Prom. Nr. 2328

DIE ANWENDUNG DER METHODE DER KONJUGIERTEN GRADIENTEN UND IHRER MODIFIKATIONEN AUF DIE LOESUNG LINEARER RANDWERTPROBLEME

von der

EIDGENÖSSISCHEN TECHNISCHEN HOCHSCHULE IN ZÜRICH

zur Erlangung
DER WÜRDE EINES DOKTORS DER
MATHEMATIK

genehmigte

PROMOTIONSARBEIT

vorgelegt von
URS HOCHSTRASSER
von Zürich und Gisikon (LU.)

Referent: Herr Prof. Dr. E. Stiefel
Korreferent: Herr Priv. Doz. Dr. H. Rutishauser

Druck: E. Truninger, Urania, Zürich 1954

INHALTSVERZEICHNIS

EINLEITUNG

Eines der häufigsten Probleme, das in den verschiedensten Zusammenhängen sich dem angewandten Mathematiker stellt, ist die Auflösung von linearen Gleichungssystemen. Diese Aufgabe scheint einfach zu sein und wurde theoretisch schon vor langer Zeit gelöst. Die numerische Behandlung des Problems bietet jedoch, sobald die Zahl der Gleichungen gross ist, erhebliche Schwierigkeiten, obwohl die Aufgabe nur die Anwendung der elementarsten Rechenoperationen verlangt. Die Schwierigkeiten liegen eigentlich nur in der enormen Häufung von Operationen, die einen grossen Zeitaufwand verursacht und die Genauigkeit der Resultate erheblich beeinträchtigen kann.

Das älteste praktische Verfahren ist die Eliminationsmethode, die von Gauss in systematische Form gebracht wurde und deshalb meist als Gauss'scher Algorithmus bezeichnet wird. Seither sind noch einige andere Methoden entwickelt worden, jedoch hat sich dieser Algorithmus für die Auflösung kleinerer Gleichungssysteme als einfacher und rascher erwiesen. Einzig bei speziellen Gleichungssystemen, deren Koeffizientenmatrix ausserhalb der Hauptdiagonalen eine grosse Zahl von verschwindenden Elementen besitzt, haben sich die besonders in den letzten Jahren stark entwickelten iterativen Methoden (Relaxationsrechnung) überlegen gezeigt und deshalb eine breitere Anwendung gefunden.

Mit dem Bau von programmgesteuerten Rechenmaschinen mit grosser Rechengeschwindigkeit haben sich allerdings die Gesichtspunkte, unter denen die Lösungsmethoden beurteilt werden, verschoben, da die Zahl der Rechenoperationen nicht mehr eine solche Rolle spielt. Hingegen wirkt bei der Anwendung des Gauss'schen Algorithmus auf grössere Gleichungssysteme die Notwendigkeit des Mitführens einer grossen Zahl von Stellen zur Erzielung einer genügenden Genauigkeit sehr hinderlich. Dazu kommt als weiterer Nachteil dieser Methode beim Gebrauch von programmgesteuerten Rechenmaschinen, dass sie nicht in einfacher Weise zyklisch ist, d.h. die Lösungen werden nicht durch mehrmalige Anwendung desselben Rechenprogrammes gefunden. Die erwähnten Nachteile führten in letzter Zeit zur Entwicklung eines neuen Verfahrens, das sich speziell für die Verwendung auf programmgesteuerten Rechenmaschinen eignet. Dieses unter dem Namen "Methode der konjugierten Gradienten" bekannte Verfahren wurde gleichzeitig von Hestenes[1] und Mitarbeitern, sowie unabhängig von diesen, von Stiefel[2] gefunden und später in einer gemeinsamen Arbeit[3] ausführlich diskutiert. Die von C. Lanczos entwickelte Methode der "minimalisierten Iterationen"[4] ist damit nahe verwandt. Die Methode der konjugierten Gra-

1) M.R. Hestenes, Iterative Methods for Solving linear Equations, NAML Report 52-9 National Bureau of Standards, Los Angeles, 1951

2) E. Stiefel, Ueber einige Methoden der Relaxationsrechnung, ZAMP 3, 1952

3) M.R. Hestenes und E. Stiefel, Method of conjugate Gradients for solving linear Systems NBS J. Res. 49, 409-436, (Dez. 1952)

4) C. Lanczos, Solution of Systems of Linear Equations by Minimized Iterations NBS J. Res. 49, 33-53 (Jul. 1952)

dienten stellt in gewissem Sinne eine Verallgemeinerung der von G. Temple[5] diskutierten Methode des steilsten Abstieges dar. Im Gegensatz zur letzteren benützt sie jedoch in jedem Stadium die ganze Vorgeschichte der Relaxationsrechnung so, dass man theoretisch nur endlich viele Iterationen zur Bestimmung der exakten Lösung eines linearen Gleichungssystemes braucht. Zudem kommt man in jedem Schritt dem Lösungspunkt näher. In günstigen Fällen muss man deshalb nicht alle der theoretisch notwendigen Iterationen ausführen, um eine brauchbare Lösung zu erhalten. Andererseits kann man durch wenige zusätzliche Iterationen Lösungen mit grossen Rundungsfehlern verbessern. Zudem kann durch geeignete Vorbehandlung der Ausgangsdaten erreicht werden, dass die Methode stabil bezüglich der Ausbreitung von Rundungsfehlern ist. Die Methode der konjugierten Gradienten hat sich dank dieser Eigenschaften bei der Auflösung sehr umfangreicher linearer Gleichungssysteme bewährt und eignet sich besonders für die Verwendung mit programmgesteuerten Rechenmaschinen. Als Beispiel haben wir die von uns auf der programmgesteuerten Rechenmaschine des Institutes für angewandte Mathematik der E.T.H. durchgeführte Berechnung einer parallelogrammförmigen Scheibe in die vorliegende Arbeit aufgenommen.

Bei der numerischen Lösung linearer Randwertprobleme treten Differenzengleichungssysteme auf, deren Koeffizientenmatrix ausserhalb der Hauptdiagonalen zum grössten Teil Nullen aufweist. In einem solchen Fall stellt die Berechnung der beiden Parameter in jedem Schritt im Verhältnis zu den übrigen Operationen einen erheblichen Aufwand dar, da sie die Bildung von mindestens zwei Skalarprodukten erfordert. Deshalb haben wir eine Modifikation der Methode untersucht, bei der die Parameter festgehalten werden. Die Verwendung von zwei festen Parametern wurde schon von Frankel[6] in Betracht gezogen, eine ausführliche Diskussion und richtige Würdigung des Verfahrens fehlte jedoch bis jetzt. Bei dem modifizierten Verfahren gehen natürlich viele Eigenschaften der Methode der konjugierten Gradienten verloren. Vor allem erhalten wir nicht mehr in endlich vielen Schritten die exakte Lösung. Praktisch ist dies von keiner grossen Bedeutung. Es genügt, wenn das Verfahren eine gute Konvergenz gegen die exakte Lösung aufweist, da man die Lösung in der angewandten Mathematik immer nur auf eine beschränkte Zahl von Stellen genau kennen muss.

Das modifizierte Verfahren, das wir "Frankel'sches Verfahren" nennen wollen, verlangt eine ungefähre Kenntnis des grössten und kleinsten Eigenwertes der Gleichungsmatrix. Wenn diese nur mit grossem Zeitaufwand bestimmbar sind, wird man mit Vorteil zuerst die Methode der konjugierten Gradienten anwenden und nach einigen Iterationen mit Hilfe der so gesammelten Informationen auf das Frankel'sche Verfahren übergehen. Dieses Vorgehen wird im § 6 an einem Beispiel näher erläutert werden. Der Uebergang ist auch bei der Verwendung von

5) G. Temple, The general Theory of Relaxation Methods applied to Linear Systems, Proc. Roy.Soc. ser A, 169 476-500 (1939)

6) S.P. Frankel, Convergence Rates of Iterative Treatments of Partial Differential Equations, Math. Tables and other Aids f.Comput. 4,30 (1950) Frankel bezeichnet in dieser Arbeit diese Methode als "Second order Richardson's method"

programmgesteuerten Rechenmaschinen leicht zu bewerkstelligen, da das Rechenprogramm nur kleiner Abänderungen bedarf.
Eine spezielle Eigenheit des Frankel'schen Verfahrens besteht im Auftreten von Schwingungen, die sich unter Umständen zur Verbesserung der Konvergenz des Verfahrens verwenden lassen. Allerdings wird man bei der praktischen Anwendung oft nicht genügend Informationen besitzen, um ohne grossen Rechenaufwand den Schwingungscharakter ausnützen zu können. Ein weiterer Vorteil dieses Verfahrens gegenüber der Methode der konjugierten Gradienten besteht vom rechentechnischen Standpunkt aus in der Möglichkeit, die Rechnung so anzulegen, dass man neben der Matrix D des Gleichungssystemes zu keiner Zeit mehr als zwei Vektoren gespeichert halten muss. Im Hinblick auf die bei vielen Maschinen beschränkte Speicherkapazität ist dies vielleicht ein nicht unwesentlicher Punkt. Gegenüber dem Gesamtschrittverfahren[7], das als Spezialfall des Frankel'schen Verfahrens betrachtet werden kann, besitzt das Frankel'sche Verfahren den Vorteil, rascher zu konvergieren. Die Konvergenz ist speziell bei schlecht konditionierten Matrizen D, d.h. falls die Eigenwerte von D weit voneinander liegen, wesentlich besser. Deshalb eignet sich das Frankel'sche Verfahren auch für umfangreiche Differenzengleichungssysteme, die bei der Wahl eines Differenzennetzes mit sehr kleiner Maschenweite entstehen und die wegen der Proportionalität des kleinsten Eigenwertes zur Maschenweite eine schlecht konditionierte Koeffizientenmatrix besitzen.
Am nützlichsten wird sich wahrscheinlich das Frankel'sche Verfahren in Verbindung mit der Methode der konjugierten Gradienten zur Abkürzung der Rechnung erweisen.

7) für eine Diskussion dieses Verfahrens siehe:
L.F. Richardson, The Approximate Arithmetical Solution by finite Differences of physical Problems, Phil. Trans. Roy. Soc. ser. A 210 307-357 (1911)

§ 1 DIE METHODE DER KONJUGIERTEN GRADIENTEN

Die Methode der konjugierten Gradienten zusammen mit der Methode des steilsten Abstieges und dem Gesamtschrittverfahren, stellt den Ausgangspunkt unserer Betrachtungen dar. Wir wollen deshalb diese Verfahren kurz rekapitulieren, wobei wir für die Beweise auf die schon zitierten Arbeiten (1), (2) hinweisen.
Es sei das symmetrische, lineare Gleichungssystem mit der positiv definiten symmetrischen Koeffizientenmatrix (d_{ik}) und den absoluten Gliedern l_i

$$(1.1) \quad \sum_{k=1}^{n} d_{ik} x_k + l_i = 0 \qquad (i = 1,2 \cdots n)$$

nach den Unbekannten x_k aufzulösen. Wir verwenden im folgenden die Vektorschreibweise, x_k und l_i seien die Komponenten der Vektoren x und l in einem n – dimensionalen Raum R^n, die Koeffizientenmatrix (d_{ik}) fassen wir als Operator D auf, der den Vektor x in einen neuen Vektor $y = Dx$ transformiert. In dieser Schreibweise lautet Gleichung (1.1)

$$(1.1a) \qquad Dx + l = 0$$

Die Aufgabe, dieses Gleichungssystem zu lösen, ist aequivalent zum Problem, das Minimum der quadratischen Funktion F(x):

$$(1.2) \qquad F(x) = \tfrac{1}{2} \sum_{i,k=1}^{n} d_{ik} x_i x_k + \sum_{i=1}^{n} l_i x_i$$

zu bestimmen. Definieren wir als skalares Produkt (x,y) der Vektoren x und y die Summe:

$$(x,y) = \sum_{k=1}^{n} x_k y_k$$

so lautet (1.2) in Vektorschreibweise:

$$(1.2a) \qquad F(x) = \tfrac{1}{2} (x,Dx) + (l,x)$$

Für einen Näherungsvektor $v=(v_1 \ldots v_n)$ zum unbekannten Lösungsvektor x steht auf der rechten Seite der Gleichung (1.1a) im allgemeinen ein Vektor, dessen Komponenten von Null verschieden sind und Residuen genannt werden;

$$(1.3) \qquad Dv + l = r$$

Der Residuenvektor r kann auch als Gradient der quadratischen Funktion F(v) aufgefasst werden:

$$r = \text{grad } F(v)$$

Die Aequivalenz zwischen dem linearen Gleichungssystem und einem Variationsproblem legt es nahe, in Analogie zum Ritz'schen Verfahren die Versuchsvektoren v auf eine lineare Schar zu beschränken, indem wir einen Gewichtsvektor p einführen und ausgehend von einem Versuchsvektor v mit Hilfe dieses Gewichtsvektors neue Versuchsvektoren v' bestimmen:

$$(1.4)\quad v' = v + \lambda p$$

Die quadratische Funktion $F(v') = F(\lambda)$ wird dann als Funktion von λ minimalisiert, falls wir die Wahl treffen:

$$(1.5)\quad \lambda = -\frac{(r,p)}{(p,Dp)} \qquad \text{wo r das zu v gehörige Residuum darstellt.}$$

Bei der Methode der konjugierten Gradienten wird ein System von n linear unabhängigen Gewichtsvektoren $p_0 \cdots p_{n-1}$ *) sukzessive in n Schritten so konstruiert, dass am Schluss die exakte Lösung des Problemes als Entwicklung nach den Gewichtsvektoren p_i vorliegt. Die Gewichtsvektoren p_i sind dabei zueinander konjugiert, d.h. $(p_i, Dp_k) = 0$ für $i \neq k$. Die nach jedem Zyklus nach der Vorschrift (1.4) gebildeten Näherungslösungen v_k

$$(1.6)\quad v_k = v_{k-1} + \lambda_{k-1}\, p_{k-1} = v_o + \sum_{j=0}^{k-1} \lambda_j p_j$$

approximieren zudem die exakte lösung immer besser in dem Sinne, dass der v_k darstellende Punkt näher am Lösungspunkt liegt, als alle vorhergehenden. Ausgangspunkt bildet irgend eine Versuchslösung v_o (z.B. der Nullvektor) mit zugehörigem Residuum r_o

$$(1.7)\quad Dv_o + l = r_o$$

Der erste Zyklus unterscheidet sich von allen andern, indem hier der erste Gewichtsvektor p_o wie bei dem sogenannten Verfahren des stärksten Abstieges, entgegengesetzt gleich dem Residuenvektor r_o genommen wird:

$$p_o = -r_o$$

Beim allgemeinen Zyklus geht man von den schon berechneten Vektoren r_{k-1} und p_{k-1} und den von ihnen abgeleiteten Grössen Dp_{k-1} und λ_{k-1} aus.

Der zur Näherungslösung v_k gehörende Residuenvektor r_k berechnet sich dann aus der Formel

$$(1.8)\quad r_k = r_{k-1} + \lambda_{k-1} Dp_{k-1} \quad \text{und ist orthogonal zu allen andern } r_j$$

Der auf p_{k-1} folgende Gewichtsvektor p_k ist durch die Rekursionsformel gegeben:

$$(1.9)\quad p_k = -r_k + \varepsilon_{k-1} p_{k-1}$$

* Die Indizes bedeuten in Zukunft eine Numerierung von Vektoren, falls nicht ausdrücklich erwähnt wird, dass sie Vektorkomponenten bezeichnen.

wobei ε_{k-1} aus der Forderung, dass p_k konjugiert zu p_{k-1} sein soll, $(p_k, Dp_{k-1}) = 0$ zu

$$(1.10) \quad \varepsilon_{k-1} = \frac{(r_k, r_k)}{(r_{k-1}, r_{k-1})}$$

bestimmt wird. Damit ist man beim Ausgangspunkt zum nächsten Zyklus angelangt. Da die Residuenvektoren nicht nur orthogonal sondern auch zueinander konjugiert sind, mit Ausnahme derjenigen Vektoren, deren Indizes unmittelbar benachbart sind,

$$(r_k, Dr_j) = 0 \qquad j \neq k, k-1, k+1$$

können wir die Formel (1.5) für λ_k umformen:

$$(1.5a) \quad \lambda_k = \frac{(r_k, r_k)}{(p_k, Dp_k)}$$

Zusammenfassend besitzt die Methode der konjugierten Gradienten folgende wesentliche Eigenschaften:

1. Das Verfahren führt nach höchstens n Zyklen zur exakten Lösung des gegebenen Gleichungssystems.
2. In jedem Zyklus gelangt man näher zum Lösungspunkt.
3. Die dabei konstruierten Gewichtsvektoren bilden ein konjugiertes System.
4. Die Residuenvektoren bilden ein orthogonales System und sind zueinander mit Ausnahme der unmittelbar benachbarten konjugiert.

Das Verfahren kann auch ohne weiteres auf nichtsymmetrische Matrizen übertragen werden unter Erhaltung der Eigenschaften 1-4 (siehe dazu die in der Fussnote 3 der Einleitung zitierte Arbeit).

Bei der praktischen Anwendung der Methode ist die geeignete Wahl der Versuchslösung v_o von grosser Wichtigkeit für die rasche Konvergenz und die Stabilität gegenüber Rundungsfehlern. Bei Gleichungssystemen mit symmetrischer und positiv definiter Koeffizientenmatrix kann man zeigen, dass falls im Anfangsresiduenvektor r_o der Anteil des zum kleinsten Eigenwert gehörigen Eigenvektors stark überwiegt, die Methode stabil gegenüber Rundungsfehlern ist und zudem rasch konvergiert. Es ist deshalb von Vorteil mit Hilfe einer Relaxationsmethode, die die Tendenz hat, die Anteile der höhern Eigenvektoren am stärksten aus den Residuen zu eliminieren, die Versuchslösung vorzubehandeln.

Als Beispiel für die Anwendung der Methode der konjugierten Gradienten auf ein umfangreiches Gleichungssystem ist am Ende dieser Arbeit die von uns auf der programmgesteuerten Rechenmaschine des Institutes für angewandte Mathematik der E.T.H. durchgeführte Berechnung der Deformationen in einer parallelogrammförmigen Scheibe wiedergegeben.

Wie schon in der Einleitung erwähnt, kann die Methode des stärksten Abstieges aus der Methode der konjugierten Gradienten abgeleitet werden, falls nämlich die Parameter ε_k alle Null gesetzt werden. Man wählt also in diesem Falle die Gewichtsvektoren gerade entgegengesetzt gleich zu den Residuenvektoren, die Parameter λ_k stellen dann gerade das Reziproke des Rayleigh'schen Quotienten der Residuenvektoren r_k dar. Die Methode des stärksten Abstieges ist somit durch die Formeln gegeben

$$r_k = r_{k-1} - \lambda_{k-1} Dr_{k-1}$$

$$v_k = v_o - \sum_{j=0}^{k-1} \lambda_j r_j \qquad (1.11)$$

$$\lambda_{k-1} = \frac{(r_{k-1}, r_{k-1})}{(r_{k-1}, Dr_{k-1})}$$

Auch bei diesem Verfahren wird die quadratische Funktion F in Richtung des jeweiligen Gewichtsvektors minimalisiert, die Lösung kann jedoch hier in endlich vielen Schritten nur approximiert werden.
Setzt man auch noch λ_k konstant so gelangt man zum Gesamtschrittverfahren, das also durch die folgende Rekursionsformel gegeben ist:

$$v_k = v_{k-1} - \lambda r_{k-1} \qquad (1.12)$$

oder $\quad \Delta v_k = -\lambda r_{k-1}$

Die Konvergenz dieses Verfahrens kann auf einfache Weise mit Hilfe der Eigenwerte μ_i $(i = 1,\ldots,n)$ der n-reihigen Matrix D und der zugehörigen Eigenvektoren y_i diskutiert werden. Ist die Entwicklung des nullten Residuenvektors nach den Eigenvektoren

$$r_o = \sum_{i=1}^{n} d_i y_i$$

so überlegt man sich leicht, dass dann beim Gesamtschrittverfahren die Entwicklung des k-ten Residuenvektors nach den Eigenvektoren lautet:

$$r_k = \sum_{i=1}^{n} d_i (1 - \lambda \mu_i)^k y_i \qquad (1.13)$$

Das Verfahren konvergiert also dann und nur dann, wenn

$$0 < \lambda < \frac{2}{\mu_n} \qquad (1.14)$$

wobei μ_n der grösste Eigenwert ist.

§ 2 DAS FRANKEL'SCHE VERFAHREN.

Analog zum Uebergang von der Methode des steilsten Abstieges auf das Gesamtschrittverfahren kann man aus der Methode der konjugierten Gradienten ein Iterationsverfahren mit zwei festen Parametern (Frankel'sches Verfahren) ableiten indem man für λ_k und ε_k in Formeln (1.8) und (1.9) feste Werte wählt. Damit geht allerdings die Orthogonalität der Residuen r_k und das System konjugierter Gewichtsvektoren p_k verloren, sodass in endlich vielen Schritten die Lösung nur approximiert werden kann. Die Vorteile des Verfahrens liegen darin, dass die Berechnung der Skalarprodukte, die bei der Methode der konjugierten Gradienten zur Bestimmung der Parameter notwendig ist, wegfällt und dass die Gewichtsvektoren p_k eliminiert werden können. Diese Eigenschaften ermöglichen, dass der Verlauf der Rechnung auf einfache Weise diskutiert werden kann. Um diese Diskussion durchzuführen, muss zuerst der Gewichtsvektor mit Hilfe der Gleichungen (1.8) und (1.9) eliminiert werden. (1.9) in (1.8) eingesetzt ergibt

$$(\ \lambda_k = \lambda = \text{konst.} \qquad \varepsilon_k = \varepsilon = \text{konst.})$$

$$r_k = r_{k-1} - \lambda D r_{k-1} + \varepsilon \lambda D p_{k-2}$$

und wenn Gleichung (1.8) nochmals benützt wird:

$$(2.1)\quad r_k = (1 + \varepsilon) r_{k-1} - \lambda D r_{k-1} - \varepsilon r_{k-2}$$

Dabei wird wie bei der Methode der konjugierten Gradienten festgesetzt, dass

$$p_o = -r_o \quad \text{und deshalb} \quad r_1 = r_o - \lambda D r_o \quad \text{ist.}$$

Für die Näherungslösung v_k gewinnt man durch Kombination von (1.6) mit (1.9) die Rekursionsformel:

$$(2.2)\quad v_k = v_{k-1} - \lambda r_{k-1} + \varepsilon(v_{k-1} - v_{k-2})$$

$$\text{oder} \quad \Delta v_k = -\lambda r_{k-1} + \varepsilon \Delta v_{k-1}$$

Für $\varepsilon = 0$ erhalten wir die Rekursionsformel (1.12) für das Gesamtschrittverfahren. Das Konvergenzverhalten des Frankel'schen Verfahrens kann auf analoge Weise wie beim Gesamtschrittverfahren diskutiert werden. Wir bezeichnen mit $y_1, y_2 \cdots y_n$ die Eigenvektoren der positiv definiten, symmetrischen Matrix D, denen die positiven, der Grösse nach geordneten, Eigenwerte $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$ zugeordnet sind. Die Residuenvektoren r_k seien wiederum nach den Eigenvektoren entwickelt:

$$r_k = \sum_{i=1}^{n} d_k^i y_i$$

Aus der Rekursionsformel (2.1) folgt dann für die Koeffizierten d_k^i die Gleichung:

$$(2.3)\quad d_k^i - (1 + \varepsilon - \lambda\mu_i)\, d_{k-1}^i + \varepsilon d_{k-2}^i = 0 \qquad (i = 1 \cdots n)$$

Gleichung (2.3) kann mit dem Ansatz erfüllt werden:

$$d_k^i = a^k(\mu_i)$$

Dabei ist a eine Konstante, für die man durch Einsetzen in (2.3) die folgende quadratische Gleichung erhält:

$$(2.4)\quad a^2 - (1 + \varepsilon - \lambda\mu_i)\, a + \varepsilon = 0$$

Diese besitzt die zwei Lösungen

$$(2.5)\quad a_{1/2} = \tfrac{1}{2}\left[1 + \varepsilon - \lambda\mu_i \pm \sqrt{(1 + \varepsilon - \lambda\mu_i)^2 - 4\varepsilon}\right]$$

Damit lautet die allgemeine Lösung der Rekursionsformel (2.1)

$$(2.6)\quad r_k = \sum_{i=1}^{n} \left[b_i a_1^k(\mu_i)\, y_i + c_i a_2^k(\mu_i)\, y_i\right]$$

Die Koeffizienten b_i und c_i lassen sich aus den zwei Anfangsbedingungen bestimmen, die frei vorgebbar sind. Im vorliegenden Falle ist das Residuum r_o, d.h. eigentlich die nullte Versuchslösung v_o, beliebig, während r_1 durch die Annahme

$$(2.7)\quad r_1 = r_o - \lambda D r_o$$

schon festgelegt ist. Die Entwicklung von r_o nach den Eigenvektoren laute:

$$(2.8)\quad r_o = \sum_{i=1}^{n} d_i y_i$$

Aus (2.6) und (2.7) folgt dann für die Koeffizienten b_i und c_i,

$$b_i + c_i = d_i$$

$$a_1 b_i + a_2 c_i = (1 - \lambda\mu_i)\, d_i$$

Nach b_i und c_i aufgelöst, erhält man nach einer trivialen Umformung:

$$b_i = \frac{d_i}{a_1 - a_2}(a_1 - \varepsilon)$$

$$c_i = \frac{d_i}{a_1 - a_2}(\varepsilon - a_2)$$

In (2.6) eingesetzt erhält man, wenn man noch die Beziehung $a_1 . a_2 = \varepsilon$ zur Umformung benützt, für das Residuum r_k die Formel

$$(2.9)\quad r_k = \sum_{i=1}^{n} \frac{d_i y_i}{a_1(\mu_i) - a_2(\mu_i)} \left[(1 - a_2(\mu_i))\, a_1^{k+1}(\mu_i) - (1 - a_1(\mu_i))\, a_2^{k+1}(\mu_i)\right]$$

Da dieser Ausdruck sicher gegen 0 konvergiert, wenn a_1 und a_2 dem Betrage nach kleiner als 1 sind, gilt der Satz:

Satz 1: Für die Konvergenz des Frankel'schen Verfahrens ist es hinreichend, wenn die Ungleichungen

$$(2.10)\quad |a_1(\mu_i)| < 1 \quad \text{und} \quad |a_2(\mu_i)| < 1$$

für alle Eigenwerte μ_i erfüllt sind.

Dann konvergiert nämlich jeder Term der Summe einzeln mit zunehmendem Index k gegen Null.
Aus Gleichung (2.5) folgt, dass je nach der Wahl von λ und ε die Wurzeln $a_1(\mu_i)$, $a_2(\mu_i)$ reell oder komplex sein können. Die Grenze zwischen den reellen und komplexen Wurzeln ist durch das Verschwinden der Diskriminante der quadratischen Gleichung (2.4) gekennzeichnet, d.h. λ und ε müssen dann der Gleichung genügen:

$$(2.11)\quad (1 + \varepsilon - \lambda\mu_i)^2 = 4\varepsilon$$

Geometrisch interpretiert stellt diese Gleichung für jeden Eigenwert μ_i eine Parabel dar, falls λ und ε die kartesischen Koordinaten eines Punktes in einer Ebene sind. Eine eingehendere Disskussion von Gleichung (2.11) ergibt das folgende Bild

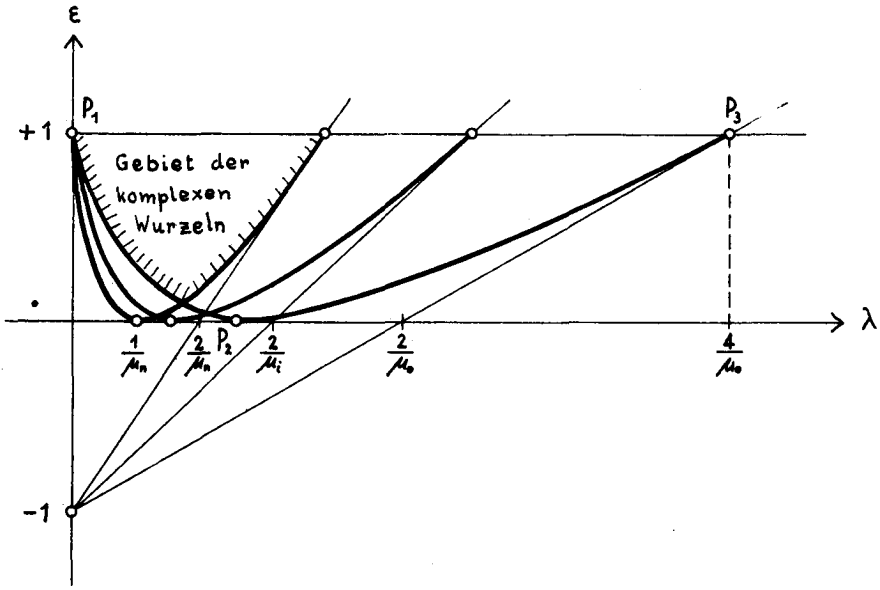


Fig. 1

Die zu einem Eigenwert μ_i gehörigen Parabeln tangieren sich untereinander alle im gleichen Punkte P_1 auf der ε-Achse und berühren zudem noch die λ-Achse.

Wegen $a_1 . a_2 = \varepsilon$ ist eine der beiden Wurzeln a_1, a_2 dem Betrage nach grösser oder gleich $\sqrt{|\varepsilon|}$. Sie sei fortan mit a_1 bezeichnet. Für die genauere Untersuchung des Konvergenzverhaltens des Frankel'schen Verfahrens erweist es sich als zweckmässig, Gleichung (2.9) durch Elimination der dem Betrage nach kleineren Wurzel a_2 auf die Form zu bringen

$$(2.12)\quad r_k = \sum_{i=1}^{n} d_i y_i \, a_1^k \left[(1-a_1) \sum_{l=1}^{k} \left(\frac{\varepsilon}{a_1^2} \right)^l + 1 \right]$$

wobei nach Voraussetzung und wegen der Konvergenzbedingung (2.10) $|\varepsilon| \leq |a_1|^2 \leq 1$ sein muss.

Die Konvergenzgeschwindigkeit, mit der der Anteil der i-ten Eigenfunktion y_i eliminiert wird, ist also durch den Quotienten ergeben:

$$(2.13)\quad \frac{q_{k+1}}{q_k} = a_1 + \frac{a_1 (1-a_1) \left(\frac{\varepsilon}{a_1^2} \right)^{k+1}}{(1-a_1) \sum_{l=1}^{k} \left(\frac{\varepsilon}{a_1^2} \right)^l + 1} \approx a_1 (\mu_i)$$

wobei $q_k = \left[(1-a_1) \sum_{l=1}^{k} \left(\frac{\varepsilon}{a_1^2} \right)^l + 1 \right] a_1^k$ ist.

Der Nenner wird mit wachsendem k sehr gross, der Zähler hingegen nimmt wegen der Voraussetzung $|a_1| > \sqrt{|\varepsilon|}$ sicher nicht zu, deshalb ist für grosse Indizes k die Konvergenzgeschwindigkeit praktisch durch $a_1 (\mu_i)$ allein gegeben. Da die Grösse der $a_1 (\mu_i)$ von der Wahl von λ und ε abhängen, wollen wir als Nächstes die Parameterwerte λ, ε bestimmen, für die $a_1 (\mu_i)$ den Betrag η hat, wo η eine gegebene positive Zahl ist. Dieses Problem werden wir für die komplexen und reellen Wurzeln getrennt behandeln:

a) komplexe Wurzeln:

Da die Koeffizienten der quadratischen Gleichung (2.4) reell sind, ist $a_1 (\mu_i)$ konjugiert komplex zu $a_2 (\mu_i)$ und der Betrag

$$|a_1|^2 = a_1 . a_2 = \varepsilon$$

Für alle Punkte, die im Gebiete der komplexen Wurzeln auf der Geraden

$$(2.14)\quad \varepsilon = \eta^2$$

liegen, haben also die Wurzeln $a_1 (\mu_i)$ den gleichen Betrag.

b) reelle Wurzeln:

Hier ist a_1 entweder

$$a_1 = +\eta \qquad \text{oder} \qquad a_1 = -\eta$$

Nach den Vieta'schen Wurzelsätzen folgen dann für λ und ε die Beziehungen:

$$(2.15) \quad \varepsilon = \frac{\eta\mu_i}{\eta - 1}\lambda + \eta \qquad \text{für } a_1 = +\eta$$

$$(2.16) \quad \varepsilon = \frac{\eta\mu_i}{\eta + 1}\lambda - \eta \qquad \text{für } a_1 = -\eta$$

Geometrisch stellen diese beiden Gleichungen zwei Geraden dar, die die zum entsprechenden Eigenwert gehörende Parabel (2.11), in den Schnittpunkten dieser Kurve mit der Geraden $\varepsilon = \eta^2$ tangieren. Weil $|a_1| > \sqrt{|\varepsilon|}$ darf $|\varepsilon|$ auch im Gebiet der reellen Wurzeln nicht grösser als η^2 werden. Fassen wir diese Resultate mit denen für das Gebiet der komplexen Wurzeln zusammen, so folgt der Satz:

Satz 2: Auf dem durch die Geraden (2.14), (2.15) und (2.16) gegebenen Dreiecksrand wird der Anteil des i-ten Eigenvektors im Ausgangsresiduenvektor r_0 bei allen Verfahren mit konstanten Parametern gleich rasch eliminiert.

Aus der Kombination der Sätze 1 und 2 geht hervor, dass man die Grenzen des Gebietes, in dem sicher Konvergenz herrscht, erhält, falls man in den Geradengleichungen (2.14) bis (2.16) $\eta = 1$ und $\mu_i = \mu_n$, dem grössten Eigenwert, setzt, da das von diesen Geraden gebildete Dreieck in allen zum Wert $\eta = 1$ und den übrigen Eigenwerten gehörigen Dreiecken enthalten ist. Wir halten dieses Resultat in dem Satz fest:

Satz 3: Das Frankel'sche Verfahren konvergiert innerhalb des Dreieckes, das durch die drei Geraden

$$(2.17) \quad \begin{aligned} \lambda &= 0 \\ \varepsilon &= \frac{\mu_n}{2}\lambda - 1 \\ \varepsilon &= 1 \end{aligned}$$

gegeben ist, wo μ_n der grösste Eigenwert der Gleichungsmatrix D ist.

Die Konvergenzbedingung (1.14) für das Gesamtschrittverfahren ist in diesem Satz enthalten, da dieses aus dem Frankel'schen Verfahren einfach durch Nullsetzen von ε abgeleitet werden kann.
Aus Satz 2 ergibt sich auch eine Aussage über die Konvergenzgeschwindigkeit des Gesamtschrittverfahrens im Vergleich zu den Frankel'schen Verfahren:

Satz 4: Wünscht man bei einem Gesamtschrittverfahren mit gegebenem Parameter λ eine Konvergenzbeschleunigung durch ein Frankel'sches Verfahren mit demselben Wert λ, so muss der zweite Parameter ε positiv gewählt werden.

Dieser Satz kann leicht anschaulich bewiesen werden, falls man für verschiedene Werte η und Eigenwerte μ_i die Dreiecke von Satz 2 graphisch darstellt, wie das in Fig. 2 getan wurde

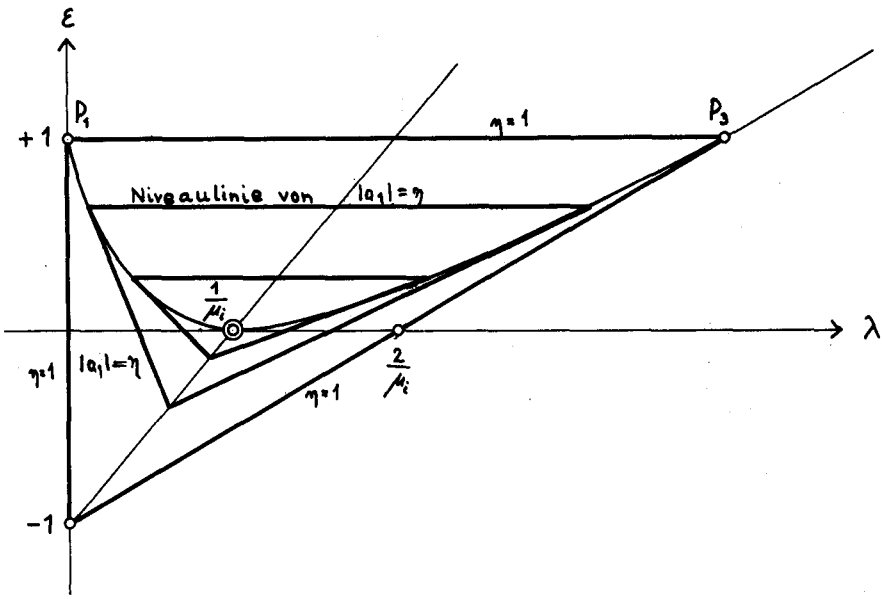


Fig. 2

Aus dieser Figur ist ersichtlich, dass für einen bestimmten Eigenwert μ_i die Dreiecke mit abnehmendem Betrag η konzentrisch auf den Punkt $(\lambda = \frac{1}{\mu_i}, \varepsilon = 0)$ zusammenschrumpfen. Will man also von einem Punkt der λ-Achse aus, der einem bestimmten Gesamtschrittverfahren mit einem Konvergenzfaktor η entspricht, parallel zur ε- Achse zu Niveaulinien mit kleineren η gelangen, so kann das nur in Richtung der positiven ε erreicht werden.
Aus Satz 2 kann weiterhin noch gefolgert werden, dass für diejenigen Punkte der obern (λ, ε) -Halbebene, die allen zu den n Eigenwerten und einem festen Betrag η gehörenden Dreiecksrändern gemeinsam sind, gleich rasche Konvergenz der Anteile aller Eigenfunktionen am Anfangsresiduenvektor r_o besteht. Da für eine bestimmte Grösse von η die zum kleinsten und grössten Eigenwert gehörenden Dreiecke am stärksten gegeneinander verschoben liegen, weist das Frankel'sche Verfahren für denjenigen Punkt die beste allgemeine Konvergenz auf, der gerade gemeinsamer Eckpunkt zweier solcher Dreiecke ist. Nach den Ueberlegungen zu Satz 2 liegen in der obern Halbebene alle Eckpunkte der zu einem Eigenwert μ_i gehörenden Dreiecke auf der durch (2.11) gegebenen Parabel, sodass der gesuchte Eckpunkt der Schnittpunkt mit der kleinsten Ordinate der zum grössten und kleinsten Eigenwert gehörenden Parabeln ist:

Satz 5: Das Frankel'sche Verfahren besitzt die beste allgemeine Konvergenz für die Parameterwerte

$$(2.18)\quad \lambda_s = \frac{4}{(\sqrt{\mu_n} + \sqrt{\mu_1})^2}\ ;\ \varepsilon_s = \left[\frac{\sqrt{\mu_n} - \sqrt{\mu_1}}{\sqrt{\mu_n} + \sqrt{\mu_1}}\right]^2\ ,\ a_1 = a_2 = \varepsilon_s$$

Für diese Wahl der Parameterwerte lautet Gleichung (2.9), falls wir die komplexen Wurzeln $a_1(\mu_i)$, $a_2(\mu_i)$ $(i \neq 1, n)$ zu reellen Grössen zusammenfassen:

$$(2.19)\quad r_k = (1 - \frac{2\sqrt{\mu_1}}{\sqrt{\mu_1} + \sqrt{\mu_n}})^k \Big[\sum_{j=1}^{k-1} d_j y_j \frac{\sqrt{\mu_1 \mu_n} - \mu_j}{\sqrt{(\mu_j - \mu_1)(\mu_n - \mu_j)}} \sin k\varphi_j + \cos k\varphi_j \Big]$$
$$+ d_1 y_1 \left(2k \frac{\sqrt{\mu_1}}{\sqrt{\mu_1} + \sqrt{\mu_n}} + 1 \right) + (-1)^k d_n y_n \left(2k \frac{\sqrt{\mu_n}}{\sqrt{\mu_1} + \sqrt{\mu_n}} + 1 \right) \Big]$$

wobei $\cos \varphi_j = 1 - 2 \frac{\mu_j - \mu_1}{\mu_n - \mu_1}$

Liegen die zum grössten und kleinsten Eigenwert benachbarten μ_i nicht sehr nahe bei diesen, so zeigt Formel (2.19), dass in diesem Falle die Koeffizienten des ersten und n-ten Eigenvektors in der Entwicklung des Residuenvektors r_k mit zunehmendem k immer grösser werden relativ zu den übrigen Koeffizienten. Für grosse k wird der Residuenvektor also im Wesentlichen nur noch die erwähnten beiden Eigenvektoren enthalten. Dieses Verhalten ist analog zu demjenigen des Gesamtschrittverfahrens, bei dem λ den Wert hat:

$$\lambda = \frac{2}{\mu_1 + \mu_n}$$

Dieses Verfahren führt auch auf eine Residuenverteilung, die praktisch aus einer Superposition des ersten und n-ten Eigenvektors besteht. Zum Vergleich der beiden Methoden wollen wir die Zahl n der Iterationen berechnen, die man braucht, um den Betrag des Residuenvektors r_k mindestens um einen Faktor 10^{-N} zu verkleinern. Nach Formel (2.13), die für beide Methoden gilt, muss also für das absolut grösste a, die folgende Ungleichung erfüllt sein.

$$a_1^{\,n} \leq 10^{-N}$$

oder falls wir von beiden Seiten den Logarithmus nehmen und nach n auflösen:

$$n \geq \frac{N \log 10}{\log \frac{1}{a_1}}$$

Im Falle des Gesamtschrittverfahrens ($\varepsilon = 0$) folgt, falls $M = \frac{\mu_1}{\mu_n}$ klein gegen 1 ist, mit Hilfe der Formel (2.5) und einigen Umformungen

$$(2.20) \qquad n \geq \frac{N \log 10}{2M} \quad (1 - M)$$

Für das Frankel'sche Verfahren hingegen erhält man

$$(2.21) \qquad n \geq \frac{N \log 10}{2\sqrt{M}} \quad (1 - \sqrt{M})$$

Da in diesem Falle $\sqrt{M}$ statt nur M im Nenner auftritt, konvergiert das Frankel'sche Verfahren besonders bei starker Streuung der Eigenwerte wesentlich besser als das Gesamtschrittverfahren.

Bei der bisherigen Diskussion des Konvergenzverhaltens wurde meist vorausgesetzt, dass die Zahl der durchgeführten Iterationen gross sei, weshalb vom anfänglichen Verlauf der Folge der Residuenvektoren abgesehen werden könne. Die bisherigen Resultate schliessen deshalb die Möglichkeit nicht aus, dass für andere Parameterwerte als diejenigen von Satz 5 die Beträge der Residuenvektoren anfänglich rascher abnehmen. In dieser Hinsicht sind speziell die Parameterwerte von Interesse, für welche die Wurzeln $a_1(\mu_i)$, $a_2(\mu_i)$ komplex sind, da dort, falls man die konjugiert komplexen Wurzeln zusammenfasst, der Ausdruck (2.9) für die Residuenvektoren die Form hat:

$$(2.22) \quad r_k = 2\,\varepsilon^{k/2} \sum_{j=1}^{n} \sqrt{\frac{\mu_j}{4\varepsilon - (1 + \varepsilon - \lambda\mu_j)^2}} \sin(k\varphi_j + \theta_j)\, d_j y_j$$

wobei $\cos\varphi_j = \frac{1}{2\sqrt{\varepsilon}}(1 + \varepsilon - \lambda\mu_j)$; $\cos\theta_j = \frac{1 - \varepsilon - \lambda\mu_j}{2\sqrt{\varepsilon\lambda\mu_j}}$

Aus Formel (2.22) ist ersichtlich, dass unter gewissen Bedingungen die Folge der Residuenvektoren wegen des Auftretens von trigonometrischen Funktionen anfänglich rascher konvergieren kann, als es der die allgemeine Konvergenz bestimmende Faktor $\varepsilon^{k/2}$ vermuten lässt. Der Schwingungscharakter der Koeffizienten der Eigenvektoren kann allerdings nur dann auch im Residuenvektor r_k selbst zum Ausdruck kommen, wenn im Anfangsresiduenvektor r_0 nur solche Eigenvektoren vertreten sind, deren zugehörigen Eigenwerte sehr nahe beieinander liegen. Zudem muss ε und λ so bestimmt werden können, dass für alle im Residuenvektor vertretenen Eigenvektoren y_j

$$(2.23) \qquad \varphi_j \cong \frac{\pi - \theta_j}{m} \qquad \text{wo m eine möglichst kleine ganze Zahl ist.}$$

Diese Bedingungen verlangen also die Kenntnis zum mindesten der Eigenwerte der Gleichungsmatrix D. Da die Bestimmung aller Eigenwerte mit einem erheblichen Arbeitsaufwand verbunden ist, wird man den Schwingungscharakter der Koeffizienten in der Entwicklung der Residuenvektoren nach Eigenvektoren

praktisch nur in denjenigen Fällen benützen, wo alle Eigenwerte nahe beieinander liegen. Um entscheiden zu können, ob ein solcher Fall vorliegt, müssen dann nur der grösste und kleinste Eigenwert bestimmt werden, die auch bei Anwendung des Satzes 5 benötigt werden.

Abschätzungen für diese beiden Eigenwerte können mit Hilfe der Gerschgorin'schen Sätze *) erhalten werden, oft geben auch physikalische Ueberlegungen einigen Aufschluss über die Lage dieser Werte. Sind diese Abschätzungen zu ungenau, so kann man eine der Methoden zur iterativen Bestimmung von Eigenwerten **) zur Verbesserung benützen. In manchen Fällen wird man am einfachsten Informationen für die geeignete Wahl der Parameter λ und ε gewinnen, wenn man zuerst einige Iterationen der Methode der konjugierten Gradienten ausführt. Die Rayleigh'schen Quotienten $\sigma_{2,k}$ der dabei berechneten Residuenvektoren, sowie die Werte für die Paramter λ_k (siehe Formel (1.5)),können bei nicht zu spezieller Wahl der ersten Versuchslösung einen, wenn vielleicht auch sehr ungenauen, Aufschluss über die Lage der beiden extremalen Eigenwerte geben, da die Ungleichungen gelten ***)

$$(2.24) \qquad \begin{aligned} &\mu_1 < \sigma_{2k} < \mu_n \qquad \left(\sigma_{2k} = \frac{(r_k, Dr_k)}{(r_k, r_k)}\right) \\ &\mu_1 < \frac{1}{\lambda_k} < \mu_n \end{aligned}$$

Das genauere Vorgehen bei diesem Verfahren werden wir an Hand eines Beispieles im § 6 demonstrieren. Das Frankel'sche Verfahren wird also in diesem Falle zur Abkürzung der Methode der konjugierten Gradienten verwendet. In dieser Kombination wird es sich wahrscheinlich am nützlichsten erweisen.

Zusammenfassend besteht also das Vorgehen beim Frankel'schen Verfahren aus den folgenden Punkten:

1) Approximative Bestimmung des grössten und kleinsten Eigenwertes μ_n, μ_1
2) Wahl eines Versuchsvektors v_o
3) Berechnung von λ und ε nach Formel (2.19) eventuell Korrektur dieser Werte falls $\frac{\mu_n}{\mu_1}$ nahe bei 1 so, dass die Beziehungen (2.22) und (2.23) gelten.
4) Berechnung des ersten Schrittes $v_1 = v_o - \lambda (Dv_o + l)$

* Gerschgorin: Abgrenzung der Eigenwerte einer Matrix, Bull. Acad.Sciences de l'URSS., Classe mathémat. 7-c série, 1931

** Siehe Collatz: Eigenwertprobleme und ihre numerische Behandlung

*** Siehe dazu die in Fussnote 3) der Einleitung zitierte Arbeit.

5) Durchführung des Iterationsverfahrens. In der k-ten Iteration wird mit Hilfe der aus den beiden vorhergehenden Iterationen bekannten Approximationen v_{k-1} und v_{k-2} zum Lösungsvektor eine neue Approximation v_k nach folgendem Rechenschema gefunden:

a) Berechnung von $r_{k-1} = Dv_{k-1} + l$

b) $v_k = (1 + \varepsilon) v_{k-1} - \lambda r_{k-1} - \varepsilon v_{k-2}$

Sobald der Betrag von r_{k-1} sehr klein ist, kann die Rechnung abgebrochen werden. Im § 4 dieser Arbeit haben wir als kleine Illustration die Anwendung des Frankel'schen Verfahrens zur Lösung einer Dirichlet'schen Randwertaufgabe aufgeführt.

§ 3. DIE ANALOGIE ZWISCHEN DEN GRADIENTENVERFAHREN UND GEWISSEN ANFANGSWERTPROBLEMEN.

Im folgenden betrachten wir im Falle von linearen Randwertaufgaben analoge Probleme zum Frankel'schen- und Gesamtschritt-Verfahren, wo an Stelle der diskontinuierlich fortschreitenden Schrittnummer die kontinuierliche Variable Zeit und an Stelle des Differenzengleichungssystemes das entsprechende Randwertproblem im Kontinuierlichen tritt. Mit Hilfe dieser analogen Probleme lassen sich einige besondere Eigenheiten der beiden Verfahren verdeutlichen, da die Ergebnisse für das Kontinuierliche unter gewissen Voraussetzungen eine Approximation des analogen diskontinuierlichen Falles darstellen.
Als Analogon zum Gesamtschrittverfahren kann das Wärmeleitungsproblem angesehen werden, das durch die Differentialgleichung:

$$(3.1)\quad \frac{\partial v(z,t)}{\partial t} = a\,(\Delta\, v(z,t) + g(z))$$

und die Anfangs- und Randbedingungen:

$$(3.2)\quad v(z,t=0) = v_o(z)$$

$$v(z=z_o,t) = c_o \;;\; v(z=z_n,t) = c_n$$

gegeben ist. Dabei ist v(z,t) eine Funktion der Zeit und je nach Problem einer oder mehrerer Ortskoordinaten z, Δ der dem Operator des Differenzengleichungssystemes entsprechende Operator im Kontinuierlichen und g(z) eine zeitunabhängige Funktion. Approximiert man nämlich die Differentialgleichung durch Einführung von Differenzenquotienten in einem Differenzennetz mit endlicher Maschenweite, so entsteht das folgende Differenzengleichungssystem:

$$(3.3)\quad \frac{1}{h}(v_{k+1,i} - v_{k,i}) = \frac{a}{s^2h^2} Dv_{k,i} + al_k \quad \text{für } k > 0$$

wenn h die Maschenweite in Zeitrichtung, sh die Maschenweite für die Ortskoordinate, $v_{k,i} = v(z=ksh, t=ih)$, $l_k = g(z=ksh)$ und D der Differenzenoperator ist, der dem Differentialoperator Δ unter Berücksichtigung der Randbedingungen entspricht. Damit die Lösung der Differenzengleichungen gegen die Lösung der Differentialgleichung konvergiert, muss bekanntlich das Verhältnis s der Maschenweiten in Orts- und Zeitrichtung die Ungleichung erfüllen:

$$(3.4)\quad \frac{1}{s^2h} < \frac{2}{\mu_n} \quad \text{wo } \mu_n \text{ der grösste Eigenwert von D}$$

Die Differenzengleichungen (3.3) stimmen mit den Rekursionsformeln (1.12) überein, falls man formal setzt:

$$(3.5)\quad \lambda = \frac{a}{s^2h} \;;\; l = -s^2h\{\, l_i \,\}$$

wenn $\{l_i\}$ ein Vektor mit den l_i als Komponenten bedeutet.

Da bei Beachtung der Bedingung (3.4) für genügend kleine h die Lösung des Differenzengleichungssystems die Lösung der Differentialgleichung approximiert, gibt umgekehrt die Lösung der Differentialgleichung das Verhalten der Lösung des Differenzengleichungssystemes (3.3) und damit auch des Gesamtschrittverfahrens approximativ wieder. Speziell kann man auf diese Weise einigen Aufschluss über das Verhalten der Residuenvektoren bekommen. Wir wollen dies an einem konkreten Beispiel zeigen:
Als lineares Randwertproblem nehmen wir das folgende Potentialproblem im Eindimensionalen:

$$(3.6) \qquad \frac{\partial^2 u}{\partial z^2} + \delta(z-c) = 0$$

$$u = 0 \text{ für } z = 0$$

und $$u = \text{endlich für } z = \infty$$

Dabei ist $\delta(z-c)$ die singuläre Deltafunktion von Dirac. Dann lautet das dem Gesamtschrittverfahren für dieses Randwertproblem analoge Anfangswertproblem:

$$(3.7) \qquad \frac{\partial v}{\partial t} = a\left(\frac{\partial^2 v}{\partial z^2} + \delta(z-c)\right)$$

wobei die Randbedingungen dieselben wie für u sind. Als Anfangsbedingungen wählen wir

$$v(t=0) = 0$$

Dieses Anfangswertproblem kann auf einfache Weise mit Hilfe der Laplacetransformation gelöst werden *. Transformieren wir die Differentialgleichung (3.7) nach Laplace, so erhalten wir im Bildgebiet die gewöhnliche Differentialgleichung:

$$(3.8) \qquad \frac{d^2 V}{dz^2} - \frac{p}{a} V = -\frac{\delta(z-c)}{p}$$

für die Transformierte $V(p,z) = \int_0^\infty e^{-pt}\, v(z,t)\, dt$

Die Lösung dieser inhomogenen Differentialgleichung lautet, wenn man die Randbedingungen berücksichtigt:

$$V(p,z) = \begin{cases} \frac{1}{2p}\sqrt{\frac{a}{p}}\left(e^{-\sqrt{\frac{p}{a}}(c-z)} - e^{-\sqrt{\frac{p}{a}}(c+z)}\right); & z < c \\ \frac{1}{2p}\sqrt{\frac{a}{p}}\left(e^{-\sqrt{\frac{p}{a}}(z-c)} - e^{-\sqrt{\frac{p}{a}}(c+z)}\right); & z > c \end{cases}$$

*) Für nähere Einzelheiten siehe z.B. Doetsch, Tabellen zur Laplace-Transformation, Springer 1947

Die Transformation dieser Lösung ins Originalgebiet ergibt die Lösung der Differentialgleichung (3.6):

$$v(z,t) = \frac{1}{\sqrt{\pi}}\left\{\sqrt{\frac{t}{a}}\left[e^{-\frac{(c-z)^2}{4at}} - e^{-\frac{(c+z)^2}{4at}}\right] - (c-z)\int_{\frac{c-z}{2\sqrt{at}}}^{\frac{c+z}{2\sqrt{at}}} e^{-\xi^2}\,d\xi + \right.$$

(3.9)

$$\left. + 2z\int_{\frac{c+z}{2\sqrt{at}}}^{\infty} e^{-\xi^2}\,d\xi\right)\Big\};\ z < c$$

$$v(z,t) = \frac{1}{\sqrt{\pi}}\left\{\sqrt{\frac{t}{a}}\left[e^{-\frac{(c-z)^2}{4at}} - e^{-\frac{(c+z)^2}{4at}}\right] - (z-c)\int_{\frac{z-c}{2\sqrt{at}}}^{\frac{z+c}{2\sqrt{at}}} e^{-\xi^2}\,d\xi + 2a\int_{\frac{z+c}{2\sqrt{at}}}^{\infty} e^{-\xi^2}\,d\xi\right\};\ z > c$$

Wenn die Zeit t gegen Unendlich strebt, muss v(z,t) gegen die Lösung des Randwertproblemes konvergieren. Nach Durchführung des Grenzüberganges erhält man:

$$v(z,\infty) = z \quad z < c$$

$$v(z,\infty) = c \quad z > c$$

d.h. die Green'sche Funktion für das Potentialproblem, wie es auch nicht anders zu erwarten war.

Setzt man beim Gesamtschrittverfahren eine Näherungslösung v_k mit endlichem Index k in das Differenzengleichungssystem ein, so wird man entsprechend der Definition (1.3) im allgemeinen einen vom Nullvektor verschiedenen Residuenvektor r_k erhalten. Analog dazu wird, wenn man die Lösung v(z,t) des Anfangswertproblemes für eine bestimmte endliche Zeit t in die Differentialgleichung (3.6) des linearen Randwertproblemes einsetzt, im allgemeinen ein von Null verschiedener Wert herauskommen, der von der gewählten Zeit t abhängt. Wir bezeichnen die Gesamtheit dieser Werte als Residuenfunktion.

$$(3.10)\quad r(z,t) = \frac{\partial^2 v(z,t)}{\partial z^2} + \delta(z-c)$$

Setzt man v(z,t) ein, so folgt:

$$(3.11)\quad r(z,t) = \frac{1}{z\sqrt{a\pi}} \cdot \frac{1}{\sqrt{t}}\left[e^{-\frac{(z-c)^2}{4at}} - e^{-\frac{(z+c)^2}{4at}}\right]$$

Dieser Formel entspricht beim Gesamtschrittverfahren die Ausbreitung eines einzelnen Residuums. Aus ihr ersieht man, dass die am Anfang vorhandene Singularität im Punkte c für endliche Zeiten durch eine mit der Zeit sich verflachende Glockenkurve ersetzt wird, die jedoch relativ nur langsam gegen die z Achse konvergiert. In Figur 3 ist für einige Zeiten die Form von r(z,t) festgehalten.

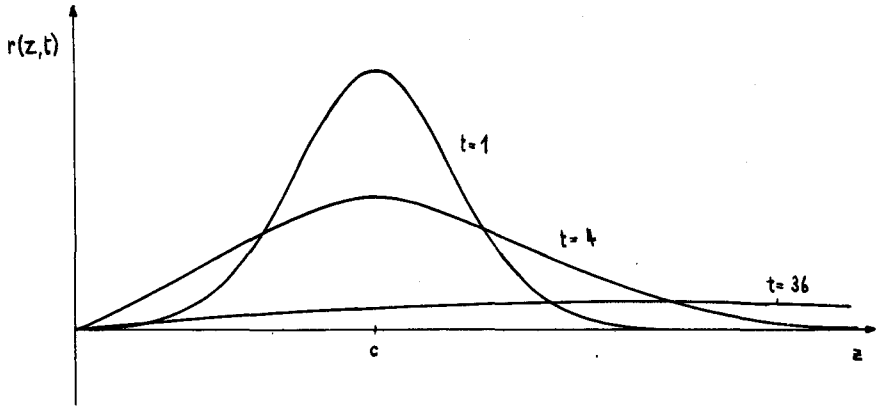


Fig. 3

Die relativ langsame Konvergenz ist analog zur sog. "Käfigbildung" beim Gesamtschrittverfahren, d.h. dem Auftreten einer Residuenverteilung, die proportional zur ersten Eigenfunktion ist und deren Betrag von Iteration zu Iteration nur wenig abnimmt.
Beim Frankel'schen Verfahren kann das folgende Anfangswertproblem als Analogon im Kontinuierlichen betrachtet werden:
Gesucht die Lösung v(z,t) der partiellen Differentialgleichung:

$$(3.12)\quad \frac{\partial^2 v}{\partial t^2} + b\,\frac{\partial v}{\partial t} = a(\Delta v + g(z))$$

für die Anfangsbedingungen:

$$(3.13)\quad \left.\frac{\partial v}{\partial t}\right|_{t=0} = 0 \;;\; v(z,t=0) = v_0(z)$$

und die Randbedingungen

$$v(z = z_o, t) = c_o, \quad v(z = z_n, t) = c_n$$

a und b sind Konstanten, b sei positiv (der Ausbreitungsvorgang soll gedämpft sein). Im übrigen sind die verwendeten Bezeichnungen die gleichen, wie beim Wärmeleitungsproblem. Die partielle Differentialgleichung (3.12) tritt auch bei der Beschreibung der Ausbreitung elektrischer Signale in Kabeln ohne Leitungsverlust auf. Approximiert man die Differentialgleichung und die Anfangsbedingung (3.13) mit Hilfe von Differenzenquotienten entsprechend dem Vorgehen bei der Wärmeleitungsgleichung, so erhält man das folgende Differenzengleichungssystem:

$$\frac{1}{h^2}(v_{k+1,i} - 2v_{k,i} + v_{k-1,i}) + \frac{b}{2h}(v_{k+1,i} - v_{k-1,i}) = -\frac{a}{s^2 h^2} Dv_{k,i} + a l_k$$

(3.14)

$$v_{o,i} - v_{-1,i} = 0 \quad , \quad v_{o,i} = v_o \text{ (ish)}$$

Die Bezeichnungen sind die gleichen, wie für die Formel (3.3). Die Zuordnung zwischen Differenzengleichungssystem und Differentialgleichung ist in diesem Falle nicht ganz eindeutig, da bei Benützung von drei in der Zeitrichtung unmittelbar benachbarten Funktionswerten die Ableitung $\frac{\partial v}{\partial t}$ auf verschiedene Weisen durch Differenzen von diesen Funktionswerten dargestellt werden kann. Die Differenzengleichungen (3.14) sind jedoch insofern ausgezeichnet als sie in dieser Näherung die Differentialgleichung im Mittel am besten approximieren.
Man muss auch hier beachten, dass die Lösung der Differenzengleichung nicht immer mit abnehmender Maschenweite des Differenzennetzes gegen die Lösung der Differentialgleichung konvergiert. Für den Fall, dass Δ der Laplace'sche Operator ist, wurde die Konvergenz schon in einer Arbeit von Courant, Friedrichs und Lewy * untersucht und die folgende Konvergenzbedingung gefunden:

$$(3.15) \quad a < s^2$$

Vergleicht man die Differenzengleichungen (3.14) mit der Rekursionsformel (2.2) so stimmen die Gleichungen überein, falls man formal setzt

$$(3.16) \quad \varepsilon = \frac{2 - bh}{2 + bh} \quad ; \quad \lambda = \frac{2a}{s^2(2 + bh)} \quad l = -s^2 h \{l_i\}$$

Die die Anfangsbedingungen (3.13) approximierenden Differenzengleichungen entsprechen dem Start des Frankel'schen Verfahren.

* Courant, Friedrichs, Lewy, Math.Ann. 100, 32, 1928

Die Forderung (3.15) lautet dann mit Hilfe von λ und ε ausgedrückt:

$$\lambda < \frac{1}{2} (1 + \varepsilon)$$

Da die Eigenwerte μ_i des dem Laplace'schen Operator entsprechenden Differenzenoperators D nicht grösser als 4 sein dürfen, weil mit μ_i auch $4 - \mu_i$ ein Eigenwert sein muss und alle μ_i positiv sind, liegen die durch diese Ungleichung abgegrenzten λ-Werte nach (2.17) innerhalb des Konvergenzgebietes des Frankel'schen Verfahrens. Da zudem vorausgesetzt wurde, dass b positiv ist, ist $\varepsilon < 1$, sodass nach (2.17) das zum Anfangswertproblem analoge Frankel'sche Verfahren konvergent ist.

Es ist aufschlussreich, das dem Frankel'schen Verfahren entsprechende Anfangswertproblem für dasselbe Potentialproblem wie im vorhergehenden Beispiel zu lösen, d.h. eine Lösung v(z,t) zu bestimmen, die die partielle Differentialgleichung:

$$(3.17)\quad \frac{\partial^2 v}{\partial t^2} + b \frac{\partial v}{\partial t} = a \left(\frac{\partial^2 v}{\partial z^2} + \delta(z-c) \right)$$

und die Anfangsbedingungen:

$$(3.18)\quad \left.\frac{\partial v}{\partial t}\right|_{t=0} = 0 \quad , \quad \left. v \right|_{t=0} = 0$$

sowie die Randbedingungen:

$$v(z=0,t) = 0 \quad ; \quad v(z=\infty, t) = \text{endlich}$$

erfüllt.

Zur Lösung des Anfangswertproblemes verwenden wir dieselben Methoden wie beim ersten Beispiel dieses Paragraphen, weshalb wir auf die Wiedergabe der einzelnen Stationen des Lösungsweges verzichten und gerade das Schlussresultat für v(z,t) angeben, falls $b > 0$ ist.

$$v(z,t) = 0 \quad , \quad \sqrt{a}t < |z-c|$$

$$(3.19)\quad v(z,t) = \frac{\sqrt{a}}{2} \int_{\frac{|z-c|}{\sqrt{a}}}^{t} e^{-\frac{b}{2}\tau} I_o\left(\frac{b}{2\sqrt{a}} \sqrt{a\tau^2 - (z-c)^2}\right) d\tau \quad ; \quad |z-c| < \sqrt{a}t < z + c$$

$$v(z,t) = \frac{\sqrt{a}}{2} \left\{ \int_{\frac{|z-c|}{\sqrt{a}}}^{t} e^{-\frac{b}{2}\tau} I_o\left(\frac{b}{2\sqrt{a}} \sqrt{a\tau^2 - (z-c)^2}\right) d\tau - \int_{\frac{z+c}{\sqrt{a}}}^{t} e^{-\frac{b}{2}\tau} I_o\left(\frac{b}{2\sqrt{a}} \sqrt{a\tau^2 - (z+c)^2}\right) \right\};$$

$$\text{für } z + c < \sqrt{a}t$$

wobei $I_o(\xi)$ die nullte Besselfunktion mit imaginären Argument ist.

Sie wird zum Beispiel durch die Integraldarstellung

$$I_n(\xi) = \frac{(-i)^n}{\pi} \int_0^{\pi} e^{-\xi \cos\varphi} \cos n\varphi \, d\varphi$$

für $n = 0$ gegeben.

Die Lösung hat offensichtlich den Charakter eines Ausbreitungsvorganges, wobei $\sqrt{a}$ die Ausbreitungsgeschwindigkeit gibt. Auch hier können wir in Analogie zum Residuenvektor im Diskontinuierlichen eine Residuenfunktion $r(z,t)$ im Kontinuierlichen definieren. Da das Randwertproblem in diesem Beispiel dasselbe ist, wie bei der Wärmeleitungsaufgabe, fällt die Definitionsgleichung für $r(z,t)$ mit (3.10) zusammen. Setzt man $v(z,t)$ ein, so sieht man folgendes:
Die Residuenfunktion hat je nachdem, ob der linke Randpunkt vom Ausbreitungsvorgang schon erreicht wurde, oder nicht, verschiedene Gestalt:

wenn $\sqrt{a}t < c$

$$r(z,t) = \frac{\sqrt{a}}{2} e^{-\frac{b}{2}t} \delta\left(t - \frac{|c-z|}{\sqrt{a}}\right) \quad ; \text{ für } \sqrt{a}t \leq |z-c|$$

$$(3.20) \quad r(z,t) = \frac{b}{4\sqrt{a}} e^{-\frac{b}{2}t} \left\{ I_0\left(\frac{b}{2}\sqrt{t^2 - \frac{(z-c)^2}{a}}\right) + \frac{\sqrt{a}t}{\sqrt{t^2 - \frac{(z-c)^2}{a}}} I_1\left(\frac{b}{2}\sqrt{t^2 - \frac{(z-c)^2}{a}}\right) \right\}$$

für $|c-z| < \sqrt{a}\, t$

wobei $I_0(\xi)$, bzw. $I_1(\xi)$ die nullte, bzw. erste Besselfunktion mit imaginärem Argument ist.

Für $\sqrt{a}\, t > c$ (Nach der Reflexion am linken Rande)

$$(3.21) \quad r(z,t) = \frac{b}{4\sqrt{a}} e^{-\frac{b}{2}t} \left\{ I_0\left(\frac{b}{2}\sqrt{t^2 - \frac{(z-c)^2}{a}}\right) + \frac{t}{\sqrt{t^2 - \frac{(z-c)^2}{a}}} I_1\left(\frac{b}{2}\sqrt{t^2 - \frac{(z-c)^2}{a}}\right) \right\} -$$

$$- \frac{\sqrt{a}}{2} e^{-\frac{b}{2}t} \delta\left(t - \frac{z+c}{\sqrt{a}}\right) \qquad \text{falls } |c-z| < \sqrt{a}t < z+c$$

$$r(z,t) = \frac{b}{4\sqrt{a}} e^{-\frac{b}{2}t} \left\{ I_0\left(\frac{b}{2}\sqrt{t^2 - \frac{(z-c)^2}{a}}\right) - I_0\left(\frac{b}{2}\sqrt{t^2 - \frac{(z+c)^2}{a}}\right) + \right.$$

$$\left. + t \left[\frac{I_1\left(\frac{b}{2}\sqrt{t^2 - \frac{(z-c)^2}{a}}\right)}{\sqrt{t^2 - \frac{(z-c)^2}{a}}} - \frac{I_1\left(\frac{b}{2}\sqrt{t^2 - \frac{(z+c)^2}{a}}\right)}{\sqrt{t^2 - \frac{(z+c)^2}{a}}} \right] \right\} \text{für } \sqrt{a}\, t > z + c$$

$$r(z,t) = \frac{\sqrt{a}}{2} e^{-\frac{b}{2}t} \quad \delta\left(t - \frac{z-c}{\sqrt{a}}\right) \qquad \text{für } \sqrt{a}t < z-c \; ; \; z > c$$

Eine am Anfang vorhandene Singularität breitet sich also im Laufe der Zeit nach beiden Seiten aus und wird am linken Rande unter Umkehrung des Vorzeichens ihrer Amplitude reflektiert. Wenn die Zeit gegen Unendlich strebt, konvergiert die Residuenfunktion wegen des Dämpfungsfaktors $e^{-bt/2}$ gegen Null. Wir haben in Figur 4 einige Phasen des zeitlichen Verlaufes der Residuenfunktion festgehalten.

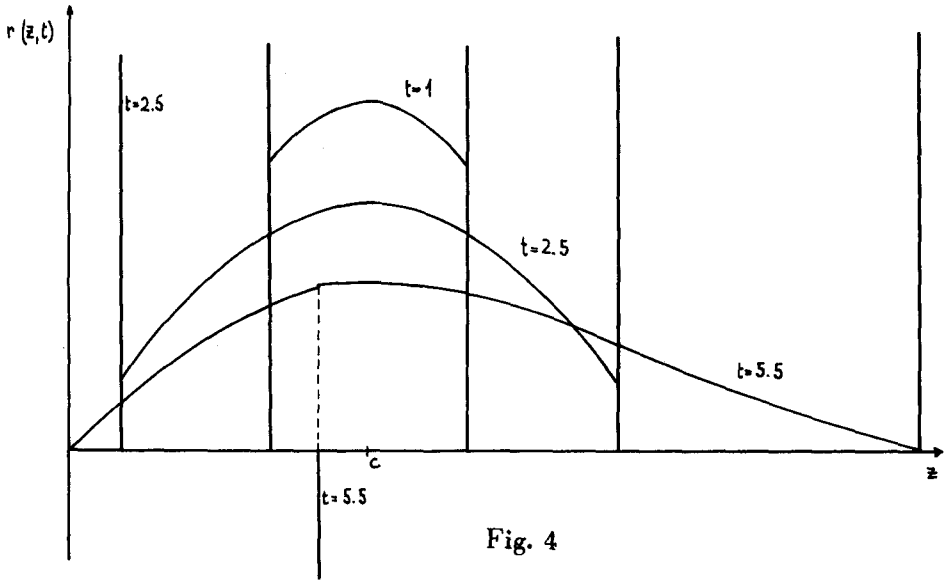


Fig. 4

Auch in diesem Beispiel kann die Lösung des Anfangswertproblemes nur eine qualitative Aussage über die Ausbreitung eines einzelnen Residuums geben, umsomehr als das Residuum von endlicher Grösse durch die singuläre Diracfunktion im Kontinuierlichen ersetzt worden ist. Das hat zur Folge, dass der Wert der Residuenfunktion im Rücken der Wellenfront im kontinuierlichen Fall um eine Grössenordnung kleiner als die Wellenfront selbst ist, während im Diskontinuum die entsprechenden Werte von der gleichen Grössenordnung sind. Das verschiedene Verhalten des Frankel'schen Verfahrens und des Gesamtschrittverfahrens bezüglich der Reflexion am Rande wird jedoch durch die beiden Beispiele im Kontinuum gut illustriert.

Da bei der Ableitung der Analogie zwischen den beiden Gradientenmethoden und den beiden Anfangswertproblemen keinerlei Voraussetzungen über die Dimension des zu Grunde liegenden Randwertproblemes gemacht werden musste, besteht diese Analogie auch bei Randwertproblemen in der Ebene oder im Raum. Nur verursacht dann die Bestimmung der Lösung des Anfangswertproblemes im allgemeinen erheblich mehr Schwierigkeiten.

§ 4 EIN BEISPIEL FUER DAS FRANEKL'SCHE VERFAHREN.

Als numerisches Beispiel für die im § 2 entwickelte Methode lösen wir die Dirichlet'sche Randwertaufgabe für ein rechteckiges Grundgebiet. Wir wählen dieses Beispiel, weil bei ihm alle zu einer vollständigen Diskussion notwendigen Grössen, exakte Lösung usw. leicht bestimmbar sind.
Das Rechteck besitze ein Seitenverhältnis von 3:2. Als Gitter für die Uebersetzung des Problems in die Differenzenrechnnug benützen wir ein solches quadratisches Netz, bei dem die Randpunkte des Rechteckes Gitterpunkte sind. Die Maschenweite h betrage ¼ der Länge der kürzern Seite, sodass im Innern des Rechteckes 15 Gitterpunkte liegen. Beim Dirichlet'schen Randwertproblem ist dann eine Gitterfunktion gesucht, die in jedem innern Gitterpunkt das arithmetische Mittel der Werte in den vier Nachbarpunkten ist und die auf dem Rande vorgegebene Werte annimmt. Der Operator D des Differenzengleichungssystems hat also die Form

$$(4.1) \qquad D = -\begin{bmatrix} G & -E & O \\ E & G & -E \\ O & -E & G \end{bmatrix}$$

wobei E = fünfreihige Einheitsmatrix , O = Nullmatrix,

$$G = -\begin{bmatrix} 4 & -1 & 0 & 0 & 0 \\ 1 & 4 & -1 & 0 & 0 \\ 0 & -1 & 4 & -1 & 0 \\ 0 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & -1 & 4 \end{bmatrix}$$

Die Eigenfunktionen von D sind Produkte trigonometrischer Funktionen. Wenn ξ und η die Koordinaten in einem kartesischen System sind, in dessen ersten Quadranten das Rechteck mit dem einen Eckpunkt im Nullpunkt des Systems und den längern Seiten parallel der ξ Achse liegt, so lauten die 15 Eigenfunktionen

$$(4.2)\ y_{mn} = \sin \frac{m\pi}{6h}\,\xi \cdot \sin \frac{n\pi}{4h}\,\eta \qquad \begin{cases} m = 1,2,\cdots 5 \\ n = 1,2,3 \end{cases}$$

Die zugehörigen Eigenwerte μ findet man durch Einsetzen in das Differenzengleichungssystem zu:

$$(4.3)\ \mu_{mn} = 4\left[\sin^2 \frac{m\pi}{12} + \sin^2 \frac{n\pi}{8}\right]$$

Der kleinste Eigenwert wird für $m = n = 1$ erreicht und beträgt

$$\mu_1 = 0.854$$

Den grössten Eigenwert bekommt man am einfachsten, wenn man die Tatsache benützt, dass beim Dirichletproblem in der Ebene mit μ_i auch 8- μ_i ein Eigenwert des Problemes ist. *) Es wird also

$$\mu_n = 7.146$$

Nach Formel (2.18) wird damit:

$$(4.4) \qquad \lambda_s = 0.309 \qquad\qquad \varepsilon_s = 0.236$$

Für die Festlegung der Randwerte wurde eine differenzharmonische Funktion verwendet, d.h. eine Funktion, deren Wert in einem Gitterpunkt des quadratischen Differenzennetzes das arithmetische Mittel der Funktionswerte in den vier unmittelbar benachbarten Punkten ist. Die Randwerte wurden gerade gleich den Werten der folgenden differenzharmonischen Funktion in den Randpunkten gewählt:

$$(4.5) \qquad \begin{array}{ccccccc} 5 & 10 & 15 & 20 & 25 & 30 & 35 \\ 4 & 8 & 12 & 16 & 20 & 24 & 28 \\ 3 & 6 & 9 & 12 & 15 & 18 & 21 \\ 2 & 4 & 6 & 8 & 10 & 12 & 14 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array}$$

Diese Tabelle stellt somit gerade auch die exakte Lösung des Problemes dar.

Für die nullte Näherung v_0 haben wir den Nullvektor verwendet. Das ergab für das nullte Residuum die Verteilung:

$$(4.6) \qquad r_0 = Dv_0 = \begin{array}{ccccc} -14 & -15 & -20 & -25 & -58 \\ -3 & 0 & 0 & 0 & -21 \\ -4 & -3 & -4 & -5 & -20 \end{array}$$

In diesem Residuenvektor ist die höchste Eigenfunktion die nach (4.2) bezüglich ihrer Vorzeichen eine schachbrettartige Anordnung aufweist, im nullten Residuum stark vertreten.

Da nach Gleichung (2.19) der Koeffizient der höchsten Eigenfunktion am Anfang für die Parameterwerte λ_s, ε_s sogar anwachsen kann, rundeten wir ε_s zur Dämpfung dieser Erscheinung auf den Wert $\varepsilon = 0.24$ auf. In diesem Falle waren alle Wurzeln komplex und die Frequenzen verteilten sich von $\varphi_1 = 4^\circ 59'$ für den kleinsten Eigenwert bis auf $\varphi_{15} = 171^\circ 30'$ für den höchsten Eigenwert ziemlich gleichmässig. Aus diesem Grunde zeigte auch der Verlauf der Residuen von Iteration zu Iteration keinen ausgesprochenen Schwingungscharakter. Mit Absicht haben wir hier nicht die optimalen Bedingungen für die Methode gewählt, da man gewöhnlich ohne einen unverhältnismässigen Arbeitsaufwand nicht so viele Informationen zur Verfügung haben wird.

Die Rechnung wurde auf dem "Card Programmed Calculator" des "Institute for Numerical Analysis" N.B.S. in Los Angeles ausgeführt. Diese von der International Business Machines Co., konstruierte programmgesteuerte Lochkartenmaschine kann mit Hilfe von Schaltbrettern für Rechnungen mit festen oder gleitendem Komma eingerichtet werden. Am genannten Institut existierten nur Schal-

*) Siehe die in der Fussnote 2) der Einleitung zitierte Arbeit.

tungen für festes Komma und entweder acht- oder zehnstellige Zahlen. Bei der Rechnung mit zehnstelligen Zahlen besass die Maschine für 38 solche Zahlen und eine siebenstellige Zahl Speichermöglichkeiten. Das vorliegende Beispiel wurde deshalb mit zehnstelligen Zahlen ohne Aenderung des Dezimalpunktes durchgeführt. Insgesamt brauchte es 24 Iterationen bis der Betrag des Residuenvektors auf den 10^{-7} ten Teil des Betrages des Anfangsresiduenvektors hinabgesunken war. Im 24. Schritt war die Lösung auf 7 bis 8 Stellen genau vorhanden. Schon die Approximation der Lösung im sechzehnten Schritt besass zwischen 4 und 5 richtige Stellen.

Zum Vergleich haben wir auch das Gesamtschrittverfahren auf das gleiche Problem angewendet. Da der kleinste Eigenwert μ_1 für dieses Gleichungssystem die Ungleichung erfüllt $\mu_1 < \frac{1}{2}\mu_{15}$ kann man den Parameter λ nicht so variieren, dass die Anteile aller Eigenfunktionen exakt eliminiert werden. Um die zu den kleinern Eigenwerten gehörenden Eigenfunktionen stark zu eliminieren, sind wir mit λ hart an den Wert $\frac{2}{\mu_{15}}$ herangegangen und verwendeten

$$\lambda = 0.279$$

Das hatte allerdings zur Folge, dass etwa vom zehnten Schritte an die Residuenverteilung immer mehr proportional zur höchsten Eigenfunktion wurde, sodass die Konvergenz schlecht wurde. Wir brachen deshalb das Verfahren ab, da man im allgemeinen Fall in einer solchen Situation durch eine Pauschalkorrektur in Richtung des Residuenvektors die Konvergenz verbessert hätte.

In Tabelle 3 haben wir die Quadrate der Beträge der Residuenvektoren für beide Fälle zusammengestellt.

Schritt k	0	1	2	3	4	5
Frankel Verf. $\lvert r_k \rvert^2$	5726	1235	488.2	157.5	51.49	15.07
Ges. Verf. $\lvert r_k \rvert^2$	5726	1287	594	322.7	186.8	112.9

Schritt k	6	7	8	9	10	11
Frankel Verf. $\lvert r_k \rvert^2$	4.774	1.259	.3566	0.0992	0.0261	0.0068
Ges.verf. $\lvert r_k \rvert^2$	71.33	47.6	33.9	25.9	21.3	18.6

Schritt k	12	13	14	15	16	24
FrankelVerf. $\lvert r_k \rvert^2$	0.0017	$4.40.10^{-4}$	$1.08.10^{-4}$	$2.68.10^{-5}$	$6.29.10^{-6}$	$4.05.10^{-11}$

Tab.1

Untersucht man das Verhältnis der Beträge aufeinanderfolgender Residuen, so sieht man, dass bei dem Frankel'schen Verfahren gegen das Ende hin eine geringe Beschleunigung der Konvergenz auftritt, die auf den Anteil der höchsten und kleinsten Eigenfunktion zurückzuführen ist, weil nach Gleichung (2.19) der Anteil derjenigen Eigenfunktionen, deren Frequenz φ_j klein ist, relativ zu den übrigen Anteilen anfänglich praktisch linear anwächst. Nach einigen Iterationen tritt jedoch der Charakter der trigonometrischen Funktionen wieder in Erscheinung und verlangsamt dann dieses Anwachsen erheblich, was sich in einer Verkleinerung des Verhältnisse $(\frac{(r_k, r_k)}{(r_{k-1}, r_{k-1})})$ ausdrückt. In unserem Beispiel liegt $k\varphi_1$ um den 10. Schritt herum in der Nähe von 90^0, während $k\varphi_{15}$ um den 20. Schritt diesen Wert erreicht. Wie man aus der Tabelle 1 sieht, wird tatsächlich die Konvergenz gegen den Schritt 10 und dann nochmals gegen den Schritt 20 zu beschleunigt. Das Gesamtschrittverfahren zeigt eher das umgekehrte Verhalten. Bei ihm wird im allgemeinen die Konvergenz mit zunehmender Zahl von Iterationen schlechter, da die Verteilung der Residuen immer mehr proportional zu der Eigenfunktion wird, deren Konvergenzfaktor $(1- \lambda\mu_i)$ dem Betrage nach am grössten ist.

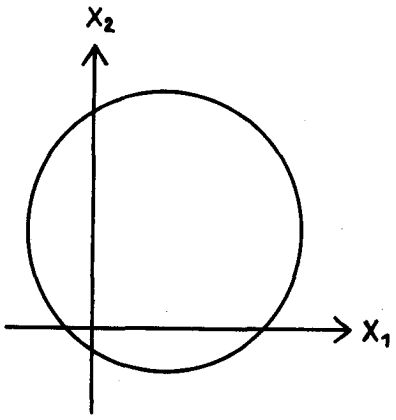
§ 5. ANSATZ EINES ELASTIZITAETSPROBLEMES MIT HILFE DER VARIATIONSRECHNUNG

In der Elastizitätstheorie führen schon die Platten- und Scheibenaufgaben auf Randwertprobleme, die sich nur für ganz einfache Formen des Körpers analytisch lösen lassen. Hingegen gibt die Differenzenrechnung in Verbindung mit der Methode der konjugierten Gradienten eine einfache und zugleich wirksame Methode, auch für kompliziertere Probleme approximative Lösungen zu finden. Vom Standpunkt der Technik aus genügt aber eine approximative Lösung, weil in den meisten Fällen die mathematische Formulierung schon eine starke Idealisierung der physikalischen Gegebenheiten darstellt.

Die Uebersetzung der Probleme in die Differenzenrechnung kann mit Hilfe der Differentialgleichungen und der Randbedingungen erfolgen. Die Formulierung der Randwertaufgabe als Variationsproblem – sofern sie möglich ist – ist jedoch als Ausgangspunkt vorzuziehen, weil sich hier der Uebergang leichter und übersichtlicher gestaltet. Auf diesem zweiten Weg erhält man direkt die zu minimalisierende quadratische Funktion F, aus der dann das Differenzengleichungssystem durch einmalige Ableitung folgt.

Bei den Fragen des stabilen Gleichgewichtes, zu denen auch die statischen Probleme der Elastizitätstheorie gehören, stellt das Prinzip vom Minimum der potentiellen Energie die Formulierung der Variationsaufgabe dar. Das übliche Vorgehen bei Platten- und Scheibenproblemen besteht darin, die potentielle Energie als Funktion der Airy'schen Spannungsfunktion darzustellen und dann aus der Variation dieses Ausdruckes die Differenzengleichungen zu gewinnen. Da die Spannungen erst durch die zweiten Ableitungen der Airy'schen Funktion dargestellt werden, erweist es sich aber vom numerischen Standpunkt aus, besonders bei Aufgaben, die auch die Bestimmung der Spannungen verlangen, günstiger, das Variationsproblem in den Verschiebungen anzusetzen.

Als Beispiel wollen wir den allgemeinen Ansatz für Scheibenprobleme, bei denen ein ebener Spannungszustand vorliegt, mit Hilfe der Variationsrechnung durchführen. Bezeichnen wir mit $\sigma_1, \sigma_2, \tau_{12}$ die Spannungen, mit ε_1, ε_2, γ_{12} die Verzerrungen in Richtung eines kartesischen Koordinatensystems, das in der Mittelebene der Scheibe liegt, dann lauten die aus dem Elastizitätsgesetz für den ebenen Spannungszustand folgenden Beziehungen zwischen diesen Grössen:



$$\sigma_1 = \frac{m^2 E}{m^2-1}\left(\varepsilon_1 + \frac{1}{m}\varepsilon_2\right)$$

(5.1) $$\sigma_2 = \frac{m^2 E}{m^2-1}\left(\varepsilon_2 + \frac{1}{m}\varepsilon_1\right)$$ (m = Poisson'sche Konstante, E = Elastizitätsmodul)

$$\tau_{12} = \frac{mE}{2(m+1)}\gamma_{12}$$

Die Verzerrungen sind mit den Komponenten U,V des Verschiebungsvektors $\vec{W}$ bezüglich desselben Koordinatensystems durch die kinematischen Bedingungen verbunden.

(5.2) $$\varepsilon_1 = \frac{\partial U}{\partial x_1} = U_{x_1},\ \varepsilon_2 = \frac{\partial V}{\partial x_2} = V_{x_2},\ \gamma_{12} = \frac{\partial U}{\partial x_2} + \frac{\partial V}{\partial x_1} = U_{x_2} + V_{x_1}$$

Wir werden im Folgenden Ableitungen nach x_1, bezw. x_2 mit dem Index x_1 bezw. x_2 bezeichnen.
Wird die Scheibe durch äussere, in ihrer Mittelebene liegende Kräfte deformiert – die pro Längeneinheit angreifende Kraft werde mit dem Vektor $\vec{p}$ bezeichnet –, so setzt sich ihre potentielle Energie A aus der Formänderungsenergie und dem Potential der äussern Kräfte zusammen:
($(\vec{p},\vec{W})$ = Skalarprodukt der Vektoren $\vec{p},\vec{W}$)

(5.3) $$A = \frac{1}{2}\iint_F (\sigma_1 \varepsilon_1 + \sigma_2 \varepsilon_2 + \tau_{12}\gamma_{12})\,df - \oint ds\,(\vec{p},\vec{W})$$

wo das erste Integral über den Querschnitt F, das zweite über die gesamte Umrandung der Scheibe zu erstrecken ist. Um die potentielle Energie als Funktion der Verschiebungen allein zu erhalten, müssen wir die Verzerrungen aus den Gleichungen (5.1) mit Hilfe von (5.2) eliminieren:

$$\sigma_1 = \frac{m^2 E}{m^2-1}\left(U_{x_1} + \frac{1}{m}V_{x_2}\right)$$

(5.4) $$\sigma_2 = \frac{m^2 E}{m^2-1}\left(V_{x_2} + \frac{1}{m}U_{x_1}\right)$$

$$\tau_{12} = \frac{mE}{2(m+1)}\left(U_{x_2} + V_{x_1}\right)$$

und dies, sowie Gleichungen (5.2) in (5.3) einsetzen.

(5.5) $$A = \frac{m^2 E}{4(m^2-1)}\iint_F \Big[(U_{x_1} + V_{x_2})^2 + \frac{m-1}{2m}\left\{(U_{x_2}-V_{x_1})^2 - 4(U_{x_1}V_{x_2} - U_{x_2}V_{x_1})\right\}\Big]\,df - \oint ds\,(\vec{p},\vec{W})$$

Wir können diesen Ausdruck noch umformen, da $(U_{x_1} V_{x_2} - U_{x_2} V_{x_1})$ eine Funktionaldeterminante darstellt, weshalb:

$$(5.6) \quad \iint_F (U_{x_1} \cdot V_{x_2} - U_{x_2} \cdot V_{x_1}) \, df = \frac{1}{2} \oint (UdV - VdU)$$

Führen wir den Normalenvektor $\vec{n}$ zum Randelement ds ein:

$$\vec{n} = \left\{ \frac{dx_2}{ds}, -\frac{dx_1}{ds} \right\}$$

so lautet (5.6) in Vektorschreibweise

$$\iint (U_{x_1} \cdot V_{x_2} - U_{x_2} \cdot V_{x_1}) \, df = \tfrac{1}{2} \oint (\vec{W}, \{ \vec{n} \operatorname{div} \vec{W} - (\vec{n}, \operatorname{grad} \vec{W}) \}) \, ds$$

(grad $\vec{W}$ ist ein Tensor zweiter Stufe.)

Damit können wir die potentielle Energie in invarianter Form schreiben:

$$(5.7) \quad A = \frac{m^2 E}{2(m^2-1)} \iint_F \left\{ (\operatorname{div} \vec{W})^2 + \frac{m-1}{2m} (\operatorname{rot} \vec{W})^2 \right\} df -$$

$$- \oint \left(\left\{ \frac{mE}{2(m+1)} \left[\vec{n} \operatorname{div} \vec{W} - (\vec{n}, \operatorname{grad} \vec{W}) \right] + \vec{p} \right\}, \vec{W} \right) ds$$

Wir wollen noch verifizieren, dass man mit Hilfe dieses Ausdruckes die Differentialgleichungen und Randbedingungen für den ebenen Spannungszustand ableiten kann. Nach dem Prinzip vom Minimum der potentiellen Energie ist im Falle des Gleichgewichtes der äussern und innern Kräfte die potentielle Energie ein Minimum. Deshalb muss die durch die Variation der Verschiebungskomponenten erhaltene Aenderung der potentiellen Energie Null sein:

$$(5.8) \quad \delta A = \frac{m^2 E}{m^2-1} \iint_F \left\{ \operatorname{div} \vec{W} \operatorname{div} \delta \vec{W} + \frac{m-1}{2m} (\operatorname{rot} \vec{W}, \operatorname{rot} \delta \vec{W}) \right\} df -$$

$$- \delta \left(\oint \left(\left(\frac{mE}{2(m+1)} \left\{ \vec{n} \operatorname{div} \vec{W} - (\vec{n}, \operatorname{grad} \vec{W}) \right\} + \vec{p} \right), \vec{W} \right) ds \right)$$

Zur Umformung dieses Ausdruckes brauchen wir folgende Identitäten, die mit Hilfe des Nablavektors $\nabla = \left\{ \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right\}$ unmittelbar folgen:

$$(5.9) \quad \operatorname{div}(a\vec{b}) = a \operatorname{div} \vec{b} + (\vec{b}, \operatorname{grad} a)$$

$$(5.10) \quad \operatorname{div}[\vec{a}, \vec{b}] = (\vec{b}, \operatorname{rot} \vec{a}) - (\vec{a}, \operatorname{rot} \vec{b})$$

wenn wir mit $[\vec{a}, \vec{b}]$ das Vektorprodukt der Vektoren $\vec{a}, \vec{b}$ bezeichnen

$$(5.11) \quad [\operatorname{rot} \vec{a}, \vec{b}] = (\vec{b}, \operatorname{grad}) \vec{a} - (\{\operatorname{grad} \vec{a}\}, \vec{b})$$

$$(5.12) \quad \iint_F \operatorname{div} \vec{a} \, df = \oint (\vec{a}, \vec{n}) \, ds$$

Damit wird nach Formel (5.9) und (5.12)

$$\iint_F \operatorname{div}\vec{W}\operatorname{div}\delta\vec{W}\,df = \iint_F \{\operatorname{div}(\delta\vec{W}\cdot\operatorname{div}\vec{W}) - (\delta\vec{W},\ \operatorname{grad}\operatorname{div}\vec{W})\}\,df =$$

$$= \oint (\vec{n},\delta\vec{W})\operatorname{div}\vec{W}\,ds - \iint_F (\delta\vec{W},\ \operatorname{grad}\operatorname{div}\vec{W})\,df$$

Aus Formel (5.10), (5.11) und (5.12) folgt

$$\iint_F (\operatorname{rot}\vec{W},\operatorname{rot}\delta\vec{W})\,df = \iint_F \{\operatorname{div}[\delta\vec{W},\operatorname{rot}\vec{W}] + (\delta\vec{W},\ \operatorname{rot}\operatorname{rot}\vec{W})\}df =$$

$$= \oint (n,[\delta\vec{W},\operatorname{rot}\vec{W}])\,ds + \iint_F (\delta\vec{W},\operatorname{rot}\operatorname{rot}\vec{W})\,df$$

$$= \oint (\delta\vec{W},\{(\vec{n},\operatorname{grad}\vec{W}) - (\operatorname{grad}\vec{W},\vec{n})\})\,ds + \iint_F (\delta\vec{W},\operatorname{rot}\operatorname{rot}\vec{W})\,df.$$

Zur Berechnung der Variation des Randintegrals in Formel (5.8) geht man am einfachsten auf das ursprüngliche Koordinatensystem (x_1,x_2) zurück:

$$\delta\left[\oint \frac{mE}{2(m+1)}(\{\vec{n}\operatorname{div}\vec{W} - (\vec{n},\operatorname{grad}\vec{W})\},\vec{W})\,ds\right] = \frac{mE}{2(m+1)}\oint((\delta U\operatorname{grad}V + U\operatorname{grad}\delta V - \delta V\operatorname{grad}U - V\operatorname{grad}\delta U),\vec{n})\,ds$$

Wenn wir den zweiten und vierten Term partiell integrieren, folgt das Ergebnis, das wir wieder invariant schreiben können:

$$\delta\left[\oint \frac{mE}{2(m+1)}(\cdots,\vec{W})\,ds\right] = \frac{mE}{m+1}\oint(\{\delta U\operatorname{grad}V - \delta V\operatorname{grad}U\},\vec{n})\,ds =$$

$$= \frac{mE}{m+1}\oint(\delta\vec{W},\{\vec{n}\operatorname{div}\vec{W} - (\operatorname{grad}\vec{W},\vec{n})\})\,ds$$

Setzen wir diese Resultate in (5.8) ein:

$$\delta A = \frac{m^2E}{m^2-1}\iint_F(\delta\vec{W},(-\operatorname{grad}\operatorname{div}\vec{W} + \frac{m-1}{2m}\operatorname{rot}\operatorname{rot}\vec{W}))\,df +$$

$$+\oint(\{\frac{m^2E}{m^2-1}(\frac{1}{m}\vec{n}\operatorname{div}\vec{W} + \frac{m-1}{2m}\{(\vec{n},\operatorname{grad}\vec{W}) + (\operatorname{grad}\vec{W},\vec{n})\}) - \vec{p}\},\delta\vec{W})\,ds = 0$$

Aus der Forderung, dass das Flächenintergral verschwinden muss, folgt dann die Differentialgleichung im Innern des Gebietes

$$\operatorname{grad}\operatorname{div}\vec{W} - \frac{m-1}{2m}\operatorname{rot}\operatorname{rot}\vec{W} = 0$$

die wir mit Hilfe der Identität für den Laplaceoperator $\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}$

$$\Delta\vec{W} = \operatorname{grad}\operatorname{div}\vec{W} - \operatorname{rot}\operatorname{rot}\vec{W}$$

auf die übliche Form bringen können:

$$\Delta \vec{W} + \frac{m+1}{m-1} \operatorname{grad} \operatorname{div} \vec{W} = 0$$

Man kann diese Differentialgleichungen natürlich auch aus den Gleichgewichtsbedingungen für die Spannungen im Innern ableiten. Aus dem Verschwinden des Randintegrals folgen die Randbedingungen:

$$\frac{1}{m} \vec{n} \operatorname{div} \vec{W} + \frac{m-1}{2m} \left\{ (\vec{n}, \operatorname{grad} \vec{W}) + (\operatorname{grad} \vec{W}, \vec{n}) \right\} = \frac{m^2-1}{m^2 E} \vec{p}$$

womit die Aequivalenz des Variationsproblemes mit dem Randwertproblem an diesem speziellen Beispiel verifiziert ist.

Für die Uebersetzung des Problems in die Differenzenrechnung ist es nun notwendig, eine Unterteilung des Gebietes mit Hilfe eines Netzes vorzunehmen. Im Interesse einer möglichst guten Approximierung der Randwerte wird man durch eine geeignete Wahl des Netzes die Netzpunkte am Rande gleichmässig zu verteilen suchen. Allerdings wird sich dies nicht immer bei beliebiger Form des Gebietes erreichen lassen, da in der Ebene nur drei Netze existieren, bei denen jeder Netzpunkt von allen seinen Nachbarpunkten gleichen Abstand besitzt, nämlich das quadratische, das Dreieck- und das Sechsecknetz. Bei dem speziellen Beispiel, das hier im Anschluss an die allgemeinen Ausführungen behandelt werden soll, handelt es sich um eine parallelogrammförmige Scheibe von folgender Form:

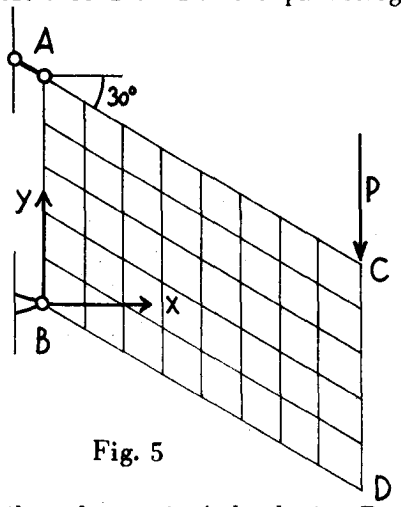


Fig. 5

In B ist die Scheibe gelagert, in A durch eine Pendelstütze befestigt. In diesem Falle war die Verwendung des Dreiecknetzes gegeben. Wegen des irrationalen Seitenverhältnisses von 5:8 besitzt das weitmaschigste Netz, das das Gebiet gerade vollständig überdeckt, insgesamt 54 Gitterpunkte.

Zur Aufstellung der Differenzengleichungen müssen die in der Gleichung (5.7) für die potentielle Energie vorkommenden ersten Ableitungen durch Differenzen ersetzt werden. Eine gleichmässige Berücksichtigung aller Punkte erreicht man auf

einfache Weise, wenn man die räumliche Vorstellung zu Hilfe nimmt. Die Grundfigur aus der sich das Dreiecknetz aufbaut, kann als die Projektion eines räumlichen rechtwinkligen Koordinatensystems aufgefasst werden:

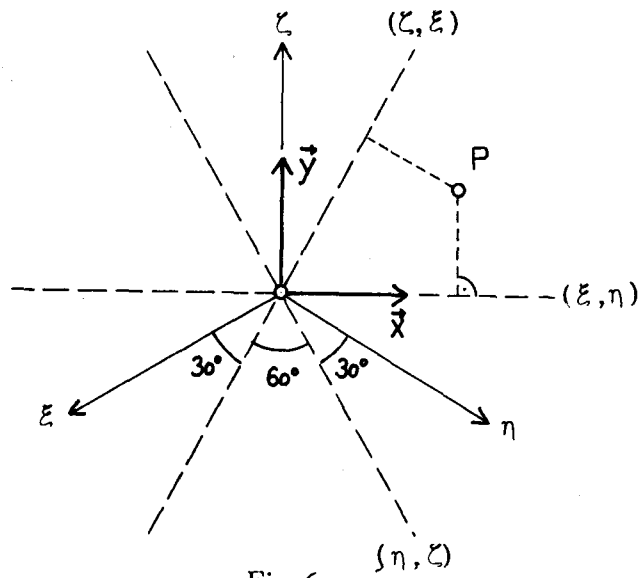


Fig. 6

Die Dreiecknetzebene sei durch die Gleichung gegeben:

$$\xi + \eta + \zeta = 0$$

Sie geht also durch den Ursprung des Koordinatensystems und ist parallel zur Zeichenebene. In der Figur 6 sind gerade noch die Schnittgeraden der Koordinatenebenen mit der Dreiecknetzebene eingezeichnet. Die räumlichen Koordinaten eines Punktes P der Netzebene lassen sich damit ihrer relativen Grösse nach direkt aus der Zeichnung als Abstände von diesen Schnittgeraden ablesen. Das in der Figur 7 eingezeichnete Dreieck mit den Eckpunkten 0,1,2, gehöre zum Dreiecknetz:

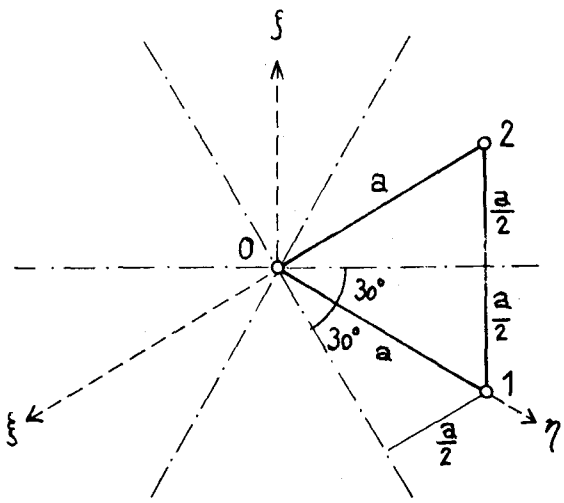


Fig. 7

Wie aus der Figur ersichtlich erfüllen die Koordinaten des Punktes 1 folgende Beziehungen:

$$\xi^2 + \eta^2 + \zeta^2 = a^2$$

$$\xi : \eta : \zeta = (-\frac{a}{2}) : a : (-\frac{a}{2})$$

Daraus folgt

$$\xi = -\frac{a}{\sqrt{6}}; \quad \eta = \frac{2a}{\sqrt{6}}; \quad \zeta = -\frac{a}{\sqrt{6}}$$

Die Komponenten des Verschiebungsvektors $\vec{W}$ im räumlichen Koordinatensystem (ξ,η,ζ) seien u,v,w. Sie sind stetige und differenzierbare Funktionen von ξ,η,ζ sodass sie in eine Taylorreihe entwickelbar sind. Begnügt man sich mit den Gliedern erster Ordnung, so lautet die Taylorentwicklung für irgendeine der Verschiebungskomponenten um den 0-Punkt herum

$$f(\xi,\eta,\zeta) = f(0) + \xi f_\xi(0) + \eta f_\eta(0) + \zeta f_\zeta(0)$$

wo f entweder gleich u oder v oder w ist.

Für den Funktionswert im Punkte 1 erhalten wir damit als approximativen Wert

$$f(1) = f(0) + \frac{a}{\sqrt{6}}(-f_\xi(0) + 2f_\eta(0) - f_\zeta(0))$$

Da u,v,w nur in der Netzebene variieren, steht der Normalenvektor $\vec{m}$ der Netzebene senkrecht auf dem Gradienten dieser Funktion:

$$(\vec{m}, \operatorname{grad} f) = 0, \text{ d.h. } f_\xi + f_\eta + f_\zeta = 0$$

Damit folgt

$$f_\eta(0) = \frac{\sqrt{6}}{3a}(f(1) - f(0))$$

Auf dieselbe Weise findet man des weiteren:

$$f_\xi(0) = \frac{\sqrt{6}}{3a}(f(0) - f(2))$$

$$f_\zeta(0) = \frac{\sqrt{6}}{3a}(f(2) - f(1))$$

Der Ausdruck für die potentielle Energie als Funkton der räumlichen Verschiebungskomponenten kann unmittelbar aus Formel (5.7) gewonnen werden, da diese invariant geschrieben ist.

$$A = \frac{m^2 E}{2(m^2-1)} \iint_F df\{(u_\xi + v_\eta + w_\zeta)^2 + \frac{m-1}{2m}((u_\eta - v_\xi)^2 + (w_\eta - v_\zeta)^2 +$$

$$+ (u_\zeta - w_\xi)^2)\} - \oint\{\frac{mE}{2(m+1)}((un_1 + vn_2 + wn_3)(u_\xi + v_\eta + w_\zeta) - u(n_1 u_\xi + n_2 v_\xi + n_3 w_\xi) -$$

$$- v(n_1 u_\eta + n_2 v_\eta + n_3 w_\eta) - w(n_1 u_\zeta + n_2 v_\zeta + n_3 w_\zeta)) + p_1 u + p_2 v + p_3 w\} ds$$

wo n_1, n_2, n_3 die räumlichen Komponenten des Normalenvektors längs des Randes sind,

p_1, p_2, p_3 die räumlichen Komponenten der pro Längeneinheit angreifenden Kraft $\vec{p}$. Jedes Netzdreieck (0,1,2) liefert also, wenn man sich auf die ersten Differenzen beschränkt, folgenden Beitrag zur potentiellen Energie:
(Zur Abkürzung verwenden wir die Schreibweise f_i für den Funktionswert von f im Punkte i)

$$\Delta A = \frac{m^2 E}{2(m^2-1)} \frac{\sqrt{3}}{4} a^2 \frac{2}{3a^2} \left[(u_0 - u_2 + v_1 - v_0 + w_2 - w_1)^2 + \frac{m-1}{2m}((u_1 - u_0 + v_2 - v_0)^2 + \right.$$

$$\left. + (w_1 - w_0 + v_1 - v_2)^2 + (u_2 - u_1 + w_2 - w_0)^2)\right]$$

Die potentielle Energie A wird durch die Summe dieser Ausdrücke für alle Netzdreiecke und den Beiträgen, die vom Randintegral herrühren, approximiert. Sie ist somit eine Funktion der Variabeln u_i, v_i, w_i, wobei der Index i nur von 1 bis 53 läuft, da der Scheibenpunkt B festgehalten ist, sodass dort immer $u=v=w=0$ gilt. Im Gleichgewichtszustand ist sie minimal d.h. sie genügt dann folgenden 3 x 53 Bedingungen:

$$\frac{\partial A}{\partial u_i} = \frac{\partial A}{\partial v_i} = \frac{\partial A}{\partial w_i} = 0 \qquad (i = 1,2,\cdots 53)$$

In jedem Gitterpunkt bestehen drei Differenzengleichungen, die allerdings nicht voneinander unabhängig sind, da der Verschiebungsvektor in der Netzebene liegt, also

$$u + v + w = 0 \tag{5.13}$$

und

$$u_\xi + u_\eta + u_\zeta = 0$$

$$v_\xi + v_\eta + v_\zeta = 0 \tag{5.14}$$

$$w_\xi + w_\eta + w_\zeta = 0$$

Wir verzichten hier, das Differenzengleichungssystem in u_i, v_i, w_i, anzuschreiben, da es zu umfangreich ist. Eine erhebliche Reduktion des Systems wird erreicht, wenn man wieder auf die Verschiebungskomponenten U, V in der Netzebene übergeht. Zudem werden die Differenzengleichungen auf diese Weise symmetrisch.
Ist $\vec{x}$ ein Einheitsvektor in Richtung der x-Achse, $\vec{y}$ ein solcher in Richtung der y-Achse (siehe Fig. 6), so bestehen die Beziehungen

$$U = (\vec{W}, \vec{x}) \quad ; \quad V = (\vec{W}, \vec{y})$$

Aus Fig. 6 folgt auch, indem man wieder die Tatsache benützt, dass die Abstände von den Schnittgeraden (ξ, η), (ξ, ζ), (η, ζ) proportional den Koordinaten sind, für die Einheitsvektoren $\vec{x}, \vec{y}$

$$\vec{x} = \frac{1}{\sqrt{2}}\{-1,1,0\} \quad ; \quad \vec{y} = \frac{1}{\sqrt{6}}\{-1,-1,2\}$$

Berücksichtigt man noch Gleichung (5.13), so wird

$$U = \frac{1}{\sqrt{2}}(-u+v) \quad ; \quad V = -\sqrt{\frac{3}{2}}(u+v)$$

Für die numerische Rechnung ist es von Vorteil, wenn die Koeffizienten in den Differenzengleichungen mindestens rational sind. Deshalb wurde für die Verschiebungskomponente V eine Masstabänderung vorgenommen:

$$V' = \frac{1}{\sqrt{3}}\, V$$

Sodann wurden aus den gleichen Gründen alle Gleichungen mit $2\sqrt{6}\,\frac{m^2-1}{mE}$ multipliziert. Numerieren wir die Netzpunkte in der in Fig. 8 angegebenen Weise, so lauten dann die Differenzengleichungen:

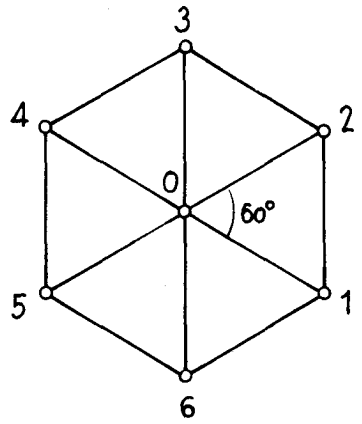


Fig. 8

a) Falls der Punkt 0 im Innern der Scheibe liegt:

$$(3m-1)U_o + \frac{1}{6}(3-m)(U_3 + U_6) - \frac{2m}{3}(U_1 + U_2 + U_4 + U_5) + (m+1)(V'_1 - V'_2 + V'_4 - V'_5) = 0$$

$$3(3m-1)V'_o - \frac{1}{2}(5m+1)(V'_3 + V'_6) - (m-1)(V'_1 + V'_2 + V'_4 + V'_5) + (m+1)(U_1 - U_2 + U_4 - U_5) = 0$$

b) Der Punkt 0 liegt auf dem Rande AC: (Eckpunkte ausgeschlossen)

$$\frac{1}{2}(3m-1)U_o - \frac{1}{6}(m-3)U_6 - \frac{m}{3}(U_1 + U_4 + 2U_5) + \frac{1}{2}(m-1)V'_1 - \frac{1}{2}(m+1)V'_5 + V'_4 = 0$$

$$\frac{3}{2}(3m-1)V'_o - \frac{1}{2}(m-1)(V'_1 + V'_4 + 2V'_5) - \frac{1}{2}(5m+1)V'_6 + \frac{1}{2}(m-1)U_4 - \frac{1}{2}(m+1)U_5 + U_1 = 0$$

c) Der Punkt 0 liegt auf dem Rande CD (ohne Eckpunkte):

$$\frac{1}{2}(3m-1)U_o - \frac{1}{12}(m-3)(U_3 + U_6) - \frac{2}{3}m(U_4 + U_5) + \frac{1}{4}(m-3)(V'_6 - V'_3) + \frac{1}{2}(m+1)(V'_4 - V'_5) = 0$$

$$\frac{3}{2}(3m-1)V'_o - \frac{1}{4}(5m+1)(V'_3 + V'_6) - (m-1)(V'_4 + V'_5) + \frac{1}{4}(m-3)(U_3 - U_6) + \frac{1}{2}(m+1)(U_4 - U_5) = 0$$

d) Der Punkt 0 fällt mit dem Punkte D zusammen:

$$\frac{1}{12}(5m-3)U_o - \frac{1}{12}(m-3)U_3 - \frac{m}{3}U_4 - \frac{1}{4}(m+1)V'_o - \frac{1}{4}(m-3)V'_3 + \frac{1}{2}(m-1)V'_4 = 0$$

$$\frac{1}{4}(7m-1)V'_o - \frac{1}{4}(5m+1)V'_3 - \frac{1}{2}(m-1)V'_4 - \frac{1}{4}(m+1)U_o + \frac{1}{4}(m-3)U_3 + U_4 = 0$$

e) Im Punkte C lauten die Differenzengleichungen, falls eine Kraft von 1 Kg in Richtung der negativen y-Achse angreift:

$$\frac{1}{12}(13m-3)U_o - \frac{m}{3}(U_4+2U_5) - \frac{1}{12}(m-3)U_6 + \frac{1}{4}(m+1)V'_o + \frac{1}{4}(m-3)V'_6 + V'_4 -$$

$$-\frac{1}{2}(m+1)V'_5 = 0$$

$$\frac{1}{4}(11m-5)V'_o - \frac{1}{4}(5m+1)V'_6 - \frac{1}{2}(m-1)(V'_4+2V'_5) + \frac{1}{4}(m+1)U_o + \frac{1}{2}(m-1)U_4 - \frac{1}{2}(m+1)U_5 -$$

$$-\frac{1}{4}(m-3)U_6 = -2\frac{m^2-1}{mE}$$

Die Differenzengleichungen für den übrigen Rand erhält man durch Drehung der Figur 8 um den Winkel π mit dem Punkte 0 als Drehzentrum, d.h. man hat einfach die Indizes in folgender Weise zu vertauschen:

$$1 \longrightarrow 4, 2 \longrightarrow 5, \quad 3 \longrightarrow 6, \quad 4 \longrightarrow 1, \quad 5 \longrightarrow 2, \quad 6 \longrightarrow 3$$

Zu diesen Differenzengleichungen kommen noch die Festhaltebedingungen:

in A: $U_o = V'_o$ wobei angenommen wird, die Pendelstütze sei nicht deformierbar.

in B: $U_o = V'_o = 0$

Die Koeffizienten des Differenzengleichungssystems werden noch wesentlich von der Wahl der Poisson'schen Konstanten m beeinflusst, da die Deformationen vom elastischen Verhalten des Stoffes abhängig sind. In der vorliegenden Rechnung wurde m = 3 gewählt. Dieser Wert liegt in der Nähe desjenigen von Stahl und hat zudem den Vorteil, dass im Differenzengleichungssystem dann einige Koeffizienten Null und die übrigen ganzzahlig sind.

§ 6 DIE BERECHNUNG DER VERSCHIEBUNGEN UND SPANNUNGEN IN EINER PARALLELOGRAMMFOERMIGEN SCHEIBE.

Die theoretischen Grundlagen zu diesem Problem sind in den Paragraphen 1 und 5 entwickelt worden. Wir wollen deshalb hier nur die numerische Auflösung des linearen, symmetrischen Gleichungssystemes mit 106 Unbekannten, das am Schlusse des § 5 aufgeführt ist, behandeln. Bei diesem Problem hatten wir für die Poisson'sche Konstante den Wert m =3 gewählt, für das Elastizitätsmodul E setzen wir den Wert E= 21000 kg/mm^2 ein.

Als nullte Versuchslösung nahmen wir die Verschiebungen Null an. Spezielle Aufmerksamkeit wurde dem Einbau von Rechenkontrollen gewidmet. Für je 6 Verschiebungskomponenten in derselben Richtung, die zu Punkten gehörten, welche denselben Abstand von den kürzern Seiten hatten, wurde eine Summenkontrolle in der Rechnung mitgeführt, d.h. es wurden für die Summe dieser Komponenten dieselben Operationen wie für diese selbst ausgeführt und dann dieses Resultat mit der Summe der Resultate verglichen. Diese Summenkontrolle diente speziell zur Prüfung der Berechnung des Vektors Dp_{k-1} im k-ten Zyklus. Die Berechnung von λ_{k-1} und r_k nach den Formeln (1.5a) und (1.8) wurde mit Hilfe des Skalarproduktes (r_{k-1}, r_k) geprüft, da dieses wegen der Orthogonalität der Residuenvektoren nur sehr klein sein durfte. Bei der Rechnung mit einer endlichen Ziffernzahl wird man wegen den Rundungsfehlern die exakte Erfüllung solcher Beziehungen nicht erwarten können. Der Parameter ε_{k-1} und mit ihm der Gewichtsvektor p_k wurden dadurch geprüft, dass man mit dem Skalarprodukt (p_k, Dp_{k-1}) feststellte, ob p_k tatsächlich zu p_{k-1} konjugiert war. Diese Kontrolle war besonders wichtig, da beabsichtigt war, weitere Belastungsfälle für dieselbe Scheibe mit Hilfe des Systems konjugierter Gewichtsvektoren p_k zu berechnen. Deshalb wurde auch die folgende Korrekturformel verwendet:

Wenn das im k-ten Schritt berechnete Skalarprodukt (p_k, Dp_{k-1}) nicht mehr sehr klein war, wurde im nächsten Schritt ein Korrekturfaktor d_k eingeführt, der so bestimmt wurde, dass $(r_k, r_{k+1}) = 0$ innerhalb der Rechengenauigkeit exakt erfüllt war. Dieser Korrekturfaktor bewirkte eine Aenderung der Parameter λ_k und ε_k im (k+1) ten Schritt entsprechend den Formeln

$$\bar{\lambda}_k = \frac{\lambda_k}{d_k} \qquad \bar{\varepsilon}_k = \varepsilon_k \, d_k$$

wo λ_k, ε_k die nach den Formeln (1.5a) und (1.10) berechneten Grössen sind, während der Korrekturfaktor d_k sich nach der Formel

$$d_k = 1 - \varepsilon_{k-1} \cdot \frac{(p_k, Dp_{k-1})}{(p_k, Dp_k)}$$

bestimmt. Diese Korrekturformel wurde nur verwendet, wenn die Orthogonalitätsrelation $(p_k, Dp_{k-1}) = 0$ wirklich schlecht erfüllt war.

Die Rechnung wurde auf der programmgesteuerten Zuse-Rechenmaschine des Institutes für angewandte Mathematik an der ETH durchgeführt. Diese Relais-Maschine arbeitet im Dualsystem mit gleitendem Dezimalpunkt. Bei der Rechnung

werden etwas mehr als sechs Dezimalstellen mitgeführt. Das mechanische Speicherwerk kann 64 Zahlen speichern. Da in diesem Problem die Vektoren p_k und r_k insgesamt mit den Summenkontrollen 124 Komponenten besassen, mussten in ausgiebigem Masse Zahlen durch Lochung in Filmstreifen gespeichert werden. Diese Art der Speicherung verlangsamte die Rechengeschwindigkeit, sodass ein Zyklus etwa 140 Minuten dauerte. Insgesamt wurden 90 Iterationen durchgeführt. Nach dem 90. Schritt war der Betrag des Residuenvektors auf ungefähr den 10^4 ten Teil der Werte am Anfang der Rechnung hinabgesunken. Da infolge der Rundungsfehler der Residuenvektor im 90-ten Schritt nur noch schlecht mit dem mit Hilfe der Funktionswerte v_{90} bestimmten Residuenvektor übereinstimmte, wurde die Rechnung abgebrochen. Es wurde dann eine Abschätzung des mittleren Fehlers δ mit Hilfe der folgenden Beziehungen vorgenommen *)

$$\delta = \frac{\rho}{\sigma_{\frac{1}{2}}}$$

wo $\rho = \sqrt{\frac{(r,r)}{n}}$ das mittlere quadratische Residuum und

$$\sigma_{\frac{1}{2}} = \sqrt{\frac{(r,r)}{(f,f)}}$$

ist, wobei f den Fehlervektor darstellt. $\sigma_{\frac{1}{2}}$ kann angenähert mit Hilfe der Schwarz'schen Quotienten σ_2 und σ_3 durch logarithmische Extrapolation berechnet werden:

$$\log \sigma_{\frac{1}{2}} = \log \sigma_2 - \frac{3}{2} (\log \sigma_3 - \log \sigma_2)$$

Die beiden Schwarz'schen Quotienten können direkt aus den Parametern λ_k und ε_k berechnet werden:

$$\sigma_{2,k} = \frac{(r_k, Dr_k)}{(r_k, r_k)} = \frac{1}{\lambda_k} + \frac{\varepsilon_{k-1}}{\lambda_{k-1}}$$

$$\sigma_{3,k} = \frac{(Dr_k, Dr_k)}{(r_k, Dr_k)} = \sigma_2 + \frac{1}{\sigma_2} \left(\frac{\varepsilon_k}{\lambda_k^2} + \frac{\varepsilon_{k-1}}{\lambda_{k-1}^2} \right)$$

Im 90. Schritt war das so berechnete $\delta \sim 6.5$ .Dieser Wert scheint zu klein zu sein, da die Funktionswerte Zahlen in der Grössenordnung um 10^6 herum waren und die 7. Stelle bei dieser Rechenmaschine schon unsicher ist.
Die Orthogonalität des Systemes konjugierter Gewichtsvektoren p_k wird durch den Winkel β_{ik} geprüft, wo

$$\cos \beta_{ik} = \frac{(p_i, Dp_k)}{\sqrt{(p_i, Dp_i) \cdot (p_k, Dp_k)}}$$

*) Siehe dadür die in der 2. Fussnote der Einleitung erwähnte Arbeit.

ist. Stichproben ergaben Abweichungen bis zu etwa 7° vom rechten Winkel. Da zudem das System nicht vollständig war, musste auf die Berechnung weiterer Lastfälle mit Hilfe der Gewichtsvektoren p_k verzichtet werden.

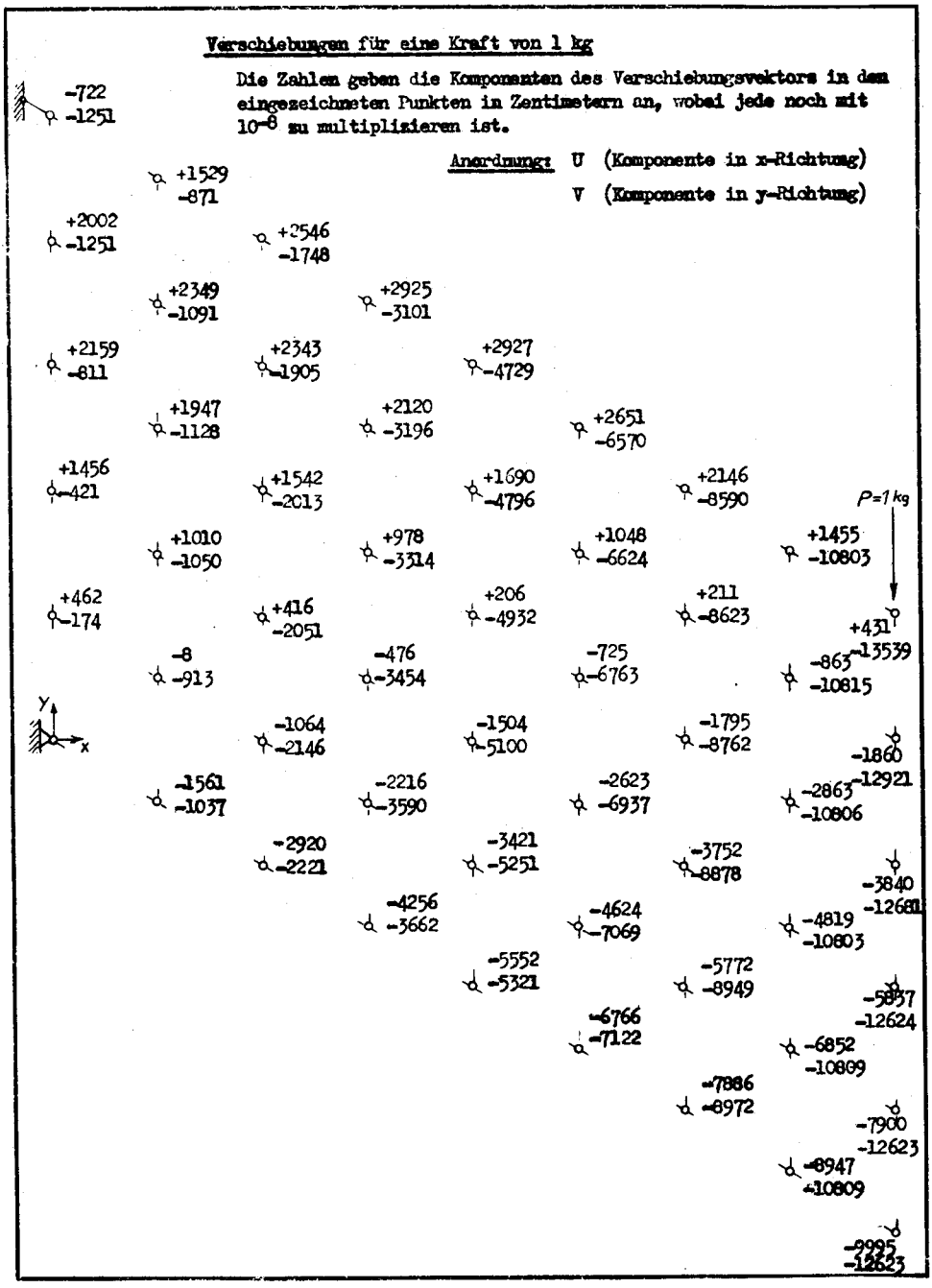
Der Rayleigh'sche Quotient $\sigma_{2,k}$ der Residuenvektoren schwankte während der ganzen Rechnung nur innerhalb der Werte 10.166 und 29.072. Da $\sigma_{2,k}$ zwischen dem grössten und kleinsten Eigenwert liegen muss, geben die aufgeführten Werte eine untere Grenze für das Verhältnis des grössten zum kleinsten Eigenwert. Diese Grenze kann noch verbessert werden, wenn man die für die Reziproken des Parameters λ_k geltende Ungleichung (2.24) heranzieht. Die Reziproken von λ_k schwankten zwischen 4.13 und 17.7. Das Verhältnis des grössten zum kleinsten Eigenwert muss also grösser als 7 sein. Die trotz der weniger als 7-stelligen Rechnung relativ gute Konvergenz lässt vermuten, dass diese Grenze wenigstens grössenordnungsmässig das Verhältnis richtig wiedergibt, da falls $\frac{\mu_n}{\mu_1}$ viel grösser wäre, die Rundungsfehler die Konvergenz wahrscheinlich stark beeinträchtigen würden.

Benützen wir auf Grund dieser Zahlen für den kleinsten Eigenwert den Wert $\mu_1 = 4$, für den grössten $\mu_n = 30$, so erhalten wir als optimale Werte für das Frankel'sche Verfahren nach der Formel (2.18):

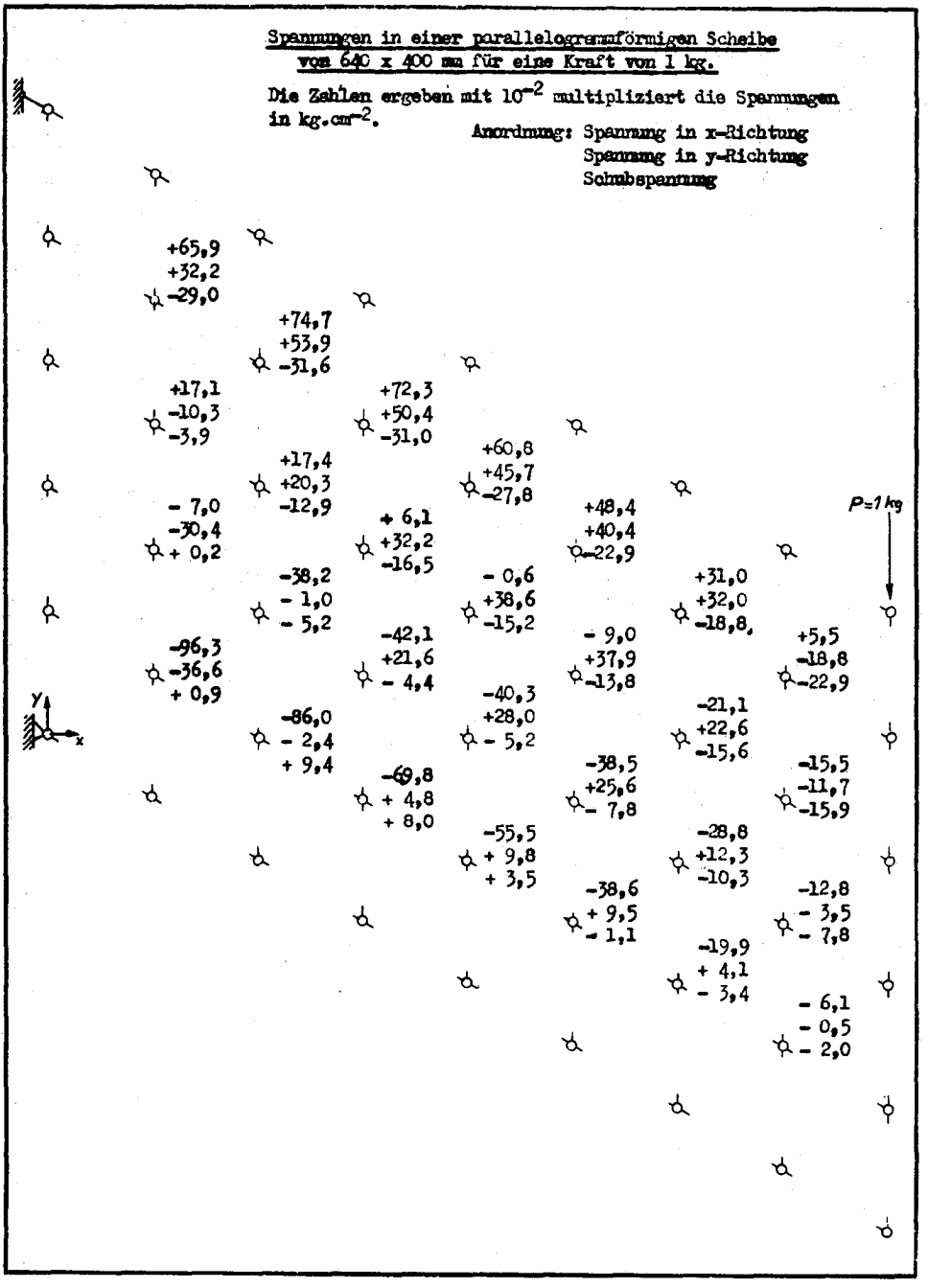
$$\lambda_s = 0.084 \qquad \varepsilon_s = 0.22$$

Untersucht man die Quadrate der Residuen, so findet man, dass diese immer dann stark abnahmen, wenn λ_k und ε_k in der Nähe von λ_s und ε_s waren, und immer dann zunahmen wenn ε_k grösser als 1 war. Dieses Verhalten steht also im Einklang mit der Diskussion im Paragraphen 2 und zeigt gleichzeitig auch eine Möglichkeit, die Methode der konjugierten Gradienten abzukürzen, indem man, nachdem man sich in einigen Iterationen genügend Informationen über die grössten und kleinsten Eigenwerte gesammelt hat, zum Frankel'schen Verfahren übergeht. In einem solchen Fall wird man zu Vermeidung grosser Umstellungen in den Rechenplänen die Gewichtsvektoren p_k beibehalten und nur die Berechnung von ε_k und λ_k eliminieren.

Die Verschiebungen für eine Kraft von 1 kg, die in der Scheibenmittelebene in der in Fig. 5 eingezeichneten Richtung angreift, sind in der nachfolgenden Tabelle 2 aufgeführt. Zur Veranschaulichung haben wir die deformierte Scheibe beim Angriff einer auf 40to vergrösserten Kraft in Figur 9 dargestellt. Die Spannungen wurden nach den Formeln (5.4) durch numerische Differentiation aus den Verschiebungen bestimmt und sind in Tabelle 3 aufgeführt. Die Werte am Rande wurden dabei weggelassen, da diese durch unsere Rechnung nur sehr schlecht approximiert werden. Dies ist ohne weiteres erklärlich, wenn man sich vor Augen hält, dass man durch die Verwendung eines Dreiecknetzes die homogene Scheibe physikalisch durch ein Fachwerk ersetzt hat. Dieses approximiert in den innern Punkten, wo die Stäbe symmetrisch angreifen, die homogene Scheibe ziemlich gut, während auf dem Rande infolge der fehlenden Symmetrie die Approximation schlecht ist. Eine Verbesserung dieser Randwerte könnte vielleicht dadurch erreicht werden, dass man den Rand aufteilt und für die einzelnen Stücke die Lösung des Scheibenproblems für die unendliche Halbebene und den Keil benutzen würde. Da die Randwerte der Spannungen nicht gebraucht wurden, wurde diese Vermutung nicht weiter untersucht.


Verschiebungen für eine Kraft von 1 kg
Die Zahlen geben die Komponenten des Verschiebungsvektors in den eingezeichneten Punkten in Zentimetern an, wobei jede noch mit 10^{-8} zu multiplizieren ist.
Anordnung: U (Komponente in x-Richtung)
V (Komponente in y-Richtung)
P=1 kg
-722
-1251
+1529
-871
+2002
-1251
+2546
-1748
+2349
-1091
+2925
-3101
+2159
-811
+2343
-1905
+2927
-4729
+1947
-1128
+2120
-3196
+2651
-6570
+1456
-421
+1542
-2013
+1690
-4796
+2146
-8590
+1455
-10803
+1010
-1050
+978
-3314
+1048
-6624
+462
-174
+416
-2051
+206
-4932
+211
-8623
+431
-13539
-8
-913
-476
-3454
-725
-6763
-865
-10815
-1064
-2146
-1504
-5100
-1795
-8762
-1860
-12921
-1561
-1037
-2216
-3590
-2623
-6937
-2863
-10806
-2920
-2221
-3421
-5251
-3752
-8878
-3840
-12681
-4256
-3662
-4624
-7069
-4819
-10803
-5552
-5321
-5772
-8949
-5837
-12624
-6766
-7122
-6852
-10809
-7886
-8972
-7900
-12623
-8947
-10809
-8995
-12623
y
x

Tab. 2


Spannungen in einer parallelogrammförmigen Scheibe
von 640 x 400 mm für eine Kraft von 1 kg.
Die Zahlen ergeben mit 10^{-2} multipliziert die Spannungen in $kg.cm^{-2}$.
Anordnung: Spannung in x-Richtung
Spannung in y-Richtung
Schubspannung
P=1 kg
+65,9
+32,2
-29,0
+74,7
+53,9
-31,6
+17,1
-10,3
-3,9
+72,3
+50,4
-31,0
+17,4
+20,3
-12,9
+60,8
+45,7
-27,8
- 7,0
-30,4
+ 0,2
+ 6,1
+32,2
-16,5
+48,4
+40,4
-22,9
-38,2
- 1,0
- 5,2
- 0,6
+38,6
-15,2
+31,0
+32,0
-18,8
-96,3
-36,6
+ 0,9
-42,1
+21,6
- 4,4
- 9,0
+37,9
-13,8
+5,5
-18,8
-22,9
-86,0
- 2,4
+ 9,4
-40,3
+28,0
- 5,2
-21,1
+22,6
-15,6
-69,8
+ 4,8
+ 8,0
-38,5
+25,6
- 7,8
-15,5
-11,7
-15,9
-55,5
+ 9,8
+ 3,5
-28,8
+12,3
-10,3
-38,6
+ 9,5
- 1,1
-12,8
- 3,5
- 7,8
-19,9
+ 4,1
- 3,4
- 6,1
- 0,5
- 2,0
y
x

Tab. 3

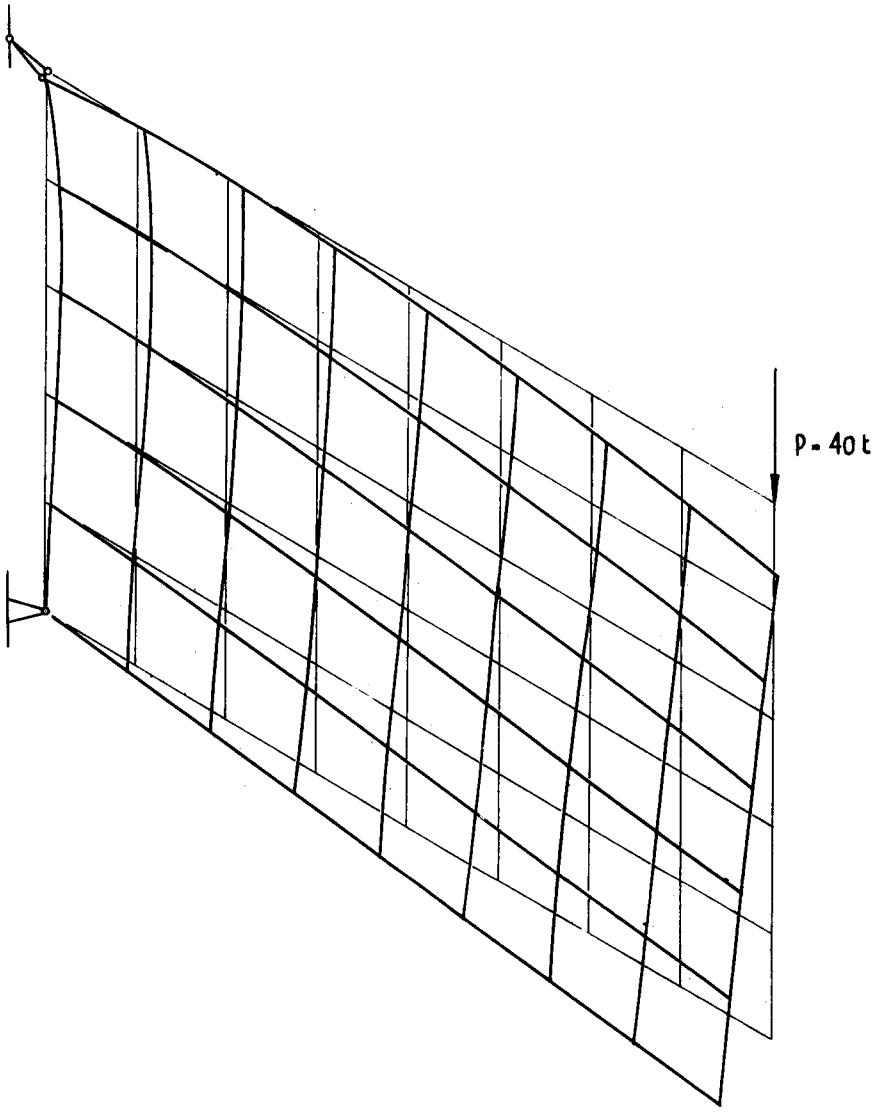


Fig. 9

Lebenslauf

Geboren am 12. Januar 1926 in Zürich, verbrachte ich meine ganze Kindheit in dieser Stadt. Nach Absolvierung der Primarschule besuchte ich das kantonale Gymnasium, wo ich im Herbst 1944 mit der Maturität vom Typus B abschloss. Anschliessend trat ich in die Abteilung für Mathematik und Physik der E. T. H. ein. Im Dezember 1948 wurde mir das Diplom als Physiker verliehen. Nach anderthalb Jahren weiterer Ausbildung in theoretischer Physik bei Herrn Prof. Pauli wurde ich im Frühjahr 1950 Assistent für Mechanik bei Herrn Prof. Ziegler. Vom Herbst 1950 an arbeitete ich auch noch am Institut für angewandte Mathematik der E. T. H., aus welcher Tätigkeit unter der Leitung von Herrn Prof. Stiefel die vorliegende Dissertation entstanden ist. Im Herbst 1951 gab ich meine Assistentenstelle auf und folgte meinem Lehrer Herrn Prof. Stiefel nach Kalifornien, wo ich auf Grund einer Fellowship für ein Jahr am Institute for Numerical Analysis N. B. S., University of California Los Angeles tätig war. Nach meiner Rückkehr aus Amerika im September 1952 nahm ich eine Stelle bei den Flug- und Fahrzeugwerken A.-G. Altenrhein an und bin seither für diese Firma als angewandter Mathematiker in Zürich tätig.