Ein Mehrgitterverfahren zur Berechnung der Eigenschwingungen von abgeschlossenen Wasserbecken

Diplomarbeit
von
Stefan Sauter

Betreuung
Prof. Dr. R. Verfürth

Universität Heidelberg
Institut für angewandte Mathematik
November 1989

Inhalt

§1 Einleitung

Oftmals werden neue numerische Methoden zunächst auf Modellprobleme angewendet und dabei deren Effizienz getestet. Obwohl es dann häufig naheliegend ist, daß sich die Algorithmen dann "im Prinzip" auch auf verwandte komplexere Probleme übertragen lassen, ist die Realisierung sowohl theoretisch (Übertragen der Sätze und Beweise) als auch programmiertechnisch sehr aufwendig.
In der vorliegenden Arbeit wird versucht, ein reales physikalisches Problem aus der Strömungsmechanik mit modernen numerischen Methoden zu lösen. Sie wurde angeregt durch einen Vortrag von Dr. E. Bäuerle über die Eigenschwingungen von Seen, insbesondere des Bodensees. So berichtet der Chronist Schulthaiss aus Konstanz im Jahre 1549 über das „Wunder anloffen wassers" [18], wie der Rhein bei Konstanz damals mit einer Periode von etwa fünfzehn Minuten einigemale flußaufwärts und dann wieder flußabwärts floß. Bäuerle hat in [2] versucht, diese Eigenschwingungen, auch seiches genannt, zu berechnen, indem er ein Differenzenverfahren zur Diskretisierung der relevanten partiellen Differentialgleichungen (vgl. § 2) verwendet hat und die Eigenwerte der entstandenen Matrix als Nullstellen des charakteristischen Polynoms berechnet hat. Nun sind aber Differenzenverfahren für komplizierte Gebiete ziemlich ungeeignet und die Eigenwertberechnung mit Hilfe des charakteristischen Polynoms enorm aufwendig (vgl. § 8).
Ziel dieser Arbeit war es, die Grund- und Oberschwingungen des Bodensees mit einem Mehrgitterverfahren mit ILU-Glättung für eine finite Elemente Diskretisierung zu berechnen. Die Vorteile dieser drei Bausteine sind:

Die finiten Elemente sind im Gegensatz zu Differenzenverfahren gut geeignet, sich Gebieten mit komplizierter topologischer Struktur anzupassen.

Mehrgitterverfahren sind die schnellsten Verfahren zur Lösung linearer Gleichungssysteme; der Aufwand ist proportional zur Dimension des linearen Gleichungssystems.

Das ILU-Verfahren wurde für Differenzenverfahren bei Modellproblemen auf Rechtecksgebieten von Wittum [25] systematisch getestet und zeigte dort ausgezeichnete Glättungseigenschaften. Auch erwiesen sie sich als numerisch robust (vgl.§ 6.6).

Andererseits entstehen durch die Kombination dieser Methoden wesentliche algorithmische und theoretische Probleme. Das ILU-Verfahren setzt eine Datenstruktur voraus, welche zu einfach strukturierten Matrizen führt (gleichmäßige Gitter). Dagegen ist die finite Elemente Methode gerade deshalb so effektiv für Gebiete mit „krummen Rändern", weil sie auch unstrukturierte und dadurch problemangepaßte Gitter erlaubt.Es mußte also versucht werden, strukturierte und gleichzeitig problembezogene Gitter zu erzeugen. Um den Algorithmus auf dem Computer zu realisieren, mußte dazu zunächst eine neue Datenstruktur geschaffen werden. Dadurch wurde der Programmieraufwand im Vergleich zu entsprechenden Modellproblemen und einfacheren Glättungsverfahren wesentlich höher, da so gut wie keine Routinen aus bereits existierenden Programmen übernommen werden konnten. Zusätzlich ist die Komplexität der einzelnen Routinen (z.B. Abspeicherung der Matrix, ILU-Zerlegung der Matrix, sukkzesive Anpassung des Gitters an das Gebiet durch einen speziellen Projektionsalgorithmus, Matrix-Vektor Operationen etc.) im vorliegenden Fall teilweise erheblich höher, als bei vergleichbaren Routinen für Modellprobleme.

Der Programmcode zur Berechnung der Eigenschwingungen des Bodensees umfaßt etwa 6900-, das zugehörige Graphik-Interface etwa 3900 Fortran-Zeilen und die einzulesenden Daten (Rand und Tiefe des Bodensees und Grobtriangulierung) ungefähr 2800 Zeilen.

Am interessantesten an der ganzen Problemstellung war natürlich die Frage: "Wie gut ist das Verfahren zur Berechnung der Eigenschwingungen"?
Hofmann hat in [11] die Eigenwerte und -vektoren der Plattengleichung auf dem Einheitsquadrat mit Dirichlet-Randbedingung gerechnet. Da dort auf einem regelmäßigen Gitter über einem sehr einfachen Gebiet gerechnet wurde, haben diese Ergebnisse kaum eine Aussagekraft in bezug aus das komplexe Bodenseegebiet mit einer nichtkonstanten Tiefe und einem unregelmäßigen Gitter. Für die Laplacegleichung mit Neumann-Randbedingung auf dem Bodensee erhalten wir hervorragende Konvergenzraten und können für die Eigenwertgleichung mit einer konstanten Äquivalenztiefe (§ 8) elf Eigenschwingungen berechnen. Dagegen erhalten wir bei einer punktweise eingegebenen Tiefe Konvergenz nur noch für den niedrigsten Eigenwert (vgl.§ 8).
Die mit der Äquivalenztiefe berechneten Eigenwerte stellen eine sehr gute Approximation der bisher in der Natur beobachteten dar. Man muß diese Ergebnisse jedoch mit gewissen Vorbehalten betrachten, da es fraglich ist, wie genau die Messungen und Beobachtungen, die teilweise im Jahre 1549 erfolgten, waren. Ein weiteres Problem stellen die topologischen Eingangsdaten dar. Durch Ablagerungen auf dem Grund des Sees, vor allem im Bereich von Flußmündungen, ändert sich beispielsweise die Tiefe des Sees. In diesem Zusammenhang ist zu bemerken, daß die neueste Tiefenkarte des Bodensees aus dem Jahr 1893 stammt.

Der theoretische Teil der Arbeit gliedert sich wie folgt.
In § 2 werden zunächst die relevanten Gleichungen für die Wasserschwingungen hergeleitet und mit Hilfe physikalischer Überlegungen vereinfacht. In § 3 stellen wir die kontinuierlichen Gleichungen in einen mathematischen Rahmen und erklären die wichtigsten analytischen Eigenschaften. In § 4 beschäftigen wir uns kurz mit der Diskretisierung der Gleichungen durch lineare finite Elemente und geben einige Abschätzungen für den Diskretisierungsfehler an. In § 5 wird das verwendete Mehrgitterverfahren erklärt zur Lösung linearer Eigenwertprobleme und singulärer Gleichungen. Es wird untersucht unter welchen Voraussetzungen an die Operatoren und Diskretisierungen das Verfahren konvergiert, d.h. die Approximations- und Glättungseigenschaft besitzt. Der wesentliche Teil der Arbeit stellt dann § 6 dar. Hier wird die Diskretisierung (Gittergenerierung) des Gebiets und das verwendete ILU-Verfahren entwickelt. Danach stellen wir die wichtigsten bisher bewiesenen theoretischen Eigenschaften der ILU-Zerlegung zusammen und werden feststellen, daß die Voraussetzungen bei diesen Beweisen in unserem Fall von finiten Elementen üblicherweise nicht zu erfüllen sind. Satz (6.8.22) zeigt aber, daß wir auch in unserem Fall beispielsweise mit der Stabilität der Zerlegung rechnen können. § 7 und § 8 beziehen sich auf die programmiertechnische Seite der Arbeit und auf die vom entwickelten Programm erzielten numerischen Ergebnisse.

Abschließend kann man sagen, daß zwar ein recht effektives Verfahren geschaffen wurde, man aber dabei immer aufpassen mußte, nicht zu schnell zu große Sprünge in der Komplexität der Probleme machen zu wollen.

Ich danke Herrn Professor R. Verfürth, der die Arbeit trotz seines Umzuges nach Zürich weiterhin unterstützt hat, für die Überlassung des Themas und die Betreuung.
Auch möchte ich mich bei Herrn Dr. G. Wittum bedanken, der mich mit vielen guten Ratschlägen und neuen Ideen sehr motiviert und geholfen hat, sowie Herrn Dr. H. Blum, der mir in interessanten Diskussionen wertvolle Tips gegeben hat.

§2 Herleitung der Shallow-Water-Equations

In diesem Abschnitt werden wir die Shallow-Water-Equations, auch Flachwasser- bzw. Lange-Wellen-Gleichungen genannt, direkt aus physikalischen Überlegungen herleiten (vgl. [2]). Für eine Herleitung der Gleichungen als Spezialfall der Navier-Stokes-Gleichungen siehe [12] und [14].

(2.1) Physikalische Herleitung der Gleichungen

Wir betrachten im folgenden ausschließlich abgeschlossene Wasserbecken. Für die Tiefenfunktion $h(x, y)$, die den Abstand zwischen dem Seegrund und dem ruhenden Seespiegel angibt, soll gelten: $h(x, y) \geq c_h > 0$. Das bedeutet, daß der Seeboden am Ufer senkrecht um mindestens c_h abfallen soll. Auch soll das Ufer um c_h überhöht sein, wobei c_h größer als die maximale Amplitude der Schwingung sein muß.

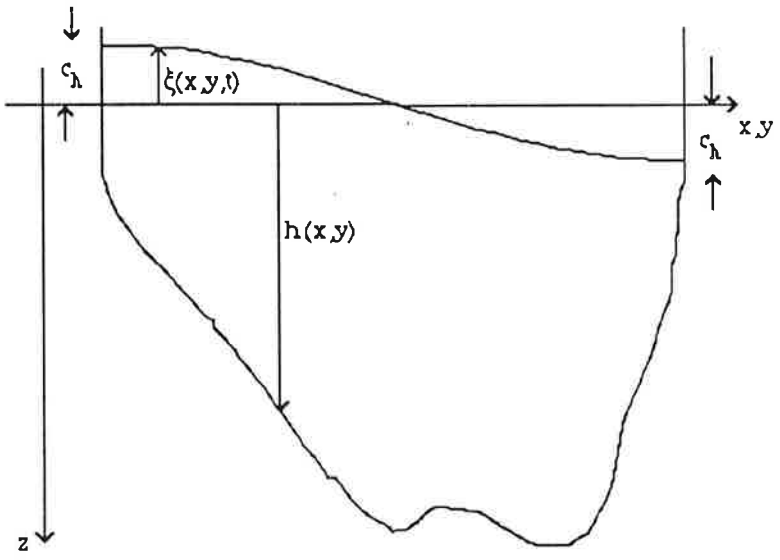


Abb.1 abgeschlossenes Wasserbecken (Querschnitt)

Wir wollen annehmen, daß die Schwingungen der Wasseroberfläche durch Druckunterschiede im Wasser verursacht werden, welche durch die Wirkung der äußeren Kräfte: der Schwerkraft G, der Corioliskraft F_{cor} und weiterer zeitlich periodischer Horizontalkräfte F (z.B. Wind) entstehen.

Seien u, v, w die Geschwindigkeitskomponenten eines Wasserteilchens in x, y, z Richtung.
Mit den Größen :

$$\boldsymbol{u} := \begin{pmatrix} u \\ v \end{pmatrix}, \ \tilde{\boldsymbol{u}} := \begin{pmatrix} \boldsymbol{u} \\ w \end{pmatrix}, \ \boldsymbol{u}^{\perp} := \begin{pmatrix} -v \\ u \end{pmatrix}$$

lauten die Kräfte, die auf ein beliebiges Volumenelement V im See wirken :

Corioliskraft :

$$F_{cor} = \rho \int_V c \begin{pmatrix} \boldsymbol{u}^{\perp} \\ 0 \end{pmatrix} dV \quad ,$$

wobei ρ die Dichte des Wassers im Volumen V ist, welche hier als konstant angenommen wird, und c den Coriolisparameter bezeichnet.

Gravitationskraft :

$$G = g \cdot \rho \int_V \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} dV \ ,$$

wobei $g = 9.81 \ m s^{-2}$ die Gravitationskonstante bezeichnet.

Trägheitskraft :

$$F_{Trägheit} = \rho \int_V \frac{\partial \tilde{\boldsymbol{u}}}{\partial t} dV \ .$$

Druckkraft :
Druckkräfte stehen senkrecht zur Oberfläche $\overrightarrow{dS}$ des betrachteten Volumenelements.

$$F_{Druck} = \int_{\partial V} p \, \overrightarrow{dS} \, ;$$

wobei p den Druck auf das Volumenelement V bezeichnet

äussere Kräfte :
Für die äußeren Kräfte nehmen wir an, daß sie sich als Integral über eine Volumendichte $\tilde{f}$ ausdrücken lassen; $\tilde{f} = (\tilde{f_1}, \tilde{f_2}, \tilde{f_3})^T$.
Da wir ausschließlich horizontale äußere Kräfte betrachten wollen, gilt : $\tilde{f_3} = 0$.

$$F = \int_V \tilde{f} \, dV$$

(2.1.1) Bemerkung
Falls der Druck p auf das hinreichend glatte Volumenelement V differenzierbar ist, gilt :

$$\int_{\partial V} p \, \overrightarrow{dS} = \int_V \operatorname{grad} p \, dV$$

❑

Beweis:

Sei $a \in \mathbb{R}^3$ fest aber beliebig.
Dann folgt aus dem Gaußschen Integralsatz:

$$\int_{\partial V} (p\,a)\,\overrightarrow{dS} = \int_V \operatorname{div}(p\,a)\,dV = \int_V < a\,,\ \operatorname{grad} p > dV \qquad \text{QED}$$

Wegen des Newtonschen Prinzips von actio et reactio gilt :

$$(2.1.2) \qquad \int_V \operatorname{grad} p\,dV + \rho \int_V \frac{\partial \tilde{u}}{\partial t}\,dV = \rho \int_V c \begin{pmatrix} u^{\perp} \\ 0 \end{pmatrix} dV + g \cdot \rho \int_V \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} dV + \int_V \tilde{f}\ dV\,.$$

In differentieller Form lautet obige Gleichung : (mit $\tilde{f}_{12} := (\tilde{f}_1, \tilde{f}_2)^T$

$$(2.1.3\ a) \qquad \frac{\partial u}{\partial t} - c\,u^{\perp} + \frac{1}{\rho} \nabla p = \tilde{f}_{12}$$

$$(2.1.3\ b) \qquad \rho \frac{\partial w}{\partial t} - g \cdot \rho + \frac{\partial p}{\partial z} = 0$$

Die Kontinuitätsgleichung lautet unter den Annahmen, daß

$$\frac{\partial \rho}{\partial t} = 0 \text{ und } \nabla \rho = 0 \text{ gilt :}$$

$$(2.1.3\ c) \qquad \operatorname{div} \tilde{u} = 0$$

Wir wollen nur Schwingungen mit kleinen Amplituden (im Vergleich zu den Abmessungen des Sees) betrachten. Nach Bäuerle[2] können wir annehmen, daß $\left|\frac{\partial w}{\partial t}\right| << g$ ist und erhalten dann, indem wir Gl.(2.1.3 b) von der Wasseroberfläche bis zu einer Tiefe z (d.h. von ξ bis z) integrieren :

$$(2.1.3\ b') \qquad p = \rho\, g\, (\xi(x,y,t) + z) + p_{Atmosphäre}\,.$$

Indem wir (2.1.3 b´) in (2.1.3 a) einsetzen, ergibt sich :

$$(2.1.3\ a') \qquad \frac{\partial u}{\partial t} - c \cdot u^{\perp} + g \nabla \xi(x,y,t) = \tilde{f}_{12}$$

z - Integration von (2.1.3 a´) von 0 bis $h(x,y)$ liefert

$$(\text{mit } U := \int_0^{h(x,y)} \frac{\partial u}{\partial t}\,dz \text{ und } f := \int_0^{h(x,y)} \tilde{f}\,dz\,) :$$

$$(2.1.4\ a) \qquad \frac{\partial U}{\partial t} - c \cdot U^{\perp} + g \cdot h(x,y) \nabla \xi = f\,.$$

Da weiter gilt :

$$\int_0^{h(x,y)} \frac{\partial w}{\partial z}\, dz = w\Big|_{z=h(x,y)} - w\Big|_{z=0} = 0 - \left(-\frac{\partial \xi}{\partial t}\right) = \frac{\partial \xi}{\partial t} ,$$

folgt für die Kontinuitätsgleichung (2.1.3 c) nach z - Integration :

(2.1.4 b)
$$\operatorname{div} \boldsymbol{U} + \frac{\partial \xi}{\partial t} = 0 .$$

Für abgeschlossene Wasserbecken sind die Randbedingungen gegeben durch :

$$\boldsymbol{U} \cdot \boldsymbol{n} = 0 ;$$

wobei $\boldsymbol{n}$ den nach außen gerichteten Normalenvektor an Ω bezeichnet.

Die Bewegungsgleichungen für die Wasserschwingungen lauten dann in Operatorform zusammengefaßt (ξ ist im folgenden ersetzt durch z .) :

(2.1.5 a)
$$\begin{pmatrix} \frac{\partial}{\partial t} & c & g \cdot h \frac{\partial}{\partial x} \\ -c & \frac{\partial}{\partial t} & g \cdot h \frac{\partial}{\partial y} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial t} \end{pmatrix} \begin{pmatrix} u \\ v \\ z \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ 0 \end{pmatrix} \qquad \text{in } \Omega$$

(2.1.5 b)
$$\boldsymbol{U} \cdot \boldsymbol{n} = 0 \qquad \text{auf } \Gamma := \partial\Omega .$$

Wir wollen ausschließlich periodische Lösungen betrachten der Form :

(2.1.6)
$$\begin{aligned} \boldsymbol{U} &= \bar{\boldsymbol{u}}(x,y) \sin(\omega t) \\ z &= \bar{z}(x,y) \cos(\omega t) , \end{aligned}$$

angeregt durch äußere Kräfte der Form :

$$\begin{aligned} f &= \bar{f}(x,y) \cos(\omega t) \\ \bar{f} &= (\bar{f}_1, \bar{f}_2, 0)^T . \end{aligned}$$

Bäuerle stellte in [2] fest, daß die Corioliskraft speziell für den Bodensee einen vernachlässigbaren Einfluß auf die Seeschwingungen hat.
Wir können also c = 0 setzen und erhalten dann mit dem Ansatz (2.1.6) das bis auf den Faktor $(g \cdot h)$ symmetrische Gleichungssystem :

(2.1.7)
$$\begin{matrix} (i) \\ (ii) \end{matrix} \begin{pmatrix} \omega I & g \cdot h \cdot \nabla \\ -\operatorname{div} & \omega I \end{pmatrix} \begin{pmatrix} \bar{u} \\ \bar{z} \end{pmatrix} = \begin{pmatrix} \bar{f} \\ 0 \end{pmatrix}$$

Dieses System läßt sich entkoppeln, und es ergibt sich dann (mit $\underline{f} := \operatorname{div} \bar{f}$) :

$$\begin{pmatrix} \omega I & g \cdot h \cdot \nabla \\ 0 & \omega^2 I + \operatorname{div}(g \cdot h \cdot \nabla) \end{pmatrix} \begin{pmatrix} \bar{u} \\ \bar{z} \end{pmatrix} = \begin{pmatrix} \bar{f} \\ \underline{f} \end{pmatrix} \tag{2.1.8}$$

Zur Randbedingung :

Wir betrachten Gl. (2.1.7 *i*) auf Γ und multiplizieren mit dem Normalenvektor $\mathbf{n}$:

$$(g \cdot h \cdot \nabla \bar{z}) \cdot \mathbf{n} = \bar{f} \cdot \mathbf{n} \ .$$

Da $h > 0$ ist, können wir definieren :

$$\frac{\partial z}{\partial n} = \nabla \bar{z} \cdot \mathbf{n} = \frac{\bar{f} \cdot \mathbf{n}}{g \cdot h} =: r(x,y) \quad \text{auf } \Gamma \ .$$

(2.1.9) Bemerkung

In dieser Herleitung wurde immer von einer homogenen Temperaturschichtung des Wassers ausgegangen. Für ein Mehrschichtenmodell siehe [2].

§3 Analyse der kontinuierlichen Gleichungen

Wir wollen nun die in §2 aus physikalischen Überlegungen gewonnenen Gleichungen

$$\begin{aligned} \omega^2 z + \operatorname{div}(g \cdot h \cdot \nabla z) &= f \quad \text{in } \Omega \\ \frac{\partial z}{\partial n} &= r \quad \text{auf } \Gamma \end{aligned} \tag{3.1}$$

in einen mathematischen Rahmen stellen, um sie dann zunächst analytisch zu untersuchen.

Zunächst benötigen wir einige Definitionen:

(3.2) Definition *Gelfand-Dreier*

Seien V,U Hilbert-Räume und $V \subset U$ stetig und dicht eingebettet. Dann ist U´ stetig und dicht in V´ eingebettet (vgl. [9; Lemma 6.3.9]). Nach dem Darstellungssatz von Riesz läßt sich U mit U´ identifizieren. Man erhät dadurch den *Gelfand-Dreier* $V \subset U \subset V'$, wobei alle Einbettungen stetig und dicht sind.

❑

(3.3) Definition

Eine Bilinearform $a(\cdot,\cdot)$: $V \times V \to \mathbb{R}$ heißt *stetig* (oder *beschränkt*), wenn ein C_S existiert, so daß

$$|a(x,y)| \le C_S \|x\|_V \|y\|_V \quad \forall x,y \in V \tag{3.4}$$

❑

(3.5) Definition

$V \subset U \subset V'$ sei ein Gelfand-Dreier. Eine Bilinearform $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ heißt *V-koerziv* , wenn sie stetig ist und wenn es $C_K \in \mathbb{R}$ und $C_E > 0$ gibt, so daß

$$a(x,x) \ge C_E \|x\|_V^2 - C_K \|x\|_U^2 \quad \forall x \in V \tag{3.6}$$

❑

(3.7) Definition

Sei $\Omega \subset \mathbb{R}^n$. Wir schreiben $\Omega \in C^{0,1}$, wenn für jedes $x \in \Gamma := \partial\Omega$ eine Umgebung $U_x \subset \mathbb{R}^n$ existiert, und es dazu eine bijektive Abbildung:

$$\Phi : U_x \rightarrow K_1(0) := \{\xi \in \mathbb{R}^n : |\xi| < 1\}$$

gibt, mit:

$$\Phi \in C^{0,1}(\overline{U_x}),\ \Phi^{-1} \in C^{0,1}(\overline{K_1(0)})$$
$$\Phi(U_x \cap \Gamma) = \{\xi \in K_1(0),\ \xi_n = 0\}$$
$$\Phi(U_x \cap \Omega) = \{\xi \in K_1(0),\ \xi_n > 0\}$$
$$\Phi(U_x \cap \mathbb{R}^n \backslash \Omega) = \{\xi \in K_1(0),\ \xi_n < 0\}.$$

❑

Bezeichnungen und Voraussetzungen:

Das Produkt $(g \cdot h)$ aus (2.1.8) wird jetzt immer mit h bezeichnet. Der Sobolevraum H^k ist definiert als diejenigen Teilmengen von L^2, deren Elemente schwache Ableitungen bis zur Ordnung k besitzen, die in L^2 liegen (vgl [9]).

Wir nehmen an :

$$\Omega \in C^{0,1},\ f \in H^{-1}(\Omega),\ h \in C^{\infty}(\Omega),\ r \in H^{-1/2}(\Gamma),\ \lambda := \omega^2.$$

Wir führen die Bilinearform :

(3.8)
$$a : H^1(\Omega) \times H^1(\Omega) \rightarrow \mathbb{R}$$
$$a(u,v) := (\nabla u, h\,\nabla v)_0$$

und das Funktional :

(3.9)
$$g \in H^{-1}(\Omega)$$
$$g(v) := -\int_\Omega (f \cdot v)\, d\Omega + \int_\Gamma (h \cdot r \cdot v)\, d\Gamma \quad v \in H^1(\Omega)$$

ein.

Damit lautet die Variationsformulierung von (3.1) :

(3.10)
$$\text{Suche } u \in H^1(\Omega), \text{ so daß}$$
$$a(u,v) = \lambda\,(u,v)_0 + g(v) \quad \text{für alle } v \in H^1(\Omega)$$

(3.11) Lemma

Einer stetigen Bilinearform kann man eineindeutig einen Operator $A \in \mathfrak{L}(V,V')$ zuordnen, so daß:

$$a(x,y) = \langle A\,x\,,\,y \rangle_{V'xV} \quad \forall\, x,y \in V$$

mit
$$\| A \|_{V'xV} \leq C_s$$

❑

Beweis:

siehe [9; Lemma 6.5.1] QED

Mit Hilfe von Lemma (3.11) können wir nun $a(\cdot,\cdot)$ den Operator A und $(\cdot\,,\cdot\,)_0$ den Operator M zuordnen, so daß (3.10) äquivalent ist, zu :

Suche $u \in H^1(\Omega)$, so daß

(3.10´) $$A u - \lambda M u = g$$

(3.12) Bemerkung

Die Bilinearform $a(\cdot,\cdot)$ und das Funktional g ist durch (3.8) bzw. (3.9) wohldefiniert.

Es gilt :

$$\|g\|_{-1} \leq C \{ \|f\|_{-1} + \|r\|_{H^{-1/2}(\Omega)} \}$$

Weiter ist die Bilinearform $a(\cdot,\cdot)$ stetig und V-koerziv.

❑

Beweis:

siehe [9; p.139/140]

QED

Die Frage nach der Existenz und Eindeutigkeit von Lösungen beantwortet eine Folgerung aus der Riesz-Schauder Theorie:

(3.13) Satz

Sei $V \subset U \subset V'$ ein Gelfand-Dreier mit kompakter Einbettung $V \subset U$. Die Bilinearform $a(\cdot,\cdot)$ sei V-koerziv mit zugehörigem Operator A. $I : V \to V'$ sei die Inklusion. Sei

$$E(\lambda) := \ker(A - \lambda I) \text{ und } E'(\overline{\lambda}) := \ker(A' - \overline{\lambda} I).$$

Behauptung:

a) Für jedes $\lambda \in C$ gilt eine der folgenden Alternativen:

(*i*) $(A - \lambda I)^{-1} \in \mathfrak{L}(V,V')$ und $(A' - \lambda I)^{-1} \in \mathfrak{L}(V',V)$

(*ii*) λ ist Eigenwert von A

b) Das Spektrum $\sigma(A)$ von A besteht aus höchstens abzählbar unendlich vielen Eigenwerten

Es gilt: $\lambda \in \sigma(A) \Leftrightarrow \overline{\lambda} \in \sigma(A')$.

Ferner ist $\dim E(\lambda) = \dim E'(\overline{\lambda}) < \infty$.

c) Für $\lambda \in \sigma(A)$ hat $Ax - \lambda x = f$ für $f \in V'$ mindestens eine Lösung $x \in V$ genau dann, wenn $f \perp E'(\lambda)$; d.h.: $\langle f, x' \rangle_{V'xV} = (f, x')_U = 0 \quad \forall x \in E'(\overline{\lambda})$.

❑

Beweis:

siehe [9; Satz 6.5.15]

QED

In unserem Fall ist $V = H^1(\Omega)$; $U = L^2\,\Omega)$ und $V' = H^{-1}(\Omega)$.
Nach dem Sobolevschen Einbettungssatz ist $H^1(\Omega)$ kompakt eingebettet in $L^2(\Omega)$ für $\Omega \subset \mathbb{R}^2$. Die Bilinearform $a(\cdot,\cdot)$ in Gleichung (3.10) ist nach Bemerkung(3.12) H^1 - koerziv; das bedeutet nach Satz (3.13) u.a.:

Sei $\lambda \in \mathbb{C}$ fest aber beliebig, und a' die zu a adjungierte Bilinearform ($a'(x,y) := a(y,x)$). Dann gilt entweder:
Die Gleichungen

$$a(x,y) - \lambda(x,y)_0 = g(y) \quad \forall y \in H^1(\Omega)$$
$$a'(x',y) - \overline{\lambda}(x',y)_0 = g(y) \quad \forall y \in H^1(\Omega)$$

sind für alle $g \in H^{-1}(\Omega)$ eindeutig lösbar,
oder :

$$1 \leq \dim E(\lambda) = \dim E'(\overline{\lambda}) < \infty;$$

m.a.W. besitzen die Gleichungen

$$a(x,y) = \lambda(x,y)_0 \quad \forall y \in H^1(\Omega)$$
$$a'(x',y) = \overline{\lambda}(x',y)_0 \quad \forall y \in H^1(\Omega)$$

nichttriviale Lösungen $x \in E(\lambda)$ und $x' \in E'(\overline{\lambda})$.

§4 Diskretisierung der kontinuierlichen Gleichungen

4.1 Diskretisierung mit linearen finiten Elementen

Für Leser, die an einer ausführlichen Theorie des Ritz-Galerkin Verfahrens bzw. der finiten Elemente interessiert sind, sei an dieser Stelle auf [9; § 8] verwiesen. Der vorliegende Abschnitt soll nur eine kurze Beschreibung der Methode sein und die für unser Problem wichtigsten Sätze zusammenfassen.

Wir wollen das Problem (3.10) mit der Ritz-Galerkin Methode diskretisieren, indem wir den unendlichdimensionalen Raum $H^1(\Omega)$, der im folgenden mit V bezeichnet wird, durch eine Folge endlichdimensionaler Teilräume $\{V_n\}$ approximieren. Dies geschiet in unserem Fall dadurch, daß das Gebiet Ω in kleine Dreiecke („finite Elemente") zerlegt wird. Die „Ansatzfunktionen" φ_i müssen stückweise linear sein (auf jedem Dreieck linear), in einem Eckpunkt eines Dreiecks (P_i) gleich eins sein und in allen anderen Eckpunkten verschwinden. $V_n := \text{span}\{\varphi_i ; 1 \le i \le n\}$, wobei n die Anzahl der Eckpunkte aller Dreiecke ist.

Wir erhalten dadurch das Problem:

(4.1.1 a) Sei $V_n \subset V$, $\dim V_n = n < \infty$

(4.1.1 b) Suche $u_n \in V_n$, so daß

$$a(u_n, v) - \lambda(u_n, v)_0 = g(v) \quad \forall v \in V_n \quad .$$

Indem man auf V_n eine Basis $\{b_1, b_2, \ldots, b_n\}$ einführt, kann man das lineare Gleichungssystem:

(4.1.2) $$K u - \lambda M u = f$$

mit den n xn Matrizen:

$$K := (K_{i,j})_{1 \le i,j \le n} \qquad K_{i,j} = a(b_j, b_i)$$

$$M := (M_{i,j})_{1 \le i,j \le n} \qquad M_{i,j} := (b_j, b_i)_0$$

und dem n - Vektor:

$$f := (f_1, f_2, \ldots, f_n) \qquad f_i := g(b_i)$$

definieren.
Die Lösung u von (4.1.2) ist über den Isomorphismus:

$$P : \mathbb{R}^n \to V_n \subset V$$

$$Pu := \sum_{i=1}^{n} u_i \, b_i$$

äquivalent ist zu u_n aus Problem (4.1.1).
Die Matrix K bezeichnet man als Steifigkeitsmatrix, M üblicherweise als Massenmatrix.

4.2 Analyse der diskreten Gleichungen

Um Aussagen über die Lösbarkeit des Problems (4.1.1) bzw. (4.1.2) machen zu können, (speziell für $n \to \infty$) und Abschätzungen zwischen kontinuierlicher und diskreter Lösung zu erhalten, müssen wir zwei Fälle unterscheiden:

(4.2.1 a) Das (kontinuierliche) Problem (3.10) ist für alle $g \in H^{-1}(\Omega)$ lösbar,

(4.2.1 b) λ ist Eigenwert von $M^{-1}K$ (M, K aus 4.1.12) .

(4.2.2) Satz

Die Bilinearform $a(\cdot\,,\cdot\,)$ sei V - koerziv und $V \subset U \subset V'$ ein Gelfand - Dreier.
Das Problem :

$$\text{Suche } u \in V \text{, so daß}$$
$$a(u,v) = g(v) \quad \forall v \in V$$

sei für alle $g \in V'$ lösbar.
Sei $V_i := V_{n_i} \subset V$ ($i \in \mathbb{N}$) eine Folge von Unterräumen mit :

$$\lim_{i\to\infty} d(u,V_i) := \lim_{i\to\infty} \inf_{w\in V_{n_i}} \| u - w \|_V = 0 .$$

Dann gilt :
a) Für hinreichend großes i ist die folgende Stabilitätsbedingung mit $\varepsilon_{n_i} \geq \varepsilon > 0$ erfüllt.

(4.2.3)
$$\inf \{ \sup \{ |a(u,v)| : v \in V_{n_i}, \|v\|_V = 1 \} : u \in V_{n_i}, \|u\|_V = 1 \} = \varepsilon_{n_i} > 0$$

b) Bedingung (4.2.3) ist äquivalent zu

$$\| A_{n_i}^{-1} \|_{V \leftarrow V'} < \frac{1}{\varepsilon_{n_i}}$$ ❑

Beweis:

a) folgt aus [9; Satz 8.2.8 und Lemma 11.2.7]

b) siehe [9; Lemma 6.5.3] QED

Der folgende Satz gibt an, wie man aus der Stabilitätsbedingung (4.2.3) und der Konstanten C_S aus (3.4) eine Abschätzung für den Diskretisierungsfehler $\|u - u_n\|_V$ bekommt (vgl. [9; Satz 8.2.1].

(4.2.4) Satz

Es gelte (4.1.1 a), (3.3), (4.2.3). $u \in V$ sei eine Lösung von Aufgabe (3.10), während $u_n \in V_n$ die Ritz-Galerkin Lösung von (4.1.1) sei. Dann gilt die Abschätzung:

$$\| u - u_n \|_V \leq (1 + C_S / \varepsilon_n) \inf_{w \in V_n} \| u - w \|_V$$

mit C_S aus (3.3) und ε_n aus (4.2.3).

❑

(4.2.5) Bemerkung

Die Konstante $(1 + C_S/\varepsilon_n)$ ist für alle $n \in \mathbb{N}$ beschränkt durch $(1 + C_S/\varepsilon)$ mit ε aus Satz (4.2.2).

❑

Wir wollen das singuläre Problem (3.10) - Situation (4.2.1 b) - nur im homogenen Fall ($g \equiv 0$) betrachten. Wie wir in § 5 sehen werden, kann man mit der Kenntnis der Eigenvektoren von $M^{-1} K$ die Aufgabe (3.10) regularisieren, indem man sie auf geeigneten Quotientenvektorräumen betrachtet.
Bei der Konvergenzuntersuchung für das Eigenwertproblem hat man folgende Schwierigkeiten zu beachten:

1.) unendlich vielen Eigenwerten und -vektoren im kontinuierlichen Fall stehen im diskreten immer nur endlich viele gegenüber. Man kann deshalb eine gleichmäßige Approximation aller Eigenwerte nicht erwarten.

2.) Die Dimension der Eigenräume der kontinuierlichen bzw. der dazugehörigen diskreten Eigenwerte braucht nicht übereinzustimmen. Wir können jedoch hoffen, daß in völlig unsymmetrischen Gebieten (Bodensee) gilt:

$$\dim E(\lambda) = \dim E(\lambda_h) = 1 .$$

Falls man aber trotzdem mit mehrfachen Eigenwerten zu tun hat, hilft einem die im folgenden ausgeführte Überlegung, für die wir zunächst einige Konvergenzbegriffe für Operatorfolgen benötigen:

(4.2.6) Definition

Seien im folgenden T_n und T Operatoren in X.
Wir wollen annehmen, daß für den Definitionsbereich (dom) aller nun vorkommenden Operatoren T_n , T gilt:

$$\mathrm{dom}\,(T_n) = \mathrm{dom}\,(T) = : D.$$

Sei $\{T_n\}_{n \in \mathbb{N}}$ eine Folge beschränkter Operatoren in D.

Man definiert:

T_n konvergiert gegen $T \in \mathfrak{L}(D)$

a) *punktweise*: $T_n \xrightarrow{p} T$ genau dann, wenn für alle $x \in D$ gilt:

$$T_n x \to Tx, \quad n \to \infty ;$$

b) *stabil*: $T_n \xrightarrow{s} T$ genau dann, wenn

(i) $T_n \xrightarrow{p} T$

(ii) $\exists M > 0; \ \exists N_0 \in \mathbb{N}$ so daß: $\forall n \in \mathbb{N}, n > N_0$ gilt:

$$T_n^{-1} \in \mathfrak{L}(D) \quad \text{und} \quad \|T_n^{-1}\| \leq M.$$

c) Für einen Eigenwert λ von T der algebraischen Vielfachheit $m < \infty$ definieren wir die *stark stabile* Konvergenz in einer Umgebung U von λ

$$T_n \xrightarrow{ss} T$$

durch:

(i) $$T_n - z \xrightarrow{s} T - z \quad \forall z \in U - \{\lambda\}$$

(ii) $$\dim P_n(X) = m ;$$

wobei P_n die mit $\sigma(T_n) \cap U$ assoziierte Spektralprojektion von T_n bezeichnet.

$$(P_n := -\frac{1}{2\pi i} \int_{\partial U} R_n(z)dz, \text{ mit dem Resolventenoperator } R_n \text{ von } T_n).$$

❑

Damit können wir nun den Satz über die Konvergenz mehrfacher Eigenwerte formulieren.

(4.2.7) Satz

Voraussetzungen wie in Definition (4.2.6 c).

Dann gilt:

a) Falls $T_n - z \xrightarrow{ss} T - z$ in U, dann enthält $\sigma(T_n) \cap U$ für hinreichend großes n exakt m Eigenwerte $\{\mu_{jn}\}_{1 \leq j \leq m}$ von T_n, gemäß ihrer Vielfachheit gezählt.

b) Die beste Approximation von λ ist der Mittelwert $\hat{\lambda}_n$ definiert durch:

$$\hat{\lambda}_n := \frac{1}{m} \sum_{j=1}^{m} \mu_{jn}$$

❑

Beweis:

a) siehe [6; Proposition 5.6]

b) siehe [6; § 6 und Satz 4.2.7] QED

Den Zusammenhang zwischen den Operatoren T_n , $T \in \mathfrak{L}(D)$ und der partiellen Differentialgleichung in schwacher Formulierung (4.1.1) erläutert folgende Bemerkung.

(4.2.8) Bemerkung

$V \subset U \subset V'$ sei ein Gelfand-Dreier mit kompakter Einbettung $V \subset U$. Die Bilinearform $a(\cdot,\cdot)$ sei V-koerziv mit zugehörigem Operator A (vgl. Lemma 3.11). $I: V \to V'$ bezeichne die Inklusion. C_K sei die Konstante aus Lemma (3.11).
Der Operator $T \in \mathfrak{L}(V)$ ist dann definiert durch:

$$T := (A + C_K)^{-1} I \ .$$

Analog definert man im diskreten Fall die Operatoren T_n.

Zwischen den Eigenwerten μ_i von T und den Eigenwerten λ_i von A besteht die Beziehung :

$$\lambda_i = C_K + 1/\mu_i \ .$$

❑

Siehe dazu [9; p.131 und p.126].

Eine hinreichende Bedingung für die stark stabile Konvergenz von T_n wird in [6; Prop. 5.17] gegeben.

Der nächste Satz gibt eine Abschätzung für den Diskretisierungsfehler an:

(4.2.9) Satz

Sei $\Omega \subset \mathbb{R}^n$ beschränkt und $\Gamma := \partial\Omega$ glatt. $H := L^2(\Omega)$, $V := H^1(\Omega)$; $a(\cdot\,,\cdot)$ sei definiert wie in (3.8).
Wir betrachten das Problem:

$$\text{Suche } u \in V \text{ , so daß}$$

$$a(u,v) = \lambda\,(u,v)_H \quad \forall\, v \in V \ .$$

Die Größe Θ ist definiert durch :
Sei H ein Hilbertraum und B, C Unterräume von H

$$\Theta_H(B,C) := \max\{\delta_H(B,C)\,,\ \delta_B(C,B)\},$$

$$\delta_H(B,C) := \sup_{\substack{v \in B \\ \|v\|_H = 1}} \mathrm{dist}_H(v,C)\ ;$$

wobei für $v \in H$, $\mathrm{dist}_H(v,B)$ definiert ist, durch

$$\mathrm{dist}_H(v,B) := \inf_{w \in B} \|v - w\|_H \ .$$

Wir wählen eine Familie $\{V_{h_n}\}$ von endlichdimensionalen Teilräumen von V, die folgendem genügt :

(i) $$V_{h_n} \subset V$$

(ii) $$\text{dist}_V(v, V_{h_n}) \leq C\, h_n^{k-1} \qquad \forall v \in V$$

Dann gilt :

$$|\lambda - \hat{\lambda}_n| = O(h_n^{2k-2})$$
$$\Theta_V\{E(\lambda), E_{h_n}(\lambda_n)\} = O(h_n^{k-1})$$
$$\Theta_H\{E(\lambda), E_{h_n}(\lambda_n)\} = O(h_n^{k}) \ ;$$

mit $\hat{\lambda}_n$ wie in Satz (4.2.7) .
In unserem Fall von linearen finiten Elementen ist k = 2.

❑

Beweis:

siehe [5]

Für den Fall, daß $a(\cdot,\cdot)$ nicht selbstadjungiert ist, sei auf [6; § 6] verwiesen.
Der Fall, daß $\partial\Omega$ nicht glatt ist (z.B. einspringende Ecken), wird in [15] behandelt.

§5 Numerisches Verfahren zur Lösung der diskreten Gleichungen

(5.1) Vorbereitung

In diesem Kapitel wollen wir zunächst für das endlichdimensionale verallgemeinerte Eigenwertproblem:

(5.1.1) $$K u = \lambda M u$$

ein effizientes numerisches Verfahren entwickeln. Dies wird in diesem Fall ein Mehrgitteralgorithmus sein, mit einem modifizierten ILU-Verfahren als Glätter.
Anschließend wird klar werden, daß dieser Algorithmus nach geringfügigen Veränderungen geeignet ist, das reguläre Problem:

$$(K - \lambda M)\, u = f \quad ; \lambda \text{ kein Eigenwert}$$

und das singuläre Problem:

$$(K - \lambda M)\, u = f \quad ; \lambda \text{ Eigenwert}$$

zu lösen.
Da in unserem Problem K und M symmetrische Matrizen sind, wollen wir uns in der folgenden Herleitung und Untersuchung des numerischen Verfahrens auf den symmetrischen Fall beschränken. Falls K und M nicht symmetrisch sind, siehe [8] und [11].

Eindeutigkeit bei singulären Gleichungen :

Wir betrachten das Problem:

(5.1.2) $$(K - \lambda M)\, u = f \ ,$$

wobei K und M selbstadjungiert sein sollen, M regulär angenommen wird, und λ im Spektrum von $M^{-1} K$ liegen soll.

(5.1.3) Satz

Sei $E(\lambda)$ der Eigenraum von Problem (5.1.1) zum Eigenwert λ;
$E(\lambda) := \ker(K - \lambda M)$. Sei $f \perp E(0)$; d.h. $\langle f, v \rangle = 0 \quad \forall\, v \in E(0)$.

Behauptung:
Dann hat (5.1.2) eine eindeutige Lösung $u \perp M(E(0)) := \{ Mv \,,\, v \in E(0) \}$. ❑

Beweis :
siehe [11; p.16]

QED

Um aus einer Eigenvektornäherung " $\tilde{e}$ " eine Approximation für den zugehörigen Eigenwert zu erhalten, benötigen wir den verallgemeinerten *Rayleigh - Quotienten* Λ:

$$\Lambda(\tilde{e}) := \frac{< K\tilde{e}, \tilde{e} >}{< M\tilde{e}, \tilde{e} >} .$$

(5.1.4) Satz

Sei λ der exakte Eigenwert zur Eigenfunktion e, und $\delta := \min_{\alpha \in \mathbb{C}} \{ < M(e - \alpha\tilde{e}), e - \alpha\tilde{e} >^{1/2} \}$.
$\tilde{e}$ sei dabei eine Approximation von e.
Dann gilt :

$$| \lambda - \Lambda(\tilde{e}) | \leq C\delta^2 .$$

Beweis :
siehe [8; Lemma 12.1.2] ❑

Damit sind wir nun in der Lage, das Mehrgitterverfahren zur Lösung des Eigenwertproblems (5.1.1) zu erklären.

(5.2) Konstruktion des Mehrgitterverfahrens

Wir wollen annehmen, eine Grobtriangulierung des Gebiets Ω n-mal verfeinert zu haben (vgl. §6) und daraus mit Hilfe von finiten Elementen wie üblich eine Sequenz von n Matrizen: $(K_\ell - \lambda_\ell M_\ell)$ $1 \leq \ell \leq n$ erzeugt zu haben.

Sei $S_\ell(u_\ell, f_\ell, \lambda_\ell)$ ein Schritt eines linearen Iterationverfahrens (Glättungsschritt) zur Lösung von:

$$(K_\ell - \lambda_\ell M_\ell)\, u_\ell = f_\ell .$$

Auf die Wahl von S_ℓ werden wir in § 6 ausführlich eingehen.
$S_\ell(\lambda_\ell) := S_\ell(u_\ell, 0, \lambda_\ell)$ ist dann ein Glättungsschritt für das Eigenwertproblem. S_ℓ muß folgende Fixpunkt-Eigenschaft besitzen:

$$S_\ell(\lambda_\ell)\, e_\ell = e_\ell \quad \forall\, e_\ell \in E_\ell(\lambda_\ell) .$$

Ist nun mit $\tilde{\lambda}_\ell$ eine Approximation für den exakten Eigenwert λ_ℓ gegeben, hat die Anwendung von ν Glättungsschritten $u_\ell \to \tilde{u}_\ell := S_\ell(\tilde{\lambda}_\ell)^\nu$ denselben Effekt wie bei regulären linearen Gleichungssystemen. Hochfrequente Anteile des Fehlers von u_ℓ werden reduziert. Die Eigenwertkomponente bleibt dabei nahezu unverändert ($\lambda_\ell \approx \tilde{\lambda}_\ell$).

Für die Grobgitterkorrektur benötigen wir den Defekt:

$$d_\ell := (K_\ell - \tilde{\lambda}_\ell M_\ell)\, \tilde{u}_\ell .$$

Um die Darstellung zu vereinfachen, wollen wir annehmen, daß gilt:

$$\dim E_\ell(\lambda_\ell) = 1 \quad \text{und} \quad < M_\ell e_\ell, e_\ell > = 1 .$$

Sei $\tilde{u}_\ell$ zerlegt in $\tilde{u}_\ell = \alpha e_\ell + v_\ell$ mit $v_\ell \perp M\{E_\ell(\lambda_\ell)\}$.
Daraus folgt :

$$d_\ell = (\lambda_\ell - \tilde{\lambda}_\ell)\, \alpha M_\ell e_\ell + (K_\ell - \tilde{\lambda}_\ell M_\ell) v_\ell .$$

Wir führen die Projektion Q_ℓ auf $E_\ell(\lambda_\ell)^\perp$ ein :

$$Q_\ell u_\ell = u_\ell - < u_\ell, e_\ell > M_\ell e_\ell .$$

Da gilt $(K_\ell - \tilde{\lambda}_\ell M_\ell) v_\ell \in E_\ell(\lambda_\ell)^\perp$, folgt :

$$d_\ell{}^\perp := Q_\ell d_\ell = (K_\ell - \lambda_\ell M_\ell) v_\ell .$$

Die exakte Korrektur von $\tilde{u}_\ell$ wäre :

$$\tilde{u}_\ell \rightarrow \tilde{u}_\ell - w_\ell ;$$

wobei w_ℓ die Lösung der Gleichung :

(5.2.1)
$$(K_\ell - \lambda_\ell M_\ell) w_\ell = d_\ell{}^\perp$$

mit $d_\ell{}^\perp \in E_\ell(\lambda_\ell)^\perp$ und $w_\ell \in M_\ell \{ E_\ell(\lambda_\ell)\}$ ist.

Auf Grund von Satz (5.1.3) hat die Gleichung (5.2.1) - falls $\tilde{\lambda}_\ell$ ein Eigenwert ist - eine eindeutige Lösung. Falls $\tilde{\lambda}_\ell$ in der Nähe eines Eigenwertes liegt, folgt die eindeutige Lösbarkeit von (5.2.1) aus einem Störungslemma (siehe [8; Lemma 12.1.7].

Da die Lösung der Defektgleichung :

$$w_\ell = (K_\ell - \tilde{\lambda}_\ell M_\ell)^{-1} d_\ell^\perp$$

mit Rundungsfehlern behaftet sein wird, sollte man w_ℓ mit Hilfe von $\tilde{Q}_\ell$ auf $M_\ell \{E_\ell(\tilde{\lambda}_\ell)\}$ projizieren; d.h. :

$$w_\ell := \tilde{Q}_\ell (K_\ell - \tilde{\lambda}_\ell M_\ell)^{-1} Q_\ell d_\ell ,$$

wobei $\tilde{Q}_l$ definiert ist durch :

$$\tilde{Q}_\ell u_\ell = u_\ell - < u_\ell, M_\ell e_\ell > e_\ell .$$

Wir erhalten also folgende Iterationsvorschrift für ein Zweigitterverfahren zur Lösung von (5.1.1) :

gegeben seien Startnäherungen $\tilde{e}_\ell$, $\tilde{\lambda}_\ell$.

(5.2.3)

Iterationsschritt :

$$\tilde{e}_\ell \rightarrow \tilde{e}_\ell - p_{\ell,\ell-1} \tilde{Q}_{\ell-1} (K_{\ell-1} - \tilde{\lambda}_{\ell-1} M_{\ell-1})^{-1} Q_{\ell-1} r_{\ell-1,\ell} (K_\ell - \tilde{\lambda}_\ell M_\ell) S_\ell(\tilde{\lambda}_\ell)^\nu \tilde{e}_\ell ;$$

wobei r und p die kanonische Restriktion und Prolongation zwischen den Gittern ℓ und ℓ-1 bezeichnet.

Für die Projektion $Q_{\ell-1}$ benötigt man den (approximativen) Eigenvektor $\tilde{e}_\ell$. Diesen erhält man, indem man das Verfahren der geschachtelten Iteration verwendet (vgl. [8; § 5 und § 12.3.3]).

(5.2.4) Algorithmus :

program nested iteration

Berechne e_0 , λ_0 (z.B. durch das QR - Verfahren)

do $\ell = 1$, L_{max} , step 1

$\tilde{e}_{\ell-1} := \tilde{e}_{\ell-1} \; / < M_{\ell-1}\tilde{e}_{\ell-1}, \tilde{e}_{\ell-1} >$

$\tilde{\lambda}_{\ell-1} := \Lambda_{\ell-1}(\tilde{e}_{\ell-1})$

$\tilde{e}_{\ell} := p_{\ell,\ell-1}\tilde{e}_{\ell-1}$

do $j = 1, I_{Abbruch}$, step 1 call emg(ℓ, $\tilde{e}_{\ell}$)

continue

end

program eigenvalue multigrid iteration (emg)

subroutine emg (ℓ, u)

$\lambda := \Lambda_{\ell}(u)$

$u := S_{\ell}(\lambda)^{\nu} u$

$d := Q_{\ell-1}\, r_{\ell-1,\ell}\, (K_{\ell} - \lambda M_{\ell})\, u$

$v := 0$

do $j = 1, \gamma$, step1 call smg(ℓ - 1, v, d)

$u := u - p_{\ell,\ell-1}\, v$

end

program singular multigrid iteration (smg)

subroutine smg(ℓ, u, f)

if ($\ell = 0$) then $u := \tilde{Q}_0\, (K_0 - \tilde{\lambda}_0 M_0)\, Q_0 f$

else

$u := S_{\ell}^{\nu}(u, f, \tilde{\lambda}_{\ell})$

$d := Q_{\ell-1}\, r_{\ell-1,\ell}\, (K_{\ell}\, u - \tilde{\lambda}_{\ell} M_{\ell}\, u - f)$

$v := 0$

do $j = 1, \gamma$, step 1 call smg (ℓ - 1, v, d)

endif

$u := \tilde{Q}_{\ell}\, (u - p_{\ell,\ell-1} v)$

end

(5.2.5) Bemerkung

Die Größe γ in Algorithmus (5.2.4) bestimmt den Zyklus des Mehrgitterverfahrens. $\gamma = 1$ ergibt einen V - Zyklus; $\gamma = 2$ einen W - Zyklus (vgl. [8; §2.5]).

❑

(5.3) Konvergenz des Verfahrens

Um die Konvergenz des in (5.2) entwickelten Verfahrens zu beweisen, zeigt man in Anlehnung an die von [8; § 6] vorgestellten Theorie, daß der Algorithmus (5.2.3) die Glättungs - und die Approximationseigenschaft besitzt.

Wir betrachten dazu zunächst den Zweigitteralgorithmus (5.2.3) und zerlegen diesen gemäß:

$$A := (K_\ell - \lambda_\ell M_\ell)^{-1} - p_{\ell,\ell-1}\, \tilde{Q}_{\ell-1}\, (K_{\ell-1} - \lambda_{\ell-1} M_{\ell-1})^{-1}\, Q_{\ell-1}\, r_{\ell-1,\ell}$$

$$G := (K_\ell - \lambda_\ell M_\ell)\, S_\ell^{\nu}(\lambda_\ell)$$

(5.2.3´)
$$\tilde{e}_\ell^{\,i+1} := A \cdot G\, \tilde{e}_\ell^{\,i} \quad .$$

Damit der Algorithmus (5.2.3´) konvergiert, muß das Produkt A·G in einer geeigneten Norm kleiner eins sein. Dazu definieren wir zunächst die folgenden Normen.

(5.3.1) Definition

Sei $u \in \mathbb{R}^n$. $\|\cdot\|$ bezeichnet die euklidische Norm ($\|u\| := \{ \sum_{i=1}^{n} u_i^2 \}^{1/2}$).

$|||\cdot|||$ sei die zugehörige Operatornorm ($|||T||| := \sup_{\substack{u \in \mathbb{R}^n \\ u \neq 0}} \frac{\|Tx\|}{\|x\|}$ für $T \in \mathfrak{L}(\mathbb{R}^n)$).

Wir werden nun noch weitere -gitterabhängige- Normen definieren:
Sei dazu $\Omega \in \mathbb{R}^2$ ein Gebiet, welches mit Hilfe der Triangulierung τ in Dreiecke zerlegt sei (vgl.§6.3); n bezeichne die Anzahl aller Gitterpunkte von τ. Mit Hilfe der finiten Elemente Ansatzfunktionen $\{b_i\}_{1 \le i \le n}$ können wir den Isomorphismus $P : \mathbb{R}^n \to V_n \subset V$ ($=H^1(\Omega)$) definieren (vgl.§4). Sei $R := P'$ die zu P adjungierte Abbildung. Mit der Massenmatrix M = R*P gilt dann:

$$(Pu\,,\, Pv)_0 := \langle u\,,\, Mv \rangle \quad .$$

Da M positiv definit ist, können wir ein gitterabhängiges Skalaprodukt und damit eine gitterabhängige Norm auf dem $\mathbb{R}^n$ definieren durch:

$$\langle u\,,\, v \rangle_M := \langle u\,,\, Mv \rangle$$

$$\|u\|_M := (\langle u\,,\, u \rangle_M)^{1/2} \quad \forall\, u,v \in \mathbb{R}^n.$$

Sei $\mathfrak{K}_n$ der Operator, welcher auf Grund von Lemma 3.11 der Bilinearform $a(\cdot,\cdot) : V_n \times V_n \to \mathbb{R}$ eineindeutig zugeordnet werden kann. $\mathfrak{K}_n$ (und nicht die Steifigkeitsmatrix K_n!) stellt die Approximation des diskretisierten Differentialoperators dar. Es gilt:

$$K_n := R\, \mathfrak{K}_n\, P$$

(vgl. [9; Lemma 8.1.7]). Um K_n in der Norm den Charakter eines Differentialoperators ($\mathfrak{K}_n$) zu geben, versehen wir den Bildraum von K_n mit der euklidischen Norm und den Urbildraum mit der Norm $\|\cdot\|_M$. Dadurch können wir die Norm $|||\cdot|||_M$ bzw. $|||\cdot|||_{-M}$ definieren, durch:

$$\|\!|K_n|\!\|_M := \sup_{\substack{u \in \mathbb{R}^n \\ u \neq 0}} \frac{\|K_n\, u\|}{\|u\|_M}$$

bzw.

$$\|\!|K_n^{-1}|\!\|_{-M} := \sup_{\substack{u \in \mathbb{R}^n \\ u \neq 0}} \frac{\|K_n^{-1}\, u\|_M}{\|u\|} \quad . \qquad \square$$

Da der Glätter S (vgl. Definition 5.3.2), auf Grund der Räume in denen er operiert, als Operator des $\mathbb{R}^n$ in sich zu interpretieren ist (siehe §6.1), werden wir S in der Norm $\|\!|\cdot|\!\|$ messen.

Man definiert dann den Begriff *Glättungseigenschaft* für Iterationsverfahren zur Lösung eines Eigenwertproblems. Der Einfachheit halber wird hier angenommen, der Glätter S sei linear (vgl. [11]).

(5.3.2) Definition

Eine Abbildung

$$S_\ell : V_\ell \times \mathbb{C} \times V_\ell$$

erfüllt die *Glättungseigenschaft* für Eigenwertprobleme genau dann, wenn folgende Bedingungen erfüllt sind :

Sei λ_ℓ der kleinste Eigenwert von (5.1.1). Dann gelte :

(i) $\exists\, \alpha > 0$, $\nu_0 \in \mathbb{N}$, eine Funktion $\eta(\nu)$ mit $\eta(\nu) \to 0$; $\nu \to \infty$
und $\bar{\nu}(h_\ell) \in \mathbb{N}$; $\bar{\nu}(h_\ell) > \nu_0$

so, daß gilt :

$$\|\,(K_\ell - \lambda_\ell M_\ell)\, S_\ell(\lambda_\ell)^\nu z_\ell\,\| \le \eta(\nu)\, h_\ell^{-\alpha}\, \|z_\ell\|_M$$
$$\forall\, z_\ell \perp M_\ell\{E_\ell(0)\} \quad \text{und} \quad \nu_0 \le \nu \le \bar{\nu}(h_\ell) \,.$$

Sei e_ℓ eine Eigenfunktion zu λ_ℓ. Dann gelte :

(ii) $$S_\ell(\lambda_\ell)\, e_\ell = e_\ell \,.$$

(iii) $$\|\,\{S_\ell^\nu(\lambda) - S_\ell^\nu(\tilde{\lambda})\}\, z\,\| \le C\,|\lambda - \tilde{\lambda}|\,\|z\| \quad \forall \lambda \in \mathbb{C},\ \tilde{\lambda} \in U(\lambda) \,.$$

(5.3.3) Bemerkung

Bedingung (*i*) ist die übliche Glättungseigenschaft, wie sie z.B. aus der Behandlung elliptischer Differentialgleichungen bekannt ist. Die Bedingungen (*ii*) und (*iii*) sichern die Verträglichkeit mit dem Eigenwertproblem. $\square$

Um die Konvergenz von (5.2.3) zu zeigen, benötigen wir noch eine Abschätzung für den Operator A aus (5.2.3´). Dieses Problem wird in [7] ausführlich behandelt. Es soll hier nur zitiert werden, daß unter geeigneten Voraussetzungen gilt:

$$\|\!|A|\!\|_M \le C h_\ell^{\alpha} \quad ;$$

mit $C := C(h_\ell / h_{\ell-1})$ und α aus Definition (5.3.2).

Man folgert nun weiter, wie bei Hofmann [11; p.31] und erhält schließlich über die Konvergenz des Zweigitterverfahrens die Konvergenz eines Mehrgitterverfahrens, welches exakte Projektionen Q_ℓ, $\tilde{Q}_\ell$ benutzt, und schließlich die Konvergenz von Algorithmus (5.2.3).

(5.3.4) Satz

Seien die durch die geschachtelte Iteration gelieferten approximativen Eigenvektoren und damit die Projektionsabbildungen Q_ℓ, $\tilde{Q}_\ell$ hinreichend genau. Das Mehrgitterverfahren zur Lösung der singulären Defektgleichung bestehe aus mehr als einer Iteration. Prolongation und Restriktion seien kanonisch definiert (vgl. [8; §3.6]. Der Glätter $S_\ell(\lambda_\ell)$ besitze die Glättungseigenschaft (5.3.2).

Dann gilt:
1) Der beschriebene Mehrgitteralgorithmus (5.2.3) ist konvergent.

2) Es gelten die Fehlerabschätzungen:

a) $$\| e_\ell^{i+1} - e_\ell \| \leq C\, \eta(\nu)\, \| e_\ell^i - e_\ell \|;$$

wobei e_ℓ die exakte Lösung von (5.1.1) bezeichnet.

b) $$| \lambda - \Lambda(e_\ell^i) | \leq C\, \Theta_i^2$$

wobei Λ den Rayleighquotienten bezeichnet, und Θ_i definiert ist durch :

$$\Theta_i := \frac{\| e_\ell^{i+1} - e_\ell \|}{\| e_\ell^i \|} .$$

❑

Beweis :

vgl [11] und [7]

QED

§6 Ein modifiziertes ILU-Verfahren zur Lösung singulärer linearer Gleichungssysteme

In diesemAbschnitt werden wir für unser Problem ein modifiziertes ILU-Verfahren zur Lösung großer singulärer linearer Gleichungssysteme (LGS) entwickeln und dessen Verhalten bei unregelmäßigen Diskretisierungen untersuchen.

(6.1) Iterationsverfahren zur Lösung linearer Gleichungssysteme

Wir betrachten das LGS:

(6.1.1) $$A\,x = b\ ,$$

wobei A eine reguläre n x n Matrix ist und $b, x \in \mathbb{R}^n$.

Ein lineares Iterationsverfahren zur Lösung von (6.1.1) läßt sich beschreiben durch:

(6.1.2) a) x_0 sei ein gegebener Startvektor

b) Iterationsvorschrift: $x_{i+1} := x_i - \tilde{A}^{-1} (Ax_i - b)$;

wobei die n x n Matrix $\tilde{A}$ regulär sein soll und A approximieren soll.

(6.1.3) Definition *Spektralradius*

Der *Spektralradius* ρ einer n x n Matrix A ist definiert durch :

$$\rho(A) := \max \{ |\lambda| : \lambda \text{ ist Eigenwert von } A \}$$

❑

(6.1.4) Satz

Behauptung:

(*i*) (6.1.2) beschreibt eine konvergente Iteration genau dann, wenn

$$\rho(I - \tilde{A}^{-1} A) < 1$$

erfüllt ist.

(*ii*) Ist (6.1.2) konvergent, so hat (6.1.2) genau die Lösung von (6.1.1) als Fixpunkt.

❑

Beweis :
(*i*) siehe [20; Theorem 1.4]

(*ii*) (6.1.2) konvergiere gegen y. Dann gilt:

$$y = y - \tilde{A}^{-1} (A y - b) \Leftrightarrow A y = b ,$$

da $\tilde{A}$ regulär ist.

QED

(6.2) Das ILU-Verfahren

Sei das LGS : Ax=b mit der sehr großen aber schwach besetzten Matrix A gegeben. Diese Situation liegt beispielsweise bei der Diskretisierung partieller Differentialgleichungen mit finiten Elementen oder finiten Differenzen vor (vgl. §4). Eine LU - Zerlegung (bzw. eine Cholesky-Zerlegung im symmetrischen Fall)

$$A = L U$$

ist in diesem Fall nicht durchführbar, da die oberen - bzw. unteren Dreiecksmatrizen L und U i.A. wegen des "fill-in" nicht mehr schwach besetzt sind.
Die Idee der unvollständigen LU-Zerlegung (ILU) ist, die Aufspaltung A = L U zu ersetzen durch

(6.2.1)
$$A = L U - C$$

mit schwach besetzten oberen bzw. unteren Dreiecksmatrizen L und U und einer Restmatrix C.

(6.2.2) Definition

Das *Besetzungsmuster* $\mathfrak{P}_A$ einer n x n Matrix $\{A_{i,j}\}_{1 \le i,j \le n}$ ist gegeben durch:

$$\mathfrak{P}_A := \{(i\,j) : A_{i,j} \neq 0 : 1 \le i\,j \le n\}$$

❑

Definiert man jetzt für L und U die Besetzungsmuster $\mathfrak{P}_L$ und $\mathfrak{P}_U$, dann berechnet man beim ILU-Verfahren diejenigen Elemente $L_{i,j}$ $(U_{i,j})$ für die gilt:

$$(i\,j) \in \mathfrak{P}_L \quad (\in \mathfrak{P}_U)$$

nach der üblichen LU-Zerlgung. Alle anderen Elemente von L bzw. U werden während der Zerlegung null gesetzt. Die Matrix C ist dann definiert durch:

$$C := A - L U .$$

Die üblichen Voraussetzungen an $\mathfrak{P}_L$, $\mathfrak{P}_U$ und $\mathfrak{P}_C$ sind:

(6.2.3) a) $$\mathfrak{P}_L \cap \mathfrak{P}_C = \mathfrak{P}_U \cap \mathfrak{P}_C = \varnothing$$

(6.2.3) b) $$\mathfrak{P}_A \subset \mathfrak{P}_L \cup \mathfrak{P}_U .$$

Wir können die Faktoren L und U normieren durch die Bedingung:

$$L_{i,i} = U_{i,i} \quad 1 \leq i \leq n.$$

Damit können wir Gleichung (6.2.1) auf die Form:

(6.2.1´) $$A = (L' + D)\, D^{-1}\, (U' + D) - C$$

bringen, wobei L´ (U´) strikte obere (untere) Dreiecksmatrizen sind und D diagonal sein soll.

Das ILU-Verfahren zur Lösung von (6.1.1) lautet dann:

(6.2.4)
a) gegeben sei ein Startvektor x_0
b) Iterationsschritt : $x_{i+1} = x_i - (LU)^{-1} (A x_i - b)$

wobei L und U wie oben definiert sein sollen.

Um eine ILU-Zerlegung bei sehr großen Matrizen durchführen zu können, ist es entscheidend, die Besetzungsstruktur der Matrix a priori zu kennen. Beispielsweise ist die Steifigkeitsmatrix K aus (4.1.2) zwar schwach bestzt, jedoch ist bei unregelmäßigen Gittern das Besetzungsmuster $\mathfrak{P}_K$ völlig unstrukturiert. Hat man jedoch feste Nachbarbeziehungen zwischen den Gitterpunkten vorliegen, z.B. 5 oder 7 Punkt Sterne (vgl. [9; § 4.2]), besitzt K Bandstruktur und die ILU- Zerlegung läßt sich optimal durchfüren.

(6.3) Gittergenerierung

In diesem Abschnitt wird eine Methode vorgestellt, um bei finite Elemente Diskretisierungen einerseits problemorientierte, andererseits strukturierte Gitter zu erhalten.

(6.3.1) Definition

$\tau := \{T_1, T_2, \ldots, T_t\}$ heißt *zulässige Triangulierung* von Ω, falls folgende Bedingungen erfüllt sind :

a) T_i sind offene Dreiecke für $1 \leq i \leq t$

b) die finiten Elemente sind disjunkt , d.h. : $T_i \cap T_j = \varnothing$ für $i \neq j$

c) $\bigcup_{1 \leq i \leq t} \overline{T}_i = \overline{\Omega}$

d) für $i \neq j$ ist $\overline{T}_i \cap \overline{T}_j$ entweder
   (*i*) leer, oder
   (*ii*) eine gemeinsame Seite der Elemente T_i und T_j , oder
   (*iii*) eine gemeinsame Ecke der Elemente T_i und T_j .

❑

Wir wollen nun drei Strategien erklären, um eine Folge verschieden feiner Gitter auf Ω zu erzeugen:

1. regelmäßige Verfeinerung

Die einfachste Möglichkeit, eine Hierarchie verschieden feiner Gitter $\tau_0, \tau_1, \ldots, \tau_N$ zu erzeugen, ist ausgehend von einer zulässigen Grobtriangulierung τ_0 sukkzesive feinere Gitter $\tau_1, \tau_2, \ldots, \tau_N$ zu gewinnen, indem man die Seitenmitten der Dreiecke einer alten Triangulierung $\tau_{\ell-1}$ miteinander verbindet und so eine feinere Triangulierung τ_ℓ erhält, welche aus 4- mal so vielen Dreiecken besteht wie $\tau_{\ell-1}$.

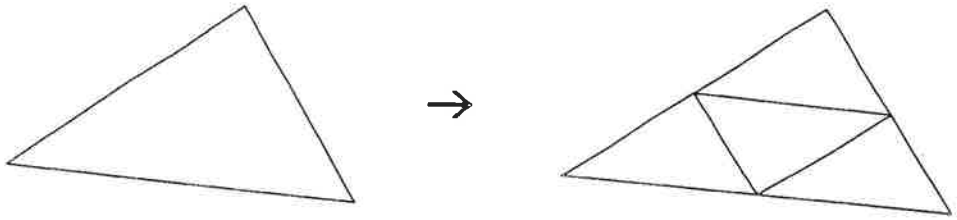


Abb. 2 regelmäßige Verfeinerung eines Dreiecks

Man prüft leicht nach, daß jede so entstandene Triangulierung τ_ℓ zulässig im Sinne von Definition (6.3.1) ist.

2. adaptive Verfeinerung

Eine verbreitete Möglichkeit, problemangepaßte Gitter zu erhalten, ist an "kritischen" Stellen des Gebietes, wie oben beschrieben, regelmäßig zu verfeinern und zu versuchen, die Verfeinerung möglichst lokal zu halten. Dies erreicht man, falls man eine Halbierung der Dreiecke zuläßt (siehe z.B. [8; §3.8.2]).

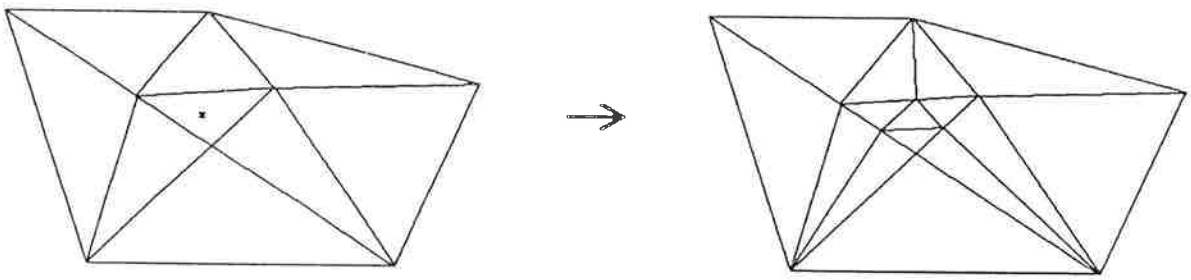


Abb.3 lokale Verfeinerung um das Dreieck "x".

Das Dreieck x wird regelmäßig verfeinert; dessen Nachbardreiecke halbiert; und die übrigen Dreiecke unverändert gelassen.

Um die kritischen Stellen automatisch zu lokalisieren, verwendet man üblicherweise Fehlerschätzer, wie sie z.B. in [21] beschrieben werden.
Dabei entstehen nun aber ungleichmäßig strukturierte Gitter, was bedeutet, daß das Besetzungsmuster der Steifigkeitsmatrix weder a priori bekannt noch regelmäßig ist.

3. problemangepaßte regelmäßige Diskretisierung

Um die Schwierigkeit von unstrukturierten Matrizen zu vermeiden, machen wir die folgende Vorüberlegung:

Die zu untersuchende partielle Differentialgleichung ist analytisch und numerisch gutartig: H^1 -koerziv (vgl §2,§3). Jedoch macht der Rand unseres Gebietes Ω (Bodensee) wegen seiner Komplexität (z.B.einspringende Ecken) Schwierigkeiten. Es wäre also günstig, ein regelmäßiges Gitter zu erzeugen, welches am Rand von Ω wesentlich feiner ist als im Innern, und dann regelmäßig zu verfeinern.

Das folgende Verfahren läßt sich anwenden im Falle beschränkter, einfach zusammenhängender, null homotoper Gebiete des $\mathbb{R}^2$ und zwar umso besser je mehr das Gebiet einem langgestreckten Rechteck ähnelt.

Gittergeneration

a) Konstruktion des Grobgitters

Wir wollen nun die Grobtriangulierung τ_0 an folgendem Testgebiet Ω erklären:

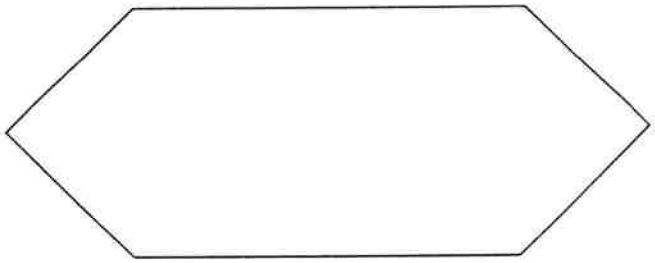


Abb. 4 Testgebiet

Man legt zunächst eine Konturlinie K in Ω, die einen näherungsweise konstanten, möglichst geringen Abstand zum Rand $\partial\,\Omega$ besitzen soll.

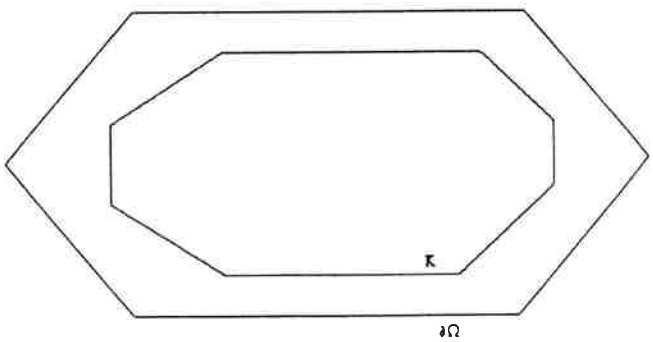


Abb. 5 Konturlinie in Ω

Danach legt man in den Zwischenraum von K und $\partial\,\Omega$ Dreiecke so nebeneinander, daß für jedes dieser Dreiecke T_i eine der folgenden Bedingungen erfüllt ist:

1.) Eine Ecke von T_i liegt auf $\partial\,\Omega$ und zwei Ecken auf K.

2.) Eine Ecke von T_i liegt auf K und zwei Ecken auf $\partial\,\Omega$.

Ferner muß gelten:

3.) Erfüllt ein Dreieck T_i die Eigenschaft (1) dann erfüllen seine Nachbardreiecke die Eigenschaft (2) bzw. umgekehrt.

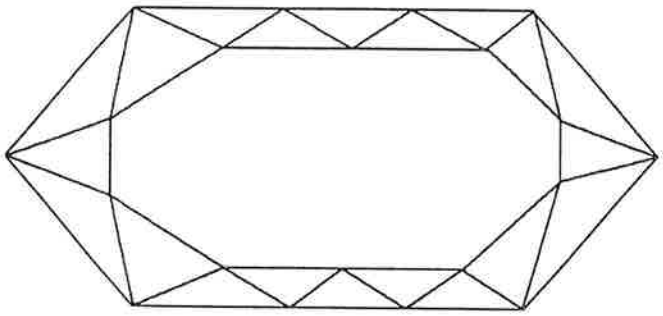


Abb. 6 Randtriangulierung

4.) Die so entstandenen Eckpunkte K_i der Randdreiecke, welche auf K liegen, müssen sich paarweise durch Strecken verbinden lassen so, daß im Innern des von K berandeten Gebiets (nicht entartete) Vierecke$\{V_i\}$ entstehen. Dabei ist darauf zu achten, daß in den entstandenen Vierecken der Quotient aus größtem und kleinstem Innenwinkel möglichst klein wird.

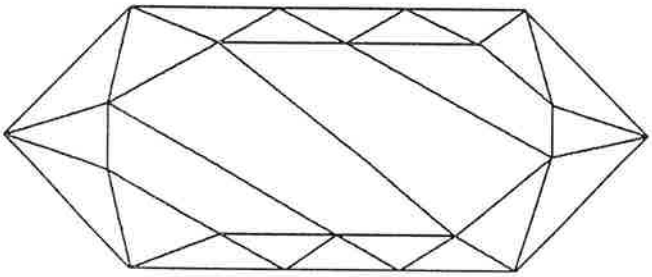


Abb.7 Unterteilung des Inneren von K in Vierecke mit **zu spitzen/stumpfen** Winkeln

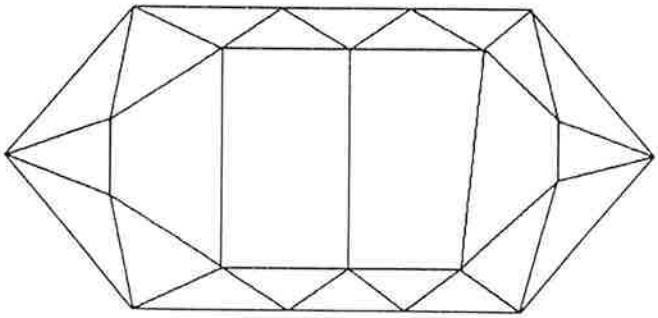


Abb.8 Unterteilung des Inneren von K in Vierecke mit **günstigen** Innenwinkeln

Man teilt nun jedes ViereckV_i in zwei Teile durch eine Diagonale S_i von V_i . Für die Strecken $\{S_i\}$ muß gelten, daß je zwei Strecken S_i und S_j $(i \neq j)$ sich nicht berühren dürfen.

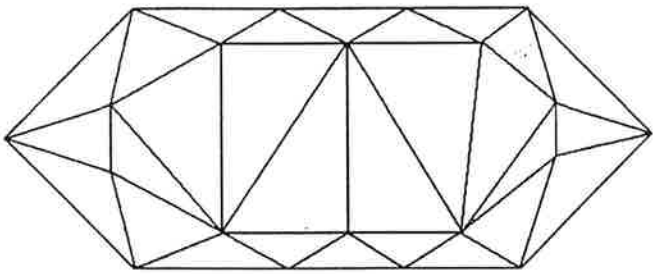


Abb.9 unzulässige Unterteilung der inneren Vierecke

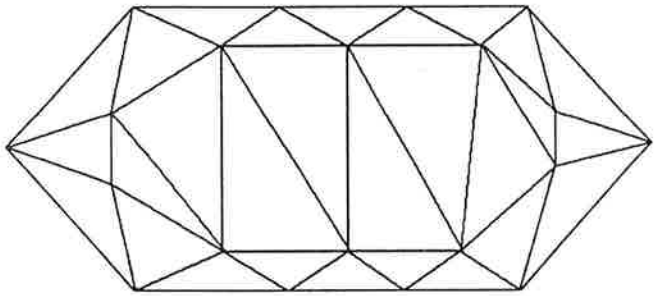


Abb.10 zulässige Grobtriangulierung

b) Verfeinerung des Grobgitters

Man erhält rekursiv aus einem Gitter $\tau_{\ell-1}$ ein feineres Gitter τ_ℓ indem man alle Dreiecke von $\tau_{\ell-1}$ regelmäßig verfeinert.

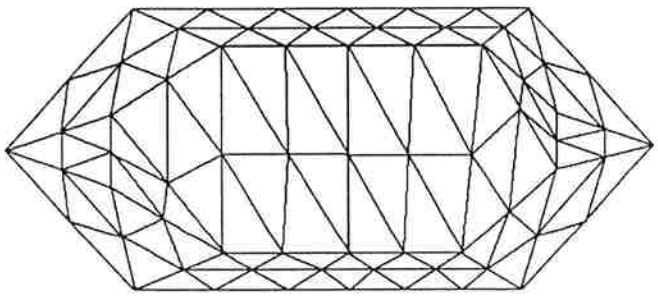


Abb.11 erste Verfeinerung der Grobtriangulierung aus Abb.10

(6.4) Numerierung der Gitterpunkte

(6.4.2) Definition *Knotenringe einer Triangulierung*

Seien NT_ℓ die Anzahl aller Dreiecke einer Triangulierung τ_ℓ und NP_ℓ die Anzahl aller Ecken dieser Dreiecke (Knotenpunkte). $T_\ell := \{T_{\ell,m}\}_{1 \le i \le NT_\ell}$ sei die Menge aller Dreiecke von τ_ℓ. $P_\ell := \{P_{\ell,i}\}_{1 \le i \le NP_\ell}$ sei die Menge aller Knotenpunkte einer Triangulierung τ_ℓ .

Sei $NR_\ell := \begin{cases} 3 \cdot 2^{\ell-1} + 1 & \text{für } \ell > 0 \\ 2 & \text{für } \ell = 0 \end{cases}$

Wir fassen nun Teilmengen von P_ℓ zusammen zu *Knotenringen* : $R_{\ell,i}$ $1 \le i \le NR_\ell$; rekursiv definiert durch :

$$R_{\ell,1} := \{P_{\ell,m} \in P_\ell \, ; \, P_{\ell,m} \in \partial\Omega \, ; \, 1 \le m \le NP_\ell\}$$

Für $1 \le i \le NR_\ell - 2$ definieren wir Ω_{i+1} durch :

$\Omega_{i+1} := \Omega \setminus \{T_{\ell,m}, 1 \le m \le NT_\ell$, mindestens ein Eckpunkt von $T_{\ell,m}$ liegt in $\bigcup_{k=1}^{i} R_{\ell,k}\}$.

$$R_{\ell,i+1} := \{P_{\ell,m} \in P_\ell \, ; \, P_{\ell,m} \in \partial\Omega_{i+1} \, ; \, 1 \le m \le NP_\ell\}$$

Der innerste Knotenring R_{ℓ,NR_ℓ} besteht aus den übrigen Punkten :

$$R_{\ell,NR_\ell} := \{P_{\ell,m} \in P_\ell ; P_{\ell,m} \notin \bigcup_{k=1}^{NR_\ell - 1} R_{\ell,k} ; 1 \le m \le NP_\ell\}$$

Sei R_ℓ die Menge aller Knotenringe einer Triangulierung τ_ℓ :

$$R_\ell := \{R_{\ell,i}\}_{1 \le i \le NR_\ell} \, .$$ ❑

Wir numerieren nun eine Triangulierung τ_l entlang von Punktringen gegen den Uhrzeigersinn von außen nach innen durch. Dies müssen wir aus technischen Gründen genau fassen, wobei dabei einige einfache geometrische Überlegungen ohne Beweis verwendet werden.

a) Numerierung der Grobtriangulierung

Für die Grobtriangulierung τ_0 existieren genau zwei Gitterpunkte A und B , für die gilt:

$\exists!$ 4 Dreiecke T_1 , ..., T_4 , für die gilt:

A bzw. B ist gemeinsame Ecke von T_1 , ..., T_4.

Einer dieser beiden Punkte erhält nun die Nummer $NP_0/2+1$. OBdA sei dies A.
Es existiert nun genau ein Dreieck T von τ_0 , für das gilt:

a) $A \in T$

b) Zwei Ecken E_1 , E_2 von T liegen auf $\partial\Omega$.

Diejenige Ecke E_i, $i \in \{1,2\}$ von T für die gilt:

$$\det(\overrightarrow{A - E_i}, \overrightarrow{E_j - E_i}) > 0 \, ; \; i \ne j \, ; \; i,j \in \{1,2\}$$

erhält die Nummer 1.

Danach numerieren wir die restlichen Punkte aus $R_{0,1}$ bzw. $R_{0,2}$ startend bei E_i und danach bei A fortlaufend gegen den Uhrzeigersinn durch.

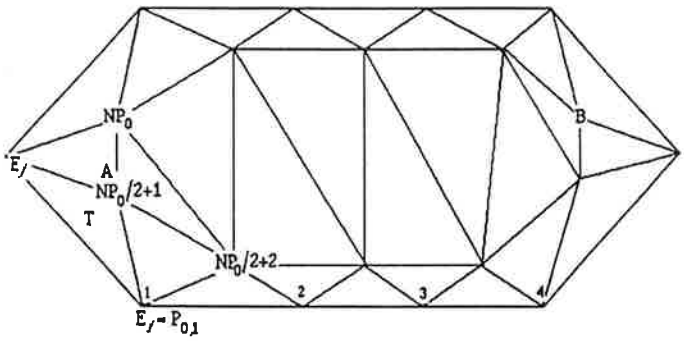


Abb.12 Numerierung der Knotenpunkte der Grobtriangulierung.

b) Numerierung der verfeinerten Grobtriangulierung

Wir betrachten nun das ℓ-mal verfeinerte Grobitter τ_ℓ ($\ell \geq 1$)
Wir legen sukkzesive denjenigen Punkt mit der niedrigsten Nummer aus den Knotenringen $R_{\ell,i}$ fest und numerieren dann die übrigen Punkte des jeweiligen Ringes gegen den Uhrzeigersinn fortlaufend durch.
Es gilt : $P_{0,1} \in P_\ell$. Dieser Punkt erhält auch auf dem Gitter τ_ℓ die Nummer 1.
Weiter gilt: $P_{0,NP0/2+1} \in P_\ell$.

Sei m ;$1 \leq m \leq NR_\ell$, so daß $P_{0,NP0/2+1} \in R_{\ell,m}$.

Wir legen dann für $0 \leq k \leq m$ den Punkt P_{ℓ,N_k} mit der niedrigsten Nummer N_k eines Ringes $R_{\ell,k}$ fest durch:

P_{ℓ,N_k} liegt auf der Geraden durch $P_{0,1}$ und $P_{0,NP0/2+1}$, und $P_{\ell,N_k} \in R_{\ell,k}$.

Für Ringe $R_{\ell,k}$ mit $k >$ m legen wir P_{ℓ,N_k} bzw. N_k fest durch :

1.) $P_{\ell,N_k} \in R_{\ell,k}$

2.) P_{ℓ,N_k} ist Ecke eines Dreiecks, welches $P_{\ell,N_{k-1}+1}$ und $P_{\ell,N_{k-1}+NQ}$ als Ecken besitzen soll; wobei NQ die Anzahl der Elemente von $R_{\ell,k-1}$ bezeichnen soll.

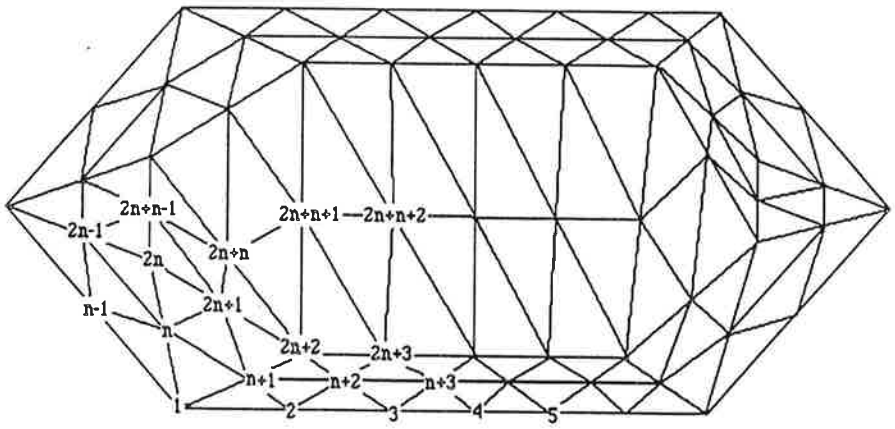


Abb. 13 Numerierung der ersten Verfeinerung von der Grobtriangulierung aus Abb. 12; wobei n-1 die Anzahl der Punkte auf dem äußersten Ring bezeichnet.

(6.5) Struktur der Steifigkeitsmatrix

Hat man nun für ein Gebiet gemäß angegebener Triangulierungsvorschrift (3) eine Folge von Gittern $\tau_0,\ldots,\tau_{\ell_{max}}$ erzeugt und anschließend die zugehörigen Knotenpunkte P_ℓ mit Hilfe der in (6.4) angegebene Numerierungsvorschrift numeriert, ist a priori die Struktur des Gitters und damit die Struktur der zugehörigen Matrix bekannt.

Dies illustriert der folgende Satz:

(6.5.1) Satz

Sei für ein Gebiet Ω gemäß angegebener Gitterkonstruktion (3) eine Folge von Gittern $\tau_0,\ldots,\tau_{\ell_{max}}$ erzeugt.
τ_0 bestehe aus NT_0 Dreiecken.

Dann gelten für die Anzahl der Grobgitterpunkte NP_0, die Anzahl der Knotenringe NR_0 und die Anzahl $NPR_{0,i}$ $(1 \le i \le NR_0)$ der Gitterpunkte auf dem Knotenring $R_{0,i}$, die Beziehungen:

$$NP_0 = (NT_0 + 2) / 3$$

$$NR_0 = 2$$

$$NPR_{0,i} = NP_0 / 2 \qquad i \in \{0,1\} \quad .$$

Sei τ_ℓ die Triangulierung die entsteht , wenn man τ_0 ℓ- mal regelmäßig verfeinert. Dann gelten für die Größen: NT_ℓ, NP_ℓ und NR_ℓ die Beziehungen:

$$NT_\ell = 4^\ell \cdot NT_0$$

$$NP_\ell = (3 \cdot 2^\ell + 1) \cdot 2^{\ell-1} \cdot \frac{NT_0 + 2}{3} - 4^\ell + 1$$

$$NR_\ell = 3 \cdot 2^{\ell-1} + 1$$

Seien die Knotenringe $R_{\ell,m}$ $(1 \le m \le NR_\ell)$ von außen nach innen fortlaufend durchnumeriert.
Sei $NPR_{\ell,m}$ $(1 \le m \le NR_\ell)$ die Anzahl der Gitterpunkte auf dem m.ten Knotenring.
Dann gilt :

$$NPR_{\ell,i} = 2^\ell \cdot \frac{NT_0 + 2}{3} \qquad (1 \le i \le 2^\ell + 1)$$

$$NPR_{\ell,i} = 2^\ell \cdot \frac{NT_0 + 2}{3} - 8 \cdot (i - 2^\ell - 1) \qquad (2^\ell + 1 \le i \le NR_\ell - 1)$$

$$NPR_{\ell,NR_\ell} = 2^{\ell-1} \cdot \frac{NT_0 - 10}{3} + 1 \ .$$

Das bedeutet :
für $1 \le i \le 2^\ell + 1$ liegt ein Gitterpunkt $P_{\ell,j}$ im i.ten Knotenring $R_{\ell,i}$ (von außen gezählt), wenn

$$(i - 1) \cdot 2^\ell \cdot \frac{NT_0 + 2}{3} + 1 \le j \le i \cdot 2^\ell \cdot \frac{NT_0 + 2}{3} \ .$$

für $2^\ell + 1 \le i \le 3 \cdot 2^{\ell-1}$ liegt ein Gitterpunkt $P_{\ell,j}$ im i.ten Knotenring $R_{\ell,i}$ (von außen gezählt), wenn

$$(i - 1) \cdot 2^\ell \cdot \frac{NT_0 + 2}{3} - 4 \cdot (i - 2^\ell - 1) \cdot (i - 2^\ell - 2) + 1 \le j \le$$

$$i \cdot 2^\ell \cdot \frac{NT_0 + 2}{3} - 4 \cdot (i - 2^\ell) \cdot (i - 2^\ell - 1)$$

Ein Gitterpunkt $P_{\ell,j}$ liegt auf dem innersten Knotenring, wenn :

$$2^{2\ell-1} \cdot (NT_0 + 2) - 2^{\ell+1} \cdot (2^{\ell-1} - 1) + 1 \le j \le (3 \cdot 2^\ell + 1) \cdot 2^{\ell-1} \cdot \frac{NT_0 + 2}{3} - 4^\ell + 1.$$

❑

Der Beweis von Satz (6.5.1) beruht auf einfachen geometrischen Überlegungen, die dann per Induktion bewiesen werden können. Er wird hier nicht ausgeführt.

(6.5.2) Bemerkung

Falls für zwei Gitterpunkte P_i und P_j $(i \neq j)$ einer Triangulierung τ_ℓ gilt :

P_i und P_j sind nicht durch eine Dreieckseite miteinander verbunden,

folgt auf Grund der in §4.1 gewählten Diskretisierung für die Massen - bzw. Steifigkeits-matrix :

$$K_{i,j} = 0 \quad \text{und} \quad M_{i,j} = 0.$$

Wir können nun mit Hilfe von Satz (6.5.1) und Bemerkung (6.5.2) die Besetzungs-struktur der Matrizen K und M angeben.

In den folgenden Schemata bedeutet ein "x" , daß das Matrixelement, welches sich an dieser Stelle befindet, i.a. nicht veschwindet. Alle anderen Matrixelemente sind null.

K=

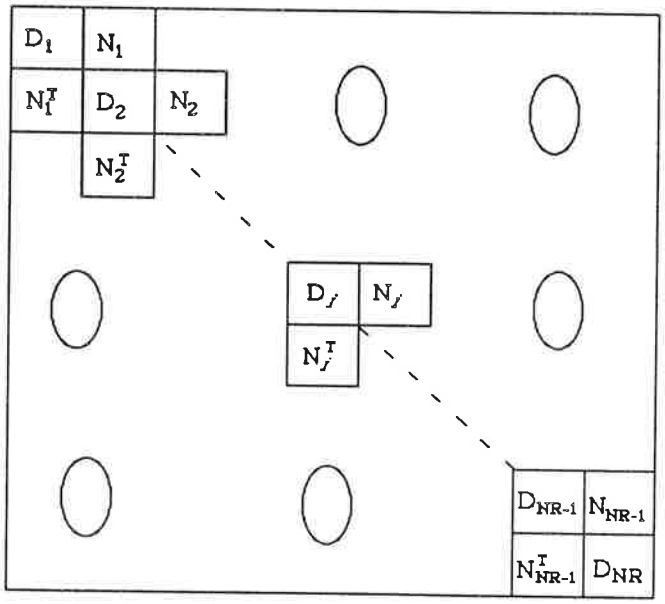
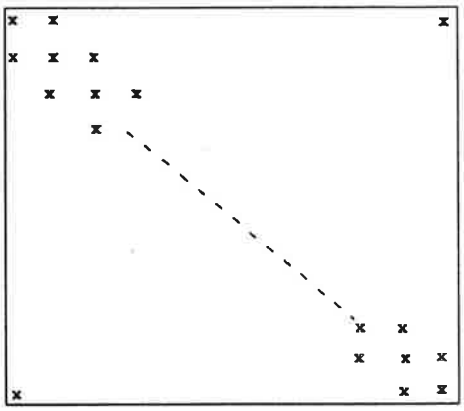


Abb.14 Struktur der Steifigkeitsmatrix

D_i sind $NPR_{\ell,i}$ x $NPR_{\ell,i}$ Matrizen $(1 \leq i \leq NR_\ell)$ und N_ℓ $NPR_{\ell,i}$ x $NPR_{\ell,i+1}$ Matrizen $(1 \leq i \leq NR_\ell - 1)$.

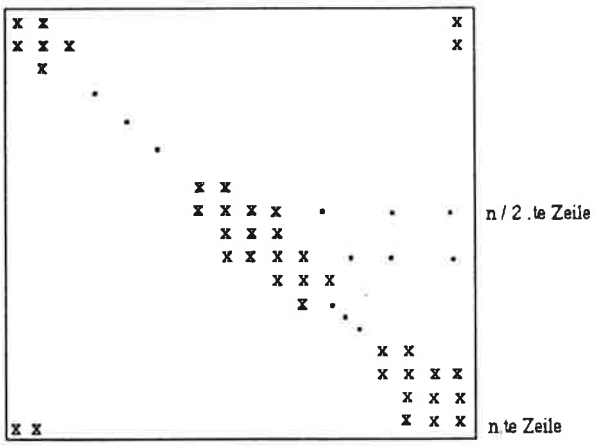
Für $1 \leq i \leq 2^{\ell}$ ist :

D_i =



Für $2^{\ell} + 1 \leq i \leq NR_{\ell} - 1$:

D_i =



$n = NR_{\ell,j}$

Der letzte Diagonalblock hat die Gestalt:

$D_{NR_{\ell}}$ =

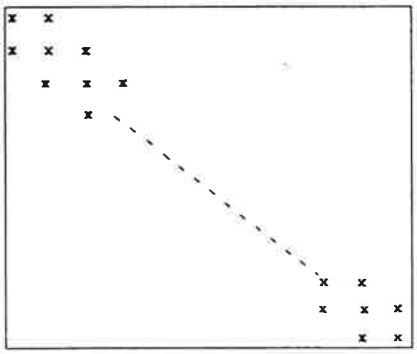
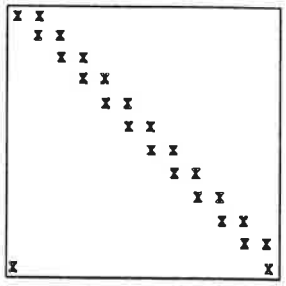


Abb.15 a-c Struktur der Diagonalblöcke der Steifigkeitsmatrix

Die Nebendiagonalblöcke N_ℓ haben die Form:

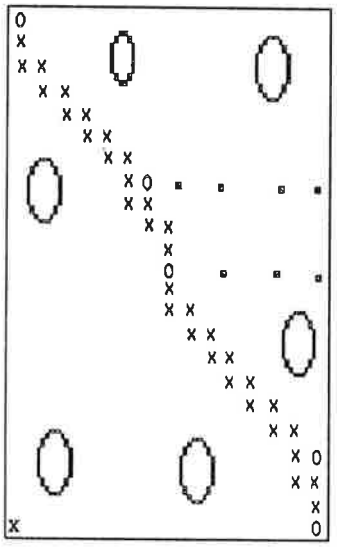
Für $1 \le i \le 2^\ell$ gilt:

$N_i =$



Für $2^{\ell-1}+1 \le i \le 3 \cdot 2^{\ell-1}-1$ gilt:

$N_i =$



Für den letzten Nebendiagonalblock erhalten wir die Struktur:

$N_{NR_\ell - 1} =$

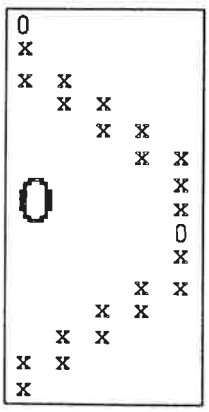


Abb.16 a-c Struktur der Nebendiagonalblöcke der Steifigkeitsmatrix

Die Unregelmäßigkeiten in der Struktur der einzelnen Blöcke entsprechen den Gitterpunkten der Triangulierung τ_ℓ die im Innern von Ω liegen und weniger als sechs Nachbarpunkte (=Punkte, die durch eine Dreiecksseite direkt verbunden sind) besitzen.

(6.5.3) Beispiel

Wir betrachten folgende Grobtriangulierung:

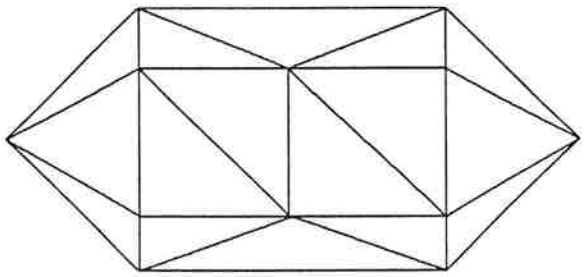


Abb.17 a Grobtriangulierung eines Testgebiets

Diese verfeinern wir zweimal regelmäßig und erhalten:

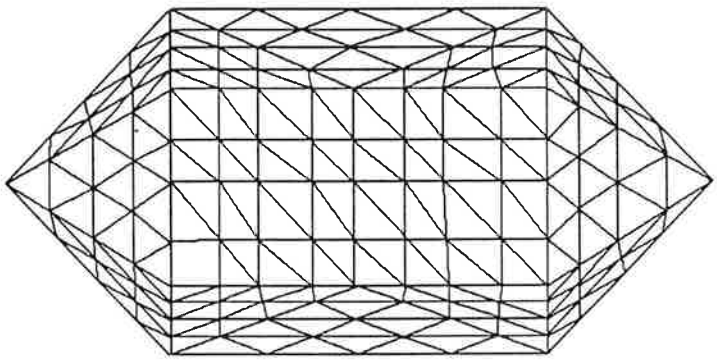


Abb.17 b Zweite Verfeinerung der Grobtriangulierung aus Abb.17 a

Die Steifigkeitsmatrix, welche zu der Triangulierung von Abb.17 b gehört, ist eine 141 x 141 Matrix mit folgender Besetzungsstruktur:

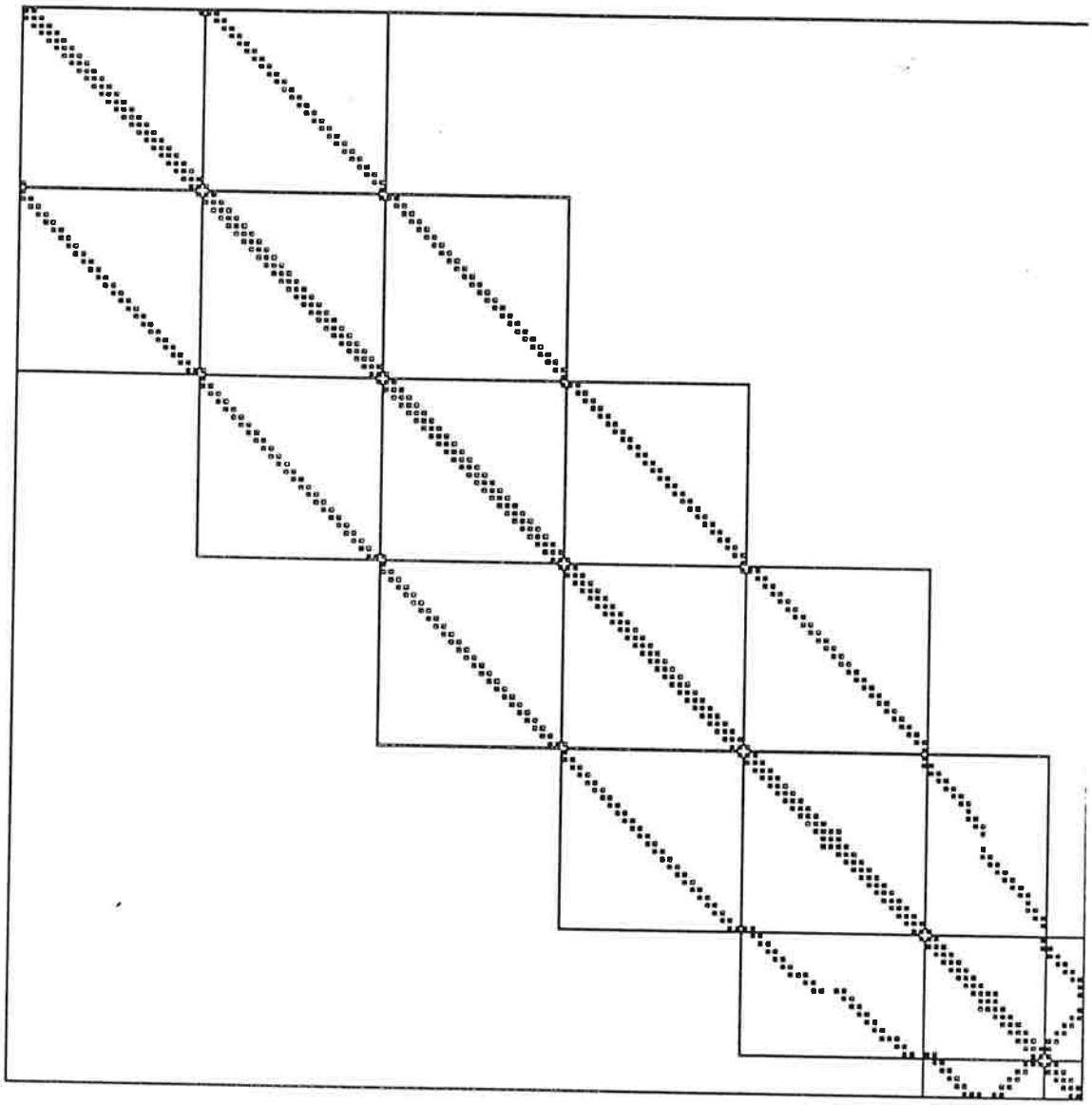


Abb.18 Die zur Triangulierung von Abb.17 b gehörende Steifigkeitsmatrix

(6.6) Herleitung des modifizierten ILU-Verfahrens

Zunächst benötigen wir eine spezielle Eigenschaft des ILU-Verfahrens: die Robustheit des Verfahrens.

Wir betrachten dazu folgendes anisotrope Modellproblem:

(6.6.1) $$-\varepsilon \cdot u_{xx} - u_{yy} = f(x,y) \text{ in } \Omega := (0,1) \times (0,1)$$
$$u = 0 \text{ auf } \Gamma := \partial\Omega \ ;$$

wobei $\varepsilon << 1$ sein soll.

Das Gebiet Ω sei mit Hilfe einer Quadratgittertriangulierung zerlegt (vgl. [9; p.160])

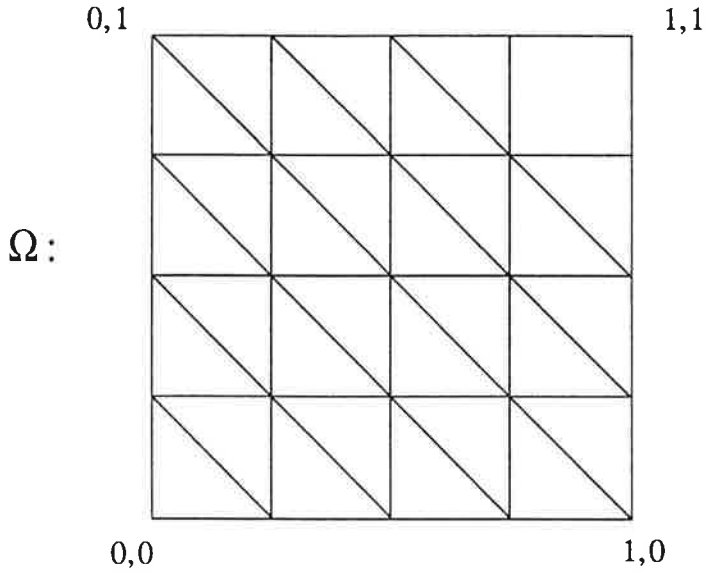


Abb.19 Quadratgittertriangulierung des Einheitsquadrats

Mit Hilfe der Transformation:

$$x = \sqrt{\varepsilon}\, x'$$
$$y = y'$$

formt man (6.6.1) um und erhält:

(6.6.1´) $$-u_{x'x'} - u_{y'y'} = f(x',y') \text{ in } \Omega' := (0, \frac{1}{\sqrt{\varepsilon}}) \times (0,1)$$
$$u(x',y') = 0 \text{ auf } \Gamma.$$

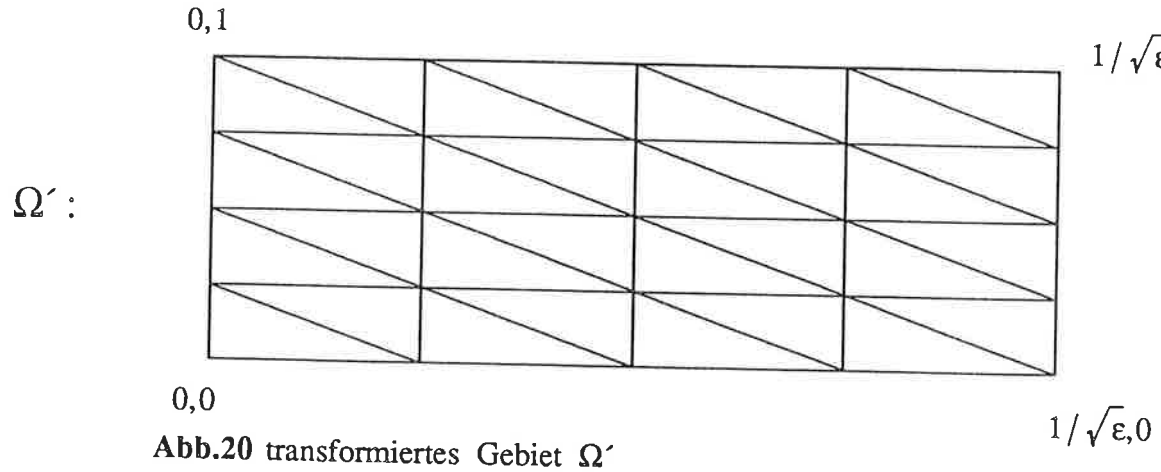


Abb.20 transformiertes Gebiet Ω'

Bei lexikographischer Numerierung (vgl.[8; p.51]) kann man die zugehörige Steifigkeitsmatrix $K(\varepsilon)$ charakterisieren durch den entarteten 7-Punkte Stern

(6.6.2)
$$\begin{bmatrix} 0 & -1 & \\ -\varepsilon & 2+2\varepsilon & -\varepsilon \\ & -1 & 0 \end{bmatrix}$$

bzw. durch den 5 - Punkte Stern :

$$\begin{bmatrix} & -1 & \\ -\varepsilon & 2+2\varepsilon & -\varepsilon \\ & -1 & \end{bmatrix}$$

(vgl. [24]).

(6.6.3) Kriterium

Sei $K(\varepsilon)$ ein singulär gestörter diskreter Operator mit dem Grenzwert:

$$K(0) = \lim_{\varepsilon \to 0} K(\varepsilon)$$

Das Glättungsverfahren sollte ein schneller (bzw.sogar exakter) Löser sein, für die Gleichung:

$$K(0) = f$$

❑

(6.6.4) Definition

Ein Glättungsverfahren heißt *robust*, falls es das Kriterium (6.6.3) erfüllt.

❑

(6.6.5) Bemerkung

Das ILU-Verfahren (6.2.4) erfüllt Kriterium (6.6.3) für K(ε) , definiert durch (6.6.2) bzw. auch für das entsprechend diskretisierte "gekippte" Problem:

$$-u_{xx} - \varepsilon u_{yy} = f \ .$$

❑

Beweis:

Für ε=0 entartet die Steifigkeitsmatrix K(ε) zu einer symmetrischen Bandmatrix, bestehend aus drei Diagonalen. Da für eine exakte Dreiecks-Zerlegung einer solchen Matrix K, bestehend aus drei Bändern gilt:

$$\mathfrak{P}_K = \mathfrak{P}_L \cup \mathfrak{P}_U \ ,$$

und wir für die Besetzungsmuster der Zerlegungsmatrizen L und U vorausgesetzt haben:

$$\mathfrak{P}_K \subset \mathfrak{P}_U \cup \mathfrak{P}_L$$

folgt, daß in diesem Fall die ILU-Zerlegung mit der exakten Dreieckszerlegung übereinstimmt.

QED

Wir definieren nun die ILU-Zerlegung für unser Problem, indem wir das Besetzungsmuster $\mathfrak{P}_L$ und $\mathfrak{P}_U$ für die Zerlegungsmatrizen L und U aus (6.2.1´) angeben.

Sei dazu $\mathfrak{P}_K$ das Besetzungsmuster der Steifigkeitsmatrix K.

(6.6.7 a) $$\mathfrak{P}_U := \{(i,j)\ ,\ 1 \le i < j \le NP\ ;\ (i,j) \in \mathfrak{P}_K\} \cup \mathfrak{P}_Z$$

(6.6.7 b) $$\mathfrak{P}_L := \{(i,j)\ ,\ 1 \le i, j \le NP\ ;\ (j,i) \in \mathfrak{P}_U\} \ ;$$

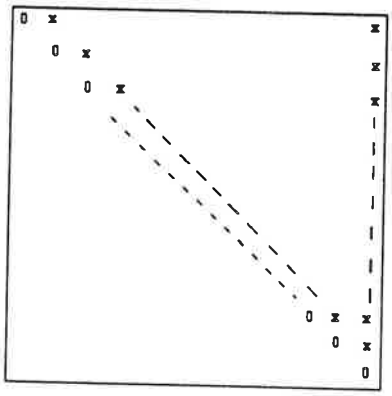
wobei $\mathfrak{P}_Z$ definiert ist durch :

(6.6.7 c) $$\mathfrak{P}_Z := \{(i,j)\ ,\ 1 \le i < j \le NP\ ;$$
$$\exists\, k,\ 1 \le k \le NR\ ;\ [P_i \in R_k \wedge j = \max_{1 \le m \le NR_k}(m;\ P_m \in R_k)\ ]\ \} \ .$$

Erläuterung zur Definition (6.6.7):

Falls man ein Gebiet Ω gemäß der in (6.3) erklärten Weise diskretisiert (und verfeinert), erhält man ähnlich geometrisch strukturierte Dreiecke wie bei der Triangulierung (6.6.1´) im Bereich zwischen der - in möglichst geringer Entfernung von ∂Ω gezogenen- Konturlinie K und ∂Ω. Hierbei ist zu beachten, daß im Gegensatz zum Modellproblem (6.6.1) derjenige Gitterpunkt mit der niedrigsten Nummer eines Ringes P_m über eine Dreiecksseite gekoppelt ist mit demjenigen Punkt von P_m, welcher die höchste Nummer besitzt. Um aber entlang eines Ringes P_m unabhängig zu dessen Entfernung zu den Nebenringen P_{m-1} und P_{m+1} gleichmäßig genau zu lösen (Robustheit!), müssen wir für die periodischen Ausnahmestellen die Dreieckszerlegung exakt durchführen. $\mathfrak{P}_Z$ besteht aus den Indexpaaren derjenigen Matrixelemente von U für die diese zusätzlichen Eliminationsschritte durchgeführt werden müssen.

Den Diagonalblöcken der Steifigkeitsmatrix K (vgl.Abb 15) würden dann - beispielsweise an gleicher Stelle bei der strikten oberen Dreiecksmatrix U - Diagonalblöcke der Struktur:



entsprechen.

(6.6.8) Definition *modifiziertes ILU-Verfahren*

Sei Ω ein Gebiet, welches per Konstruktion (3) aus § 6.3 trianguliert wurde.
Sei K die Steifigkeits- und M die Massenmatrix, die entsteht, falls man Problem (3.10), wie in (4.1) beschrieben wurde, diskretisiert.
Wir betrachten das lineare Gleichungssystem (vgl. 5.1.3):

$$(K - \lambda M)\, u = f \;; \quad u \perp M\{E(\lambda)\}; \quad f \perp E(\lambda).$$

Das *modifizierte ILU-Verfahren* zur Lösung obiger Gleichung lautet:

Sei $\beta > 0$. L, D, U seien die durch (6.6.7) definierten Zerlegungsmatrizen der Matrix $K+\beta I$ (vgl.6.2.1´); (I bezeichnet hier die Einheitsmatrix!).

(6.6.8 a) Sei x_0 ein Startvektor.

(6.6.8 b) Rekursionsvorschrift: $x_{i+1} = x_i - \omega A_\beta^{-1} \{(K - \lambda M) x_i - f\}$;

mit $A_\beta := (L´+ D) D^{-1} (U´+ D)$ und festem $\omega \in (0,1]$.
ω bezeichnet man als Dämpfungsparameter, β als Shiftparameter.

❑

(6.6.9) Bemerkung

Im Fall, daß die Steifigkeitsmatrix K symmetrisch ist, gilt :

$$L = U^T.$$

❑

(6.7) Glättungseigenschaft für das modifizierte ILU-Verfahren

In diesem Abschnitt werden wir untersuchen, inwieweit das modifizierte ILU-Verfahren die Glättungseigenschaft (5.3.1) für die singuläre Gleichung (4.1.2) besitzt. Dazu benötigen wir für die in §5 eingeführten Normen die folgenden Abschätzungen:

(6.7.1) Bemerkung

Für die Steifigkeitsmatrix K und die Massenmatrix M definiert in (4.1) gilt mit der Eliptizitätskonstante ε aus Satz (4.2.2) und $h := \max\{\text{Länge der Kanten der Dreiecke der Triangulierung von } \Omega\}$:

$$\|K\| \le C$$

$$\|M\| \le Ch^2$$

$$\|K\|_M \le Ch^{-2}$$

$$\|K^{-1}\| \le Ch^{-2} / \varepsilon$$

$$\|K^{-1}\|_{-M} \le C / \varepsilon$$

$$\|M^{-1}\| \le Ch^{-2}$$

❑

Beweis:

siehe [9; Bem.8.8.4, Satz(8.8.6) und (8.8.7)] und [8; Prop.6.3.28)]

QED

Da das ILU-Verfahren zunächst für Finite Differenzen Verfahren entwickelt und angewendet wurde (vgl.[23]) und dort Normen verwendet werden, welche der diskreten Operatornorm $\|\cdot\|_M$ bzw. $\|\cdot\|_{-M}$ in bezug auf die h -Abhängigkeit bei den Abschätzungen (6.7.1) am natürlichsten entsprechen, werden wir im folgenden Abschnitt mit dieser Norm arbeiten.

Wir werden zunächst die Glättungseigenschaft für ein reguläres Problem unter geeigneten Voraussetzungen nachweisen, um dann die singuläre Gleichung (4.1.2) als Störung dieses regulären Problem zu interpretieren. Daraus erhalten wir dann die Glättungseigenschaft auch für dieses Problem.

Wir betrachten dazu zunächst das lineare Gleichungssystem:

(6.7.2) $$K x = f$$

mit der regulären n x n Matrix K und den n-Vektoren x und f.

Ein Iterationsverfahren zur Lösung von (6.7.2) läßt sich schreiben (vgl.6.1.2):

(6.7.3)

a) x_0 sei ein gegebener Startvektor.

b) $$x_{i+1} = S x_i + N f$$

mit der Iterationsmatrix $S := I - A^{-1}K$ und $N := A^{-1}$.

Für das modifizierte ILU-Verfahren ist die Matrix A, welche K "möglichst geschickt" approximieren soll, durch (6.2.1) und (6.6.7) definiert.

(6.7.4) Satz

Wir betrachten das Iterationsverfahren (6.7.3) für die Gleichung (6.7.2).
Voraussetzungen:

(6.7.5) a) die n x n Matrix K sei symmetrisch und positiv definit.

b) K sei zerlegt in : $K = A_\beta - C$;

c) wobei für A_β gelten soll: $A_\beta = A_\beta^T$; A_β positiv definit.

d) $\|A_\beta\|_M \le C_A\, h^{-2}$

Wir definieren:

$$S_{0,\omega} := I - \omega A_0^{-1} K.$$

Dann existiert für gegebenes $\vartheta \in [0, 1)$ ein $\omega_\vartheta \in (0,1]$, so daß:

$$\|K\, S_{0,\omega}^{\nu}\|_M \le C_A\, \eta\,(\nu,\vartheta)\, h^{-2} \qquad \text{für } \omega \in (0,\omega_\vartheta\,]$$

und $\eta\,(\nu,\vartheta)$ definiert ist durch:

$$\eta(\nu,\vartheta) := \max\left\{\frac{\nu^\nu}{(\nu+1)^{(\nu+1)}}, \frac{\vartheta^\nu}{1+\vartheta}\right\}.$$

❑

Beweis:
siehe [25; Theorem 3.1.3] QED

(6.7.6) Satz

Es gelten die Voraussetzungen (6.7.5 a-d). Sei $\vartheta \in [0,1)$ gegeben und β_ϑ definiert durch:

$$\beta_\vartheta := \frac{\lambda_{min} - \vartheta}{\lambda_{min} \cdot (1+\vartheta)}$$

mit $\lambda_{min} := \min\{|\lambda|,\ \lambda \text{ ist Eigenwert von } S_{0,1}\}$.
Dann erfüllt für $\beta > \beta_\vartheta$ der modifizierte Glätter $S_{\beta,1}$ die Glättungseigenschaft für das Problem (6.7.2):

$$\|K\, S_{\beta,1}{}^{\nu}\|_M \le C_A\, \eta\,(\nu,\vartheta)\, h^{-2} \qquad \text{für } \beta \in (0,\beta_\vartheta\,].$$ ❑

Beweis:
siehe [25; Theorem 3.1.5)] QED

Wir betrachten nun das singuläre Problem:

(6.7.7) $$(K_\ell - \lambda_\ell M_\ell)\, u_\ell = f_\ell,$$

wobei wir uns die Matrizen K_ℓ und M_ℓ durch die Diskretisierung der Differentialgleichung (3.10) mit Hilfe von finiten Elementen entstanden vorstellen,wie sie in §4 beschrieben wurde.
Das Iterationsverfahren für (6.7.7) sei gegeben durch:

(6.7.8)
a) x_0 sei ein gegebener Startvektor.
b) $$x_{i+1} = S(\lambda)\, x_i + N f$$

mit der Iterationsmatrix $S(\lambda)$ definiert durch :

$$S(\lambda) := I - A^{-1}(K - \lambda M)$$

und $$N := A^{-1} .$$

A ist die durch (6.2.1) und (6.6.7) definierte approximative Inverse von K.

(6.7.9) Satz

Wir betrachten das Verfahren (6.7.8) zur Lösung von obiger singulärer Gleichung und definieren:

$$\tilde{K}_\ell := -\lambda_\ell M_\ell ; \quad S_\ell := I - A^{-1}K_\ell ; \quad \tilde{S}_\ell := \lambda_\ell A^{-1} M_\ell ; \quad S(\lambda) := S_\ell + \tilde{S}_\ell .$$

Voraussetzungen :

Es gelten die Voraussetzungen (6.7.5) und für hinreichend kleines h_0 gelte :

a) S_ℓ besitzt die Glättungseigenschaft für das reguläre Problem (6.7.2), d.h.

$$\|K_\ell S_\ell^\nu\|_M \le \eta(\nu)\, h_\ell^{-2} \quad 0 \le \nu \le \bar{\nu}(h_\ell) ;$$

$$0 < \eta(\nu) \to 0 \quad \text{für } \nu \to \infty$$

$$\bar{\nu}(h_\ell) = \infty \text{ oder } \bar{\nu}(h_\ell) \to \infty \text{ für } h_\ell \to 0.$$

b) $$\|S_\ell{}^\nu\| \le C_S(\nu) \ , \ \nu < \bar{\nu}(h_\ell)$$

c) $$\lim_{\ell \to \infty} h_\ell^2 \, \|\tilde{K}_\ell\|_M = 0$$

d) $$\lim_{\ell \to \infty} \|\tilde{S}_\ell^\nu\| = 0 .$$

Behauptung :

(i) Dann besitzt Verfahren (6.7.8) für die Gleichung (4.1.2) die Glättungseigenschaft.

(ii) Für das Eigenwertproblem ($f_\ell \equiv 0$ in Gleichung (6.7.7)) ist (5.3.1 i und ii) erfüllt.

❑

Beweis:

(i) siehe [8; Criterion(6.2.7)]

(ii) (5.3.1 ii) folgt sofort aus der Rekursionsvorschrift (6.7.7).
(5.3.1 i):
Zu zeigen ist:

$$\|S(\lambda)^\nu - S(\tilde{\lambda})^\nu\| \le C\,|\lambda - \tilde{\lambda}| .$$

Es gilt, falls h_0 hinreichend klein ist :

$$\|S(\lambda)^\nu - S(\tilde{\lambda})^\nu\| \le \sum_{i=0}^{\nu-1} \|S(\lambda)\|^i \cdot \|S(\lambda) - S(\tilde{\lambda})\| \cdot \|S(\tilde{\lambda})\|^{\nu-1-i}$$

$$\le |\lambda - \tilde{\lambda}| \sum_{i=0}^{\nu-1} \left[\{C_S(\nu) + \|\tilde{S}_\ell\|\}^i \|A^{-1} M_\ell\| \{C_S(\nu) + \|\tilde{S}_\ell^\nu\|\}^{\nu-1-i} \right]$$

$$\le |\lambda - \tilde{\lambda}| \sum_{i=0}^{\nu-1} C C_S^i(\nu) C_S^{\nu-1-i}(\nu)$$

$$\le C(\nu) |\lambda - \tilde{\lambda}|$$

QED

Es bleibt jetzt noch nachzuprüfen, inwieweit unser konkretes Problem (4.1.2) mit unserem speziellen Glättungsverfahren die Voraussetzungen der Sätze (6.7.5), (6.7.7) und (6.7.9) erfüllt.

1) Voraussetzungen (6.7.6)

a) $K = K^T$ klar.

In unserem Fall (Neumann-Randwerte) ist K lediglich positiv semidefinit, da mit:

$c_h := \inf\{h(x,y) ; x,y \in \Omega\}$ ($h(x,y)$ bezeichnet hier die Tiefenfunktion)

gilt:

$$(\nabla u , h \nabla u) \ge c_h (\nabla u , \nabla u) \begin{cases} = 0 & ; \text{falls } \nabla u = 0 \text{ (d.h. } u = \text{const.)} \\ > 0 & ; \text{sonst} \end{cases}$$

und $$(\nabla u , h \nabla u) = 0 .$$

Da durch unsere Diskretisierung (4.1) die konstanten Funktionen auf Ω exakt approximiert werden, bedeutet das für den Kern der Steifigkeitsmatrix K:

$$\ker K = \text{span}\{u \in \mathbb{R}^n , u_i = 1 \quad \text{für } 1 \le i \le n\}$$

Wenn wir nun (6.7.7) äquivalent umformulieren in:

$$(K_\ell' - \lambda_\ell' M_\ell) u_\ell = f_\ell ;$$

mit $K_\ell' := K_\ell + t \cdot M_\ell$; $\lambda_\ell' := \lambda_\ell + t$ und $t > 0$; $t \notin \sigma(M^{-1} K)$ (t fest!), erhalten wir eine positiv definite Matrix K_ℓ', die wir dann zerlegen.

c) $A_\beta = A_\beta^T$ ist nach Definition (6.6.8) klar.

Die Eigenschaft: „ A_β positiv definit " nennt man Stabilitiät. Darauf wird in (6.8) ausführlich eingegangen.

d) Man benötigt :

$$\|A\|_M \le C_A h^{-2} .$$

Diese Bedingung ist so zu interpretieren, daß das Glättungsverfahren die hohen Eigenwerte der Matrix K gut approximieren soll.
Für das Modellproblem der anisotropen Laplace-Gleichung auf dem Einheitsquadrat mit

Dirichlet-Randbedingung, welches unserem in gewisser Hinsicht teilweise gleicht (vgl.Erläuterung zur Definition 6.6.7), wurde von Wittum [24; Criterion 2.1 und Lemma 2.3] gezeigt, daß für die Restmatrix

$$N := K - A$$

gilt:

$$\|N\|_\infty \leq \varepsilon\, h^{-2}\, C$$

mit ε aus (6.6.1).
Wegen Bemerkung (6.7.1) folgt dann mit Hilfe der Dreiecksungleichung sofort die Abschätzung für A.

2) Voraussetzungen von Satz (6.7.9):

a) folgt aus Satz (6.7.4) bzw.(6.7.6)

b_i) Wir benötigen die Abschätzung :

$$\|A^{-1}\|_M \leq C h^2.$$

Dies bedeutet zusammen mit (6.7.5 d), daß A die Steifigkeitsmatrix K nur im oberen Bereich ihres Spektrums gut approximieren sollte (vgl.Abb.21) . Dies ist vor allem deshalb sinnvoll, da wir Mehrgitterverfahren effektiv verwenden wollen und die niedrigeren Eigenwerte auf den gröberen Gittern approximieren, bzw. für die niedrigsten Eigenwerte auf dem gröbsten Gitter exakt lösen. Falls A beispielsweise mit K übereinstimmen würde, wäre ein Mehrgitterverfahren überflüssig. Für ein Modellproblem auf einer Quadratgittertriangulierung wurde obige Abschätzung von Wittum [23] durch Kriterium (5.2.4), Lemma (5.2.2) und Bemerkung(5.2.4) bewiesen.

b_{ii}) Aus (b_i) folgt dann:

$$\|S_\ell^\nu\| \leq \|I - A^{-1} K_\ell\|^\nu \leq \|A^{-1} N\|^\nu \leq (C h^2 h^{-2})^\nu = C(\nu).$$

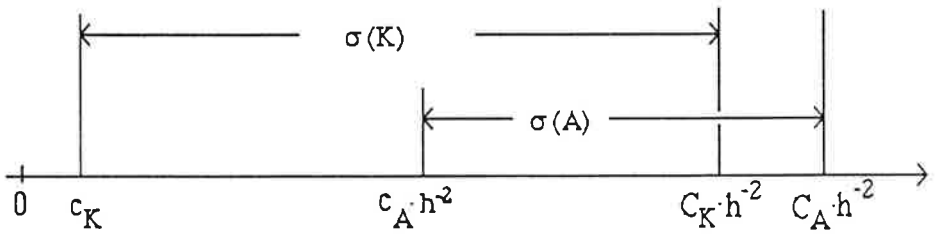


Abb. 21 Verteilung des Spektrums von A und K.

c) Für jedes feste $\lambda \in \mathbb{C}$ gilt :

$$\lim_{\ell \to \infty} h_\ell^2 \, \| \tilde{K}_\ell \|_M = \lambda \lim_{\ell \to \infty} h_\ell^2 \, \| M_\ell \|_M \leq \lambda \, C \lim_{\ell \to \infty} h_\ell^2 = 0.$$

d) Zu zeigen :

$$\lim_{\ell \to \infty} \| \tilde{S}_\ell^\nu \| = 0 .$$

Folgt aus :

$$\| \tilde{S}_\ell^\nu \| = \lambda \, \| A^{-1} M_\ell \| \leq \lambda \, \| A^{-1} \|_M \cdot \| M_\ell \|_M \leq C \lambda \, h_\ell^2 \xrightarrow{\ell \to \infty} 0.$$

(6.8) Stabilität der ILU-Zerlegung

In diesem Abschnitt werden wir den in § 6.7 kurz erwähnten Begriff der Stabilität definieren und für eine spezielle Triangulierung, welche der in § 6.3 definierten in gewisser Hinsicht ähnelt, die Stabilität beweisen.

Wir betrachten folgende Situation:

(6.8.1) gegeben sei eine symmetrische n x n Matrix K mit Besetzungsmuster $\mathfrak{P}_K$.

Sei das Besetzungsmuster der strikten oberen Dreiecksmatrix U gegeben durch $\mathfrak{P}_U$.

Analog zu (6.2) zerlegen wir dann K in:

$$K = (U + D)^T \; D^{-1} \; (U + D) - N \; ;$$

wobei D diagonal sein soll.

Weiter muß gelten:

$$\mathfrak{P}_U \cup \mathfrak{P}_D \cup \mathfrak{P}_{U^T} \supset \mathfrak{P}_K$$
$$\mathfrak{P}_U \cap \mathfrak{P}_D = \varnothing .$$

(6.8.2) Definition

Die durch (6.8.1) definierte ILU-Zerlegung heißt *stabil*, falls gilt:

$$D_{i,i} > 0 \quad \text{für } 1 \leq i \leq n .$$

❑

Es gilt:

$$K := (U + D)^T D^{-1} (U + D)$$

ist positiv definit genau dann, wenn die Zerlegung stabil ist. Wie man in § 6.7 gesehen hat, benötigt man die Stabilität der Zerlegung für den Beweis der Glättungseigenschaft. Aussagen über die Stabilität der Zerlegung existieren bisher nur, falls K eine M-Matrix ist.

(6.8.3) Definition

Eine n x n Matrix A heißt *M-Matrix*, falls gilt:

(6.8.4) a)
$$A_{i,i} > 0 \text{ , für } 1 \leq i \leq n$$
$$A_{i,j} \leq 0 \text{ , für } 1 \leq i,j \leq n \quad i \neq j$$

(6.8.4) b) A ist regulär und $A^{-1}_{i,j} \geq 0$, für $1 \leq i,j \leq n$.

❑

(6.8.5) Definition

Wir betrachten eine Triangulierung τ eines Gebietes Ω.
Ein Knotenpunkt P_i $(1 \leq i \leq NP)$ heißt *innnerer* bzw. *äußerer (Knoten-)Punkt*, falls gilt:

$$P_i \notin \partial\Omega \quad \text{bzw.} \quad P_i \in \partial\Omega.$$

Ein Knotenpunkt P_j ist *Nachbarpunkt* eines Knotenpunktes P_i, falls gilt:

P_i und P_j sind durch eine Seite eines Dreiecks $T \in \tau$ verbunden.

Ein Knotenpunkt P_i heißt *randnaher* bzw. *randferner (Knoten-) Punkt*, falls es mindestens einen bzw. keinen Nachbarpunkt P_j von P_i gibt, der ein äußerer Punkt ist.
Seien nun P_i und P_j zwei innere Nachbarknotenpunkte. Es existieren dann genau zwei Dreiecke T_s und T_t von τ die P_i und P_j als gemeinsamen Eckpunkt besitzen. α_i bzw. α_j sei derjenige Winkel in T_s bzw. T_t, welcher der Dreiecksseite $\overline{P_iP_j}$ gegenüber liegt.
Wir definieren:

$$\alpha_{ij} := \alpha_i + \alpha_j \ ;$$

$$I := \{ (i,j),\ 1 \leq i < j \leq NP \ ;\ P_i \text{ und } P_j \text{ sind innere Nachbarpunkte} \}$$

$$A_\tau := \max_{(i,j) \in I} \alpha_{ij} \ .$$

❑

Für Satz (6.8.7) benötigen wir noch folgendes technisches Lemma:

(6.8.6) Lemma

Sei $0 < \alpha,\beta \leq \pi$.

Dann gilt:

$$\cot\alpha + \cot\beta \leq 0 \ \Leftrightarrow \ \alpha + \beta \geq \pi.$$

❑

Beweis:

$$\cot\alpha + \cot\beta = \sin(\alpha+\beta)/\{\sin\alpha \sin\beta\}$$

Wegen der Voraussetzung gilt: $\sin\alpha \sin\beta \geq 0.$

Aus

$$\sin(\alpha+\beta) < 0 \Leftrightarrow \alpha+\beta > \pi \quad \text{und} \quad \sin(\alpha+\beta) = 0 \Leftrightarrow \alpha+\beta = n\pi;\ n \in \{0,1\}$$

folgt die Behauptung.

QED

(6.8.7) Satz

Sei τ eine zulässige Triangulierung des Gebietes Ω, welche NP Knotenpunkte und n innere Knotenpunkte besitzt. τ darf kein Dreieck enthalten, dessen Ecken alle auf $\partial\Omega$ liegen.
K sei die n x n Steifigkeitsmatrix, die entsteht, indem man die Gleichung:

(6.8.8)
$$-\Delta u = f \quad \text{in } \Omega$$
$$u = 0 \quad \text{auf } \partial\Omega$$

mit Hilfe von stückweise linearen, stetigen finiten Elemente über τ diskretisiert.

Behauptung:
Die folgenden Aussagen sind äquivalent:

a) K ist eine M - Matrix
b) $A_\tau \leq \pi$.

❑

Beweis:

Sei T ein Dreieck der Triangulierung τ mit den Eckpunkten :

$$P_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix}, P_j = \begin{pmatrix} x_j \\ y_j \end{pmatrix}, P_k = \begin{pmatrix} x_k \\ y_k \end{pmatrix},$$

die gegen den Uhrzeigersinn fortlaufend numeriert sein sollen.
Weiter definieren wir :

$$s_i := \begin{pmatrix} x_k - x_j \\ y_k - y_j \end{pmatrix}, s_j := \begin{pmatrix} x_i - x_k \\ y_i - y_k \end{pmatrix}, s_k := \begin{pmatrix} x_j - x_i \\ y_j - y_i \end{pmatrix};$$

$$\alpha_i := \measuredangle(s_k, s_j), \alpha_j := \measuredangle(s_i, s_k), \alpha_k := \measuredangle(s_j, s_i).$$

Die stückweise lineare Finite - Elemente Ansatzfunktion $b_i(x,y)$ auf T, für die gelten soll :

$$b_i(x_i, y_i) = 1 \;,\; b_i(x_j, y_j) = 0 \;,\; b_i(x_k, y_k) = 0 \;,$$

läßt sich in der Form schreiben :

$$b_i(x,y) = \frac{(x - x_j)(y_k - y_j) - (y - y_j)(x_k - x_j)}{(x_i - x_j)(y_k - y_j) - (y_i - y_j)(x_k - x_j)}$$
$$= \frac{(x - x_j)(y_k - y_j) - (y - y_j)(x_k - x_j)}{2|T|}.$$

Daraus folgt:

$$\nabla b_i(x,y) = \frac{1}{2|T|} s_i^{\perp} \; ;$$

wobei für einen Vektor $c = \begin{pmatrix} a \\ b \end{pmatrix}$ $c^{\perp}$ definiert ist durch :

$$\begin{pmatrix} a \\ b \end{pmatrix}^{\perp} := \begin{pmatrix} b \\ -a \end{pmatrix}.$$

Es gilt dann :

(6.8.9)
$$\int_T < \nabla b_i , \nabla b_j > = \frac{1}{4|T|} < s_i^{\perp} , s_j^{\perp} > = \frac{1}{4|T|} < s_i , s_j >$$
$$= -\frac{|s_i| \, |s_j| \cos \alpha_k}{2|s_i| \, |s_j| \sin \alpha_k} = -\frac{1}{2} \cot \alpha_k \; .$$

Analog zeigt man :

(6.8.10)
$$\int_T < \nabla b_i , \nabla b_i > = \frac{|s_i|^2}{4|T|} = \frac{\sin \alpha_i}{2 \sin \alpha_j \sin \alpha_k} \; .$$

1. Teil: a $\Rightarrow$ b

indirekt : Wir zeigen : $A_\tau > \pi \Rightarrow$ K ist keine M - Matrix.
Sei $A_\tau > \pi$.
$\Rightarrow \exists \, (i,j) \, , \, 1 \leq i,j \leq NP$; P_i und P_j sind innere Nachbarpunkte, und :

$$\alpha_{i,j} = \alpha_i + \alpha_j > \pi.$$

Aus den eben hergeleiteten Formeln folgt sofort :

$$K_{i,j} = -\frac{1}{2} (\cot \alpha_i + \cot \alpha_j) \, , \text{ mit } \alpha_i + \alpha_j > \pi.$$

Wegen Lemma (6.8.6) gilt :

$$K_{ij} > 0$$

$\Rightarrow$ K ist keine M - Matrix.

2. Teil: b $\Rightarrow$ a

Sei der Isomorphismus P wie in Abschnitt 4.1 definiert und $u \in \mathbb{R}^n$ so,daß $Pu \in H_0^1(\Omega)$.
Dann gilt :

$$< u , K u > = (\nabla P u , \nabla P u)_0 \begin{cases} > 0 \text{ für } u \neq 0 \\ = 0 \text{ für } u = 0 \end{cases} \text{ (Friedrichsche Ungleichung)}$$

$\Rightarrow$ K ist positiv definit.

Da K symmetrisch ist, sind alle Eigenwerte von K positiv.
Weiter folgt aus $A_\tau \leq \pi$, daß alle Elemente $K_{i,j}$ $(i \neq j)$ von K nicht positiv sind.
Aus Satz (6.8.11) folgt dann die Behauptung.

QED

(6.8.11) Satz

Sei A eine n x n Matrix, die den Eigenschaften (i) und (ii) genügt.

(i) $A_{i,j} \leq 0 \quad , \quad 1 \leq i,j \leq n; \; i \neq j.$

(ii) Jeder reelle Eigenwert von A ist nicht negativ.

Dann gilt : A ist eine M - Matrix.

❑

Beweis :
siehe [3; Theorem(6.4.6) Kriterium C_8] QED

(6.8.12) Bemerkung

In [19; p. 78] wird ein ähnliches Winkelkriterium wie die Bedingung „$A_\tau \leq \pi$" angegeben, welches jedoch nur hinreichend dafür ist, daß K eine M-Matrix ist.

❑

(6.8.13) Bemerkung

Wir betrachten eine zulässige Grobtriangulierung τ_0 des Gebietes Ω.

τ_ℓ bezeichnet die Triangulierung, die entsteht, indem man τ_0 ℓ-mal regelmäßig verfeinert (vgl.6.3).
α sei der größte Innenwinkel, der Dreiecke der Triangulierung τ_0.
Sei $\ell \geq 2$.
Dann gilt :
Die Steifigkeitsmatrix K_ℓ , welche zu Triangulierung τ_ℓ gehört ist eine M - Matrix.
$\Leftrightarrow \quad \alpha \leq \pi/2$

❑

Beweis:

1.Teil: "$\Rightarrow$"

indirekt : Wir zeigen : $\alpha > \pi/2 \Rightarrow K_\ell$ ist keine M - Matrix für $\ell \geq 2$.
Es genügt zu zeigen : Es existieren zwei innere Nachbarpunkte P_i und P_j für die gilt :
$\alpha_{ij} > \pi$.

Sei T ein Dreieck von τ_0 , welches einen Innenwinkel besitzt, der größer als $\pi/2$ ist.OBdA sei di α_1 .Die Bezeichnungen seien wie in Abb.22 a bzw.b. Regelmäßige Verfeinerung (Verbindung der Seitenmiten von T) ergibt vier neue Dreiecke, welche alle ähnlich zu T sind. Abb. 22 b zeigt T nac zweimaliger Verfeinerung. Aus Symmetriegründen gilt für die Punkte P_s und P_t :

P_s und P_t sind innere Nachbarpunkte und $\alpha_{st} = 2 \cdot \alpha_1 > \pi$.

Wegen Lemma (6.8.11) gilt dann : $(K_\ell)_{s,t} > 0$;

$\Rightarrow$ K_ℓ ist keine M - Matrix.

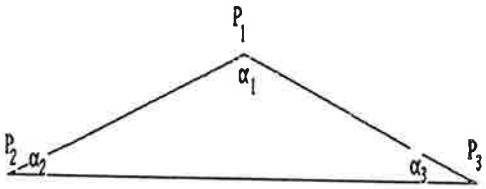


Abb .22a Dreieck mit stumpfen Winkel α_1

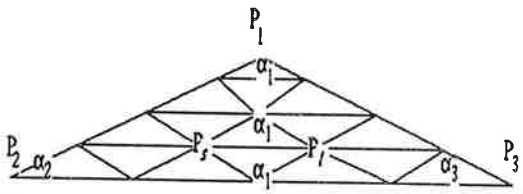


Abb. 22b Dreieck aus Abb.22 a, zweimal regelmäßig verfeinert

2.Teil: "$\Leftarrow$"

Da bei regelmäßiger Verfeinerung eines Dreiecks vier neue Dreiecke entstehen, die alle ähnlich zum alten sind, folgt aus $\alpha \leq \pi/2$, daß der größte Innenwinkel aller Dreiecke jeder Triangulierungsstufe kleiner oder gleich $\pi/2$ ist. Die Behauptung folgt dann wie im Beweis von Satz (6.8.7) Teil 2.

QED

(6.8.14) Bemerkung

Bei adaptiver Verfeinerung (Halbierung einiger Dreiecke vgl. (6.3) (2)) einer Triangulierung entstehen Dreiecke, deren größter Winkel mindestens $\pi/2$ ist.

❑

Beweis:

Folgt aus elementaren geometrischen Überlegungen.

Wir wollen nun die bekannten Ergebnisse über die Stabilität der ILU-Zerlegung von M-Matrizen kurz zitieren.

(6.8.15) Satz

Die ILU- Zerlegung einer symmetrischen M-Matrix K ist stabil, falls wir Situation (6.8.1) zu Grunde legen.

❑

Beweis:

siehe [13]

QED

Für den nächsten Satz betrachten wir folgendes Modellproblem:

$$-\Delta u = f \quad \text{in} \quad (0,1) \times (0,1)$$
$$u = 0 \qquad \text{auf } \partial\Omega;$$

diskretisiert über eine Hierarchie äquidistanter Gitter mit Schrittweite h_ℓ, $\ell \in \mathbf{N}$; $h_\ell \to 0$ für $\ell \to \infty$, indem wir den 5-Punkte Stern :

$$K_\ell = h_\ell^{-2} \begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}$$

benutzen (vgl.[9; § 4.2]).

(6.8.16) Satz

Wir verwenden die Bezeichnungen aus (6.8.1). Sei $n := 1/h_l - 1$ und $NP = n^2$.

a) Es gelte :

$$\mathfrak{P}_U \cup \mathfrak{P}_{U^T} \cup \mathfrak{P}_D = \mathfrak{P}_{K_\ell} \quad (5 \text{ Punkt - ILU})$$
$$N := K_\ell - (U + D)^T D^{-1} (U + D).$$

Behauptung :

Für jedes i $(1 \le i \le n)$ ist die Folge $\{D_{i+jn,i+jn}\}_{0 \le j \le n-1}$ monoton fallend (bezüglich j)und konvergiert für $n \to \infty$ gegen $\delta_5 = 2 + \sqrt{2}$.

b) Es gelte (mit N aus Teil a) :

$$\mathfrak{P}_U \cup \mathfrak{P}_{U^T} \cup \mathfrak{P}_D = \mathfrak{P}_{K_\ell} \cup \mathfrak{P}_N \quad (7 \text{ Punkt - ILU}).$$

Behauptung :

Für jedes i $(1 \le i \le n)$ ist die Folge $\{D_{i+jn,i+jn}\}_{0 \le j \le n-1}$ monoton fallend (bezüglich j) und konvergiert gegen $\delta_7 = 3.294\ldots$.

❑

Beweis :

a) siehe [24]

b) siehe [22]

QED

Das folgende Beispiel zeigt, daß die Bedingung: „K ist eine M-Matrix" zwar hinreichend (vgl.Satz (6.8.15)), aber keineswegs notwendig ist.

(6.8.17) Beispiel

Sei

$$P_1 := (0,0)^T, \; P_2 := (1,0)^T, \; P_3 := (0.5,0.5/c)^T, \; P_4 := (1.5,0.5/c)^T;$$

mit $1 \le c \le \infty$ (bezogen auf ein kartesisches Koordinatenkreuz).

Sei $\Omega \subset \mathbb{R}^2$ das (beschränkte) Innere des Vierecks, welches die Ecken P_i $(1 \le i \le 4)$ besitzt.
$\Gamma := \partial\Omega$

Sei $n \in \mathbb{N}$, $h := \frac{1}{2n}$; $k := n - 1$.

Die Triangulierung τ von Ω entsteht, indem man für $0 \le i \le n$

1.) die Punkte $P_i := (2ih, 0)$ und $T_i := (2ih + 0.5, 0.5/c)$ durch $S_i := \overline{P_iT_i}$,

2.) die Punkte $Q_i := (ih, ih/c)$ und $R_i := (1 + ih, ih/c)$ durch $S_i' := \overline{Q_iR_i}$,

3.) die Punkte P_i und Q_i durch $S_i'' := \overline{P_iQ_i}$,

4.) die Punkte T_i und R_i durch $S_i''' := \overline{T_iR_i}$,

verbindet.

Die Numerierung der Punkte sei lexikographisch (vgl.[8; p.51]).

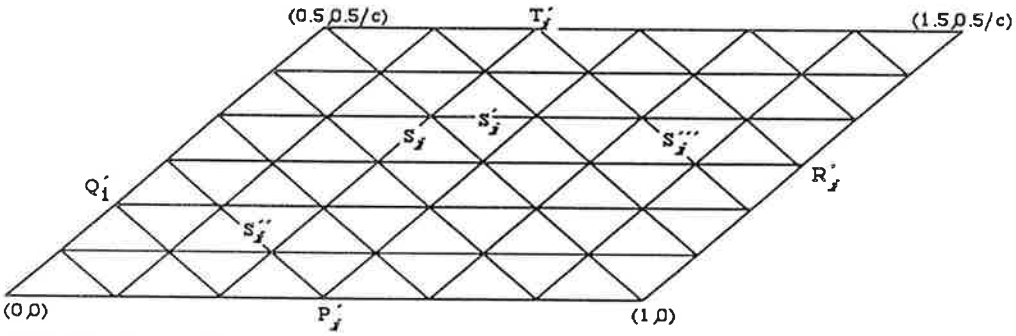
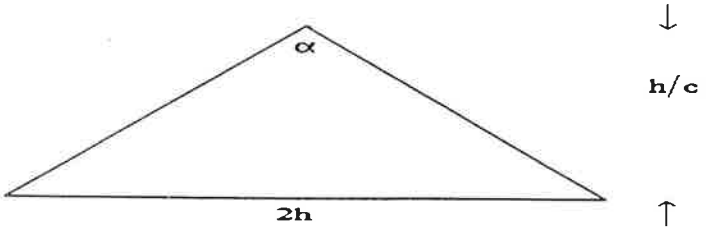


Abb.23 Ω aus Beispiel (6.8.17) mit beschriebener Triangulierung

(6.8.18) Bemerkung

Jedes Dreieck von τ hat die Gestalt:



und es gilt:

$$\cos \alpha = (1-c^2) \,/\, (1+c^2) \,.$$

Für $c = 1$ erhält man: $\alpha = \pi/2$; d.h. eine Quadratgittertriangulierung;

für $c \to \infty$ folgt $\alpha \to \infty$; d.h. man erhält eine entartete Triangulierung.

❑

Wir betrachten die Gleichung:

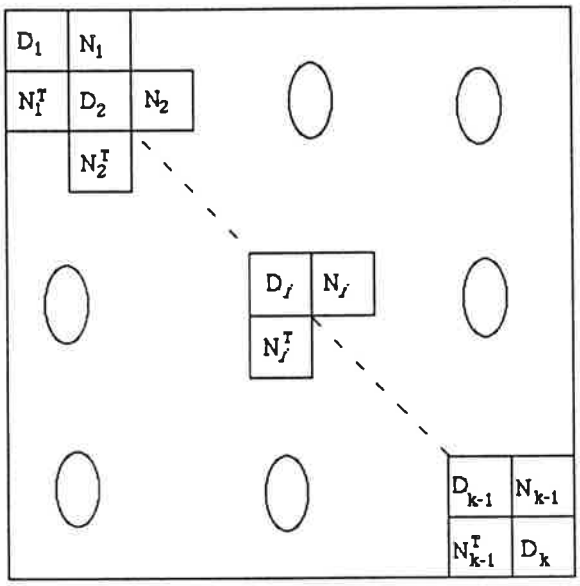
$$-\Delta u = f \quad \text{in } \Omega$$

$$u = 0 \quad \text{auf } \Gamma.$$

Diskretisierung mit stückweise linearen Finiten Elementen über τ ergibt ein LGS der Form:

$$K u = f$$

(vgl.4.1.2); wobei die $k^2 \times k^2$ Steifigkeitsmatrix K die Besetzungsstruktur besitzt:



mit den k x k Blöcken D und N:

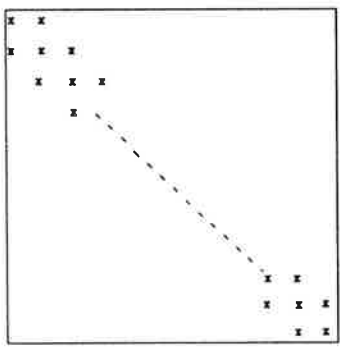
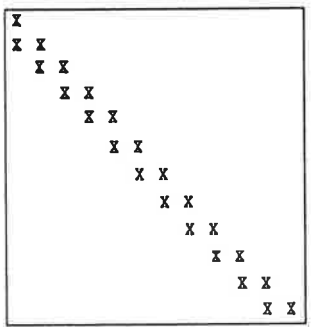
$D_j =$  $N_j =$  $1 \le j \le k$.

Abb.24 Besetzungsmuster der Steifigkeitsmatrix für die Triangulierung τ aus Beispiel(6.8.17).

Es gilt für die Elemente $(i,j) \in \mathfrak{I}_K$:

$$\text{a)} \quad K_{i,i} = 3c + \frac{1}{c}$$

$$\text{(6.8.19)} \quad \text{b)} \quad K_{i,i-1} = K_{i-1,i} = \frac{1}{2}\left(c - \frac{1}{c}\right)$$

$$\text{c)} \quad K_{i,i-k+1} = K_{i,i+k-1} = K_{i,i-k} = K_{i,i+k} = -c.$$

(6.8.20) Bemerkung

Für c > 1 ist K keine M-Matrix mehr.

❑

Beweis:

$$K_{i,i+1} = \frac{1}{2}\left(c - \frac{1}{c}\right) > 0 \; ; \quad \text{für } c > 1.$$

QED

Im folgenden werden wir die Elemente $\{K_{i,j}\}_{1 \le i,j \le k^2}$ von K mit

$$K_D := K_{i,i}$$

$$K_N := K_{i,i+1} = K_{i+1,i}$$

$$K_O := K_{i,i+k-1} = K_{i+k-1,i}$$

$$K_A := K_{i,i+k} = K_{i+k,i}$$

bezeichnen. Zur Vereinfachung der Notation bezeichnen wir die Elemente der Diagonalmatrix D aus der ILU-Zerlegung (6.8.1) mit :

$$D_i := D_{i,i} .$$

Wir geben nun die Formeln an zur Berechnung der unvollständigen Zerlegung von K.

$$D_i = K_D - \frac{U_{i-1,i}^2}{D_{i-1}} - \frac{U_{i-k+1,i}^2}{D_{i-k+1}} - \frac{U_{i-k,i}^2}{D_{i-k}} ,$$

(6.8.21)

(i) $$U_{i,i+k} = K_A ,$$

(ii) $$U_{i,i+k-1} = K_O - U_{i-1,i}\frac{K_A}{D_{i-1}} ,$$

(iii) $$U_{i,i+1} = K_N - U_{i-k+1,i}\frac{K_A}{D_{i-k+1}} .$$

Bei obigen Formeln ist zu beachten, daß alle Terme, die nicht positive Indizes enthalten, null gesetzt werden müssen.
Beachte: In den Formeln (ii) und (iii) wurde (i) bereits benutzt.

Wir wollen nun die Stabilität für die durch (6.8.21) beschriebene ILU-Zerlegung in Abhängigkeit des Parameters c untersuchen. Dabei werden wir sehen, daß die Zerlegung selbst im Entartungsfall ($c \to \infty$) stabil ist.
Die Beweisidee dabei ist, zunächst die Elemente der Zerlegungsmatrix U durch die Diagonalelemente K_D abzuschätzen (z.B. $|U_{i,i+1}| \le \text{const} \cdot K_D$), und diese Abschätzungen dann in die Formel zur Berechnung der Elemente der Diagonalmatrix D (6.8.21) einzusetzen. Damit kann man dann induktiv eine Abschätzung nach unten für die Elemente D_i gewinnen, der Form: $D_i \ge \alpha K_D$, wobei α nicht vom Störungsparameter c abhängt.
Um den Beweis, der sehr technisch ist, möglichst übersichtlich zu halten, werden die dafür nötigen Lemmata und Hilfsbehauptungen im Anschluß daran nachgeholt.

(6.8.22) Satz

Die durch (6.8.21) beschriebene ILU-Zerlegung der Matrix K ist für alle $c \ge 1$ stabil. Genauer gilt für $1 \le i \le k^2$

$$K_D \ge D_i \ge \alpha K_D \,, \qquad \text{mit } \alpha = 0.612.$$

❑

Beweis:
Die Behauptung gelte für $1 \le i \le m$; mit $m \in [1,k^2]$.
Lemma 6.8.27, Lemma 6.8.36 und Lemma 6.8.42 beweisen dann die Abschätzungen:

1)
$$|U_{i,i+k-1}| \le \frac{1}{3.6} K_D \,, \qquad \text{für } 1 \le i \le \min(m+1,k),$$
$$|U_{i,i+1}| \le \frac{1}{6} K_D \,, \qquad \text{für } 1 \le i \le k,$$

und falls $m > k$

2 a) für $1 \le c \le 1.9$:
$$|U_{i,i+k-1}| \le \frac{1}{3} K_D \,, \qquad \text{für } k+1 \le i \le \min(m+1,k^2-k),$$
$$|U_{i,i+1}| \le \frac{1}{6} K_D \,, \qquad \text{für } k+1 \le i \le \min(m+k-1,k^2-1),$$

2 b) für $c > 1.9$:
$$|U_{i,i+k-1}| \le \frac{1}{2.8997} K_D \,, \qquad \text{für } k+1 \le i \le \min(m,k^2-k),$$
$$|U_{i,i+1}| \le \frac{1}{11.65} K_D \,, \qquad \text{für } k+1 \le i \le \min(m+k-1,k^2-1),$$

sowie

für $1 \le c \le 1.9$:

3 a)
$$|U_{i,i+k}| \le 0.305 \cdot K_D \,, \qquad \text{für } 1 \le i \le k^2-k,$$

für $c > 1.9$:

3 b)
$$|U_{i,i+k}| \le \frac{1}{3} K_D \,, \qquad \text{für } 1 \le i \le k^2-k.$$

Wir können nun mit obigen Abschätzungen den Induktionsbeweis für D_i führen:

Für den ersten Block von K ($1 \le i \le k$) zeigen wir die Behauptung explizit, und folgern dann für die übrigen Blöcke induktiv.

Induktionsanfang :

Lemma (6.8.23) zeigt : Die Behauptung von Satz (6.8.22) gilt für die Elemente D_i des ersten Blocks ($1 \le i \le k$).

Induktionsschluß ($k+1 \le i \le k^2$) :

Es gelte : $D_m \ge 0.612 \cdot K_D$ für alle $m < i$.

1.Fall $1 \le c \le 1.9$

Mit den Abschätzungen für die Nebendiagonalelemente folgt :

$$D_i = K_D - \frac{U_{i-1,i}^2}{D_{i-1}} - \frac{U_{i-k+1,i}^2}{D_{i-k+1}} - \frac{U_{i-k,i}^2}{D_{i-k}}$$

$$\ge K_D - \left[\frac{1}{9} + 0.305^2 + \frac{1}{36}\right] \cdot \frac{K_D}{\beta} \overset{!}{\ge} \beta K_D$$

$$\Leftrightarrow \qquad -\beta^2 + \beta - \left[\frac{1}{9} + 0.305^2 + \frac{1}{36}\right] \ge 0 .$$

Die letzte Ungleichung ist erfüllt für $\beta = 0.6344$, d.h.

$$D_i \ge 0.6344 \cdot K_D > 0.612 \cdot K_D.$$

2. Fall $c > 1.9$

Analog wie beim ersten Fall hat man jetzt ein β zu finden, das die Ungleichung :

$$-\beta^2 + \beta - \left[\frac{1}{2.8997^2} + \frac{1}{9} + \frac{1}{11.65^2}\right] \ge 0$$

erfüllt. $\beta = 0.6121$ (> 0.612 !) genügt der Ungleichung.

Die Abschätzung der Elemente von D nach oben durch K_D ist trivial.

QED

(6.8.23) Lemma

Wir verwenden die Bezeichnungen aus Beispiel (6.8.17).
Für die Elemente D_i im ersten Block der Matrix D ($1 \le i \le k$) gilt:

$$D_i \ge \alpha \cdot K_D .$$

Beweis:
Wir beweisen die Behauptung induktiv:

Induktionsanfang:
Es gilt:

$$D_1 = K_D \ge 0.612 \cdot K_D .$$

Induktionsschluß:

Sei nun $2 \le i \le k$.

$$D_i = K_D - \frac{U_{i-1,i}^2}{D_{i-1}}$$

Wir nehmen an : $D_m \ge \beta \cdot K_D$ für alle $m < i$.

Dann folgt für D_i :

$$D_i = K_D - \frac{U_{i-1,i}^2}{D_{i-1}} \ge K_D - \frac{U_{i-1,i}^2}{\beta \cdot K_D} \overset{!}{\ge} \beta \cdot K_D$$

$$\Leftrightarrow \qquad -\beta^2 + \beta - \frac{U_{i-1,i}^2}{K_D^2} \ge -\beta^2 + \beta - \frac{1}{36} \ge 0 \qquad \text{für } \beta = 0.5 + \frac{1}{3}\sqrt{2} = 0.971\ldots$$

Das bedeutet :

$$D_i \ge 0.971 \cdot K_D > 0.612 \cdot K_D.$$

QED

Wir kommen nun zu den Abschätzungen für die Elemente der Zerlegungsmatrix U. Dazu müssen wir zunächst die Rekursionsformel (6.8.21) geeignet umformen.

(6.8.24) Lemma

In den folgenden Formeln sind diejenigen Terme null zu setzen, welche nicht positive Indizes enthalten. Dann gilt für die Elemente $\{U_{i,j}\}_{1 \le i,j \le k^2}$ von U:

(6.8.25)

a) $U_{i,i+k} = K_A$,

b) $U_{i,i+k-1} = K_O - \frac{K_A}{D_{i-1}} K_N + \frac{K_A^2}{D_{i-k} \cdot D_{i-1}} U_{i-k,i-1}$,

c) $U_{i,i+1} = K_N - \frac{K_A}{D_{i-k+1}} K_O + \frac{K_A^2}{D_{i-k} \cdot D_{i-k+1}} U_{i-k,i-k+1}$.

Indem wir die Beziehungen (6.8.19) benützen, erhalten wir :

(6.8.26)

a) $U_{i,i+k} = -c$

b) $U_{i,i+k-1} = 0$ für $i = j \cdot k + 1$; $j \in \mathbb{N}$

c) $U_{i,i+k-1} = -c + \frac{1}{2} \frac{c^2 - 1}{D_{i-1}} + \frac{c^2}{D_{i-1} \cdot D_{i-k}} U_{i-k,i-1}$ für $i \ne j \cdot k + 1$

d) $U_{i,i+1} = 0$ für $i = j \cdot k$

e) $U_{i,i+1} = \frac{1}{2}\left(c - \frac{1}{c}\right) - \frac{c^2}{D_{i-k+1}} + \frac{c^2}{D_{i-k+1}} U_{i-k,i-k+1}$ für $i \ne j \cdot k$.

❑

Beweis:
Der Beweis erfolgt durch einfaches Einsetzen und wird hier nicht ausgeführt.

Wir werden nun die im Beweis von Satz (6.8.22) verwendeten Abschätzungen für die Elemente der Matrix U beweisen.

1.) Abschätzungen für $U_{i,i+k-1}$

(6.8.27) Lemma

Es gelte Satz (6.8.22) für $1 \le i \le m$, mit $m \in [1,k^2]$. Dann gilt:

a) $$|U_{jj+k-1}| \le \frac{5}{18} K_D \qquad \text{für } 2 \le j \le \min(m+1,k),$$

und falls $m > k$:

b) $$|U_{i,i+k-1}| \le \frac{1}{2.8997} K_D \qquad \text{für } k+1 \le i \le \min(m+1,k^2 - k + 1).$$

c) Für $1 \le c \le 1.9$ kann man Ungleichung (b) verschärfen:

$$|U_{i,i+k-1}| \le \frac{1}{3} K_D \qquad \text{für } k+1 \le i \le \min(m+1,k^2 - k + 1).$$

❑

Beweis:

zu (a):

Sei $D := D_s$ mit $s < j \le \min(m+1,k)$. Dann gilt:

$$\frac{1}{2} \cdot \frac{c^2 - 1}{3c^2 + 1} \le \frac{1}{6} \qquad \text{für } c \ge 1$$

$$\Rightarrow \qquad \frac{1}{2} \cdot \frac{c^2 - 1}{3c^2 + 1} \le \alpha \qquad \text{mit } \alpha \text{ aus Behauptung.}$$

$$\Rightarrow \qquad \frac{1}{2} \cdot \frac{c^2 - 1}{c} \le \alpha\left(3c + \frac{1}{c}\right) = \alpha K_D \le D \qquad \text{(nach Voraussetzung)}$$

$$(6.8.28) \Rightarrow \qquad -c + \frac{1}{2} \cdot \frac{c^2 - 1}{D} \le 0$$

$$\Rightarrow \qquad |U_{jj+k-1}| = c - \frac{1}{2} \cdot \frac{c^2 - 1}{D} .$$

Einfache Rechnung zeigt:

$$c - \frac{1}{2} \cdot \frac{c}{3c^2 + 1} (c^2 - 1) \le \frac{5}{18}\left(3c + \frac{1}{c}\right)$$

$$\Rightarrow \qquad c - \frac{1}{2} \frac{c^2 - 1}{D} \le c - \frac{1}{2} \cdot \frac{c^2 - 1}{3c + \frac{1}{c}} \le \frac{5}{18}\left(3c + \frac{1}{c}\right)$$

$$\Rightarrow \qquad |U_{jj+k-1}| \le \frac{5}{18} K_D .$$

zu (b)
Wegen Gleichung (6.8.28), (6.8.26 c) und (6.8.19 c) gilt:

$$U_{i,i+k-1} \leq 0$$

d.h. $$|U_{i,i+k-1}| = -U_{i,i+k-1} \quad \text{für } 1 \leq i \leq \min(m+1,k^2 - k + 1).$$

Wir wollen nun induktiv Konstanten q_s bestimmen, die die Abschätzung:

(6.8.29) $$|U_{i,i+k-1}| \leq 1/q_s \cdot K_D$$

für $1 \leq i \leq \min(m+1,k^2 - k + 1)$; $s := [(i - 1)/k] + 1$ erfüllen (d.h. $U_{i,i+k-1}$ liegt im s.ten Nebendiagonalblock von oben gerechnet). $[\cdot]$ bezeichnet hier die Gaußklammer. Sei i_{max} und n_{max} definiert durch:

$$i_{max} := \min(m+1,k^2 - k + 1); \qquad n_{max} := [(i_{max} - 1)/k] + 1.$$

Dann gilt:
Die Folge $\{q_s\}_{1 \leq s \leq n_{max}}$ definiert durch:

$$q_1 = 3.6$$

(6.8.30) $$q_{s+1} = \frac{18 q_s \alpha^2}{6 q_s \alpha^2 - q_s \alpha + 2} \qquad \text{für } s \in \mathbb{N},\ 1 \leq s \leq n_{max}$$

erfüllt Abschätzung (6.8.29). Wir zeigen dies induktiv:

Induktionsanfang:

Teil (a) beweist: $q_1 = 3.6$.

Induktionsannahme:

Formel (6.8.30) sei bewiesen für $1 \leq s \leq n$, mit $n = n_{max} - 1$.
Sei also $|U_{i-k,i-1}| \leq 1/q_n \cdot K_D$ für $[(i - 1)/k] = n$ (d.h. $U_{i-k,i-1}$ liegt im n.ten Nebendiagonalblock).

Induktionsschluß:

Wir setzen die Induktionsannahme in (6.8.26 c) ein und bilden die Beträge:

$$|U_{i,i+k-1}| \leq c - \frac{1}{2} \cdot \frac{c^2 - 1}{D_{i-1}} + \frac{c^2}{D_{i-1} \cdot D_{i-k}} \cdot \frac{K_D}{q_n}$$

Wir zeigen, daß $\rho = q_{n+1}$ die Ungleichung:

(6.8.31) $$c - \frac{1}{2} \cdot \frac{c^2 - 1}{D_{i-1}} + \frac{c^2}{D_{i-1} \cdot D_{i-k}} \cdot \frac{K_D}{q_n} \leq \frac{K_D}{\rho},$$

erfüllt.

(6.8.31) ist äquivalent zu:

$$0 \geq 2 \cdot \left[\rho \cdot c^2 - (3c^2 + 1)\right] q_n \cdot D_{i-1} D_{i-k} - q_n \cdot \rho \cdot c (c^2 - 1) D_{i-k} + 2c^2 (3c^2 + 1) \rho$$

$$=: g(D_{i-1}, D_{i-k}) .$$

Indem wir die Konstanten R, S und T definieren durch:

$$
\begin{aligned}
R &:= 2\cdot\left[\rho\cdot c^2 - (3c^2+1)\right] q_n, \\
S &:= q_n\cdot\rho\cdot c\,(c^2-1), \\
T &:= 2c^2(3c^2+1)\,\rho\,,
\end{aligned}
\qquad (6.8.32)
$$

können wir g wie folgt schreiben:

$$g(y,z) = R\,y\,z - S\,y + T\ .$$

Lemma (6.8.43) beweist, daß ein $D \in I := \{\min(D_{i-1}, D_{i-k}), \max(D_{i-1}, D_{i-k})\}$ existiert mit:

$$g(D_{i-1}, D_{i-k}) \le g(D,D) =: f(D).$$

Wir zeigen: $f(D) \le 0$.

Diskussion von f:

1. Fall: $R \le 0$

Da gilt: $S \ge 0$ und $T > 0$, besitzt $f(y)$ in diesem Fall genau eine positive Nullstelle. Wir suchen also ρ so, daß $f(\alpha(3c + 1/c)) \le 0$ erfüllt ist. Da $f(y)$ für $y \in I$ monoton fallend ist, folgt daraus dann:

$$f(y) \le 0 \qquad \forall\, y \in I.$$

Lemma (6.8.44) beweist für $q_n \le 4$, daß gilt:

$$f(\alpha(3c + 1/c)) \le 0\ ,$$

für :

$$\rho = \frac{18\,q_n\cdot\alpha^2}{6\,q_n\cdot\alpha^2 - q_n\cdot\alpha + 2}\ . \qquad (6.8.33)$$

2. Fall: $R > 0$

$f(y)$ besitzt in diesem Fall höchstens zwei reelle Nullstellen, die beide positiv sein können. Da $f(y)$ in diesem Fall eine nach oben geöffnete Parabel ist, müssen wir ρ so bestimmen, daß:

i) $\quad f\{\alpha(3c + 1/c)\} \le 0\ ,$

ii) $\quad f\{3c + 1/c\} \le 0\ .$

Lemma (6.8.44) beweist für $q_n \le 4$, daß (ii) erfüllt ist, für

$$\rho \le \frac{18\,q_n}{5\,q_n + 2}\ .$$

Bedingung (i) ist dieselbe wie im Fall 1. Für $q_n \le 4$ gilt:

$$\frac{18\,q_n\cdot\alpha^2}{6\,q_n\cdot\alpha^2 - q_n\cdot\alpha + 2} \le \frac{18\,q_n}{5\,q_n + 2}.$$

Auf Grund von Lemma (6.8.44) folgt, daß die Folge $\{q_n\}$ streng monoton fallend gegen 2.8997... konvergiert. Damit ist sie insbesondere durch 3.6 nach oben beschränkt (vgl. Voraussetzungen bei Fall 1 und 2). Damit ist die Abschätzung (6.8.29) mit $\{q_s\}$ –definiert durch (6.8.30)– gezeigt.
Das bedeutet zusammenfassend:

(6.8.34)

$$\text{i)} \quad |U_{i,i+k-1}| \le \frac{5}{18} K_D \qquad \text{für } 1 \le i \le \min(m+1,k)$$

$$\text{ii)} \quad |U_{i,i+k-1}| \le \frac{1}{2.8997} K_D \qquad \text{für } k+1 \le i \le \min(m+1,k^2 - k + 1) \,.$$

zu (c)
Wir zeigen, daß für $1 \le c \le 1.9$ und $1 \le \min(m+1,k^2 - k + 1)$ gilt:

(6.8.35)

$$|U_{i,i+k-1}| \le \frac{1}{3} K_D$$

per Induktion.

Induktionsanfang:
Aus (6.8.34 i) folgt (6.8.35) für $1 \le i \le \min(m+1,k)$.

Induktionsannahme:
Sei n_{max} definiert wie beim Beweis von Formel (6.8.30) und $n := n_{max}-1$. Es gelte:

$$|U_{i-k,i-1}| \le 1/3 \cdot K_D \qquad \text{für } [(i-1)/k] = n$$

d.h. $U_{i-k,i-1}$ liegt im $(n_{max} - 1)$.ten Nebendiagonalblock.

Induktionsschluß:
Wir zeigen:

$$|U_{i,i+k-1}| \le 1/3 \cdot K_D.$$

Für $\rho = 3$ ist R -definiert durch (6.8.32)- negativ, d.h. wir haben wieder die Situation von Fall 1 vorliegen. Wir müssen zeigen, daß für $\rho = 3$ gilt: $f\{\alpha(3c + 1/c)\} \le 0$.
Sei $h(c)$ definiert durch:

$$h(c) := f\{\alpha(3c + 1/c)\}.$$

Für $\rho = 3$ und $q_n = 3$ besitzt h die Form:

$$h(c) = \Theta \cdot c^4 + \Lambda \cdot c^2 + \Psi;$$

wobei die Konstanten Θ, Λ, Ψ definiert sind durch:

$$\Theta := -9\,\alpha + 6 = 0.492,$$

$$\Lambda := -1.233792,$$

$$\Psi := -2.247264\,.$$

$h(c)$ besitzt zwei reelle Nullstellen :

$$P_1 = -1.93\ldots, \quad P_2 = 1.93\ldots,$$

und ist eine nach oben geöffnete Parabel.

$\Rightarrow$ $\qquad h(c) \le 0 \quad$ für $1 \le c \le 1.9$

QED

2.) Abschätzungen für $U_{i,i+1}$

(6.8.36) Lemma

Es gelte Satz (6.8.22) für $1 \le i \le m$, mit $m \in [1,k^2]$. Dann gilt:

(6.8.37) a) $U_{i,i+1} \ge 0 \qquad 1 \le i \le k$

und falls $m > k$

b) $U_{i,i+1} \ge -K_D/6 \qquad k+1 \le i \le \min(m+k-1,k^2-1).$

Für $c > 1.9$ können wir die Ungleichung (6.8.37 b) noch verschärfen:

c) $U_{i,i+1} \ge -K_D/16 \qquad k+1 \le i \le \min(m+k-1,k^2-1).$

$U_{i,i+1}$ läßt sich nach oben abschätzen durch:

(6.8.38) a) $U_{i,i+1} \le K_D/6 \qquad 1 \le i \le k$

b) $U_{i,i+1} \le K_D/11.65 \qquad k+1 \le i \le \min(m+k-1,k^2-1).$

❑

(6.8.39) Bemerkung

Die Ungleichungen (6.8.37) und (6.8.38) zusammengefaßt ergeben:

(6.8.40) a) $|U_{i,i+1}| \le K_D/6 \qquad 1 \le i \le \min(m+k-1,k^2-1)$

und falls $m > k$, $c > 1.9$ können wir (6.8.38 a) ab dem zweiten Block verschärfen:

b) $|U_{i,i+1}| \le K_D/11.65 \qquad k+1 \le i \le \min(m+k-1,k^2-1).$

❑

Beweis von Lemma (6.8.36):

zu (6.8.37 a):

Es gilt :

$$U_{i,i+1} = \frac{1}{2}\left(c - \frac{1}{c}\right) \ge 0 \qquad \text{für } c \ge 1 \text{ und } 1 \le i \le k.$$

zu (6.8.37 b):
Sei $1 \le c \le 1.9$. Wir zeigen (6.8.37 b) per Induktion über die Diagonalblöcke von U.

Induktionsanfang:

(6.8.37 a) beweist (6.8.37 b) für $1 \le i \le k$.

Induktionsannahme:

Sei i_{ind} definiert durch:

$$i_{ind} := \min(m+k-1,k^2-1)-k+1.$$

(6.8.37 b) gelte für $1 \le j \le i_{ind}$.

Induktionsschluß:

Auf Grund der Induktionsannahme gilt:

$$U_{j,j+1} \geq -K_D/6 \qquad \text{für } 1 \leq j \leq i_{ind}.$$

Indem wir diese Abschätzung in (6.8.26 e) einsetzen, erhalten wir für $i_{ind} + 1 \leq i \leq i_{ind} + k$:

$$U_{i,i+1} = \frac{1}{2}\left(c - \frac{1}{c}\right) - \frac{c^2}{D_{i-k+1}} + \frac{c^2}{D_{i-k+1} \cdot D_{i-k}} U_{i-k,i-k+1}$$

$$\geq \frac{1}{2}\left(c - \frac{1}{c}\right) - \frac{c^2}{D_{i-k+1}} - \frac{c^2}{D_{i-k+1} \cdot D_{i-k}} \cdot \frac{3c + \frac{1}{c}}{6}$$

$$=: \eta\,(D_{i-k}, D_{i-k+1})$$

Es genügt zu zeigen:

$$\eta(D_{i-k+1}, D_{i-k}) \geq -K_D/6.$$

Dies ist äquivalent zu:

$$(6.8.41) \qquad 0 \leq \left[\frac{1}{2}\left(c - \frac{1}{c}\right) + \frac{1}{6}\left(3c + \frac{1}{c}\right)\right] D_{i-k+1} \cdot D_{i-k} - c^2 D_{i-k} - \frac{c^2}{6}\left(3c + \frac{1}{c}\right)$$

$$=: \psi(D_{i-k+1}, D_{i-k}).$$

Wir definieren die Größen Φ, Ξ, Θ durch:

$$\Phi := \frac{1}{2}\left(c - \frac{1}{c}\right) + \frac{1}{6}\left(3c + \frac{1}{c}\right),$$
$$\Xi := c^2,$$
$$\Theta := \frac{c^2}{6}\left(3c + \frac{1}{c}\right).$$

Damit läßt sich ψ in der Form schreiben:

$$\psi(y,z) = \Phi\, y\, z - \Xi\, z - \Theta.$$

Sei $\mathfrak{m} := \alpha \cdot (3c + 1/c)$, $\mathfrak{M} := (3c + 1/c)$ und das Intervall I definiert durch $I := [\mathfrak{m}, \mathfrak{M}]$. Wegen $\Phi \geq 0$ ist ψ monoton wachsend in y. Es gilt weiter:

$$\Phi\, y - \Xi \geq \Phi\, \mathfrak{m} - \Xi \geq 0, \qquad \forall y \in I$$

da aus $c \geq 1$ folgt:

$$(9\alpha - 3)c^4 - 2\alpha \geq 0$$

$$\Rightarrow \qquad \Phi_{m} - \Xi \geq \left\{\left[\frac{1}{2}(c^2 - 1) + \frac{1}{6}(3c^2 + 1)\right]\alpha(3c^2 + 1) - c^4\right\}/c^2 \geq 0 .$$

Das bedeutet ψ ist monoton wachsend in z für $z \in I$. Daraus folgt weiter:

$$\psi(y,z) \geq \psi(m,m) \qquad \forall y,z \in I.$$

Die Behauptung folgt dann aus $\psi(m,m) \geq 0$. Dies gilt wegen:

$$\psi(m,m) \geq 0$$

$$\Leftarrow \qquad h(c) := (18\alpha^2 - 6\alpha - 1)c^4 - 2\alpha^2 \geq 0$$

$$\Leftrightarrow \qquad 2.069792\, c^4 - 0.749088 \geq 0 .$$

Da h eine nach oben geöffnete Parabel ist, folgt aus $h(1) > 0 : h(c) > 0 ; \forall\, c > 1$.

zu (6.8.37 c)
Analog erhalten wir im Fall $c > 1.9$:

Induktionsanfang:

(6.8.37 a) beweist (6.8.37 c) für $1 \leq i \leq k$.

Induktionsannahme:

(6.8.37 c) gelte für $1 \leq j \leq i_{ind}$, mit i_{ind} aus dem Beweis von (6.8.37 b).

Induktionsschluß:

Setzen wir nun die Abschätzung $U_{j,j+1} \leq -K_D/16$ in (6.8.26 e) ein, erhalten wir diesmal:

$$U_{i,i+1} \geq \frac{1}{2}\left(c - \frac{1}{c}\right) - \frac{c^2}{D_{i-k+1}} - \frac{c^2}{D_{i-k+1}\cdot D_{i-k}} \cdot \frac{3c + \frac{1}{c}}{16}$$

$$=: \overline{\eta}(D_{i-k}, D_{i-k+1})$$

Wir zeigen:

$$\overline{\eta}(D_{i-k+1}, D_{i-k}) \geq -K_D/16.$$

Dies ist äquivalent zu:

$$(6.8.42) \qquad 0 \leq \left[\frac{1}{2}\left(c - \frac{1}{c}\right) + \frac{1}{16}\left(3c + \frac{1}{c}\right)\right] D_{i-k+1}\cdot D_{i-k} - c^2 D_{i-k} - \frac{c^2}{16}\left(3c + \frac{1}{c}\right)$$

$$=: \overline{\psi}(D_{i-k+1}, D_{i-k}) .$$

Wir schließen für $\overline{\psi}$ analog wie beim Beweis von (6.8.37 b) und müssen diesmal zeigen:

(6.8.42) $$h(c) := (33\,\alpha^2 - 16\,\alpha - 1)\,c^4 - 10\,\alpha^2 c^2 - 7\,\alpha^2 \geq 0.$$

Dies ist äquvalent zu:

$$1.567952\,c^4 - 3.74544\,c^2 - 2.621808 \geq 0\,.$$

$h(c)$ ist eine nach oben geöffnete Parabel mit zwei reellen Nullstellen, einer positiven und einer negativen. Aus $h(1.9) = 4.29\ldots > 0$ folgt dann die Behauptung für $c > 1.9$.

Wir beweisen jetzt Abschätzungen nach oben für $U_{i,i+1}$.

zu (6.8.38 a)

Für den ersten Diagonalblock von U gilt:

$$U_{i,i+1} = \frac{1}{2}\left(c - \frac{1}{c}\right) \leq \frac{1}{6}\left(3c + \frac{1}{c}\right) \qquad 1 \leq i \leq k.$$

zu (6.8.38 b)

Wir führen den Beweis wieder induktiv über die Diagonalblöcke von U, wobei zu beachten ist, daß (6.8.38 b) nur gilt für $i \geq k+1$. Gelte also Satz (6.8.22) $1 \leq j \leq m$, mit $m > k$.

Induktionsanfang:

Wir zeigen die Behauptung zunächst für den zweiten Block ($k+1 \leq i \leq \min(m+k-1, 2\cdot k)$. Indem wir (6.8.38 a) in (6.8.26 e) einsetzen, erhalten wir:

$$U_{i,i+1} = \frac{1}{2}\left(c - \frac{1}{c}\right) - \frac{c^2}{D_{i-k+1}} + \frac{c^2}{D_{i-k+1}\cdot D_{i-k}}\,U_{i-k,i-k+1}$$

$$\leq \frac{1}{2}\left(c - \frac{1}{c}\right) - \frac{c^2}{D_{i-k+1}} + \frac{c^2}{D_{i-k+1}\cdot D_{i-k}} \cdot \frac{3c + \frac{1}{c}}{6}$$

$$=: \tau\,(D_{i-k+1}, D_{i-k})\ .$$

Wir wollen zeigen :

$$\tau(D_{i-k+1}, D_{i-k}) \leq \frac{3c + \frac{1}{c}}{11.65}\ .$$

Dies ist äquivalent zu :

$$0 \geq \left[\frac{1}{2}\left(c - \frac{1}{c}\right) - \frac{1}{11.65}\left(3c + \frac{1}{c}\right)\right] D_{i-k+1}\,D_{i-k} - c^2 D_{i-k} + \frac{c^2}{6}\left(3c + \frac{1}{c}\right)$$

$$=: \varphi(D_{i-k+1}, D_{i-k})\,.$$

Zur Vereinfachung definieren wir die Größen U,V und W durch:

$$U := \left\{\frac{1}{2}\left(c - \frac{1}{c}\right) - \frac{1}{11.65}\left(3c + \frac{1}{c}\right)\right\},$$
$$V := c^2,$$
$$W := \frac{c^2}{6}\left(3c + \frac{1}{c}\right).$$

φ läßt sich dann darstellen durch:

$$\varphi(y,z) = U\,y\,z - V\,z + W.$$

Man sieht sofort:

$$-V\alpha\,(3\,c + 1/c\,) + W \le 0.$$

Falls $U \le 0$ ist, folgt deshalb (6.8.38 b).

Sei nun $U > 0$ und $\mathfrak{w} := \alpha(3\,c + 1/c)$, $\mathfrak{M} := (3\,c + 1/c)$. Das Intervall I sei definiert durch $I := [\mathfrak{w}, \mathfrak{M}]$.
Da $U > 0$ ist, ist φ monoton wachsend in y. Weiter gilt: $U\mathfrak{M} - V \le 0$, da mit $\kappa = 11.65$ gilt:

$$U \cdot \mathfrak{M} - V = \left[\frac{1}{2}\left(c - \frac{1}{c}\right) - \frac{1}{\kappa}\left(3\,c + \frac{1}{c}\right)\right]\left(3\,c + \frac{1}{c}\right) - c^2 \le 0$$

$\Leftrightarrow$

$$(\kappa - 18)\,c^4 - 2\,(6 + \kappa)\,c^2 - (2 + \kappa) \le 0.$$

Das heißt: φ ist monoton fallend in z, und zusammenfassend:

Wir zeigen: $$\varphi(y,z) \le \varphi(\mathfrak{M},\mathfrak{w}).$$

$$\varphi(\mathfrak{M},\mathfrak{w}) \le 0.$$

Mit $\kappa = 11.65$ gilt:

$$\varphi(\mathfrak{M},\mathfrak{w}) = \left[\frac{1}{2}\left(c - \frac{1}{c}\right) - \frac{1}{\kappa}\left(3\,c + \frac{1}{c}\right)\right]\left(3\,c + \frac{1}{c}\right)^2\alpha - c^2\left(3\,c + \frac{1}{c}\right)\alpha + \frac{c^2}{6}\left(3c + \frac{1}{c}\right) \le 0$$

$\Leftrightarrow$

$$[3\,\kappa\,\alpha - 54\,\alpha + \kappa]\,c^4 - 6\,\alpha\,[\kappa + 6]\,c^2 - 3\,\alpha\,[2 + \kappa] \le 0$$

$\Leftrightarrow$

$$-0.0086\,c^4 - 64.8108\,c^2 - 25.0614 \le 0.$$

Damit ist der Induktionsanfang bewiesen. Um jetzt beim Induktionsschluß vom Block n auf den Block $n + 1$ zu schließen, müßte man die Induktionsannahme: $U_{j,j+l} \le K_D/11.65$ in Formel (6.8.26 e) einsetzten und dann die Aussage für $U_{j+k,j+k+l}$ beweisen. Da aber gilt $K_D/11.65 \le K_D/6$ kann man den Beweis für den Induktionsanfang wörtlich wiederholen.

QED

3.) Abschätzungen für $U_{i,i+k}$

(6.8.42) Lemma

Es gilt für $1 \le i \le k^2 - k$:

$$|U_{i,i+k}| = c\,,$$

$$|U_{i,i+k}| \le \begin{cases} 0.305 \cdot K_D & \text{für } 1 \le c \le 1.9 \\ 1/3 \cdot K_D & \text{für } c > 1.9 \end{cases}.$$

Beweis: ❑
Die Behauptung folgt aus:

$$\frac{|U_{i,i+k}|}{K_D} = \frac{c^2}{3c^2+1} \le \begin{cases} 0.305 & \text{für } 1 \le c \le 1.9 \\ 1/3 & \text{für } c > 1.9 \end{cases}.$$

QED

Es bleiben jetzt noch die technischen Lemmata für Lemma (6.8.27) nachzutragen.

(6.8.43) Lemma

Die Funktion g aus dem Beweis von Lemma (6.8.27) sei definiert durch (6.8.32):

$$g(y,z) = R\,y\,z - S\,y + T\,,$$

mit R, S und T aus (6.8.32). Sei $\mathfrak{m} := \alpha\cdot(3c + 1/c)$, $\mathfrak{M} := (3c + 1/c)$ und das Intervall I definiert durch. I:= $[\mathfrak{m}, \mathfrak{M}]$ (α aus Satz (6.8.22)). Dann exisiert ein $\xi \in I$, so daß gilt:

$$g(\xi,\xi) \ge g(y,z) \quad \forall y,z \in I.$$

Beweis:
1. Fall: Sei $R \le 0$.
$\Rightarrow$ $R\cdot\mathfrak{m} - S < 0 \Rightarrow g(y,z) \le \mathfrak{m}(R\cdot\mathfrak{m} - S) = g(\mathfrak{m},\mathfrak{m})$.

2. Fall: Sei $R > 0$ und $R\cdot\mathfrak{M} - S \ge 0$.
$\Rightarrow$ $R\cdot\mathfrak{M} - S < 0 \Rightarrow g(y,z) \le \mathfrak{M}(R\cdot\mathfrak{M} - S) = g(\mathfrak{M},\mathfrak{M})$.

3. Fall: Sei $R > 0$ und $R\cdot\mathfrak{M} - S < 0$.
Dieser Fall kann nicht eintreten, da einfache Rechnung zeigt, daß für $c \ge 1$ gilt:

$$3.6 \le \frac{2(3c^2+1)^2}{c^2(5c^2+3)}.$$

Auf Grund Lemma (6.8.44) folgt daher, daß gilt:

$$\rho \le \frac{2(3c^2+1)^2}{c^2(5c^2+3)}\,;$$

und daraus durch einfaches Umformen:

$$R\cdot\mathfrak{M} - S > 0.$$

QED

(6.8.44) Lemma

Wir verwenden die Bezeichnungen wie im Beweis von Lemma (6.8.27). Sei $x \in \{\alpha,1\}$. Es gilt für $q_n \leq 4$:

$$\rho \leq \frac{18 q_n x^2}{6 q_n x^2 - q_n x + 2} \overset{!}{\Rightarrow} f(x(3c + 1/c)) \leq 0$$

Weiter gilt: $\{q_n\}_{1 \leq n \leq n_{max}}$ ist eine positive, monoton fallende Folge, die gegen:

$$q_\infty := \frac{18 x^2 - 2}{x(6x - 1)} = \begin{cases} 2.8997\ldots & \text{für } x = \alpha \\ 16/5 & \text{für } x = 1 \end{cases}$$

konvergiert.

❑

Beweis:

$$\Leftrightarrow \quad f(x(3c + 1/c)) \leq 0$$

$$\{6 q_n (\rho - 3) x^2 - q_n \rho x + 2\rho\} c^4 +$$

$$+ q_n x \{(2x + 1)\rho - 12x\} c^2 - 2 q_n x^2 \overset{!}{\leq} 0$$

Obige Ungleichung ist erfüllt, da

a) $$6 q_n (\rho - 3) x^2 - q_n \rho x + 2\rho \leq 0 \Leftrightarrow \rho \leq \frac{18 q_n x^2}{6 q_n x^2 - q_n x + 2},$$

b) $$(2x + 1)\rho - 12x \leq 0 \Leftrightarrow \rho \leq \frac{12x}{2x + 1},$$

c) $$\frac{18 q_n x^2}{6 q_n x^2 - q_n x + 2} \leq \frac{12x}{2x + 1} \quad \text{für } q_n \leq 4.$$

Der Beweis der Monotonie und des Grenzwerts ist trivial.

QED

§7 Realisierung des Verfahrens auf dem Computer

In diesem Abschnitt werden wir uns damit beschäftigen, ein Computerprogramm zur Lösung der "Shallow-Water-Equations auf dem Bodensee mit den beschriebenen Methoden zu entwickeln.

Das Programm gliedert sich in 5 Phasen:

1.) Definition und Initialisierung aller Variablen und Steuerungsparameter,

2.) Generierung der Triangulierungen $\tau_0, \ldots, \tau_{\ell_{max}}$,

3.) Aufstellen des linearen Gleichungssystems auf jedem Gitter und unvollständige Zerlegung der Steifigkeitsmatrix,

4.) Mehrgitterverfahren zur Lösung des linearen Gleichungssystems auf dem feinsten Gitter,

5.) Auswerten der Ergebnisse.

7.1 Beschreibung der Phasen des Programms zur Berechnung der Eigenschwingungen des Bodensees.

Es werden nur die grundsätzlichen Ideen und Konzepte dargelegt, da eine spezielle detaillierte Beschreibung des Programmcodes zu technisch und unübersichtlich würde.

" zu 1 "

(i) Datenstruktur:
Sei NP_ℓ die Anzahl aller Knotenpunkte der Triangulierung τ_ℓ und:

$$SNP := \sum_{i=0}^{\ell_{max}} NP_i \ .$$

Um das ILU-Verfahren als Glätter sinnvoll benutzen zu können (§6), müssen wir die Elemente der Steifigkeitsmatrix K_ℓ $(1 \le \ell \le l_{max})$ abspeichern. K_ℓ besteht aus sieben Diagonalen und wenigen periodischen Ausnahmen (vgl. §6.5).
Sei NA_ℓ die Anzahl der periodischen Ausnahmestellen auf dem Gitter τ_ℓ und

$$SNA := \sum_{i=0}^{\ell_{max}} NA_i \ .$$

Da wir das Programm flexibel halten wollen für den Fall, daß K_ℓ nicht symmetrisch ist, speichern wir alle siebenDiagonalen ab im Feld $K_\ell(7,NP_\ell)$ und die Ausnahmeelemente im Feld $K_\ell^A(2,NA_\ell)$.

Die Felder $K_\ell(\cdot,\cdot)$ bzw. $K_\ell^A(\cdot,\cdot)$ speichern wir aus programmiertechnischen Gründen nicht alle einzeln ab, sondern schreiben sie hintereinander (bzgl. ℓ) in die Felder K(7,SNP) bzw. K^A(2,SNA).

Sei K_ℓ unvollständig zerlegt in:

$$K_\ell = (L_\ell + D_\ell)\, D_\ell^{-1}\, (U_\ell + D_\ell) - N_\ell$$

und

$$K_\ell^{ILU} := L_\ell + D_\ell + U_\ell \ .$$

Diejenigen Elemente $(K_\ell^{ILU})_{i,j}$ von K_ℓ^{ILU} für die gilt : $(i,j) \in \mathfrak{P}_{K_\ell}$ speichern wir entsprechend wie für K_ℓ erklärt ab und erhalten die Felder : $K^{ILU}(7, SNP)$ und $K^{ILU,A}(2, SNP)$.

Auf Grund der Musterdefinition (6.6.7) gilt für $\mathfrak{P}_{Z_\ell} := \mathfrak{P}_{K_\ell^{ILU}} \setminus \mathfrak{P}_{K_\ell}$:

$$|\mathfrak{P}_{Z_\ell}| = 2 \cdot NP_\ell - O(1) \ .$$

Das heißt, wir benötigen noch ein weiteres Feld $ZUS_\ell(2,NP_\ell)$ bzw. ZUS(2,SNP) zur Abspeicherung dieser zusätzlichen Elemente.
Zusammengefaßt bedeutet das, daß wir für die Abspeicherung der Steifigkeitsmatrizen und der zugehörigen ILU-Zerlegungen auf allen Gittern insgesamt 16·SNP + 4·SNA Speicherplätze benötigen.
Da die Massenmatrix M_ℓ die gleiche Besetzungsstruktur wie K_ℓ besitzt, würde man für die Abspeicherung von M_ℓ ebenfalls 7·SNP Speicherplätze benötigen. Wir verwenden statt der Massenmatrix M_ℓ die " gelumpte " Massenmatrix M_ℓ^{lump}, definiert durch:

$$(M_\ell^{lump})_{i,i} := \sum_{j=1}^{NP_l} (M_\ell)_{i,j} \qquad \text{für } 1 \le i \le NP_\ell$$

$$(M_\ell^{lump})_{i,j} := 0 \qquad \text{für } i \ne j \ ,$$

und benötigen somit nur noch 1·SNP Speicherplätze zur Abspeicherung von M_ℓ^{lump} ($1 \le \ell \le \ell_{max}$). Die Abschätzungen für den Diskretisierungsfehler aus §4.2 bleiben erhalten (vgl. [19 und den dort gegebenen Referenzen]), falls man die Gleichung:

$$(K_\ell - \lambda_\ell M_\ell)\, u = f$$

ersetzt durch:

$$(K_\ell - \lambda_\ell\ M_\ell^{lump})\, u = f \ .$$

Desweiteren benötigen wir noch Vektoren u_ℓ der Länge NP_ℓ, die die Iterierten im Glättungsverfahren auf den Gittern τ_l bezeichnen und Vektoren f_l, welche die rechten Seiten der linearen Gleichungssysteme auf den einzelnen Gittern darstellen. Auch hier fassen wir die Vektoren u_ℓ und f_ℓ zusammen, indem wir sie hintereinander in den Vektor u(SNP) und f(SNP) setzen.
Letztlich benötigen wir zur Lösung der linearen Gleichungssysteme noch einige Vektoren und Felder der Länge O(1) (z.B. zur vollen Abspeicherung der Steifigkeitsmatrix auf dem gröbsten Gitter) und mindestens einen Hilfsvektor der Länge $NP_{\ell_{max}}$ als Zwischeniterierte für das Glättungsverfahren.

(ii) Datenstruktur für die Triangulierungen $\tau_0, \ldots, \tau_{\ell_{max}}$.

X,Y seien Vektoren der Länge $NP_{\ell_{max}}$, die die Koordinaten der Gitterpunkte der Triangulierung $\tau_{\ell_{max}}$ bezeichnen.

Seien die Dreiecke T_ℓ einer Triangulierung τ_ℓ durchnumeriert von 1 bis NT_ℓ.

$$NT := \sum_{i=0}^{\ell_{max}} NT_i .$$

Um die Eckpunkte (EX_i, EY_i) $1 \le i \le 3$ für ein Dreieck zu speichern, benötigen wir einen Pointer "$itnode_\ell(3,NT_\ell)$", für den gilt:

$$itnode_\ell(i,j) = k \quad 1 \le i \le 3;\ 1 \le j \le NT_\ell\ ;\ 1 \le k \le NP_\ell$$

genau dann, wenn der i. te Eckpunkt des Dreiecks j der Gitterpunkt mit der Nummer k ist. Die Felder $itnode_\ell$ werden wie oben zusammengefaßt zu einem Feld itnode(3,NT).

Zusätzlich benötigen wir noch einen Vektor der Länge $NP_{\ell_{max}}$, der angibt, ob ein Gitterpunkt auf dem Rand des Gebietes liegt oder nicht und schließlich ein Feld der Dimension (3,NT), welches für jedes Dreieck die Nummer der Nachbardreiecke angibt. Näheres dazu im nächsten Abschnitt (zu 2).

(iii) Dimensionierung der Felder und Vektoren und Initialisierung der Steuerungsparameter.

Die Längen der Vektoren und Felder sind a priori bekannt und werden mit Hilfe der Formeln aus Satz (6.5.1) bzw. daraus abgeleiteter Formeln zuerst bestimmt, damit die einzelnen Felder u, X, Y, itnode etc. exakt dimensioniert werden können.

Zur Initialisierung der Steuerungsparameter siehe Kapitel 8.

" zu 2 "

Die Grobtriangulierung muß von Hand eingegeben werden nach dem Verfahren (3) aus Abschnitt 6.3. Die Verfeinerung und die korrekte Numerierung der Gitterpunkte (vgl. §6.3 und §6.4) erfolgt automatisch.
Zunächst erzeugt man aus dem Grobgitter τ_0 die feineren Gitter $\tau_1,\ldots,\tau_{\ell_{max}}$ durch sukkzesive regelmäßige Verfeinerung (vgl.§6.3), ohne auf die korrekte Numerierung der Gitterpunkte zu achten.

(i) Numerierung

Mit Hilfe der Kenntnis aller Nachbardreiecke jedes Dreiecks und der Information, ob ein Punkt Randpunkt ist oder nicht, können die Gitterpunkte von $\tau_{\ell_{max}}$ nun entlang der Knotenringe, wie in §6.4 beschrieben, numeriert werden. Danach wird der Pointer "itnode(3,NT)" entsprechend der neuen Numerierung modifiziert.
Falls wir auf jedem Gitter τ_ℓ $0 \le \ell \le \ell_{max}$ die Gitterpunkte gemäß §6.4 numerieren, erhalten wir auf jedem Gitter eine andere Numerierung der Knotenpunkte. Dieses Problem umgehen wir, indem wir eine Funktion "SORT" benutzen, die folgendem genügt:

> Sei (PX_i, PY_i) ein Gitterpunkt von $\tau_{\ell_{max}}$ der auch ein Gitterpunkt von τ_m ist für ein m ; $0 \le m < \ell_{max}$. Falls die Gitterpunkte von τ_m nach unserer Voschrift numeriert werden würden, hätte der Punkt (PX_i, PY_i) nicht die Nummer i sondern eine andere Nummer j. Es muß dann gelten:
>
> $$SORT(m,i) = j .$$

Seien $(PX_i, PY_i)\ \{1 \le i \le NP_{\ell_{max}}\}$ die gemäß unserer Numerierungsvorschrift sortierten Punkte von $\tau_{\ell_{max}}$ und I_m definiert durch:

$$I_m := \{\, i\,,\, i \in \mathbb{N}\,;\, (PX_i\,,\, PY_i\,) \in \tau_{\ell_{max}} \text{ und } (PX_i\,,\, PY_i\,) \in \tau_m\} \quad \text{für } 1 \le m < \ell_{max}.$$

Dann sind die Punkte:

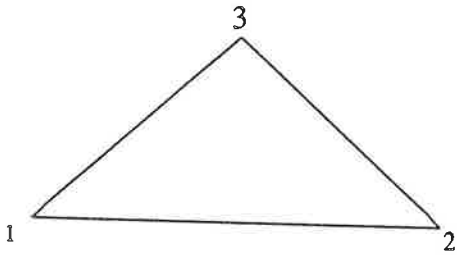
$$\{(PX_{SORT(m\,,j\,)}\,,\, PY_{SORT(m,j\,)}\,\}_{j \in I_m}$$

richtig sortiert bezüglich τ_m gemäß Vorschrift §6.4.

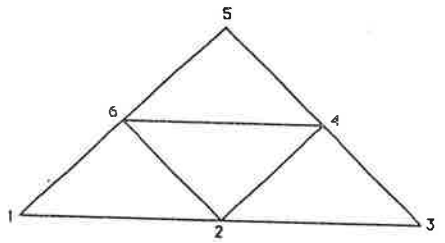
(7.1.1) Beispiel

Wir betrachten das Dreiecksgebiet:

Ω :



Als Grobgittertriangulierung wählen wir ganz Ω.
Einmal regelmäßig verfeinert ergäbe folgende Triangulierung τ_1 :



In diesem Fall würde die Funktion SORT wie folgt aussehen:

$$SORT(0,1) = 1$$

$$SORT(0,3) = 2$$

$$SORT(0,5) = 3\ .$$

(7.1.2) Beispiel

Programm zur Berechnung der Steifigkeitsmatrix für den Laplace-Operator mit Neumann-Randbedingung auf dem Gitter $\tau_m\ ; 0 < m < \ell_{max}$.
Sei $K_{i,j} = 0\ \forall\, i,j$; $b(i\,)$ diejenige Finite-Elemente Ansatzfunktion, die im Punkt mit der Nummer i den Wert eins hat und in allen allen anderen Knotenpunkten verschwindet. Sei NP definiert durch:

$$\underline{NP} := \sum_{i=0}^{m-1} NP_i \quad .$$

Algorithmus:

```
do 10 i = NT_{m-1} , NT_{m-1} + NT_m
  do 10 j_1 = 1 , 3
  k_1 = itnode(j_1,i)
  i_1 = SORT(m,k_1)
    do 10 j_2 = 1 , 3
    k_2 = itnode(j_2,i)
    i_2 = SORT(m,k_2)
    K(NP+ i_1 , NP + i_2 ) = K(NP+ i_1 , NP + i_2 ) + ∫_{T_i} (∇b(k_1 ) , ∇b(k_2 ))
10  continue
    return
    end
```

(ii) Randprojektion

Da der Rand des Bodensees (= Ω) in keinster Weise einem n-Eck (n „klein") ähnelt, können wir nicht erwarten, daß für die Dreiecke $T_{0,i}$ $(1 \le i \le NT_0)$ der Grobtriangulierung τ_0 gilt:

$$\Omega = \bigcup_{i=1}^{NT_0} T_{0,i} \quad .$$

Wir werden daher mit Gebieten Ω_i $(0 \le i \le \ell_{max})$ arbeiten, die Ω mit zunehmender Verfeinerung immer besser approximieren.

1.Schritt:
Ersetzte Ω durch ein Polygon Ω', welches Ω „hinreichend" gut approximiert und gebe die Eckpunkte von Ω' ein.
„hinreichend" heißt in diesem Fall, daß gilt:

$$| \Omega \setminus \Omega' \cup \Omega' \setminus \Omega | \le Ch_{\ell_{max}}{}^2 .$$

(Damit bleiben die Fehlerabschätzungen in unserem Fall von linearen Finiten Elementen erhalten (vgl. [19]).)

2.Schritt:
Gebe die Grobtriangulierung ein, und zwar so, daß die äußeren Randpunkte von τ_0 auf $\partial\Omega'$ liegen.

3.Schritt: Verfeinerung

Dreiecke, welche mindestens zwei Ecken besitzen, die auf $\partial\Omega'$ liegen, nennen wir Randdreiecke.

Um eine sukkzesive Anpassung des Gebietes an Ω' zu erreichen, werden Randdreiecke speziell verfeinert, die anderen alle regelmäßig.

Sei also T ein Dreieck mit Ecken E_i $(1 \le i \le 3)$ -gegen den Uhrzeigersinn fortlaufend gezählt- und E_1 , E_2 auf $\partial\Omega$. $R_i := 0.5 \cdot (E_i + E_{i \,(\mathrm{mod}\, 3)+1})$ $(1 \le i \le 3)$ bezeichnet die Seitenmitten von T. Wir ordnen nun R_1, wie unten näher beschrieben wird, einen Punkt $R_1' \in \partial\Omega'$ zu. Wir erhalten vier neue Dreiecke, indem wir die Punkte:

(i) R_1', R_2, R_3 ;

(ii) R_1', R_3, E_1 ;

(iii) R_1', E_2, R_2 ;

(iv) R_2, E_3, R_3

jeweils miteinander verbinden.

Üblicherweise erhält man R_1', indem man R_1 senkrecht zur Seite $\overline{E_1 E_2}$ auf das Stück von $\partial\Omega'$ projiziert, welches zwischen E_1 und E_2 liegt (Orthogonalprojektion).

Wir haben diese Verfahren modifiziert aus Gründen, die durch das folgende Beispiel illustriert werden.

(7.1.3) Beispiel

Sei in den folgenden Abbildungen "x" immer das zu verfeinernde Dreieck. Es wird immer die Orthogonalprojektion verwendet. Ω ist in den folgenden Abbildungen das schraffierte Gebiet. Der Übersichtlichkeit wegen wird nicht die ganze Triangulierung gezeichnet, sondern nur das Dreieck "x", bzw. dessen Nachfolger.

Probleme bei konkaven Ecken:

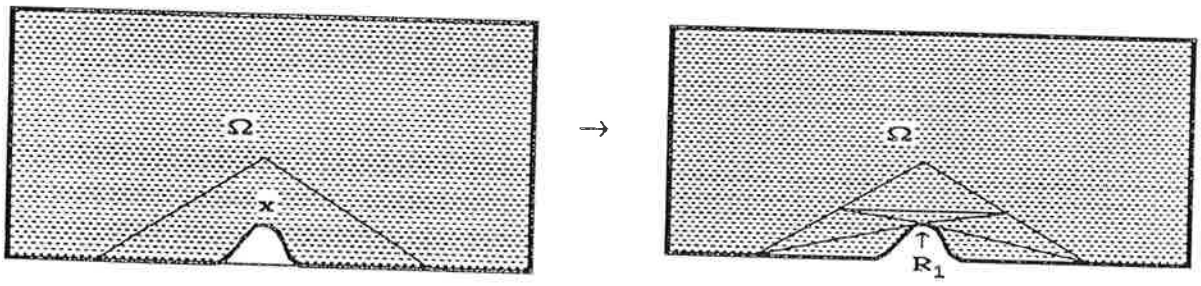


Abb.25 Bei der Verfeinerung des Dreiecks "x" entstehen Dreiecke mit sehr stumpfen Winkeln.

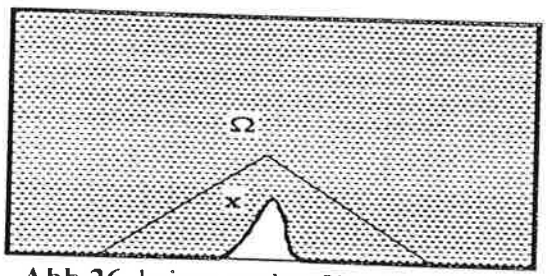


Abb.26 keine regelmäßige Verfeinerung des Dreiecks "x" möglich!

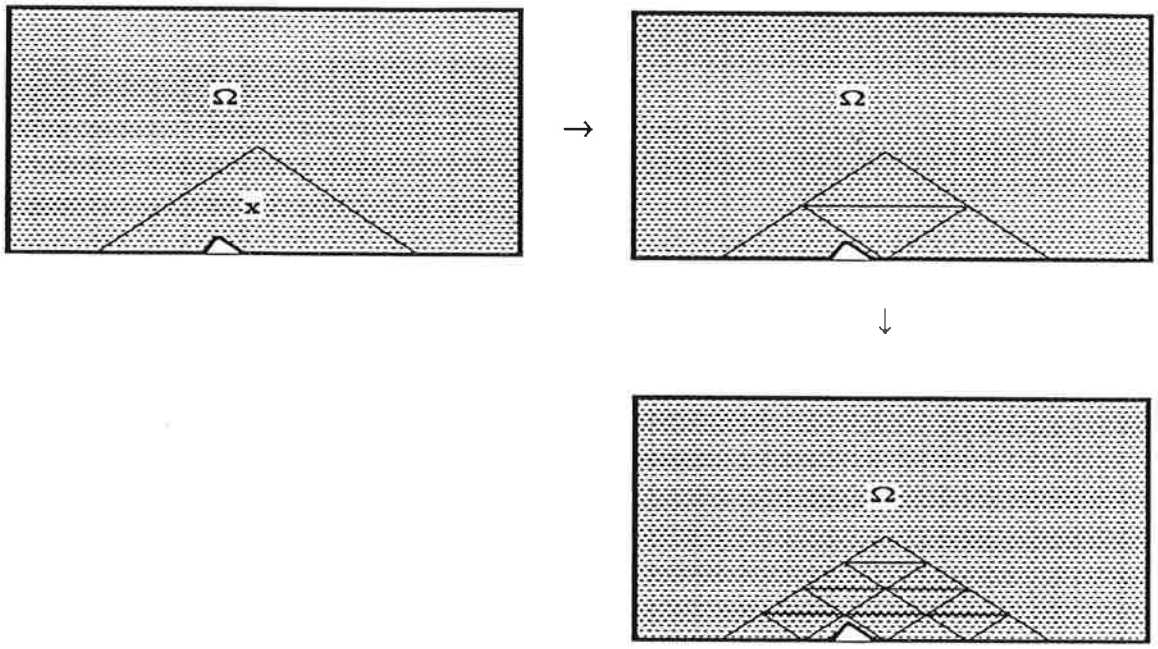


Abb.27 Die zweimalige Verfeinerung des Dreiecks "x" läßt sich nicht mehr weiter verfeinern, da man das Dreieck, welches über der einspringenden Ecke von Ω liegt, nicht mehr weiter unterteilen kann.

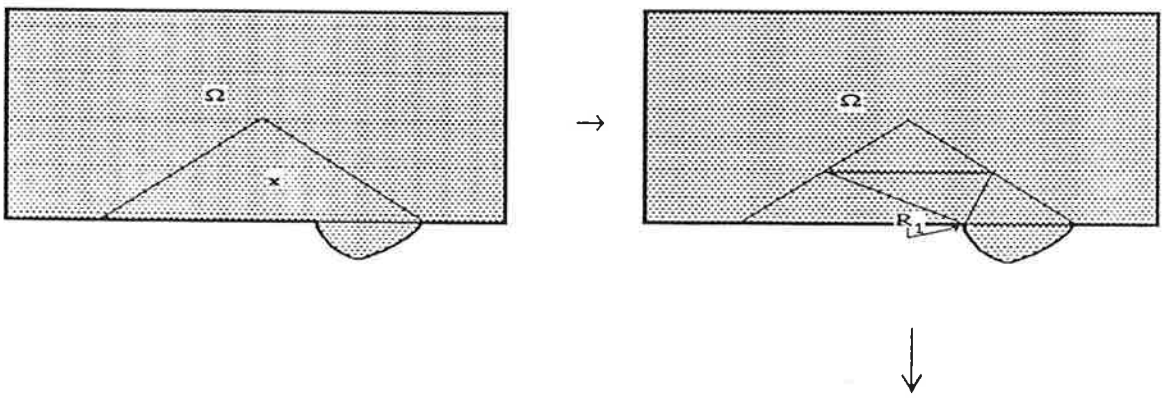
Grundidee des entwickelten Projezierverfahrens ist, einer zu projizierenden Seitenmitte R_1 eines Randdreiecks denjenigen Punkt R_1' auf $\partial\Omega'$ zuzuordnen, der am günstigsten ist, damit:

1.) die vier neu entstehenden Dreiecke weiter verfeinert werden können,

2.) ein möglichst kleiner Quotient entsteht aus größtem und kleinsten Innenwinkel der vier neuen Dreiecke.

Die erste Bedingung besagt, daß man versuchen sollte, Randdreiecke T_i so zu generieren, daß das (die) Gebiet(e) zwischen $\partial\Omega'$ und T_i konvex ist (sind).
Die zweite Bedingung besagt, daß man die erste Bedingung möglichst spät erfüllen sollte, damit die Innenwinkel der Dreiecke in der gleichen Größenordnung bleiben.
Dies veranschaulicht folgendes Beispiel:

(7.1.4) Beispiel
zur ersten Bedingung:
a)



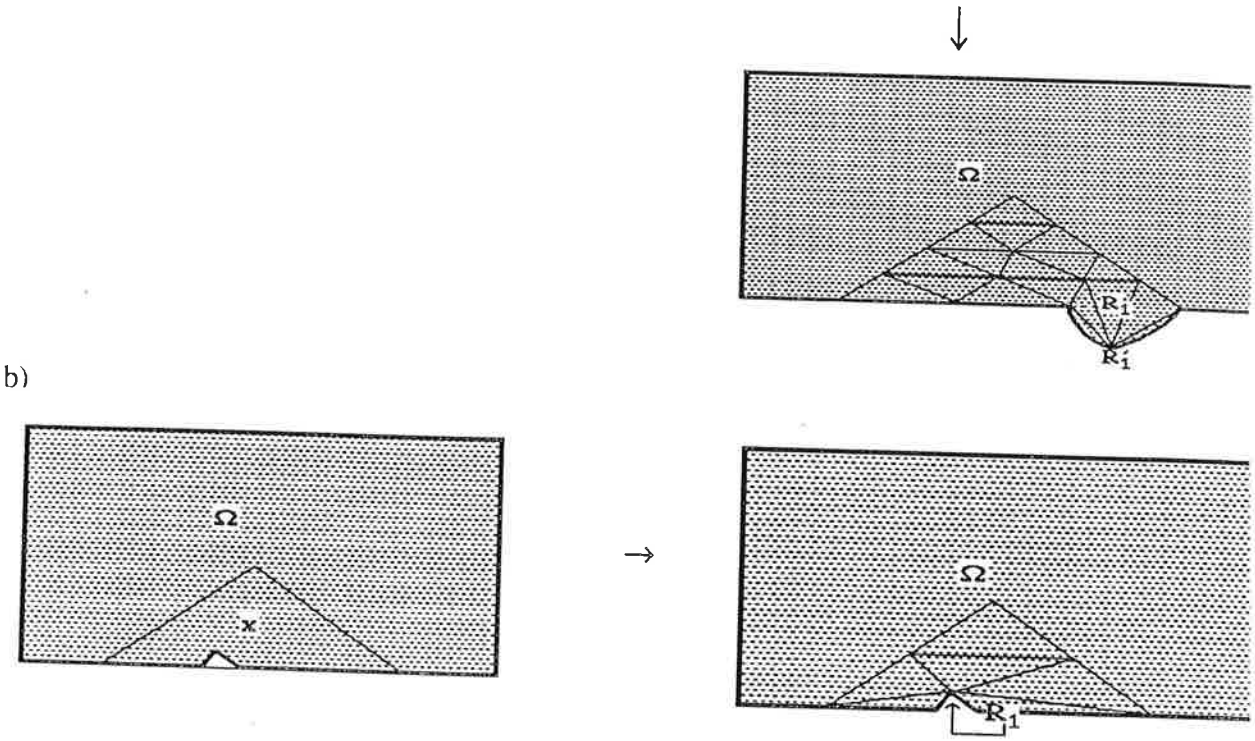
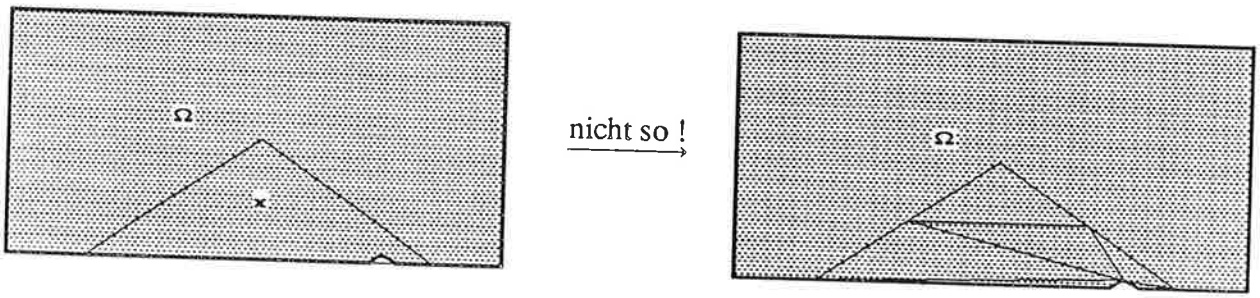


Abb.28 a,b Randprojektionen nach modifiziertem Verfahren



↓ sondern so!

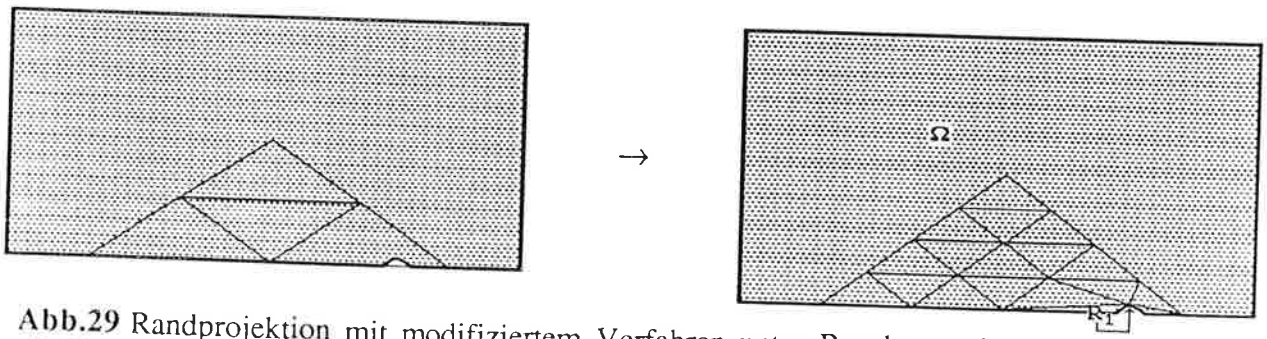


Abb.29 Randprojektion mit modifiziertem Verfahren unter Beachtung der Innenwinkel in den neuen Dreiecken.

" **zu 3** "

Die Steifigkeitsmatrix wird berechnet, wie im Beispiel (7.1.2) skizziert. Es ist aber zu beachten, daß nicht der Laplaceoperator diskretisiert werden soll, sondern der Operator: $\nabla (h \nabla \cdot)$; mit der Tiefenfunktion h des Sees. Da wir mit stückweise linearen Finiten-Elementen arbeiten, gilt $\nabla u = \text{const}$ für jede Funktion u aus dem Ansatzraum. Daraus folgt:

$$\int_T < \nabla b_i , h \nabla b_j > \; = \; < \nabla b_i , \nabla b_j > \int_T h \; ;$$

wobei b_i bzw. b_j die Finite-Elemente Basisfunktionen zum Knotenpunkt P_i bzw. P_j bezeichnen. Die Tiefe eines Sees wird in der Praxis punktweise vorliegen, deshalb werten wir das Integral über h mit Hilfe einer Quadraturformel aus. Dazu müssen wir einem beliebigem Punkt eine Tiefe zuordnen können. Das geschieht wie folgt:
Man denke sich über das Gebiet Ω ein kartesisches Gitter gelegt, dessen benachbarte Gitterpunkte einen Abstand voneinander haben sollen, der in der Größenordnung der kleinsten Entfernung zwischen den Knotenpunkten der feinsten Triangulierung von Ω liegt. Für jeden dieser Gitterpunkte gebe man dann die zugehörige Tiefe in lexikographischer Reihenfolge ein. Man benötigt nun lediglich eine Routine, welche jedem Punkt P von Ω den nächstgelegenen Gitterpunkt zuordnet, woraus man dann eine approximative Tiefe für P erhält. Falls man Approximationen höherer Ordnung erreichen will, interpoliert man die Tiefe zwischen den n nächstgelegenen Gitterpunkten von P. Im hier entwickelten Programm wird das Integral approximiert durch die folgende Formel: Seien R_i ($1 \le i \le 3$) die Seitenmitten des Dreiecks T und h_i die Tiefen der nächstgelegenen kartesischen Gitterpunkte zu R_i .

$$\int_T h(x,y)\mathrm{d}x\mathrm{d}y \approx \frac{|T|}{3} \{h_1 + h_2 + h_3\}$$

Da die Steifigkeitsmatrix diagonalenweise abgespeichert werden soll, muß für jedes Element $\{K_\ell\}_{i,j}$ durch eine recht aufwendige Fallunterscheidung- festgestellt werden in welcher Diagonalen es sich befindet.

Die ILU- Zerlegung der Steifigkeitsmatrix ist auf Grund der periodischen Ausnahmestellen im Besetzungsmuster von K_ℓ ungleich aufwendiger zu programmieren als bei regelmäßigen Bandmatrizen. Die Rechenzeit bleibt jedoch bei geschickter Programmierung und Abspeicherung der Ausnahmestellen nahezu die selbe.

" **zu 4** "

Es wurde für das Eigenwertproblem genau der Algorithmus (5.2.4) umgestzt.

Als Beispiel für eine singuläre Gleichung wurde das Problem:

(7.1.5)
$$-\Delta u = f \quad \text{in } \Omega$$
$$\partial u / \partial n = 0 \quad \text{auf } \Gamma$$

gerechnet.

Zur Normierung von u
Sei $c \in \mathbb{R}$ fest. Eine Normierung für u in (7.1.5) ist gegeben durch:

$$\int_\Omega u = c.$$

Üblicherweise wählt man $c = 0$.

Sei jetzt NP die Anzahl der Punkte auf dem Gitter τ und $u_h \in \mathbb{R}^{NP}$ eine Iterierte zur Lösung des LGS: $K x = f$.

$\tilde{u}$ sei definiert durch:

$$\sum_{i=1}^{NP} u_{h,i}\, b_i =: \tilde{u} \approx u.$$

Die Normierungsbedingung mit $c = 0$ würde dann lauten (beachte: $\ker K = \mathrm{span}\{v \in \mathbb{R}^{NP};\ v_i = 1;\ 1 \le i \le NP\}$):

$$\int_\Omega \tilde{u}^{norm} := \int_\Omega (\tilde{u} - \mathrm{const}) = 0 .$$

d.h.

$$0 = \int_\Omega (\tilde{u} - \mathrm{const}) = \int_\Omega \left(\sum_{i=1}^{NP} (u_{h,i} - \mathrm{const}) \cdot b_i \right) = -\mathrm{const} \cdot |\Omega| + \sum_{i=1}^{NP} u_{h,i} \int_\Omega b_i$$

$$= -\mathrm{const} \cdot |\Omega| + \frac{1}{3} \sum_{i=1}^{NP} u_{h,i} \cdot |\mathrm{Tr}\,(b_i)| .$$

$$\Leftrightarrow \qquad \mathrm{const} = \frac{\sum_{i=1}^{NP} u_{h,i} \cdot |\mathrm{Tr}(b_i)|}{3 \cdot |\Omega|} ;$$

wobei $\mathrm{Tr}(b_i)$ den Träger von b_i bezeichnet.

Das bedeutet weiter, daß für jede Iterierte u_h die Größe "const" wie oben beschrieben berechnet werden müßte; man hätte also insbesondere die Träger der Ansatzfunktionen b_i entweder zusätzlich abzuspeichern oder jedesmal neu zu berechnen.

Auf Grund des Zwischenwertsatzes kann man aber nun auch c so wählen, daß gilt:

$$\sum_{i=1}^{NP} u(x_i) = 0 ;$$

wobei $\{x_i\}_{1 \le i \le NP}$ die Knotenpunkte von τ bezeichnen.

Das würde bedeuten, daß für die Iterierte u_h gelten müßte :

$$\sum_{i=1}^{NP} u_{h,i} = 0 .$$

d.h.

$$0 = \sum_{i=1}^{NP} u_{h,i}^{norm} = \sum_{i=1}^{NP} (u_{h,i} - \mathrm{const}) = -NP \cdot \mathrm{const} + \sum_{i=1}^{NP} u_{h,i}$$

$$\Leftrightarrow \qquad \mathrm{const} = \frac{1}{NP} \sum_{i=1}^{NP} u_{h,i} .$$

Im Programm ist während des Iterationsverfahrens die zweite Normierungsvariante verwendet und anschließend wird, falls ein Abbruchkriterium erreicht ist, nach der ersten Möglichkeit ($c = 0$) normiert; damit besser mit eventuell bekannten Lösungen verglichen werden kann.

Zur Normierung der rechten Seite
Auf Grund des Satzes von Gauß gilt:

$$0 = \int_{\partial\Omega} \frac{\partial u}{\partial n} = \int_{\Omega} \Delta u = -\int_{\Omega} f \quad .$$

Daraus folgt, daß Gleichung (7.1.5) nur dann sinnvoll gestellt ist, falls gilt :

$$\int_{\Omega} f = 0.$$

Für kompliziertes Ω muß diese Integrabilitätsbedingung vom Rechner realisiert werden. Dies geschiet in den nun beschriebenen drei Schritten.

Sei f gegeben mit $\int_{\Omega} f \neq 0$. f_i $(1 \leq i \leq NP)$ bezeichnet die Werte von f in den Knotenpunkten und $f_h \in \mathbb{R}^{NP}$ sei die rechte Seite des LGS, welches auf Grund einer finite Elemente Diskretisierung der Gleichung (7.1.5) entsteht, wie sie in §4 beschrieben wird.

1. Schritt :

Ersetze f durch $\sum_{i=1}^{NP} f_i \, b_i$.

2. Schritt :

Bestimme c so, daß gilt :

$$\sum_{i=1}^{NP} \left(\int_{\Omega} f_i^{norm} \cdot b_i \right) := \sum_{i=1}^{NP} \left(\int_{\Omega} (f_i - c) \cdot b_i \right) = 0 \quad .$$

d.h.

$$c = \frac{\sum_{i=1}^{NP} f_i \, |Tr(b_i)|}{3 \cdot |\Omega|} \quad .$$

3.Schritt :

Berechne $f_{h,j}$ $(1 \leq j \leq NP)$ durch :

$$f_{h,j} = \sum_{i=1}^{NP} \int_{\Omega} (f_i - c) b_i b_j = \sum_{i=1}^{NP} f_i^{norm} \int_{\Omega} b_i \, b_j = \sum_{j=1}^{NP} M_{i,j} \, f_j^{norm} \quad ;$$

wobei M die (symmetrische) Massenmatrix bezeichnet.

" **zu 5** "

siehe Kapitel 8.

§ 8 Numerische Ergebnisse

Wir betrachten in diesem Abschnitt das Problem (vgl.3.10):

$$- \nabla\cdot(h\,\nabla u) - \lambda\, u = g \qquad \text{in } \Omega,$$
$$\partial u/\partial n = 0 \qquad \text{auf } \Gamma := \partial\Omega\,.$$

Als Beispiel einer singulären Gleichung haben wir den Fall: $\lambda = 0$, $g \neq 0$, bzw. als Beispiel einer Eigenwertgleichung den Fall $g \equiv 0$ gerechnet. Es wurden die Algorithmen an verschiedenen Gebieten Ω und mit unterschiedlichen Tiefenfunktionen h getestet. Beim Mehrgitterverfahren wurde immer ein V-Zyklus verwendet. Wir haben immer mit zwei Vor- und zwei Nachglättungsschritten gerechnet, was sich am effektivsten erwies. Zunächst zur singulären Gleichung.

§ 8.1 Numerische Ergebnisse für eine singuläre Gleichung

Sei $\Omega := \{x \in \mathbb{R}^2 ; x_1^2 + x_2^2 < 1\}$. Wir betrachten die Gleichung:

(8.1.1)
$$- \nabla\cdot(h\,\nabla u) = g \qquad \text{in } \Omega,$$
$$\partial u/\partial n = 0 \qquad \text{auf } \Gamma := \partial\Omega;$$

Wir setzen zunächst $g := -2\cdot(2\cdot r^2 - 1)$, wobei $r := \sqrt{x_1^2 + x_2^2}$ und $h \equiv 1$. Die exakte Lösung lautet dann : $u = 0.25\cdot r^4 - 0.5\cdot r^2 + c$.

Mit der Normierungsbedingung $\int_\Omega u = 0$ erhält man $c = 1/6$.

(8.1.2) Diskretisierung

Wir verwenden ein Grobgitter, welches sich mit Hilfe des in § 7 beschriebenen Projizierverfahrens mit zunehmender Verfeinerung immer besser dem Kreisgebiet Ω anpaßt.

Gitterdaten:

Level	# Gitterpunkte	# Dreiecke
0	24	34
1	81	136
2	297	544
3	1137	2176
4	4449	8704
5	17601	34816

Tab. 1 Gitterdaten für Kreistriangulierung, Level bezeichnet die Verfeinerungsstufe.

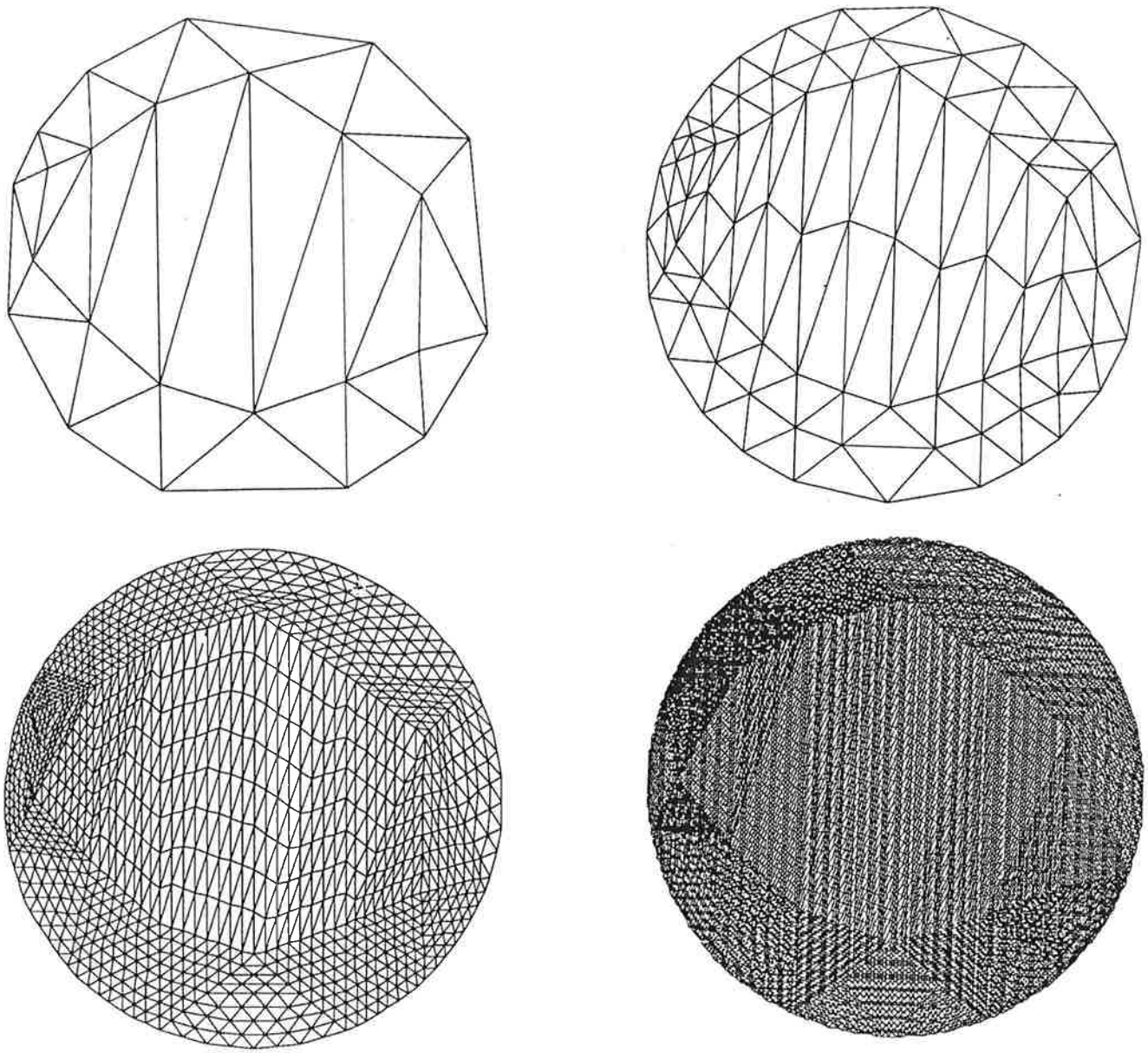


Abb. 30 Triangulierungen für das Kreisgebiet auf den Verfeinerungsstufen 0, 1, 3, 5.

Wir erhalten auf Grund unserer Diskretisierung ein lineares Gleichungssystem der Form:

$$K\,x = f,$$

welches wir mit einem ILU-Verfahren, bzw. mit einem Mehrgitterverfahren lösen. Bezeichne u die exakte Lösung von (8.1.1) und u_i die exakte Lösung in den Gitterpunkten. x^m sei die m.te Iterierte in unserem Iterationsverfahren. $D := \{D_{i,i}\}$ bezeichne die Diagonalmatrix der ILU-Zerlegung (vgl. 6.2.1'). In den folgenden Tabellen ist die Konvergenzrate κ, der maximale Approximationsfehler e_∞, der gemittelte Approximationsfehler e_2 und die Stabilitätskonstante c definiert durch:

$$\kappa := \|(Kx^m - g)\| \,/\, \|(Kx^{m-1} - g)\|,$$

$$e_\infty := \max\{\, |u_i - x_i^m| \,;\, 1 \le i \le NP\},$$

$$e_2 := \frac{1}{NP} \sum_{i=1}^{NP} | u_i - x_i^m |,$$

$$c := \min\{D_{i,i} / K_{i,i} ; 1 \le i \le NP\}.$$

Die folgende Tabelle zeigt die erwartete quadratische Konvergenz des Approximationsfehlers und illustriert, daß die Konvergenzraten κ bei zunehmender Verfeinerung von eins weg beschränkt bleiben. Dies ist in Abb. 31 graphisch dargestellt

Level	κ	# Iterationen	e_2	e_∞
2	$4 \cdot 10^{-2}$	10	$2.728 \cdot 10^{-3}$	$6.1 \cdot 10^{-3}$
3	$7.2 \cdot 10^{-2}$	12	$7.003 \cdot 10^{-4}$	$1.6 \cdot 10^{-3}$
4	$1.365 \cdot 10^{-1}$	16	$1.89 \cdot 10^{-4}$	$4.68 \cdot 10^{-4}$
5	$2.187 \cdot 10^{-1}$	20	$6.814 \cdot 10^{-5}$	$1.59 \cdot 10^{-4}$

Tab.2 Konvergenzverhalten des Mehrgitterverfahrens und Approximationsverhalten der finite Elemente Diskretisierung.

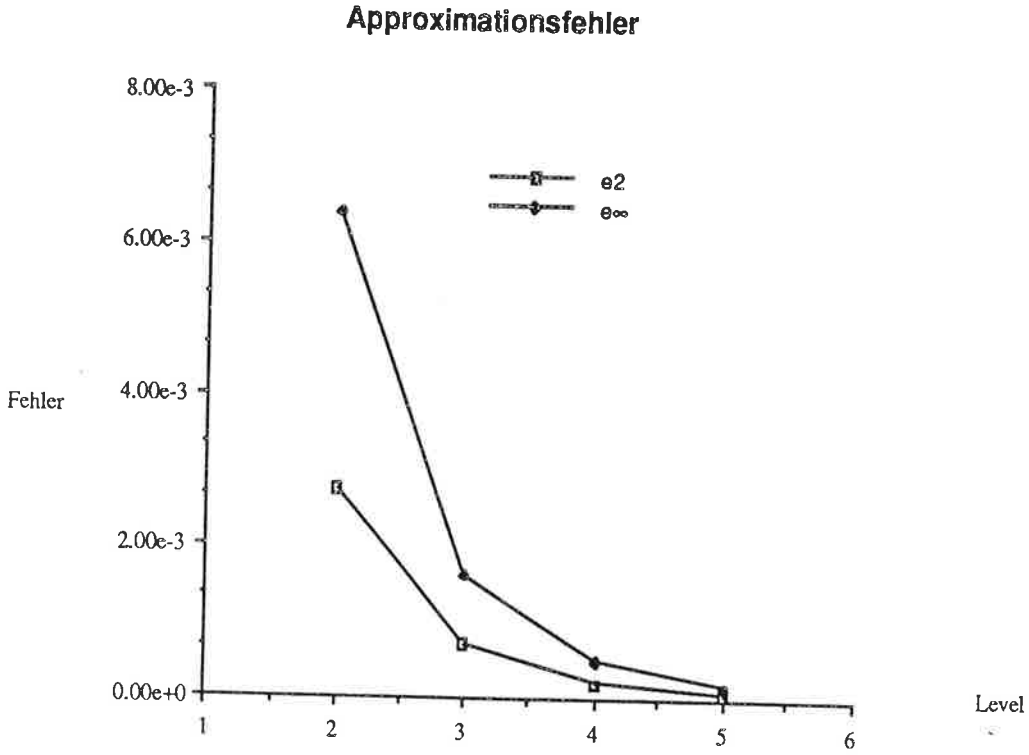


Abb.31 absoluter und gemittelter Fehler zwischen exakter und berechneter Lösung

Falls man das ILU-Verfahren ohne Mehrgitterkorrekturen verwendet, bemerkt man, daß die Konvergenz mit der Anzahl der Iterationen langsamer wird, und sie sich auch bei feineren Gittern zunehmend verschlechtert. Dies illustriert die folgende Tabelle.

# Iterationen	Level = 2	Level = 3	Level = 4
1	0.100	0.080	0.076
2	0.170	0.200	0.168
3	0.420	0.510	0.440
4	0.620	0.680	0.640
5	0.790	0.780	0.760
6	0.890	0.800	0.840
7	0.900	0.820	0.890
8	0.910	0.830	0.920
50	0.920	0.980	0.990

Tab. 3 Konvergenzraten des ILU-Verfahrens

Die berechnete Lösung besitzt die Form:

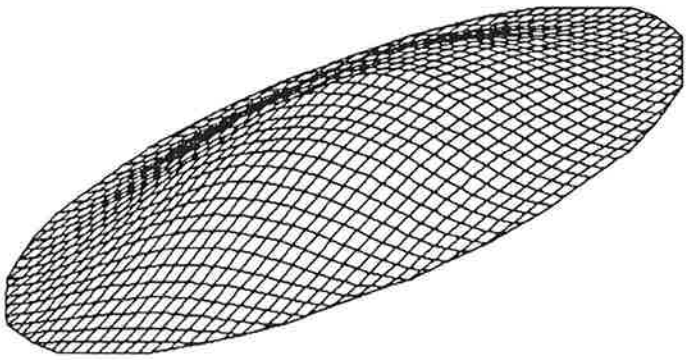


Abb.32 3-D Plot der berechneten Approximation der exakten Lösung: $u = 0.25 \cdot r^4 - 0.5 \cdot r^2 + 1/6$.

Im nächsten Schritt wollen wir den Einfluß verschiedener Tiefenfunktionen auf das Konvergenzverhalten des Mehrgitterverfahrens untersuchen.

(8.1.3) Konvergenzverhalten in Abhängigkeit der Tiefe

Wir werden sehen, daß die Tiefe bei der singulären Gleichung keinen spürbaren Einfluß auf die Konvergenz des Mehrgitterverfahrens besitzt. Wir haben das Konvergenzverhalten an den folgenden Tiefenfunktion h_1, $h_2^{\varepsilon,\delta}$:

$$h_1(x,y) := 2 - r^2, \quad h_2^{\varepsilon,\delta}(x,y) := \begin{cases} 1 + \frac{1}{(x-\varepsilon)^2 + y^2} & \text{für } (x-\varepsilon)^2 + y^2 \geq \delta \\ 1 + \frac{1}{\delta} & \text{für } (x-\varepsilon)^2 + y^2 < \delta \end{cases}$$

getestet. Die folgende Abbildung zeigt diese Tiefen.

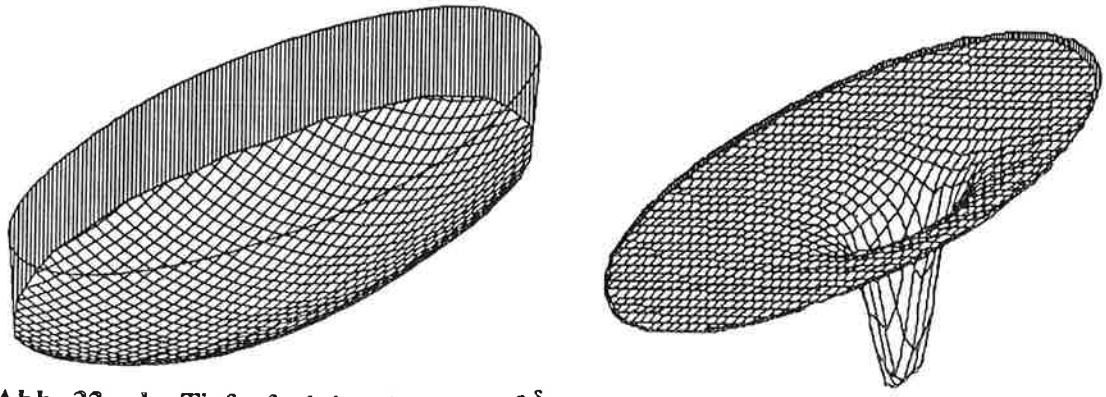
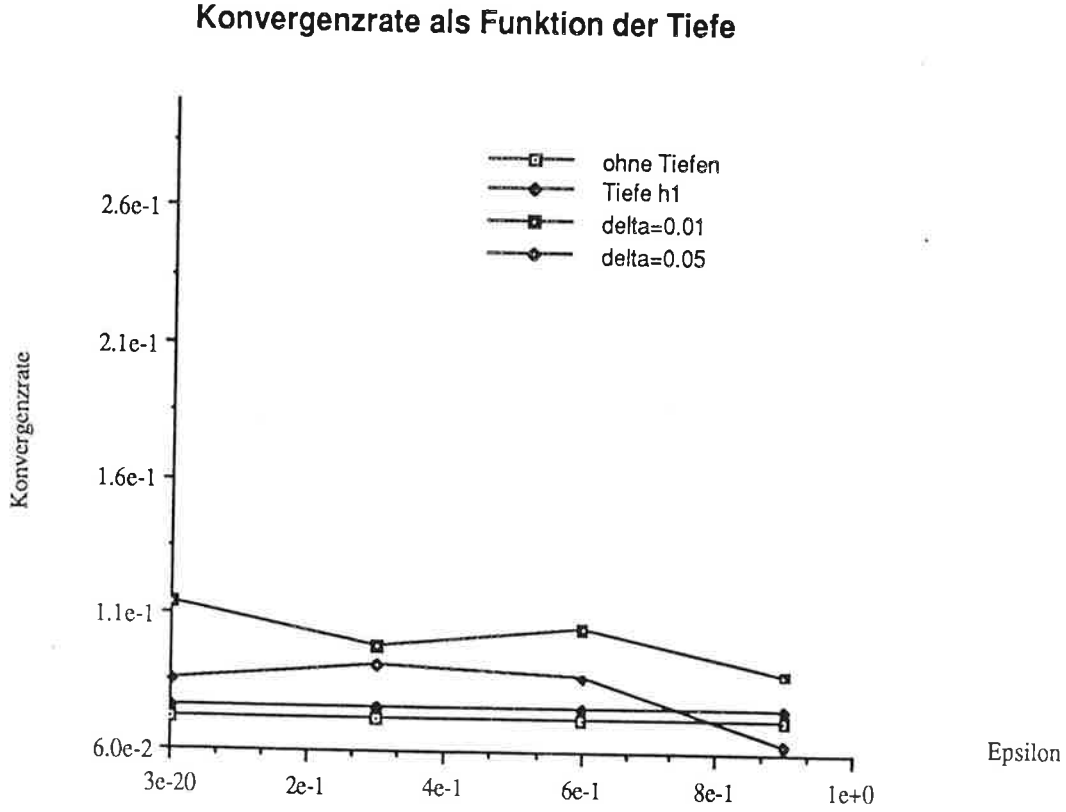


Abb. 32 a,b Tiefenfunktion h_1 und $h_2^{\varepsilon,\delta}$ mit $\varepsilon = 0.3$ und $\delta = 0.01$.

Die nächsten Graphik verdeutlicht, daß die Konvergenzraten κ nicht wesentlich von der Tiefe des Gebietes abhängen.



(8.1.4) Konvergenzverhalten in Abhängigkeit des Gebiets

Es zeigt sich, daß die Konvergenzgeschwindigkeit des Mehrgitterverfahrens sich deutlich verschlechtert, falls (z.B. durch Projektionen an den Rand) die Winkel in den Dreiecken des Gitters stumpf werden. Wir betrachten das folgende Testgebiet, welches durch den Parameter ε gestört wird.

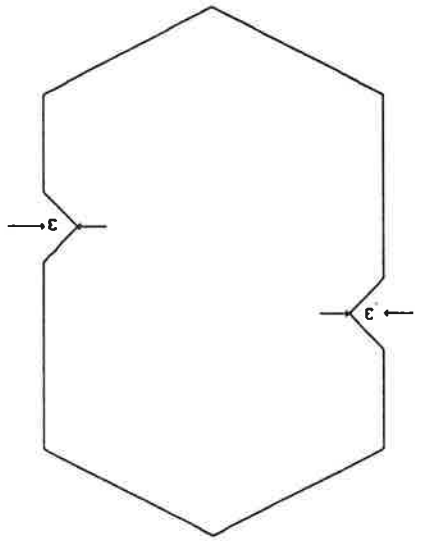


Abb. 34 Testgebiet mit Störungsparameter

Als Grobtriangulierung verwenden wir die Triangulierung aus Abb 35 a:

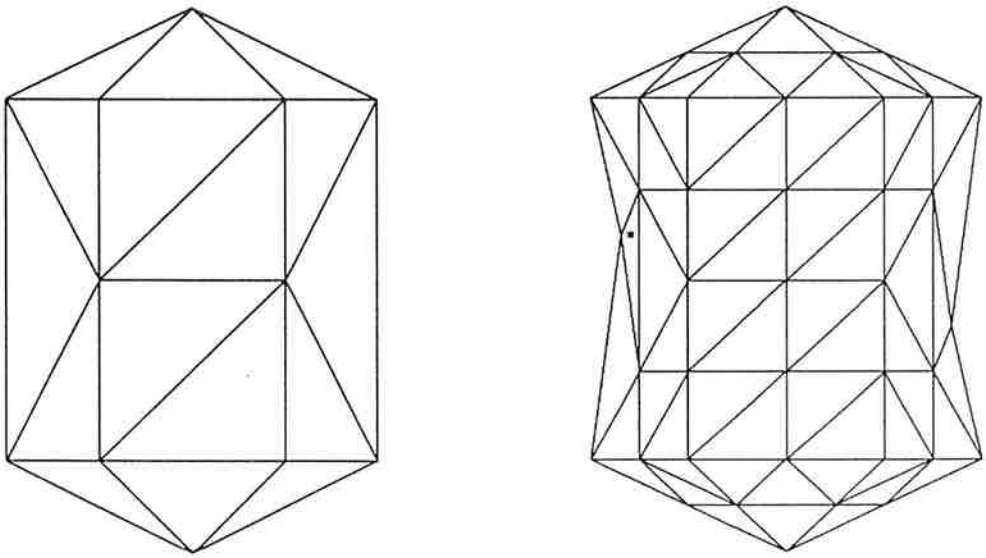


Abb.35 a,b Grobtriangulierung und erste Verfeinerung des Gebietes aus Abb.34 mit ε =0.7

In der folgenden Tabelle sind die Konvergenzraten des Mehrgitterverfahrens in Abhängigkeit von ε aufgetragen. ε=0 bedeutet, daß das Gebiet keine einspringende Ecke besitzt, ε=1 bedeutet, daß das Dreieck * in Abbildung 35 b zu einer Linie entartet. $\varepsilon < 0$ bedeutet daß die „einspringende" Ecke in Abb. 34 um $|\varepsilon|$ nach außen zeigt. Es wurde mit vier Gittern gerechnet, daß gröbste besitzt 12 Ecken und 16 Dreiecke, das feinste 537 Punkte und 1024 Dreiecke. Die Abbruchtoleranz für das Mehrgitterverfahrens war 10^{-12}.

Epsilon	alpha	beta	alpha/beta	c	kappa	# Iterationen
-15	1.516	135	89.050	0.257	0.055	11
-2	11.430	135	11.811	0.471	0.050	11
0	18.430	135	7.325	0.471	0.057	11
0.6	8.714	147	16.821	0.412	0.076	12
0.7	7.000	153	21.814	0.406	0.074	11
0.8	5.332	159	29.825	0.352	0.120	14
0.9	3.600	165	45.833	0.267	0.920	*größer* 150
0.95	1.910	172	90.052	0.186	4.841	*divergent*

Tab.4 Stabilitätskonstante c, Konvergenzrate κ, minimaler (maximaler) Innenwinkel der Dreiecke α (β) und Anzahl der Iterationen in Abhängigkeit des Parameters ε.

Tabelle 4 verdeutlicht, daß die Konvergenzraten κ des Mehrgitterverfahrens schlecht werden, falls die Triangulierung sehr stumpfe Winkel besitzt. Das liegt einerseits daran, daß mit zunehmend stumpfen Winkeln die Stabilitätskonstante c der ILU-Zerlegung klein wird, und andererseits das Mehrgitterverfahren dann keine gute Approximationseigenschaft besitzt. Auffallend ist, daß kein funktionaler Zusammenhang zwischen den kleinsten Winkeln der Triangulierung und der Konvergenzrate erkennbar ist, obwohl man üblicherweise annimmt, daß die Approximationseigenschaft vom Quotient aus größtem und kleinsten Winkel einer Triangulierung abhängt (beachte die Quotienten für negatives ε!). Die Tatsache, daß κ im wesentlichen nur vom größten Winkel abhängt, wird in [17] untersucht. Auffallend ist desweiteren, daß sich die Konvergenz in Abhängigkeit von ε nahezu sprunghaft verschlechtert ($\varepsilon_{Sprung} \approx 0.8$), und für kleinere ε nahezu konstant ist. Die Abbildungen 36 bis 38 illustrieren dies graphisch.

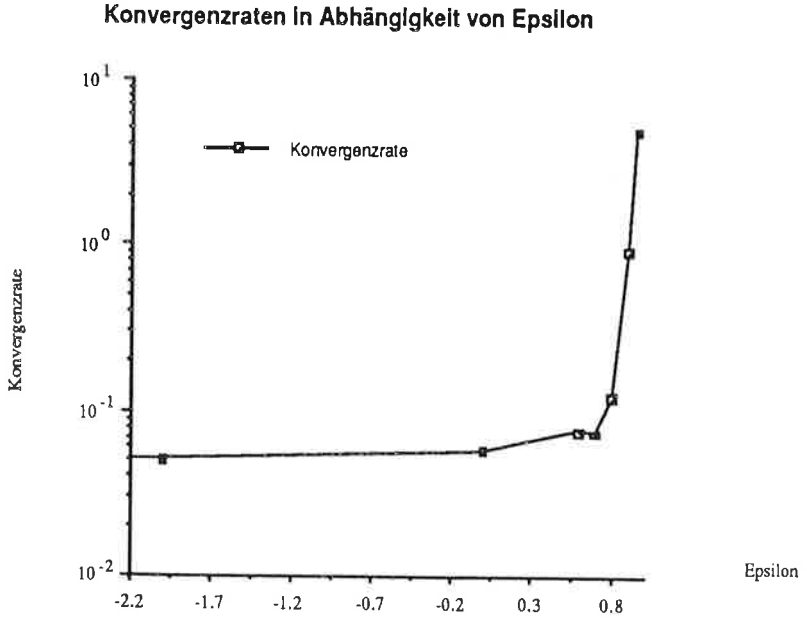


Abb.36 Konvergenzraten in Abhängigkeit vom Störungsparameter ε

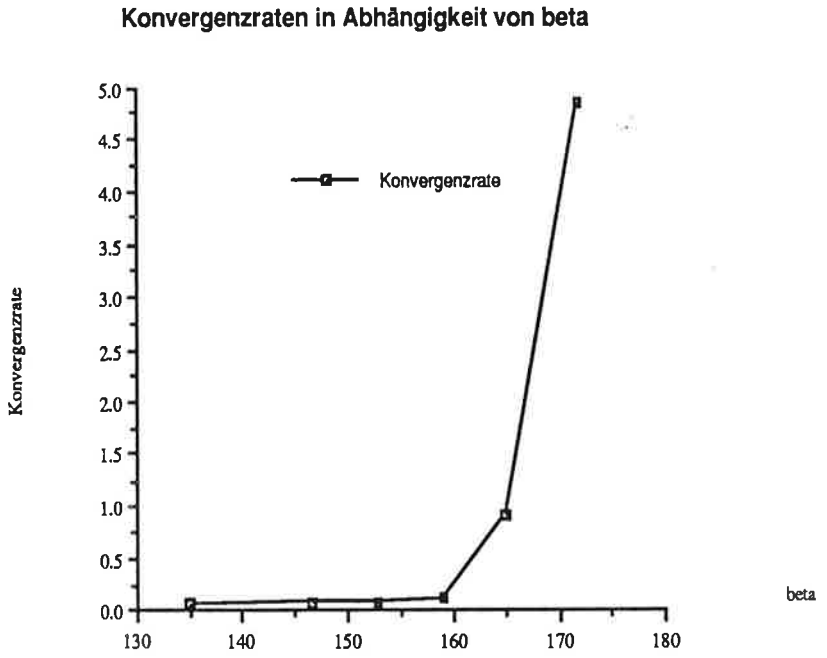


Abb.37 Konvergenzraten in Abhängigkeit vom maximalen Winkel der Triangulierung

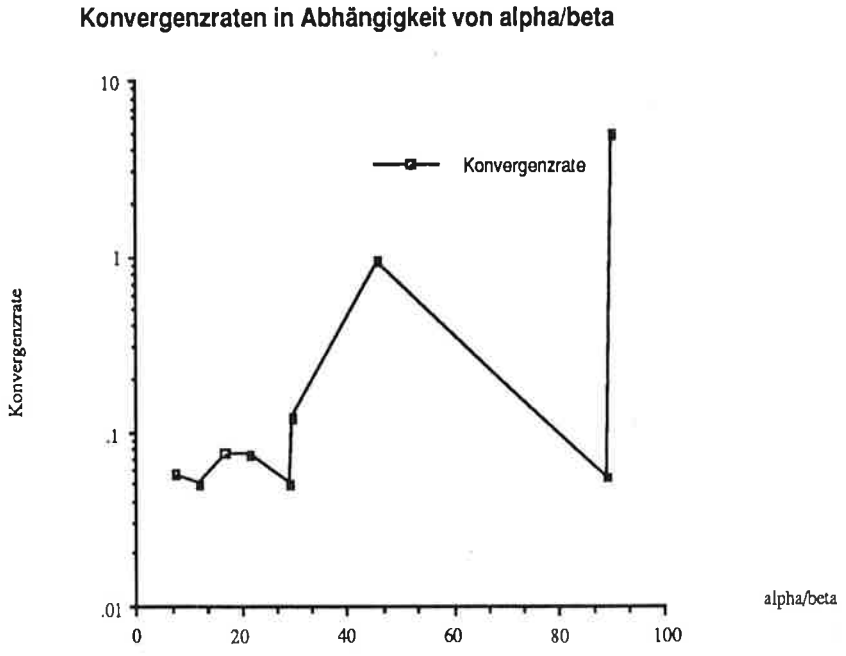


Abb.38 Konvergenzraten in Abhängigkeit vom Quotient aus größtem und kleinstem Winkel einer Triangulierung. Es ist kein einfacher funktionaler Zusammenhang erkennbar.

Abschließend wollen wir die singuläre Gleichung im Fall, daß Ω der Bodensee ist, betrachten.

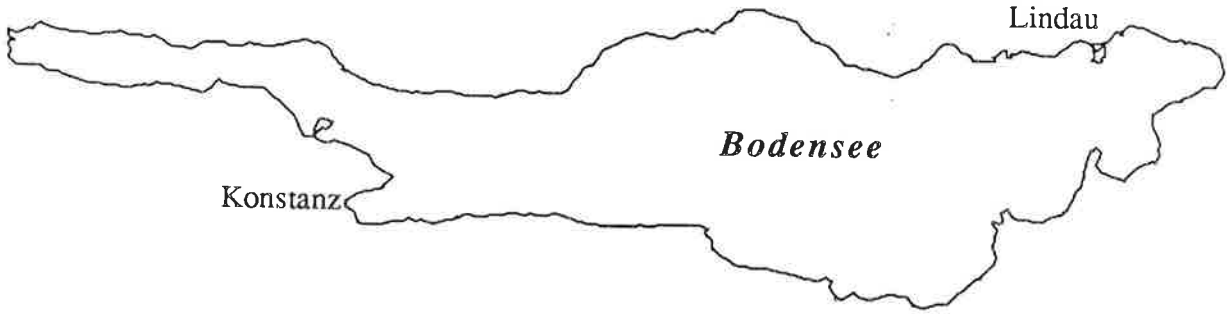
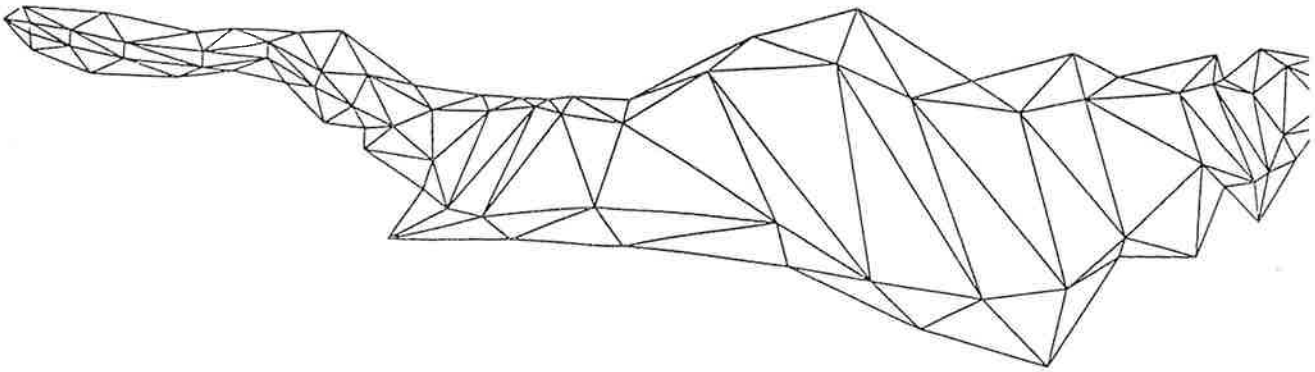
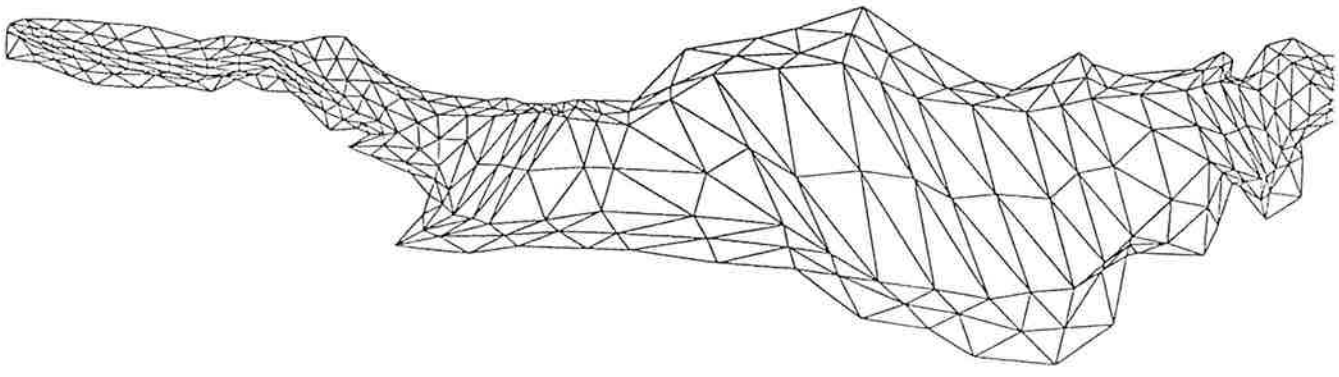


Abb.39 Bodenseegebiet

Da der Übergang des Bodensees in den Untersee bei Konstanz sehr schmal ist, erwarten wir keine Kopplung der Eigenschwingungen in diesem Bereich. In unseren Berechnungen wurde deshalb der Untersee immer weggelassen.

zur Diskretisierung:
Die folgenden Abbildungen zeigen die verschiedenen Gitter. Man beachte, wie sich das Gitter mit zunehmender Verfeinerung immer besser dem Rand aus Abb. 39 anpaßt. Das hat jedoch zur Folge, das extrem stumpfe bzw. spitze Winkel auftreten, was sich dann auch bei den Konvergenzraten bemerkbar macht.





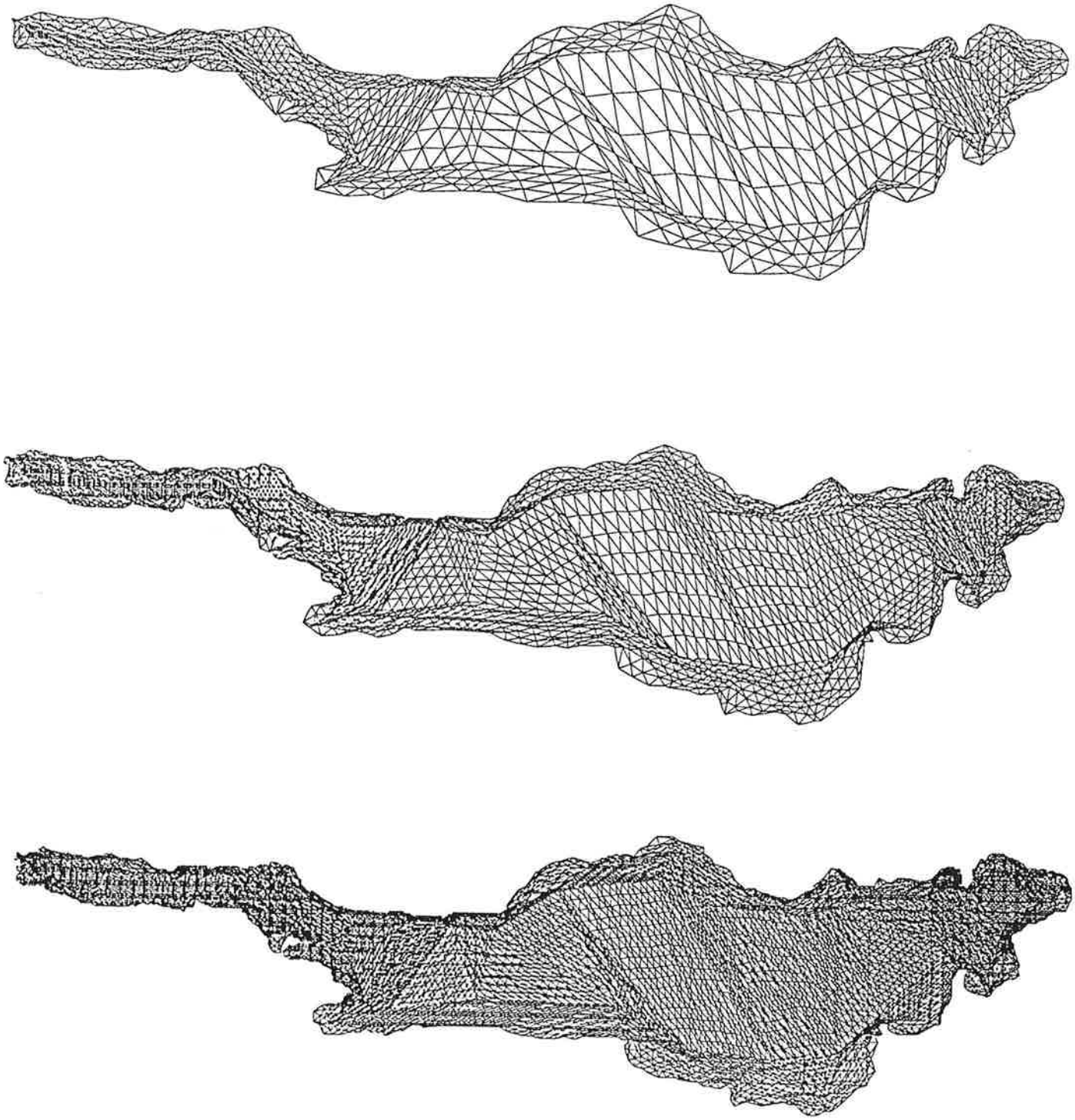


Abb. 40 Triangulierungen des Bodensees auf den Verfeinerungsstufen 0, 1, 2, 3, 4

Gitterdaten:

Level	# Gitterpunkte	# Dreiecke	α_{min}	α_{max}
0	88	130	7.56	162.04
1	305	520	1.7	173.04
2	1129	2080	0.7957	173.04
3	4337	8320	0.2565	178.78
4	16993	33280	0.2565	178.78

Tab.5 Gitterdaten für die Bodenseetriangulierungen, Level bezeichnet den Verfeinerungsgrad, α_{min} den minimalen und α_{max} den maximalen Winkel der Dreiecke in der Triangulierung

Zunächst haben wir mit konstanter Tiefe gerechnet, d.h. $h \equiv 1$ in Gleichung (8.1.1). Als rechte Seite haben wir $g := \sin(x/5)\cdot\sin(y/3) - C$ gewählt, wobei C durch das Verfahren aus § 7 Abschnitt 4 vom Programm berechnet wurde so, daß $\int_\Omega g = 0$ erfüllt ist. Auf dem Level 2 lag die Konvergenzrate bei 0.125, nach 15 Iterationen war der Rundungsfehler (10^{-12}) erreicht. Die berechnete Lösung ist in den folgenden beiden Plots dargestellt.

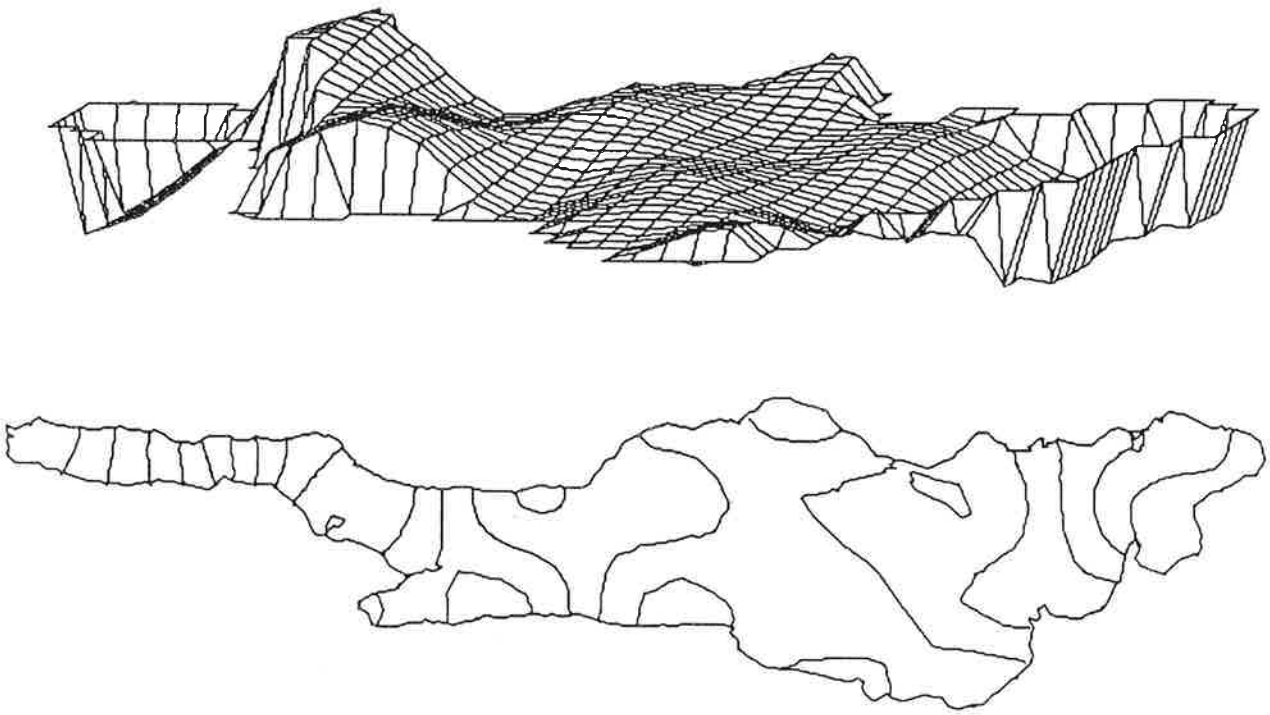


Abb.41 Höhenlinien- und 3-D Plot der berechneten Lösung der singulären Gleichung (8.1.1) auf dem Bodensee

Rechnet man nun die Gleichung mit der variablen Tiefe des Bodensees, erhält man auf der zweiten Verfeinerungsstufe nach wie vor Konvergenz, aber mit deutlich schlechterer Konvergenzrate (≈ 0.671). Diesmal ist der Rundungsfehler nach 72 Iterationen erreicht. Die folgende Abbildung zeigt einen Höhenlinien bzw. einen 3-D Plot der verwendeten Tiefe.

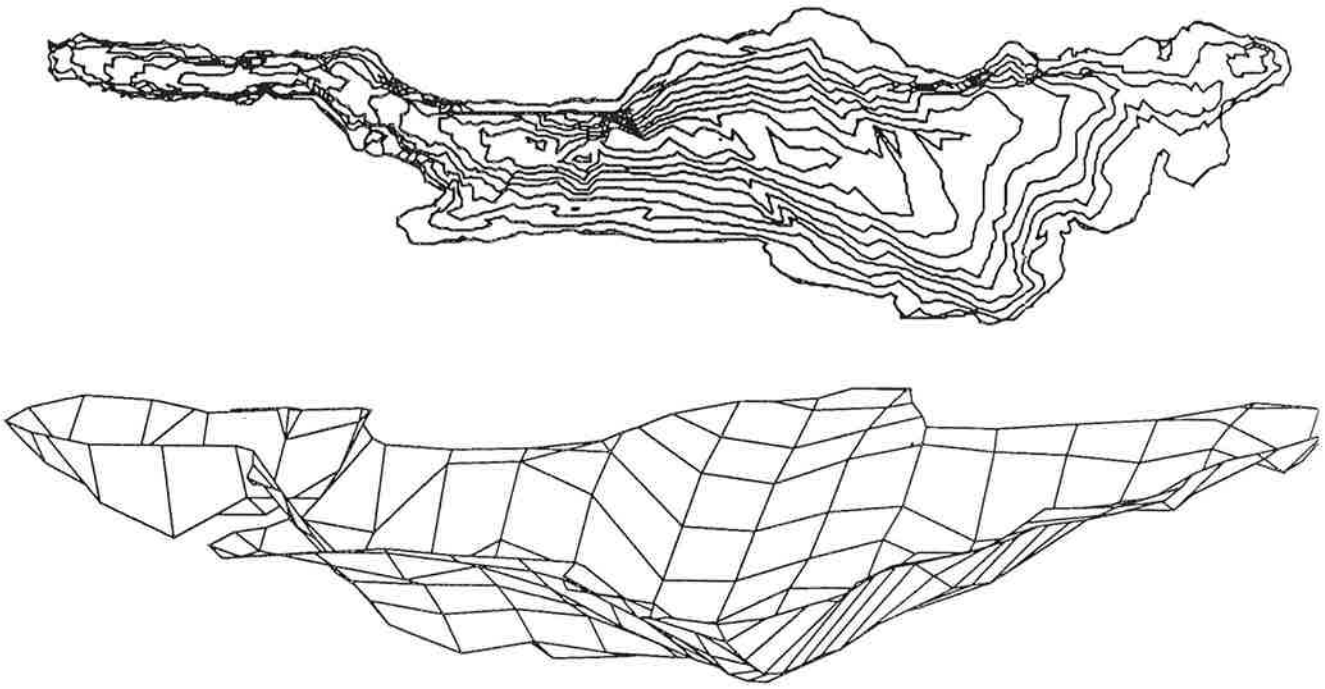


Abb. 42 Tiefe des Bodensees als Höhenlinien- und 3-D Plot

Verwendet man das nächst feinere Gitter (# Gitterpunkte = 4337) so divergiert das Verfahren sowohl mit konstanter als auch mit variabler Tiefe. Dies liegt offensichtlich am verwendeten Gitter (man beachte die Winkel aus Tabelle 5 und vergleiche mit § 8.1.4!). Betrachtet man die Entwicklung des Defektes d (d := Kx - f), so stellt man fest, daß d genau an den Stellen explodiert, wo das Gitter entartet ist. Das Gitter ist besonders im Bereich des Übergangs des Bodensees in den Obersee stark verzerrt. Bei variabler Tiefe konzentriert sich d auf den Bereich um die Insel Mainau, im Fall konstanter Tiefe auf das gegenüberliegende Ufer. Dies wird durch die folgenden Abbildungen illustriert.

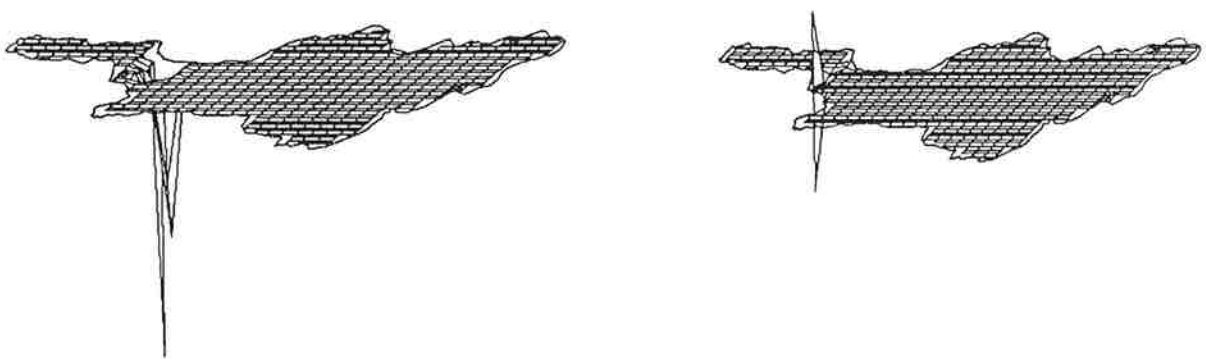


Abb. 43 a,b Defekt nach fünf Mehrgitteriterationen a) mit variabler Tiefe b) mit konstanter Tiefe

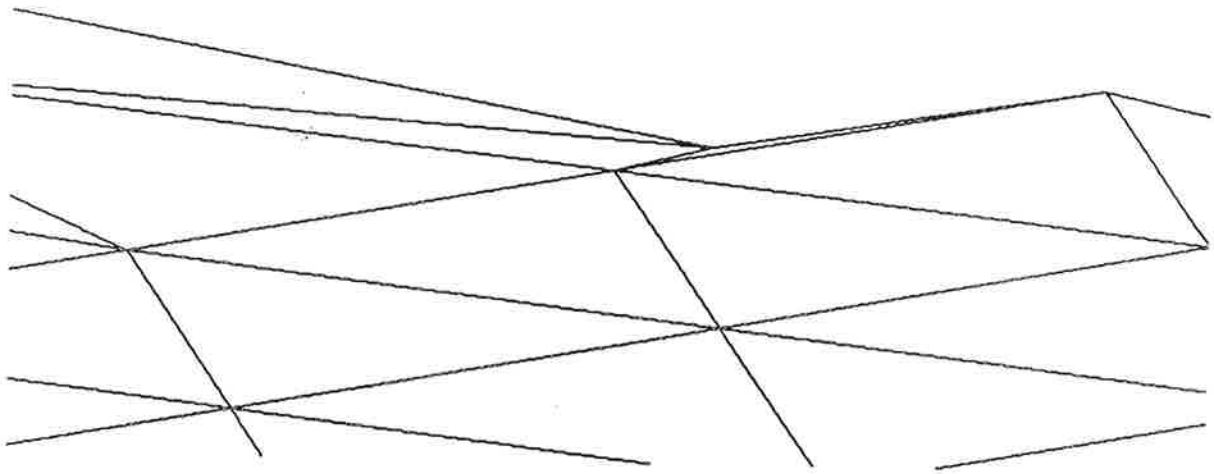


Abb. 44 Triangulierung im Bereich aus Abb. 43 b, in dem der Defekt divergiert.

§ 8.2 Numerische Ergebnisse für das Eigenwertproblem

Sei $\Omega := (0, 11) \times (0, 3)$. Wir betrachten die Gleichung:

(8.2.1)
$$- \nabla\cdot(h\,\nabla u) + \lambda\, u = 0 \quad \text{in } \Omega$$
$$\partial u/\partial n = 0 \quad \text{auf } \Gamma.$$

Wir setzen in diesem Beispiel $h \equiv 1$. Mit der Normierungsbedingung $\|u\|_0 = 1$ lauten die Eigenfunktionen bzw. Eigenwerte wie folgt:

Eigenfunktionen : Sei $k_1, k_2 \in \mathbb{N}$,

$$u(x,y) = \sqrt{\frac{C}{33}} \cdot \cos\frac{k_1 \cdot \pi \cdot x}{11} \cdot \cos\frac{k_2 \cdot \pi \cdot y}{3} \;;$$

wobei :

$$C := \begin{cases} 1 & \text{falls } k_1 = k_2 = 0 \\ 2 & \text{falls entweder } k_1 = 0 \text{ oder } k_2 = 0 \\ 4 & \text{falls } k_1 \cdot k_2 \neq 0 \end{cases};$$

zugehörige Eigenwerte :

$$\lambda(k_1, k_2) = \pi^2 \cdot \left\{ \left(\frac{k_1}{11}\right)^2 + \left(\frac{k_2}{3}\right)^2 \right\}.$$

(8.2.2) Diskretisierung und numerische Ergebnisse

Wir haben als Beispiel für das Verhalten der Konvergenzraten und des Approximationsfehlers bei einem Eigenwertproblem die zweite Oberschwingung in obigem Problem berechnet.

Gitterdaten:

Level	# Gitterpunkte	# Dreiecke	α_{min}	α_{max}
0	28	40	26.56	112.62
1	95	160	26.56	112.62
2	349	640	26.56	112.62
3	1337	2560	26.56	112.62
4	5233	10240	26.56	112.62

Tab.6 Gitterdaten für die Triangulierung, α_{min} bzw. α_{max} bezeichnen den minimalen Winkel der Triangulierungen.

Konvergenzraten:

Level	Shift 1	Shift 2	Konvergenzrate	# Iterationen bis Rundungsfehler
1	0.4	0.7	0.085	10
2	0.4	0.7	0.06	9
3	0.4	0.7	0.16	12
4	0.4	0.7	0.28	14

Tab.7 Konvergenzraten auf den Verfeinerungsstufen „Level". Shift 1 ist der Shiftparameter für das ILU-Verfahren auf der ersten Verfeinerungsstufe, Shift 2 ist der Shiftparameter auf den feineren Gittern. Die angegebenen Konvergenzrate ist die Summe aller Konvergenzraten einer Verfeinerungsstufe, dividiert durch die Anzahl der Iterationen. Der Rundungsfehler lag bei 10^{-11} .

Diskretisierungsfehler:

Level	e_∞	e_2	$\lvert\tilde{\lambda} - \lambda\rvert$	$\tilde{\lambda}$
0	$1.02 \cdot 10^{-1}$	$4.652 \cdot 10^{-2}$	$1.598 \cdot 10^{-2}$	0.310288
1	$2.02 \cdot 10^{-2}$	$1.557 \cdot 10^{-2}$	$4.293 \cdot 10^{-3}$	0.321975
2	$4.03 \cdot 10^{-3}$	$1.790 \cdot 10^{-3}$	$1.066 \cdot 10^{-3}$	0.325202
3	$1.22 \cdot 10^{-3}$	$5.270 \cdot 10^{-4}$	$2.646 \cdot 10^{-4}$	0.326003
4	$3.37 \cdot 10^{-4}$	$1.352 \cdot 10^{-4}$	$6.595 \cdot 10^{-5}$	0.326202

Tab.8 Diskretisierungsfehler; e_∞, e_2 sind wie in § 8.1.2 definiert; $\tilde{\lambda}$ bezeichnet die Eigenwertnäherung auf dem jeweiligen Gitter und $\lambda = 0.3262679$ den exakten Eigenwert.

Tabelle 8 zeigt die quadratische Konvergenz des Fehlers zwischen exakter Lösung und berechneter Näherungslösung in den entsprechenden Normen. Auch kann man die erwartete quadratische Konvergenz des Eigenwertes beobachten. Dies ist in der folgenden Abbildung graphisch dargestellt.

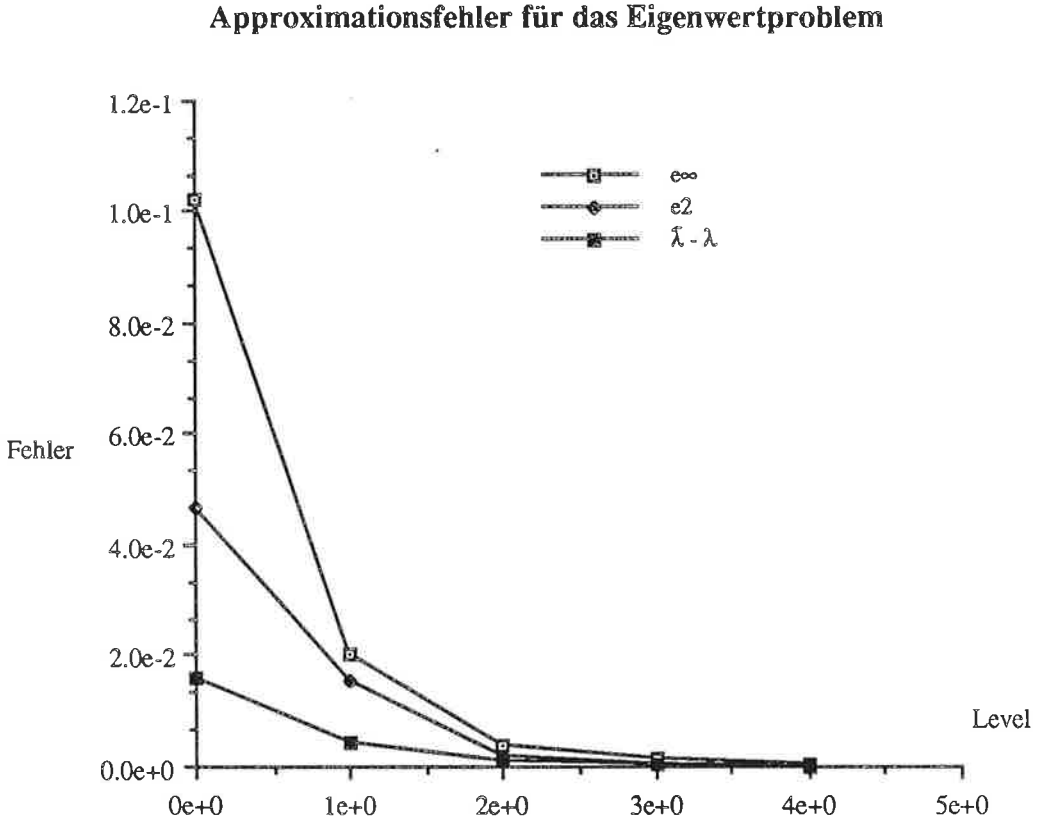


Abb.45 Graphische Darstellung des Fehlers zwischen kontinuierlicher und berechneter diskreter Lösung

Die folgende Abbildung zeigt die berechnete Eigenschwingung.

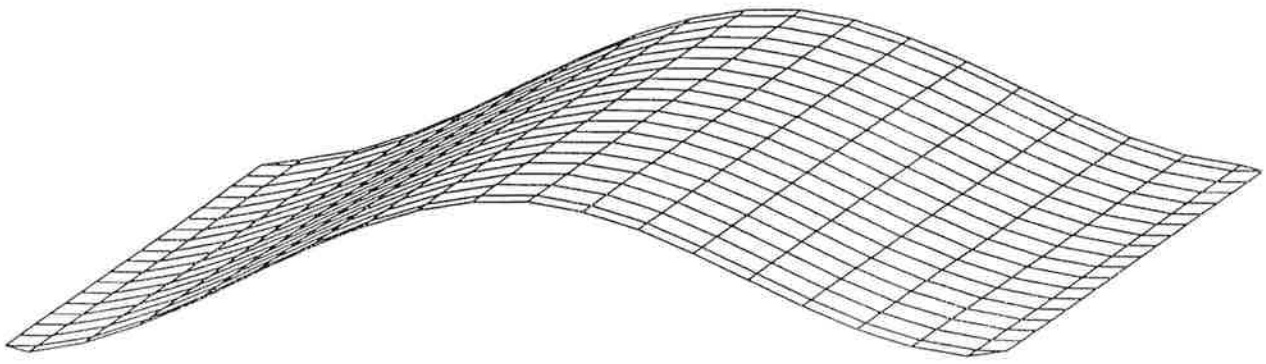


Abb.46 3-D Plot der berechneten zweiten Oberschwingung auf dem Rechteck (0,11) x (0,3)

§ 8.2.3 Das Eigenwertproblem für den Bodensee

Abschließend wollen wir die Eigenwertgleichung (8.2.1) für den Bodensee betrachten. Zunächst haben wir eine konstante „Äquivalenztiefe" von 102,78 Meter verwendet (siehe [2]). Das grobe Gitter besitzt 88 Knotenpunkte (vgl. Tab.5). Wir konnten mit den –aus dem Grobgitter resultierenden– Startwerten für die geschachtelte Iteration elf Eigenwerte auf der ersten Verfeinerung berechnen und sieben auf der zweiten Verfeinerungsstufe. Mit der variablen Tiefe des Bodensees erhalten wir nur noch Konvergenz für den niedrigsten Eigenwert. Dies liegt vor allem daran, daß wir mit sehr wenig Gitterpunkten auf dem groben Gitter und Dreiecken mit nahezu entarteten Winkeln rechnen. Hofmann [11] verwendet für das Einheitsquadrat 256 Punkte auf dem groben Gitter und erhält daraus 46 Eigenwerte. Bank berechnet die Eigenwerte des Oberen Sees (USA) mit dem Programm PLTMG [1] und startet dabei mit einem groben Gitter mit über 1000 Gitterpunkten. Unsere Ergebnisse sind unter diesen Gesichtspunkten völlig zufriedenstellend. Um die Ergebnisse noch weiter zu verbessern, müsste man sich überlegen, wie man ein feineres Grobgitter verwenden könnte. Eine Möglichkeit wäre, die erste Verfeinerungsstufe als Grobgitter zu verwenden. Damit würde die Besetzungsstruktur der Steifigkeitsmatrix erhalten bleiben, und man hätte etwa viermal so viele Knotenpunkte auf dem gröbsten Gitter wie in unserem Fall.

Bisher wurden die Eigenschwingungen des Bodensees von Bäuerle [2], Hamblin/ Hollan [10] und Rao [16] berechnet. Die folgende Tabelle zeigt einen Vergleich dieser Ergebnisse mit denen hier berechneten und den bisher gemessenen.

	gemessener Eigenwert	*Hamblin/Hollan*	*Bäuerle*	*Rao*	*Sauter*
1.	55.64	53.7	53.83	53.87	56.47
2.	37.72	35.7	35.95	35.96	38.77
3.	28.28	27.2	27.01	27.03	27.83
4.	19.13	19.4	19.83	19.84	20.74
5.		18.6			15.19
6.		16.4			15.00
7.	15.0	14.8			13.71
8.	14.6	14.3			13.20
9.		12.5			12.42
10.		12.0			11.65
11.	11.6	11.3			

Tab.9 Vergleich der Ergebnisse von Hamblin/Hollan, Bäuerle, Rao und Sauter mit den tatsächlich gemessenen. Die Eigenwerte sind in Minuten angegeben.

§ 8.2.4 Plots der berechneten Eigenschwingungen

Wir stellen auf den nächsten Seiten die berechneten Eigenschwingungen graphisch als 3-D Plot und als Höhenlinienplot dar. Die konstante Grundschwingung (Null-Schwingung) ist nicht dargestellt.

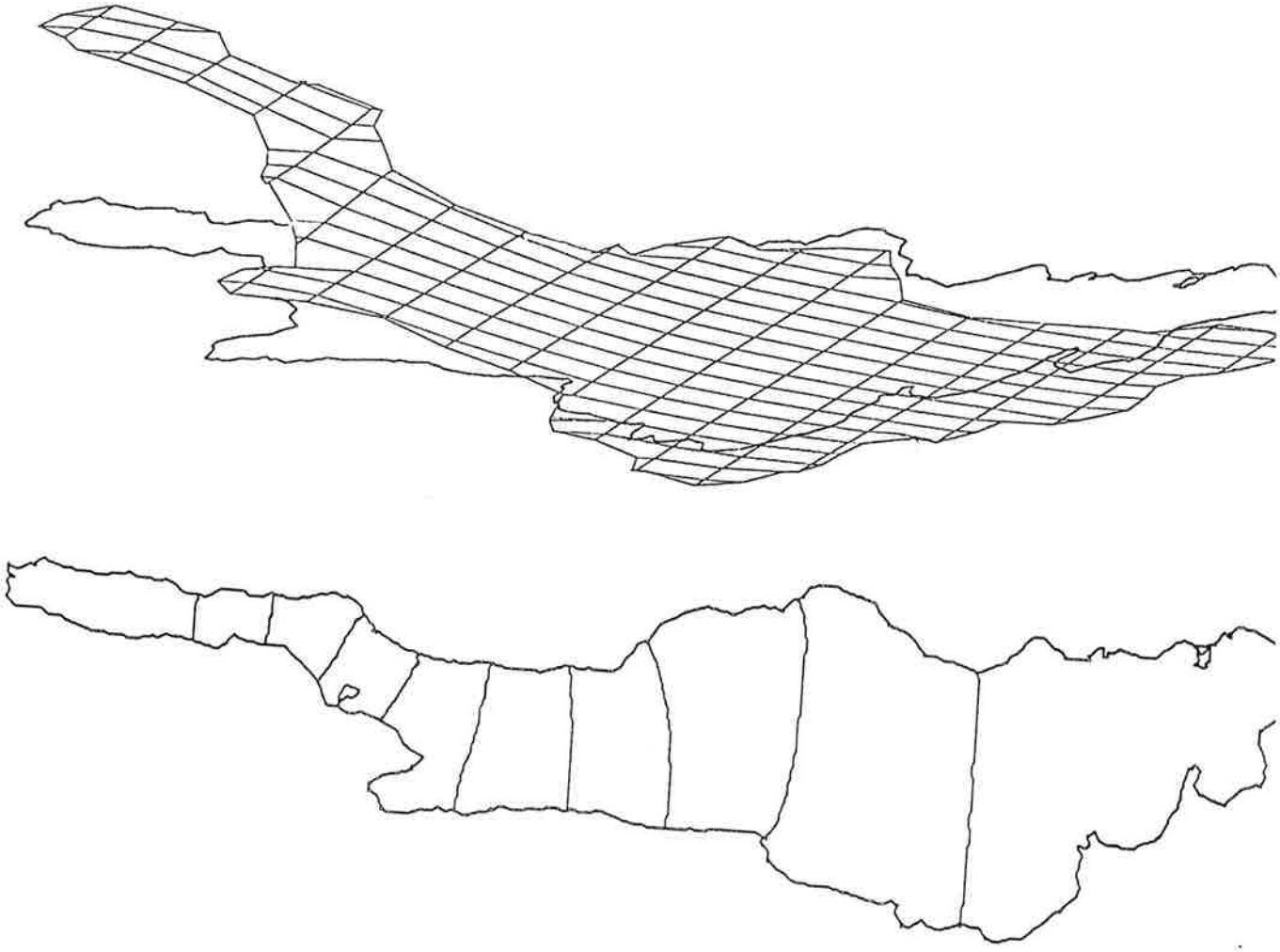


Abb. 47 1.Oberschwingung, Konvergenzraten: 0.6 auf der ersten und 0.86 auf der zweiten Verfeinerungsstufe

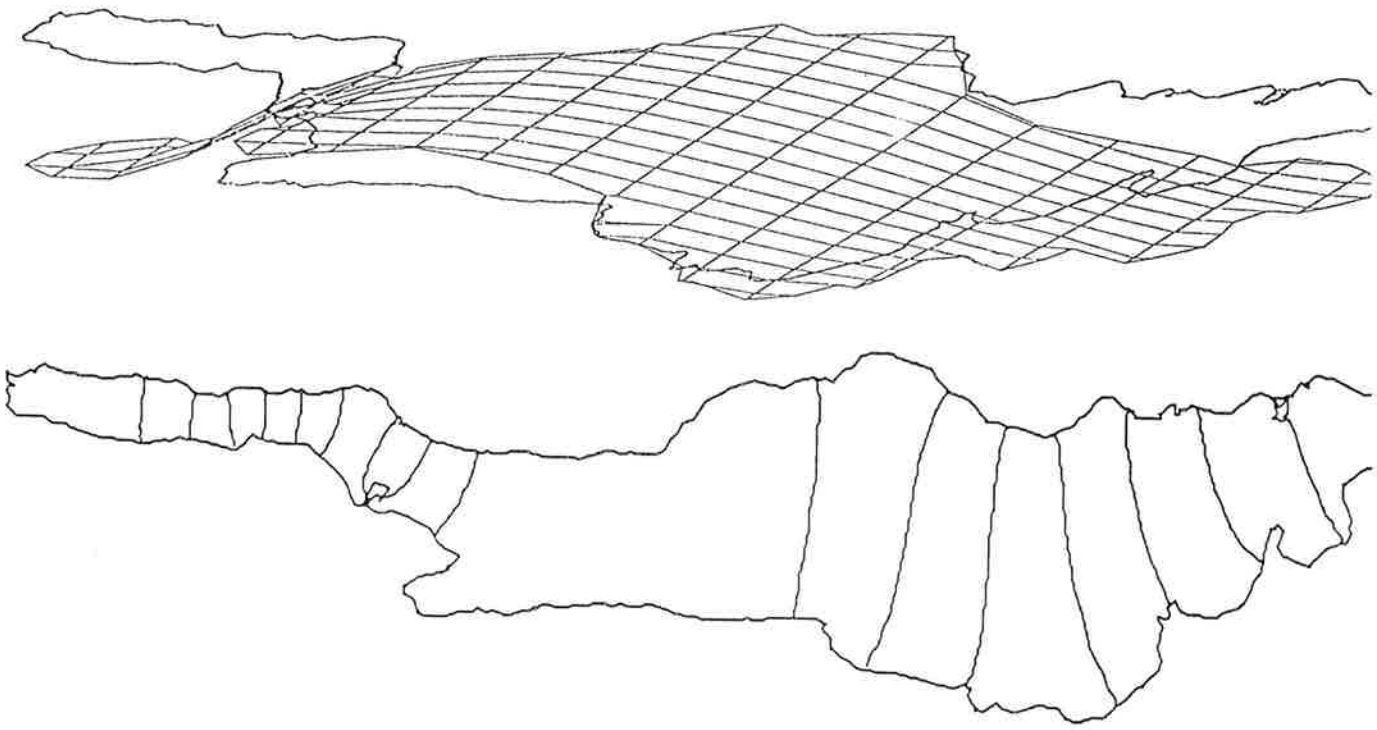


Abb. 48 2.Oberschwingung, Konvergenzraten: 0.6 auf der ersten und 0.858 auf der zweiten Verfeinerungsstufe

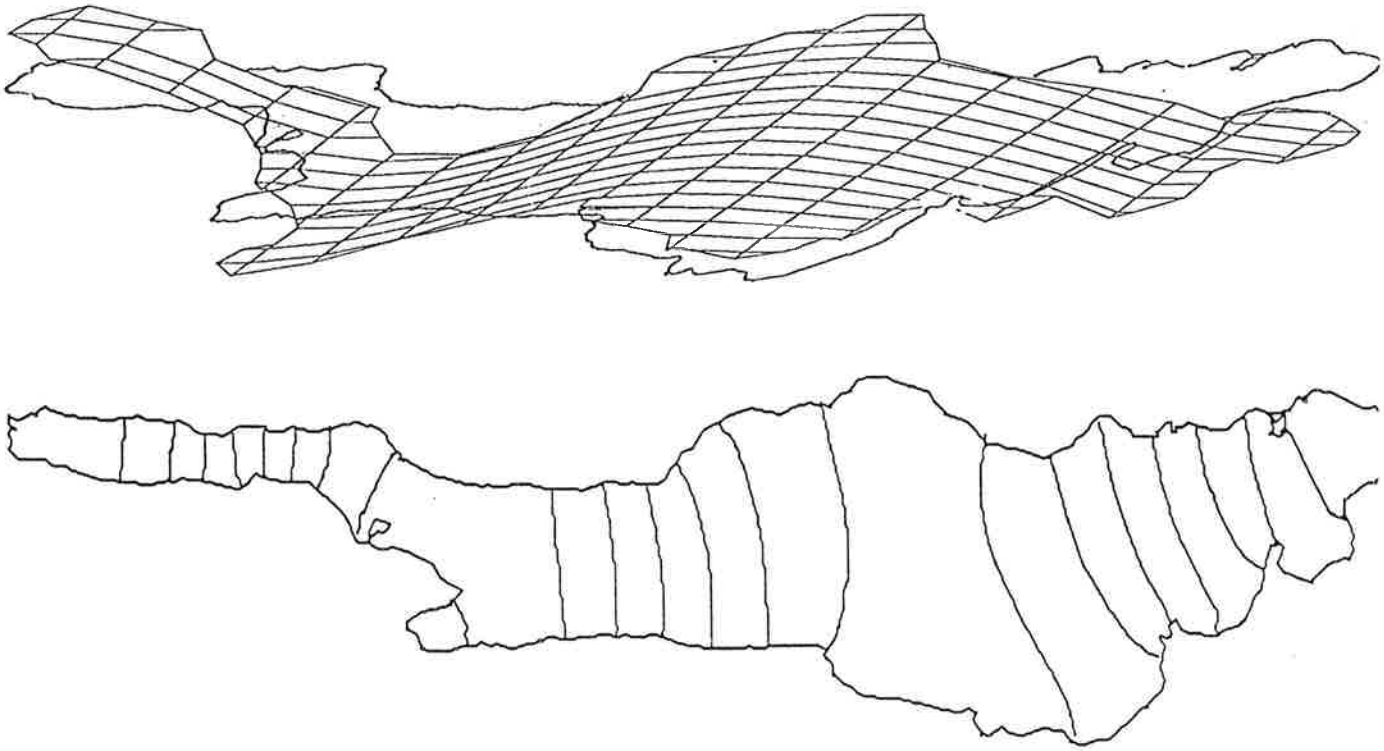


Abb. 49 3.Oberschwingung, Konvergenzraten: 0.6 auf der ersten und 0.858 auf der zweiten Verfeinerungsstufe

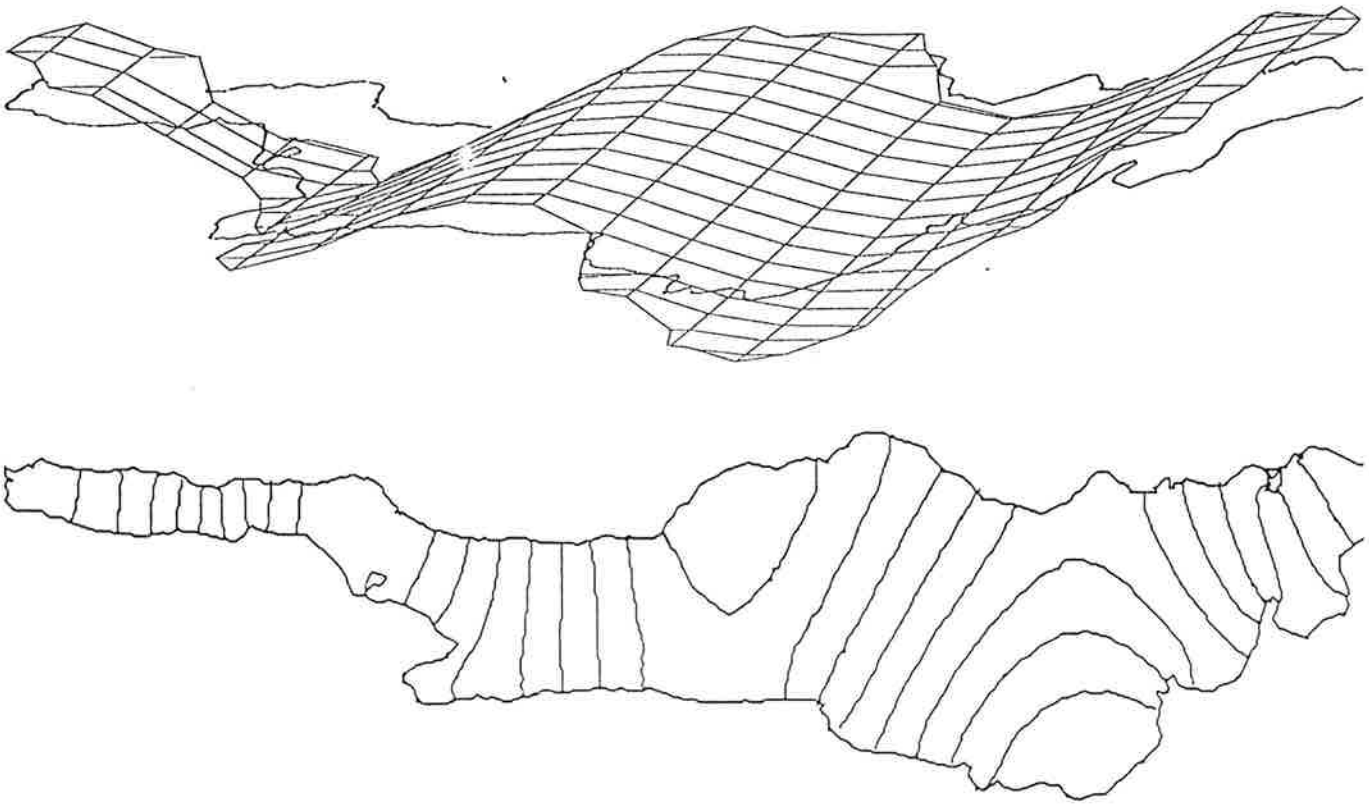


Abb. 50 4.Oberschwingung, Konvergenzraten: 0.6 auf der ersten und 0.865 auf der zweiten Verfeinerungsstufe

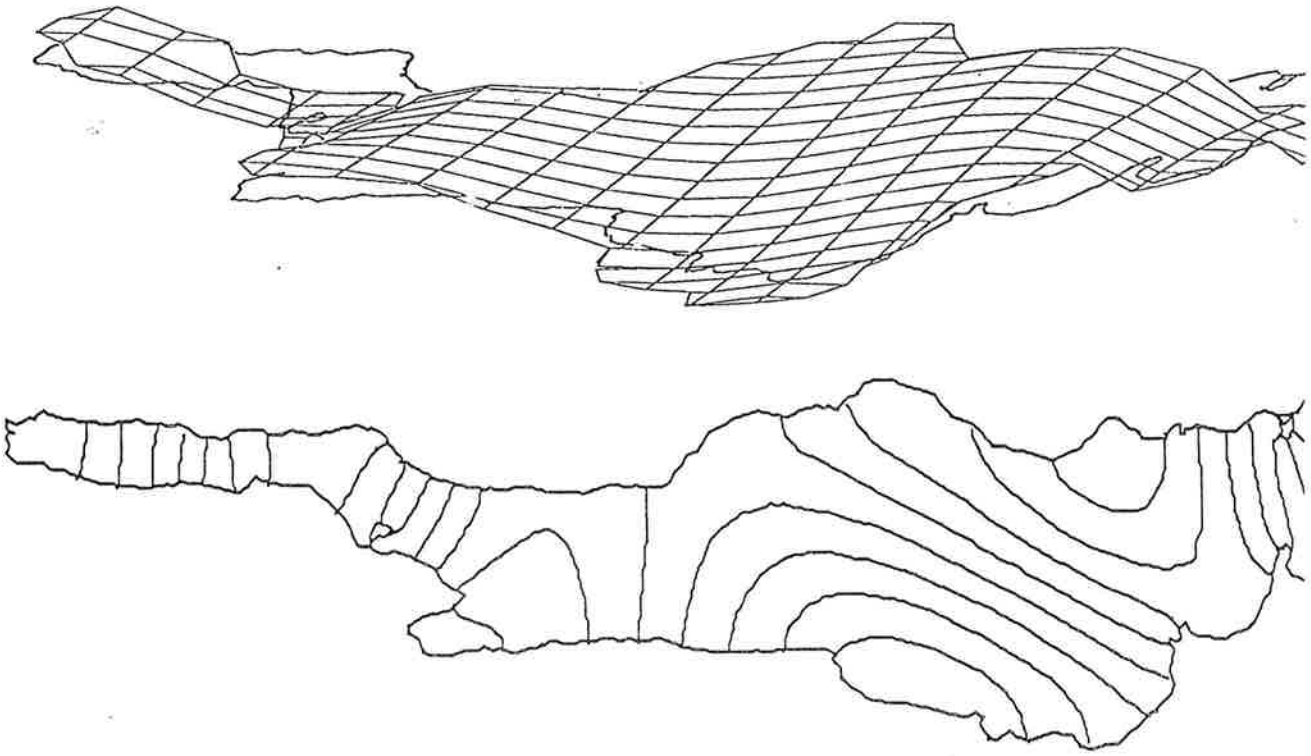


Abb. 51 5.Oberschwingung, Konvergenzraten: 0.58 auf der ersten und divergent auf der zweiten Verfeinerungsstufe

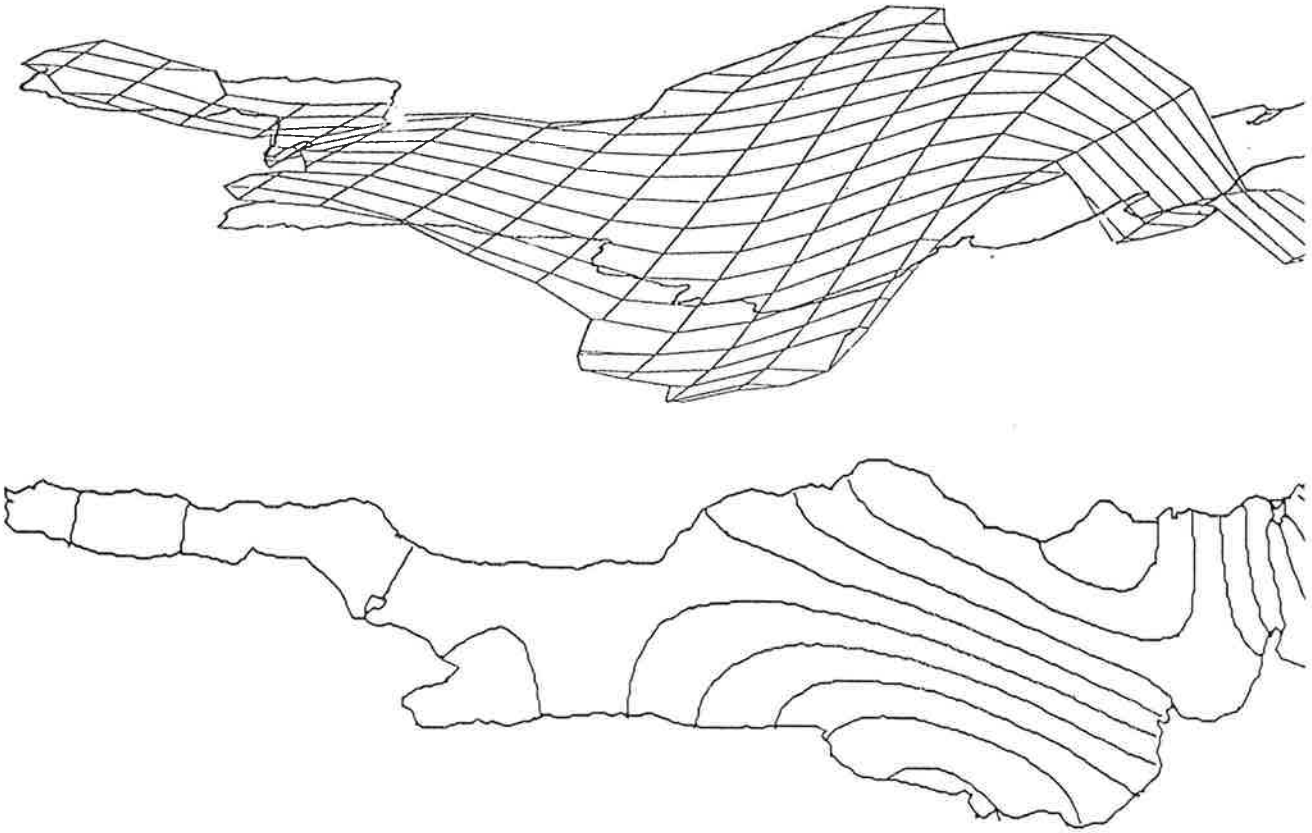


Abb. 52 6.Oberschwingung, Konvergenzraten: 0.85 auf der ersten und 0.76 auf der zweiten Verfeinerungsstufe

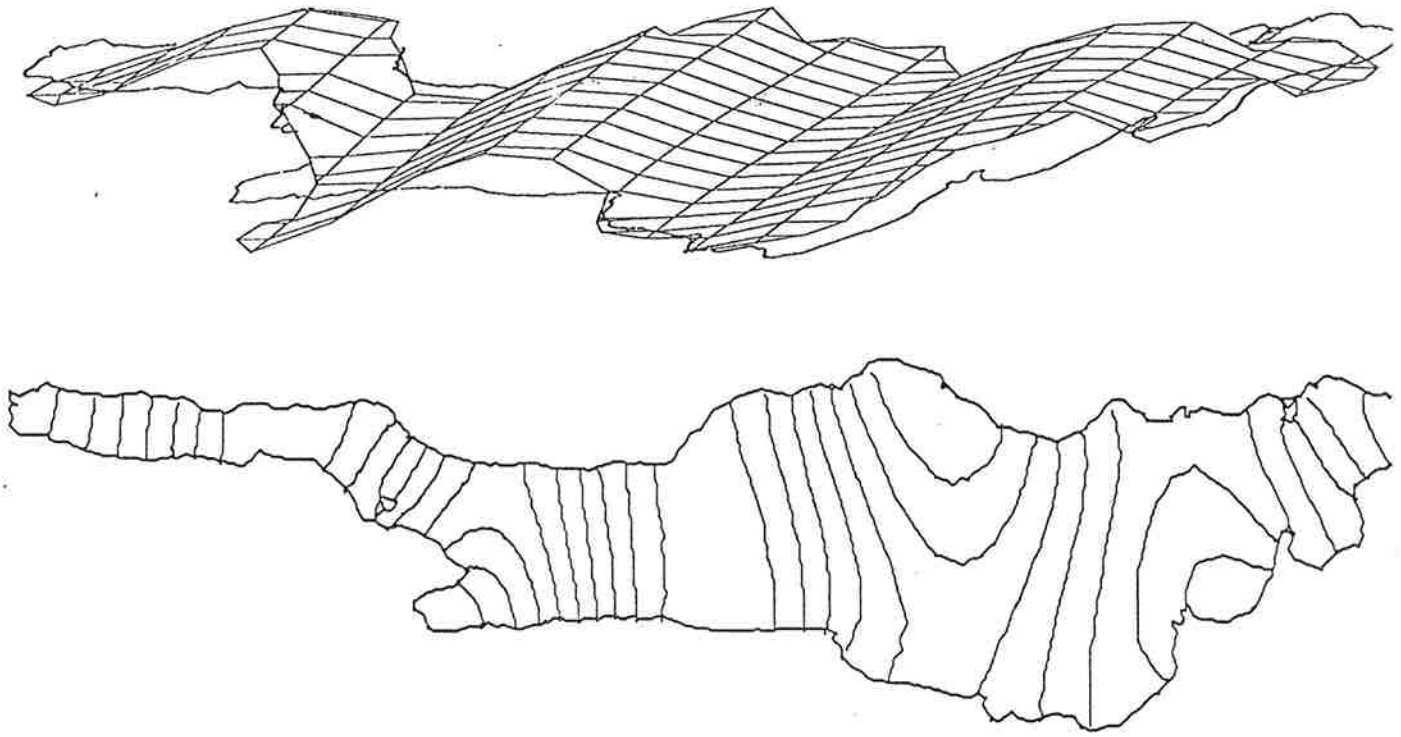


Abb. 53 7.Oberschwingung, Konvergenzraten: 0.8 auf der ersten und 0.78 auf der zweiten Verfeinerungsstufe

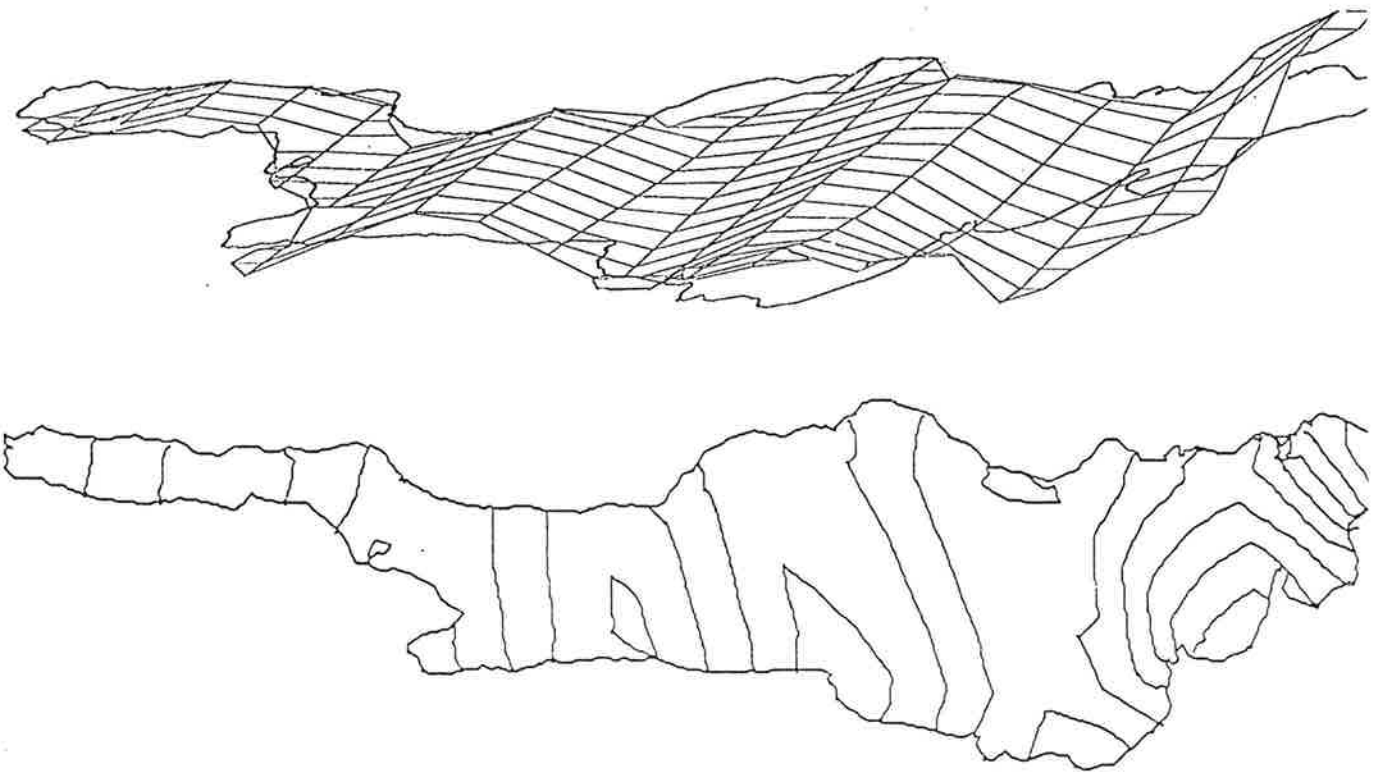


Abb. 54 8.Oberschwingung, Konvergenzraten: 0.74 auf der ersten und divergent auf der zweiten Verfeinerungsstufe

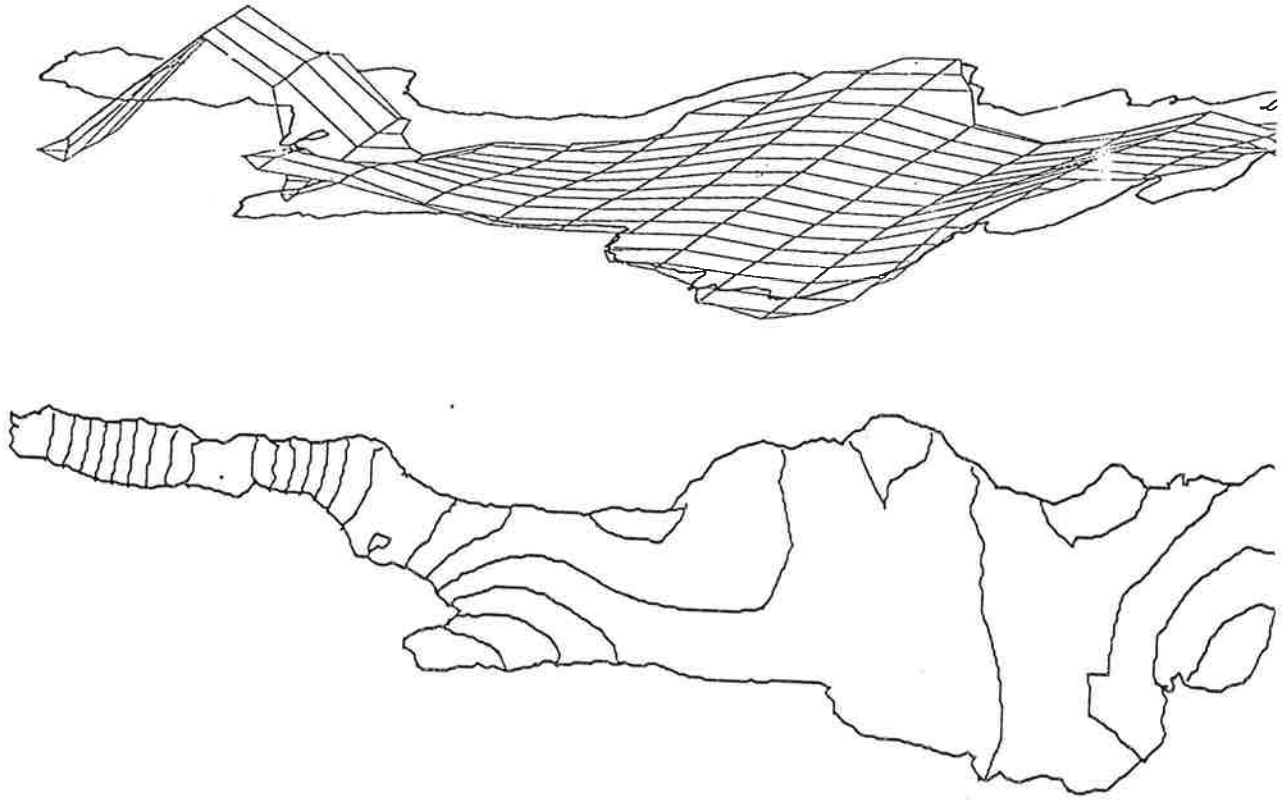


Abb. 55 9.Oberschwingung, Konvergenzraten: 0.8 auf der ersten und divergent auf der zweiten Verfeinerungsstufe

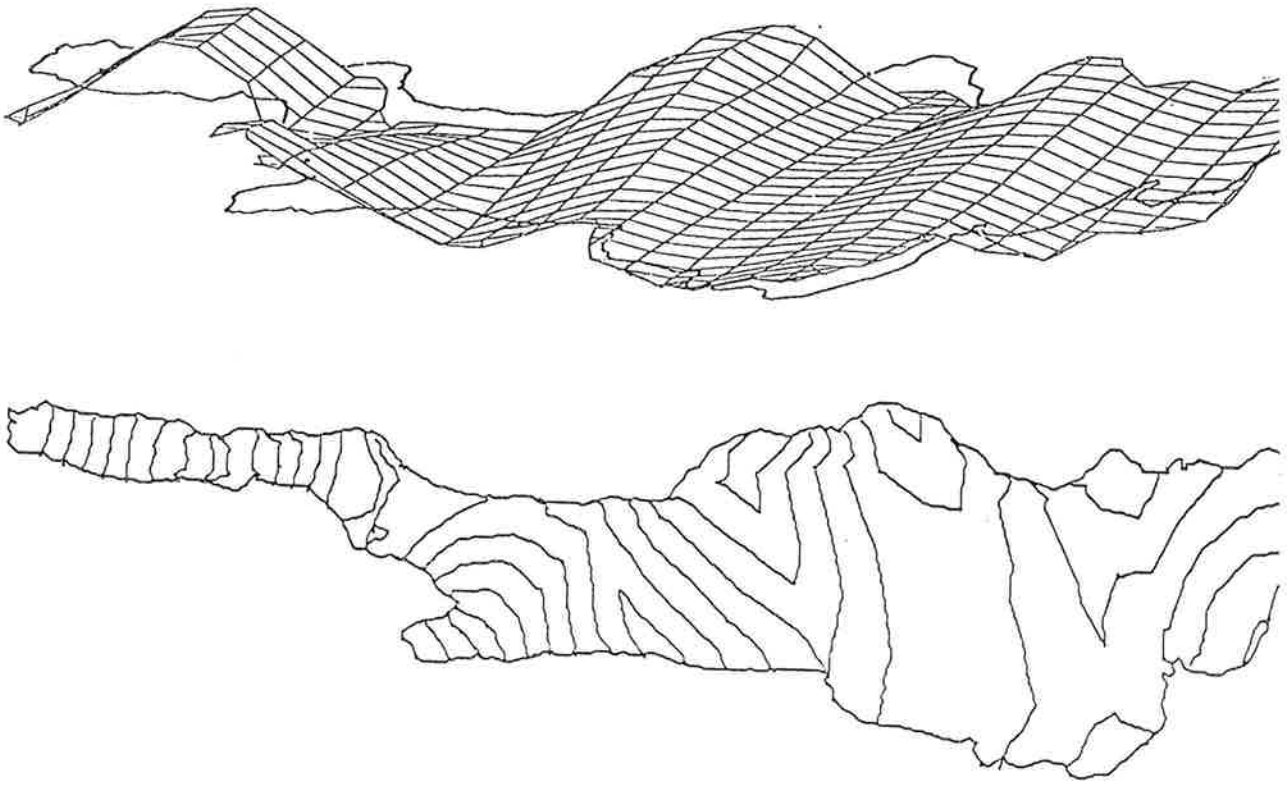


Abb. 56 10.Oberschwingung, Konvergenzraten: 0.8 auf der ersten und divergent auf der zweiten Verfeinerungsstufe

Notationen:

div, $\nabla\cdot$	Divergenzoperator
grad	Gradientenoperator in drei Dimensionen
∇	Gradientenoperator in zwei Dimensionen
U'	Dualraum zu U
$(\cdot,\cdot)_0$, $(\cdot,\cdot)_U$	L^2 - Skalarprodukt, Skalarprodukt im Hilbertraum U
$<\cdot,\cdot>$	Skalarprodukt im $\mathbb{R}^n$
$\|\cdot\|_0$, $\|\cdot\|_U$	L^2 - Norm, Norm auf dem Raum U
$\|\cdot\|_M$, $\|\cdot\|_{-M}$	gitterabhängige Normen, siehe p. 22
$\|\|\|\cdot\|\|\|_o$, $\|\|\|\cdot\|\|\|_M$, $\|\|\|\cdot\|\|\|_{U\leftarrow V}$	Operatornormen, siehe p.22

Literaturverzeichnis

[1] Bank, R. E.: PLTMG Users´ Guide - Edition 5.0. Department of Mathematics, University of California at San Diego, La Jolla 1988.

[2] Bäuerle, E.: Die Eigenschwingungen abgeschlossener, zweigeschichteter Wasserbecken bei variabler Bodentopographie. Berichte aus dem Institut für Meereskunde an der CAU Kiel, Nr. 85, 1981.

[3] Berman, A., Plemmons, R.
Nonnegative matrices in the mathematical sciences. Academic Press 1979.

[4] Bramble, J. H., Schatz, A. H.: Rayleigh-Ritz-Galerkin methods for Dirichlet´s problem using subspaces without boundary conditions. Comm. Pure Appl. Math. 23, p.633-675, 1970.

[5] Bramble, J. H., Osborn, J. E.: Rate of convergence estimates for nonself-adjoint eigenvalue approximations. Math. Comp. 27, 525-549.

[6] Chatelin, F.: Spectral approximation of linear operators. Academic Press New York 1983.

[7] Hackbusch, W.: On the computation of approximate eigenvalues and eigenfunctions of elliptic operators by means of a multi-grid method. SIAM J Numer Anal 16, 201-215, 1979.

[8] Hackbusch, W.: Multi-grid methods and applications.Springer, Berlin, Heidelberg, 1985.

[9] Hackbusch, W.: Theorie und Numerik elliptischer Differentialgleichungen.Teubner, Stuttgart, 1986.

[10] Hamblin, P. F., Hollan, E.: On the gravitational seiches of Lake Constance and their generation. Schweiz. Z. Hydrol. 40, 119-154, 1978.

[11] Hofmann, G.: Analyse eines Mehrgitterverfahrens zur Berechnung von Eigenwerten elliptischer Differentialoperatoren. Dissertation, Kiel 1985.

[12] Krauss, W.: Methods and results of theoretical oceanography. Gebr. Borntträger, Berlin Stuttgart 1973.

[13] Meijerink, J. A, Van der Vorst, H. A.: An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix. Math. Comp. 31, 148-162, 1977.

[14] Pätsch,J. :Ein Mehrgitterverfahren zur Berechnung der Eigenwerte und der Eigenvektoren der Oberflächenschwingungen eines abgeschlossene Wasserbeckens. Diplomarbeit, Institut für Informatik und Praktische Mathematik, CAU Kiel.

[15] Rannacher, R., Blum, H.: Finite element eigenvalue computation on domains with reentrant corners using Richardson extrapolation. Fachbereich Mathematik, Universität des Saarlandes, Saarbrücken.

[16] Rao, D. B., Hollan, E., Bäuerle, E.: Free surface oscillations in Lake Constance with an interpretation of the „Wonder of the rising Water" at Konstanz in 1549. Arch. Met. Geoph. Biokl., Ser. A, 29, 301-325,1980.

[17] Schmitt, M.: Finite Elemente Diskretisierungen auf degenerierenden Gittern. Diplomarbeit, Universität Saarbrücken, 1989.

[18] Schulthaiss, C.: Wunder anloffen des wassers. Collectaneen, Vol.VI, 80-81, 1549.

[19] Strang, G., Fix, G. J.: An analysis of the finte element method. Prentice-Hall, Englewood Cliffs 1973.

[20] Varga, R. S.: Matrix iterative analysis. Prentice Hall, 1962.

[21] Verfürth, R.: A posteriori error estimators for the Stokes equation. Universität Heidelberg, Sonderforschungsbereich 123, Preprint 445, 1987.

[22] Wesseling, P.: Theoretical and practical aspects of a multigrid method. SIAM J. Sci. Statist. Comp. 3 , 387-407, 1982.

[23] Wittum, G.: Distributive Iterationen für indefinite Systeme. Dissertation, Universität Kiel, 1986.

[24] Wittum, G.: On the robustness of ILU-smoothing. Universität Heidelberg, Sonderforschungsbereich 123, Preprint 451, 1988. To appear in SISSC.

[25] Wittum, G.: Linear Iterations as Smoothers in Multigrid Methods: Theory with Aplications to Incomplete Decompositions. Impact of computing in science and engineering 1, 180-215, 1989.