\Omega_e}_{\mathbf{k}^{SD}\ :\ \text{symm.}}\boldsymbol{u} = \boldsymbol{v}^t\mathbf{k}^{SD}\boldsymbol{u}
$$

$$
(3a)\ \alpha\int_T u\ \delta\boldsymbol{b}\cdot\nabla v\,\mathrm{d}\Omega_e \qquad \rightsquigarrow \qquad \boldsymbol{v}^t\underbrace{\int_T \alpha\delta\mathbf{B}^t\boldsymbol{b}^t\boldsymbol{N}\,\mathrm{d}\Omega_e}_{\alpha\delta\mathbf{g}^t\ :\ \text{unsymm.}}\boldsymbol{u} = \boldsymbol{v}^t\alpha\delta\mathbf{g}^t\boldsymbol{u}
$$

$$
(4a)\ \int_T f\ \delta\boldsymbol{b}\cdot\nabla v\,\mathrm{d}\Omega_e \qquad \rightsquigarrow \qquad \boldsymbol{v}^t\underbrace{\int_T \delta\ \mathbf{B}^t\boldsymbol{b}^t f\,\mathrm{d}\Omega_e}_{\boldsymbol{f}^{SD}} = \boldsymbol{v}^t\boldsymbol{f}^{SD}
$$

Im Anhang 1 befindet sich eine Zusammenstellung der entstehenden Elementmatrizen. Der Zusammenbau der lokalen Größen, in n_{el} Elementen, zu den globalen Größen erfolgt wie oben formal mittels der Boolschen Matrizen und führt auf die Form:

$$\bigcup_{e=1}^{n_{el}} \boldsymbol{v}^t \underbrace{\{\mathbf{k}+\mathbf{g}+\alpha\mathbf{m}+\mathbf{k}^{SD}+\alpha\delta\mathbf{g}^t\}}_{\text{lokale Größen}} \boldsymbol{u} \;=\; \bigcup_{e=1}^{n_{el}} \boldsymbol{v}^t\{\boldsymbol{f}+\boldsymbol{f}^{SD}\}$$

$$\rightarrow\; \boldsymbol{V}^t \underbrace{\{\mathbf{K}+\mathbf{G}+\alpha\mathbf{M}+\mathbf{K}^{SD}+\alpha\delta\mathbf{G}^t\}}_{\text{globale Größen}} \boldsymbol{U} \;=\; \boldsymbol{V}^t\bar{\mathbf{F}} \qquad \forall\, \boldsymbol{V}$$

Die einzelnen Matrizen werden wiederum zu einer insgesamt unsymmetrischen Matrix $\bar{\mathbf{A}}$ zusammengefaßt

$$\bar{\mathbf{A}} \;=\; \mathbf{K}+\mathbf{G}+\alpha\mathbf{M}+\mathbf{K}^{SD}+\alpha\delta\mathbf{G}^t = \bar{\mathbf{E}}+\bar{\mathbf{M}}$$

$$\begin{aligned} \text{mit} \qquad & \bar{\mathbf{E}} = \mathbf{K}+\alpha\mathbf{M}+\mathbf{K}^{SD} : \text{symmetrisch} \\ & \bar{\mathbf{M}} = \mathbf{G}+\alpha\delta\mathbf{G}^t : \text{unsymmetrisch.} \end{aligned}$$

Damit erhält man das globale Gleichungssystem $\bar{\mathbf{A}}\bar{\boldsymbol{U}} = \bar{\boldsymbol{F}}$.

Die Arbeit zur mathematischen Analysis der SDFEM wurde von Johnson und seinen Mitarbeitern geleistet. Sie zeigten, daß es sich bei dieser Methode um eine Methode höherer Genauigkeit handelt, d.h. der Fehler ist $\mathcal{O}(h^{k+1/2})$, wobei k die Ordnung des approximierenden Polymons ist. Desweiteren wurde die Stabilität für die Ableitung in Stromrichtung gezeigt [34].

3.2.1 SUPG-Parameter δ

Die explizite Wahl des Faktors δ ist nicht offensichtlich und in einer Vielzahl von Veröffentlichungen werden unterschiedliche Ansätze gewählt. Eine umfangreiche Zusammenstellung findet sich in [12]. In einigen Veröffentlichungen wird der Parameter δ als 'intrinsic time scale' bezeichnet.
Allen Definitionen ist gemeinsam, daß der Faktor wächst, wenn ϵ fällt, und für die Wahl $\delta = 0$ das SUPG-Verfahren in die Standard-Galerkin Methode übergeht. Für den rein konvektiven Fall ($\epsilon = 0$) wird in [36] der Wert $\delta \sim h/|\boldsymbol{b}|$ angegeben, dieser Wert wird u.a. auch in [22] verwendet. In dem Buch [34] wird für den allgemeinener Fall ($\epsilon > 0$) der folgende Vorschlag gemacht:

$$\delta_e = \begin{cases} ch_e, & \text{wenn } \epsilon < h_e, \text{ mit } c > 0, \text{ klein} \\ 0, & \text{wenn } \epsilon \geq h_e \end{cases}$$

Diese Vorgehensweise ist nicht geeignet, wenn $|\boldsymbol{b}|$ variabel ist und wenn nicht-quasi-uniforme Triangulierungen $\mathcal{T}_h$ vorliegen. Für diese Fälle empfiehlt sich

$$\delta_e = \begin{cases} \frac{ch_e}{|\boldsymbol{b}|} & \text{auf } T \in \mathcal{T}_h, \quad \text{wenn } \epsilon < h_e|\boldsymbol{b}| \\ 0, & \text{wenn } \epsilon \geq h_e|\boldsymbol{b}| \end{cases} .$$

Ein weiterer Ansatz wird in [35] von der Form

$$\delta_e = \frac{c_1}{|\boldsymbol{b}|} \max\{h_e - \frac{\epsilon}{|\boldsymbol{b}|}, 0\} \ , \quad \text{mit z.B. } c_1 = 0.5 \quad \text{(siehe in Tab.3.1 Zeile 7)}$$

gegeben. Es zeigt sich, daß der Faktor δ ortsabhängig ist, zum einen von der Elementgröße und zum anderen von den lokalen Parametern wie der Geschwindigkeitsverteilung und Diffusivität. Insbesondere läßt sich δ als Funktion der lokalen Peclet-Zahl $\mathrm{Pe}^h = \frac{|\boldsymbol{b}|h_e}{2\epsilon}$ als

$$\delta_e = \frac{h_e}{2|\boldsymbol{b}|} \ \zeta(\mathrm{Pe}^h)$$

definieren. Wesentlich für die Wahl der Funktion $\zeta(\mathrm{Pe}^h)$ ist das vorausgesetzte asymptotische Verhalten für δ_e:

$$\delta_e = \mathcal{O}(\frac{h_e}{|\boldsymbol{b}|}) \quad , \quad \text{wenn } \mathrm{Pe}^h \to \infty$$

$$\delta_e = \mathcal{O}(\frac{h_e^2}{\epsilon}) \quad , \quad \text{wenn } \mathrm{Pe}^h \to 0$$

In dieser Form sind einige in der Literatur genannten Varianten für $\zeta(\mathrm{Pe}^h)$ in der Tabelle 3.1 zusammengefaßt.

Die Wahl der charakteristischen Länge h_e ist nicht von unerheblichem Einfluß auf δ_e, [12]. Verschiedene Ansätze sind in der Tabelle 3.2 wiedergegeben.
Die erste Wahl entspricht der üblicherweise in der Theorie getroffenen, nämlich der längsten Kante des speziellen Dreiecks. Der zweite Vorschlag geht auf Mizukami [46] und Tezduyar [54] zurück. Es wird die maximale Länge des Schnittes zwischen dem Dreieck und einer zum Geschwindigkeitsvektor $\boldsymbol{b}$ parallelen Linien durch den Schwerpunkt berechnet. Dieser Ansatz eignet sich besonders für gestreckte Dreiecke. Die dritte Definition entspricht dem Durchmesser eines flächengleichen Kreises, wobei A_e den Flächeninhalt des Elements angibt. Verwendung findet diese Definition vorallem bei isotropen Dreiecken. Der letzte Ausdruck stammt von Harari [24], wobei $\boldsymbol{x}_i$ die Knotenkoordinaten bezeichnet und $\boldsymbol{x}_S$ die Schwerpunktkoordinaten.
Im Rahmen der Beispiele innerhalb dieser Arbeit wurde, sofern nicht anders angegebenen, für die Funktion $\zeta(\mathrm{Pe}^h)$ des Parameters δ die Variante 7 nach Johnson und für h_e die längste Kante gewählt.

3.3 Shock-Capturing Modifikationen

Die SDFEM erzielt zwar stabile Lösungen, aber ein Über- und Unterschwingen der numerischen Lösung nahe von Sprung-Diskontinuitäten der exakten Lösung kann auch

	$\zeta(\mathrm{Pe}^h)$	Bezeichnung	Literatur
1.	$2c$	s.o. Hansbo	[22]
2.	$2c$, wenn $\epsilon < h_e\|\boldsymbol{b}\|$, sonst 0	s.o.	[34]
3.	$\coth(\mathrm{Pe}^h) - \frac{1}{\mathrm{Pe}^h}$	optimal	[8]
4.	$\min\{1, \frac{\mathrm{Pe}^h}{3}\}$	doppelt asymptotisch	[31]
5.	$\frac{\mathrm{Pe}^h}{1+\mathrm{Pe}^h}$	Mizukami	[46]
6.	$\max\{0, 1 - \frac{1}{\mathrm{Pe}^h}\}$	kritisch	[8]
7.	$\max\{0, 1 - \frac{1}{2\mathrm{Pe}^h}\}$	Johnson	[35]
8.	$\frac{\mathrm{Pe}^h}{3}\left(1 + \left(\frac{\mathrm{Pe}^h}{3}\right)^2\right)^{-1/2}$	Hughes ,Tezduyar	[26], [54]
9.	$\mathrm{Pe}^h\left(1 + \left(\mathrm{Pe}^h\right)^2\right)^{-1/2}$	Tezduyar	[55]

Tabelle 3.1: Funktionen $\zeta(\mathrm{Pe}^h)$

	h_e
1.	$h_e = \max\ h_i\ ,\ i = 1,2,3$
2.	$h_e^{-1} = \frac{1}{2}\sum_{i=1}^{3}\left\|\frac{\boldsymbol{b}}{\|\boldsymbol{b}\|}\cdot\nabla N_i\right\|$
3.	$h_e = \sqrt{\frac{4A_e}{\pi}}$
4.	$h_e^{-1} = \frac{1}{4A_e}\left[3\sum_{i=1}^{3}\|\boldsymbol{x}_i - \boldsymbol{x}_S\|^2\right]^{1/2}$

Tabelle 3.2: Charakteristische Längen h_e

durch diese Methode nicht völlig ausgeschlossen werden. Dies erklärt sich aus der Tatsache, daß diese Methode keine *monotone* und auch keine *monotonieerhaltende* Methode ist. Eine monotone Methode erhält von einem Zeitschritt zum nächsten die Vorzeichen der numerischen Lösung an allen Knoten des räumlichen Netzes. Im monotonieerhaltenden Fall wird die Monotonie der Anfangsdaten erhalten. Die Entwicklung einer monotonen Methode höherer Genauigkeit ist nicht trivial. Mit Godunovs Theorem gelangt man zu einer linearen, monotonieerhaltenden Methode, die aber im besten Fall von erster Ordnung genau ist [11]. Einen sinnvollen Weg, um hohe Genauigkeit in Regionen mit glatter numerischer Lösung zu erzielen und Oszillationen an Grenzschichten zu vermeiden, stellt die Konstruktion einer nichtlinearen Methode dar, also eines Schemas, das selbst von der numerischen Lösung abhängt. Die grundlegende Idee dabei ist, die Dissipation in der Nähe von Grenzschichten zu erhöhen.

3.3.1 Shock-Capturing Modifikation nach Hughes

Die Oszillationen, welche bei der Benutzung der SDFEM beobachtet werden können, liegen in Richtung normal zu dem Gradienten der transportierten Größe bzw. der Konzentration. Aus diesem Grund entwickelten Hughes und seine Mitarbeiter [47], [27] eine Modifikation der SDFEM, die weitere Diffusion in diese Richtung einführt. Dieser neue Diffusionsterm wird in die Wichtungsfunktion eingearbeitet und als *Discontinuity-Capturing Modifikation* bezeichnet. Um konsistent zu sein, muß die neu eingeführte Dissipation proportional zum Elementresiduum sein und zugunsten der Genauigkeit muß sie schnell in den Gegenden verschwinden, in denen die Lösung glatt ist und der konvektive Term des Residuums klein ist. Bei Johnson und seinen Mitarbeitern wird sie als *Shock-Capturing Modifikation* (SC) der SDFEM bezeichnet. Die theoretische Begündung zur Einführung der Shock-Capturing Modifikation sind in [53] und [41] gegeben.
Die Testfunktion nimmt mit dem modifizierten Term die folgende Gestalt an.

$$w = v + \delta\ \boldsymbol{b} \cdot \nabla v + \bar{\delta}\ \bar{\boldsymbol{b}} \cdot \nabla v$$

Der zusätzliche Term in der Testfunktion kann wiederum als eine Form von künstlicher Diffusion erster Ordnung betrachtet werden, wobei der Diffusionskoeffizient von dem Residuum der FE-Lösung abhängt. Die Robustheit der SDFEM wird insbesondere bei der Präsenz von Diskontinuitäten, wie Schockfronten, erhöht, da die numerische Diffusion dort eingeführt wird, wo das Residuum groß ist, aber nur kleine Werte in glatten Gegenden gibt.
Das modifizierte Geschwindigkeitsfeld $\bar{\boldsymbol{b}}$ stellt eine Projektion von $\boldsymbol{b}$ auf ∇u^h dar.

$$\bar{\boldsymbol{b}} = \begin{cases} 0 & \text{falls } \nabla u^h = 0 \\ \dfrac{\boldsymbol{b} \cdot \nabla u^h}{\left|\nabla u^h\right|^2} \nabla u^h & \text{falls } \nabla u^h \neq 0 \end{cases}$$

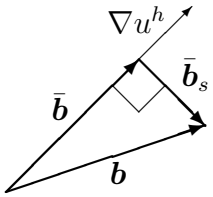


Abbildung 3.2: Modifiziertes Geschwindigkeitsfeld

Da $\bar{\boldsymbol{b}}$ von der unbekannten diskreten Lösung abhängt, führt diese Modifikation auf eine nichtlineare Methode, obwohl das zugrundeliegende Problem linear ist. Bei der Berechnung wird zunächst ein Schritt mit der SDFEM ohne die SC-Modifikation vorgenommen und die ermittelte Lösung für das Shock-Capturing verwendet. Die so gewonnene Lösung kann als Startwert für eine "äußere" Fixpunkt-Iteration verwendet werden:

$$\begin{aligned} \text{Startwert} \ : &\quad x_1 = f(x_0) \\ \text{Iteration} \ \ i = 1, n \ : &\quad x_{i+1} = f(x_i) \\ \text{Abbruch, wenn} \ : &\quad |x_{i+1} - x_i| < tol. \end{aligned}$$

Es wird also solange iteriert, bis keine nennenswerte Verbesserung mehr erzielt werden kann.

Die Wahl des modifizierten Faktors $\bar{\delta}$ kann analog zu der oben beschriebenen Wahl von δ getroffen werden, wobei die modifizierte Geschwindigkeit zu berücksichtigen ist. Im einfachsten Fall z.B.:

$$\bar{\delta} = \frac{h}{|\bar{\boldsymbol{b}}|} \quad \text{oder} \quad \bar{\delta} = ch$$

Das Geschwindigkeitsfeld $\boldsymbol{b}$ läßt sich, wie oben dargestellt, in die Komponenten $\bar{\boldsymbol{b}}$ und $\bar{\boldsymbol{b}}_s = \boldsymbol{b} - \bar{\boldsymbol{b}}$ zerlegen. Man erhält die Beziehung

$$\bar{\boldsymbol{b}} \cdot \nabla u^h = \boldsymbol{b} \cdot \nabla u^h.$$

Bei der Diskretisierung entstehen der SC-Modifikation entsprechende Elementmatrizen (mit $\bar{\delta}$, $\bar{\boldsymbol{b}}$), diese sind ebenfalls im Anhang 1 aufgeführt.
Setzt man die um den SC-Term erweiterte Testfunktion in die schwache Form ein und nutzt die obige Beziehung, so erkennt man, daß durch die Modifikation eine verbesserte Stabilität erreicht wurde [27]. Durch den Zusammenhang $\bar{\boldsymbol{b}} \cdot \nabla u^h = \boldsymbol{b} \cdot \nabla u^h$ erhält man unter dem Integral der schwachen Form u.a. die Terme:

$$.... \; v\boldsymbol{b} \cdot \nabla u^h + \nabla v \cdot \delta \; \boldsymbol{b}\boldsymbol{b}^t \nabla u^h + \nabla v \cdot \bar{\delta} \; \bar{\boldsymbol{b}}\bar{\boldsymbol{b}}^t \nabla u^h ....$$

Der letzte dieser Terme gibt die Kontrolle über die Gradienten in Richtung von ∇u^h.
Bei der oben vorgeschlagenen Wahl für $\bar{\delta}$ ergibt sich bei $\bar{\boldsymbol{b}} = \boldsymbol{b}$ eine Verdopplung von δ entlang der Stromlinien. Um dies zu umgehen, kann man die Komponente $\delta \boldsymbol{b}\boldsymbol{b}^t$ in Richtung von $\bar{\boldsymbol{b}}\bar{\boldsymbol{b}}^t$ durch die Wahl

$$\bar{\delta} = \max\{0, \delta(\bar{\boldsymbol{b}}) - \delta(\boldsymbol{b})\}$$

umgehen. Dabei wird schließlich das Maximum zwischen der durch die SDFEM und die SC-Modifikation eingeführte Diffusion als Dissipation in Richtung der Stromlinie in die Rechnung eingebracht.

3.4 Numerische Beispiele

In diesem Abschnitt soll anhand von Beispielen die Vorteile und Nachteile der zuvor beschriebenen Methoden veranschaulicht werden. In allen Rechnungen wurden lineare FE-Ansätze verwendet. Zur Lösung der entstehenden Gleichungssysteme wurde das im nächsten Kapitel beschriebene BiCG-Stab-Verfahren mit ILU-Vorkonditionierung gewählt. Die Netze entstanden aus uniformer Verfeinerung einer Ursprungstriangulierung von 2 Dreieckselementen über mehrere Verfeinerungsstufen.

3.4.1 Gekrümmte innere Schockfront

Als erstes numerisches Beispiel wird ein Konvektions-Diffusionsproblem auf dem Einheitsquadrat $\Omega = (0,1) \times (0,1)$ mit dominanter Konvektion, d.h. kleinem Diffusionskoeffizienten $\epsilon = 10^{-5}$, ohne Adsorption $\alpha = 0$ und Quellen oder Senken $f = 0$ betrachtet. Es seien auf dem Einflußrand Dirichlet-Bedingungen, $\Gamma_D = \{(x,y) \in \partial\Omega | x = 0$

oder $y = 1\}$, und auf dem Ausflußrand homogene Neumann-Bedingungen, $\Gamma_N = \partial\Omega\backslash\Gamma_D$, gegeben. Aufgrund des vorliegenden stationären Strömungsfeldes $\boldsymbol{b}(x, y) = (y, -x)$ stellt sich ein Konzentrationsplateau um den Koordinatenursprung ein.
Die folgende Abbildung zeigt oben die dreidimensionale Darstellung der numerischen Lösungen aus der Galerkin Methode und der SDFEM. Es wurde $\zeta(\mathrm{Pe}^h)$ nach Johnson gewählt und $h_e = \max\ h_i, i = 1, 2, 3$. Die untere Zeile zeigt Ergebnisse der SDFEM mit Shock-Capturing, links wurde $\bar{\delta}$ entsprechend δ gewählt und rechts wurde für $\bar{\delta}$ die angegebene max-Bedingung gefordert.

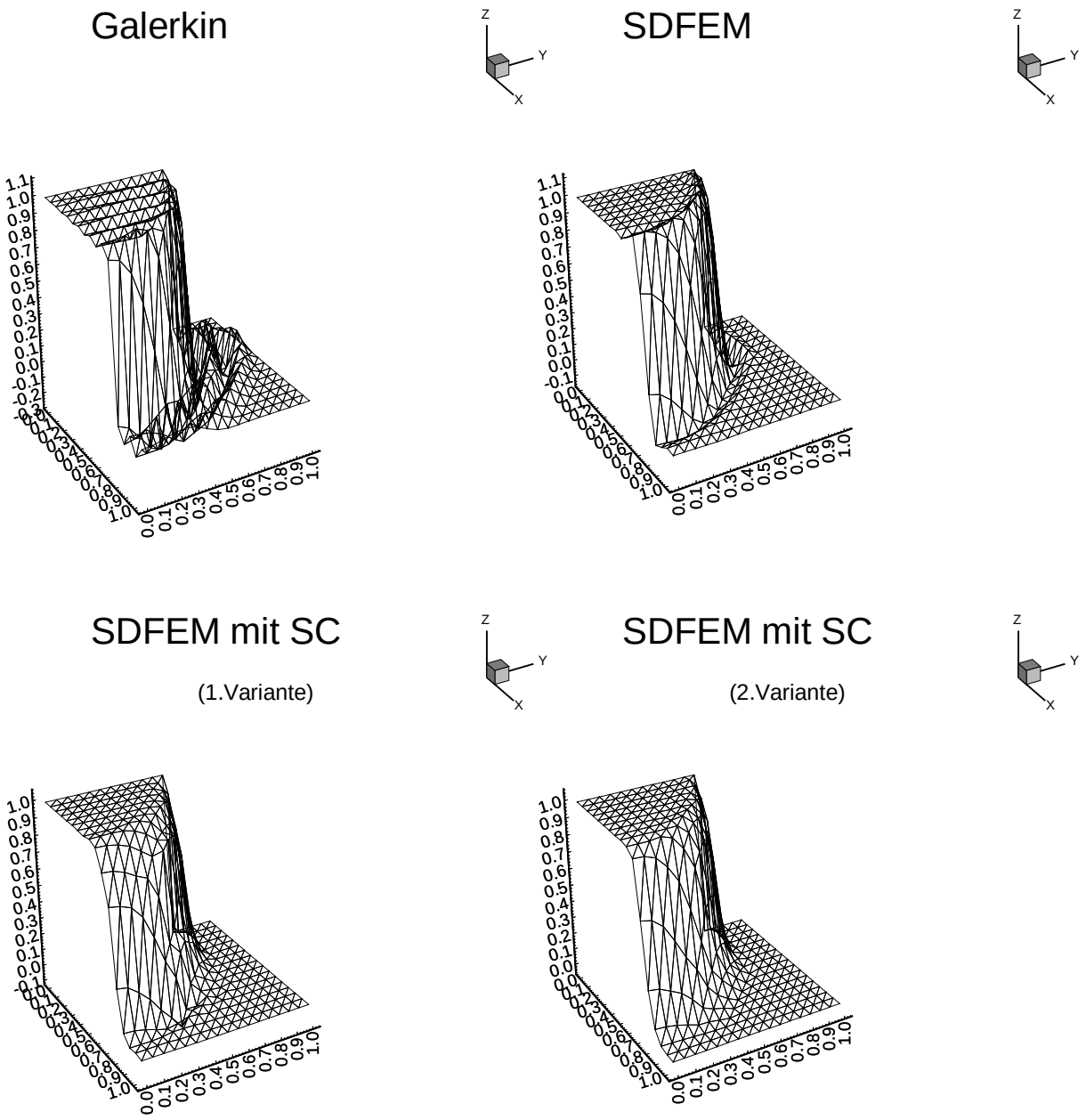


Abbildung 3.3: Gekrümmte innere Schockfront

Anhand dieses Beispiels sollen auch die verschiedenen Möglichkeiten zur Wahl des SUPG-Faktors δ verglichen werden. Es wurde wiederum uniform über 5 Verfeinerungsstufen verfeinert (=2048 Elemente). Die charakteristische Länge wurde zu $h_e = \max\ h_i, i = 1, 2, 3$ für alle Alternativen von δ gewählt. Die folgende Tabelle 3.3 zeigt für die verschiedenen Faktoren den globalen Fehler der numerischen Lösung. Da eine analytische Lösung für das betrachtete Problem nicht vorlag, wurde als "exakte" Lösung zur Ermittlung des globalen Fehlers die Lösung des hyperbolischen Grenzfalls ($\epsilon = 0$) herangezogen.

	δ	globaler Fehler
1.	$h/2\|\boldsymbol{b}\|$, wenn $\epsilon < h\|\boldsymbol{b}\|$	0.068660
2.	$h/\|\boldsymbol{b}\|$	0.073945
3.	$h\min(h/\epsilon, 1)$	0.071085
4.	$\zeta(\mathrm{Pe}^h)$: optimal	0.068656
5.	$\zeta(\mathrm{Pe}^h)$: doubl. asy.	0.068660
6.	$\zeta(\mathrm{Pe}^h)$: Mizukami	0.068656
7.	$\zeta(\mathrm{Pe}^h)$: kritisch	0.068656
8.	$\zeta(\mathrm{Pe}^h)$: Johnson	0.068658
9.	$\zeta(\mathrm{Pe}^h)$: Hughes	0.068660
10.	$\zeta(\mathrm{Pe}^h)$: Tezduyar	0.068660

Tabelle 3.3: Vergleich der δ

Der Vergleich der verschiedenen h_e, siehe Tabelle 3.4, erfolgte analog zum vorherigen Vergleich mit der Wahl von δ nach Johnson für alle Varianten von h_e.

	h_e	globaler Fehler
1.	h_{max}	0.068658
2.	$0.7h_{max}$	0.066399
3.	$\sqrt{(4A_e)/\pi}$	0.065254
4.	h_e^{-1}	0.066455

Tabelle 3.4: Vergleich der h_e

Die Wahl verschiedener Versionen für den Faktor δ zeigt, abgesehen von den sehr simplen Varianten 2 und 3 verhältnismäßig wenig Einfluß auf den Fehler der numerischen Lösung. Die Wahl der charakteristischen Länge h_e hingegen beeinflußt die Größe des Fehlers weit deutlicher.

3.4.2 Innere Schockfront und Randgrenzschicht

Auch dieses Beispiel ist auf dem Einheitsquadrat $\Omega = (0,1) \times (0,1)$ gestellt. Es wird wiederum ein Konvektions-Diffussionsproblem mit dominanter Konvektion, d.h. $\epsilon = 10^{-5}$, ohne Quellen oder Senken $f = 0$, aber nun mit einer Adsorption $\alpha = 1$ betrachtet. Es seien auf dem gesamten Rand Dirichlet-Bedingungen gegeben, auf dem unteren und dem rechten Rand mit dem Wert $u = 1$, sonst $u = 0$. Wegen des Strömungsfeldes $\boldsymbol{b}(x, y) = (2, 1)$ stellt sich entlang der Stromlinien durch den Ursprung eine innere Schockfront ein. Am rechten Rand tritt ein Grenzschicht auf. Dort springen die Werte von 0 auf 1.

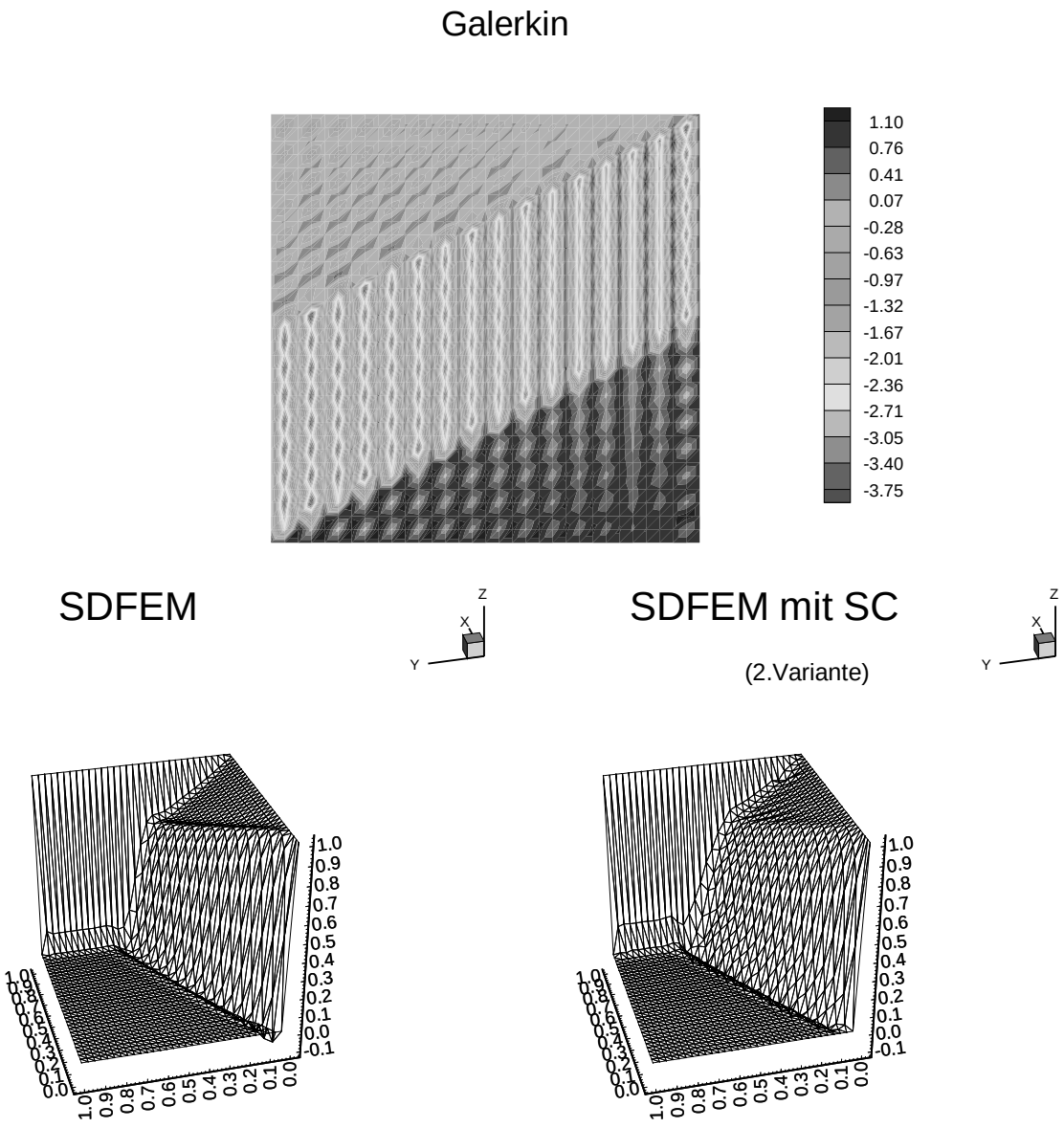


Abbildung 3.4: Schockfront und Randgrenzschicht

Die Abbildung 3.4 zeigt oben die zweidimensionale Kontur-Darstellung der Lösung der

Bobnov-Galerkin Methode. Aufgrund der großen auftretenden Oszillationen ist eine andere Darstellung nicht möglich, bzw. nicht anschaulich. Darunter ist links die dreidimensionale Darstellung der numerischen Lösungen aus der SDFEM zu finden. Es wurde $\zeta(\mathrm{Pe}^h)$ nach Johnson gewählt und $h_e = \max\ h_i, i = 1, 2, 3$. Unten rechts ist das Ergebnis der SDFEM mit Shock-Capturing mit der Forderung der angegebenen max-Bedingung für $\bar{\delta}$ ebenfalls dreidimensional dargestellt.

3.5 Zusammenfassung

Es ist deutlich zu sehen, daß für die vorgestellten Probleme mit dominanter Konvektion Oszillationen in den Lösungen aus dem Galerkin Verfahren auftreten. Die SDFEM vermag diese bis auf leichte Über- und Unterschwingungen nahe der Schockfronten zu glätten. Die verschiedenen Alternativen für δ, abgesehen von den ungünstigeren Varianten 2 und 3, sind dabei, wie im ersten Beispiel gezeigt, als ungefähr gleichwertig zu betrachten. Die Wahl der charakteristischen Länge h_e hingegen zeigt etwas mehr Einfluß auf den Fehler.
Durch Einführung von Shock-Capturing Modifikationen werden auch die leichten Schwingungen in den Lösungen aus der SDFEM eliminiert, wobei die zweite Variante der ersten etwas überlegen zu sein scheint. Allerdings wird die Vermeidung von Über- und Unterschwingungen mit einer leichten Ausschmierung der numerischen Lösung an diesen Stellen bezahlt. Außerdem kann der Rechenaufwand mit vielen SC-Iterationen stark zunehmen. Wenn der Diffusionskoeffizient nicht zu klein ist, $\epsilon > 10^{-3}$, d.h. die Konvektion nicht stark dominant, und damit ein starkes Über- und Unterschwingen der numerischen Lösung nicht zu erwarten, kann es praktikabel sein, auf ein Shock-Capturing zu verzichten.

Kapitel 4

Iterative Gleichungslösung

Bei der Diskretisierung partieller Differentialgleichungen treten meist sehr große Gleichungssysteme auf. *Direkte* Verfahren, welche nach endlich vielen Rechenschritten die (bis auf Rundungsfehler) exakte Lösung liefern, eignen sich wegen ihres relativ großen Rechenaufwandes, typischerweise werden n^2 Operationen benötigt, nur bedingt zur Lösung solcher Gleichungssysteme. Zumal die Systemmatrizen, die aus einer Diskretisierung entstanden sind, i.d.R. schwachbesetzt sind. Diese Gleichungssysteme eignen sich deshalb besonders gut zur Lösung mittels *iterativer* Verfahren.
Die erste Iterationsmethode für lineare Gleichungssysteme stammt von Carl Friedrich Gauß. Seine Methode der kleinsten Fehlerquadrate führte ihn auf größere Gleichungssysteme, die mit der direkten Methode der Gauß-Elimination nicht mehr bequem berechnet werden konnten. Das von ihm beschriebene Iterationsverfahren ist eine Variante des Gauß-Seidel-Verfahrens.
Seitdem elektronische Rechner zur Lösung von Gleichungssystemen herangezogen werden können, ist die Anzahl der Gleichungen um eine weitere Größenordnung gestiegen. Die obigen Verfahren erwiesen sich als zu langsam. In den 30er und 40er Jahren experimentierte Southwell mit Varianten der Gauß-Seidel-Methode. Durch eine wesentliche Beschleunigung des Gauß-Seidel-Verfahrens mit der sogenannten SOR-Methode gelang Young 1950 ein Durchbruch.
In der folgenden Zeit entstanden noch viele andere, noch effektivere Verfahren, wie das Verfahren der konjugierten Gradienten und das Mehrgitterverfahren [21].
Das Verfahren der konjugierten Gradienten (CG) hat sich für positiv definite Matrizen als sehr effizient erwiesen. Basierend auf ihm wurden Verfahren zur Anwendung auf indefinite oder unsymmetrische Probleme verallgemeinert. Diese Verfahren liefern nach spätestens n Schritten die exakte Lösung und sind insofern als iterative Verfahren zu verstehen, als daß sie bereits nach wenigen Iterationsschritten gute Annäherungen an die exakte Lösung geben. Im Rahmen dieser Arbeit wird insbesondere das BiCG-Stab-Verfahren Anwendung finden, da es auch auf indefinite und unsymmetrische Probleme, wie das hier vorliegende, anwendbar ist.

4.1 Conjugate Gradient Verfahren

Gegeben sei ein lineares reguläres Gleichungssystem $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$, wobei $\mathbf{A} \in \mathbb{R}^{n\times n}$ und $\boldsymbol{b} \in \mathbb{R}^n$. Die Verfahren basierend auf dem CG-Verfahren arbeiten nach folgendem Schema: Gegeben seien ein beliebiger Startvektor $\boldsymbol{x}_0 \in \mathbb{R}^n$ und das zugehörige Startresiduum $\boldsymbol{r}_0 = \boldsymbol{b} - \mathbf{A}\boldsymbol{x}_0$. Im folgenden wird mit $\boldsymbol{r}_i$ das Residuum $\boldsymbol{b} - \mathbf{A}\boldsymbol{x}_i$ bezeichnet.
Durch Aufaddieren eines Richtungsvektors $\boldsymbol{p}_i$ auf die bisherige Iterierte $\boldsymbol{x}_i$ wird die neue Iterierte $\boldsymbol{x}_{i+1}$ gewonnen:

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \alpha_i \boldsymbol{p}_i \ ,$$

wobei α_i die Schrittweite bestimmt. Die Bezeichnung Gradientenverfahren folgt daraus, daß für symmetrische Matrizen $\mathbf{A}$ das Residuum $\boldsymbol{r}_i = \boldsymbol{b} - \mathbf{A}\boldsymbol{x}_i$ als Suchrichtung gewählt wird, welches der Gradient der Funktion $F(\boldsymbol{x}) = \frac{1}{2}\langle \mathbf{A}\boldsymbol{x}, \boldsymbol{x}\rangle - \langle \boldsymbol{b}, \boldsymbol{x}\rangle$ an der Stelle $\boldsymbol{x}_i$ ist. Wie sich zeigt, löst die Minimalstelle von F das Gleichungssystem $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$.
In dem Verfahren der konjugierten Gradienten sind die Richtungsvektoren bezüglich des euklidischen Skalarprodukts konjugiert bzw. A-orthogonal, d.h. es gilt

$$\langle \boldsymbol{p}_i, \mathbf{A}\boldsymbol{p}_j\rangle = 0 \ , \quad j = 0, 1, ...i-1.$$

Für die Konvergenz des CG-Verfahrens ist die Positiv-Definitheit der Matrix $\mathbf{A}$ Voraussetzung. Dabei wird eine Matrix $\mathbf{A}$ als positiv-definit bezeichnet, wenn sie symmetrisch ist und $\boldsymbol{x}^t\mathbf{A}\boldsymbol{x} > 0 \quad \forall \boldsymbol{x} \neq 0$ gilt.
Bezüglich der erzeugten Iterierten zeichnet das CG-Verfahren zwei Eigenschaften aus:

- sie werden durch eine Minimierungseigenschaft im zugrundeliegenden Krylov-Raum charakterisiert,
- sie können mit wenig Aufwand und Speicherplatzbedarf pro Iteration berechnet werden.

Es wäre für ein CG-ähnliches Verfahren anwendbar auf indefinite, unsymmetrische Probleme wünschenswert, wenn es eben diese Eigenschaften aufweisen würde. Dies ist, wie in [16] gezeigt wird, nicht möglich. Meist wird nur eine der obigen Eigenschaften erfüllt.

4.2 BiCG-Stab-Verfahren

Das BiCG-Stab-Verfahren (Biconjugate Gradients Stabilized) ist ein auf indefinite, unsymmetrische Probleme anwendbares, sogenanntes CG-ähnliches Verfahren. Es wurde aus dem BiCG-Verfahren entwickelt, bei welchem die Minimierung des Residuums aufgegeben wurde. Bei der Entwicklung des BiCG-Stab-Verfahrens wurden Ideen des CGS-Algorithmus (Conjugate Gradients Squared) genutzt, so daß nicht die Transponierte der Systemmatrix benötigt wird, wie im BiCG-Verfahren. Allerdings besitzen beide Verfahren die gleichen breakdown-Schwachstellen (breakdown: Abbruch des Iterationsverfahrens ohne Berechnung einer Näherungslösung, z.B. Division durch Null). Da im BiCG-Stab-Algorithmus das Residuum lokal minimiert wird, glättet sich das unregelmäßige Konvergenzverhalten der BiCG- und CGS-Algorithmen. Die exakte Lösung wird nach spätestens n Schritten berechnet [19]. Zum Abbruch der Iteration ist es sinnvoll, eine

Gegeben seien $\boldsymbol{b}, \boldsymbol{x}_0 \in \mathbb{R}^n$, eine beliebige Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ und ein Präkonditionierer $\mathbf{W} \in \mathbb{R}^{n\times n}$, $\mathbf{W}$ regulär. Es wird $\boldsymbol{x} \in \mathbb{R}^n$ mit $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$ berechnet.

$\boldsymbol{r}_0 = \boldsymbol{b} - \mathbf{A}\boldsymbol{x}_0$, $\bar{\boldsymbol{r}}_0$ beliebig, so daß $\langle \bar{\boldsymbol{r}}_0, \boldsymbol{r}_0 \rangle \neq 0$
$\rho_{-1} = \alpha_{-1} = \omega_{-1} = 1$
$\boldsymbol{v}_{-1} = \boldsymbol{p}_{-1} = \mathbf{0}$

Iteration: $i = 0, 1, ...$ (bis $i \geq i_{max}$ oder $\|\boldsymbol{r}_i\| \leq tol$)

$\rho_i = \langle \bar{\boldsymbol{r}}_0, \boldsymbol{r}_i \rangle \neq 0$
$\beta_{i-1} = \frac{\rho_i \alpha_{i-1}}{\rho_{i-1}\omega_{i-1}}$ breakdown möglich
$\boldsymbol{p}_i = \boldsymbol{r}_i + \beta_{i-1}(\boldsymbol{p}_{i-1} - \omega_{i-1}\boldsymbol{v}_{i-1})$
$\boldsymbol{y}_i = \mathbf{W}^{-1}\boldsymbol{p}_i$
$\boldsymbol{v}_i = \mathbf{A}\boldsymbol{y}_i$
$\alpha_i = \frac{\rho_i}{\langle \bar{\boldsymbol{r}}_0, \boldsymbol{v}_i \rangle}$ breakdown möglich
$\boldsymbol{s}_i = \boldsymbol{r}_i - \alpha_i \boldsymbol{v}_i$
$\boldsymbol{z}_i = \mathbf{W}^{-1}\boldsymbol{s}_i$
$\boldsymbol{t}_i = \mathbf{A}\boldsymbol{z}_i$
$\omega_i = \frac{\langle \boldsymbol{t}_i, \boldsymbol{s}_i \rangle}{\langle \boldsymbol{t}_i, \boldsymbol{t}_i \rangle}$
$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \alpha_i \boldsymbol{y}_i + \omega_i \boldsymbol{z}_i$
$\boldsymbol{r}_{i+1} = \boldsymbol{s}_i - \omega_i \boldsymbol{t}_i$

Tabelle 4.1: BiCG-Stab-Algorithmus

Toleranz *tol* für das Residuum vorzugeben bzw. die maximale Anzahl der Iterationsschritte zu beschränken.

Die Präkonditionierungmatrix $\mathbf{W}$ bewirkt eine Transformation, die auf ein Gleichungssystem mit gleicher Lösung führt, dessen Eigenwertverteilung bezüglich der Konvergenzrate iterativer Methoden jedoch günstiger ist. Als Präkonditionierer kann man entweder eine Matrix verwenden, die $\mathbf{A}$ approximiert, aber leichter als $\mathbf{A}$ zu invertieren ist, oder eine Matrix $\mathbf{W}$, die $\mathbf{A}^{-1}$ approximiert, so daß nur eine Multiplikation mit $\mathbf{W}$ benötigt wird. Ersteres geschieht im allgemeinen mit den Iterationsmatrizen der (klassischen) linearen Iterationsverfahren, letzteres durch die Auswertung eines Polynoms in der Matrix $\mathbf{A}$.
Speicherplatzbedarf und Kosten eines Verfahrens, sowohl in der Initialisierung als auch pro Iterationsschritt, können durch die Anwendung eines Präkonditionierers erhöht werden. Aus diesem Grund sollte zwischen dem Aufwand zur Konstruktion und der Anwendung eines Präkonditionierers und der Verbesserung der Konvergenzgeschwindigkeit abgewogen werden.
Die Präkonditionierungmatrix $\mathbf{W}$ läßt sich wie folgt aufspalten $\mathbf{W} = \mathbf{W}_1\mathbf{W}_2$. Wenn

$\mathbf{W}_1 = \mathbf{I}$, erhält man eine Präkonditionierung von rechts, für $\mathbf{W}_2 = \mathbf{I}$ eine Präkonditionierung von links, und im allgemeinen Fall $\mathbf{W}_1 = \mathbf{L}, \mathbf{W}_2 = \mathbf{U}$ erhält man eine Präkonditionierung von beiden Seiten. In dem Schema spielen $\mathbf{W}_1$ und $\mathbf{W}_2$ keine explizite Rolle.

4.2.1 SSOR-Präkonditionierung

Da SSOR-Präkonditionierer direkt aus der Matrix $\mathbf{A}$ abgeleitet werden können, benötigen sie weder Speicherplatz noch erzeugen sie Initialisierungskosten. Die ursprüngliche Matrix $\mathbf{A}$ wird zerlegt in $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{U}$, wobei $\mathbf{D}$ der Diagonalanteil, $\mathbf{L}$ der untere Dreiecksanteil und $\mathbf{U}$ der obere Dreiecksanteil von $\mathbf{A}$ sind. Der SSOR-Präkonditionierer ist dann definiert als

$$\mathbf{W}(A)(\omega) = \frac{1}{\omega}(2-\omega)^{-1}(\mathbf{D} - \omega\mathbf{L})\,\mathbf{D}^{-1}(\mathbf{D} - \omega\mathbf{U})$$

mit dem Relaxationsparameter $\omega,\ 0 < \omega < 2$. Durch die optimale Wahl des Parameters ω kann die Ordnung des Verfahrens reduziert werden. Als gute Wahl hat sich $\omega = 1.3$ gezeigt. Eine eventuell vorhandene Symmetrie der Matrix $\mathbf{A}$ überträgt sich auf $\mathbf{W}$.
Die Vorkonditionierungsmatrix $\mathbf{W}$ wird in der praktischen Anwendung nicht explizit berechnet. Stattdessen wird ein Gleichungssystem $\mathbf{A}\boldsymbol{x} = \boldsymbol{g}_k$ approximativ mit einem SSOR-Schritt bzw. 2 Halbschritten und dem Startvektor $\boldsymbol{x}_0 = \mathbf{0}$ gelöst. Durch

$$\boldsymbol{x} \longmapsto (\mathbf{D} - \omega\mathbf{L})^{-1}(\omega\boldsymbol{b} - \omega\mathbf{U}\boldsymbol{x} - (\omega-1)\mathbf{D}\boldsymbol{x})$$

wird ein Iterationsschritt in Vorwärtsrichtung definiert. Ebenso wird die Relaxation in Rückwärtsrichtung durch

$$\boldsymbol{x} \longmapsto (\mathbf{D} - \omega\mathbf{U})^{-1}(\omega\boldsymbol{b} - \omega\mathbf{L}\boldsymbol{x} - (\omega-1)\mathbf{D}\boldsymbol{x})$$

definiert. Der erste iterative Halbschritt führt mit dem Startvektor $\boldsymbol{x}_0 = \mathbf{0}$ und $\boldsymbol{b} = \boldsymbol{g}_k$ auf

$$\boldsymbol{x}^{1/2} = \omega(\mathbf{D} - \omega\mathbf{L})^{-1}\boldsymbol{g}_k.$$

Der zweite iterative Halbschritt liefert unter Berücksichtigung der Relation

$$\omega\boldsymbol{g}_k + \omega\mathbf{L}\boldsymbol{x}^{1/2} - \mathbf{D}\boldsymbol{x}^{1/2} = \mathbf{0}$$

$$\boldsymbol{x}^{1,SSOR} = (2-\omega)(\mathbf{D} - \omega\mathbf{U})^{-1}\mathbf{D}\boldsymbol{x}^{1/2}.$$

4.2.2 ILU-Präkonditionierung

Mit dem Gaußschen Eliminationsverfahren erhält man eine sogenannte LU-Zerlegung (*lower-upper*): $\mathbf{A} = \mathbf{LU}$. Wenn $\mathbf{A}$ symmetrisch ist, vererbt sich die Symmetrieeigenschaft auf die Zerlegung $\mathbf{A} = \mathbf{LDL}^t$. Bei großen Systemen, wie sie häufig in der Praxis der FEM auftreten, ist die Matrix meist dünn besetzt, d.h. in jeder Zeile sind nur einige Elemente von Null verschieden. Diese Eigenschaft geht innerhalb des Gaußschen Eliminationsprozeß bereits nach wenigen Schritten verloren. Die Matrix $\mathbf{L}$ ist also wesentlich

dichter besetzt als $\mathbf{A}$.
Um die Besetzungsstruktur zu erhalten, führt man bei der unvollständigen LU-Zerlegung (*incomplete*=ILU) den Prozeß nur näherungsweise durch. Man unterdrückt im einfachsten Fall alle Matrixelemente für die Indexpaare, für welche $A_{ij} = 0$. Damit erhält man eine Zerlegung $\mathbf{A} = \tilde{\mathbf{L}}\tilde{\mathbf{U}} + \mathbf{R}$, wobei in der Fehlermatrix $\mathbf{R}$ nur $R_{ij} \neq 0$, sofern $A_{ij} = 0$ ist. Damit ergibt sich eine approximierte Vorkonditionierungsmatrix $\mathbf{W} = \tilde{\mathbf{L}}\tilde{\mathbf{U}}$. Aufgrund der Zerlegung mit schwach besetzter Matrix $\tilde{\mathbf{L}}$ sind Gleichungen der Form $\mathbf{W}\boldsymbol{y} = \boldsymbol{b}$ schnell auflösbar.
Die Qualität eines solchen Vorkonditionieres hängt naturgemäß davon ab, wie gut $\mathbf{W}$ die ursprüngliche Matrix $\mathbf{A}$ approximiert. Für Spezialfälle wurde eine Verbesserung der Konditionszahl bewiesen, $\kappa(\mathbf{W}^{-1}\mathbf{A}) \approx \sqrt{\kappa(\mathbf{A})}$. Über das beschriebene Vorgehen hinaus gibt es modifizierte ILU-Verfahren, in denen die Diagonale abgeschwächt oder gestärkt wird.

4.3 Gewähltes Verfahren

Eine allgemeine Regel dafür, in welchen Fällen die Vorkonditionierung mit SSOR oder ILU besser ist, scheint es nicht zu geben, aber in vielen Fällen wurde eine Überlegenheit der ILU-Vorkonditionierung beobachtet.
Bei ihrer Verwendung innerhalb des BiCG-Stab-Verfahrens muß vor Beginn der Iteration die Matrix $\mathbf{W}$ aufgestellt werden, was im Gegensatz zur SSOR-Vorkonditionierung Initialisierungskosten verursacht und Anforderungen an Speicherplatz stellt.
Die in dieser Arbeit vorgestellten Beispiele wurden mit dem BiCG-Stab-Algorithmus und einer ILU-Vorkonditionerung gelöst, sofern nichts anderes angegeben wird.

Kapitel 5

Fehlerschätzer

Adaptive Techniken haben sich als wertvolle Hilfsmittel bei der numerischen Lösung von partiellen Differentialgleichungen bewährt. In der Methode der Finiten Elemente werden diese Differentialgleichungen diskretisiert, d.h. die tatsächliche Lösung wird in einem endlich dimensionalen, auf einem Rechner gut zu handhabenden Raum approximiert. An dieser Stelle sollen Methoden zur Abschätzung des durch die Diskretisierung entstehenden Fehlers gezeigt werden.
Insbesondere in Fällen, bei denen die Genauigkeit der numerischen Approximation durch lokale Singularitäten, wie einspringende Ecken, Grenzschichten oder scharfe Schockfronten, gestört wird, d.h. wo die Lösung weniger regulär ist, hat sich die Verfeinerung der Diskretisierung dieser kritischen Regionen als sinnvoll erwiesen. Um eine optimale Genauigkeit zu erhalten, gilt es, diese Regionen zu identifizieren und eine gute Balance zwischen den verfeinerten und unverfeinerten Teilgebieten zu finden. Darüberhinaus ergibt sich das Problem, eine zuverlässige Abschätzung der berechneten numerischen Lösung angeben zu können.
Die Fehler, die nicht aus der Diskretisierung entstehen, sollen hier nicht betrachtet werden. Der Modellfehler kann bei der Gewinnung des zugrundeliegenden Randwertproblems entstehen, da es durch vereinfachende Annahmen die physikalische Wirklichkeit nicht vollständig wiedergeben kann. Ferner können Fehler bei der Lösung des Gleichungssystems auftreten, z.B. als Abbruchfehler im BiCG-Stab-Verfahren.

Es sei eine partielle Differentialgleichung in dem Gebiet Ω mit der Lösung u gegeben. Sei ferner $u_{h,p}$ die Finite Element Lösung der diskretisierten partiellen Differentialgleichung korrespondierend zu einer Netzfunktion $h(x)$ und stückweise polynominalen Approximationen des Grades p.
Es soll eine Diskretisierung und eine diskrete Lösung $u_{h,p}$ gefunden werden, für die mit einer gegebenen Toleranz TOL und einer gegebenen Norm $\|\cdot\|$ die Fehlergrenze

$$\|u - u_{h,p}\| \leq TOL$$

erfüllt wird, wobei möglichst wenig rechnerischer Aufwand nötig sein soll.

Das Finden eines Netzes mit minimaler Anzahl von Unbekannten ist ein komplexes Optimierungsproblem. Es wird in der Praxis nährungsweise gelöst. Hierfür werden adaptive Methoden herangezogen. Es gibt verschiedene adaptive Methoden, z.B.

- adaptive Wahl des Netzes (h-Methode)
- adaptive Wahl des Grades der lokalen Approximation (p-Methode)
- Verschiebung von Knoten (r-Methode)
- Netz-Alignment, Netz-Orientierung, Netz-Stretching
- Kombinationen

In dieser Arbeit werden zunächst Aspekte der h-Methode betrachtet. In Kapitel 7 wird zusätzlich ein Netzglättungsalgorithmus, welcher zur Verschiebung von Knoten unter Beibehaltung der Netztopologie dient, vorgestellt.
Zur Verdichtung der Diskretisierung werden Informationen über die Verteilung des Fehlers benötigt. Diese werden mittels sogenannter Fehlerschätzer gewonnen. *A priori* Fehlerschätzer benutzen dabei nur Werte der exakten, aber unbekannten Lösung. Sie erweisen sich oft als ungenügend, da sie Informationen über das asymptotische Verhalten des Fehlers geben und Anforderungen an die Regularität der Lösung stellen, die bei Auftreten der oben genannten Singularitäten nicht unbedingt erfüllt werden können.
A posteriori Fehlerschätzer geben den Fehler in Termen der berechneten Lösung und der Daten des Ausgangsproblems an. An diese Schätzer werden die Anforderungen gestellt, daß sie lokal sein sollen und auf zuverlässige obere und untere Schranken für den tatsächlichen Fehler führen sollen. Globale obere Schranken garantieren, daß der Fehler unterhalb einer vorgegebenen Toleranz bleibt und somit die Zuverlässigkeit gesichert wird. Lokale untere Schranken sichern eine korrekte Gitterverfeinerung, so daß die numerische Lösung bei nahezu minimaler Anzahl von Gitterpunkten die gegebene Toleranz unterschreitet. Damit ist die Effizienz eines adaptiven Algorithmus gewährleistet.

Im Rahmen adaptiver Methoden sollen Fehlerschätzer die folgenden Aufgaben erfüllen:

1. Indikation:
   Die Regionen, in denen große Fehler auftreten, sollen gefunden werden. Sei η_{lok} ein lokaler Fehlerschätzer, dann sichert die Ungleichung

   $$\eta_{lok} \leq c\|u - u_{h,p}\|_{lok} \quad , \; c > 0,$$

   daß ein lokaler Fehler existiert, wenn η_{lok} einen lokalen Fehler indiziert.

2. Abbruchkriterium:
   Als Abbruchkriterium für den Rechenprozeß kann ein Fehlerschätzer benutzt werden, wenn die Abschätzung

   $$\|u - u_{h,p}\| \leq C\eta_{glob} \quad , \; C > 0$$

   erfüllt wird, wobei η_{glob} die Abschätzung des globalen Fehlers ist und C näherungsweise bekannt und nicht zu groß ist.

Ein Maß für die Effektivität eines Fehlerschätzers liefert der sogenannte Effektivitätsindex, der das Verhältnis von geschätzem Fehler zu tatsächlichem Fehler angibt:

$$\frac{\eta_{glob}}{e_{glob}}, \quad \frac{\eta_{lok}}{e_{lok}}.$$

Strebt dieser Index im Verlauf des adaptiven Prozesses gegen 1, so spricht man von einem *asymptotisch exakten* Fehlerschätzer.

Neben den Fehlerschätzern gibt es die sogenannten Fehlerindikatoren. Sowohl die Fehlerschätzer als auch die Fehlerindikatoren können die Regionen, in denen große Fehler auftreten, auffinden, aber nur Fehlerschätzer treffen eine Aussage über die Größe des Fehlers und können somit als Abbruchkriterium dienen. Im allgemeinen sind Indikatoren leicht und ohne großen Aufwand zu implementieren.

Es läßt sich eine Klassifizierung der *a posteriori* Fehlerschätzer bzw. Indikatoren folgender Art treffen:

1. Residuelle Fehlerschätzer:
   Der Fehler der numerischen Lösung wird durch ihr Residuum in einer passenden Norm mit Bezug auf die starke Form der Differentialgleichung abgeschätzt, d.h. es wird gleichsam elementweise kontrolliert, ob die Differentialgleichungen erfüllt werden.

2. Fehlerschätzer basierend auf lokalen Hilfsproblemen:
   Es werden lokale diskrete Probleme ähnlicher Form gelöst, die aber einfacher als das Ausgangsproblem sind. Die lokalen Lösungen dienen in einer geeigneten Norm der Fehlerabschätzung.

3. Hierarchische Ansätze:
   Zunächst wird das Residuum einer FE-Lösung mit anderem FE-Raum (höhere Ansätze, feineres Netz) berechnet und mit der Lösung des Ausgangsproblems verglichen.

4. Indikatoren aus Mittelungstechniken:
   In diesem Fall werden lokale Extrapolation oder Mittelungstechniken verwendet. Es sei bemerkt, daß diese sich im wesentlichen für lineare Probleme eignen.

An dieser Stelle sei auf das Buch [59] hingewiesen.

5.1 Notation

Die in diesem Kapitel verwendete Notation soll kurz vorgestellt werden. Die Lösung des analytischen Problems sei mit u und die des diskreten Problems mit u_h bezeichnet, der aus den bekannten Daten abzuschätzende Fehler sei $e = u - u_h$. $\Omega \subset \mathbb{R}^n, n = 2$ sei ein beschränktes zusammenhängendes Gebiet, dessen Rand Γ Lipschitz-stetig sei, d.h. er sei stückweise glatt und die Eckpunkte der Seiten besitzen Tangenten, deren eingeschlossene Winkel größer Null sind. Im speziellen sei Γ hier als polygonal angenommen und auf ihm die oben eingeführten Dirichlet- und Neumann-Randbedingungen definiert.

Setze den Raum X zu

$$X := H^1_{D_0}(\Omega) = \left\{ u \in H^1(\Omega) : u = 0 \ \text{ auf } \Gamma_D \right\}.$$

Es werden die üblichen L_2- bzw. H^1-Normen

$$\begin{aligned} \|u\|_0 &:= \|u\|_{0,2,T} := \left\{ \int_T |u|^2 \, \mathrm{d}x \right\}^{1/2} \\ \|u\|_1 &:= \|u\|_{1,2,T} := \left\{ \int_T |u|^2 + |\nabla u|^2 \, \mathrm{d}x \right\}^{1/2} \end{aligned}$$

sowie eine wie folgt definierte Energienorm

$$|\!|\!| u |\!|\!| := \left\{ \epsilon \|\nabla u\|_0^2 + \|u\|_0^2 \right\}^{1/2}$$

benutzt werden.

Bezüglich der Diskretisierung sei $\mathcal{T}_h$ eine Familie von Zerlegung von Ω in Dreieckselemente. Viereckselemente werden im Rahmen dieser Arbeit nicht betrachtet. Sei T ein Element, h_T sein Durchmesser, E eine Kante des Elements und h_E die Länge dieser Kante. Außerdem sei $\boldsymbol{n}_E$ ein Normaleneinheitsvektor auf der Kante E. Wenn die Kante E auf dem Gebietsrand $\Gamma = \partial\Omega$ liegt, so sei $\boldsymbol{n}_E$ der äußere Normalenvektor an Ω. Der Sprung einer Funktion u über die Kante E in Richtung der Normalen $\boldsymbol{n}_E$ sei mit $[u]_E$ bezeichnet, d.h.

$$[u]_E(\boldsymbol{x}) := \lim_{t \to +0} u(\boldsymbol{x} + t\boldsymbol{n}_E) - \lim_{t \to +0} u(\boldsymbol{x} - t\boldsymbol{n}_E) \qquad \forall \boldsymbol{x} \in E.$$

Die Triangulierungen sollen die folgenden Voraussetzungen erfüllen:

1. Zulässigkeit
   $\boldsymbol{x} \in T, T \in \mathcal{T}_h$ sei ein Eckpunkt von T und $\boldsymbol{x} \in T'$
   $\Rightarrow$ $\boldsymbol{x}$ ist auch Eckpunkt von T'
   Damit sollen hängende Knoten ausgeschlossen werden und die Stetigkeit der Funktionen über die Ränder gesichert sein.

2. Regularität
   Der kleinste Winkel der Triangulierung soll von Null weg beschränkt sein, d.h. im Zweidimensionalen soll das Verhältnis des Durchmessers des Umkreises $h_T := \operatorname{diam} T = \sup_{\boldsymbol{x},\boldsymbol{y} \in T} |\boldsymbol{x} - \boldsymbol{y}|$ zu dem des Innenkreis $\rho_T := \max\{h_B : B \subset T : \text{Kreis}\}$ gleichmäßig in $T \in \mathcal{T}_h$ und $h > 0$ beschränkt sein

   $$c_{\mathcal{T}} := \sup_{h>0} \sup_{T \in \mathcal{T}_h} \frac{h_T}{\rho_T} < \infty.$$

   Diese Bedingung erlaubt die Verwendung von lokal verfeinerten Netzen.

Für die Knoten, Kanten und deren Umgebung werden folgende Schreibweisen gewählt:

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

$$\begin{array}{lll} \mathcal{N}(T) = \{\text{Eckpunkte von } T\} & , & \mathcal{E}(T) = \{\text{Kanten von } T\} \\ \mathcal{N}_h := \cup_{T \in \mathcal{T}_h} \mathcal{N}(T) & , & \mathcal{E}_h := \cup_{T \in \mathcal{T}_h} \mathcal{E}(T) \\ \mathcal{N}_h = \mathcal{N}_{h,\Omega} \cup \mathcal{N}_{h,D} \cup \mathcal{N}_{h,N} & , & \mathcal{E}_h = \mathcal{E}_{h,\Omega} \cup \mathcal{E}_{h,D} \cup \mathcal{E}_{h,N} \end{array}$$

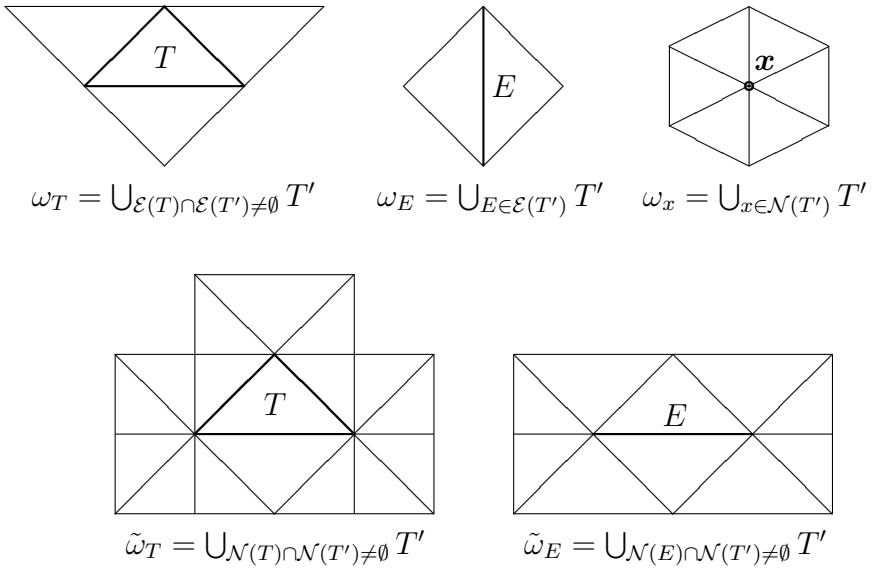


Abbildung 5.1: Umgebungen $\omega_T, \omega_E, \omega_x, \tilde{\omega}_T, \tilde{\omega}_E$

Abbildung 5.1 zeigt Definitionen verschiedener Umgebungen eines Knotens x, einer Kante E bzw. eines Dreiecks T.

Wegen der Regularität gilt $h_T/h_{T'} \leq c(c_{\mathcal{T}}),\ \forall T, T',\ T \cap T' \neq \emptyset$. Insbesondere können sich die Winkel innerhalb eines Gebietes ω_x nicht zu drastisch ändern, und die Anzahl der Elemente in ω_x bleibt beschränkt.

Desweiteren sollen die lokalen Abschneidefunktionen oder *bubble functions* eingeführt werden. Es sei $\hat{T}$ das Referenzelement mit den Eckpunkten $(0,0)$, $(1,0)$ und $(0,1)$ und $\hat{E} = \hat{T} \cap \{x = 0\}$ eine Referenzkante. Mit

$$b_{\hat{T}} \in P_3\ ;\ 0 \leq b_{\hat{T}} \leq 1\ ;\ \max_{\boldsymbol{x}\in\hat{T}} b_{\hat{T}}(\boldsymbol{x}) = 1\ ;\ b_{\hat{T}} = 0 \text{ auf } \partial\hat{T}$$

sei die Abschneidefunktion $b_{\hat{T}}$ auf $\hat{T}$, die gerade auf dem Rand des Referenzdreiecks verschwindet, definiert. Für ein Dreickselement ergibt sich die Funktion aus dem Produkt der Schwerpunktskoordinaten

$$b_{\hat{T}}(x,y) = 27xy(1-x-y)$$

bzw. in Flächenkoordinaten zu

$$b_{\hat{T}}(\lambda) = 27\lambda_1\lambda_2\lambda_3\ .$$

Definitionen zu den Flächenkoordinaten sind im Anhang 1 zu finden. Die Abschneidefunktion $b_{\hat{E}}$ auf der Referenzkante $\hat{E}$ verschwindet auf $\partial\hat{T}$, aber nicht auf der Referenzkante $\hat{E}$:

$$b_{\hat{E}} \in P_2\ ;\ 0 \leq b_{\hat{E}} \leq 1\ ;\ \max_{\boldsymbol{x}\in\hat{E}} b_{\hat{E}}(\boldsymbol{x}) = 1\ ;\ b_{\hat{E}} = 0 \text{ auf } \partial\hat{T}\backslash\hat{E}\ .$$

Im Falle eines Dreieckselements hat sie die Form:

$$b_{\hat{E}}(x,y) = 4x(1-x-y)$$

bzw. in Flächenkoordinaten für die Kante $\hat{E}_i$, die zwischen den Knoten $i+1$ und $i+2$ liegt,

$$b_{\hat{E}_i}(\lambda) = 4\lambda_{i+1}\lambda_{i+2} \ .$$

Die bubble-Funktionen haben für beliebige $T \in \mathcal{T}_h$ bzw. $E \in \mathcal{E}_h$ die folgenden Eigenschaften:

$$\begin{aligned}
&c_1 h_T^2 \le \int_T b_T = \frac{9}{20}|T| \le c_2 h_T^2 \\
&\int_E b_E = \frac{2}{3} h_E \\
&c_3 h_E^2 \le \int_{T'} b_E = \frac{1}{3}|T'| \le c_4 h_E^2, \qquad \forall T' \subset \omega_E \\
&\|\nabla b_T\|_{0,2;T} \le c_5 h_T^{-1} \|b_T\|_{0,2;T} \\
&\|\nabla b_E\|_{0,2;T'} \le c_6 h_E^{-1} \|b_E\|_{0,2;T'} \qquad \forall T' \subset \omega_E,
\end{aligned} \tag{5.1}$$

dabei ist $|T|$ der Flächeninhalt von T und die Konstanten $c_1, ..., c_6$ hängen nur vom kleinsten Winkel der Triangulierung ab.

P_k seien die Polynome mit einem Grad von höchstens k, $k \in N$. Setze

$$\begin{aligned}
S_h^{k,-1} &:= \{u : \Omega \to R : u|_T \in P_k \ \forall T \in \mathcal{T}_h\} \\
S_h^{k,0} &:= \left\{S_h^{k,-1} \cap C(\overline{\Omega})\right\}, \ k \ge 1 \\
S_{h,D}^{k,0} &:= \left\{u \in S_h^{k,0} : u = 0 \text{ auf } \Gamma_D\right\} \ k \ge 1 \ .
\end{aligned}$$

Durch $I_h : X \to X_h$ sei der Interpolationsoperator nach Clément bezeichnet. Es seien $u \in X$ und $\boldsymbol{x} \in \mathcal{N}_h$ gegeben, dann stellt $\pi_x u$ die $L^2(\omega_x)$-Projektion von u auf P_1 dar, d.h.

$$\int_{\omega_x} uv = \int_{\omega_x} \pi_x uv \qquad \forall v \in P_1.$$

$I_h u$ ist dann eindeutig wie folgt definiert :

$$\begin{aligned}
I_h u(\boldsymbol{x}) &= (\pi_x u)(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in \mathcal{N}_{h,\Omega} \cup \mathcal{N}_{h,N}, \\
I_h u(\boldsymbol{x}) &= 0 \qquad\qquad \forall \boldsymbol{x} \in \mathcal{N}_{h,D}.
\end{aligned}$$

Für den Operator I_h gelten für $T \in \mathcal{T}_h$, $E \in \mathcal{E}_h$ und $u \in X$ die folgenden lokalen Abschätzungen:

$$\begin{aligned}
\|u - I_h u\|_{0,2;T} &\le c_{I_1} h_T \|u\|_{1,2;\tilde{\omega}_T} \\
\|u - I_h u\|_{2;E} &\le c_{I_2} h_E^{1/2} \|u\|_{1,2;\tilde{\omega}_E}.
\end{aligned} \tag{5.2}$$

Hieraus ergeben sich insbesondere die Abschätzungen [60]:

$$\begin{aligned}
\|u - I_h u\|_{0,2;T} &\le c_{I_3} \min\{h_T \epsilon^{-1/2}, 1\} |\!|\!| u |\!|\!|_{\tilde{\omega}_T} \\
\|u - I_h u\|_{2;E} &\le c_{I_4} \epsilon^{-1/4} \min\{h_E \epsilon^{-1/2}, 1\}^{-1/2} |\!|\!| u |\!|\!|_{\tilde{\omega}_E} \\
|\!|\!| I_h u |\!|\!|_T &\le c_{I_5} |\!|\!| u |\!|\!|_{\tilde{\omega}_T}.
\end{aligned} \tag{5.3}$$

Dabei sind die Konstanten c_{I_i} unabhängig von der Netzweite und dem Parameter ϵ.

5.2 Indikator aus Mittelungstechnik

Der hier vorgestellte Fehlerschätzer ist in der Ingenieur-Literatur als der Zienkiewicz-Zhu-Schätzer oder Z^2-Fehlerschätzer bekannt .
Sei u die exakte Lösung des Problems und u_h die aktuelle numerische Näherung. Es ist eine Abschätzung für $\|\nabla u - \nabla u_h\|_{0,2;T}$ gesucht. Es wird angenommen, daß eine Näherung $G(u_h)$ für ∇u durch Glättung oder Mittelung aus u_h konstruiert werden kann, die genauer ist:

$$\|\nabla u - G(u_h)\|_{0,2;T} \leq \gamma \|\nabla u - \nabla u_h\|_{0,2;T}, \qquad 0 \leq \gamma < 1.$$

Es gilt

$$\|G(u_h) - \nabla u_h\|_{0,2;T} \geq | \, \|\nabla u - G(u_h)\|_{0,2;T} - \|\nabla u - \nabla u_h\|_{0,2;T} \, |$$

für jedes Element T. Wegen $0 \leq \gamma < 1$ ergibt sich hieraus die Schranke

$$\frac{1}{1+\gamma} \, \|G(u_h) - \nabla u_h\|_{0,2;T} \leq \|\nabla u - \nabla u_h\|_{0,2;T} \leq \frac{1}{1-\gamma} \, \|G(u_h) - \nabla u_h\|_{0,2;T}$$

Damit erhält man eine Näherung für den Fehler und wählt $\|G(u_h) - \nabla u_h\|_{0,2;T}$ als Fehlerschätzer

$$\eta_{Z,T} = \|G(u_h) - \nabla u_h\|_{0,2;T} \qquad \forall T.$$

Da ∇u_h ein stückweise konstantes Vektorfeld ist, kann man erwarten, daß die L^2-Projektion von ∇u_h auf ein kontinuierliches stückweise lineares Vektorfeld die obige erste Ungleichung erfüllt. Diese Projektion ist allerdings in ihrer Berechnung fast so teuer wie das Lösen des grundlegenden FE-Problems. Aus diesem Grunde ersetzt man das L^2-Skalarprodukt durch ein einfacheres.
Zur Erläuterung des Vorgehens soll ein wenig ausgeholt werden: Sei W_h der Raum der stückweise linearen Vektorfelder und $V_h := W_h \cap C(\Omega, \mathbb{R}^2)$. Mit $(.,.)_h$ sei ein netzabhängiges Skalarprodukt auf W_h definiert, wobei $|T|$ der Flächeninhalt des Elements T sei:

$$(\boldsymbol{\varphi}, \boldsymbol{\psi})_h = \sum_{T \in \mathcal{T}_h} \frac{|T|}{3} \left\{ \sum_{\boldsymbol{x} \in \mathcal{N}_h} \boldsymbol{\varphi} \mid_T (\boldsymbol{x}) \cdot \boldsymbol{\psi} \mid_T (\boldsymbol{x}) \right\}, \quad \forall \boldsymbol{\varphi}, \boldsymbol{\psi} \in W_h, \; T \in \mathcal{T}_h, \; \boldsymbol{x} \in \mathcal{N}_h,$$

mit

$$\boldsymbol{\varphi} \mid_T (\boldsymbol{x}) := \lim_{\boldsymbol{y} \to \boldsymbol{x}, \; \boldsymbol{y} \in T} \boldsymbol{\varphi}(\boldsymbol{x}).$$

Die Quadraturformel

$$\int_T \phi \approx \frac{|T|}{3} \sum_{\boldsymbol{x} \in \mathcal{N}_h} \phi(\boldsymbol{x})$$

ist für lineare Funktionen exakt, weshalb für $\boldsymbol{\varphi}, \boldsymbol{\psi} \in W_h$ und mindestens eine von ihnen konstant gilt:

$$(\boldsymbol{\varphi}, \boldsymbol{\psi})_h = \int_\Omega \boldsymbol{\varphi} \cdot \boldsymbol{\psi}.$$

Außerdem läßt sich nachrechnen, daß

$$\begin{aligned}\frac{1}{4}\|\boldsymbol{\varphi}\|_{0,2}^2 &\leq (\boldsymbol{\varphi},\boldsymbol{\varphi})_h \leq \|\boldsymbol{\varphi}\|_{0,2}^2 , \qquad \forall \boldsymbol{\varphi} \in W_h \\ (\boldsymbol{\varphi},\boldsymbol{\psi})_h &= \frac{1}{3}\sum_{\boldsymbol{x}\in\mathcal{N}_h} |\omega_x|\ \boldsymbol{\varphi}(\boldsymbol{x})\cdot\boldsymbol{\psi}(\boldsymbol{x}) , \quad \forall \boldsymbol{\varphi},\boldsymbol{\psi} \in W_h\end{aligned}$$

ist. Es sei nun $G(u_h) \in V_h$ die $(.,.)_h$-Projektion von ∇u_h auf V_h

$$(G(u_h),\boldsymbol{\varphi}_h)_h = (\nabla u_h,\boldsymbol{\varphi}_h)_h , \qquad \forall \boldsymbol{\varphi}_h \in V_h.$$

Die obigen Gleichungen implizieren, daß gilt

$$G(u_h)(\boldsymbol{x}) = \sum_{T\subset\omega_x} \frac{|T|}{|\omega_x|} \nabla u_h\mid_T (\boldsymbol{x}) , \qquad \boldsymbol{x} \in \mathcal{N}_h,$$

mit

$$\nabla u_h\mid_T (\boldsymbol{x}) := \lim_{\boldsymbol{y}\to\boldsymbol{x},\ \boldsymbol{y}\in T} \nabla u_h(\boldsymbol{x}).$$

$G(u_h)$ berechnet sich also aus der lokalen Mittelung von ∇u_h, wobei ω_x die Vereinigung aller Elemente ist, die $\boldsymbol{x}$ als Knoten haben. Der Gradient des Elements $T \subset \omega_x$ wird mit $|T|/|\omega_x|$ gewichtet, dadurch finden die möglicherweise verschiedenen Größen von Elementen Berücksichtigung [57]. Diese Methode stellt höhere Anforderungen an die Speicherung als eine Berechnung des arithmetischen Mittels der Gradienten aller Dreiecke, die $\boldsymbol{x}$ als Knoten haben, wie u.a. in [33] vorgeschlagen.
Der lokale Fehlerschätzer

$$\eta_{Z,T} := \|G(u_h) - \nabla u_h\|_{0,2;T}$$

bzw. das globale Pendant

$$\eta_Z := \left\{\sum_{T\in\mathcal{T}_h} \eta_{Z,T}^2\right\}^{1/2}$$

sind für beide Arten der Wichtung problemunabhängig, d.h. sie nehmen für jedes Modellproblem dieselbe Form an. In [59] wird für $\epsilon = 1$, $\boldsymbol{b} = \boldsymbol{0}$ gezeigt, daß es sich bei $\eta_{Z,T}$ um einen Fehlerschätzer handelt, der Beweis sei dort nachzulesen.

5.3 Ein "klassischer" residueller Fehlerschätzer

Im Gegensatz zu dem vorherbeschriebenen Fehlerindikator wird in diesem Kapitel ein problemabhängiger Fehlerschätzer vorgestellt. Mittels eines residuellen Fehlerschätzers wird der Fehler durch das Residuum der numerischen Lösung in einer passenden Norm abgeschätzt. Für das betrachtete Modellproblem ergibt sich dieser residuelle Fehlerschätzer schließlich in der Form:

$$\eta_R = \left(\sum_{T\in\mathcal{T}_h} \alpha_T^2 \|Res_T(u_h)\|_{0,2;T}^2 + \sum_{E\in\mathcal{E}_h} \beta_E \|Res_E(u_h)\|_{2;E}^2\right)^{1/2}.$$

Im Verlauf des folgenden Abschnittes soll auf die Bedeutung und den Ursprung der einzelnen Terme eingegangen werden.
Zur Veranschaulichung der Herleitung eines residuellen Fehlerschätzers soll zunächst der Einfachheit halber ein entsprechender Fehlerschätzer für die Poisson-Gleichung vorgeführt werden. Im Anschluß daran wird in analoger Weise der Fehlerschätzer für das gesamte Modellproblem dargestellt.

5.3.1 Ein residueller Fehlerschätzer für die Poisson-Gleichung

Die elliptische Poisson-Gleichung (als Vereinfachung des Modellproblems $\epsilon = 1, \boldsymbol{b} = \mathbf{0}, \alpha = 0$) mit zugehörigen Dirichlet und Neumann Randbedingungen lautet:

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega \\ u &= u_D \qquad \text{auf } \Gamma_D \\ \frac{\partial u}{\partial \boldsymbol{n}} &= t_N \qquad \text{auf } \Gamma_N. \end{aligned}$$

Das Gebiet Ω sei ein polygonal begrenztes Gebiet im $\mathbb{R}^2$ mit dem Rand $\Gamma = \Gamma_D \cup \Gamma_N$, $\Gamma_D \cap \Gamma_N = \emptyset$, wobei Γ_D eine positive Länge habe. Die Funktionen f und t_N seien quadratisch integrierbar in Ω bzw. auf Γ_N. Mit dem Standard Galerkin Verfahren erhält man folgende Finite Element Diskretisierung : Finde $u_h \in X_h$ so, daß

$$a_P(u_h, v_h) := \int_\Omega \nabla u_h \cdot \nabla v_h = \int_\Omega f\ v_h + \int_{\Gamma_N} t_N\ v_h =: l_P(v_h) \qquad \forall v_h \in X_h.$$

Im folgenden soll zur Abkürzung der Index P für das Poisson-Problem weggelassen werden.

Zunächst soll eine variationelle Formulierung für den Fehler $e = u - u_h$ aus den Formulierungen des kontinuierlichen und diskreten Problems gefunden werden.

$$\begin{aligned} a(u, v) &= l(v) \qquad \forall v \in X \\ a(u_h, v_h) &= l(v_h) \qquad \forall v_h \in X^h \end{aligned}$$

Durch Einsetzen der Beziehung $e = u - u_h$ in das kontinuierliche Problem erhält man direkt eine implizite Darstellung des Residuums:

$$\begin{aligned} a(e, v) &= a(u - u_h, v) = a(u, v) - a(u_h, v) \\ &= l(v) - a(u_h, v) \qquad \forall v \in X \end{aligned}$$

bzw.

$$\int_\Omega \nabla(u - u_h) \cdot \nabla v = \int_\Omega f\ v + \int_{\Gamma_N} t_N\ v - \int_\Omega \nabla u_h \cdot \nabla v \qquad \forall v \in X. \tag{5.4}$$

Die Bilinearform a sei stetig, d.h.

$$|a(e, v)| \leq C_a \|e\|_X \|v\|_X \qquad \forall e, v \in X.$$

In die Konstante C_a gehen Parameter des Problems ein, insbesondere der größte Eigenwert der Diffusivitätsmatrix. Außerdem hängt die Wahl der Norm vom jeweiligen Problem ab.

Damit läßt sich eine Abschätzung des Fehlers nach unten angeben:

$$\sup_{v\in X\setminus\{0\}} \frac{l(v)-a(u_h,v)}{\|v\|_X} \leq C_a\|e\|_X.$$

Eine Abschätzung nach oben erhält man aus der Forderung, daß die Bilinearform die *inf-sup*-Bedingung erfülle:

$$c_a\|e\|_X \leq \sup_{v\in X\setminus\{0\}} \frac{a(e,v)}{\|v\|_X} = \sup_{v\in X\setminus\{0\}} \frac{l(v)-a(u_h,v)}{\|v\|_X} \quad \forall v\in X.$$

Der Faktor c_a hängt, ebenso wie in C_a, von den Materialeigenschaften etc. ab, hier allerdings vom kleinsten Eigenwert der Diffusivitätsmatrix. C_a und c_a messen die Empfindlichkeit des analytischen Problems gegenüber Störungen und damit seine Stabilität. Je weiter sie auseinanderliegen, desto sensibler ist das System. Dem läßt sich aber nicht durch die Diskretisierung entgegensteuern. Bei *a posteriori* Abschätzungen geht die Stabilität des diskreten Problems nicht ein.
Da die obigen *sup*-Ausdrücke in dieser Form nicht berechenbar sind, ist es das Ziel, berechenbare Ausdrücke zu finden.

Gilt $X^h \subset X$, so ist der Fehler orthogonal zu X^h :

$$a(e,v_h) = a(u-u_h,v_h) = 0 \qquad \forall v_h \in X^h.$$

Die Konsistenz wird von der Standard-Galerkin Methode erfüllt.

Setzt man nun die entsprechenden Ausdrücke in die Beziehung für $a(e,v)$ ein, erhält man

$$l(v)-a(u_h,v) = \int_\Omega fv + \int_{\Gamma_N} t_N v - \int_\Omega \nabla u_h\cdot\nabla v.$$

Die Integrale über das Gebiet Ω sind als Summe der Integrale über die einzelnen Elemente zu verstehen. Wendet man den Gaußschen Integralsatz auf den Term

$$\int_T \nabla u_h\cdot\nabla v = \int_{\partial T} \boldsymbol{n}_T\cdot\nabla u_h\; v - \int_T \Delta u_h\; v$$

an und setzt dies wiederum ein, so nimmt die Gleichung unter Berücksichtigung von $\Delta u_h = 0$ auf allen Elementen $T\in\mathcal{T}_h$ die folgende Gestalt an:

$$\begin{aligned} a(e,v) &= \sum_{T\in\mathcal{T}_h}\int_T fv + \int_{\Gamma_N} t_N v + \sum_{T\in\mathcal{T}_h}\left\{\int_T \Delta u_h v - \int_{\partial T}\boldsymbol{n}_T\cdot\nabla u_h v\right\} \\ &= \sum_{T\in\mathcal{T}_h}\int_T f\; v + \sum_{E\in\mathcal{E}_{h,\Omega}}\int_E -[\boldsymbol{n}_E\nabla u_h]_E\, v + \sum_{E\in\mathcal{E}_{h,N}}\int_E (t_N-\boldsymbol{n}_E\nabla u_h)\, v. \end{aligned} \tag{5.5}$$

Um etwas mehr Übersichtlichkeit zu gewinnen, werden abkürzende Schreibweisen für die residuellen Anteile aus den Elementgebieten und die aus den Anteilen über die Elementkanten eingeführt:

$$a(e,v) = \sum_{T\in\mathcal{T}_h}\int_T Res_T(u_h)v + \sum_{E\in\mathcal{E}_h}\int_E Res_E(u_h)v,$$

dabei sind

$$Res_T(u_h) := f + \Delta u_h$$

$$Res_E(u_h) := \begin{cases} -[\boldsymbol{n}_E \cdot \nabla u_h]_E \ , & \text{wenn E } \in \mathcal{E}_{h,\Omega} \\ (t_N - \boldsymbol{n}_E \nabla u_h) \ , & \text{wenn E } \in \mathcal{E}_{h,N} \\ 0 \ , & \text{wenn E } \in \mathcal{E}_{h,D}. \end{cases}$$

Der residuelle Anteil auf dem Elementinneren entspricht dabei der starken Form der Differentialgleichung.
Diese variationelle Form für den globalen Fehler könnte man unter Anwendung des üblichen FE-Vorgehens mit Elemente eines höheren Ansatzes als den zuvor angenommenen lösen. Diese Verfahrensweise ist allerdings teuer, da ein Gleichungssystem gelöst werden müßte, das mindestens so groß ist wie jenes, welches zur Gewinnung von u_h diente. Der Vergleich einer z.B. durch lineare Elemente gewonnen Lösung u_h mit einer Lösung aus höheren Elementansätzen entspricht dem Vorgehen bei hierarchischen Fehlerschätzern, die aber aufgrund ihres hohen Rechenaufwandes im Rahmen dieser Arbeit nicht berücksichtigt wurden. Ein Übergang auf sogenannte Patches oder einzelne Elemente führt auf Fehlerschätzer, die auf der Lösung lokaler Hilfsprobleme basieren. Diese Fehlerschätzer sollen im nächsten Kapitel näher betrachtet werden.

Um zu einem expliziten residuellen Fehlerschätzer zu gelangen, wird die Galerkin Orthogonalität ausgenutzt und der Interpolationsoperator I_h nach Clément eingeführt:

$$\begin{aligned} a(e,v) &= l(v) - a(u_h, v) = l(v - I_h v) - a(u_h, v - I_h v) \\ &= \sum_{T \in \mathcal{T}_h} \int_T Res_T(u_h)\{v - I_h v\} + \sum_{E \in \mathcal{E}_h} \int_E Res_E(u_h)\{v - I_h v\}. \end{aligned}$$

Mit der Cauchy-Schwarzschen Ungleichung

$$\int_S \phi\psi \le \left\{\int_S \phi^2\right\}^{1/2} \left\{\int_S \psi^2\right\}^{1/2} = \|\phi\|_{0,2;S} \|\psi\|_{0,2;S}$$

und den oben angegebenen lokalen Abschätzungen (5.2) für I_h gelangt man auf die folgende Abschätzung

$$\begin{aligned} a(e,v) &\le \sum_{T \in \mathcal{T}_h} \|Res_T(u_h)\|_{0,2;T} \ \|\{v - I_h v\}\|_{0,2;T} + \sum_{E \in \mathcal{E}_h} \|Res_E(u_h)\|_{2;E} \ \|\{v - I_h v\}\|_{0,2;T} \\ &\le \sum_{T \in \mathcal{T}_h} \|Res_T(u_h)\|_{0,2;T} \ c_{I_1} h_T \|v\|_{1,2;\tilde{\omega}T} + \sum_{E \in \mathcal{E}_h} \|Res_E(u_h)\|_{2;E} \ c_{I_2} h_E^{1/2} \|v\|_{1,2;\tilde{\omega}E} \\ &\le \max\{c_{I_1}, c_{I_2}\} \left\{ \sum_{T \in \mathcal{T}_h} h_T \|Res_T(u_h)\|_{0,2;T} \|v\|_{1,2;\tilde{\omega}T} \right. \\ &\qquad \left. + \sum_{E \in \mathcal{E}_h} h_E^{1/2} \|Res_E(u_h)\|_{2;E} \|v\|_{1,2;\tilde{\omega}E} \right\}. \end{aligned}$$

Eine Umordnung der Terme erlaubt wiederum die Cauchy-Schwarzsche Ungleichung

$$\sum_i a_i b_i \le \{\sum_i a_i^2\}^{1/2} \{\sum_i b_i^2\}^{1/2},$$

womit

$$a(e,v) \leq \max\{c_{I_1}, c_{I_2}\} \qquad \left\{ \sum_{T\in\mathcal{T}_h} h_T^2 \|Res_T(u_h)\|_{0,2;T}^2 + \sum_{E\in\mathcal{E}_h} h_E \|Res_E(u_h)\|_{2;E}^2 \right\}^{1/2}$$
$$\left\{ \sum_{T\in\mathcal{T}_h} \|v\|_{1,2;\tilde{\omega}T}^2 + \sum_{E\in\mathcal{E}_h} \|v\|_{1,2;\tilde{\omega}E}^2 \right\}^{1/2}.$$

Der letzte Klammerausdruck erfüllt die Ungleichung

$$\left\{ \sum_{T\in\mathcal{T}_h} \|v\|_{1,2;\tilde{\omega}T}^2 + \sum_{E\in\mathcal{E}_h} \|v\|_{1,2;\tilde{\omega}E}^2 \right\}^{1/2} \leq c_{\mathcal{T}} \|v\|_1 = c_{\mathcal{T}} \|v\|_X$$

wobei $c_{\mathcal{T}}$ vom kleinsten Winkel der Triangulierung anhängt. Die konstanten Faktoren werden zu einem Faktor c zusammengefaßt, und die Ungleichung durch $\|v\|_X$ dividiert. Damit folgt

$$\frac{a(e,v)}{\|v\|_X} \leq c \left\{ \sum_{T\in\mathcal{T}_h} h_T^2 \|Res_T(u_h)\|_{0,2;T}^2 + \sum_{E\in\mathcal{E}_h} h_E \|Res_E(u_h)\|_{2;E}^2 \right\}^{1/2} =: c \left\{ \sum_{T\in\mathcal{T}_h} \eta_{R,T}^2 \right\}^{1/2}.$$

Die Norm der Testfunktion v taucht nun nicht mehr auf, und der Fehler kann durch die residuellen Ausdrücke aus der berechneten Lösung global nach oben abgeschätzt werden

$$c_a \|e\|_X \leq \sup_{v\in X\setminus\{0\}} \frac{l(v) - a(u_h, v)}{\|v\|_X} \leq c' \left\{ \sum_{T\in\mathcal{T}_h} \eta_{R,T}^2 \right\}^{1/2}.$$

Für das vereinfachte Modellproblem der Poisson-Gleichung erhält man den folgenden Fehlerschätzer für ein beliebiges Element $T \in \mathcal{T}_h$

$$\begin{aligned} \eta_{R,T} \;=\; & \Bigg(h_T^2 \, \|f_T\|_{0,2;T}^2 \; + \frac{1}{2} \sum_{E\subset\partial T\cap\Omega} h_E \, \|[\boldsymbol{n}_E \cdot \nabla u_h]_E\|_{2;E}^2 \\ & + \sum_{E\subset\partial T\cap\Gamma_N} h_E \, \|(t_{N,E} - \boldsymbol{n}_E \cdot \nabla u_h)\|_{2;E}^2 \Bigg)^{1/2}. \end{aligned}$$

Dabei sind f_T und $t_{N,E}$ für $T \in \mathcal{T}_h$ und $E \in \mathcal{E}_{h,N}$ definiert als

$$f_T := \frac{1}{|T|} \int_T f \;, \qquad t_{N,E} := \frac{1}{h_E} \int_E t_N.$$

Der Fehlerschätzer $\eta_{R,T}$ enthält neben den bekannten Daten f und t_N und der Lösung u_h des diskreten Problems nur geometrische Daten der Triangulierung. In diesem Sinne handelt es sich um einen *a posteriori* Fehlerschätzer. Die Integration der allgemeinen Funktionen f und t_N können sich in vielen Fällen schwierig oder gar unmöglich gestalten. Daher sollten sie z.B. durch eine passende Quadratur approximiert werden. Die obigen Näherungen f_T und $t_{N,E}$ stellen dabei denkbar einfache Möglichkeiten dar. Sie ersetzen f bzw. t_N durch deren L^2-Projektionen auf Räume stückweiser konstanter Funktionen

bezüglich $\mathcal{T}_h$ bzw. $\mathcal{E}_{h,N}$.
Unter Berücksichtigung, daß die Innenkanten doppelt gezählt werden, und mit der Dreiecksungleichung folgt eine globale obere Schranke für den Fehler

$$\|u-u_h\|_{1,2} \le c\left\{\sum_{T\in\mathcal{T}_h}\eta_{R,T}^2+\sum_{T\in\mathcal{T}_h}h_T^2\|f-f_T\|_{0,2;T}^2+\sum_{E\in\mathcal{E}_{h,N}}h_E\|t_N-t_{N,E}\|_{2;E}^2\right\}^{1/2}.$$

Dabei sind die beiden letzten Ausdrücke Störungsterme höherer Ordnung.

Um eine Abschätzung des Fehler nach unten geben zu können, wird zunächst ein beliebiges Element $T\in\mathcal{T}_h$ betrachtet und dafür

$$w_T := f_T b_T$$

definiert. Mit (5.1) folgt

$$\int_T f_T w_T = \frac{9}{20}|T||f_T|^2 = \frac{9}{20}\|f_T\|_{0,2;T}^2.$$

Da supp $w_T\subset T$ und die Gleichungen (5.4) und (5.5) gültig sind, folgt

$$\begin{aligned}\int_T f_T w_T &= \int_T f\ w_T+\int_T(f_T-f)w_T\\ &= \int_\Omega f\ w_T+\int_{\Gamma_N}t_N w_T-\int_\Omega\nabla u_h\cdot\nabla w_T+\int_T(f_T-f)w_T\\ &= \int_T\nabla(u-u_h)\cdot\nabla w_T+\int_T(f_T-f)w_T\\ &\le \|u-u_h\|_{1,2;T}\|\nabla w_T\|_{0,2;T}+\|f-f_T\|_{0,2;T}\|w_T\|_{0,2;T}.\end{aligned}$$

Unter Ausnutzung der Eigenschaften (5.1) folgt

$$\begin{aligned}\int_T f_T w_T &\le \|u-u_h\|_{1,2;T}c_5h_T^{-1}|f_T|\|b_T\|_{0,2;T}+\|f-f_T\|_{0,2;T}|f_T|\|b_T\|_{0,2;T}\\ &\le \|u-u_h\|_{1,2;T}c_5h_T^{-1}|f_T|\left\{\int_T b_T\right\}^{1/2}+\|f-f_T\|_{0,2;T}|f_T|\left\{\int_T b_T\right\}^{1/2}\\ &\le \sqrt{\frac{9}{20}}\|f_T\|_{0,2;T}\left(c_5h_T^{-1}\|u-u_h\|_{1,2;T}+\|f-f_T\|_{0,2;T}\right).\end{aligned}$$

Faßt man die Gleichung und die Ungleichung für $\int_T f_T w_T$ zusammen, so erhält man

$$\|f_T\|_{0,2;T}\le\sqrt{\frac{20}{9}}\left(c_5h_T^{-1}\|u-u_h\|_{1,2;T}+\|f-f_T\|_{0,2;T}\right). \tag{5.6}$$

Als nächstes soll eine beliebige Innenkante $E\in\mathcal{E}_{h,\Omega}$ betrachtet werden. Man definiert

$$w_E := [\boldsymbol{n}_E\cdot\nabla u_h]_E b_E.$$

Mit (5.1) gilt

$$\int_E[\boldsymbol{n}_E\cdot\nabla u_h]_E w_E=\frac{2}{3}h_E|[\boldsymbol{n}_E\cdot\nabla u_h]_E|^2=\frac{2}{3}\|[\boldsymbol{n}_E\cdot\nabla u_h]_E\|_{2;E}^2.$$

Da supp $w_E \subset \omega_E$ und die Gleichungen (5.4) und (5.5) gültig sind, folgt

$$\begin{aligned}\int_E [\boldsymbol{n}_E \cdot \nabla u_h]_E w_E &= \sum_{T' \subset \omega_E} \int_{T'} f\ w_E - \int_\Omega f\ w_E - \int_{\Gamma_N} t_N w_E + \int_\Omega \nabla u_h \cdot \nabla w_E \\ &= \sum_{T' \subset \omega_E} \int_{T'} f\ w_E - \int_{\omega_E} \nabla(u - u_h) \cdot \nabla w_E \\ &\leq \|f\|_{0,2;\omega_E} \|w_E\|_{0,2;\omega_E} + \|u - u_h\|_{1,2;\omega_E} \|\nabla w_E\|_{0,2;\omega_E}.\end{aligned}$$

Mit (5.1) und der Ungleichung (5.6) folgt

$$\begin{aligned}\int_E [\boldsymbol{n}_E \cdot \nabla u_h]_E w_E &\leq \|f\|_{0,2;\omega_E} |[\boldsymbol{n}_E \cdot \nabla u_h]_E| \|b_E\|_{0,2;\omega_E} \\ &\quad + \|u - u_h\|_{1,2;\omega_E} |[\boldsymbol{n}_E \cdot \nabla u_h]_E| c_6 h_E^{-1} \|b_E\|_{0,2;\omega_E} \\ &\leq \|f\|_{0,2;\omega_E} |[\boldsymbol{n}_E \cdot \nabla u_h]_E| \left\{ \int_{\omega_E} b_E \right\}^{1/2} \\ &\quad + \|u - u_h\|_{1,2;\omega_E} |[\boldsymbol{n}_E \cdot \nabla u_h]_E| c_6 h_E^{-1} \left\{ \int_{\omega_E} b_E \right\}^{1/2} \\ &\leq \sqrt{2c_4} \|[\boldsymbol{n}_E \cdot \nabla u_h]_E\|_{2;E} \left(h_E^{1/2} \|f\|_{0,2;\omega_E} + c_6 h_E^{-1/2} \|u - u_h\|_{1,2;\omega_E} \right) \\ &\leq \sqrt{2c_4} \|[\boldsymbol{n}_E \cdot \nabla u_h]_E\|_{2;E} \left(h_E^{1/2} \sum_{T' \subset \omega_E} \|f - f_{T'}\|_{0,2;T'} \right. \\ &\quad \left. + h_E^{1/2} \sum_{T' \subset \omega_E} \|f_{T'}\|_{0,2;T'} + c_6 h_E^{-1/2} \|u - u_h\|_{1,2;\omega_E} \right) \\ &\leq \hat{c}\ \|[\boldsymbol{n}_E \cdot \nabla u_h]_E\|_{2;E} \left(h_E^{1/2} \sum_{T' \subset \omega_E} \|f - f_{T'}\|_{0,2;T'} + h_E^{-1/2} \|u - u_h\|_{1,2;\omega_E} \right)\end{aligned}$$

mit

$$\hat{c} = \sqrt{2c_4} \max \left\{ 1 + \sqrt{\frac{20}{9}}\ ,\ c_6 + \sqrt{\frac{20}{9}} c_5 \max_{T' \subset \omega_E} h_E / h_{T'} \right\}.$$

Kombiniert man die obige Gleichung und Ungleichung für $\int_E [\boldsymbol{n}_E \cdot \nabla u_h]_E w_E$, so erhält man

$$\|[\boldsymbol{n}_E \cdot \nabla u_h]_E\|_{2;E} \leq \frac{2}{3} \hat{c} \left(h_E^{1/2} \sum_{T' \subset \omega_E} \|f - f_{T'}\|_{0,2;T'} + h_E^{-1/2} \|u - u_h\|_{1,2;\omega_E} \right). \tag{5.7}$$

Nun soll eine beliebige Kante auf dem Neumann-Rand $E \in \mathcal{E}_{h,N}$ betrachtet werden. Man definiert analog

$$w_E := (t_{N,E} - \boldsymbol{n}_E \cdot \nabla u_h) b_E.$$

Mit (5.1) gilt entsprechend

$$\int_E (t_{N,E} - \boldsymbol{n}_E \cdot \nabla u_h) w_E = \frac{2}{3} \|t_{N,E} - \boldsymbol{n}_E \cdot \nabla u_h\|_{2;E}^2$$

und

$$\begin{aligned}
\int_E (t_{N,E} - \boldsymbol{n}_E \cdot \nabla u_h) w_E &= \int_E (t_N - \boldsymbol{n}_E \cdot \nabla u_h) w_E + \int_E (t_{N,E} - t_N) w_E \\
&= \int_\Omega f \; w_E + \int_{\Gamma_N} t_N w_E - \int_\Omega \nabla u_h \cdot \nabla w_E \\
&\quad - \int_{\omega_E} f \; w_E + \int_E (t_{N,E} - t_N) w_E \\
&= \int_{\omega_E} \nabla (u - u_h) \cdot \nabla w_E - \int_{\omega_E} f \; w_E + \int_E (t_{N,E} - t_N) w_E \\
&\leq \|t_{N,E} - \boldsymbol{n}_E \cdot \nabla u_h\|_{2;E} \left(\sqrt{c_4} c_6 h_E^{-1/2} \|u - u_h\|_{1,2;\omega_E} \right. \\
&\quad \left. + \sqrt{c_4} h_E^{1/2} \|f\|_{0,2;\omega_E} + \sqrt{\frac{2}{3}} h_E^{1/2} \|t_N - t_{N,E}\|_{2;E} \right) \\
&\leq \check{c} \, \|t_{N,E} - \boldsymbol{n}_E \cdot \nabla u_h\|_{2;E} \left(h_E^{-1/2} \|u - u_h\|_{1,2;\omega_E} \right. \\
&\quad \left. + h_E^{1/2} \|f - f_{T'}\|_{0,2;T'} + h_E^{1/2} \|t_N - t_{N,E}\|_{2;E} \right),
\end{aligned}$$

wobei

$$\check{c} = \max \left\{ \sqrt{\frac{2}{3}} \; , \; \sqrt{c_4} \left(1 + \sqrt{\frac{20}{9}} \right) \; , \; \sqrt{c_4} \left(c_6 + \sqrt{\frac{20}{9}} c_5 h_E / h_{T'} \right) \right\}.$$

Kombiniert man die Gleichung und Ungleichung für $\int_E (t_{N,E} - \boldsymbol{n}_E \cdot \nabla u_h) w_E$ erhält man

$$\begin{aligned}
\|t_{N,E} - \boldsymbol{n}_E \cdot \nabla u_h\|_{2;E} \leq \frac{3}{2} \check{c} \; & \left(h_E^{-1/2} \|u - u_h\|_{1,2;\omega_E} \right. \\
& \left. + h_E^{1/2} \|f - f_{T'}\|_{0,2;T'} + h_E^{1/2} \|t_N - t_{N,E}\|_{2;E} \right). \qquad (5.8)
\end{aligned}$$

Mit den Ungleichungen (5.6), (5.7) und (5.8) gelangt man schließlich auf eine lokale untere Schranke für den Fehler

$$\eta_{R,T} \leq \tilde{c} \left\{ \|u - u_h\|_{1,2;\omega_T}^2 + \sum_{T' \subset \omega_T} h_{T'}^2 \|f - f_{T'}\|_{0,2;T'}^2 + \sum_{E \in \mathcal{E}(T) \cap \mathcal{E}_{h,N}} h_E \|t_N - t_{N,E}\|_{2;E}^2 \right\}^{1/2},$$

die gültig für alle $T \in \mathcal{T}_h$ ist. Die Konstante $\tilde{c}$ hängt nur vom kleinsten Winkel der Triangulierung ab. Die beiden letzten Ausdrücke sind Störungsterme höherer Ordnung.

5.3.2 Ein residueller Fehlerschätzer für das Modellproblem

Das betrachtete Modellproblem

$$\begin{aligned}
-\nabla \cdot (\mathbf{D} \cdot \nabla u) + \boldsymbol{b} \cdot \nabla u + \alpha u &= f \qquad \text{in } \Omega \\
u &= u_D \quad \text{auf } \Gamma_D \\
\boldsymbol{n} \cdot \mathbf{D} \cdot \nabla u &= t_N \quad \text{auf } \Gamma_N
\end{aligned}$$

besitzt, wie bereits im Kapitel 3 beschrieben, für die Standard Galerkin Methode die folgende Bilinear- und Linearform:

$$\begin{aligned} a(u,v) &:= \int_\Omega (\mathbf{D}\cdot\nabla u)\nabla v + \int_\Omega \boldsymbol{b}\cdot\nabla uv + \int_\Omega \alpha uv \\ l(v) &:= \int_\Omega fv + \int_{\Gamma_N} t_N v. \end{aligned}$$

Für die stabilisierten Methoden lautet das diskrete Problem : Finde $u_h \in S^{k,0}_{h,D}$ so, daß

$$a_\delta(u_h, v_h) = l_\delta(u_h) \qquad \forall v_h \in S^{k,0}_{h,D}.$$

Dabei stellen sich die erweiterte Bilinear- und Linearform wie folgt dar:

$$\begin{aligned} a_\delta(u_h,v_h) &:= a(u_h,v_h) + \sum_{T\in\mathcal{T}_h} \delta\left(-\nabla\cdot(\mathbf{D}\cdot\nabla u_h) + \boldsymbol{b}\cdot\nabla u_h + \alpha u_h \ , \ \boldsymbol{b}\cdot\nabla v_h\right)_T \\ l_\delta(u_h) &:= l(u_h) + \sum_{T\in\mathcal{T}_h} \delta\left(f \ , \ \boldsymbol{b}\cdot\nabla v_h\right)_T. \end{aligned}$$

Bekanntermaßen geht diese SUPG Diskretisierung für $\delta = 0$ in die Standard-Galerkin Approximation über.
Bei Anwendung eines Interpolationsoperators I_h gilt für die Bilinearform $a(e,v)$ mit beliebigem $v \in X$:

$$a(u-u_h, v) = a(u-u_h, v-I_h v) + a(u-u_h, I_h v).$$

Die elementweise partielle Integration führt für alle $\bar{v} \in X$ auf den Ausdruck

$$\begin{aligned} a(e,\bar{v}) &= \sum_{T\in\mathcal{T}_h}\int_T f\bar{v} + \int_{\Gamma_N} t_N\bar{v} \\ &\quad - \sum_{T\in\mathcal{T}_h}\{-\int_T \nabla\cdot(\mathbf{D}\cdot\nabla u_h)\bar{v} + \int_T \boldsymbol{b}\cdot\nabla u_h\bar{v} + \int_T \alpha u_h\bar{v} + \int_{\partial T} \boldsymbol{n}_T\cdot\mathbf{D}\cdot\nabla u_h\bar{v}\} \\ &= \sum_{T\in\mathcal{T}_h}\int_T \left(f + \nabla\cdot(\mathbf{D}\cdot\nabla u_h) - \boldsymbol{b}\cdot\nabla u_h - \alpha u_h\right)\bar{v} \\ &\quad + \sum_{E\in\mathcal{E}_{h,\Omega}}\int_E -\left[\boldsymbol{n}_E\cdot\mathbf{D}\cdot\nabla u_h\right]_E\bar{v} + \sum_{E\in\mathcal{E}_{h,N}}\int_E \left(t_N - \boldsymbol{n}_E\cdot\mathbf{D}\cdot\nabla u_h\right)\bar{v} \\ &= \sum_{T\in\mathcal{T}_h}\int_T Res_T(u_h)\bar{v} + \sum_{E\in\mathcal{E}_h}\int_E Res_E(u_h)\bar{v}. \end{aligned}$$

Dabei sind $Res_T(u_h)$ der residuelle Anteil aus dem Elementgebiet und $Res_E(u_h)$ der Anteil über die Elementkanten

$$Res_T(u_h) := f_h + \nabla\cdot(\mathbf{D}\cdot\nabla u_h) - \boldsymbol{b}\cdot\nabla u_h - \alpha u_h$$

$$Res_E(u_h) := \begin{cases} -\left[\boldsymbol{n}_E\cdot\mathbf{D}\cdot\nabla u_h\right]_E \ , & \text{wenn E } \in \mathcal{E}_{h,\Omega} \\ \left(t_N - \boldsymbol{n}_E\cdot\mathbf{D}\cdot\nabla u_h\right) \ , & \text{wenn E } \in \mathcal{E}_{h,N} \\ 0 \ , & \text{wenn E } \in \mathcal{E}_{h,D}. \end{cases}$$

Der residuelle Anteil auf dem Elementinneren entspricht dabei der starken Form der Differentialgleichung.

Setzt man nun $\bar{v} = v - I_h v$, berücksichtigt (5.3) und mit der Annahme von isotroper, homogener Diffusion, erhält man unter Verwendung der Cauchy-Schwarz-Ungleichung

$$a(e, v - I_h v) \leq c_b \left\{ \sum_{T \in \mathcal{T}_h} \alpha_T^2 \|Res_T(u_h)\|_{0,2;T}^2 + \sum_{E \in \mathcal{E}_h} \beta_E \|Res_E(u_h)\|_{0;E}^2 \right\}^{1/2}$$

mit

$$\begin{aligned} \alpha_T &= \min\{h_T \epsilon^{-1/2}, 1\} \\ \beta_E &= \epsilon^{-1/2} \min\{h_E \epsilon^{-1/2}, 1\}. \end{aligned}$$

Für die Bilinearform gilt

$$a(e, v_h) = - \sum_{T \in \mathcal{T}_h} \delta \int_T Res_T(u_h) \boldsymbol{b} \cdot \nabla v_h$$

für alle $v_h \in S_{h,D}^{k,0}$. Ein einfaches Skalierungsargument zeigt für alle $v_h \in S_{h,D}^{k,0}$, daß [60]

$$\|\boldsymbol{b} \cdot \nabla v_h\|_{0;T} \leq c_c \|\boldsymbol{b}\|_{L^\infty(T)} h_T^{-1} \min\{h_T \epsilon^{-1/2}, 1\} |\!|\!| v_h |\!|\!|_T.$$

Mit diesen Beziehungen und (5.3) sowie der Annahme, daß $\delta \leq h_T$, folgt

$$a(e, I_h v) \leq c_d \left\{ \sum_{T \in \mathcal{T}_h} \alpha_T^2 \|Res_T(u_h)\|_{0,2;T}^2 \right\}^{1/2}.$$

Damit erhalten wir für den Fehler eine obere Schranke und einen Fehlerschätzer $\eta_{R,T}$:

$$\frac{a(e, v)}{\|v\|_X} \leq c \left\{ \sum_{T \in \mathcal{T}_h} \alpha_T^2 \|Res_T(u_h)\|_{0,2;T}^2 + \sum_{E \in \mathcal{E}_h} \beta_E \|Res_E(u_h)\|_{2;E}^2 \right\}^{1/2} =: c \left\{ \sum_{T \in \mathcal{T}_h} \eta_{R,T}^2 \right\}^{1/2}.$$

Die Herleitung einer unteren Schranke für das Modellproblem erfolgt in analoger Weise zum Poisson-Problem (s. [60]).

5.3.3 Modifikationen

Im Falle von geringer Konvektion und Adsorption kann eventuell der Term des Elementresiduums vernachlässigt werden. Damit erhält man eine Modifikation des Fehlerschätzers $\eta_{R,T}$ der Form :

$$\eta_{J,T} = \left(\frac{1}{2} \sum_{E \subset \partial T \cap \Omega} \beta_E \left\| [\boldsymbol{n}_T \cdot \mathbf{D} \cdot \nabla u_h]_E \right\|_{2;E}^2 + \sum_{E \subset \partial T \cap \Gamma_N} \beta_E \left\| (t_N - \boldsymbol{n}_T \cdot \mathbf{D} \cdot \nabla u_h) \right\|_{2;E}^2 \right)^{1/2}.$$

Bei Verwendung anderer Interpolationsoperatoren in der Herleitung (s. (5.2)) können die Faktoren α_T und β_E in $\eta_{R,T}$ durch

$$\begin{aligned} \alpha_T &= h_T \\ \beta_E &= h_E. \end{aligned}$$

ersetzt werden. Dies ist die klassische Form wie sie für das Poisson-Problem hergeleitet wurde. Die Annahme von isotroper, homogener Diffusion ist dabei nicht nötig.
Eine weitere Möglichkeit zur Bereitstellung eines Netzverfeinerungskriteriums stellt die getrennte Verwendung der beiden Residuenterme dar

$$\begin{aligned} \eta_{R,T}^2 &= \alpha_T^2 \|Res_T(u_h)\|_{0,2;T}^2 + \sum_{E\in\partial T} \beta_E \|Res_E(u_h)\|_{2;E}^2 \\ &= \left(\eta_T^{(1)}\right)^2 + \left(\eta_T^{(2)}\right)^2. \end{aligned}$$

Die Anteile werden unabhängig voneinander als Verfeinerungsindikatoren benutzt und in die später beschriebenen Markierungsstrategien gespeist.

5.4 Fehlerschätzer basierend auf lokalen Hilfsproblemen

Eine weitere Klasse von Fehlerschätzern stellen diejenigen dar, die auf der Lösung von sogenannten Hilfsproblemen basieren. Es wird geprüft, inwieweit sich die Lösung ändert, wenn die Ordnung der Ansatzpolynome erhöht wird. Dies geschieht in Rücksicht auf die Rechenintensität nicht auf dem gesamten Gebiet, sondern auf kleinen Teilgebieten oder einzelnen Elementen. Es gibt verschiedene Möglichkeiten, Hilfsprobleme zu stellen:

- Hilfsproblem mit Dirichlet Randbedingungen auf ω_x
- Hilfsproblem mit Dirichlet Randbedingungen auf ω_T
- Hilfsproblem mit Neumann Randbedingungen auf T.

Generell werden dabei die folgenden Bedingungen an das Hilfsproblem gestellt:

1. Um Informationen über das lokale Verhalten des Fehlers zu erhalten, sollte das Hilfsproblem auf einem möglichst kleinen Bereich gestellt sein.

2. Die Finite Element Methode für das Hilfsproblem sollte hinreichend genau sein, d.h. genauer als die original gewählte Methode (z.B. Übergang von linearen Elementen zu quadratischen).

3. Um die rechnerische Arbeit klein zu halten, sollte die FEM des Hilfsproblems möglichst einfach sein, d.h. möglichst wenig Freiheitsgrade enthalten.

4. Zu jeder Kante und zu jedem Dreieck sollte mindestens ein Freiheitsgrad in mindestens einem Hilfsproblem gehören.

5.4.1 Neumann Randbedingungen für ein Hilfsproblem auf T

Dieser Fehlerschätzer geht auf Bank und Weiser [3] für die Laplace-Gleichung zurück. Er basiert auf der Lösung eines lokalen, diskreten Hilfsproblems mit Neumann-Randbedingungen auf Elementniveau, d.h. genau auf einem Element. Für jedes Element T wird mit V_T der Raum der bubble-Funktionen, die entweder im Schwerpunkt oder in einem

Kantenmittelpunkt den Wert 1 annehmen und sonst auf dem Rand verschwinden, für jedes Element $T \in \mathcal{T}_h$ definiert

$$V_T := \text{span}\{b_T, b_E : E \in \mathcal{E}(T) \backslash \mathcal{E}_{h,D}\}.$$

Zu lösen ist das folgende Hilfsproblem : Finde $v_T \in V_T$, so daß

$$\int_T \epsilon \nabla v_T \cdot \nabla w + \int_T \boldsymbol{b} \cdot \nabla v_T w + \int_T \alpha v_T w = \int_T Res_T(u_h) w + \sum_{E \subset \partial T} \int_E Res_E(u_h) w \quad \forall w \in V_T.$$

Dabei ist v_T die Finite Element-Approximation der Lösung u_T eines lokalen Konvektions-Diffusionsproblems mit den Elementresiduen als Quellterm und Kantenresiduen als Randfluß

$$\begin{aligned} -\epsilon \Delta u_T + \boldsymbol{b} \cdot \nabla u_T + \alpha u_T &= Res_T(u_h) && \text{in } T \\ \epsilon \, \partial_{\boldsymbol{n}_T} u_T &= Res_E(u_h) && \text{auf } \partial T \backslash \Gamma_D \\ u_T &= 0 && \text{auf } \partial T \cap \Gamma_D. \end{aligned}$$

In der Energienorm bildet v_T den Fehlerschätzer $\eta_{N,T}$

$$\eta_{N,T} := |\!|\!| v_T |\!|\!|_\alpha := \left\{ \epsilon \|\nabla v_T\|^2_{0,2;T} + \alpha \|v_T\|^2_{0,2;T} \right\}^{1/2}$$

Zur Berechnung von v_T ist in jedem Element T die Lösung eines 4×4 Gleichungssystems notwendig. Die entstehenden Elementmatrizen zur Lösung des Hilfsproblems sind im Anhang 1 zu finden.

5.5 Numerische Beispiele

Zur Durchführung der folgenden adaptiven Beispielrechnungen wurden im Vorgriff auf das folgende Kapitel eine Markierungsstrategie (Strategie 2) und die reguläre Verfeinerungsmethode verwendet. Aspekte dieser und anderer Markierungsstrategien und Verfeinerungsmethoden werden im Kapitel 6 eingehend betrachtet.

5.5.1 Laplace-Problem

Diesem ersten Beispiel liegt die Laplace-Gleichung zugrunde, d.h. $\epsilon = 1$, $\boldsymbol{b} = (0,0)$, $\alpha = 0$. Betrachtet wird das Gebiet $\Omega = (-1,1) \times (-1,1)$, auf dessen gesamten Rand homogene Dirichlet Bedingungen gelten. Die Funktion f der rechten Seite sei gegeben mit

$$f = -\frac{4C_2}{C_1^2} \left(\frac{x^2+y^2}{C_1^2} - 1 \right) e^{\frac{-(x^2+y^2)}{C_1^2}}.$$

Für dieses Problem ist die analytische Lösung bekannt:

$$u(x,y) = C_2 e^{\frac{-(x^2+y^2)}{C_1^2}}.$$

Hier wurden $C_1 = 1$ und $C_2 = 10$ gewählt. Für die Diskretisierung wurde wegen der fehlenden Konvektion das Standard-Galerkin Verfahren herangezogen. Die Verfeinerung

erfolgt regulär mittels der Markierungsstrategie 2 ($c_{ref} = 0.5$, $p = 0.15$), siehe Kapitel 6. Die folgende Abbildung zeigt die verfeinerten Netze nach 5 Verfeinerungsstufen mit dem Z^2-Indikator, dem residuellen Fehlerschätzer (min-Bedingungen für Vorfaktoren) und dem Neumann-Fehlerschätzer.

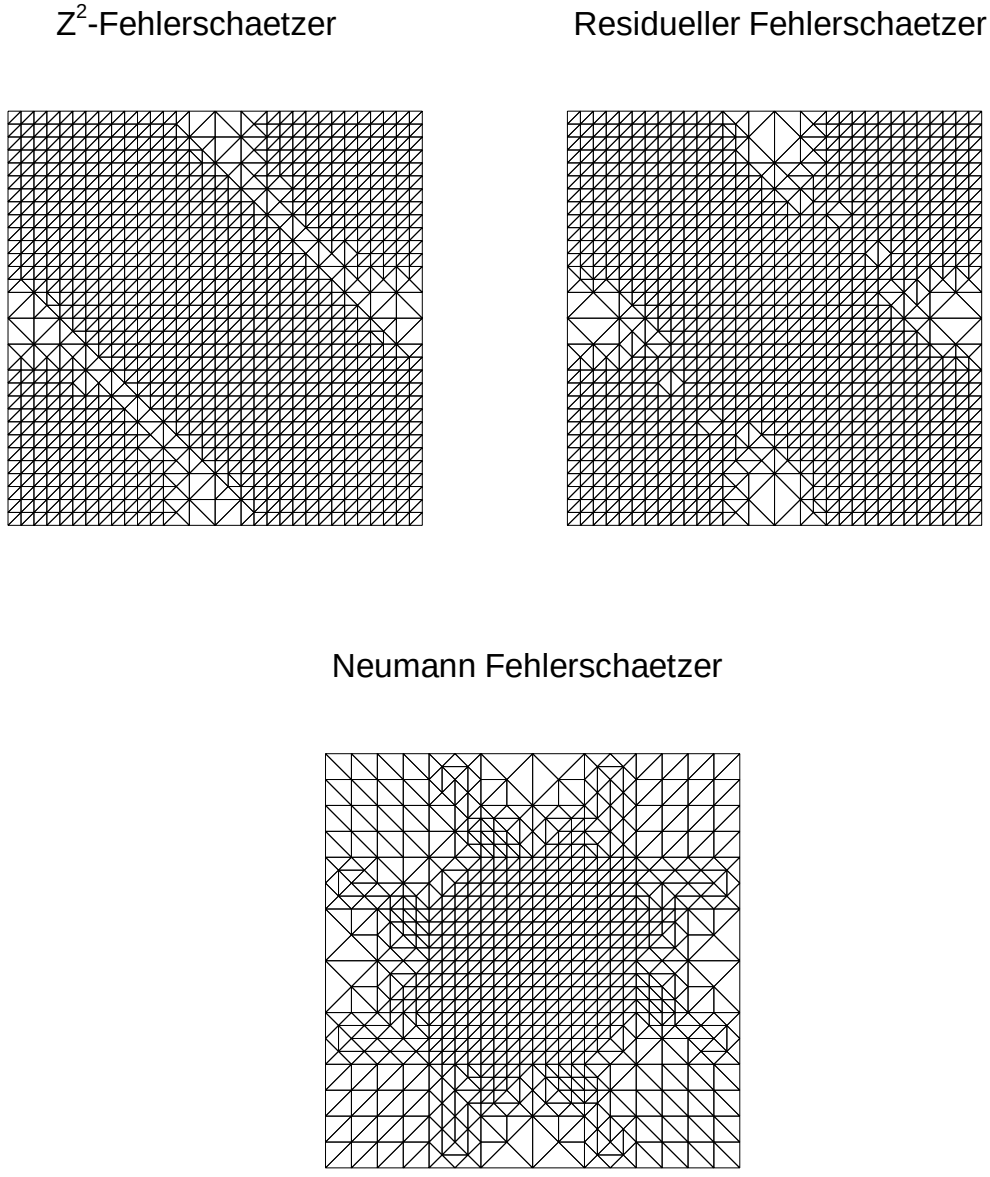


Abbildung 5.2: Laplace-Problem : Verschiedene Fehlerschätzer

Für das vorliegende symmetrische Problem liefern alle drei Fehlerschätzer symmetrische Netze. Mit 1884 Elementen bzw. 1856 Elementen sind die Netze des Z^2-Fehlerschätzers und des residuellen Fehlerschätzers erheblich dichter als das des Neumann Fehlerschätzers (1136 Elemente). Die Netze von Z^2- und residuellem Fehlerschätzer sind abgesehen von wenigen Aussparung sehr nahe an dem entsprechenden uniform verfeinerten Netz (2048 Elemente).

In den folgenden Tabellen 5.1-5.3 sind neben der Anzahl der Elemente und der Unbekannten jeweils der globale ($e = \|u - u_h\|$) und der geschätzte Fehler (η), sowie der Effektivitätsindex (η/e) über die einzelnen Verfeinerungsstufen angegeben.

Level	N_{el}	N_{free}	e	η	eff
1	8	9	8.5014	7.61008239166	0.89516
2	32	25	7.2777	5.43261726004	0.74647
3	128	81	6.9520	2.93680193527	0.42244
4	500	281	6.8597	1.46561971817	0.21366
5	1884	1001	6.8403	0.74988228421	0.10963

Tabelle 5.1: Laplace-Problem : Z^2-Fehlerschätzer

Level	N_{el}	N_{free}	e	η	eff
1	8	9	8.5014	21.5263459622	2.5321
2	32	25	7.2777	18.7983402689	2.5830
3	128	81	6.9520	10.8805481144	1.5651
4	488	273	6.8524	5.6910983291	0.8305
5	1856	985	6.8419	2.9250186560	0.4275

Tabelle 5.2: Laplace-Problem : Residueller Fehlerschätzer

Level	N_{el}	N_{free}	e	η	eff
1	8	9	8.5014	26.0881782385	3.0687
2	28	21	7.2772	27.4596892361	3.7734
3	80	49	7.2793	21.6081740933	2.9684
4	296	165	6.9451	10.6712345064	1.5365
5	1136	597	6.8501	5.5379228759	0.8084

Tabelle 5.3: Laplace-Problem : Neumann Fehlerschätzer

Hinsichtlich der Abnahme der globalen Fehlers liefern alle drei Fehlerschätzer adäquate Ergebnisse. Bemerkenswert ist dabei, daß hier der Neumann Schätzer mit ungefähr zwei Dritteln der Elemente der anderen Schätzern auskommt.

Die numerische Lösung ist für alle drei Fälle glatt und von entsprechender Form wie für den hier beispielhaft dargestellten Fall des Neumann Fehlerschätzers, Abbildung 5.3.

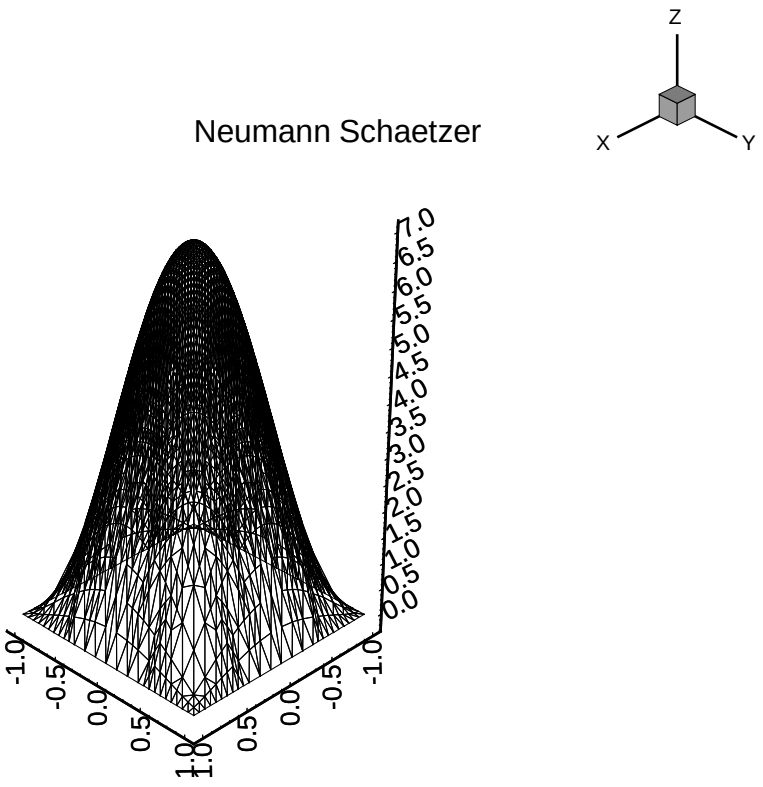


Abbildung 5.3: Numerische Lösung des Laplace-Problems

5.5.2 Gekrümmte innere Schockfront

Das folgende Beispiel wurde bereits in Kapitel 3 beschrieben. Zur Berechnung wurde die SDFEM mit den ebenfalls in Kapitel 3 angegebenen Parametern verwandt. Zur Verfeinerung der Elemente dienten die Markierungsstrategie 2 ($c_{ref} = 0.5$, $p = 0.15$) und eine reguläre Verfeinerung.
Die folgende Abbildung 5.4 zeigt die verfeinerten Netze jeweils nach 7 Verfeinerungsstufen bei Verwendung des Z^2-Indikators, des residuellen Fehlerschätzers (min-Bedingungen für Vorfaktoren) und des Neumann Fehlerschätzers.

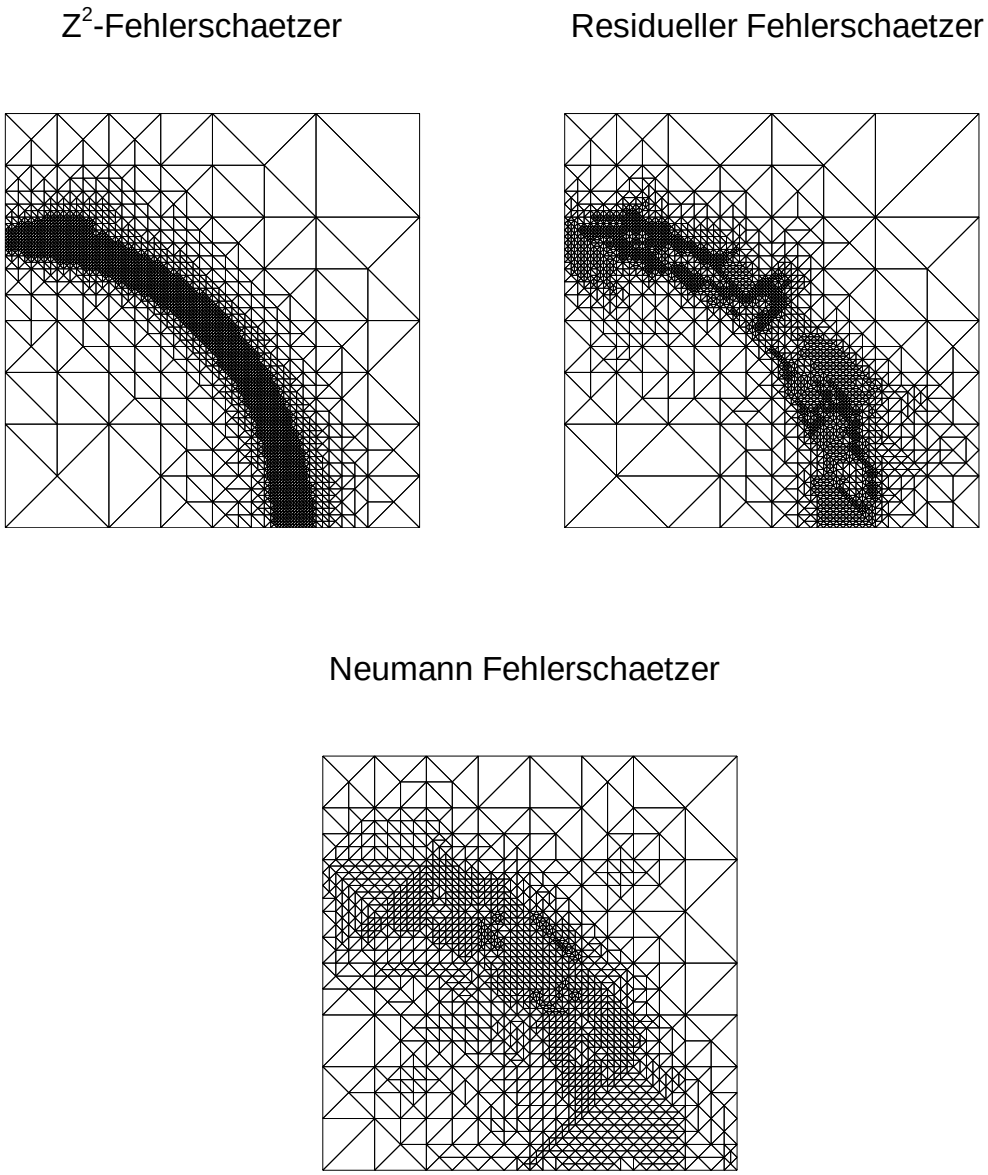


Abbildung 5.4: Gekrümmte Schockfront : Verschiedene Fehlerschätzer

Das Netz des Z^2-Fehlerschätzers ist dem Verlauf der schmalen Schockfront am besten angepaßt und in diesem Bereich sehr dicht. Auch das Netz des residuellen Fehlerschätzers

folgt der Schockfront und ist im wesentlichen in deren Bereich verdichtet, allerdings erstreckt es sich in größerer Breite. Die Netzdichte des Netzes des Neumann Schätzers ist deutlich geringer, aber der Verlauf der Schockfront wird wiedergegeben.

In Tabelle 5.4 ist die Abnahme des globalen Fehlers der numerischen Lösungen bei Verwendung der verschiedenen Fehlerschätzer zusammengefaßt. Da eine exakte analytische Lösung für dieses Beispiel nicht zur Verfügung stand, wurde die jeweilige numerische Approximation mit der bekannten Lösung des hyperbolischen Grenzproblems verglichen.

Level	Z^2-Fehlerschätzer	Residueller Fehlerschätzer	Neumann Fehlerschätzer
1	0.24586	0.24586	0.24586
2	0.09333	0.08589	0.21612
3	0.09712	0.16400	0.21263
4	0.09799	0.11719	0.16784
5	0.06859	0.10670	0.12059
6	0.06142	0.07666	0.10712
7	0.05396	0.07023	0.07691

Tabelle 5.4: Globaler Fehler verschiedener Fehlerschätzer

Die deutlich besten Ergebnisse erzielt hier der Z^2-Fehlerschätzer. Der Neumann Schätzer ergibt bezüglich der Reduzierung des Fehlers e die schlechtesten Ergebnisse, allerdings werden dabei weniger Elemente verwendet. Außerdem zeigt er im Verlauf des Verfeinerungsprozeß ein stetig abnehmendesVerhalten, während die beiden anderen in den Schritten 3 und 4 Sprünge zeigen.

5.5.3 Innere Schockfront und Randgrenzschicht

Auch dieses Beispiel wurde schon in Kapitel 3 vorgestellt. Die Randbedingungen und Parameter werden übernommen, nur der Diffusionskoeffizient wird hier kleiner gewählt, $\epsilon = 10^{-10}$. Als Diskretisierungsmethode dient die SDFEM mit den üblichen Parametern. Zur Verfeinerung werden die Markierungsstrategie 2 ($c_{ref} = 0.5$, $p = 0.15$) und die reguläre Verfeinerung benutzt.
Abbildung 5.5 zeigt die verfeinerten Netze nach jeweils 6 Verfeinerung unter Zugrundelegung der verschiedenen Fehlerschätzer.
In allen drei Fällen werden Elemente in den schmalen Bereich der Randgrenzschicht gebracht, da dort die Lösung von 0 auf 1 springt. Der Sprung innerhalb des Gebietes nimmt wegen der Adsorption beginnend vom Ursprung ab, deshalb wird auch nahe des Ursprungs verdichtet. Die Bedeutung des Gradienten der Lösung innerhalb der Z^2-Abschätzung wird hier sehr deutlich, da entlang der gesamten Schockfront verdichtet wird. Der residuelle Fehlerschätzer und der Neumann Fehlerschätzer investieren in diesem Bereich weniger Elemente.

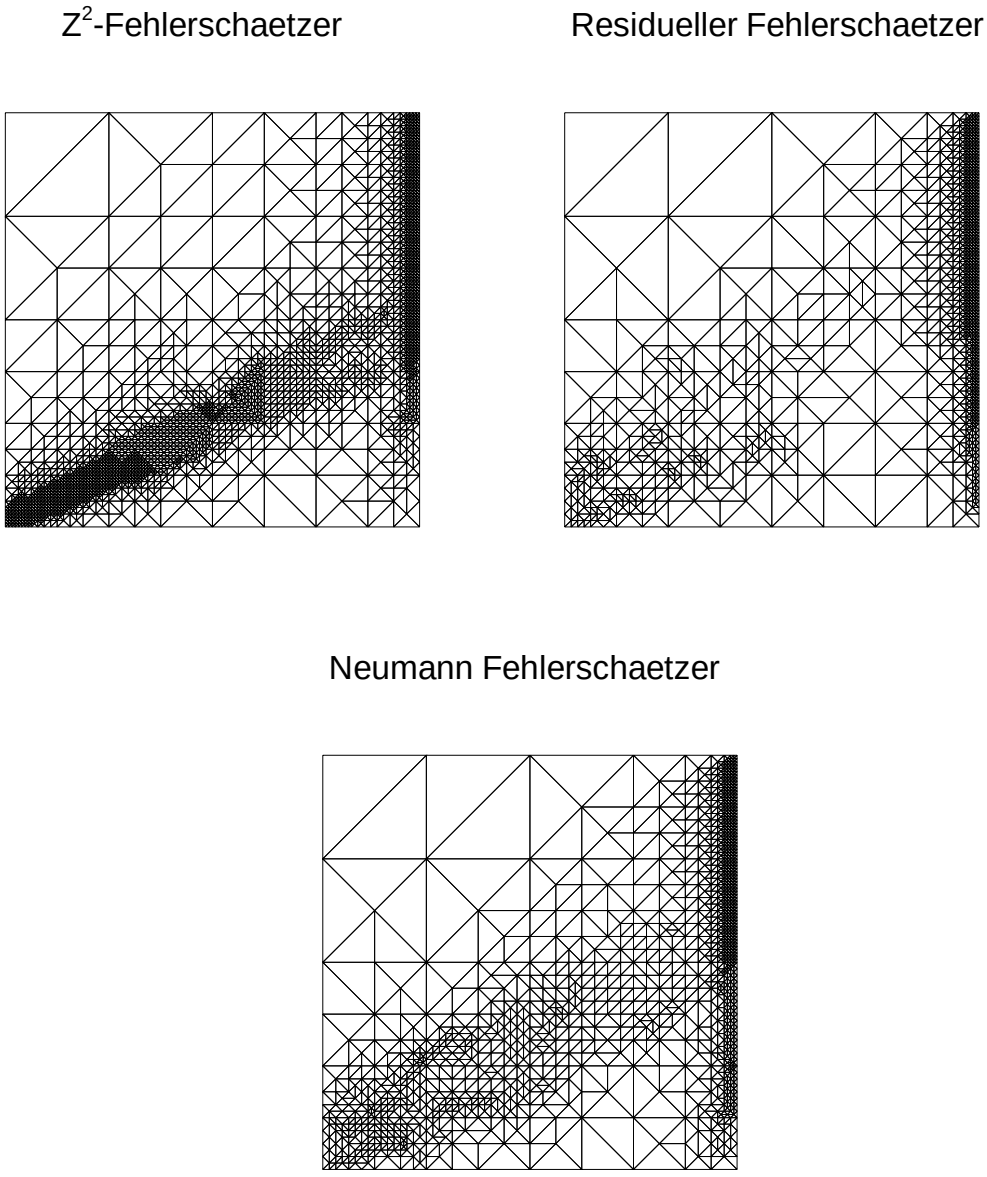


Abbildung 5.5: Z^2- , Residueller , Neumann Fehlerschätzer, $\epsilon = 10^{-10}$

Um Aussagen über die Größe des Fehlers der numerischen Lösungen treffen zu können, wird eine Vergleichslösung auf einem uniform verfeinerten Netz berechnet. Die folgenden Tabellen 5.5-5.7 geben jeweils den globalen, den relativen ($\|u - u_h\|/\|u\|$) und den geschätzten Fehler, sowie den Effektivitätsindex (η/e) für die verschiedenen Fehlerschätzer wieder.

Level	N_{el}	N_{free}	e	$\bar{e}$	η	eff
1	32	25	0.12936	25.14039 %	0.75749720713	5.8556
2	111	69	0.09336	19.63500 %	0.95080713453	10.184
3	289	164	0.08232	18.25361 %	1.23335058057	14.983
4	708	384	0.06452	14.62367 %	1.59606775951	24.738
5	1714	902	0.04944	11.36505 %	2.09245377921	42.320
6	3964	2055	0.03855	8.92580 %	2.80376113890	54.274

Tabelle 5.5: Z^2-Fehlerschätzer, $\epsilon = 10^{-10}$

Level	N_{el}	N_{free}	e	$\bar{e}$	η	eff
1	32	25	0.12936	25.14039 %	2.66614808028	20.610
2	83	53	0.14669	30.79115 %	3.47223918942	23.671
3	192	113	0.11352	24.96747 %	5.09567248880	44.887
4	460	257	0.11184	25.36955 %	7.00795195195	62.662
5	956	521	0.08561	19.62437 %	10.1384023411	118.42
6	1848	998	0.07114	16.46008 %	14.4386484321	202.95

Tabelle 5.6: Residueller Fehlerschätzer, $\epsilon = 10^{-10}$

Level	N_{el}	N_{free}	e	$\bar{e}$	η	eff
1	32	25	0.12936	25.14039 %	0.209065046017	1.6161
2	85	53	0.15704	33.57044 %	0.154707933533	0.9851
3	211	121	0.10612	23.32342 %	0.126631403459	1.1932
4	494	271	0.11106	24.98255 %	0.090953655894	0.8189
5	1098	587	0.08336	19.09005 %	0.068287727824	0.8192
6	2308	1219	0.06034	13.95732 %	0.049306174243	0.8171

Tabelle 5.7: Neumann Fehlerschätzer, $\epsilon = 10^{-10}$

Der globale Fehler e nimmt beim residuellen Schätzer in absoluten Zahlen am wenigstens ab. Außerdem zeigt er wie auch der Neumann Schätzer ein unregelmäßiges Verhalten, während der Z^2 stetig und stark abnimmt. Der geschätzte Fehler η des residuellen Schätzers steigt im Verlauf der Verfeinerung ebenso wie der des Z^2. Der Neumann

Fehlerschätzer hingegen schätzt den Fehler von Verfeinerung zu Verfeinerung kleiner. Eine Erklärung für das Verhalten des residuellen Schätzers mag dessen Anfälligkeit für Schwingungen in der numerischen Lösung sein. Da bei einer geringen Diffusion von $\epsilon = 10^{-10}$ und der Verwendung der SDFEM in der numerischen Lösung Über- und Unterschwingungen auftreten, soll noch der Fall $\epsilon = 10^{-2}$ betrachtet werden.
Die Abbildung 5.6 zeigt die verfeinerten Netze mit den verschiedenen Fehlerschätzern für diesen Fall nach 6 jeweils Verfeinerungsstufen.

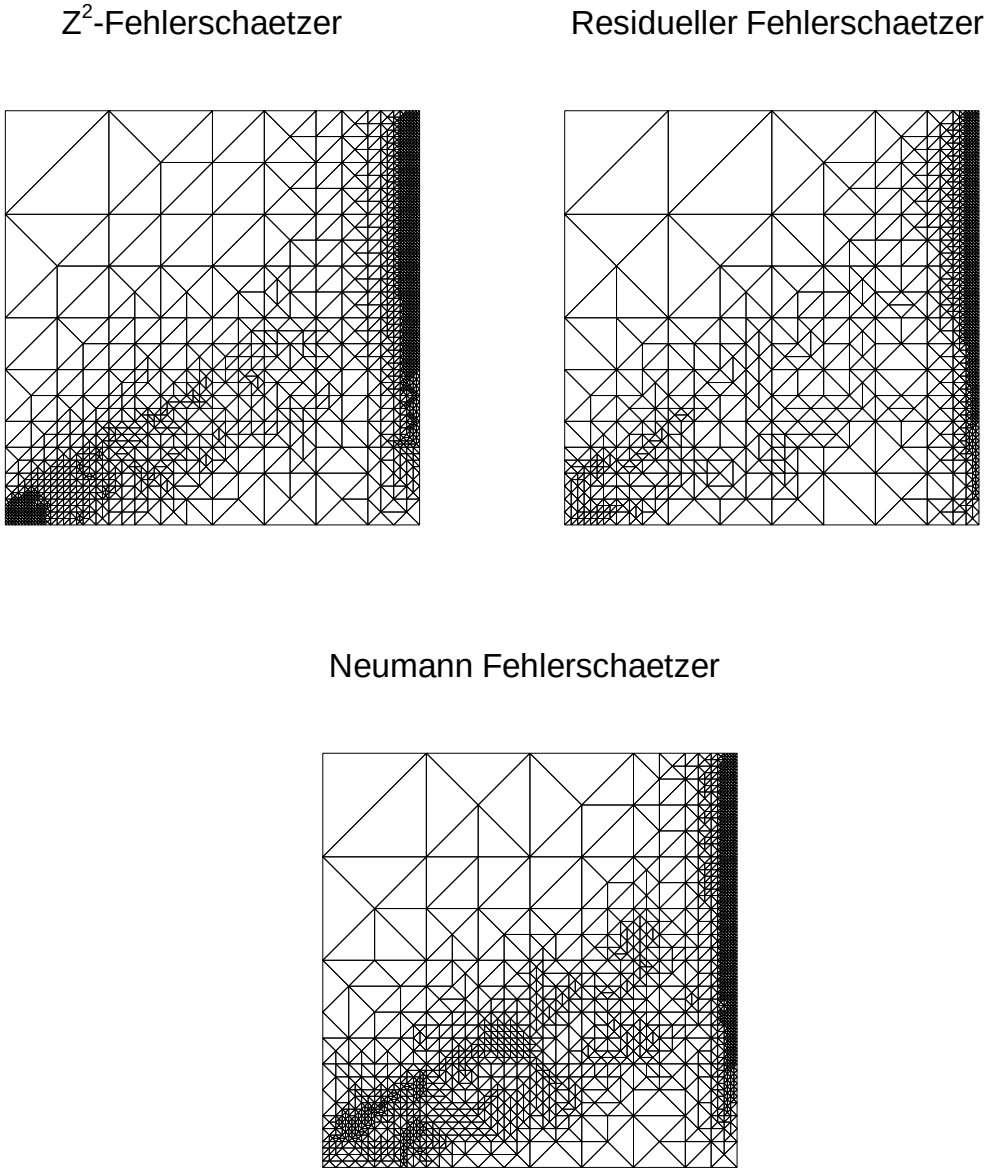


Abbildung 5.6: Z^2- , Residueller , Neumann Fehlerschätzer, $\epsilon = 10^{-2}$

Alle drei Fehlerschätzer verdichten die Netze nahe des Ursprungs und insbesondere im Bereich der Randgrenzschicht am Ausflußrand, da dort die größten Sprünge vorliegen. Dies entspricht dem Trend der Verdichtung im Fall von $\epsilon = 10^{-10}$. Wegen der deut-

lich größeren Diffusion und damit bedingten Ausschmierung der inneren Schockfront verdichtet nun auch der Z^2-Schätzer nicht mehr sehr stark im Inneren des Gebietes.

Die Tabellen 5.8-5.10 zeigen die üblichen numerischen Ergebnisse. Der residuell geschätzte Fehler nimmt nun im wesentlichen im Verlauf des Verfeinerungsprozesses ab, ebenso wie der des Z^2-Schätzers .

Level	N_{el}	N_{free}	e	$\bar{e}$	η	eff
1	32	25	0.16393	30.27679 %	0.73878185111	4.5066
2	111	69	0.12345	23.16426 %	0.87381793726	7.0782
3	298	170	0.09943	17.73874 %	1.03808535258	10.440
4	677	369	0.07835	12.77379 %	1.21716462291	15.534
5	1448	773	0.01554	2.36140 %	1.40479326518	90.419

Tabelle 5.8: Z^2-Fehlerschätzer, $\epsilon = 10^{-2}$

Level	N_{el}	N_{free}	e	$\bar{e}$	η	eff
1	32	25	0.16393	30.27679 %	2.63927065001	16.100
2	83	53	0.21089	39.80078 %	3.38819543665	16.066
3	176	105	0.14058	25.55004 %	2.48210865699	17.657
4	418	236	0.17083	28.26516 %	1.65998714821	9.717
5	931	509	0.10116	15.54908 %	1.17421126005	11.607

Tabelle 5.9: Residueller Fehlerschätzer, $\epsilon = 10^{-2}$

Level	N_{el}	N_{free}	e	$\bar{e}$	η	eff
1	32	25	0.16393	30.27679 %	0.840181015592	5.125
2	94	58	0.21570	40.98799 %	0.812783113515	3.768
3	218	126	0.17061	30.79727 %	1.65851360959	9.721
4	544	301	0.11899	19.61428 %	1.12183582624	9.428
5	1278	685	0.05134	7.85728 %	0.810633916634	15.788

Tabelle 5.10: Neumann Fehlerschätzer, $\epsilon = 10^{-2}$

Abschließend zu diesem Beispiel soll Abbildung 5.7 das Auftreten von Über- und Unterschwingungen bei der Verwendung der SDFEM im Falle von $\epsilon = 10^{-10}$ darstellen. Links die numerische Lösung auf dem durch den residuellen Fehlerschätzer verfeinerten Netz für $\epsilon = 10^{-10}$ und rechts entsprechend für $\epsilon = 10^{-2}$.

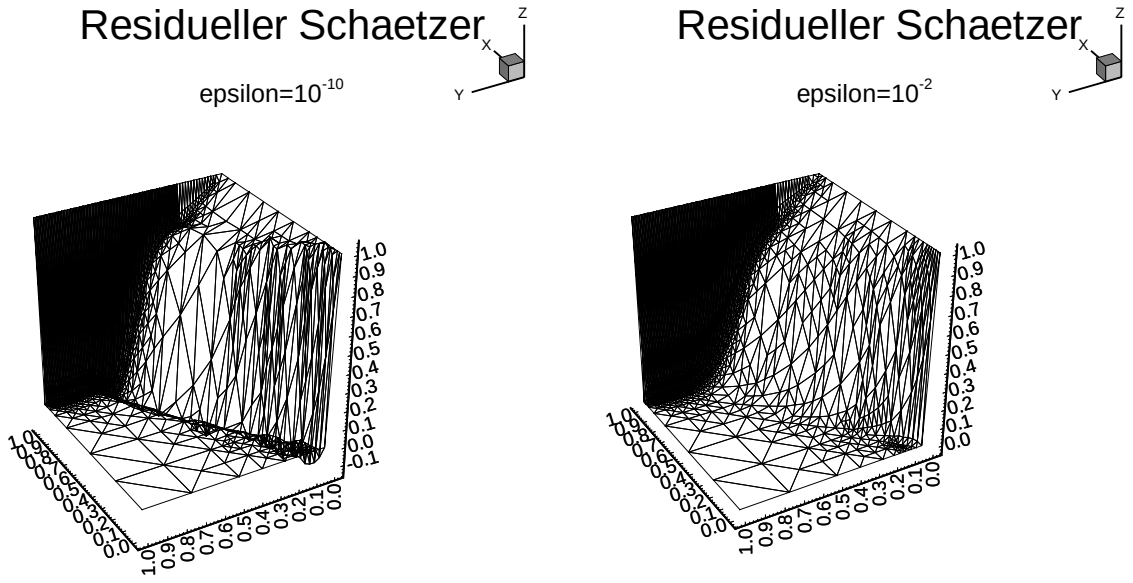


Abbildung 5.7: Lösung für residuellen Fehlerschätzer: $\epsilon = 10^{-10}$, $\epsilon = 10^{-2}$

5.5.4 Dirichlet Randbedingungen mit Sprung

Bei diesem Beispiel handelt es sich um ein Konvektions-Diffusionsproblem auf dem Einheitsquadrat mit konstantem Geschwindigkeitsvektor $\boldsymbol{b} = (1, 0)$, ohne Adsorption und Senken bzw. Quellen. Der Diffusionskoeffizient sei $\epsilon = 10^{-3}$. Bezeichnend für dieses Beispiel sind Sprünge in den Dirichlet Randbedingungen auf dem Einflußrand

$$\begin{aligned} \epsilon \partial_{\boldsymbol{n}} u &= 0 \ , \text{ wenn } x = 1 \\ u_D(x,y) &= 0 \ , \quad \text{wenn } \begin{cases} y = 0 & \text{und} \quad 0 < x < 1 \\ y = 1 & \text{und} \quad 0 < x < 1 \\ x = 0 & \text{und} \quad |y - 0.5| > 0.05 \end{cases} \\ u_D(x,y) &= 1 \quad \text{sonst.} \end{aligned}$$

Die analytische Lösung dieses Problems ist

$$u(x,y) = \sum_{k=1}^{\infty} \varphi_k(x)\psi_k(y) = \sum_{k=1}^{\infty} (\gamma_k \varphi_{1k}(x) + \beta_k \varphi_{2k}(x))\psi_k(y),$$

wobei

$$\begin{aligned} \psi_k(y) &= \sin(k\pi y) \\ \varphi_{1k}(x) &= e^{x/(2\epsilon)} \frac{\sinh(g_k(1-x))}{\sinh(g_k)} \\ \varphi_{2k}(x) &= e^{(x-1)/(2\epsilon)} \frac{\sinh(g_k x)}{\sinh(g_k)} \end{aligned}$$

mit

$$g_k = \sqrt{\frac{1}{(2\epsilon)^2} + (k\pi)^2}.$$

Die Konstanten lauten

$$\begin{aligned}\gamma_k &= \frac{2}{k\pi}(\cos(0.45k\pi) - \cos(0.55k\pi)) \\ \beta_k &= \gamma_k \left(\frac{g_k e^{1/(2\epsilon)}}{\sinh(g_k)}\right) / \left(\frac{1}{2\epsilon} + g_k \frac{\cosh(g_k)}{\sinh(g_k)}\right).\end{aligned}$$

Zur Diskretisierung diente die SDFEM. Die Verfeinerung erfolgte regulär mit der Markierungsstrategie 2 ($c_{ref} = 0.5$, $p = 0.10$).

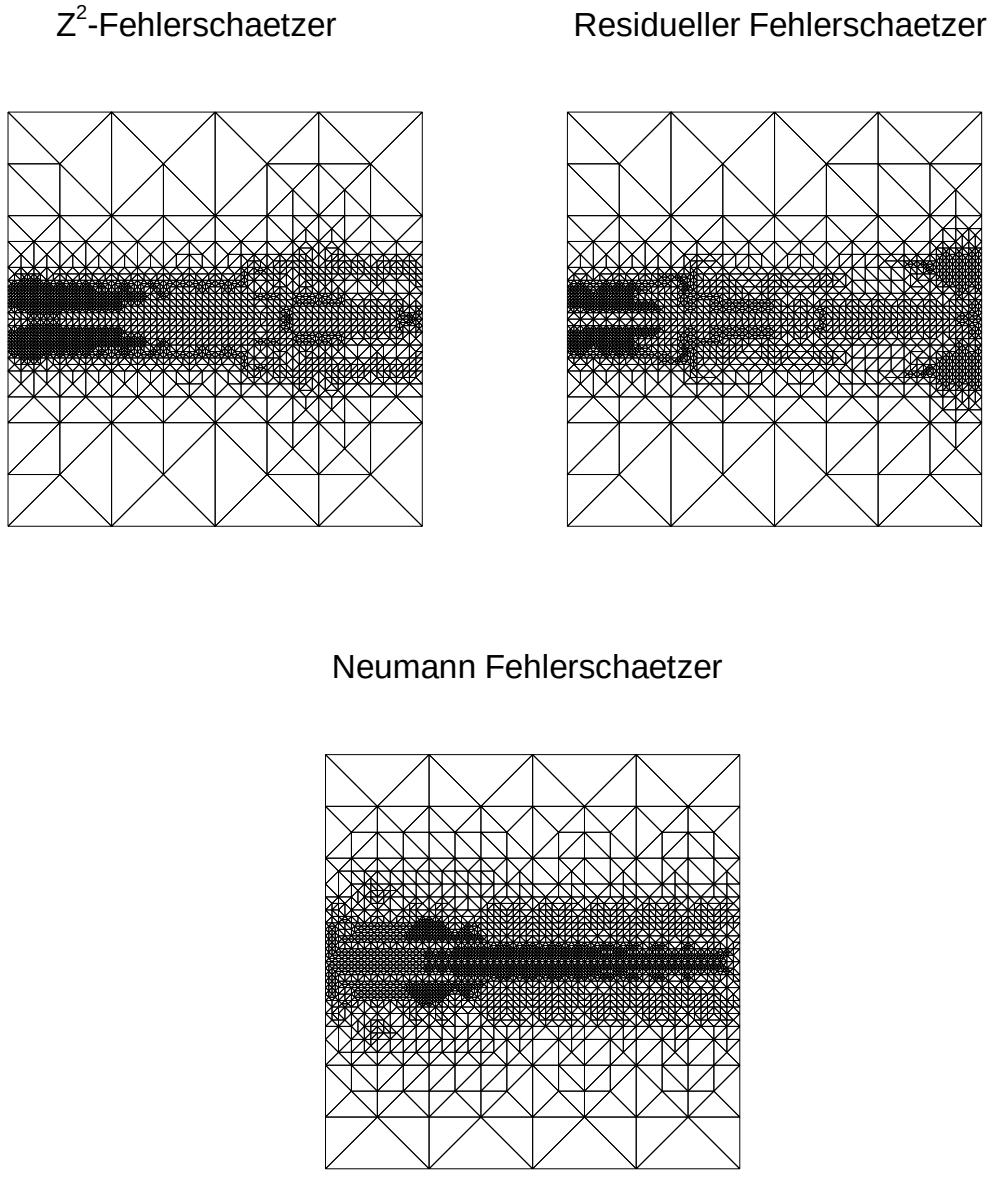


Abbildung 5.8: Dirichlet RB.: Z^2-, Residueller, Neumann Fehlerschätzer

Die Abbildung 5.8 zeigt die verfeinerten Netze unter Verwendung der verschiedenen Fehlerschätzer. In der folgenden Abbildung 5.9 sind die zugehörigen numerischen Lösungen dreidimensional dargestellt.

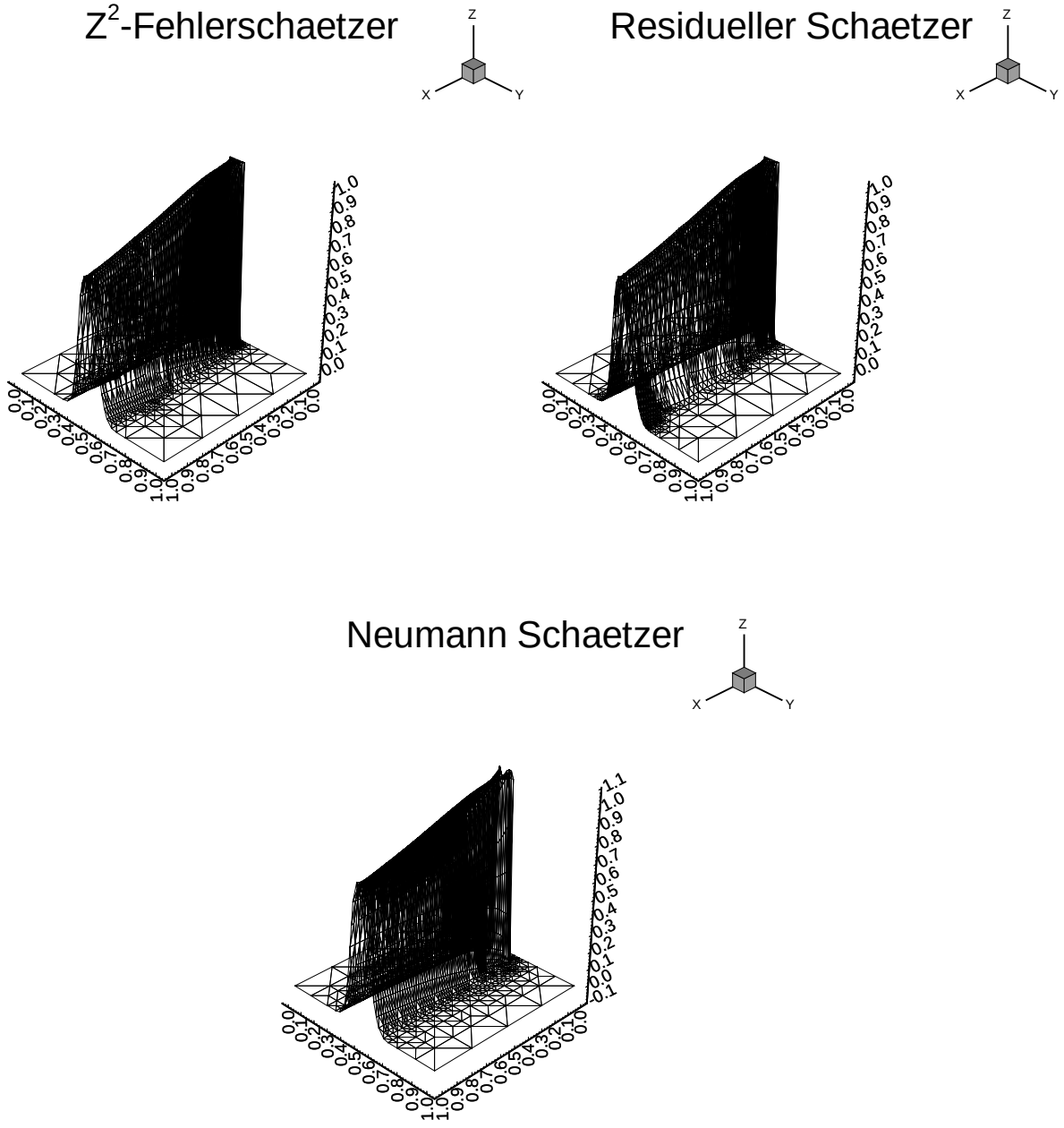


Abbildung 5.9: Dirichlet RB.: Z^2-, Residueller, Neumann Fehlerschätzer

Bei diesem Beispiel handelt es sich wieder um ein symmetrisches Problem. Alle drei Fehlerschätzer liefern symmetrische Netze. Die Netze aus der Verwendung des residuellen und des Z^2-Fehlerschätzers sind stärker im Bereich der Sprünge am Einflußrand und am Ausflußrand verdichtet. Der Neumann Schätzer hingegen konzentriert mehr Elemente in den mittleren Bereich des abfallenden Kammes. Aus diesem Grund ergeben sich bei der numerischen Lösung mit dem Neumann Schätzer leichte Überschwingungen

am Einflußrand. Diese schlagen sich in den Ergebnissen für den globalen bzw. relativen Fehler nieder.

Level	N_{el}	N_{free}	e	$\bar{e}$	η	eff
1	32	25	0.055034	14.95694 %	2.19249434065	39.839
2	106	65	0.027872	9.94667 %	2.65469933061	95.245
3	252	141	0.077698	29.98419 %	2.51171073768	32.327
4	521	273	0.034233	11.79560 %	1.97089694995	57.574
5	1418	732	0.039016	13.21034 %	1.49993032137	38.444
6	3670	1865	0.014987	5.01080 %	1.17344575848	78.300

Tabelle 5.11: Dirichlet RB.: Z^2-Fehlerschätzer

Level	N_{el}	N_{free}	e	$\bar{e}$	η	eff
1	32	25	0.055034	14.95694 %	0.006009265176	1.0919
2	80	51	0.054719	19.35733 %	0.237140578900	4.3338
3	204	115	0.079759	30.74257 %	0.249433934681	3.1274
4	510	273	0.025365	8.74507 %	0.269606390863	10.629
5	1246	647	0.037014	12.53711 %	0.196177329365	5.3000
6	3280	1675	0.014000	4.68091 %	0.135536086535	9.6810

Tabelle 5.12: Dirichlet RB.: Residueller Fehlerschätzer

Level	N_{el}	N_{free}	e	$\bar{e}$	η	eff
1	32	25	0.05503	14.95694 %	0.647019408060	11.757
2	104	62	0.22112	77.75015 %	0.457711074707	2.0700
3	334	181	0.08887	34.11302 %	0.393923489175	4.4324
4	766	401	0.12440	42.94433 %	0.345887686771	2.7805
5	1714	879	0.04694	15.90435 %	0.339364305857	7.2302
6	4542	2299	0.04896	16.39173 %	0.284985751012	5.8211

Tabelle 5.13: Dirichlet RB.: Neumann Fehlerschätzer

Die Tabellen 5.11-5.13 geben in gewohnter Weise die numerischen Ergebnisse für den globalen, den relativen und den geschätzten Fehler sowie den Effektivitätsindex wieder. Die numerischen Lösungen wurden mit der entsprechenden näherungweisen analytischen Lösung verglichen, wobei die Summen bis $k = 100$ berechnet wurden.
Der Z^2-Fehlerschätzer zeigt die Tendenz den Fehler zu überschätzen, während sowohl der residuelle Fehlerschätzer als auch der Neumann-Fehlerschätzer den Fehler deutlich unterschätzen.

5.6 Zusammenfassung

Dieses Kapitel soll ein Vergleich der vorgestellten Fehlerschätzer und Indikatoren abschließen.
Der Z^2-Indikator ist problemunabhängig und aufgrund seiner einfachen Struktur sehr leicht in ein Programm zu integrieren. Der Z^2 berücksichtigt den Gradienten der numerischen Lösung. Da dieser an Schockfronten und Grenzschichten entscheidend ist, werden diese kritischen Regionen gut erfaßt. Die entsprechend verfeinerten Gitter liefern gute Ergebnisse. Phänomene höherer Ordnung wie numerische Oszillationen an den Grenzschichten, z.B. bei Verwendung der SDFEM, können vom Z^2 allerdings nicht gesehen werden. Damit stellt sich die Frage, inwieweit Informationen bei sehr komplizierten Problemen verlorengehen. Der Z^2 eignet sich als Indikator, aber eine quantitative Aussage über den Fehler kann er nicht geben.
Die Fähigkeit, das Auftreten von Unter- und Überschwingungen in der numerischen Lösung zu lokalisieren, besitzen der residuelle Fehlerschätzer und der Neumann-Fehlerschätzer. In ihrer Umsetzung in einem Programm sind beide aufwendiger, insbesondere der Neumann-Fehlerschätzer, da für jedes Element der Triangulierung ein lokales Problem gelöst werden muß. Aber beide ermöglichen eine quantitative Aussage über den Fehler.
Für dominierende Konvektion treten in der Approximation Oszillationen auf. Der residuelle Fehlerschätzer kann diese finden und berücksichtigen. Der Einfluß der Oszillationen scheint sich im adaptiven Prozeß zu verstärken. Dies zeigen die Tabelle 5.6, η steigt im Wert, während der Fehler e sinkt.
Zur Verteilung der lokalen Effektivitätsindizes ist zu bemerken, daß für das dritte Beispiel in den meisten Bereichen $\eta/e \approx 1$, aber entlang der Grenzschicht bzw. Schockfront größere Werte (von 3 bis zu 16) erreicht werden. Ebenso finden sich für die lokalen Werte des Fehlerschätzers die größeren Werte entlang der Fronten. Aus diesem Grund wird auch dort verfeinert, wo die großen Gradienten auftreten. Eine wesentliche Rolle spielt dabei der Term "$\boldsymbol{b} \cdot \nabla u_h$".
Nun noch eine Bemerkung zu der Minimum-Bedingung der Vorfaktoren des residuellen Fehlerschätzers $\alpha_T = \min\{h_T\epsilon^{-1/2}, 1\}$ und $\beta_E = \epsilon^{-1/2}\min\{h_E\epsilon^{-1/2}, 1\}$: Abhängig von der Größe von ϵ springen die Bedingungen im Verlauf des Verfeinerungsprozesses früher oder später an. So fällt die Bedingung für α_T bei $\epsilon = 10^{-10}$ erst bei $h_T < 10^{-5}$ ins Gewicht, während bei $\epsilon = 1$, nur $h_T < 1$ erfüllt sein muß und somit die Bedingung sofort gültig ist. Insbesondere zeichnet sich das Anspringen der Bedingung in dem Abfall der Werte von η ab. Der fehlende Einfluß des über die Verfeinerungen kleiner werdenden h_T schlägt sich bei kleinen Werten ϵ im Anstieg von η nieder.
Das letzte Beispiel zeigt deutlich die Problematik der Ausrichtung der Elemente. Im ersten Beispiel des elliptischen Laplace-Problems, liegt keine Richtungsabhängigkeit im Problem vor. Die Ausrichtung der Elemente verursacht dort keinerlei Schwierigkeiten. In der Konvektions-Diffusionsgleichung des letzten Beispiels ist eine charakteristische Richtung durch das Geschwindigkeitsfeld vorgegeben. Nicht in jedem Verfeinerungsschritt werden die Element wegen der regulären Unterteilung der Elemente für die vorgegebene Richtung geeignet plaziert. Dies schlägt sich in den schwankenden Werten des Fehlers nieder.

Kapitel 6

Verfeinerungsstrategien

Nach der Ermittlung einer Fehlerabschätzung für jedes Element einer Triangulierung ist zu entscheiden, welche Elemente zu verfeinern bzw. zu vergröbern sind. Diese Prozedur bezeichnet man als Markierungsmethode. Es gibt eine Vielzahl von Methoden, von denen hier nur einige dargestellt werden sollen. Einen weiteren Teil der Verfeinerungsstrategie stellt die eigentliche Verfeinerung bzw. auch Vergröberung der ausgewählten Elemente dar. Im Rahmen der stationären Betrachtung werden strukturierte, hierarchische Netze verwendet. Diese Netze werden durch reguläre Verfeinerung, d.h. die Unterteilung eines Dreiecks in vier, oder die Bisektion, d.h. die Teilung in zwei Dreiecke, aus einem vorgegebenen Grundnetz entwickelt.
Mit diesen Bausteinen kann der Ablauf eines adaptiven Algorithmus für ein stationäres Problem nach folgendem Schema erfolgen:

1. Anfangsnetz $\mathcal{T}_0$, setze $k = 0$
2. Löse diskretes Problem auf $\mathcal{T}_k$
3. Berechne für jedes $T \in \mathcal{T}_k$ *a posteriori* Fehlerschätzer
4. Wenn der geschätzte globale Fehler hinreichend klein ist: STOP, sonst Markierung der zu verfeinernden Elemente und konstruiere neues Netz $\mathcal{T}_{k+1}$, setze $k = k + 1$ und gehe zu Punkt 2.

Tabelle 6.1: Ablaufschema eines adaptiven Algorithmus

Bei instationären Problemen müssen die folgenden Änderungen im Algorithmus Beachtung finden:

- Die Genauigkeit der numerischen Lösung muß problemabhängig nach einigen Zeitschritten abgeschätzt werden, und eine Zeitschrittweitensteuerung muß vorgenommen werden.
- Der Verfeinerunsprozeß des räumlichen Gebiets sollte mit einer Zeitschrittkontrolle gekoppelt werden.
- Stellenweise Vergröberungen des Netzes könnten nötig sein.
- Eine vollständige Neuvernetzung könnte wünschenswert sein.

6.1 Markierungsmethoden

Den Fehlerindikator η_T eines Elements T mit einer Konstanten zu vergleichen und das Element zu verfeinern, falls der Indikator größer als diese Konstante sein sollte, stellt eine scheinbar einfache Vorgehensweise dar. Diese gestaltet sich aber nur dann als sinnvoll, wenn eine solche Konstante in sinnvoller Weise gewählt werden kann, was im allgemeinen unmöglich ist.

6.1.1 Strategie 1 - Maximalwert-Methode, Globalfehler-Methode

Eine weitverbreitete Methode speichert während der Berechnung des Fehlerindikators oder Fehlerschätzers den maximal berechneten Wert unter allen Elementwerten. Dieser Wert sei hier $\eta_{max} = \max\{\eta_T\}$ genannt. Ein Element T wird verfeinert, wenn es

$$\eta_T \geq c_{ref} * \eta_{max} \ , \qquad 0 < c_{ref} < 1$$

erfüllt. Diese Strategie soll im folgenden Maximalwert-Methode (Strategie 1) genannt werden.
Eine andere Strategie berechnet den globalen Fehler $(\sum \eta_T^2)^{1/2}$ und dividiert durch die Anzahl der Elemente. Dadurch erhält man einen gemittelten Fehler $\bar{\eta}_T$. Ein Element T wird verfeinert, wenn es

$$\eta_T \geq c_{ref} * \bar{\eta}_T \ , \qquad 0 < c_{ref}$$

erfüllt.

Wenn ein Fehlerschätzer verwendet wird, kann man auch einen möglichst kleinen Satz S von Elementen suchen, so daß

$$\left(\sum_{T \in S} \eta_T^2 \right)^{1/2} \leq c_{ref} * \left(\sum_{T} \eta_T^2 \right)^{1/2} \ , \qquad 0 < c_{ref} < 1 \ .$$

Die Elemente in S sind für einen bestimmten Teil des globalen Fehlers verantwortlich. Diese Strategie soll als Globalfehler-Methode bezeichnet werden.
In der Praxis verwendet man zunächst die Maximalwert-Methode mit einem festen c_{ref0}, wobei man einen Satz S_0 von Elementen erhält, der nicht die Bedingung in der Globalfehler-Methode erfüllen muß. In einem solchen Fall würde man c_{ref0} verkleinern und fortfahren bis man schließlich einen Satz von Elementen findet, der die Bedingung erfüllt.

6.1.2 Strategie 2 - Sortierungsmethode

Um mögliche Ausreißer des maximal berechneten Wertes η_{max} zu umgehen, kann man einen abgeschwächten Wert verwenden. Dazu wählt man die folgende pragmatische Vorgehensweise. Die berechneten Werte η_T werden ihrer Größe nach geordnet, so daß

$$\eta_{T_1^*} \geq \eta_{T_2^*} \geq .... \geq \eta_{T_{N_k}^*} \geq 0,$$

wobei N_k die Anzahl der Elemente in der Triangulierung $\mathcal{T}_k$ ist. In der Maximalwert-Methode soll nun der Maximalwert η_{max} durch den Wert $\eta_{T^*_{k_0}}$ ersetzt werden, der sozusagen eine untere Beschränkung der oberen 5 oder 10 Prozent der geordneten Elemente darstellt:

$$k_0 := p * N_k \qquad \text{mit z.B. } p = 0.05 \text{ oder } 0.1$$

Diese Vorgehensweise soll mit Sortierungsmethode (Strategie 2) bezeichnet werden.

6.1.3 Strategie 3 - Prozentsatzmethode

Wenn man in jedem Schritt nur eine sehr geringe Anzahl von Elementen zur Verfeinerung auswählt, so erwartet man, ein nahezu optimales Netz zu erhalten. Dies führt allerdings auf viele, möglicherweise teuere Iterationsschritte im Lösungsprozeß und generell auf lange Rechenzeiten. Wenn man hingegen sehr viele Elemente auswählt, so erreicht man wenige Iterationsschritte, aber eine sehr große Anzahl von Unbekannten. Man ist also gezwungen, Kompromisse einzugehen, um eine optimale Zeit zur Lösung des Problems zu erreichen.
Um das erste Phänomen zu verhindern, kann man einen gewissen Anstieg der Anzahl der Unbekannten nach der Verfeinerung fordern. Wenn mit der Maximalwert-Methode und einem gegebenen c_{ref0} der geforderte Prozentsatz nicht erreicht wird, muß c_{ref} verkleinert werden. Die Markierung ist dann erneut vorzunehmen. Wird die Globalfehler-Methode verwendet, dann wird der Satz S von Elementen solange erhöht bis sowohl die entsprechende Bedingung erfüllt, als auch der geforderte Prozentsatz erreicht ist.
Statt den Toleranzfaktor c_{ref} zu verkleinern kann man auch den entsprechenden Maximalwert der Forderung nach einer Mindestanzahl von neuen Elementen anpassen. Wenn der geforderte Prozentsatz nicht in einem ersten Schritt erreicht wird, wird unter den verbleibenden, d.h. noch nicht markierten Elementen, deren Maximalwert bestimmt. Anschließend wird die Markierung erneut nach Strategie 1 vorgenommen. Auf diese Methode wird später als Prozentsatzmethode (Strategie 3) zurückgegriffen werden.
Ihr Ablauf gestaltet sich mit den gegebenen Daten $\bar{\eta} := \max_{T\in\mathcal{T}_k}\{\eta_T\},\ c_{ref},\ \tilde{\mathcal{T}}_k = \mathcal{T}_k$ wie folgt

$$\begin{aligned}
&\forall T \in \tilde{\mathcal{T}}_k \ : \\
&\qquad \diamondsuit \quad \eta_T \geq c_{ref} * \bar{\eta} \ \rightarrow \ \text{verfeinere } T \text{ und } \tilde{\mathcal{T}}_k := \tilde{\mathcal{T}}_k \backslash T \\
&\qquad\qquad \text{Frage : } \quad N_k \geq (1+p)N_k \qquad ?? \\
&\qquad\qquad\qquad \text{nein : } \bar{\eta} := \max_{T\in\tilde{\mathcal{T}}_k}\{\eta_T\} \ \rightarrow \ \text{goto } \diamondsuit \\
&\qquad\qquad\qquad \text{ja : akzeptiere } \tilde{\mathcal{T}}_{k+1} \text{ und verfeinere}
\end{aligned}$$

6.1.4 Strategie 4 - Gleichverteilungsmethode

Eine weitere Strategie zum Finden einer Triangulierung $\mathcal{T}_{k+1}$ mit möglichst wenigen Elementen, so daß, mit gegebener Toleranz ToL,

$$\left\{ \sum_{T\in\mathcal{T}_k} \eta_T^2 \right\}^{1/2} \leq ToL$$

erfüllt wird, beruht auf der Idee der Gleichverteilung des Fehlers

$$\eta_T \sim \frac{ToL}{\sqrt{N_k}}.$$

Es wird angenommen, daß das Gitter optimal ist, wenn alle Elemente den gleichen Beitrag zum Gesamtfehler leisten. Deshalb soll sie als Gleichverteilungsmethode (Strategie 4) bezeichnet werden. Sie geht auf Babuška zurück. Man hat Kriterien sowohl zur Verfeinerung als auch zur Vergröberung, welche insbesondere bei transienten Problemen wichtig sind:

$$\text{Wähle} \qquad 0 < \theta_2 < \theta_1 \leq 1 \ :$$

$$\eta_T \geq \frac{\theta_1 \ ToL}{\sqrt{N_k}} \ \longrightarrow \ \text{verfeinere}$$

$$\eta_T \leq \frac{\theta_2 \ ToL}{\sqrt{N_k}} \ \longrightarrow \ \text{vergröbere}$$

Ein in der Praxis häufig gewählter Wert für θ_1 ist 0.5. Die Wahl der Parameter erlaubt die Einflußnahme auf den Charakter der Triangulierung. Für kleine Werte von θ_2/θ_1 erhält man eher uniforme Netze. Für größere Werte von θ_2/θ_1 wird der Gradient der lokalen Netzweite $h(\boldsymbol{x})$ steiler [1]. Die Toleranz ToL kann im Verfeinerungsprozeß adaptiv gewählt werden, derart, daß sie in jeder Verfeinerungsstufe proportional zur Netzweite h_{max} ist.

6.1.5 Strategie 5 - Prognosemethode

Es ist auch möglich eine Vorhersage über den Fehler auf dem nächsten Level durch lokale Extrapolation zu treffen und daraus eine Entscheidung über die Verfeinerung zu fällen. Dieses Vorgehen soll als Prognosemethode (Strategie 5) referiert werden.
Der Fehler in einem Element T verhält sich ungefähr wie $c \ h_T^\beta$, wobei c und β unbekannte Konstanten sind. Mit berechneten Fehlerschätzern für zwei verschiedene Triangulierungen $\mathcal{T}_0$ und $\mathcal{T}_1$, die eine sei z.B. durch regelmäßige Verfeinerung aus der anderen entstanden, lassen sich die Konstanten bestimmen. Damit läßt sich anschließend eine Aussage über den Fehler in einem Element T^*, das durch eine weitere Unterteilung von T gewonnen wird, treffen. Es gilt also

$$\begin{aligned} \eta_{T_0} &= c \ h_{T_0}^\beta \\ \eta_{T_1} &= c \ h_{T_1}^\beta = c \ 2^{-\beta} h_{T_0}^\beta. \end{aligned}$$

Daraus kann man die lokal gleichen Konstanten zu

$$\beta = \frac{\ln\left(\frac{\eta_{T_0}}{\eta_{T_1}}\right)}{\ln 2}$$

$$c = \eta_{T_0} \ h_{T_0}^{-\beta}$$

berechnen und mit ihnen nun eine Prognose über den zu erwartenden Fehler bei einer weiteren Verfeinerung angeben

$$\eta_{T_2} = c\ h_{T_2}^{\beta} = c\ 4^{-\beta} h_{T_0}^{\beta} = 4^{-\beta}\eta_{T_0}$$

Die Verfeinerung wird tatsächlich vorgenommen, wenn der prognostizierte Wert folgende Bedingung erfüllt

$$\eta_{T_2} \ll \eta_{T_1}$$

bzw. der Faktor β einen Wert nahe Null animmt.

6.2 Verfeinerung

Nach der Markierung stellt sich die Frage nach der Durchführung der Unterteilung der entsprechenden Elemente. Dabei ist zu beachten, daß das Gitter geometrisch regulär sein soll, d.h. der kleinste Winkel von Null weg beschränkt ist, und daß das entstehende Gitter zuverlässig sein soll, d.h. es sind keine sogenannten hängenden Knoten erlaubt. Diese Knoten grenzen an Dreiecke, die nicht entsprechend unterteilt wurden.

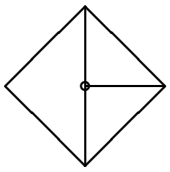


Abbildung 6.1: Hängender Knoten

An diesen Punkten werden die Kompatibilitätsbedingungen für FE-Diskretisierungen verletzt. Dem kann man entgehen, indem man entweder die Unbekannten an diesen Knoten als unechte Freiheitsgrade einführt und ihre Werte als passende Interpolation der Unbekannten an den Nachbarpunkten annimmt oder man unterteilt weitere Elemente derart, daß die hängenden Knoten auflöst werden und die Regularitätsanforderung an das Netz erfüllt werden.

Bei Dreieckselementen gibt es 3 Möglichkeiten zur Unterteilung, wobei die neuentstehenden Knoten sich in den Kantenmittelpunkten befinden. Die reguläre Unterteilung, auch als *rote* Unterteilung bezeichnet, verbindet die drei Kantenmittelpunkte und erzeugt so vier Dreiecke in der nächsten Generation. Dieses Vorgehen führt auf geometrisch ähnliche Dreiecke und das Verhältnis h_T/ρ_T ändert sich nicht, die Winkelbedingung wird also erfüllt.
Bei der Bisektion, auch *grüne* Unterteilung, wird ein Kantenmittelpunkt mit dem gegenüberliegenden Knoten verbunden. Dadurch bleiben zwei Winkel erhalten und sowohl spitzere als auch stumpfere Winkel werden eingeführt.
Die *blaue* Unterteilung verfährt wie die grüne, es wird aber zusätzlich eins der neu entstandenen Dreiecke nochmals grün unterteilt, indem der beim ersten Unterteilen benutzte Kantenmittelpunkt nun als Knoten mit der gegenüberliegenden Kante verbunden wird. Die blaue Unterteilung ist insbesondere beim Auflösen bzw. Vermeiden von

hängenden Knoten von Bedeutung. Die folgenden Abbildung stellt die verschieden Arten der Dreiecksunterteilung beispielhaft dar.

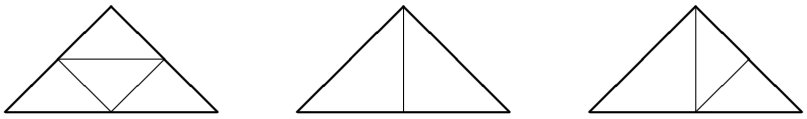
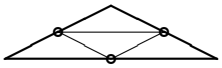


Abbildung 6.2: Rote, güne, blaue Dreiecksunterteilung

Zur Verfeinerung von Triangulierungen bieten sich verschiedene Vorgehensweisen an:

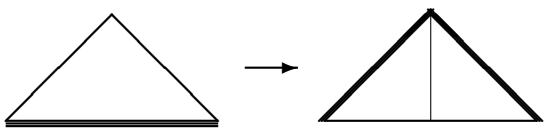
- Reguläre Verfeinerung
  Bei der regulären Unterteilung werden die markierten Elemente nach der roten Art verfeinert. Die eventuell in Nachbarelementen entstehenden hängenden Knoten werden mittels Zusatzregeln aufgelöst:

  1. Ein hängender Knoten 
  2. Zwei hängende Knoten 
  3. Drei hängende Knoten 

  Dabei können in angrenzenden Dreiecken durchaus weitere hängende Knoten entstehen, die wiederum aufgelöst werden müssen.

- Bisektion
  a) Teilung der längsten Kante
  b) Teilung der markierten Kante

  In beiden Fällen hat das Ursprungsdreieck genau eine Kante, die halbiert wird. Folgendes Bild zeigt, wie die Markierung von einem Level auf den nächsten übergeht.

  Für die Behandlung der hängenden Knoten wird nach den obigen Zusatzregeln vorgegangen.

Sowohl die reguläre Unterteilungsstrategie als auch die auf der Bisektion basierende führen mit den Zusatzregeln auf die geforderten zulässigen und geometrisch regulären Triangulierungen.

6.3 Vergröberung

Die lokale Vergröberung von Netzen erfolgt in umgekehrter Weise zur Verfeinerung, d.h. die Vergröberung entspricht dem Zurückgehen im binären Baum, der durch den Prozeß der Verfeinerung entstanden ist. Im Falle der Bisektion verschmelzen bei der Vergröberung also zwei Elemente der aktuellen Generation wieder zu dem Element der vorherigen Generation, aus dem sie hervorgegangen waren.
Um die Vergröberung auf lokale Gebiete zu beschränken, sind einige Vorschriften zu beachten. Im folgenden wird die Vorgehensweise bei der Bisektion betrachtet werden. Zunächst einige Definitionen:

1. Ein Element $T \in \mathcal{T}$ hat das Niveau l, wenn es nach l Verfeinerungsschritten entstand.

2. Ein Element T hat das lokal feinste oder höchste Niveau, wenn alle seine Nachbarn ein geringeres oder gleiches Niveau haben.

3. Sei $T \in \mathcal{T}$ und $T^\star$ das Eltern-Element von T. Der Knoten, der zur Bisektion von $T^\star$ eingefügt wurde wird "Vergröberungsknoten" von T genannt.

4. Sei K eine Kante der Triangulierung $\mathcal{T}$ und $K^\star$ die zugehörige Eltern-Kante mit dem Mittelpunkt Q. Setze $G := \{T \in \mathcal{T} : T \cap K^\star \neq \emptyset\}$. Wenn Q der Vergröberungsknoten aller $T \in G$ ist, dann heißt G "auflösbares Patch".

Die Abbildung 6.3 zeigt ein Beispiel für ein "auflösbares Patch".

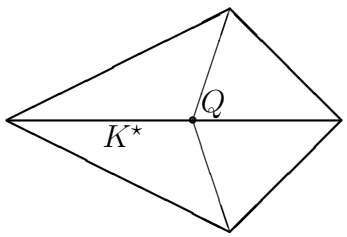


Abbildung 6.3: Auflösbares Patch

Man sieht leicht, daß, wenn G ein auflösbares Patch ist, alle $T \in G$ vergröbert werden können, ohne daß davon andere Elemente der Triangulierung außerhalb von G beeinflußt werden. Auflösbare Patches sind also die Konfiguationen, die vergröbert werden dürfen, da damit garantiert ist, daß die Vergröberung lokal begrenzt bleibt.
Die Auskunft über die Notwendigkeit der Vergröberung bzw. auch Verfeinerung einzelner Element liefert die Markierungsstrategie 4. Nach der Markierung der Elemente wird die Zugehörigkeit der zuvergröbernden Elemente zu einem auflösbaren Patch geprüft. Diese Patches dürfen dann aufgelöst werden, wenn

1. kein Element T im Patch enthalten ist, das verfeinert werden soll,

2. mindestens ein Element T vergröbert werden soll.

Im instationären Prozeß wird nur die Triangulierung des aktuellen Zeitschrittes gespeichert und die h-Methode auf jeden Zeitschritt angewendet. Damit gestaltet sich die Markierungsstrategie 4 mit gegebenem ToL im k-ten Zeitschritt wie folgt:

$$\text{Wähle} \qquad 0 < \theta_2 < \theta_1 \leq 1 \ :$$

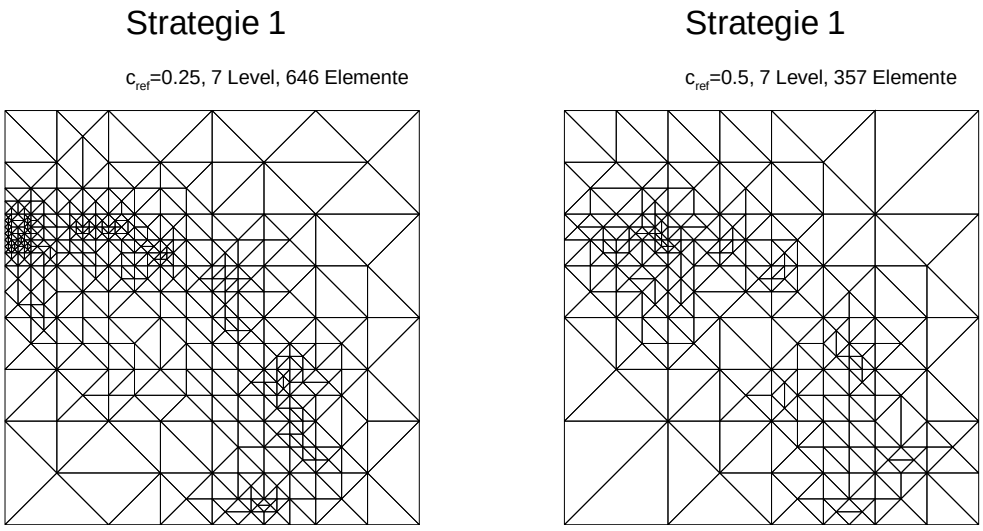
$$\eta_T \geq \frac{\theta_1 \ ToL}{\sqrt{N_k}} \longrightarrow \text{verfeinere}$$

$$\eta_T \leq \frac{\theta_2 \ ToL}{\sqrt{N_k}} \longrightarrow \text{vergröbere}$$

6.4 Beispiele

6.4.1 Gekrümmte innere Schockfront

Dieses Beispiel entspricht dem in Kapitel 3 und Kapitel 5 beschriebenen ($\epsilon = 10^{-5}$). Zur Berechnung wurde wiederum die SDFEM ohne Shock-Capturing verwendet. Die Abschätzung des Fehlers in den Elementen erfolgte mit dem im letzten Kapitel dargestellten residuellen Fehlerschätzer. Die folgende Abbildung 6.4 zeigt die verfeinerten Netze nach Anwendung verschiedener Markierungsmethoden und regulärer Verfeinerung. Die ersten beiden Bilder zeigen Ergebnisse der 1.Markierungsstrategie, links mit einer Verfeinerungsschwelle $c_{ref} = 0.25$ und rechts mit $c_{ref} = 0.5$. Das dritte Bild ist das Resultat der 2.Strategie, bei der die Werte sortiert und die obersten 15 Prozent aus der Betrachtung entfernt wurden. Die folgenden Bilder demonstrieren das Verhalten der 3.Strategie, links wurde die Verfeinerung von mindestens 25 Prozent der Elemente gefordert und rechts 40 Prozent. Die letzten Bilder zeigen schließlich Netze aus der Verwendung der 4.Markierunsstrategie. In beiden Fällen wurden $\theta_1 = 0.5$ und $\theta_2 = 0$ gewählt, links wurde eine Toleranz $ToL = 0.015$ und rechts $ToL = 0.05$ vorgesehen.



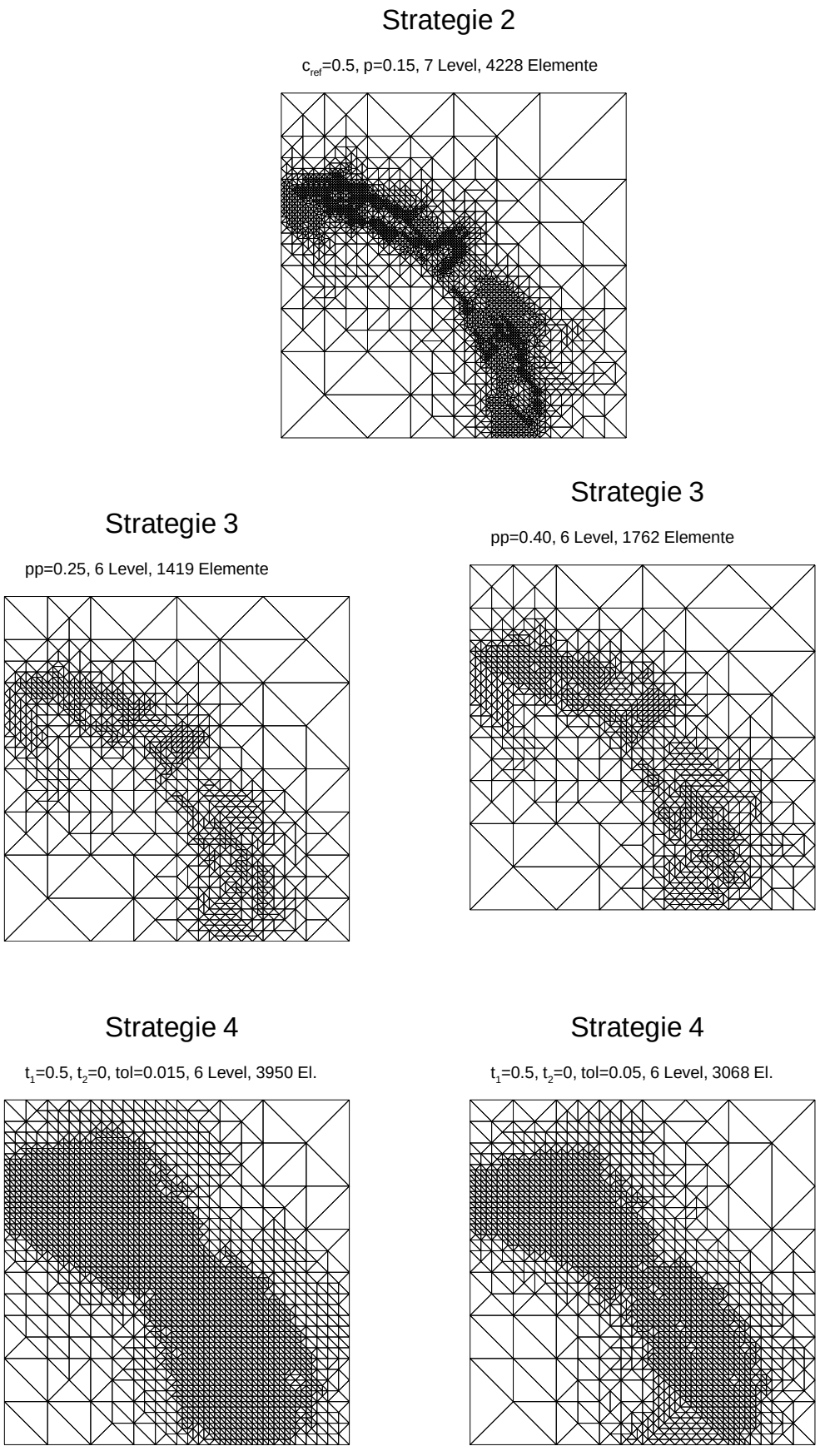


Abbildung 6.4: Verschiedene Markierungsstrategien

Die zwei ersten Netzen der Strategie 1 zeigen die Schwierigkeit der Wahl einer passenden Toleranzschwelle c_{ref} ohne Vorabkenntnisse. Außerdem verdeutlichen sie den Einfluß des Maximalwertes. Diese Schwierigkeit wird durch die Strategie 2 umgangen. Das Resultat ist ein gut an den Verlauf der Schockfront angepaßtes Netz. Die Netze der Strategie 3 liefern abhängig von dem gewählten Prozentsatz zu verfeinernder Elemente dichtere oder weniger dichte Netze. Gleichmaßen beinflußt die Wahl der Toleranz bzw. der Parameter θ_1 und θ_2 in Strategie 4 die Dichte der Netze. Auffallend ist hier, daß die Netze sehr homogen im Bereich der Schockfront verdichtet werden.

6.4.2 Dirichlet Randbedingungen mit Sprung

Dieses Beispiel ist aus Kapitel 5 bekannt, es handelt sich um ein Konvektions-Diffusionsproblem auf dem Einheitsquadrat mit konstantem Advektionsvektor $\boldsymbol{b} = (1, 0)$, ohne Adsorption und Senken bzw. Quellen. Der Diffusionskoeffizient sei $\epsilon = 10^{-3}$. Bezeichnend für dieses Beispiel sind Sprünge in den Dirichlet Randbedingungen auf dem Einflußrand. Die Diskretisierung erfolgte mit der SDFEM. Zur Fehlerabschätzeung diente der residuelle Fehlerschätzer. Die Markierung wurde mittels der Strategie 2 ($c_{ref} = 0.5$, $p = 0.15$) vorgenommen. Anhand dieses Beispiels soll die unterschiedliche Wirkung von regulärer Verfeinerung und Bisektion betrachtet werden.
Die folgende Abbildung 6.5 zeigt die verfeinerten Netze aus regulärer Verfeinerung und Bisektion nach 6 bzw. nach 12 Verfeinerungsstufen.

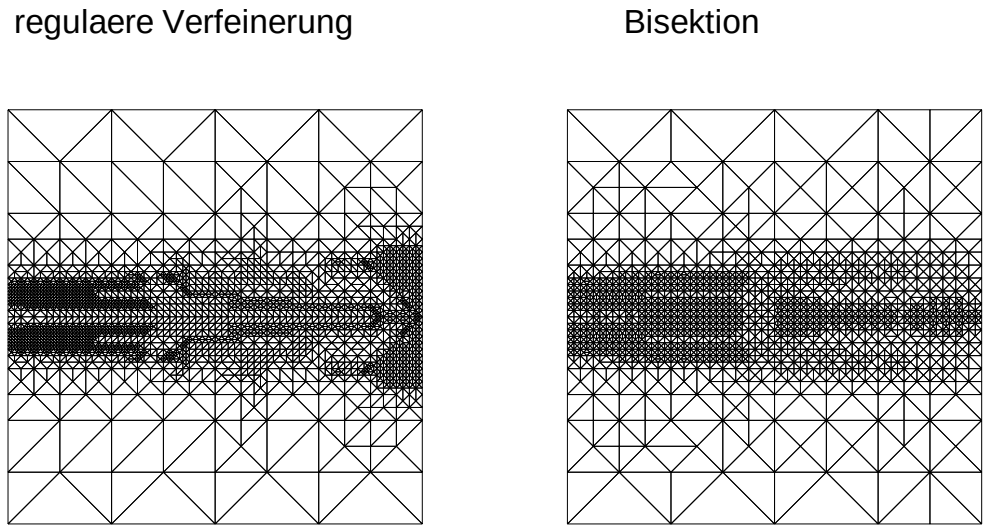


Abbildung 6.5: Reguläre Verfeinerung, Bisektion

Sowohl bei der regulären Verfeinerung als auch bei der Bisektion werden Elemente an den Sprüngen am Einflußrandes konzentriert. Im Bereich des Ausflußrandes wird allerdings deutlich stärker durch die reguläre Verfeinerung verdichtet.
In Abbildung 6.6 sind die zugehörigen numerischen Lösungen dreidimensional dargestellt.

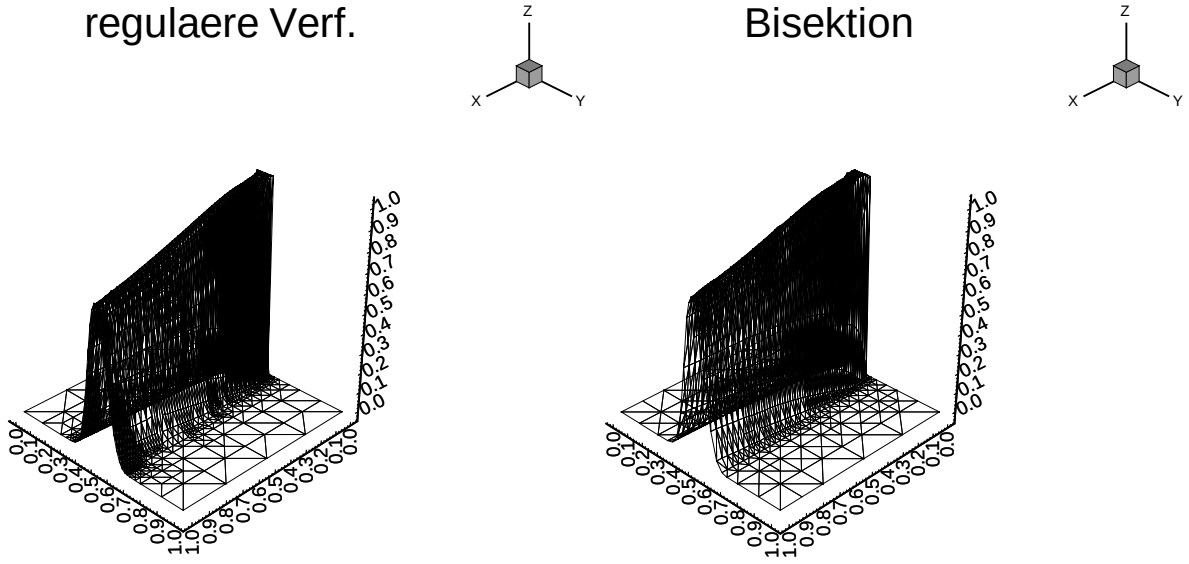


Abbildung 6.6: Reguläre Verfeinerung, Bisektion

Level	N_{el}	N_{free}	e	$\bar{e}$	η	eff
1	32	25	0.055034	14.95694 %	0.06009265176499	1.0919
2	92	57	0.054748	19.36774 %	0.237144515991	4.3315
3	274	151	0.079797	30.74938 %	0.249538499604	3.1272
4	588	313	0.024018	8.28194 %	0.267043616952	11.119
5	1626	839	0.037185	12.56145 %	0.183247890123	4.9281
6	4040	2055	0.013908	4.64835 %	0.130328602534	9.3706

Tabelle 6.2: Reguläre Verfeinerung

In den Tabellen 6.2 und 6.3 sind die Anzahl der Elemente, die Anzahl der Unbekannten, der Fehler e, der relative Fehler $\bar{e}$, der geschätze Fehler η und der Effektivitätsindex für den Fall der regulären Verfeinerung und der Bisektion angegeben.
Hinsichtlich der Genauigkeit liefern beide Vorgehensweisen ungefähr gleich gute Ergebnisse (siehe Level 11 bei Bisektion und Level 5 bei regulärer Verfeinerung). Allerdings sind bei der Verfeinerung nach regulärer Art wesentlich weniger Verdichtungsstufen nötig, um eine gewisse Zahl von Unbekannten zu erhalten.

Level	N_{el}	N_{free}	e	$\bar{e}$	η	eff
1	16	13	0.24221	59.06570 %	0.025936906484	0.1071
2	22	17	0.24153	58.63605 %	0.024862696531	0.1029
3	50	33	0.17553	56.84863 %	0.276358038725	1.5744
4	76	49	0.19347	64.30006 %	0.059658185103	0.3084
5	130	77	0.21029	72.07293 %	0.060409568355	0.2873
6	242	135	0.12612	47.23609 %	0.423651819427	3.3592
7	296	164	0.07875	30.06099 %	0.269092425983	3.4173
8	436	235	0.11061	39.34059 %	0.222167886902	2.0086
9	624	329	0.11121	38.65626 %	0.167542167653	1.5065
10	974	507	0.04138	14.09040 %	0.261820023961	6.3275
11	1732	890	0.03516	11.89259 %	0.193053747642	5.4903
12	3076	1565	0.05032	16.83742 %	0.133209850452	2.6472

Tabelle 6.3: Bisektion

6.5 Zusammenfassung

Anhand der gezeigten Beispiele ist der enorme Einfluß der Auswahl der Markierungsstrategie auf die Gestalt des Netzes offensichtlich. Bei den Strategien 1 und 4 tritt die Problematik der Wahl einer passenden Verfeinerungsschwelle bzw. Toleranz zutage. Die Wahl erfordert entweder viel Fingerspitzengefühl oder einige Testrechnungen. Bei Strategie 3 bedarf die Wahl des mindest zu verfeinernden Prozentsatzes von Elementen ebenso großer Sorgfalt. Wird dieser zu groß gewählt, werden zu viele Elemente verfeinert. Als besonders praktikabel erwies sich Strategie 2, da mit ihr ohne viele Vorabrechnungen sehr gute Ergebnisse erreicht wurden.
Das zweite Beispiel zeigt den Unterschied zwischen Bisektion und regulärer Verfeinerung. In Hinblick auf die Netzverdichtung wird mit der regulären Verfeinerung zusätzlich zum Einflußgebiet auch stark im Bereich des Ausflußrandes verdichtet. Bei der Verfeinerung nach regulärer Art sind aufgrund der Teilung eines Elementes in vier Elemente wesentlich weniger Verdichtungsstufen nötig, um eine gewisse Zahl von Unbekannten zu erhalten. Beide Methoden liefern bezüglich der Genauigkeit vergleichbar gute Ergebnisse.

Kapitel 7

Netzglättung

In diesem Kapitel soll ein Algorithmus zur Optimierung der Plazierung der Knoten einer Triangulierung nach [2] vorgestellt werden. Dabei soll eine Triangulierung $\mathcal{T}_h$, die durch lokale Netzverfeinerung gewonnen wurde, mit fester Konnektivitätsstruktur der Knoten gegeben sein. Desweiteren seien eine lokale Abschätzung des Fehlers und die zugehörige Finite Element Lösung u_h bezüglich dieser Triangulierung bekannt. Der folgende Algorithmus bestimmt die Position, d.h. die Koordinaten der Knoten so, daß der Fehler minimiert wird. Es wird also eine Triangulierung $\bar{\mathcal{T}}_h$ mit derselben Topologie wie $\mathcal{T}_h$ ermittelt, so daß der Fehler in der Lösung $\bar{u}_h$ auf dem geglätteten Netz minimiert wird. Der vorgestellte Glättungsalgorithmus ist nicht auf die Anwendung eines speziellen Fehlerschätzers beschränkt. Im Falle der Verwendung von linearen Elementen, wird lediglich gefordert, daß der Fehlerschätzer eine lokale Aussage über den Fehler in Form eines stückweise quadratischen Polynoms gibt. Die grundlegende Strategie ist, die so gewinnbaren lokalen Abschätzungen der 2.Ableitungen der exakten Lösung u als konstant anzusetzen und lokale Minimierungsprobleme für die Plazierung der Netzknoten zu lösen. Sollte diese Methode auf stückweise Polynome der Grades p verallgemeinert werden, so würden Fehlerschätzer, die auf stückweise Polynome des Grades $p+1$ führen, benötigt werden.
Bemerkenswert an dieser Methode ist, daß darüberhinaus keine Details der Berechnungen der Differentialgleichung oder der Fehlerschätzer benötigt werden. Es werden lediglich Ergebnisse vorangegangener Berechnungen benutzt. Damit ist diese Methode vielseitig anwendbar.
Im Rahmen von Netzadaption ist diese Methode nicht als grundlegende oder universelle Methode zu verstehen, da sie nicht in die Topologie bzw. Konnektivität der Triangulierung eingreift. Dies ist aber in vielen Fällen, wie in den vorhergehenden Kapiteln gezeigt, entscheidend für den Erfolg einer Netzanpassung. Vielmehr stellt die Netzglättung in diesem Zusammenhang eine sinnvolle Ergänzung dar. Das in Kapitel 6 gezeigte Vorgehen der Netzverfeinerung basierend auf der regulären Verfeinerung eines Elements in 4 gleichartige Elemente kann zwar gut Schockfronten und dergleichen auflösen, aber die exakte Postion der Knoten hängt von der Geometrie des Elementes der letzten Generation, und damit von der Struktur der Ausgangstriangulierung ab. Im allgemeinen resultiert daraus kein Problem, aber in Fällen von z.B. sehr scharfen Schockfronten kann schon eine geringe Verschiebung von Knoten an der Front bzw. die Ausrichtung des Netzes dort den Fehler deutlich reduzieren. Daher sollte man zunächst Methoden zur Verfeinerung bzw. Vergröberung wählen, um ein Netz mit ausreichender Dichte der

Knoten und sinnvoller Topologie zu gewinnen, und anschließend zur 'Feineinstellung' eine Netzglättung nachschieben. Die Knotenpunkte werden i.a. nicht sehr weit gerückt und schon einige wenige Iterationen können mit geringen Veränderungen einen drastischen Effekt liefern [2].

7.1 Bemerkungen zur Geometrie

Mittels des Flächeninhaltes A und der Kantenvektoren $\boldsymbol{l}_i$ (siehe Anhang 1) kann die Qualität eines Dreieckes T, d.h. seine Regelmäßigkeit in Form eines Quotienten

$$q(T) = \frac{4\sqrt{3}A}{|\boldsymbol{l}_1|^2 + |\boldsymbol{l}_2|^2 + |\boldsymbol{l}_3|^2}$$

angegeben werden. Für gleichseitige Dreiecke nimmt der Quotient den Wert 1 an. Bei Dreiecke mit sehr kleinen Winkeln strebt $q(T)$ gegen 0.
Der Flächeninhalt A berechnet sich bei Knotennumerierung im mathematisch positiven Sinne wie folgt:

$$2A = (x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1).$$

Ändert sich die Orientierung der Numerierung, z.B. im Verlaufe der Knotenoptimierung, so erhält die Fläche ein negatives Vorzeichen und kennzeichnet so die Umorientierung.

Im Prozeß der Netzglättung ist es wichtig zu wissen, welche Bewegungsmöglichkeiten oder Freiheitsgrade jeder einzelne Knoten besitzt. Dazu werden die Knoten in 3 Gruppen aufgeteilt, jene mit 0, 1 oder 2 Freiheitsgraden. Eckknoten besitzen keinen Freiheitsgrad, d.h. sie dürfen nicht verschoben werden, da die zugrundeliegende Geometrie des Gebietes davon beeinflußt werden würde. Randknoten bestimmen ebenso die äußere Gestalt des Gebietes, aber sie dürfen entlang der vorgegebenen Gebietsbegrenzung bewegt werden, sie besitzen somit einen Freiheitsgrad. Alle anderen Knoten sind sogenannte innere Knoten, die in alle Richtungen bewegt werden können und deshalb mit 2 Freiheitsgraden charakterisiert werden.

7.2 Netzglättung basierend auf a posteriori Fehlerschätzern

Es soll eine Methode zur Knotenplazierung dargestellt werden, die auf auf der näherungsweisen Minimierung von

$$\min_{\mathcal{T}\in\mathcal{F}} \int_\Omega |\nabla(u - u_h)|^2 \, \mathrm{d}x \tag{7.1}$$

basiert. Dabei ist $\mathcal{F}$ eine Familie von Triangulierungen von Ω mit gleicher Konnektivität. Für die einzelnen Mitglieder von $\mathcal{F}$ gelten die gleichen Bedingungen an die äußere Begrenzung des Gebietes. Sie unterscheiden sich untereinander nur durch die Position der Knoten, aber nicht durch die Struktur der Knotenzusammenhänge. Der Fehler $u - u_h$ wird durch einen a posteriori Fehlerschätzer wie in Kapitel 5 vorgestellt approximiert.

Mit $\mathbf{M}_T$ sei die 2×2 Matrix der 2. Ableitungen von u auf T gegeben

$$\mathbf{M}_T = -\frac{1}{2} \begin{bmatrix} u_{xx} & u_{xy} \\ u_{xy} & u_{yy} \end{bmatrix}.$$

Es wird angenommen, daß die Matrix $\mathbf{M}_T$ auf T konstant sei. Mit $\bar{b}_i$ seien bubble-Funktionen auf den Kanten bezeichnet, die von denen der Standarddefinition durch einen Faktor 4 abweichen

$$\bar{b}_i(x, y) = \lambda_{i+1}(x, y)\lambda_{i+2}(x, y).$$

Auf einem Element T hat der a posteriori Fehlerindikator die Form

$$e_T = e_1\bar{b}_1(x, y) + e_2\bar{b}_2(x, y) + e_3\bar{b}_3(x, y),$$

mit den Werten e_i an den Knoten. In bezug auf den Glättungsalgorithmus ist eine Darstellung als

$$e_T = \boldsymbol{l}_1^t\mathbf{M}_T\boldsymbol{l}_1\bar{b}_1 + \boldsymbol{l}_2^t\mathbf{M}_T\boldsymbol{l}_2\bar{b}_2 + \boldsymbol{l}_3^t\mathbf{M}_T\boldsymbol{l}_3\bar{b}_3$$

sinnvoll. Setzt man die Koeffizienten der letzten beiden Gleichung gleich und nutzt man die Definition von $\mathbf{M}_T$, so erhält man ein Gleichungssystem mit 3 Unbekannten

$$\begin{bmatrix} (x_3 - x_2)^2 & 2(x_3 - x_2)(y_3 - y_2) & (y_3 - y_2)^2 \\ (x_1 - x_3)^2 & 2(x_1 - x_3)(y_1 - y_3) & (y_1 - y_3)^2 \\ (x_2 - x_1)^2 & 2(x_2 - x_1)(y_2 - y_1) & (y_2 - y_1)^2 \end{bmatrix} \begin{bmatrix} u_{xx} \\ u_{xy} \\ u_{yy} \end{bmatrix} = -2 \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}$$

zur Bestimmung der Elemente von $\mathbf{M}_T$. Dieses Gleichungssystem kann leicht gelöst werden, sofern das Dreieck T nicht degeneriert ist, d.h. $q(T) \neq 0$ ist. Die Matrix $\mathbf{M}_T$ wird als konstant angesehen und daher während der Netzglättung nur einmal berechnet.
Es ist an dieser Stelle nicht unbedingt sofort ersichtlich, warum die Matrix $\mathbf{M}_T$ überhaupt berechnet wird. Wie später zu sehen sein wird, geht der Ausdruck $\boldsymbol{l}_j^t\mathbf{M}_T\boldsymbol{l}_j$ in das zu minimierende Funktional ein. Diese Größe ist über die e_j des Fehlerschätzers bekannt. Würde man die e_j direkt nutzen wollen, so wäre man gezwungen, diese für jedes verformte Element bereitzustellen, d.h. aktuell den Fehler auf dem entsprechenden Element abzuschätzen. In diesem Sinne fungiert $\boldsymbol{l}_j^t\mathbf{M}_T\boldsymbol{l}_j$ mit der aus den e_j berechneten Matrix $\mathbf{M}_T$ als kostengünstiges Modell für die T-Abhängigkeit des Fehlerschätzers.
Es läßt sich zeigen, daß

$$\int_T |\nabla(u - u_h)|^2 \,\mathrm{d}x = \boldsymbol{s}^t\mathbf{B}\boldsymbol{s},$$

wobei

$$\boldsymbol{s} = \begin{bmatrix} \boldsymbol{l}_1^t\mathbf{M}_T\boldsymbol{l}_1 \\ \boldsymbol{l}_2^t\mathbf{M}_T\boldsymbol{l}_2 \\ \boldsymbol{l}_3^t\mathbf{M}_T\boldsymbol{l}_3 \end{bmatrix} \qquad \text{und} \qquad B_{ij} = \int_T \nabla\bar{b}_i \cdot \nabla\bar{b}_j \,\mathrm{d}x.$$

Durch direktes Ausrechnen erhält man

$$\mathbf{B} = \frac{1}{48A} \begin{bmatrix} \boldsymbol{l}_1^t\boldsymbol{l}_1 + \boldsymbol{l}_2^t\boldsymbol{l}_2 + \boldsymbol{l}_3^t\boldsymbol{l}_3 & 2\boldsymbol{l}_1^t\boldsymbol{l}_2 & 2\boldsymbol{l}_1^t\boldsymbol{l}_3 \\ 2\boldsymbol{l}_2^t\boldsymbol{l}_1 & \boldsymbol{l}_1^t\boldsymbol{l}_1 + \boldsymbol{l}_2^t\boldsymbol{l}_2 + \boldsymbol{l}_3^t\boldsymbol{l}_3 & 2\boldsymbol{l}_2^t\boldsymbol{l}_3 \\ 2\boldsymbol{l}_3^t\boldsymbol{l}_1 & 2\boldsymbol{l}_3^t\boldsymbol{l}_2 & \boldsymbol{l}_1^t\boldsymbol{l}_1 + \boldsymbol{l}_2^t\boldsymbol{l}_2 + \boldsymbol{l}_3^t\boldsymbol{l}_3 \end{bmatrix}.$$

Die Matrix $\mathbf{B}$ ist positiv definit und spektral äquivalent zu ihrer Diagonalen. Die Diagonaleinträge von $\mathbf{B}$ sind konstant und können durch $1/(4\sqrt{3}q(T))$ ausgedrückt werden. Mit dieser Näherung ergibt sich

$$\int_\Omega |\nabla(u-u_h)|^2 \,\mathrm{d}x \approx \frac{(\boldsymbol{l}_1^t \mathbf{M}_T \boldsymbol{l}_1)^2 + (\boldsymbol{l}_2^t \mathbf{M}_T \boldsymbol{l}_2)^2 + (\boldsymbol{l}_3^t \mathbf{M}_T \boldsymbol{l}_3)^2}{4\sqrt{3}q(T)}.$$

Dabei fungiert die Qualitätsfunktion $q(T)$ als natürliche Schranke, die verhindert, daß das Dreieck T degeneriert, wenn Knoten der Triangulierung bewegt werden. Abkürzend wird für den Zähler die Funktion

$$s(T) = (\boldsymbol{l}_1^t \mathbf{M}_T \boldsymbol{l}_1)^2 + (\boldsymbol{l}_2^t \mathbf{M}_T \boldsymbol{l}_2)^2 + (\boldsymbol{l}_3^t \mathbf{M}_T \boldsymbol{l}_3)^2$$

definiert. Sie enthält Informationen über die Größe und Richtung der zweiten Ableitungen.

In dem Glättungsalgorithmus wird zur Lösung von (7.1) eine Gauß-Seidel ähnliche Iteration angewendet, d.h. es wird über alle Knoten iteriert, indem die lokale Plazierung eines Knotens optimiert wird, während alle anderen Knoten festgehalten werden. Für jeden Knoten der Triangulierung muß also ein lokales Optimierungsproblem gelöst werden. Die Knoten werden wie oben beschrieben nach ihren Freiheitsgraden unterschieden. Es seien zunächst die inneren Knoten betrachtet, für die Randknoten ergibt sich ein analoger Algorithmus mit entsprechenden einschränkenden Bedingungen.
Das lokale Problem für den Knoten $\boldsymbol{x}_i = (x_i, y_i)$ lautet somit

$$\min_{x_i, y_i} \sum_{T \in \Omega_i} \frac{s(T)}{q(T)}.$$

Dabei bezeichnet Ω_i die Menge der Elemente, die den Knoten $\boldsymbol{x}_i$ enthalten, in der Notation von Kapitel 5 entspricht dies $\omega_{\boldsymbol{x}_i}$. Da angenommen wurde, daß $\mathbf{M}_T$ konstant sei, ist $s(T)$ ein quadratisches Polynom von x_i und y_i. Und $q(T)$ ist eine rationale Funktion von quadratischen Polynomen. Das lokale Minimierungsproblem kann mit einem gedämpften Newton Verfahren gelöst werden. Im Anhang 2 sind die Zusammenhänge zur Lösung der lokalen Minimierungsprobleme aufgeführt.

7.2.1 Neumann-Fehlerschätzer aus Kapitel 5

In Kapitel 5 wurde ein Fehlerschätzer basierend auf einem lokalen Neumann-Problem, zu dessen Lösung eine weitere bubble-Funktion im Inneren des Elements verwendet wurde, vorgestellt. Um diese zusätzliche Information im Netzglättungsalgorithmus nutzbar zu machen, bietet sich eine Fehlerquadrat-Minimierung für die Berechnung der Matrix $\mathbf{M}_T$ an.
Für den oben zugrundegelegten Fehlerschätzer nach Bank gilt mit den bubble-Funktionen $\bar{b}_i, i = 1, 2, 3$

$$e_T = e_1 \bar{b}_1(x, y) + e_2 \bar{b}_2(x, y) + e_3 \bar{b}_3(x, y).$$

Der Fehlerschätzer aus Kapitel 5 sei hier entsprechend mit e_T^N bezeichnet. Er läßt sich mit den bubble-Funktionen $b_i, i = 1, 2, 3, 4$ als

$$e_T^N = e_1^N b_1(x, y) + e_2^N b_2(x, y) + e_3^N b_3(x, y) + e_4^N b_4(x, y)$$

darstellen. Die bubble-Funktion b_4 entspricht dabei derjenigen, welche ihren Maximalwert im Schwerpunkt des Dreieck annimmt. Damit gestaltet sich die Fehlerquadrat-Minimierung bezüglich $\mathbf{M}_T$ wie folgt

$$\int_T \frac{1}{2} \left(e_T - e_T^N\right)^2 \, \mathrm{d}A \;\rightarrow\; \mathrm{MIN}$$

bzw.

$$\int_T \left(e_T - e_T^N\right) \frac{\partial e_T}{\partial \mathbf{M}_T} \, \mathrm{d}A = 0.$$

Um daraus einen leicht berechenbaren Ausdruck zur Bestimmung der Matrix $\mathbf{M}_T$ zu entwickeln, wird von der Beziehung zwischen e_T und $\mathbf{M}_T$ ausgegangen

$$e_T = s_1 \bar{b}_1 + s_2 \bar{b}_2 + s_3 \bar{b}_3,$$

dabei steht $s_i, i = 1, 2, 3$ abkürzend für

$$s_i = \boldsymbol{l}_i^t \mathbf{M}_T \boldsymbol{l}_i = [\boldsymbol{l}_i \otimes \boldsymbol{l}_i] : \mathbf{M}_T.$$

Mit der Anordnung der Elemente der Matrix $\mathbf{M}_T$ in einem Vektor $\boldsymbol{m}$ und der Definition eines Vektors $\boldsymbol{L}_i$ zur Darstellung des dyadischen Produkts

$$\boldsymbol{m} = \begin{bmatrix} u_{xx} \\ u_{xy} \\ u_{yy} \end{bmatrix}, \qquad \boldsymbol{L}_i = \begin{bmatrix} l_{xi}^2 \\ 2 l_{xi} l_{yi} \\ l_{yi}^2 \end{bmatrix}$$

läßt sich s_i als

$$s_i = \boldsymbol{L}_i \cdot \boldsymbol{m}$$

und damit e_T als

$$e_T = \left(\bar{b}_1 \boldsymbol{L}_1 + \bar{b}_2 \boldsymbol{L}_2 + \bar{b}_3 \boldsymbol{L}_3\right) \cdot \boldsymbol{m} =: \boldsymbol{\Lambda} \cdot \boldsymbol{m}$$

darstellen. Mit diesen Bezeichungen folgt für die Stationaritätsbedingung, wobei $\partial e_T / \partial \mathbf{M}_T$ in $\partial e_T / \partial \boldsymbol{m} = \boldsymbol{\Lambda}$ übergeht,

$$\begin{aligned} \int_T e_T \left(\bar{b}_1 \boldsymbol{L}_1 + \bar{b}_2 \boldsymbol{L}_2 + \bar{b}_3 \boldsymbol{L}_3\right) \mathrm{d}A &= \int_T e_T^N \left(\bar{b}_1 \boldsymbol{L}_1 + \bar{b}_2 \boldsymbol{L}_2 + \bar{b}_3 \boldsymbol{L}_3\right) \mathrm{d}A \\ \int_T e_T \boldsymbol{\Lambda} \, \mathrm{d}A &= \int_T e_T^N \boldsymbol{\Lambda} \, \mathrm{d}A \\ \int_T (\boldsymbol{\Lambda} \cdot \boldsymbol{m}) \boldsymbol{\Lambda} \, \mathrm{d}A &= \int_T e_T^N \boldsymbol{\Lambda} \, \mathrm{d}A \end{aligned}$$

und schließlich der sehr übersichtliche Zusammenhang

$$\int_T \boldsymbol{\Lambda} \otimes \boldsymbol{\Lambda} \, \mathrm{d}A \cdot \boldsymbol{m} = \int_T e_T^N \boldsymbol{\Lambda} \, \mathrm{d}A.$$

Damit ist wiederum ein 3×3 Gleichungssystem zur Bestimmung von $\mathbf{M}_T$ zu lösen. Die Integration bedeutet keinen großen zusätzlichen Aufwand, da analytische Lösungen zur Verfügung stehen.

7.2.2 Andere Fehlerschätzer aus Kapitel 5

Bei einer Netzglättung ausgehend von Informationen eines residuellen Fehlerschätzers können die Werte $e_i, i = 1, 2, 3$ gleich, d.h. entsprechend dem Wert des auf dem Dreieck ermittelten Wert des Schätzers $\eta_{R,T}$ gewählt. Konzeptionell lassen sich die Werte aus dem Z^2-Indikator ebenso verwerten. Da die Motivation des Glättungsalgorithmus durch den Neumann Fehlerschätzer gegeben wurde, soll ein Beispiel unter seiner Verwendung gezeigt werden.

7.3 Beispiel

Die Wirkungsweise des Netzglättungsalgorithmus soll ebenfalls am Beispiel der gekrümmten Schockfront ($\epsilon = 10^{-5}$) veranschaulicht werden. In gewohnter Weise wurden die SDFEM, der Neumann-Fehlerschätzer, Markierungsstrategie 2 ($\gamma = 0.5$, $p = 0.25$) und Bisektion verwendet. Das Ausgangsnetz für die Glättung entstand aus 11 Verfeinerungsstufen (1371 Elemente).

Für die Wurzel der zu minimierende Funktion Υ und die Gesamtverschiebungen innerhalb der iterierten Netze dx

$$\Upsilon = \left(\sum_T \frac{s(T)}{4\sqrt{3}q(T)} \right)^{1/2}, \qquad dx = \sum_{\boldsymbol{x}} |\Delta \boldsymbol{x}|$$

lassen sich folgende Konvergenzverläufe beobachten:

Iteration	Υ	dx
1	24.5936280819	0.234231145989
2	23.2907499181	0.148909227680
3	22.5963720585	0.113596252294
4	22.0852539380	0.092940266308
5	21.6730730728	0.079018547824
6	21.3284694953	0.069091518959
7	21.0347936611	0.061452094610
8	20.7811948677	0.055268132378
9	20.5599194342	0.050099667595
10	20.3651624017	0.045688108570

Tabelle 7.1: Konvergenz von Υ und dx

Die folgende Abbildung 7.1 zeigt das Ausgangsnetz und die Netze nach einer, fünf und zehn Glättungsiterationen für ein Vorgehen mit Fehlerquadratminimierung und einmaliger Bestimmung von $\mathbf{M}_T$ vor der ersten Netzglättung.

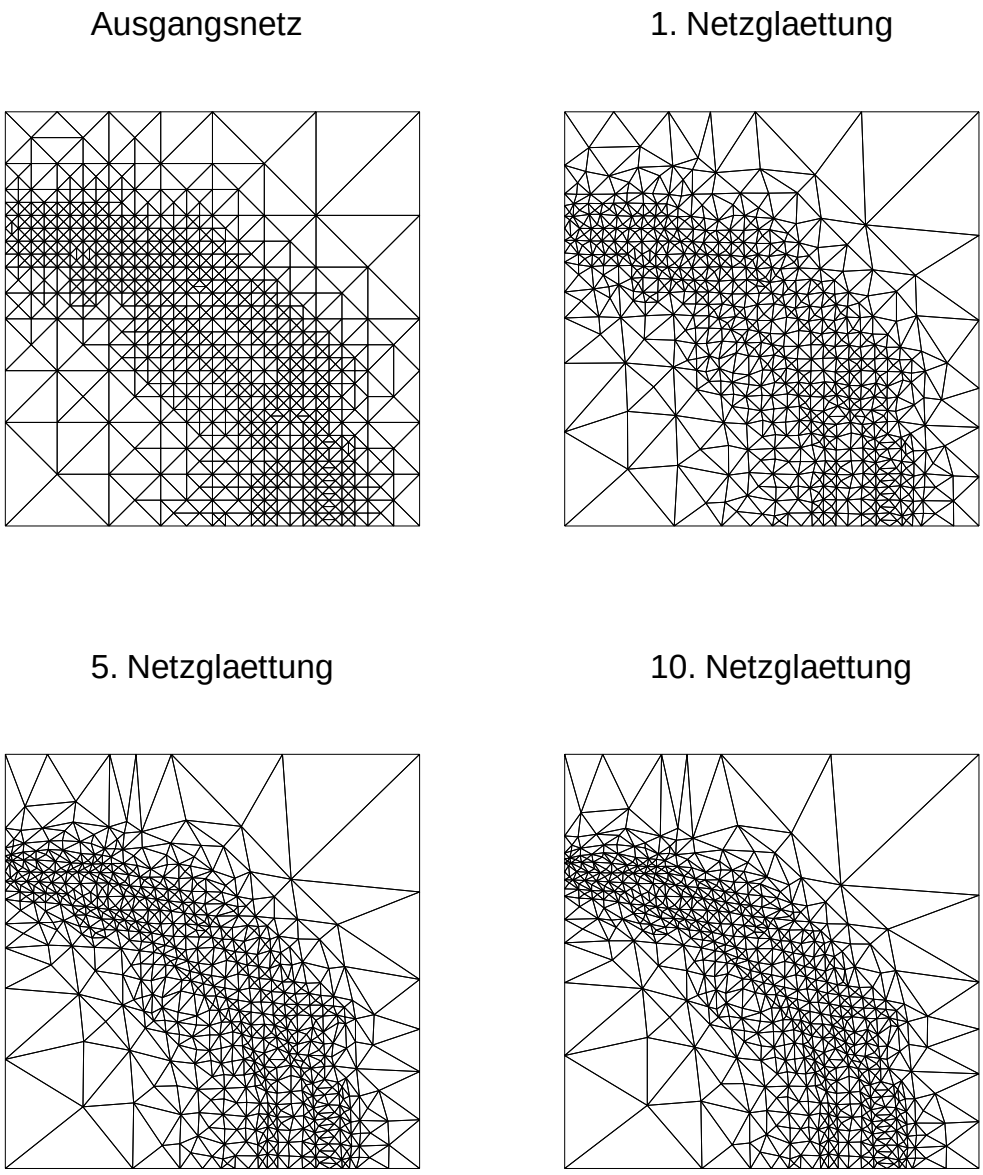


Abbildung 7.1: Ausgangsnetz und geglättete Netze

Um eine Aussage über das Verhalten des Fehlers treffen zu können, wurde nach jeder Glättungsiteration das Problem auf dem modifizierten Netz gelöst und das Ergebnis mit der hyperbolischen Grenzfall-Lösung verglichen. Die Matrix $\mathbf{M}_T$ wurde aber zunächst konstant für alle Iterationen gehalten.
Die Tabelle 7.2 zeigt die Werte des globalen Fehlers e und des relativen Fehlers $\bar{e}$ für den Fall mit Fehlerquadratminimierung (links) und ohne. In der ersten Zeile ist der entsprechende Wert nach der 11.Verfeinerungsstufe angegeben.

Es zeigt sich, daß sich die Gestalt der Netze im Verlauf des Glättungsprozesses stark ändert. Daher ist es mitunter sinnvoll, die Matrizen $\mathbf{M}_T$ nicht während des gesamten Ablaufs konstant zu lassen, da diese lediglich auf den Informationen des Ausgangsnetzes beruhen. Zur Veranschaulichung des Einflusses der Matrizen $\mathbf{M}_T$ auf die geglätteten

Iteration	e	$\bar{e}$	e	$\bar{e}$
0	0.072425	11.74856 %	0.072425	11.74856 %
1	0.070759	11.49128 %	0.071122	11.55016 %
2	0.069702	11.31020 %	0.070629	11.43295 %
3	0.069280	11.24171 %	0.072179	11.68298 %
4	0.068811	11.16159 %	0.070674	11.46509 %
5	0.068373	11.08703 %	0.066251	10.69911 %
6	0.067995	11.02288 %	0.065202	10.54871 %
7	0.068253	11.07007 %	0.068861	11.14960 %
8	0.068004	11.02777 %	0.071872	11.64589 %
9	0.067814	10.99550 %	0.074242	12.02896 %
10	0.069015	11.19619 %	0.075101	12.14632 %

Tabelle 7.2: Fehlerdaten im Glättungsprozeß mit und ohne Fehlerquadratminimierung

Netze wurden sie nach jeder Iteration neu bestimmt. Die Abbildung 7.2 zeigt das Netz nach fünf und zehn Glättungsiterationen.

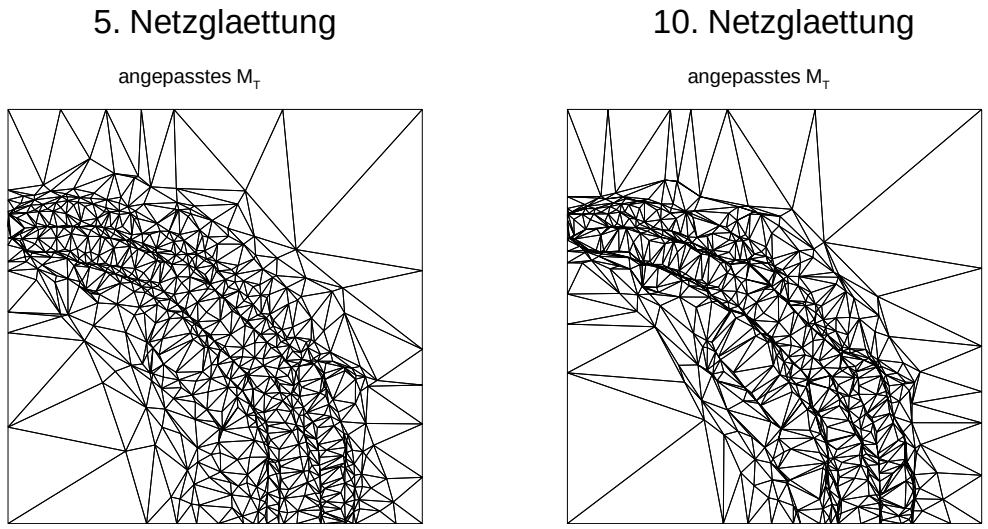


Abbildung 7.2: Geglättetes Netz mit angepaßtem $\mathbf{M}_T$

In Tabelle 7.3 sind die Werte des deutlich verbesserten globalen Fehlers und des relativen Fehlers für dieses Vorgehen und unter Verwendung der Fehlerquadratminimierung angegeben.

Iteration	e	$\bar{e}$
0	0.072425	11.74856 %
1	0.070759	11.49128 %
2	0.069257	11.20801 %
3	0.066803	10.85728 %
4	0.065856	10.74264 %
5	0.063748	10.37783 %
6	0.063684	10.35382 %
7	0.062219	10.16229 %
8	0.060941	9.98131 %
9	0.059989	9.84450 %
10	0.061140	10.05245 %

Tabelle 7.3: Fehlerdaten im Glättungsprozeß bei angepaßtem $\mathbf{M}_T$

Die Abbildung 7.3 zeigt links das Netz nach einer weiteren Verfeinerung im Anschluß an die Netzglättung mit angepaßtem $\mathbf{M}_T$ und rechts das Netz nach einer normalen weiteren Verfeinerung, d.h. nach insgesamt 12 Verfeinerungen, ohne irgendeine Netzglättung. Der globale Fehler auf dem linken Netz mit 2261 Elementen beträgt $e = 0.070984$, der relative Fehler $\bar{e} = 11.54950\%$. Das rechte Netz enthält 2223 Elemente der globale Fehler sank durch die weitere Verfeinerung von $e_{11} = 0.072425$ auf $e_{12} = 0.064563$ und der relative Fehler von $\bar{e}_{11} = 11.74856\%$ auf $\bar{e}_{12} = 10.44160\%$.

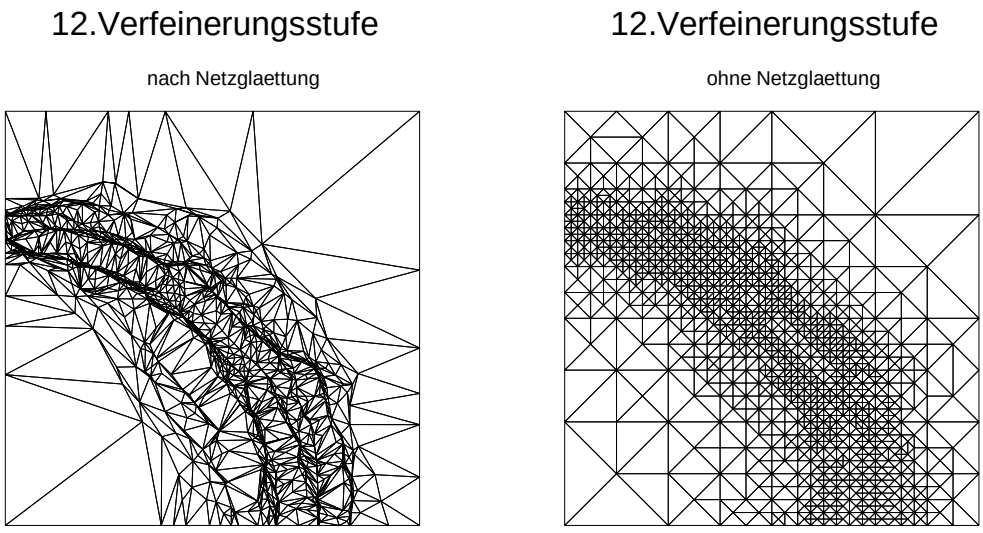


Abbildung 7.3: 12.Verfeinerungsstufe nach Netzglättung bzw. ohne Netzglättung

7.4 Zusammenfassung

Anhand der Beispiele wird deutlich, daß die Netzglättung konvergiert, siehe Υ und dx. Der Fehler der numerischen Lösung nimmt auf den geglätteten Netzen ab. Die Abnahme des Fehlers ist insbesondere stärker, wenn zur Bestimmung der Matrix $\mathbf{M}_T$ eine Fehlerquadratminimierung vorgenommen wird. Der Fehler fällt im Verlauf der Netzglättung zudem schneller, wenn die Matrix $\mathbf{M}_T$ nicht nur zu Beginn der ersten Iteration bestimmt und über alle weiteren Iterationen konstant gehalten wird, sondern wenn nach einigen oder jeder Iteration auf dem geglätteten Netz eine Fehlerabschätzung vorgenommen wird und daraus $\mathbf{M}_T$ neu bestimmt wird. Dieses Vorgehen geht durch die Lösung des Problems auf einem verbesserten Netz einher mit höherem Rechenaufwand. Dennoch kann die Neubestimmung von $\mathbf{M}_T$ nach einigen wenigen Iterationen sinnvoll sein, da zu Beginn der Netzglättung die größten Veränderungen am Netz zu beobachten sind. Die so erzielbaren Verbesserungen sind durchaus mit der weiteren Netzverfeinerung über eine Stufe zu vergleichen. Allerdings sind die Kosten durch die Lösung des Problems auf den geglätteten Netzen und das Aufstellen der $\mathbf{M}_T$ erheblich höher.
In den vorliegenden Beispielen wurde eine Kopplung von Netzglättung und Netzverfeinerung, d.h. ein Einbinden des Glättungsalgorithmus zwischen die Verfeinerungsstufen, nicht gesondert dargestellt. Wenn man z.B. nach der 7.Verfeinerungsstufe eine Netzglättung vornimmt, läßt sich beobachten, daß sich die Lösungen auf den geglätteten Netzen dieser Stufe verbessern. Auf dem anschließend verfeinerten Netz wird die numerische Lösung allerdings verhältnismäßig schlechter. Ebenso wird auch, wie oben dargestellt, bei einer nachgestellten 12.Verfeinerungsstufe nach der Netzglättung die numerische Lösung schlechter. Dies begründet sich aus den zum Teil sehr stark verformten Elementen.
Zur Diskretisierung des Problems wurde in allen Fällen die SDFEM verwendet. Für $\epsilon = 10^{-5}$ sind leichte Über- und Unterschwingungen in der numerischen Lösung vorhanden. Bei allen hier vorgestellten geglätteten Netzen zeigt sich deutlich eine Häufung der Elemente in den Bereichen der auftretenden Über- und Unterschwingungen.
Vor dem Hintergrund dieser Beobachtungen ist die Netzglättung in ihrer Form als nachgeschobene Ergänzung zur Netzverfeinerung als sinnvoll zu bewerten. Zudem bietet die Netzglättung bei Beschränkung des Speicherplatzes, die eine weitere Verfeinerung nicht zulassen würde, eine Möglichkeit zur weiteren Verbesserung der numerischen Lösung.

Kapitel 8

Instationäre Erweiterung

Im vorangegangenen Teil wurde ausschließlich die stationäre Konvektions-Diffusions-Gleichung betrachtet. Das zeitabhängige Problem ist im Gegensatz zum stationären parabolisch, wenn der Koeffizient des Diffusionsterms positiv ist. Abhängig von der Größe des Diffusionskoeffizienten nimmt die Differentialgleichung einen mehr oder weniger hyperbolischen Charakter an. Die meisten FE-Methoden zur Lösung dieses Gleichungstyps beruhen auf einer *Semidiskretisierung*. Die Knotenunbekannten werden als zeitabhängig angenommen und das vom stationären Fall bekannte Vorgehen wird angewandt. Dies führt auf ein System von Differentialgleichungen mit Anfangsbedingungen, wobei die Zeit meist durch Differenzenquotienten diskretisiert wird. Ebenso können Finite Elemente zur Diskretisierung der Zeit herangezogen werden. Ein Ansatz beruhend auf einer FE-Diskretisierung in Raum und Zeit wurde in der jüngeren Vergangenheit unter dem Namen *Discontinuous Galerkin (DG)* entwickelt. Bei dieser Methode dürfen die Approximationslösungen in der Zeit diskontinuierlich sein. Die Kontinuität über die Zeitstreifen wird nur im schwachen Sinne erfüllt. Diese Methode wurde durch Lesaint & Raviart für hyperbolische Gleichungen erster Ordnung entwickelt und durch Johnson und Mitarbeiter [36] zur Zeitdiskretisierung der Konvektions-Diffusions-Gleichung herangezogen. In [39] wird gezeigt, daß die Konvergenzeigenschaften dieser Formulierung denen entsprechen, die für die SUPG-Methode im stationären Fall erzielt werden, d.h. der Ordnung $h^{m+1/2}$ mit einer FE-Diskretisierung der Größe h und stückweise polynominalen Interpolationen des Grades m.
Im Gegensatz zu anderen zeitkontinuierlichen Methoden mit Stabilitätsbedingungen, ist die Diskontinuierliche Galerkin-Methode bedingungslos stabil. Darüberhinaus weist die Methode weitere Vorteile auf. Unstrukturierte Netze, sowohl in Raum als auch in der Zeit, können leicht gehandhabt werden. Dadurch, daß die Methode auf einer Integralform der Differentialgleichung basiert, sind einfachere Konvergenzbeweise und Fehlerabschätzungen möglich.
Die Nachteile von Schemata mit zentralen Differenzen, die im stationären Fall beobachtet wurden, wenn die ersten Ableitungen im Raum dominant waren, zeigen sich ebenso im instationären Problem. Diese können mit jedem der oben angedeuteten Verfahren gehandhabt werden.
Verwendet man ein 2-Level-FD-Schema, so können mittels des Rückwärts-Euler-Verfahrens, welches numerische Dissipation in der Zeit einführt, die Oszillationen auf einfachem Wege verhindert werden. Es sei bemerkt, daß diese Methode nur eine Genauigkeit erster Ordnung aufweist und die Ergebnisse gedämpft sind.

Yu & Heinrich [61] entwickelten einen in der Zeit kontinuierlichen Petrov-Galerkin FE-Ansatz mit Wichtungsfunktionen, die ebenfalls von der Zeitdiskretisierung abhängen. Die erzielten Ergebnisse erweisen sich als sehr genau, aber der rechnerische Aufwand erscheint unverhältnismäßig hoch.
In der Discontinuous Galerkin-Methode werden zur Bewältigung der Oszillationen dieselben Maßnahmen wie für den stationären Fall angewendet. Es wird im folgenden eine Formulierung vorgestellt, deren Zeitdiskretisierung auf einem FD-Schema basiert. Dabei handelt es sich um ein θ-Schema oder auch die sogenannte generalisierte Trapezregel. Diese einfache Form bietet trotz der genannten Einschränkungen einige Vorteile. Eine hohe Genauigkeit läßt sich in Gegenden glatter Lösung leicht erzielen. Der Crank-Nicolson-Algorithmus gibt Genauigkeit zweiter Ordnung, er ist leicht zu programmieren und weniger kostspielig im Vergleich zu der linearen Zeitinterpolation der DG-Methode, wenn vergleichbare Genauigkeit erwünscht ist.

8.1 Modellproblem

Die Notation der vorangegangenen Kapiteln wird beibehalten. Sei $[0,T]$ ein gegebenes Zeitintervall mit $T > 0$ und $t \in I = (0,T)$ ein bestimmter Zeitpunkt. Das als Modellproblem gewählte instationäre Konvektions-Diffusionsproblem gestaltet sich wie folgt: Finde die Funktion $u = u(\boldsymbol{x}, t)$, so daß

$$\begin{aligned} \frac{\partial u}{\partial t} - \epsilon \Delta u + \boldsymbol{b} \cdot \nabla u + \alpha u &= f && \text{in } \Omega \times (0,T) \\ u &= u_D && \text{auf } \Gamma_D \times (0,T) \\ \epsilon \frac{\partial u}{\partial \boldsymbol{n}} = \epsilon \nabla u \cdot \boldsymbol{n} &= t_N && \text{auf } \Gamma_N \times (0,T) \\ u &= u^0 && \text{in } \Omega \times \{0\} \end{aligned}$$

Das Geschwindigkeitsfeld $\boldsymbol{b}$, der Diffusionskoeffizient ϵ und die gegebenen Funktionen f, u_D und t_N können von der Zeit abhängen. Der Einfachheit halber werden $\boldsymbol{b}$ und ϵ als konstant in der Zeit und $\boldsymbol{b}$ als divergenzfrei angenommen. Die Funktion u^0 stellt eine gegebene Anfangsbedingung dar.

8.2 Diskretisierung im Raum

Das Vorgehen entspricht dem für das stationäre Problem. Die schwache Form erhält man wie in Kapitel 3. Allgemein sei die Wichtungsfunktion hier mit w bezeichnet.

$$\int_\Omega \frac{du}{dt} w \,\mathrm{d}\Omega + \epsilon \int_\Omega \nabla u \cdot \nabla w \,\mathrm{d}\Omega + \int_\Omega \boldsymbol{b} \cdot \nabla u \; w \,\mathrm{d}\Omega + \alpha \int_\Omega u w \,\mathrm{d}\Omega \; = \int_\Omega f w \,\mathrm{d}\Omega + \int_{\Gamma_N} t_N w \,\mathrm{d}\Gamma$$

Für die Zeitableitung $\frac{du}{dt} = \dot{u}$ wird ein linearer Ansatz entsprechend des isoparametrischen Konzepts gewählt.

$$\dot{u}^h = \sum_{k=1}^{n_{nod}} N^k \dot{u}_k = \boldsymbol{N}\dot{\boldsymbol{u}} \qquad \text{in } T$$

Für die Galerkin-Methode führt damit der in der schwachen Formulierung hinzugekommene erste Term auf den Beitrag

$$\int_\Omega \dot{u} w \,\mathrm{d}\Omega \quad \rightsquigarrow \quad \boldsymbol{w}^t \underbrace{\int_T \mathbf{N}^t\mathbf{N} \,\mathrm{d}\Omega_e}_{\mathbf{m} \;:\; \text{symm}} \dot{\boldsymbol{u}} = \boldsymbol{w}^t \mathbf{m} \dot{\boldsymbol{u}}$$

mit einer symmetrischen Fundamentalmatrix der Ansatzfunktionen.

Bei der SUPG-Methode mit einer Wichtungsfunktion $w = v + \delta \boldsymbol{b} \cdot \nabla v$ und linearen Ansätzen entstehen aus dem ersten Term, analog zu dem Vorgehen von Kapitel 3, die Anteile:

$$\int_\Omega \dot{u} w \,\mathrm{d}\Omega = \int_\Omega \dot{u} v \,\mathrm{d}\Omega + \sum_T \int_T \dot{u} \delta \boldsymbol{b} \cdot \nabla v \,\mathrm{d}\Omega.$$

Dabei liefert der erste Term den Elementmatrizenbeitrag analog zum Galerkin-Verfahren und der zweite den folgenden:

$$\int_T \dot{u} \delta \boldsymbol{b} \cdot \nabla v \,\mathrm{d}\Omega \quad \rightsquigarrow \quad \boldsymbol{v}^t \underbrace{\int_T \delta \mathbf{B}^t \boldsymbol{b} \mathbf{N} \,\mathrm{d}\Omega_e}_{\delta\mathbf{g}^t \;:\; \text{unsymm}} \dot{\boldsymbol{u}} = \boldsymbol{w}^t \delta \mathbf{g}^t \dot{\boldsymbol{u}}.$$

Durch den Zusammenbau der Matrizen ergibt sich schließlich ein globales System der Form

$$\bar{\mathbf{M}}\dot{\boldsymbol{u}} + \bar{\mathbf{A}}\boldsymbol{u} = \mathbf{F},$$

wobei $\bar{\mathbf{A}}$ der in Kapitel 3 beschriebenen globalen Matrix entspricht und $\bar{\mathbf{M}}$ die Anteile $\mathbf{M}$ aus Galerkin Verfahren und zusätzlich $\delta\mathbf{G}^t$ aus der SDFEM enthält.

8.3 Zeitdiskretisierung

Die Zeitdiskretisierung erfolgt mittels eines θ-Schemas. Sei $\theta \in [0,1]$ gegeben und Δt die Zeitschrittweite der Partitionierung des Intervalls $[0,T]$. Der Zeitschritt kann konstant oder auch abhängig von dem Zeitintervall $[t^n, t^{n+1}]$, mit $n = 0,1,2,..(N-1)$ und $t^N = T$, sein. Alle Terme, die zu dem Zeitpunkt $t = t^n$ berechnet werden, werden mit einem Kopfzeiger n versehen. In dieser Weise steht u^n für die Approximation $u^h(t^n)$ unter Fortlassung der Kopfzeiger h für das räumlich diskretisierte System.
Das θ-Schema hat für obiges Modellproblem dann die Form

$$\bar{\mathbf{M}}\frac{\boldsymbol{u}^{n+1} - \boldsymbol{u}^n}{\Delta t} + \bar{\mathbf{A}}\boldsymbol{u}^{n+\theta} = \mathbf{F}^{n+\theta}$$

Dabei sei

$$u^{n+\theta} = (1-\theta)u^n + \theta u^{n+1}.$$

Mit kleinen Umformungen erhält man

$$(\bar{\mathbf{M}} + \Delta t\theta\bar{\mathbf{A}})\boldsymbol{u}^{n+1} = \Delta t\theta\mathbf{F}^{n+1} + \Delta t(1-\theta)\mathbf{F}^n + (\bar{\mathbf{M}} - \Delta t(1-\theta)\bar{\mathbf{A}})\boldsymbol{u}^n$$

bzw. abgekürzt

$$\hat{\bar{\mathbf{A}}}\boldsymbol{u}^{n+1} = \hat{\mathbf{F}}.$$

Die Eigenschaften dieses Algorithmus bezüglich Stabilität und Konvergenz sind bekannt und können in vielen Bücher nachgelesen werden, z.B. im Zusammenhang mit Finiten Elementen für das reine Diffusionsproblem in [25], [34] und für das Konvektions-Diffusions-Problem z.B. in [7], [62]. Fehlerabschätzungen für das Diffusionsproblem unter Verwendung der Galerkin-Methode sind in [34] angegeben.
Es ist weithin bekannt, daß der Algorithmus in der Zeit erster Ordnung genau ist, sofern $\theta \in [0,1]$. Genauigkeit zweiter Ordnung wird nur erzielt, wenn $\theta = 1/2$, was dem Crank-Nicolson-Schema entspricht. Der Algorithmus ist bedingungslos stabil für $\theta \geq 1/2$. Für kleinere Werte von θ gelten hinsichtlich der Stabilität Zeitschrittweitenbeschränkungen. Von besonderem Interesse sind neben dem Crank-Nicolson-Schema die Fälle $\theta = 1$ (Rückwärts-Euler) und $\theta = 0$ (Vorwärts-Euler).
Das implizite Rückwärts-Euler-Schema ist wie oben beschrieben uneingeschränkt stabil und führt numerische Dämpfung ein. Diese Eigenschaft ist insbesondere wichtig, wenn die Lösung scharfe Gradienten im Raum-Zeit-Gebiet enthält, denn diese werden dann automatisch geglättet und somit Oszillationen ausgeschlossen. Die Verwendung dieses Schemas empfiehlt sich besonders für die ersten Zeitschritte, denn, wie sich mit einer Fourier-Reihenentwicklung zeigen läßt, besitzt die Lösung des kontinuierlichen Problems 'rapidly oscillating harmonics' mit hoher Frequenz, die mit fortlaufender Zeit schnell gedämpft werden. Viele dieser harmonics können nicht mit der Zeitdiskretisierung reproduziert werden, unabhängig davon wie klein der Zeitschritt gewählt wird. Mit der Wahl von $\theta = 1$ können diese jedoch gedämpft werden [34].
Das Vorwärts-Euler-Schema, $\theta = 0$, ist explizit, in dem Sinne, daß, wenn die Matrix $\mathbf{M}$ durch eine Diagonalmatrix approximiert wird, die Lösung der Gleichung trivial ist und kein Löser für das algebraische System benötigt wird. Allerdings sind hinsichtlich der Stabilität für dieses Verfahren Zeitschrittweitenbeschränkungen zu beachten.

Nun verbleibt noch die Diskussion der Wahl des SUPG-Parameters δ_e. Eine Möglichkeit besteht in der Übernahme der Form aus dem stationären Problem.

$$\delta_e = \frac{h_e}{2|\boldsymbol{b}_e|}\zeta(Pe^e)$$

Eine andere Wahl berücksichtigt die Courant-Zahl

$$\mathrm{Co} := \frac{|\boldsymbol{b}|\Delta t}{h},$$

die ebenso auf jedem Element definiert werden kann. Statt δ_e wird nun

$$\delta_e^t = C_\delta \delta_e$$

gesetzt, wobei C_δ von der Courant-Zahl abhängt. Für den eindimensionalen reinen Konvektionsfall mit linearen Elementen wurde die beste Zeitgenauigkeit mit dem Ausdruck

$$C_\delta = \frac{2}{\sqrt{15}} + \left(1 - \frac{2}{\sqrt{15}}\right) \mathrm{Co}$$

gefunden [56]. Für $C_\delta = \frac{2}{\sqrt{15}}$ und das Vorwärts-Euler-Schema wurde eine Phasengenauigkeit vierter Ordnung erzielt [50]. Diese Wahl wurde auch für allgemeine transiente Probleme z.B. in [6], [32] getroffen. In Fällen von Auftreten von Diffusion oder bei Verwendung von quadratischen Elementen wurde festgestellt, daß die Wahl von $C_\delta = 1$ optimal ist, sofern der Fehler in der diskreten L^2-Norm gemessen wird [11]. Somit empfiehlt sich die Verwendung der gleichen Testfunktion für das transiente Problem wie für das stationäre Problem.

Die Courant-Zahl kann auch als Kriterium zur Diskretisierung des Zeitschritts dienen. In der Form

$$\mathrm{Co} = \frac{|\boldsymbol{b}|\Delta t_{Co}}{h} \leq 1$$

wird garantiert, daß während eines Zeitschritts Δt_{Co} die Konzentrationsdifferenz in einem Element nicht größer werden kann als die durch advektive Stoffströme. Ein weiteres Kriterium zur Beschränkung des Zeitschritts stellt das sogenannte Neumann-Kriterium

$$\mathrm{Ne} = \frac{\epsilon \Delta t_{Ne}}{h^2} \leq \frac{1}{2}$$

dar [44]. Es soll verhindern, daß Konzentrationsgradienten in einem Element während eines Zeitschritts Δt_{Ne} allein durch diffusive Stoffströme umgekehrt werden. In der praktischen Anwendung können beide Kriterien in sinnvoller Weise Niederschlag finden, indem das kleinere Δt aus beiden Ausdrücken als Zeitschrittweite gewählt wird [42]

$$\Delta t = \min\{\Delta t_{Co}, \Delta t_{Ne}\}.$$

Ist das zu berechnende Problem durch starke Konvektion geprägt, so wird i.a. das Courant-Kriterium ausschlaggebend sein.

8.4 Netzanpassung

Bei instationären Berechnungen muß im Gegensatz zu stationären Berechnungen davon ausgegangen werden, daß sich die Bereiche, in denen große Fehler auftreten, während des Rechenprozesses ändern. Durchwandert z.B. eine Schockfront ein eindimensionales Gebiet, so wird bei der Verwendung einer reinen Verfeinerungsstrategie das gesamte Gebiet gleichmäßig verfeinert sein, wenn die Schockfront das Ende erreicht. Der Rechenaufwand ist in diesem Fall unnötig hoch, da lediglich in der näheren Umgebung der Schockfront kleinere Elemente gebraucht werden. Aus diesem Grunde ist es sinnvoll, zusätzlich zu den Netzverfeinerungsstrategien der adaptiven stationären Berechnung Vergröberungsstrategien bereitzustellen.
Im folgenden sollen drei unterschiedliche Methoden zur Netzanpassung im instationären Rechenprozeß dargestellt werden.

8.4.1 Verfeinerung ausgehend von einem Grundnetz

Die erste Verfahrensweise umgeht die Forderung nach Vergröberungstechniken, indem in jedem Zeitschritt ausgehend von einem relativ feinen uniformen Grundnetz über eine vorgegebene Anzahl von Schritten adaptiv verfeinert wird. Dies entspricht für jeden einzelnen Zeitschritt dem Vorgehen im stationären Fall. Die Anfangsbedingungen für jeden neuen Zeitschritt werden vom letzten verfeinerten Netz des vorhergehenden Zeitschrittes auf das Grundnetz interpoliert. Da dieses im Vergleich zum verfeinerten Netz des alten Zeitschritts relativ grob ist, können wichtige Informationen verloren gehen. Somit stellt dieses Verfahren eine recht einfach aus dem stationären Berechnungsprogramm zu verwirklichende Variante dar, wenn insbesondere keine Vergröberungsmethoden verfügbar sind. Allerdings muß die günstige Umsetzbarkeit mit einem Verlust an Genauigkeit bezahlt werden.

8.4.2 Vergröberungsmaßnahmen

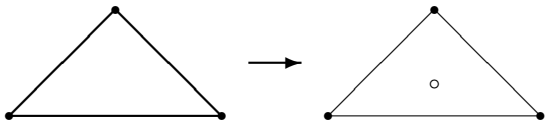
Im Falle strukturierter hierarchischer Netze, wie sie bisher im Rahmen dieser Arbeit betrachtet wurden, lassen sich einzelne Elemente mit Hilfe der in Kapitel 6 beschriebenen Markierungsstrategie 4 und unter Beachtung des Vorliegens auflösbarer Patches auf ihre Vorfahrenelemente zurückführen. Die Durchführung von Vergröberungen innerhalb eines FE-Programms bedarf allerdings einer geschickten Datenhaltung, da große Netze, langsame konvektive Prozesse und Speicherplatzbeschränkungen Schwierigkeiten bei der Bereitstellung von netzgeschichtlichen Daten bereiten können. Ein wesentlicher Vorteil dieser Methode besteht darin, daß i.a. von Zeitschritt zu Zeitschritt, insbesondere bei kleinen Geschwindigkeiten, nur wenige Elemente verfeinert bzw. vergröbert werden müssen. Damit liegen die benötigten Anfangswerte für den jeweils neuen Zeitschritt bereits an den meisten Netzknoten vor und nur wenige müssen interpoliert werden.
In dem im Rahmen dieser Arbeit verwendeten Programm konnte eine Vergröberung über eine Elementgeneration verwirklicht werden. Es zeigt sich allerdings, daß dies mitunter nicht ausreichend ist und eine tiefere Vergröberung, d.h. über mehrere Generationen wünschenswert wäre.

8.4.3 Delaunay-Triangulierungen

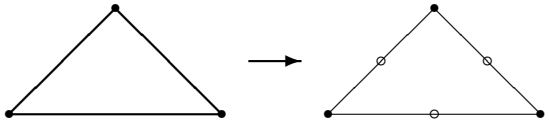
Eine alternative Methode basiert auf der Nutzung unstrukturierter sogenannter Delaunay-Triangulierungen. Abweichend von der bisher verwendeten Erzeugung hierarchischer Netze werden Delaunay-Triangulierungen aus einem gegebenen beliebigen Satz von Knotenkoordinaten entwickelt. Zur Verfeinerung einer Triangulierung werden in einem entsprechend markierten Element zusätzliche Knoten angeordnet. Bei der darauffolgenden Neuvernetzung der vorliegenden Knoten können neue Dreieckszusammenhänge, insbesondere durch "edge swapping" entstehen. In entsprechender Weise werden bei zur Vergröberung markierten Elementen Knoten entfernt, die dann bei der Neuvernetzung nicht mehr auftauchen. Details zum Konstruktionsalgorithmus von Delaunay-Triangulierungen können in [52] nachgelesen werden.

Zur Unterteilung eines Elements sind verschiedene Varianten denkbar, u.a.:

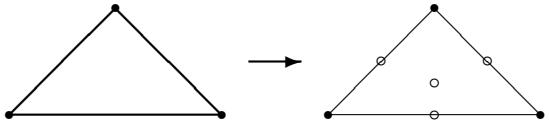
1. Einfügen eines neuen Knotens im Schwerpunkt des Dreiecks



2. Einfügen neuer Knoten in den Kantenmittelpunkten



3. Einfügen neuer Knoten in den Kantenmittelpunkten und im Schwerpunkt



Zusatzregeln müssen darüberhinaus nicht beachtet werden, da der Triangulierungsalgorithmus die vorgegebenen Knoten selbständig in eindeutiger Weise vernetzt.

Auch für die Vergröberung von einzelnen Elementen sind unterschiedliche Möglichkeiten denkbar:

1. Wegnahme der drei Eckknoten und Einsatz eines neuen Knotens im Schwerpunkt des Dreiecks

2. Wegnahme der Eckknoten ohne Einfügung eines neuen Knotens unter Beachtung einiger fixierter Grundknoten, die eine grobe Triangulierung garantieren

Im Fall der Vergröberung kann es sinnvoll sein, zu kontrollieren, ob ein zu vergröberndes Element an ein zu verfeinerndes Element grenzt. Dann sollten die betroffenen gemeinsamen Knoten nicht durch die Vergröberung entfernt werden.

8.5 Beispiel

8.5.1 Rotierender Sinoid

Als Beispiel soll hier das folgende Konvektionsproblem in zwei Raum-Dimensionen mit hinreichend glatter Anfangsbedingung dienen

$$\begin{aligned} \frac{\partial u}{\partial t} + \boldsymbol{b} \cdot \nabla u &= 0 && \text{in } \Omega \times (0,T) \\ u = u_D &= 0 && \text{auf } \partial\Omega \times (0,T) \\ u &= u^0 && \text{in } \Omega \times \{0\}. \end{aligned}$$

Das Problem ist auf dem Gebiet $\Omega = (-1,1) \times (-1,1)$ gestellt. Das Geschwindigkeitsfeld $\boldsymbol{b} = r(-\sin\theta, \cos\theta)$ beschreibt eine Rotation entgegen dem Uhrzeigersinn um den Ursprung. Die Anfangsbedingung ist durch

$$u^0 = \frac{1}{2}\left(1 + \cos\left(\frac{5}{2}\pi R\right)\right), \quad \text{wenn} \quad R = \sqrt{x^2 + \left(y - \left(-\frac{1}{2}\right)\right)^2} \leq \frac{2}{5}$$

gegeben. In den ersten 5 Zeitschritten wurde der Euler-Rückwärts-Algorithmus ($\theta = 1$) verwendet, danach das Crank-Nicolson-Verfahren ($\theta = 1/2$).

Quasi-stationäre Behandlung und sukzessive Verfeinerung

Für jeden einzelnen Zeitschritt, mit konstanter Schrittweite $\Delta t = 0.025$, wurde ausgehend von einem uniformen Grundnetz in gewohnter Weise adaptiv mit dem Z^2-Indikator und regulärer Verfeinerung (Markierungsstrategie 2, $\gamma = 0.5$, $p = 0.15$) über 4 Stufen verfeinert. Als Anfangsbedingung für den neuen Zeitschritt wurde jeweils die Lösung vom feinsten Netz des vorangegangen Zeitschritts auf das neue Netz interpoliert.

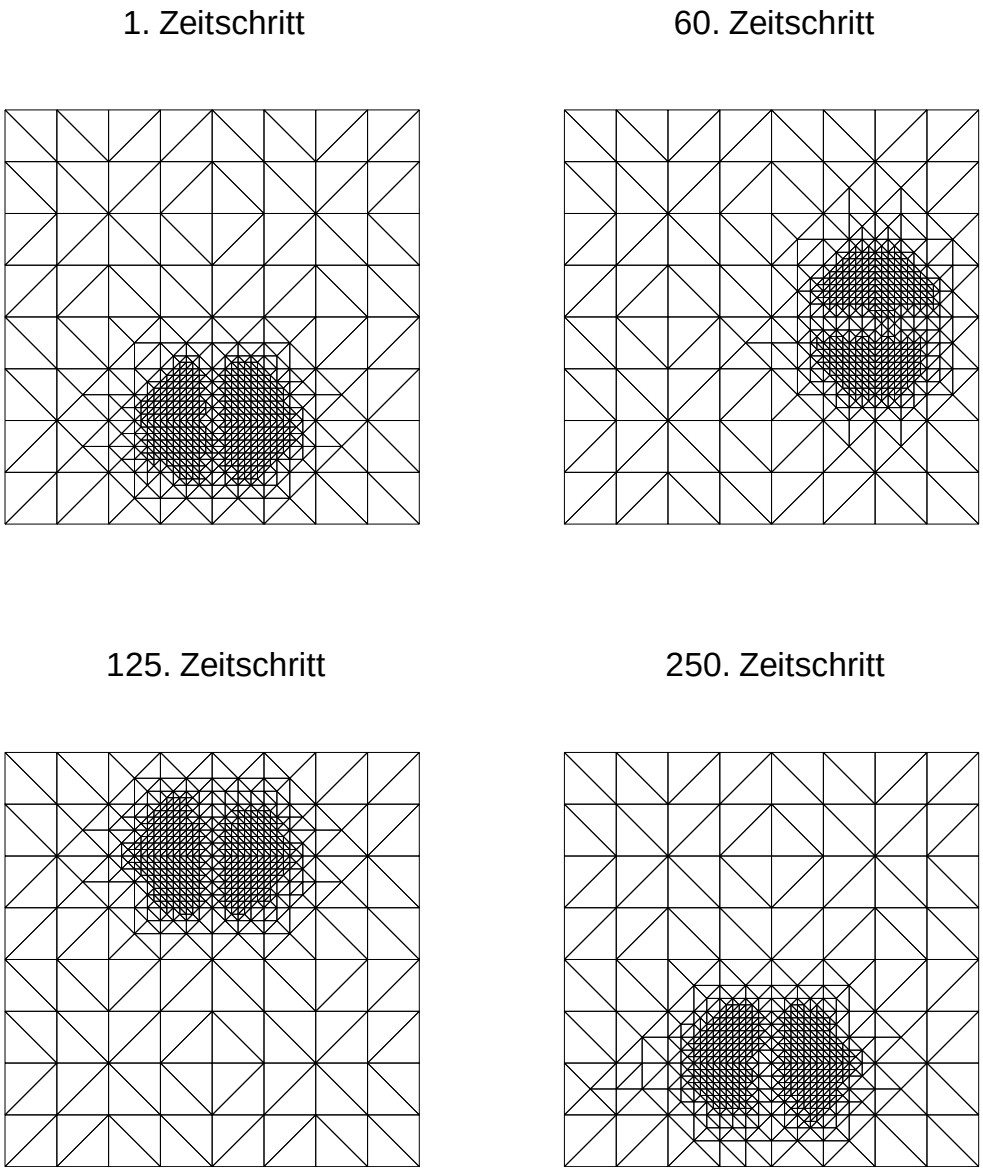


Abbildung 8.1: Rotierender Sinoid : verfeinerte Netze (quasi-stationär)

Die Abbildung 8.1 zeigt die verfeinerten Netze nach dem 1. Zeitschritt, dem 60. Zeitschritt, gleichbedeutend mit einer Rotation um ungefähr 90°, dem 125. Zeitschritt und dem 250. Zeitschritt, ensprechend den Rotationen um 180° bzw. 360°. Die zugehörigen Lösungen sind in Abbildung 8.2 dreidimensional dargestellt.

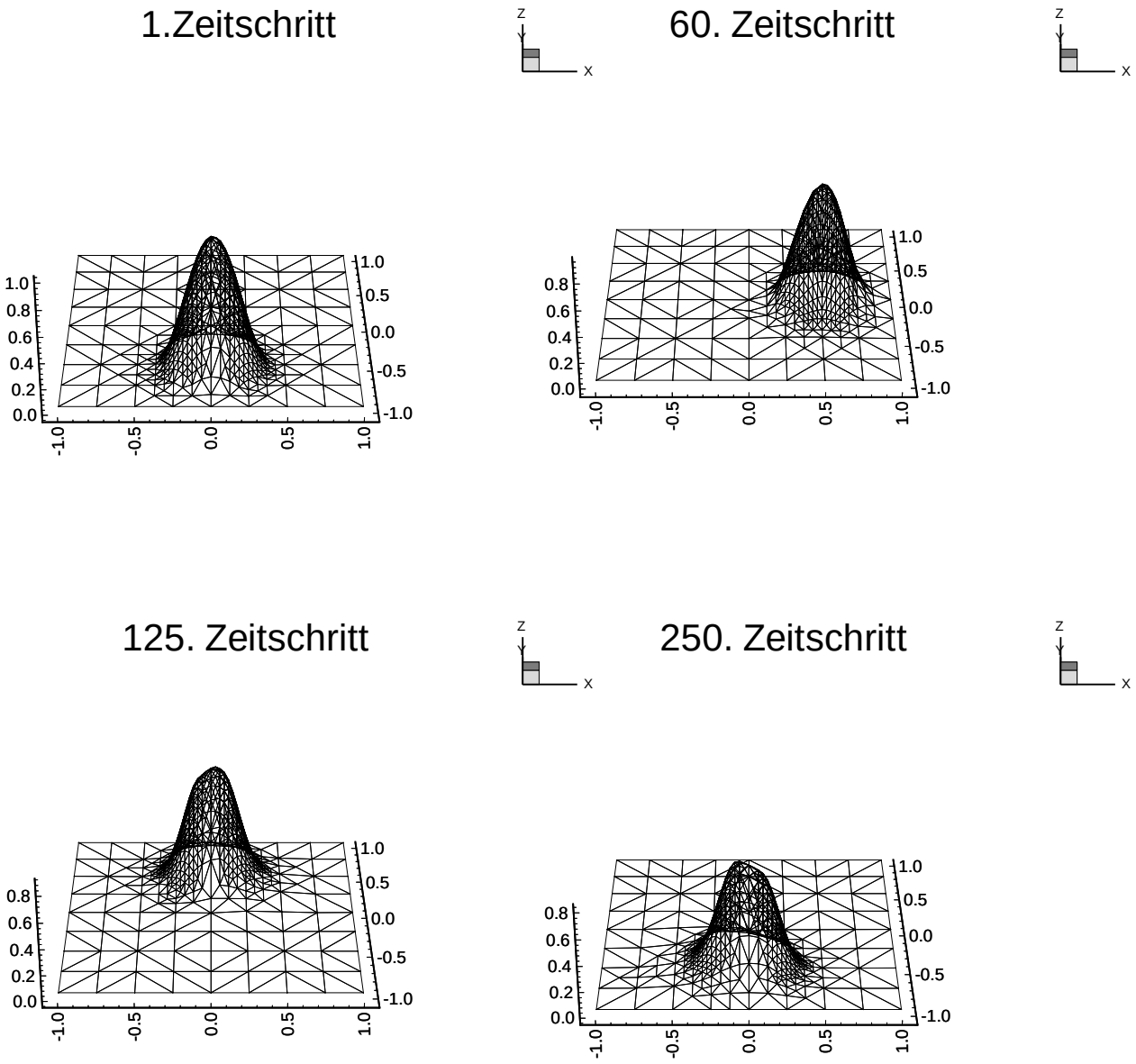


Abbildung 8.2: Rotierender Sinoid : Lösungen (quasi-stationär)

Da diese Vorgehensweise auf einen großen Verfeinerungs- und damit Rechenaufwand führt, soll nun eine Strategie zur sukzessiven Verfeinerung bzw. Vergröberung vorgeführt werden. Im ersten Zeitschritt wird zunächst bis zu einer gewünschten Verfeinerungstiefe verfeinert und nachfolgend basierend auf dieser Triangulierung in jedem weiteren Zeitschritt weiter verfeinert bzw. vergröbert.

Für die folgenden Bilder wurde die Markierungsstrategie 4 mit $\theta_1 = 0.7$, $\theta_2 = 0.0$ und $TOL = 10^{-3}$ und Bisektion der Elemente gewählt. Die übrigen Parameter wurden von oben übernommen.

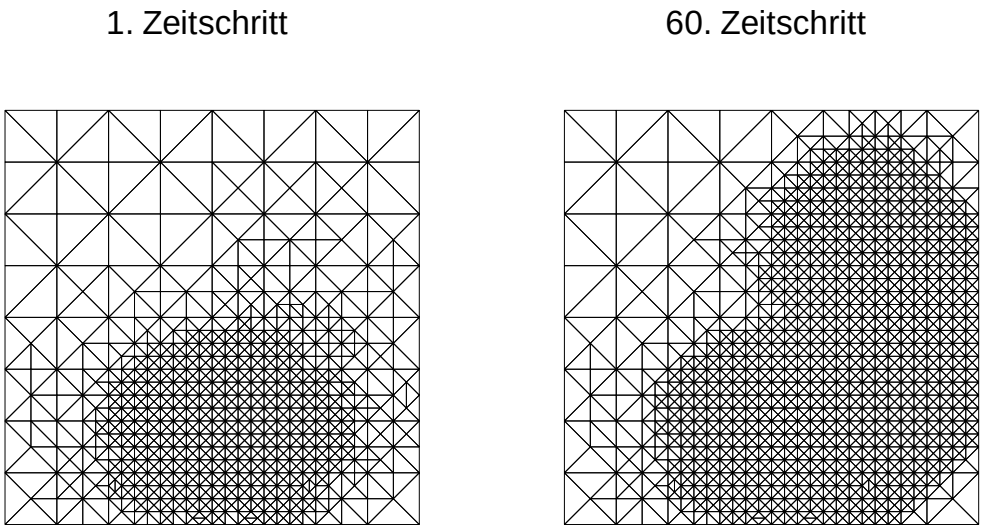


Abbildung 8.3: Rotierender Sinoid : verfeinerte Netze (sukzessive Verfeinerung)

Abbildung 8.3 zeigt die verfeinerten Netze für den 1. und den 60. Zeitschritt, Abbildung 8.4 die zugehörigen Lösungen in dreidimensionaler Darstellung.
Im Vergleich zu dem quasi-stationären Vorgehen zeigt sich hier in der numerischen Lösung nach einer Viertelrotation keine Abnahme des Maximalwertes.
Durch die Wahl $\theta_2 = 0.0$ wird in diesem Fall keine Vergröberung des Netzes vorgenommen. Nach einer vollen Rotation wird ein nahezu gleichmäßig verfeinertes Netz vorliegen. Da auch dies einen unnötig hohen Rechenaufwand nach sich zieht, ist eine teilweise Netzvergröberung unabdingbar.
Im verwendeten Programm wurde eine Vergröberung über eine Elementgeneration ermöglicht. Für das vorliegende Beispiel zeigen sich bei einer Wahl von $\theta_2 = 0.05$ keine signifikanten Unterschiede in den Netzen. Um zu einer deutlichen Abnahme der Netzdichte in den Bereichen hinter den großen Konzentrationen zu gelangen, ist eine tiefere Vergröberung nötig. Diese ist allerdings im verwendeten Programm nicht umgesetzt worden.

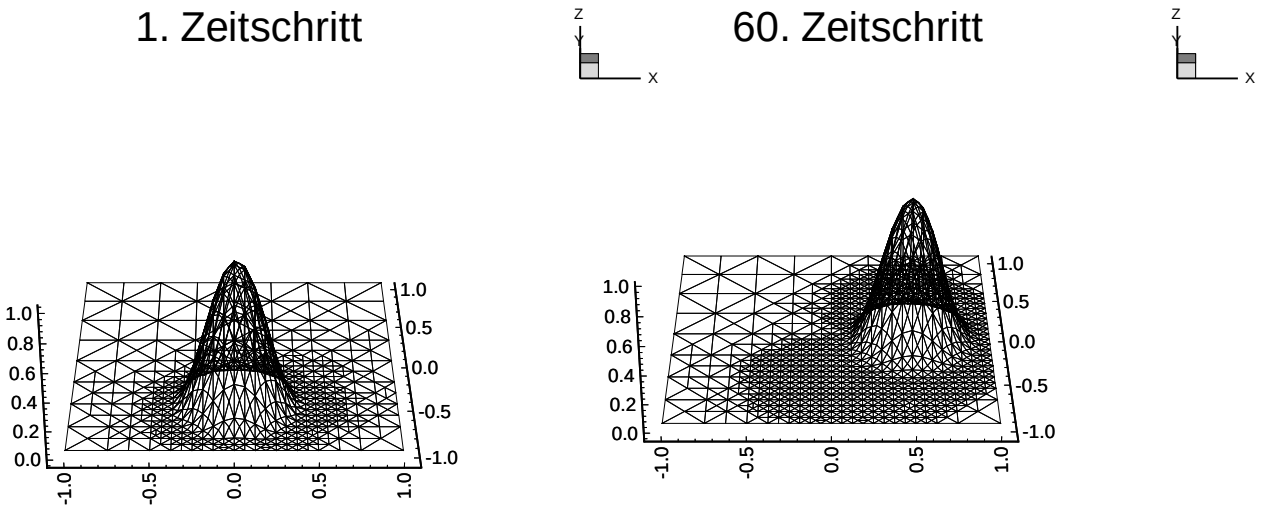


Abbildung 8.4: Rotierender Sinoid : Lösungen (sukzessive Verfeinerung)

Delaunay-Triangulierungen

Bei der Verwendung von Delaunay-Triangulierungen wird zunächst im ersten Zeitschritt ausgehend von einem Grundnetz, basierend auf einem kleinen Satz von Knoten, über einige Stufen in gewohnter Weise verfeinert bzw. vergröbert. In dem feinsten Netz des ersten Zeitschritts wird der Flächeninhalt des kleinsten Elements A_{min} festgestellt und gespeichert. Dieser dient später zur Beschränkung der Verfeinerungstiefe. In den folgenden Zeitschritten wird ein Element zur Verfeinerung nur dann markiert, wenn der geschätzte Fehler in ihm die vorgegebene Toleranz überschreitet und zusätzlich sein Flächeninhalt größer als das Dreifache von A_{min} ist. Ohne diese Bedingung könnten extrem kleine Elemente entstehen, die wiederum durch das Courant- bzw. Neumann-Kriterium auf sehr kleine Zeitschrittweiten führen würden.
In den folgenden Beispielrechnungen wurden Elemente durch Einfügen zusätzlicher Knoten in den Elementschwerpunkt verfeinert. Die Vergröberung erfolgte durch Wegnahme der Elementknoten unter Berücksichtigung eines festen Grundnetzes, das nicht vergröbert werden kann. Die Verfeinerung innerhalb des ersten Zeitschrittes lief über 6 Verfeinerungstufen. Die Parameter der Markierungsstrategie 4 wurden zu $TOL = 0.1$, $\theta_1 = 0.45$ und $\theta_2 = 0.01$ gewählt. Als Fehlerschätzer diente der residuelle Schätzer. Die erste Zeitschrittweite wurde mit $\Delta t_1 = 0.0625$ vorgegeben, danach wurde eine Zeitschrittweitensteuerung nach den oben genannten Kriterien vorgenommen, d.h. aufgrund fehlender Diffusion steuerte das Courant-Kriterium die Schrittweite. Alle weiteren Parameter wurden wie im vorigen Beispiel belassen.
Die Abbildung 8.5 zeigt die Netze zu den Zeitpunkten $t_1 = 0.0625$, $t_{50} = 1.4729$, $t_{100} = 2.9120$ und $t_{210} = 6.0247$. In Abbildung 8.6 sind die zugehörigen dreidimensional dargestellten Lösungen zu finden.

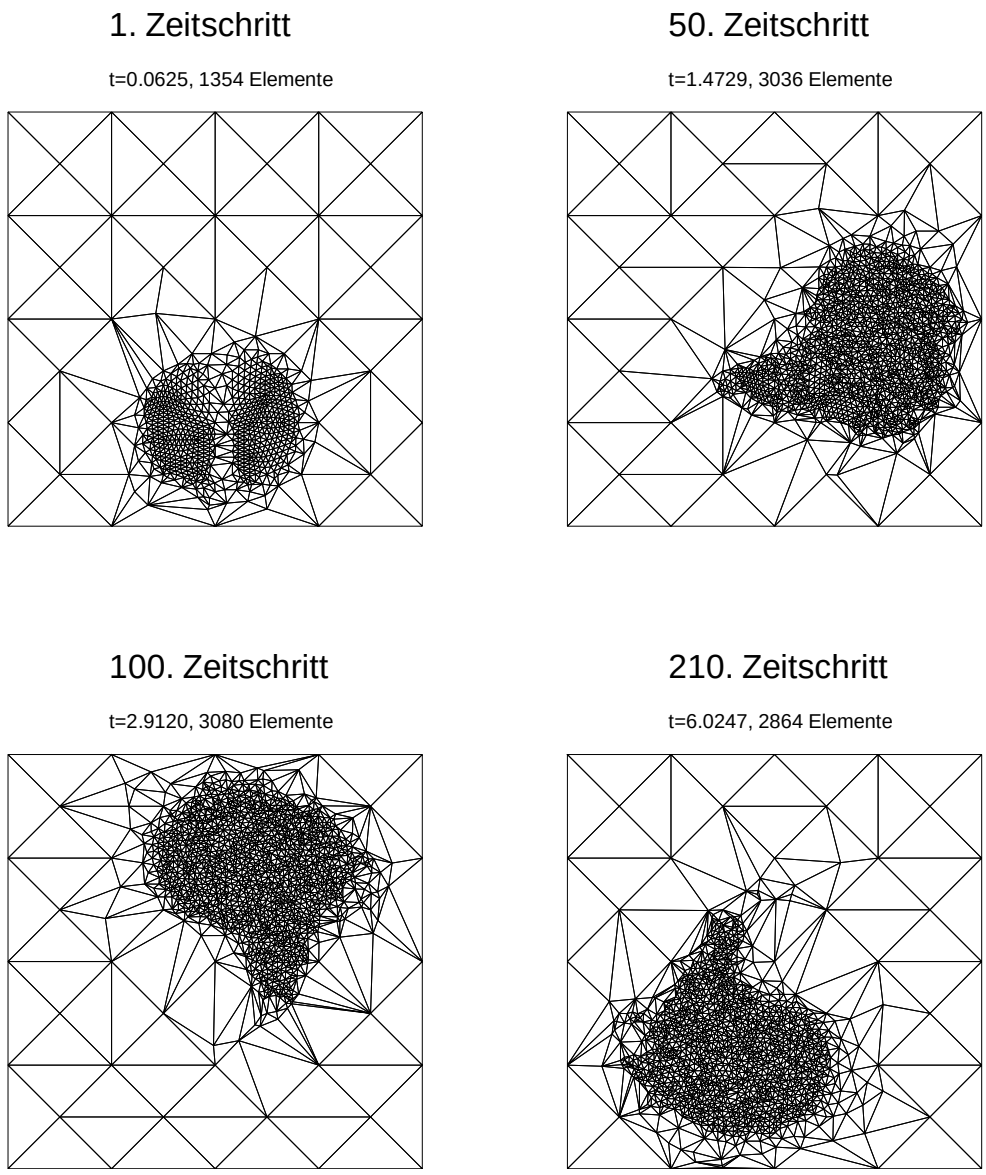


Abbildung 8.5: Rotierender Sinoid : Netze (Delaunay)

In den Netzen zeigt sich eine deutliche Zunahme der Elementzahl im Verlauf des Prozesses. Außerdem ist zu beobachten, daß eine kleine Fahne verfeinerter Elemente nachgezogen wird. Diese Phänomene erklären sich aus der bereits in Kapitel 6 erwähnten Schwierigkeit der Wahl der Markierungsparameter. Bei anderen Konstellationen von θ_1, θ_2, TOL treten diese Begleiterscheinungen nicht oder weniger gravierend auf. Insgesamt erfolgt die Verfeinerung ebenso wie die Vergröberung sehr zuverlässig und der numerischen Lösung angepaßt.
Die Ausgangsgestalt der numerischen Lösung wird bei dieser Vorgehensweise über alle Zeitschritte erhalten. Erst gegen Abschluß einer vollen Rotation erfährt die Lösung eine sehr leichte Dämpfung.

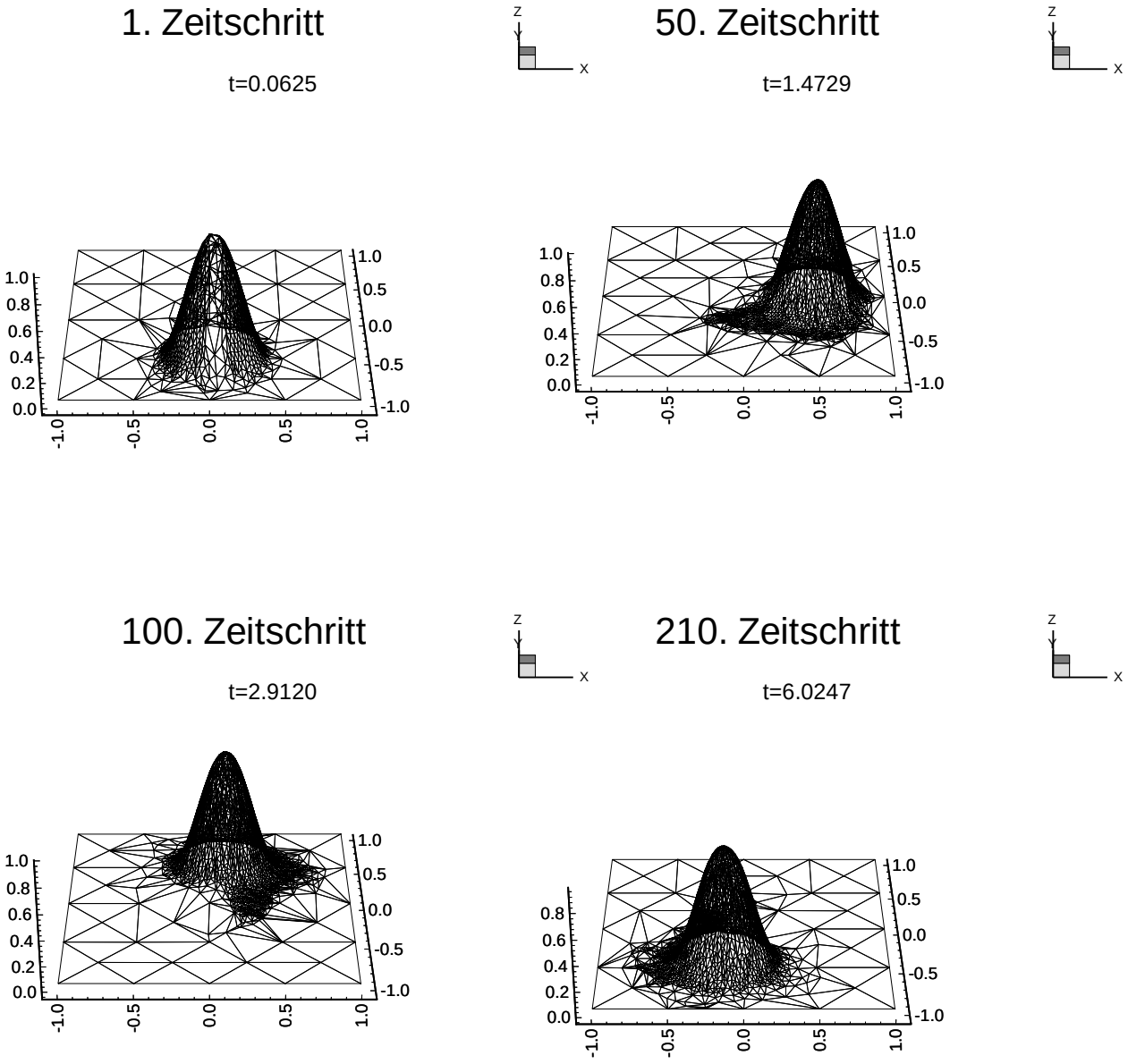


Abbildung 8.6: Rotierender Sinoid : Lösungen (Delaunay)

8.6 Zusammenfassung

Das hier vorgestellte Vorgehen zur Erweiterung der Methoden für stationäre Probleme auf instationäre Probleme mittels eines θ-Schemas erweist sich als sehr praktikabel. Durch die Wahl von $\theta = 1$ für die ersten Zeitschritte werden Oszillationen gedämpft. Die anschließende Wahl von $\theta = 1/2$ garantiert neben Stabilität eine Genauigkeit zweiter Ordnung.
Insbesondere ist bei instationären Problemen unverzichtbar, Rücksicht auf die mögliche Veränderung der Gebiete mit großem Fehler zu nehmen. Eine denkbare Vorgehensweise stellt die sukzessive Verfeinerung in jedem Zeitschritt ausgehend von einem Grundnetz dar, eine andere die Anwendung von Vergröberungsstrategien und eine dritte die Verwendung von Delaunay-Triangulierungen.
Das Beispiel des rotierenden Sinoids zeigt für die quasi-stationäre Behandlung in jedem Zeitschritt über die Zeit eine deutliche Dämpfung der Lösung. Außerdem leidet die Genauigkeit durch die Interpolation der Anfangsbedingungen für jeden neuen Zeitschritt von einem feinen Vorgängernetz auf das relativ grobe Grundnetz. Diese Methode stellt eine leicht umsetzbare Erweiterung des Vorgehens im stationären Fall dar, die aber einige Mängel aufweist.
Die sukzessive Verfeinerung bzw. die damit gekoppelte Vergröberung weist demgegenüber den Vorteil auf, von Zeitschritt zu Zeitschritt nur relativ kleine Veränderungen am Netz vorzunehmen. Damit werden bei der Interpolation der Anfangsbedingungen Schwierigkeiten umgangen. Außerdem werden Rechenkosten gespart, da nicht in jedem Zeitschritt ein feines Netz aus gröberen Netzen entwickelt werden muß und damit mehrere Probleme gelöst werden müssen. Eine Vergröberung über eine Elementgeneration ist zur Auflösung feiner Netzteile in Gebieten hinter einer Schockfront nicht ausreichend. Durch verbleibende Gebiete mit großer Netzdichte, aber geringem Fehler, wird der Aufwand der Berechnung unnötig gesteigert.
Mit der Verwendung von Delaunay-Triangulierungen geht man von hierarchischen und strukurierten Netzen zu einer knotenbasierten Netzentwicklung über. Die Problematik der Verfügbarkeit von netzgeschichtlichen Daten stellt sich bei diesem Vorgehen nicht, da jedes Gitter automatisch aus einem Satz von Knoten entwickelt wird. Mit geeigneten Verfeinerungs- bzw. Vergröberungskriterien läßt sich durch Einfügen bzw. Wegnehmen von Knoten Einfluß auf die Netzgestalt nehmen. Die Netze im Beispiel zeigen eine sehr gute Anpassung an die jeweilige numerische Lösung.
Da kein Beispiel mit exakter Lösung zur Verfügung stand, konnte leider kein quantitativer Vergleich zwischen den verschiedenen Verfahren angestellt werden. Eine Überlegenheit der Verfahrensweise mit Delaunay-Triangulierungen gegenüber der quasi-stationären Behandlung der Zeitschritte und der sukzessiven Verfeinerung ohne ausreichende Vergröberung zeichnet sich dennoch ab.
Wie sich zeigt, bieten instationäre Probleme ein weites Feld für adaptive Methoden und ihre einzelnen Teilaspekte. Neben effektiven Netzanpassungsverfahren können numerische Verfahren wie die diskontinuierliche Galerkin Methode als Ausgangspunkt für weitere Untersuchungen dienen.

Kapitel 9

Zusammenfassung

In dieser Arbeit wurden Aspekte der Theorie und Numerik adaptiver Finiter Element Methoden für Konvektions-Diffusionsprobleme behandelt.
Zunächst wurden die bei Konvektions-Diffusionsproblemen auftretenden charakteristischen Prozesse erläutert und das stationäre Modellproblem formuliert. Anschließend wurde die numerische Behandlung der Differentialgleichung mittels Finiter Element Methoden diskutiert. Ausgehend von der Bubnov-Galerkin Methode, die bei dominanter Konvektion unphysikalische numerische Lösungen liefert, wurden stabile Finite Element Methoden mit zufriedenstellenden Konvergenzeigenschaften zur Lösung eines Konvektions-Diffusionsproblems auch mit dominanter Konvektion bereitgestellt. Insbesondere wurden spezielle Aspekte der SU/PG bzw. SDFEM wie die Wahl des Parameters δ und Shock-Capturing Modifikationen dargestellt.

Zur iterativen Lösung der aus den Finiten Element Diskretisierungen entstehenden unsymmetrischen Gleichungssysteme wurde das BiCG-Stab Verfahren vorgestellt.

Auf dem Wege zu einer effizienten adaptiven Methode, welche basierend auf Informationen über den Fehler eine Approximation gewünschter Genauigkeit mit nahezu minimalem Aufwand liefert, wurden verschiedene Möglichkeiten zur *a posteriori* Abschätzung des Diskretisierungsfehlers bereitgestellt. Zunächst wurde der Z^2-Indikator aus einer Mittelungstechnik dargestellt. Ausgehend von einer klassischen Formulierung eines residuellen Fehlerschätzers wurden modifizierte Vorfaktoren der Residuenterme entwickelt. Desweiteren wurde ein Fehlerschätzer basierend auf einem lokalen Konvektions-Diffusionsproblem mit Neumann Randbedingungen eingeführt. Numerische Beispiele dienten dem Vergleich der verschiedenen Fehlerschätzer.

Zur Durchführung einer Netzverfeinerung wurden eine Vielzahl von Markierungsmethoden und Verfeinerungsstrategien gegenübergestellt. Anhand von Beispielen wurden diese verglichen.

Desweiteren wurde ein Netzglättungsalgorithmus zur Optimierung der Plazierung der Netzknoten, der auf Informationen eines *a posteriori* Fehlerschätzers beruht, vorgestellt. Bemerkenswert an dieser Methode ist, daß lediglich Ergebnisse vorangegangener Berechnungen benutzt werden und keine Details der Lösung der Differentialgleichung oder der Fehlerschätzer benötigt werden. Insbesondere wurde zur Einbindung des in dieser Arbeit

eingeführten Neumann-Fehlerschätzers eine Fehlerquadrat-Minimierung zur Berechnung der Matrix $\mathbf{M}_T$ vorgestellt. Diese Netzanpassungsmethode trägt zur Minimierung des Fehlers jedoch nicht vergleichbar stark bei wie eine Netzverfeinerung. Dennoch stellt sie insbesondere bei Speicherplatzbeschränkungen eine sinnvolle Ergänzung zu ihr dar.

Schließlich erfolgte die Erweiterung der Betrachtung auf ein instationäres Modellproblem. Dabei wurde die Zeit mittels eines θ-Schemas diskretisiert und eine Zeitschrittweitensteuerung ermöglicht. Die räumliche Diskretisierung wurde aus der Behandlung des stationären Problems übernommen. Zur Abschätzung des Diskretisierungsfehlers stehen in jedem Zeitschritt die vorgestellten Fehlerschätzer zur Verfügung. Basierend auf diesen Informationen wurden zur Anpassung der Netze verschiedene Verfahrensweisen vorgeschlagen. Neben einer Verfeinerung in jedem Zeitschritt ausgehend von einem Grundnetz und einer fortschreitenden Verfeinerung bzw. Vergröberung erwies sich die Verwendung von Delaunay-Triangulierungen als wirkungsvoll.

Kapitel 10

Anhang

10.1 Anhang 1

10.1.1 Dreieckselemente

- Einführung von Flächenkoordinaten

Außerhalb eines Elementes werden die globalen kartesischen Koordinaten x, y verwendet. Im Element wird das lokale System der Flächenkoordinaten $\lambda_1, \lambda_2, \lambda_3$ benutzt. Diese ergeben sich durch das Verhältnis der Teilflächen A_i, welche durch Verbindung eines Punktes P innerhalb des Dreiecks mit den Eckpunkten entstehen, zu der Gesamtfläche.

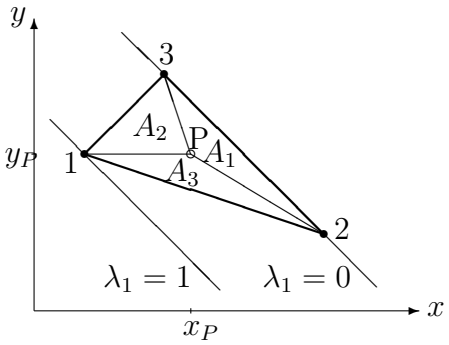


Abbildung 10.1: Lokales Koordinatensystem

Ein Punkt innerhalb des Dreieckes ist festgelegt durch die drei Flächenkoordinaten

$$\lambda_1 = \frac{A_1}{A} \ , \quad \lambda_2 = \frac{A_2}{A} \ , \quad \lambda_3 = \frac{A_3}{A} \ ,$$

$$\text{wobei} \quad \lambda_1 + \lambda_2 + \lambda_3 = 1$$

Es existieren also nur zwei unabhängige Flächenkoordinaten. Dieses Koordinatensystem ist unabhängig von der Lage das globalen Systems und der Größe der Querschnittsfläche A.

• Gesamtfläche A

Die Gesamtfläche eines Dreiecks, bestehend aus den Teilflächen A_1, A_2, A_3 berechnet sich wie folgt aus dem Kreuzprodukt der Kantenvektoren:

$$\begin{aligned} A &= \frac{1}{2} \parallel (\boldsymbol{x}_2 - \boldsymbol{x}_1) \times (\boldsymbol{x}_3 - \boldsymbol{x}_1) \parallel \\ &= \frac{1}{2}\{(x_2 - x_1)(y_3 - y_1) - (y_2 - y_1)(x_3 - x_1)\} \\ &= \frac{1}{2} \det \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix} \end{aligned}$$

• Teilflächen A_i

$$\begin{aligned} A_1 &= \frac{1}{2} \parallel (\boldsymbol{x}_2 - \boldsymbol{x}_P) \times (\boldsymbol{x}_3 - \boldsymbol{x}_P) \parallel \\ &= \frac{1}{2} \det \begin{bmatrix} 1 & x_P & y_P \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix} \\ &= \frac{1}{2} \left(\underbrace{\det \begin{bmatrix} x_2 & y_2 \\ x_3 & y_3 \end{bmatrix}}_{A_1^0} - x_P \underbrace{\det \begin{bmatrix} 1 & y_2 \\ 1 & y_3 \end{bmatrix}}_{-\alpha_1} + y_P \underbrace{\det \begin{bmatrix} 1 & x_2 \\ 1 & x_3 \end{bmatrix}}_{\beta_1} \right) \\ &= \frac{1}{2} \left(A_1^0 + x_P \alpha_1 + y_P \beta_1 \right) \end{aligned}$$

Analog erhalten wir die Teilflächen A_2, A_3 nach der allgemeinen Formulierung

$$A_i = \frac{1}{2}(A_i^0 + x_P \alpha_i + y_P \beta_i)$$

mit den Abkürzungen

$$\begin{aligned} A_i^0 &= \det \begin{bmatrix} x_{i+1} & y_{i+1} \\ x_{i+2} & y_{i+2} \end{bmatrix} \\ \alpha_i &= y_{i+1} - y_{i+2} \\ \beta_i &= x_{i+2} - x_{i+1} \qquad \text{(mit zyklischer Vertauschung, } i = 1, 2, 3) \end{aligned}$$

• Kantenvektoren

Die Kanten können auch durch die folgenden Kantenvektoren repräsentiert werden:

$$l_1 = \begin{bmatrix} x_3 - x_2 \\ y_3 - y_2 \end{bmatrix}, \qquad l_2 = \begin{bmatrix} x_1 - x_3 \\ y_1 - y_3 \end{bmatrix}, \qquad l_3 = \begin{bmatrix} x_2 - x_1 \\ y_2 - y_1 \end{bmatrix}.$$

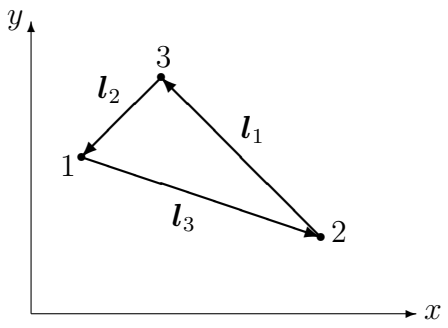


Abbildung 10.2: Kantenvektoren

• Lineare Ansatzfunktionen

Die Flächenkoordinaten berechnen sich also aus den globalen kartesischen Koordinaten zu

$$\lambda_i = \frac{A_i}{A} = \frac{1}{2A}\left(A_i^0 + x_P\alpha_i + y_P\beta_i\right) \ .$$

Für lineare Ansätze stimmen die Ansatzfunktionen mit den Flächenkoordinaten überein

$$N^i = \lambda_i \ ,$$

in der folgenden Abbildung sind diese veranschaulicht.

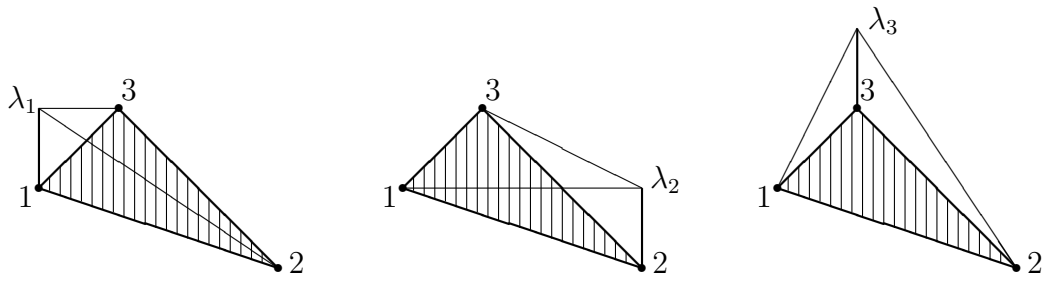


Abbildung 10.3: Lineare Ansatzfunktionen

• Zusammenhang Flächenkoordinaten-kartesische Koordinaten

Die Flächenkoordinaten ergeben sich aus den kartesischen Koordinaten durch :

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} = \frac{1}{2A}\begin{bmatrix} A_1^0 & \alpha_1 & \beta_1 \\ A_2^0 & \alpha_2 & \beta_2 \\ A_3^0 & \alpha_3 & \beta_3 \end{bmatrix}\begin{bmatrix} 1 \\ x_P \\ y_P \end{bmatrix}$$

Die inverse Darstellung der kartesischen Koordinaten durch die Flächenkoordinaten lautet

$$\begin{bmatrix} 1 \\ x_P \\ y_P \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix}$$

Die Koordinaten des aktuellen Punktes im kartesischen System lassen sich also durch das Multiplizieren der Koordinaten der Eckpunkte mit den entsprechenden Flächenkoordinaten berechnen.

$$\begin{aligned} x_P &= x_1\lambda_1 + x_2\lambda_2 + x_3\lambda_3 \\ y_P &= y_1\lambda_1 + y_2\lambda_2 + y_3\lambda_3 \end{aligned}$$

- Integrationsregel

Inbesondere kann man sich bei der Zusammenstellung der Matrizen für die Diskretisierung folgende Integrationsregel zunutze machen

$$\int_\Omega \lambda_1^a \lambda_2^b \lambda_3^c \ \mathrm{d}A \ = \ \frac{a\,!\ b\,!\ c\,!}{(a+b+c+2)!} \cdot 2A$$

Diese Regel gilt für Dreiecke mit geraden Kanten. Bei gekrümmten Seiten müßte auf numerische Integration zur Auswertung zurückgegriffen werden.

- Ableitungen

Für die partiellen Ableitungen der Flächenkoordinaten λ_i nach den kartesischen Koordinaten x, y gilt insbesondere :

$$\frac{\partial \lambda_i}{\partial x} = \frac{\alpha_i}{2A} \quad , \qquad \frac{\partial \lambda_i}{\partial y} = \frac{\beta_i}{2A}$$

10.1.2 Elementmatrizen für lineare Ansätze und verschiedene Methoden

Im folgenden sind die Elementmatrizen, die aus den verschiedenen FE-Methoden (siehe Kapitel 3) entstanden sind, zur Übersicht aufgeführt.
Zunächst seien der Vektor der linearen Ansatzfunktionen, die sogenannte B-Matrix der partiellen Ableitung der linearen Ansatzfunktionen, der Vektor des Geschwindigkeitsfeldes und ein zur Abkürzung eingeführter Ausdruck dargestellt.

$$\mathbf{N} = [N^1, N^2, N^3] = [\lambda^1, \lambda^2, \lambda^3]$$

$$\mathbf{B} = \begin{bmatrix} N^1_{,x} & N^2_{,x} & N^3_{,x} \\ N^1_{,y} & N^2_{,y} & N^3_{,y} \end{bmatrix} = \frac{1}{2A} \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ \beta_1 & \beta_2 & \beta_3 \end{bmatrix}$$

$$\boldsymbol{b} = [b_1, b_2]$$

$$\hat{\mathbf{B}} = \boldsymbol{b} \cdot \mathbf{B} = \frac{1}{2A}[(b_1\alpha_1 + b_2\beta_1), (b_1\alpha_2 + b_2\beta_2), (b_1\alpha_3 + b_2\beta_3)] = \frac{1}{2A}[\hat{B}_1, \hat{B}_2, \hat{B}_3]$$

Elementmatrizen aus dem Standard-Galerkin Verfahren

Es folgen die Elementmatrizen aus dem Standard-Galerkin Verfahren.

$$\mathbf{k} = \int_T \epsilon \, \mathbf{B}^t \cdot \mathbf{B} \; \mathrm{d}\Omega_e = \frac{\epsilon}{4A} \begin{bmatrix} \alpha_1\alpha_1 + \beta_1\beta_1 & ... & \text{symm.} \\ \alpha_1\alpha_2 + \beta_1\beta_2 & \alpha_2\alpha_2 + \beta_2\beta_2 & ... \\ \alpha_1\alpha_3 + \beta_1\beta_3 & \alpha_2\alpha_3 + \beta_2\beta_3 & \alpha_3\alpha_3 + \beta_3\beta_3 \end{bmatrix}$$

$$\alpha\mathbf{m} = \int_T \alpha \, \mathbf{N}^t \cdot \mathbf{N} \; \mathrm{d}\Omega_e = \frac{\alpha A}{12} \begin{bmatrix} 2 & ... & \text{symm.} \\ 1 & 2 & ... \\ 1 & 1 & 2 \end{bmatrix}$$

$$\mathbf{g} = \int_T \boldsymbol{N}^t \cdot \boldsymbol{b} \cdot \mathbf{B} \; \mathrm{d}\Omega_e = \frac{1}{6} \begin{bmatrix} \hat{B}_1 & \hat{B}_2 & \hat{B}_3 \\ \hat{B}_1 & \hat{B}_2 & \hat{B}_3 \\ \hat{B}_1 & \hat{B}_2 & \hat{B}_3 \end{bmatrix}$$

$$\boldsymbol{f} = \int_T \boldsymbol{N}^t f \; \mathrm{d}\Omega_e + \int_{\Gamma_N} \boldsymbol{N}^t t_N \; \mathrm{d}\Gamma_e = (\frac{A}{3} f + \frac{L}{2} t_N) \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Elementmatrizen aus der SDFEM

Bei der Stromlinien-Diffusions-Methode kommen zu den vorhergegangenen Matrizenanteilen die folgenden hinzu :

$$\mathbf{k}^{SD} = \int_T \delta \mathbf{B}^t \cdot \boldsymbol{b}^t \boldsymbol{b} \cdot \mathbf{B} \ \mathrm{d}\Omega_e = \int_T \delta \hat{\mathbf{B}}^t \cdot \hat{\mathbf{B}} \ \mathrm{d}\Omega_e = \frac{\delta}{4A} \begin{bmatrix} \hat{B}_1\hat{B}_1 & ... & \text{symm.} \\ \hat{B}_1\hat{B}_2 & \hat{B}_2\hat{B}_2 & ... \\ \hat{B}_1\hat{B}_3 & \hat{B}_2\hat{B}_3 & \hat{B}_3\hat{B}_3 \end{bmatrix}$$

$$\alpha\delta \ \mathbf{g}^t = \int_T \alpha\delta \ \mathbf{B}^t \cdot \boldsymbol{b}^t \cdot \boldsymbol{N} \ \mathrm{d}\Omega_e = \alpha\delta \ \frac{1}{6} \begin{bmatrix} \hat{B}_1 & \hat{B}_1 & \hat{B}_1 \\ \hat{B}_2 & \hat{B}_2 & \hat{B}_2 \\ \hat{B}_3 & \hat{B}_3 & \hat{B}_3 \end{bmatrix}$$

$$\boldsymbol{f}^{SD} = \int_T \delta \ \mathbf{B}^t \cdot \boldsymbol{b}^t f \ \mathrm{d}\Omega_e = f \frac{\delta}{2A} \begin{bmatrix} \hat{B}_1 \\ \hat{B}_2 \\ \hat{B}_3 \end{bmatrix}$$

Elementmatrizen aus der Shock-Capturing Modifikation

Durch die Erweiterung der Testfunktion im Falle der Shock-Capturing Modifikation kommen in der schwachen Formulierung des Problems vier weitere den SDFEM-Termen entsprechende Terme hinzu. In diesen treten an die Stelle des Faktors δ und des Geschwindigkeitsvektors $\boldsymbol{b}$ die modifizierten Analoga $\bar{\delta}$ und $\bar{\boldsymbol{b}}$. Die Elementmatrizenanteile nehmen mit der entsprechend zur oben getroffenen abkürzenden Schreibweise

$$\hat{\mathbf{B}}_{mod} = \bar{\boldsymbol{b}} \cdot \mathbf{B} = \frac{1}{2A}[(\bar{b}_1\alpha_1 + \bar{b}_2\beta_1), (\bar{b}_1\alpha_2 + \bar{b}_2\beta_2), (\bar{b}_1\alpha_3 + \bar{b}_2\beta_3)] = \frac{1}{2A}[\hat{B}_1^{mod}, \hat{B}_2^{mod}, \hat{B}_3^{mod}]$$

die folgende Form an :

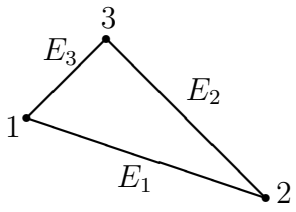
$$\mathbf{k}^{SC} = \int_T \bar{\delta} \ \mathbf{B}^t \cdot \bar{\boldsymbol{b}}^t \boldsymbol{b} \cdot \mathbf{B} \ \mathrm{d}\Omega_e = \int_T \bar{\delta} \ \hat{\mathbf{B}}^t_{mod} \cdot \hat{\mathbf{B}} \ \mathrm{d}\Omega_e = \frac{\bar{\delta}}{4A} \begin{bmatrix} \hat{B}_1\hat{B}_1^{mod} & ... & \text{symm.} \\ \hat{B}_1\hat{B}_2^{mod} & \hat{B}_2\hat{B}_2^{mod} & ... \\ \hat{B}_1\hat{B}_3^{mod} & \hat{B}_2\hat{B}_3^{mod} & \hat{B}_3\hat{B}_3^{mod} \end{bmatrix}$$

$$\alpha\bar{\delta} \ \mathbf{g}^{SCt} = \int_T \alpha\bar{\delta} \ \mathbf{B}^t \cdot \bar{\boldsymbol{b}}^t \cdot \boldsymbol{N} \ \mathrm{d}\Omega_e = \alpha\bar{\delta} \ \frac{1}{6} \begin{bmatrix} N^1\hat{B}_1^{mod} & N^2\hat{B}_1^{mod} & N^3\hat{B}_1^{mod} \\ N^1\hat{B}_2^{mod} & N^2\hat{B}_2^{mod} & N^3\hat{B}_2^{mod} \\ N^1\hat{B}_3^{mod} & N^2\hat{B}_3^{mod} & N^3\hat{B}_3^{mod} \end{bmatrix}$$

$$\boldsymbol{f}^{SC} = \int_T \bar{\delta} \ \mathbf{B}^t \cdot \boldsymbol{b}^t f \ \mathrm{d}\Omega_e = f \frac{\bar{\delta}}{2A} \begin{bmatrix} \hat{B}_1^{mod} \\ \hat{B}_2^{mod} \\ \hat{B}_3^{mod} \end{bmatrix}$$

10.1.3 Elementmatrizen des lokalen Neumann-Hilfsproblems

In diesem Abschnitt sollen die Elementmatrizen, die bei der Lösung des lokalen Neumann-Hilfsproblems entstehen, dargestellt werden (siehe Kapitel 5.4.1).



Der Vektor der Ansatzfunktionen setzt sich aus den sogenannten bubble-Funktionen, die auf den Elementkanten und im Elementinneren definiert sind, zusammen.

$$\mathbf{N} = [N^1, N^2, N^3, N^4] = [b_{E_1}, b_{E_2}, b_{E_2}, b_T] = \begin{bmatrix} 4\lambda_1\lambda_2 \\ 4\lambda_2\lambda_3 \\ 4\lambda_3\lambda_1 \\ 27\lambda_1\lambda_2\lambda_3 \end{bmatrix}^t$$

Entsprechend stellt sich die B-Matrix der partiellen Ableitung der Ansatzfunktionen wie folgt dar.

$$\mathbf{B} = \begin{bmatrix} N^1_{,x} & N^2_{,x} & N^3_{,x} & N^4_{,x} \\ N^1_{,y} & N^2_{,y} & N^3_{,y} & N^4_{,y} \end{bmatrix} = \frac{1}{2A} \begin{bmatrix} B_{11} & B_{12} & B_{13} & B_{14} \\ B_{21} & B_{22} & B_{23} & B_{24} \end{bmatrix}$$

Dabei sind mit $i = 1, 2, 3$ und $i + 1$ aus zyklischer Vertauschung

$$\begin{aligned} \frac{1}{2A} B_{1i} = N^i_{,x} &= \frac{2}{A}(\lambda_i \alpha_{i+1} + \alpha_i \lambda_{i+1}) \\ \frac{1}{2A} B_{2i} = N^i_{,y} &= \frac{2}{A}(\lambda_i \beta_{i+1} + \beta_i \lambda_{i+1}) \\ \frac{1}{2A} B_{14} = N^4_{,x} &= \frac{27}{2A}(\alpha_1\lambda_2\lambda_3 + \alpha_2\lambda_3\lambda_1 + \alpha_3\lambda_1\lambda_2) \\ \frac{1}{2A} B_{24} = N^4_{,y} &= \frac{27}{2A}(\beta_1\lambda_2\lambda_3 + \beta_2\lambda_3\lambda_1 + \beta_3\lambda_1\lambda_2). \end{aligned}$$

Aus dem Diffusionsanteil entsteht die symmetrische Elementmatrix $\mathbf{k}$

$$\mathbf{k} = \frac{\epsilon}{120A} \begin{bmatrix} k_{11} & k_{12} & k_{13} & k_{14} \\ ... & k_{22} & k_{23} & k_{24} \\ ... & ... & k_{33} & k_{34} \\ \text{symm.} & ... & ... & k_{44} \end{bmatrix}.$$

Mit $i = 1, 2, 3$ und $i + 1$, $i + 2$ aus zyklischer Vertauschung und den Abkürzungen

$$\begin{aligned} \alpha_q &= \alpha_1^2 + \alpha_2^2 + \alpha_3^2 \\ \beta_q &= \beta_1^2 + \beta_2^2 + \beta_3^2 \\ \alpha_p &= \alpha_1\alpha_2 + \alpha_2\alpha_3 + \alpha_3\alpha_1 \\ \beta_p &= \beta_1\beta_2 + \beta_2\beta_3 + \beta_3\beta_1 \end{aligned}$$

gilt für die Komponenten von $\mathbf{k}$

$$\begin{aligned} k_{ii} &= 80(\alpha_i^2 + \alpha_i\alpha_{i+1} + \alpha_{i+1}^2 + \beta_i^2 + \beta_i\beta_{i+1} + \beta_{i+1}^2) \\ k_{44} &= 243(\alpha_q + \alpha_p + \beta_q + \beta_p) \end{aligned}$$

$$\begin{aligned} k_{12} &= 40(\alpha_p + \alpha_3\alpha_1 + \alpha_2^2 + \beta_p + \beta_3\beta_1 + \beta_2^2) \\ k_{13} &= 40(\alpha_p + \alpha_2\alpha_3 + \alpha_1^2 + \beta_p + \beta_2\beta_3 + \beta_1^2) \\ k_{23} &= 40(\alpha_p + \alpha_1\alpha_2 + \alpha_3^2 + \beta_p + \beta_1\beta_2 + \beta_3^2) \\ k_{i4} &= 108(\alpha_p + \alpha_i^2 + \alpha_{i+1}^2 + \beta_p + \beta_i^2 + \beta_{i+1}^2) \end{aligned}$$

Die unsymmetrische Elementmatrix $\mathbf{g}$ entsteht aus dem Term der Konvektion

$$\mathbf{g} = \begin{bmatrix} g_{11} & g_{12} & g_{13} & g_{14} \\ g_{21} & g_{22} & g_{23} & g_{24} \\ g_{31} & g_{32} & g_{33} & g_{34} \\ g_{41} & g_{42} & g_{43} & g_{44} \end{bmatrix},$$

dabei sind die einzelnene Komponenten

$$\begin{aligned} g_{ii} &= \frac{4}{15}\left(b_1(\alpha_i + \alpha_{i+1}) + b_2(\beta_i + \beta_{i+1})\right) \\ g_{44} &= \frac{81}{140}\left(b_1(\alpha_1 + \alpha_2 + \alpha_3) + b_2(\beta_1 + \beta_2 + \beta_3)\right) \end{aligned}$$

$$\begin{aligned} g_{12} &= \frac{2}{15}\left(b_1(\alpha_2 + 2\alpha_3) + b_2(\beta_2 + 2\beta_3)\right) \\ g_{13} &= \frac{2}{15}\left(b_1(\alpha_1 + 2\alpha_3) + b_2(\beta_1 + 2\beta_3)\right) \end{aligned}$$

$$\begin{aligned} g_{21} &= \frac{2}{15}\left(b_1(\alpha_2 + 2\alpha_1) + b_2(\beta_2 + 2\beta_1)\right) \\ g_{23} &= \frac{2}{15}\left(b_1(\alpha_3 + 2\alpha_1) + b_2(\beta_3 + 2\beta_1)\right) \end{aligned}$$

$$\begin{aligned} g_{31} &= \frac{2}{15}\left(b_1(\alpha_1 + 2\alpha_2) + b_2(\beta_1 + 2\beta_2)\right) \\ g_{32} &= \frac{2}{15}\left(b_1(\alpha_3 + 2\alpha_2) + b_2(\beta_3 + 2\beta_2)\right) \end{aligned}$$

$$\begin{aligned} g_{i4} &= \frac{9}{30}\left(b_1(\alpha_i + \alpha_{i+1} + 2\alpha_{i+2}) + b_2(\beta_i + \beta_{i+1} + 2\beta_{i+2})\right) \\ g_{4i} &= \frac{9}{30}\left(b_1(\alpha_i + \alpha_{i+1}) + b_2(\beta_i + \beta_{i+1})\right) \end{aligned}$$

Der Adsorptionsterm liefert die symmetrische Elementmatrix $\mathbf{m}$

$$\alpha\mathbf{m} = \alpha A \begin{bmatrix} 16/90 & 8/90 & 8/90 & 12/70 \\ ... & 16/90 & 8/90 & 12/70 \\ ... & ... & 16/90 & 12/70 \\ \text{symm.} & ... & ... & 81/280 \end{bmatrix}.$$

Aus diesen Elementmatrizen setzt sich die Systemmatrix $\mathbf{A}$ des zu lösenden Gleichungssystems zusammen:

$$\mathbf{A} = \mathbf{k} + \mathbf{g} + \alpha\mathbf{m}.$$

Der Vektor der rechten Seite wird aus dem Vektor des Quellterms des Elementresiduums und dem Vektor der Randbedingungen, d.h. den Residuen auf den Elementkanten, gebildet.

$$\boldsymbol{F} = \int_T Res_T(u_h)\boldsymbol{N}^t \ \mathrm{d}\Omega_e + \sum_{E\subset\partial T} \int_E Res_E(u_h)\boldsymbol{N}^t \ \mathrm{d}E$$

Unter der Annahme, daß $Res_T(u_h)$ und $Res_E(u_h)$ in T konstant sind, gilt

$$\boldsymbol{F} = \frac{A}{3} Res_T(u_h) \begin{bmatrix} 1 \\ 1 \\ 1 \\ 27/20 \end{bmatrix} + \frac{2L}{3} Res_E(u_h) \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix}.$$

Dieses 4×4 Gleichungssystem $\mathbf{A}\boldsymbol{v}_T = \mathbf{F}$ zur Bestimmung von $\boldsymbol{v}_T$ kann leicht mit einem Gauß-Algorithmus oder der Cramerschen Regel gelöst werden.

10.2 Anhang 2

10.2.1 Lokales Minimierungsproblem zur Netzglättung

Wie in Kapitel 7 beschrieben wird zur Glättung eines Netzes eine äußere Gauß-Seidel ähnliche Iteration vorgenommen. Dabei wird für jeden Knoten $\boldsymbol{x}_i = (x_i, y_i)$ ein lokales Minimierungsproblems gelöst, während alle anderen Knoten der Triangulierung festgehalten werden. Das lokale Minimierungsproblem für den Knoten $\boldsymbol{x}_i$ lautet

$$\min_{x_i,y_i} F := \min_{x_i,y_i} \sum_{T\in\omega_{\boldsymbol{x}_i}} F(T) := \min_{x_i,y_i} \sum_{T\in\omega_{\boldsymbol{x}_i}} \frac{s(T)}{q(T)}.$$

Dabei bezeichnet $\omega_{\boldsymbol{x}_i}$ die Menge der Elemente, die den Knoten $\boldsymbol{x}_i$ enthalten. Der Zähler des Quotienten enthält die Funktion

$$s(T) = (\boldsymbol{l}_1^t \mathbf{M}_T \boldsymbol{l}_1)^2 + (\boldsymbol{l}_2^t \mathbf{M}_T \boldsymbol{l}_2)^2 + (\boldsymbol{l}_3^t \mathbf{M}_T \boldsymbol{l}_3)^2 \quad \text{mit} \quad \mathbf{M}_T = -\frac{1}{2}\begin{bmatrix} u_{xx} & u_{xy} \\ u_{xy} & u_{yy} \end{bmatrix} = \text{const.}$$

Der Nenner entspricht der Qualitätsfunktion des Dreiecks

$$q(T) = \frac{4\sqrt{3}A}{|\boldsymbol{l}_1|^2 + |\boldsymbol{l}_2|^2 + |\boldsymbol{l}_3|^2}.$$

Zur Lösung des Minimierungsproblems für jeden Knoten wird das Newton-Verfahren herangezogen. Bei dieser inneren Iteration ist die Nullstelle des sogenannten Residuums, d.h. der ersten partiellen Ableitungen von F nach dem Knoten $\boldsymbol{x}_i$,

$$\begin{aligned} \boldsymbol{f} := \partial_{\boldsymbol{x}_i} F &= \sum_{T\in\omega_{\boldsymbol{x}_i}} \boldsymbol{f}(T) \\ &= \sum_{T\in\omega_{\boldsymbol{x}_i}} \frac{1}{q}\left(\partial_{\boldsymbol{x}_i} s(T) - F(T)\partial_{\boldsymbol{x}_i} q(T)\right) \end{aligned}$$

zu bestimmen. Dafür wird die Hesse-Matrix von F, d.h. die zweite partielle Ableitung von F benötigt

$$\partial_{\boldsymbol{x}_i}\boldsymbol{f} = \sum_{T\in\omega_{\boldsymbol{x}_i}} \frac{1}{q}\left(\partial^2_{\boldsymbol{x}_i\boldsymbol{x}_i} s(T) - F(T)\partial^2_{\boldsymbol{x}_i\boldsymbol{x}_i} q(T) - \partial_{\boldsymbol{x}_i} q(T) \otimes \boldsymbol{f}(T) - \boldsymbol{f}(T) \otimes \partial_{\boldsymbol{x}_i} q(T)\right).$$

Die auftretenden partiellen Ableitungen der Funktionen $s(T)$ und $q(T)$ werden im folgenden bereit gestellt. Zur Vereinfachung der Terme werden werden zunächst einige Abkürzungen eingeführt:

$$\begin{aligned} s(T) &= \sum_{j=1}^{3} \left(\boldsymbol{l}_j^t \mathbf{M} \boldsymbol{l}_j\right)^2 = \sum_{j=1}^{3} s_j^2 \\ p(T) &= \sum_{j=1}^{3} \boldsymbol{l}_j^t \boldsymbol{l}_j. \end{aligned}$$

Mit j sei in jedem Dreieck T die lokale Knotennumerierung angegeben, also 1, 2 oder 3, und $j+1$, $j+2$ seien die entsprechenden zugehörigen zyklischen Knotennummern. Damit ergeben sich die 1. und 2. partielle Ableitung von $s(T)$ nach dem Knoten $\boldsymbol{x}_j$ zu

$$\begin{aligned}\partial_{\boldsymbol{x}_j} s(T) &= 4\left(s_{j+1}\boldsymbol{l}_{j+1}^t - s_{j+2}\boldsymbol{l}_{j+2}^t\right)\mathbf{M}_T \\ \partial^2_{\boldsymbol{x}_j\boldsymbol{x}_j} s(T) &= 8\left(\mathbf{M}_T\boldsymbol{l}_{j+1}\otimes\boldsymbol{l}_{j+1}^t\mathbf{M}_T + \mathbf{M}_T\boldsymbol{l}_{j+2}\otimes\boldsymbol{l}_{j+2}^t\mathbf{M}_T + \frac{1}{2}\left(s_{j+1}+s_{j+2}\right)\mathbf{M}_T\right).\end{aligned}$$

Mit den Bezeichnungen $\boldsymbol{x}_S$ für den Schwerpunkt des Dreiecks T, $\mathbf{1}$ für den Einheitstensor und $\mathbf{W}$ für den schiefsymmetrische Einheitstensor

$$\mathbf{W} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

gilt für die partiellen Ableitungen der Qualitätsfunktion

$$\begin{aligned}\partial_{\boldsymbol{x}_j} q(T) &= \frac{1}{p(T)}\left(2\sqrt{3}\;\boldsymbol{l}_j^t\mathbf{W} - 6q(T)(\boldsymbol{x}_j - \boldsymbol{x}_S)\right) \\ \partial^2_{\boldsymbol{x}_j\boldsymbol{x}_j} q(T) &= \frac{6}{p(T)}\left((\boldsymbol{x}_S-\boldsymbol{x}_j)\otimes\partial_{\boldsymbol{x}_j}q(T) + \partial_{\boldsymbol{x}_j}q(T)\otimes(\boldsymbol{x}_S-\boldsymbol{x}_j) - \frac{2}{3}q(T)\mathbf{1}\right).\end{aligned}$$

Mit Kenntnis dieser Größen läßt sich das Newton-Verfahren anwenden. In klassischer Form stellt es sich wie folgt dar:

Gegeben seien $\boldsymbol{f}$ und eine Näherungslösung $\boldsymbol{x}_0$.
Gesucht wird $\boldsymbol{x}$ mit $\boldsymbol{f}(\boldsymbol{x}) = \mathbf{0}$.

Iteration : $k = 0, 1, 2, ...$
(bis $k = k_{max}$ oder $|\boldsymbol{f}| \leq$ tol)

Bestimme $\boldsymbol{\Delta x} = \boldsymbol{\Delta x}^k$ durch Lösen
des linearen Gleichungssystems

$$\partial_{\boldsymbol{x}_i}\boldsymbol{f}\,\boldsymbol{\Delta x} = -\boldsymbol{f}$$

und setze

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \boldsymbol{\Delta x}$$

Tabelle 10.1: Newton-Verfahren

Das obige Newton-Verfahren konvergiert in der Regel in der Nähe von Lösungen quadratisch. Wird der Startwert allerdings in zu weiter Entfernung der Lösung gewählt, kann

die Folge divergieren. Um dies zu vermeiden, wird in das klassische Newton-Verfahren eine Dämpfung der Suchrichtung $\boldsymbol{\Delta x}$, in welcher die Funktion $\boldsymbol{f}$ abnimmt, eingeführt. Damit wird nur ein Teil des Schrittes in die Richtung von $\boldsymbol{\Delta x}$ ausgeführt. Bei passender Wahl des Dämpfungsparameters ψ erwartet man eine Verbesserung des Verfahrens.

Gegeben seien $\boldsymbol{f}$ und eine Näherungslösung $\boldsymbol{x}_0$.
Gesucht wird $\boldsymbol{x}$ mit $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$.

Setze $k = 0$

1. Bestimme $\boldsymbol{\Delta x} = \boldsymbol{\Delta x}^k$ durch Lösen des linearen Gleichungssystems

$$\partial_{\boldsymbol{x}_i} \boldsymbol{f} \boldsymbol{\Delta x} = -\boldsymbol{f}$$

2. setze $\boldsymbol{x}^\star = \boldsymbol{x}^k + \psi \boldsymbol{\Delta x}$ und prüfe, ob

$$\|\boldsymbol{f}(\boldsymbol{x}^\star)\| \leq \left(1 - \frac{\psi}{4}\right) \|\boldsymbol{f}(\boldsymbol{x}^k)\|$$

ja : $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \psi \boldsymbol{\Delta x}$ und gehe zu 3.
nein : $\psi = \psi/2$; wenn $\psi \geq 2^{-10}$ gehe zu 2., sonst Abbruch

3. Falls $k = k_{max}$ oder $\|\boldsymbol{f}(\boldsymbol{x}^{k+1})\| \leq \text{tol} \;\rightarrow\;$ Ende,
sonst neue Iteration : $k = k + 1$ und gehe zu 1.

Tabelle 10.2: Gedämpftes Newton-Verfahren

10.3 Notation

Physikalische Größen

u	Konzentration eines gelösten Stoffes
$\boldsymbol{b}$	Geschwindigkeitsfeld
f	Quellterm
D, ϵ	konstanter Diffusionskoeffizient
$\mathbf{D}$	Dispersionstensor (zweiter Stufe)
N	Massenstromdichte
n_e	effektive Porosität
$\boldsymbol{q}$	spezifischer Durchfluß
ρ	Dichte
ϕ	äquvalente Piezometerhöhe
$\mathbf{K}$	Permeabilitätsmatrix
α, K_D	Koeffizient linearer Adsorptionsisotherme
Pe	$=\|\boldsymbol{b}\|L/\epsilon$ Peclet-Zahl

Finite Element Methoden

Ω	offene Menge im $\mathbb{R}^n$
$\Gamma = \partial\Omega$	polygonale Berandung von Ω
Γ_D	Teil des Randes, auf dem Dirichlet-Bedingungen (u_D) vorgegeben sind
Γ_N	Teil des Randes, auf dem Neumann-Bedingungen (t_N) vorgegeben sind
$\boldsymbol{n}$	nach außen gerichteter Normaleneinheitsvektor
$\mathcal{T}_h$	Zerlegung von Ω
T	Dreieckelement in $\mathcal{T}_h$
$L^2(\Omega)$	Raum der über Ω quadrat-integrierbaren Funktionen
$(\cdot,\cdot)$	innere L^2-Produkt über Ω
$H^1(\Omega)$	Sobolev-Raum von L^2-Funktionen mit quadrat-integrierbaren Ableitungen bis zur Ordnung 1
$H_0^1(\Omega)$	Unterraum von H^1 mit verallgemeinerten Nullrandbedingungen
$\|\cdot\|_1$	Sobolev-Norm der Ordnung 1
$\|\cdot\|_1$	Sobolev-Seminorm der Ordnung 1
$a(\cdot,\cdot)$	Bilinearform
$l(\cdot)$	Linearform, lineares Funktional
Δ	Laplace-Operator
$\partial_{\boldsymbol{n}}$	partielle Ableitung in Richtung der äußeren Normalen $\partial/\partial\boldsymbol{n}$
∇f	$(\partial f/\partial x_1, \partial f/\partial x_2, ..., \partial f/\partial x_n)$
δ_e	SUPG-Faktor, intrinsic time scale
$\zeta(\mathrm{Pe}^h)$	Funktion der lokalen Peclet-Zahl
h_e	charakteristische Länge in einem Dreieck
$\bar{\delta}$	Shock-Capturing modifizierter Faktor
$\bar{\boldsymbol{b}}$	Shock-Capturing modifizierter Geschwindigkeitsvektor

Iterative Verfahren

$\mathbf{A}$	Systemmatrix $\in R^{n\times n}$
$\boldsymbol{x}$	Lösungsvektor
$\boldsymbol{b}$	Vektor der rechten Seite $\in R^n$
$\boldsymbol{r}$	Residuenvektor
$\boldsymbol{p}$	Richtungvektor
$\mathbf{W}$	Vorkonditionierungsmatrix
$\mathbf{L}, \mathbf{U}, \mathbf{D}$	Anteile der zerlegten Matrix $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{U}$
	$\mathbf{L}$ unterer und $\mathbf{U}$ oberer Dreiecksanteil, $\mathbf{D}$ Diagonalanteil
ω	Relaxationsparameter
κ	Konditionszahl
$\mathbf{R}$	Fehlermatrix

Fehlerschätzer

e	Fehler $u - u_h$
TOL	gegebene Toleranz für den Fehler e
η	Fehlerschätzer
$[\cdot]_E$	Sprung über die Kante E
$\boldsymbol{n}_E$	Normaleneinheitsvektor auf der Kante E
b_T, b_E	Abschneide- oder bubble-Funktionen
λ_i	Flächenkoordinaten
I_h	Interpolationsoperator nach Clément
$\|\!\|\!\| \cdot \|\!\|\!\|$	Energienorm
$\eta_{Z,T}$	Z^2-Indikator auf Dreieck T
$\|T\|$	Fläche eines Dreiecks T
$G(u_h)$	Näherung für ∇u durch Mittelung
$\eta_{R,T}$	residueller Fehlerschätzer
$Res_T(u_h)$	Elementresiduum
α_T	Faktor vor Elementresiduum
$Res_E(u_h)$	Kantenresiduum
β_E	Faktor vor Kantenresiduum
$\eta_{N,T}$	Fehlerschätzer basierend auf lokalem Neumann Problem
v_T	FE-Approximation des lokalen Konvektions-Diffusionsproblems
u_T	Lösung des lokalen Konvektions-Diffusionsproblems

Markierungsstrategien

c_{ref}	Verfeinerungsschwelle
η_{max}	maximaler Wert $\max_T\{\eta_T\}$
p	abzuschneidender oberster Prozentsatz in Strategie 2
N_k	Anzahl der Elemente in $\mathcal{T}_k$
θ_1, θ_2	konstante Parameter in Strategie 4
c, β	Konstanten in Strategie 5

Netzglättung

$q(T)$	Qualitätsfunktion des Dreiecks T
A	Fläche eines Dreiecks
$\boldsymbol{l}_i$	Kantenvektoren
$\mathbf{M}_T$	konstante Matrix der 2.Ableitungen von u auf T
$\bar{b}_i$	bubble-Funktionen
e_T	Fehlerindikator auf T
$s(T)$	Funktion des Zählers des lokalen Minimierungsproblems
Υ^2	Quadrat der zu minimierenden Funktion
dx	Gesamtverschiebungen innerhalb des Netzes

Instationäre Erweiterung

$[0, T]$	gegebenes Zeitintervall
$\dot{u} = \partial u/\partial t$	Zeitableitung der Funktion u
$\theta \in [0, 1]$	Parameter des θ-Schemas
Δt	Zeitschrittweite
Co	$= \lvert\boldsymbol{b}\rvert \Delta t/h$ Courant-Zahl
Ne	$= \epsilon \Delta t/h^2$ Neumann-Zahl

Literaturverzeichnis

[1] E.Bänsch : An adaptive finite-element strategy for the three-dimensional time-dependent Navier-Stokes equations, J. Comput. Appl. Math. 36, 3-28, 1991

[2] R.E.Bank, R.K.Smith : Mesh smoothing using a posteriori error estimates, SIAM J. Numer. Anal., Vol. 34, No. 3, 979-997, 1997

[3] R.E.Bank, A.Weiser : Some a posteriori error estimators for elliptic partial differential equations, Math. Comput. 44, 283-301, 1985

[4] D.Braess : Finite Elemente, Springer Verlag, 1992

[5] J.Bear : Dynamics of fluids in porous media, American Elsevier, 1972

[6] A.N.Brooks, T.J.R.Hughes : Streamline Upind/ Petrov-Galerkin formulations for convective dominated flows with particular emphasis on the incompressible Navier-Stokes equations, Comput. Meths. Appl. Mech. Engrg., Vol. 32, 199-259, 1982

[7] G.F.Carey, J.T.Oden : Finite elements : Fluid mechanics, The Texas Finite Element Series, Vol. VI, Prentice Hall, 1986

[8] I.Christie, D.F.Griffiths, A.R.Mitchell, O.C.Zienkiewicz : Finite element methods for second order differential equations with significant first derivatives, J. for Numer. Meths. in Engrg., Vol. 10, 1389-1396, 1976

[9] P.G.Ciarlet : The finite element method for elliptic problems, North-Holland, 2.Auflage 1979

[10] P.Clément : Approximation by finite element functions using local regularization, RAIRO Anal. Numér. 9, 77-84, 1975

[11] R. Codina : A finite element formulation for the numerical solution of the convection-diffusion equation, Monograph CIMNE no. 14, Jan. 1993

[12] T. De Mulder : Stabilized finite element methods (SUPG, GLS,..) for incompressible flows, Lecture Series 1997-02, von Karman Intitute for Fluid Dynamics, March 1997

[13] K.Eriksson, C.Johnson : Adaptive streamline diffusion finite element methods for convection-diffusion problems, Publ. 1990-18, ISSN 0347-2809, Chalmers University of Technology, Dept. of Mathematics, Göteborg, 1990

[14] K.Eriksson, D.Estep, P.Hansbo, C.Johnson : Computational differential equations, Acta Numerica, Studentlitteratur, Lund, 1996

[15] K.Eriksson, D.Estep, P.Hansbo, C.Johnson : Introduction to Adaptive methods for differential equations, Acta Numerica, 105-158, 1995

[16] Faber, Manteuffel : Necessary and sufficient conditions for the existence of a conjugate gradient method, SIAM J. Numer. Anal., Vol. 21, No. 2, 1984

[17] L.P.Franca, S.L.Frey, T.J.R.Hughes : Stabilized finite element methods : I. Application to the advective-diffusive model, Comput. Meths. Appl. Mech. Engrg., Vol. 95, 253-376, 1992

[18] A.C.Galeao, E.G.Dutra do Carmo : A consistent approximate upwind Petrov-Galerkin method for convection-dominated problems, Comput. Meths. Appl. Mech. Engrg., Vol. 68, 83-95, 1988

[19] S.Gutsch : Ein Vergleich CG-ähnlicher Verfahren zur Lösung indefiniter Probleme, Diplomarbeit, Chr.-Albrechts-Universität Kiel, 1994

[20] W.Hackbusch : Theorie und Numerik elliptischer Differentialgleichungen, Verlag Teubner, 1986

[21] W.Hackbusch : Iterative lösung großer schwachbesetzter Gleichungssysteme, Verlag Teubner, 2.Aufl., 1993

[22] P.Hansbo : Adaptivity and streamline diffusion procedures in the finite element method, Thesis, Publication 89:2, Chalmers University of Technology, Dept. of Structural Mechanics, Göteborg, 1989

[23] P.Hansbo : The characteristic streamline diffusion method for convection-diffusion problems, Comput. Meths. Appl. Mech. Engrg., Vol. 96, 239-253, 1992

[24] I.Harari, T.J.R.Hughes : What are C and h ?: Inequalities for an analysis and design of finite element methods, Comput. Meths. Appl. Mech. Engrg., Vol. 97, 157-192, 1992

[25] T.J.R.Hughes : The finite element method. Linear static and dynamic analysis, Prentice Hall, 1987

[26] T.J.R.Hughes : Finite element methods for fluids, in Proc. der AGARD-VKI LS "Unstructured grid methods for advection dominated flows", AGARD Report 787, 2.1-2.22, Mai 1992

[27] T.J.R.Hughes, M.Mallet, A.Mizukami : A new finite element formulation for computational fluid dynamics : II. Beyond SUPG, Comput. Meths. Appl. Mech. Engrg., Vol. 54, 341-355, 1986

[28] T.J.R.Hughes, M.Mallet : A new finite element formulation for computational fluid dynamics : III. The generalized streamline operator for multidimensional advective-diffusive systems, Comput. Meths. Appl. Mech. Engrg., Vol. 58, 305-328, 1986

[29] T.J.R.Hughes, L.P.Franca, M.Mallet : A new finite element formulation for computational fluid dynamics : VI. Convergence analysis of the generalized SUPG formulation for linear time-dependent multidimensional advective-diffusive systems, Comput. Meths. Appl. Mech. Engrg., Vol. 63, 97-112, 1987

[30] T.J.R.Hughes, A.N.Brooks : A theoretical framework for Petrov-Galerkin methods with discontinuous weighting functions : application to the streamline upwind procedure, Finite Elements in Fluids, Vol. 4, 46-65, 1982

[31] T.J.R.Hughes, A.N.Brooks : A multi-dimensional upwind sheme with no crosswind diffusion, in T.J.R.Hughes (ed.) : FEMs for convection dominated flows, Vol. 34, 19-35, ASME AMD, New York, 1979

[32] T.J.R.Hughes, T.E.Tezduyar : Finite element methods for first-order hyperbolic systems with particular emphasis on the compressible Euler equations, Comput. Meths. Appl. Mech. Engrg., Vol. 45, 217-284, 1984

[33] V.John : A comparison of some error estimators for convection-diffusion problems on a parallel computer, Preprint Nr. 12, Fakultät für Mathematik der Universität Magdeburg, 1994

[34] C.Johnson : Numerical solution of partial differential equations by the finite element method, Studentlitteratur, 1987

[35] C.Johnson : Adaptive finite element methods for diffusion and convection problems, Comput. Meths. Appl. Mech. Engrg., Vol. 82, 301-322, 1990

[36] C.Johnson, U.Nävert, J.Pitkäranta : Finite element methods for linear hyperbolic problems, Comput. Meths. Appl. Mech. Engrg., Vol. 45, 285-312, 1984

[37] C.Johnson, A.H.Schatz, L.B.Wahlbin : Crosswind smear and pointwise errors in streamline diffusions finite element methods, Math. Comput., Vol. 49, Nr. 179, 25-38, 1987

[38] C.Johnson, U.Nävert : An analysis of some finite element methods for advection-diffusion problems, in S.Axelsson, L.S.Frank, A.van der Sluis (eds.) : Analytical and Numerical Approaches to Asymptotic Problems in Analysis, North-Holland, 99-116, 1981

[39] C.Johnson, J.Pitkäranta : An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation, Math. Comput., Vol. 46, 1-26, 1986

[40] C.Johnson, A.Szepessy : On the convergence of a finite element method for a hyperbolic conservation law, Math. Comput., Vol.49, Nr.180, 427-444, 1987

[41] C.Johnson, A.Szepessy, P.Hansbo : On the convergence of shockcapturing streamline diffusion methods for hyperbolic conservation laws, Preprint 1987:21, Chalmers University of Technology, Math. Dept., Göteborg, 1987

[42] R.Kaiser : Fehlerindikatoren für die adaptive Netzgenerierung zur Lösung der Transportgleichung, Bericht, Institut für Strömungsmechanik und Elektron. Rechnen im Bauwesen, Universität Hannover, Juli 1996

[43] W.Kinzelbach : Numerische Methoden zur Modellierung des Transports von Schadstoffen im Grundwasser, Oldenbourg Verlag, 2.Aufl., 1992

[44] O.Kolditz : Stoff- und Wärmetransport im Kluftgestein, Bericht Nr. 47/1996, ISSN 0177-9028, Institut für Strömungsmechanik und Elektron. Rechnen im Bauwesen, Universität Hannover, 1996

[45] J.Medina, M.Picasso, J. Rappaz : Error estimates and adaptive finite elements for nonlinear diffusion-convection problems, Math. Models and Meth. in Appl. Sci. 6, 689-712, 1996

[46] A.Mizukami : An implementation of the Streamline-Upwind/Petrov-Galerkin method for linear triangular elements, Comput. Meths. Appl. Mech. Engrg., Vol. 49, 357-364, 1985

[47] A.Mizukami, T.J.R.Hughes : A Petrov-Galerkin finite element method for convection-diffusion dominated flows : an accurate upwinding technique for satisfying the maximum principle, Comput. Meths. Appl. Mech. Engrg., Vol. 50, 181-193, 1985

[48] K.W.Morton : Numerical solutions of convection-diffusion problems, Chapman & Hall, 1996

[49] U.Nävert : A finite element method for convection-diffusion problems, Thesis, Chalmers University of Technology, Dept. of Computer Sciences, Göteborg, 1982

[50] W.H.Raymond, A.Garder : Selective damping in a Galerkin method for solving wave problems with variable grids, Monthly Weather Review, Vol. 104, 1583-1591, 1976

[51] H.R.Schwarz : Methode der Finiten Elemente, Verlag Teubner, 3.Aufl., 1991

[52] S.W.Sloan : A fast algorithm for constructing Delaunay triangulations in the plane, Adv. Eng. Software, Vol. 9, No.1, 34-55, 1987

[53] A.Szepessy : A streamline diffusion finite element method for conservation laws, Technical report, Chalmers University of Technology, Math. Dept., Göteborg, 1988

[54] T.E.Tezduyar, S.Mittal, S.E.Ray, R.Shih : Incompressible flow computations with stabilized bilinear and linear equal-order interpolation velocity-pressure elements, Comput. Meths. Appl. Mech. Engrg., Vol. 95, 221-242, 1992

[55] V.Kalro, S.Alibadi, W.Gerrard, T.E.Tezduyar : Parallel finite element simulations or large ram-air parachutes, Preprint 96-003, AHPCRC / University of Minnesota, 1996

[56] T.E.Tezduyar, D.K.Ganjoo : Petrov-Galerkin formulations with weighting functions dependent upon spatial and temporal discretization : application to transient convection-diffusion problems, Comput. Meths. Appl. Mech. Engrg., Vol. 59, 49-71, 1986

[57] R.Verfürth : *A posteriori* error estimation and adaptive mesh-refinement techniques, J.Comput.Appl.Math. 50, 67-83, 1994

[58] R.Verfürth : Robust *a posteriori* error estimators for the singularly perturbed reaction-diffusion equation, Numer. Math., Vol. 78, 479-493, 1998

[59] R.Verfürth : *A posteriori* error estimation and adaptive mesh-refinement techniques, Teubner-Wiley, Stuttgart, 1996

[60] R.Verfürth : *A posteriori* error estimators for convection-diffusion equations, Numer. Math., to appear, 1998

[61] C.C.Yu, J.C.Heinrich : Petrov-Galerkin methods for time-dependent convective transport equation, Int. J. Numer. Meth. Engrg., Vol.23, 883-901, 1986, und
Int. J. Numer. Meth. Engrg., Vol.24, 2201-2215, 1987

[62] O.C.Zienkiewicz, R.L.Taylor : The finite element method, 4.ed., Vol. 1 und 2, McGraw-Hill, 1989

Lebenslauf

- Name :
  Areti Papastavrou

- geboren :
  14. April 1971 in Hanau

- Staatsangehörigkeit : deutsch

- Eltern :
  Dr. med. Nikolas Papastavrou
  Liselotte Papastavrou, geb. Schöndube

- Schulische Ausbildung :
  1977-1979 Wilhelm-Neuhaus-Schule Bad Hersfeld
  1979-1981 Linggschule Bad Hersfeld
  1981-1987 Gesamtschule Obersberg Bad Hersfeld
  1987-1990 Modellschule Obersberg Bad Hersfeld
  1990 Abitur

- Studium :
  10/1990 - 01/1996 Bauingenieurwesen, Universität Hannover
  - 04/1992 Vordiplom
  - 01/1996 Diplom in der Vertiefung "Statik/Mechanik"

- Beruf :
  seit 01/1996 Promotionsstipendium des Graduiertenkollegs "Computational Structural Dynamics", Ruhr-Universität Bochum, Lehrstuhl für Numerik, Prof. Dr. R. Verfürth