Mathematisches Forschungsinstitut
Oberwolfach

Tagungsbericht

Numerische Behandlung von Problemen der linearen Algebra

8. bis 12. Juni 1964.

Die Tagung fand von Montag, den 8. Juni bis Freitag, den 12. Juni in Oberwolfach statt. Sie wurde geleitet von
Professor Dr. F.L. Bauer (München) und
Professor Dr. A. Ostrowski (Basel).

Es nahmen teil:

J. Albrecht, Hamburg
F.L. Bauer, München
W. Börsch-Supan, Mainz
R. Nicolovius, Hamburg
W. Niegel, München
A. Ostrowski, Basel
H. Rutishauser, Zürich
H.J. Schneider, Stuttgart
A. Schönhage, Köln
H.R. Schwarz, Zürich
H. Sprenger, Hamburg
J. Stoer, München
G.W. Veltkamp, Eindhoven

Die Anzahl der Teilnehmer war verhältnismäßig gering. Umso lebhafter konnte die Fühlungnahme und Diskussion unter den Teilnehmern sein. Es herrschte keine Überfüllung von Vorträgen, und es blieb deshalb Zeit, die nachfolgenden Problemkreise ausführlich zu besprechen:

A. Nachprüfbarkeit der Lösung eines linearen Gleichungssystems
B. Fragen um die Pivotwahl
C. Numerische Berechnung von Eigenwerten nichtsymmetrischer und nichtnormaler Matrizen
D. Probleme um den Normbegriff
E. Fragen um die LR-Transformation
F. Diskussion über die optimale Organisation eines Recheninstituts im Idealfall.

Alle Teilnehmer waren vom Verlauf der Tagung sehr befriedigt, und es wurde einstimmig der Wunsch laut, es möchten doch jedes Jahr auch die numerische Mathematik berücksichtigende Tagungen in Oberwolfach abgehalten werden. Insbesondere wurden folgende Themen vorgeschlagen:

a) Numerische Behandlung gewöhnlicher Differentialgleichungen

b) Numerische Behandlung bestimmter Typen partieller Differentialgleichungen

c) Numerische Methoden der Approximation unter Berücksichtigung der linearen und nichtlinearen Optimierung

d) Moderne Entwicklungen zur Theorie der Interpolation und Extrapolation

e) Numerische Behandlung von allgemeinen Gleichungen und Gleichungssystemen

f) Iterative Behandlung von Funktionalgleichungen.

Insbesondere wurden auch die folgenden geschlossenen Vorträge gehalten, die in der Regel sehr ausführlich diskutiert wurden:

W. Börsch-Supan: Defektabschätzungen, Genauigkeitsbeurteilung der Nullstellen von Polynomen.

R. Nicolovius: Konvergenzverbesserung durch Extrapolation nach der Norm bei Iterationsverfahren.

A. Schönhage: Unitäre Triangulierung beliebiger Matrizen.

H. Rutishauser: Zum Eigenwertproblem für nichtnormale Matrizen.
QR-Transformation und Eigenwertquadrierung.
Treppeniteration an Funktionswerten.

G.W. Veltkamp: Nachiteration der Matrixinversen.

J. Stoer: Charakterisierung der Operatornormen.

F.L. Bauer: Fragen der Pivotwahl bei Eliminationsprozessen.

H.R. Schwarz: Die Reduktion einer symmetrischen Bandmatrix auf tridiagonale Form.

H. Rutishauser: Eigenwerte normaler Matrizen.
Die LR-Transformation für unendliche, positiv definite symmetrische Matrizen.

F.L. Bauer: QR-Transformation mit Nullpunktverschiebung.

F.L. Bauer (München)
A. Ostrowski (Basel)

Mathematisches Forschungsinstitut
Oberwolfach

Tagung über
Numerische Behandlung von Problemen der linearen Algebra
vom 8. - 12. Juni 1964

Protokollführer: Dr. H.R. Schwarz, Zürich

Montag, den 8. Juni

Eröffnung der Tagung um 10.00 Uhr. Herr Prof. Bauer begrüßt kurz die anwesenden Teilnehmer, erörtert sodann Organisationsfragen und gibt eine kurze Übersicht über die zu erwartenden Vorträge. Ferner skizziert er die Problemkreise, welche in den folgenden Tagen im besonderen zur Diskussion stehen sollen. Es sind dies:

A) Auflösung von Gleichungssystemen. Im Vordergrund sollen Fehlerabschätzungen stehen ohne explizite Benützung der Inversen.

B) Fragen der optimalen Pivotwahl bei nicht positiv definiten Matrizen. Die Kriterien haben bezüglich den Gleitkomma-Operationen invariant zu sein. Zur Diskussion sollen das Equilibrieren, optimale Skalierung und minimale Kondition kommen.

C) Eigenwerte von wesentlich nicht symmetrischen Matrizen. Speziell von Interesse ist der Fall, daß nicht sämtliche Eigenwerte reell sind. Adaptierungen des Jacobi-Verfahrens scheinen mögliche Ansätze zu sein.

D) Probleme um den Normbegriff. Für die Operatornormen soll eine axiomatische, innere Charakterisierung erfolgen.

E) Fragen um die LR-Transformation. Wahl der Nullpunktverschiebungen, im besonderen die nicht lokalen Verschiebungen.

Auf besonderen Wunsch von Herrn Prof. Ostrowski wird noch folgendes Thema zur Diskussion vorgeschlagen: Wie organisiert man ein zentrales Recheninstitut? Die lineare Algebra soll bei der allgemeinen Diskussion als Modellfall dienen und das Problem der qualifizierten Kräfte gestreift werden. Selbstverständlich sind Standardprobleme umgehend innert kürzester Frist zu erledigen durch Bereitstellung entsprechender Programme. Doch sofort erhebt sich die Frage nach einer Fehleranalyse, da dem Recheninstitut die Verpflichtung auferlegt werden muß, über die Genauigkeit der Resultate Angaben zu machen. Zusammen mit den Resultaten ist eine Aussage über die Anzahl der richtigen Stellen mitzuliefern. In kritischen Fällen sind die Schwierigkeiten zu analysieren und spezielle Methoden

bereitzustellen, welche von Konditionierungen Gebrauch machen. Schon hier können sich für den Benützer Kostenfragen ergeben. In besonderen Fällen sind den Benützern der Rechenanlage Hinweise dahin zu geben, welche ihrer Messdaten und in welchem Umfang mit erhöhter Genauigkeit zu beschaffen sind, damit überhaupt eine numerische Aussage auf Grund der Messdaten möglich wird. Endlich hat ein Recheninstitut auch solche Probleme zu meistern, welche wissenschaftliche Untersuchungen zur Entwicklung von neuen Verfahren nötig machen.

Die vorgeschlagenen Diskussionsthemata werden von den Teilnehmern ohne weitere Wünsche akzeptiert.

Herr Prof. Bauer richtet an die Teilnehmer einige Worte über den Sinn der Tagung. Der ungenügend entwickelte Stand der numerischen Mathematik in Relation zu den Automaten erfordert Intensivierung der Anstrengungen und Forschungen, wobei das Hauptgewicht auf Fehleranalysis gelegt werden soll.

Nach Festlegung des allgemeinen Stundenplanes betreffend der täglichen Zeiteinteilung in Vorträge plus Diskussionen, bzw. der freien Tätigkeit, sowie nach Abklärung der zuerst zu behandelnden Probleme, wird zum ersten Vortragsbeitrag von Herrn Börsch-Supan übergegangen.

Defektabschätzungen. Genauigkeitsbeurteilung der Nullstellen von Polynomen. von Herrn Börsch-Supan

Problemstellung: Es sei für irgendeine Aufgabe, die sich als Gleichung oder Gleichungssystem formulieren läßt, eine Näherungslösung bekannt, wobei es uninteressant ist, auf welche Weise die Näherung gewonnen wurde. Gesucht ist eine Abschätzung des Fehlers dieser Näherung. Hierzu soll benutzt werden der Defekt (Residuum), welcher definiert ist und berechnet wird als Differenz zwischen beiden Seiten der Gleichungen bei einer Einsetzprobe mit der gegebenen Näherung. Nach Skizzierung einiger spezieller Fehlerabschätzungen mit Hilfe des Defektes (Nullstellen von Funktionen, lineare Gleichungssysteme und Inversion, Matrixeigenwertproblem, Anfangswertproblem bei gewöhnlichen Differentialgleichungen erster Ordnung), betrachtet der Vortragende im besonderen die Fehlerabschätzungen für sämtliche Nullstellen eines Polynoms.

Sind Näherungen ξ_i für sämtliche Nullstellen x_i $(i = 1,2,\ldots,n)$ eines Polynoms $f(x)$ bekannt, so läßt sich dieses bei lauter einfachen Nullstellen mit Hilfe der Lagrangeschen Interpolations-

formel durch die Defekte $f(\xi_i)$ ausdrücken. Zur Eingrenzung einer bestimmten Nullstelle x_j benutzt man die Darstellung

$$f(x) = \prod_{\substack{i=1 \\ i\neq j}}^{n} (x - \xi_i)\,[(x - \xi_j)(1 + \sum_{\substack{i=1 \\ i\neq j}}^{n} \frac{c_i}{x - \xi_i}) + c_i]$$

mit $c_i = f(\xi_i) \prod_{\substack{\nu=1 \\ \nu\neq i}}^{n} (\xi_i - \xi_\nu)^{-1}$, wobei diese Nullstelle nach dem Satz von Rouché im Kreis $|x - \xi_j| \leq \varrho_j$ liegt, falls es eine Größe r_j gibt, für die gilt

(i) $\forall\, i$ mit $i \neq j$: $r_j < |\xi_i - \xi_j|$

(ii) $\eta_j = \sum_{\substack{i=1 \\ j\neq i}}^{n} \frac{|c_i|}{|\xi_i - \xi_j| - r_j} < 1$

(iii) $\varrho_j := |\frac{c_j}{1 - \eta_j}| < r_j$

Zur Bestimmung von r_j kann eine iterative Vorgehensweise verwendet werden, bei der für eine Folge von Versuchswerten die Bedingungen (i) bis (iii) nachgeprüft und nachkorrigiert werden. Eine solche Größe r_j existiert nicht, wenn die Fehler irgendwelcher Näherungen nicht mehr hinreichend klein gegenüber den Abständen zu den anderen Näherungen sind. Doch ist eine Verallgemeinerung der Überlegungen auf Haufen von Wurzeln möglich.

An einem Zahlenbeispiel wird die Güte der erzielten Abschätzungen illustriert und gleichzeitig gezeigt, unter welchen Bedingungen die Abschätzung der Einzelwurzeln zweckmäßig durch eine Abschätzung von Wurzelhaufen zu ersetzen ist.

Der Vortrag sellt einen wertvollen Beitrag dar zum Diskussionsthema über die optimale Organisation eines Recheninstituts.

<u>Diskussion:</u> Obwohl die Methode als eine reine a posteriori - Fehlerabschätzung konzipiert ist, kristallisiert sich im Verlauf der nachfolgenden Diskussion heraus, daß die Aussagen über den Defekt in engen Zusammenhang gebracht werden können mit der Methode von Newton, indem die Werte c_i im wesentlichen Newtonkorrekturen darstellen.

Es erhebt sich eine längere Diskussion um die beste Wahl eines Parameters in der Formel für die iterative Verbesserung der r_j. Es stellt sich heraus, daß die Zielsetzungen verschieden sind, indem Herr Börsch-Supan durch eine kleine Wahl des Parameters wenn möglich in einem einzigen Iterationsschritt durchzukommen versucht, während Herr Bauer die gegenteilige Auffassung vertritt, indem er die kritischen Fälle ins Feld führt. Herr Rutishauser rechnet schließlich einen optimalen Wert aus und stellt ihn zur Diskussion.

Herr Bauer wirft die Frage auf, was man neben a posteriori-Abschätzungen aus der verwendeten Methode zur Fehlerabschätzung gewinnen kann. Falls man Newton anwendet, so weiß man etwas über die Größenordnung der c_i. Ferner liefert die Analysis der Rundungsfehler eine Schranke, welche das Ossilieren in der Methode von Newton richtig erfaßt. Die daraus resultierende stürmische Diskussion mündet aus in Betrachtungen, wann mit doppelter Genauigkeit zu rechnen sei, über das Konvergenzverhalten bei benachbarten Nullstellen, über Kriterien, die höhere Ableitungen benützen.

Herr Ostrowski bemerkt, daß sein Satz über die relative Stetigkeit der Wurzeln einer algebraischen Gleichung in Funktion der Koeffizienten eine theoretisch vollständige Lösung des vom Vortragenden gestellten Problems ist, allerdings sind die Konstanten notwendigerweise zu pessimistisch.

Dienstag, den 9. Juni

Der ganze Vormittag dient der Diskussion von Problemkreis A): Herr Ostrowski stellt folgende Fragen zur Diskussion: a) Wie gut ist die Approximation eines gefundenen Vektors ξ als Lösung des Gleichungssystems $A\xi = \eta$? Eine gangbare Möglichkeit besteht in der Berechnung von A^{-1} (Aufwand $\omega \sim n^3$). Gibt es Möglichkeiten, ohne A^{-1} die Fehlerabschätzung durchzuführen (mit einem Aufwand von $\Theta\omega$) ? Wie groß ist dann Θ ?

b) Es erscheint sehr attraktiv, das GLeichungssystem ohne Inversion anzuführen unter Berücksichtigung des Aufwandes. Läßt sich die Fehlerabschätzung dennoch durchführen? Prinzipielle Frage: Hat der Mathematiker eine Methode zu verwerfen, welche mehr Aufwand erfordert, wenn er die Fehleranalyse auf alle Fälle durchzuführen hat?

c) Gibt es im allgemeinen Fall eine Möglichkeit, die Genauigkeit

der Lösung abzuschätzen, ohne A^{-1} oder deren Norm abzuschätzen?
Herr Rutishauser: Das Gauss'sche Verfahren mit Dreieckszerlegung enthält die benötigte Information im wesentlichen. Es stellt sich nur die Frage nach der Abschätzung der Norm der Inversen.
Herr Bauer skizziert den symmetrisch definiten Fall. Aus der Cholesky-Zerlegung von $A = LL^T$ und aus $||L^{-1}||_2$ gewinnt man eine obere Abschätzung für $\text{lub}_2\ (A^{-1}) \leq ||L^{-1}||_2^2$. Der Verlust in der Abschätzung ist nicht wesentlich. Gegenüberstellung mit dem allgemeinen Fall $A = LR$: Hinweis, daß man mit der entsprechenden Abschätzung $\text{lub}_2\ (A^{-1}) \leq ||L^{-1}||_2\ ||R^{-1}||_2$ viel mehr verlieren kann.
Herr Ostrowski: Abschätzung der Inversen bedeutet einen Rechenaufwand von $\frac{1}{3}\omega$, total also $\frac{2}{3}\omega$.
Herr Bauer: Die Schranke ist zu pessimistisch. Im positiv definiten Fall ist die Überschätzung höchstens ein Faktor n, im allgemeinen Fall hingegen kann der Faktor beliebig groß sein. Ein Mehraufwand muß in Kauf genommen werden, um genügend scharfe Schranken zu erhalten.
Herr Ostrowski: Während der Rechnung könnten beim Vorliegen von R und L laufend die Rundungsfehler abgeschätzt werden.
Herr Bauer: Wilkinson hat das untersucht; Schranken sind schlecht; die erzielte Genauigkeit ist besser als die erzielbare auf Grund von solchen Abschätzungen. Der Vorschlag, über die Rundungsfehler laufend Buch zu führen, muß abgelehnt werden auf Grund von Erfahrungen mit der PERM, da viel zu pessimistische Aussagen resultieren, ein dämpfender Einfluß auf frühere Rundungsfehler unberücksichtigt bleibt.
Herr Rutishauser: Es ist streng zu unterscheiden zwischen parallel laufenden und sich kumulierenden Fehlern. So spielen die Pivotelemente die Rolle von Schlüsselelementen, in denen die Rundung und Auslöschung ungedämpft bleiben. Augenmerk darauf richten.
Herr Ostrowski: A priori - Abschätzungen mit zunächst unbekannten Faktoren, welche sich im Verlauf der Rechnung ergeben, und fortlaufend in Rechnung gestellt werden können.
Herren Börsch-Supan und Bauer: Wilkinson hat solche absolute Abschätzung bereits durchgeführt. Die Norm der Inversen ist die einzige Unbekannte.
Herr Rutishauser formuliert neuen Gesichtspunkt: Genügt es, dem Kunden zu sagen, man habe ein ξ derart gefunden, daß innerhalb der Rechengenauigkeit $A\xi = \eta$ gilt.
Herr Bauer: Für den Physiker kann eine solche Lösung annehmbar

sein, da er dann die Fehler als innerhalb der durch die Messdaten gegebenen Genauigkeit bezeichnen könnte.

Herr Rutishauser: Nachiteration als n^2-Prozeß ist eine Methode, eine annehmbare Lösung zu erzeugen.

Herr Bauer: Unter dem Gesichtspunkt, daß sowohl die Matrix als auch der Konstantenvektor mit Messfehlern behaftet sind, ist das Problem der annehmbaren Lösung von Prager untersucht worden, im Sinne einer Rückwärtsanalysis.

Postulat: Jedem Auftraggeber sollte die Frage gestellt werden, was er unter seiner Lösung verstanden haben will, damit er durch die Beantwortung dieser Frage die Methode selbst wählt.

Herr Bauer: Falls der Einfluß von Messungenauigkeiten in den Koeffizienten von A auf die Lösung verlangt wird, dann ist A^{-1} unbedingt notwendig, um bestmögliche Schranken zu haben. Mit $\|L^{-1}\| \, \|R^{-1}\|$ wird zu schlecht abgeschätzt. Darauf folgt eine Diskussion über die Gegenüberstellung der verschiedenen Varianten der Berechnung von A^{-1}.

Zusammenfassung: Die Dreieckszerlegung soll zur Lösung des Gleichungssystems benutzt werden. Invertierung von L und R.
Weg A: $L^{-1} \, R^{-1} =: A^{-1}$, $\|A^{-1}\|$ bestmögliche Abschätzung des Fehlers.
Weg B: Keine Ausmultiplikation von L^{-1} und R^{-1}. Falls $\|L^{-1}\| \, \|R^{-1}\|$ zu schlechten Wert liefert, dann Nachiteration (n^2-Prozeß).
Im Standardfall ist die ungefähre Inverse X zu bilden und mit $\|I - X A\| = \eta < 1$ zu prüfen. Ohne eine solche Überprüfung kann keine verantwortungsbewußte Garantie gegeben werden. Der Mehraufwand kann nicht umgangen werden. Auf Grund des heutigen Standes konnte keine Möglichkeit aufgezeigt werden, diese Absicherung anders vorzunehmen.

Konvergenzverbesserung durch Extrapolation nach der Norm bei Iterationsverfahren. von Herrn Nicolovius

Wenn bei konvergenter Iteration mittels des Einzelschritt- oder Gesamtschrittverfahrens der Quotient der Normen zweier aufeinanderfolgenden Korrekturen gegen einen Grenzwert zustrebt, (größter Eigenwert, falls dieser reell ist und Eigenvektoren entsprechend der Vielfachheit besitzt), kann man durch einen Schritt nach dem Aitken-Steffenson'schen Verfahren aus drei aufeinanderfolgenden Quotienten nach vorsichtiger Prüfung einen verbesserten Quotienten

q bilden und gemäß

$$\tilde{x} = x_{n-1} + \frac{1}{1-q}(x_n - x_{n-1})$$

extrapolieren. Es wurden praktische Erfahrungen mit aus Differenzenverfahren für $\Delta u = 0$ sich ergebenden, bis zu 500-reihigen Matrizen mitgeteilt: Bei kleinen Matrizen ($n \approx 20$) sinkt die Anzahl der erforderlichen Schritte auf etwa 1/3, bei größeren ($n \approx 500$) auf etwa 1/8, wobei der Genauigkeitsgewinn durch die einzelne Korrektur mit der Größe etwas abnimmt. Es treten bis zu fünf verschiedene Korrekturfaktoren $1/(1-q)$ auf.
Diskussion: In der Diskussion wurde bemerkt, daß der Wynn'sche ε-Algorithmus bessere Ergebnisse als die obige Methode liefern müßte. Ferner wurde auch die Anwendung des vektoriellen Wynn-Algorithmus empfohlen, wobei allerdings mehr als zwei Speichersätze benötigt werden. Es wurde vermutet, daß eventuell auch bei Überrelaxation mit $\omega < \omega_{opt}$ die genannten Verfahren anwendbar sind.

Es folgen Beiträge zum Poblemkreis C.

Unitäre Triangulierung beliebiger Matrizen.

von Herrn Schönhage

Die Methode von Greenstadt, die Quadratsumme der subdiagonalen Elemente vermittels unitären Transformationen gegen Null zu bringen, indem in jedem Schritt ein Element zu Null gemacht wird, braucht nicht zu konvergieren. (Gegenbeispiel). Die Idee von Schönhage besteht darin, die einzelnen Quadratsummen der subdiagonalen Kolonnenelemente derart zu gewichten, daß ein Gefälle von links nach rechts entsteht. Mit

$$\tau_\mu = \sum_{\nu=\mu+1}^{n} |a_{\nu,\mu}|^2$$

verwendet er als Maß für die Abweichung von oberer Dreiecksform

$$\tau(A) = \sum_{\mu=1}^{n-1} \gamma^{\mu-1} \tau_\mu$$

mit einem Gewichtsfaktor $0 < \gamma < 1$. Durch sukzessive Anwendung von zweidimensionalen unitären Transformationen von der Form

$$U_{k,k+1} = (u_{\nu,\mu}) \quad \text{mit}$$

$$u_{kk} = u_{k+1,k+1} = \cos\varphi \qquad u_{\nu,\mu} = \delta_{\mu},_{\mu} \quad \text{sonst}$$

$$u_{k,k+1} = \ell^{-i\alpha} \sin\varphi = -\bar{u}_{k+1,k} ,$$

wird das Maß $\tau(A)$ auf Grund eines induktiven und zugleich konstruktiven Beweises des Satzes von Schur verkleinert. Beginnend mit $A = A_o$ wird in einem Durchgang jeweils

$$A_{k+1} = U_{12}^{*} U_{23}^{*} \cdots U_{n-1,n}^{*} A_k U_{n-1,n} \cdots U_{23} U_{12}$$

gebildet, wobei die Parameter α und φ in den $U_{k,k+1}$ jedesmal $\tau(A)$ minimalisieren.

Trotz des eingeführten Gefälles im Maß $\tau(A)$ scheint die Eindeutigkeit der Dreiecksmatrix nicht gesichert zu sein, da sie sehr wohl möglich von der Wahl von μ abhängig sein kann.

Mittwoch, den 10. Juni.

Im Detail wird dargelegt, wie die unitären Matrizen bestimmt werden, so daß die Abnahme von τ maximal wird. Hierin lassen sich bei geeigneter Fallunterscheidung statt der optimalen Parameter spezielle α und φ angeben, die eine Mindestabnahme garantieren. Das sichert die Konvergenz.

Das Verfahren wird sodann am Gegenbeispiel illustriert und die numerischen Erfahrungen an drei weiteren Beispielen diskutiert. Das Konvergenzverhalten ist leider ziemlich schlecht, auch wenn es nicht gar so schlecht ist, wie eine allzu pessimistische Abschätzung befürchten ließe. Bei normalen Matrizen läßt sich mehr als lineare Konvergenz feststellen, sonst aber wirkt sich der Schurüberschuß $\sum |a_{\nu,\mu}|^2 - \sum |\lambda_i|^2$ nach einigen Durchgängen stark bremsend aus. Das Verfahren zeigt die Tendenz zu stagnieren, bzw. in einen Käfig hineinzulaufen, indem die Abnahme von τ zunächst gut voranschreitet, dann praktisch stehen bleibt, um nach einer Zeit wieder weiterzugehen.

Die Abhängigkeit vom Gewicht μ ist noch nicht weiter untersucht worden.

Diskussion: In der Diskussion wird vorgeschlagen, durch Skalierung dem Stagnieren vorzubeugen. Die Herren Bauer und Stoer sind damit nicht einverstanden. Sie schlagen eine Entzerrung des Eigenvektorsystems vor. Das Maß der Nichtnormalität der Matrix A soll dadurch verkleinert werden. Für die praktische Durchführung dieser Idee

kann allerdings nichts Näheres vorgeschlagen werden. Die diesbezüglichen Normabschätzungen scheinen mehr akademisch-theoretischen Wert zu haben.

Es folgen drei verschiedene Beiträge von Herrn Rutishauser.

Zum Eigenwertproblem für nichtnormale Matrizen.
von Herrn Rutishauser

Es wird folgendes Verfahren zur Bestimmung der Eigenwerte einer Matrix A vorgeschlagen: In einem ersten Schritt werden durch geeignete Skalierungen, d.h. Ähnlichkeitstransformation mit einer Diagonalmatrix D, die Zeilenlängen gleich den entsprechenden Spaltenlängen gemacht. (Dabei wird praktisch $D^{-1}A\,D = A_1$ durch eine Folge von Elementarskalierungen berechnet, von denen jede eine Zeilenlänge gleich der entsprechenden Spaltenlänge macht). Alsdann wird die Matrix A_1 einer solchen Orthogonal- (unitär) Transformation $A_2 = U^T A_1 U$ unterworfen, daß der Kommutator $C = A_1A_1^T - A_1^TA_1$ auf Diagonalform transformiert wird: $C_1 \to C_2 = U^T C_1 U$. Alsdann wird A_2 wieder skaliert, $\to A_3$, dann wieder $C_3 = A_3A_3^T - A_3^TA_3$ diagonalisiert, etc. Bei diesem Prozeß strebt die Spektralnorm von C gegen 0, d.h., A wird immer normaler, was darauf beruht, daß der Schurüberschuß von A_{2k} durch die Skalierung um mindestens (Spektralnorm von $C_{2k})^2/2$ (Schurüberschuß von A_{2k}), (sofern $K \neq 0$). Die schließlich entstehende angenähert normale Matrix kann mit bekannten Methoden behandelt werden. Numerische Experimente mit diesem Verfahren weisen auf eine langsame Konvergenz hin, was auf eine gewisse Schwäche des Schurüberschusses als Maß für die Nichtnormalität hinzuweisen scheint.

Diskussion: In der Konvergenz wird ebenfalls eine Resistenzerscheinung festgestellt. Es werden Zweifel geäußert, ob das Verfahren praktisch überhaupt brauchbar sei. Herr Bauer stellt die prinzipielle Frage, warum denn stets am Jacobi-Verfahren herumgedoktert werde, um ein beliebig langsam konvergentes Verfahren zu entwickeln. Er verweist auf die sicheren Verfahren mit der Hessenbergform, QR-Transformation, Methode des HYMAN-Vektors zur Berechnung der Determinante.

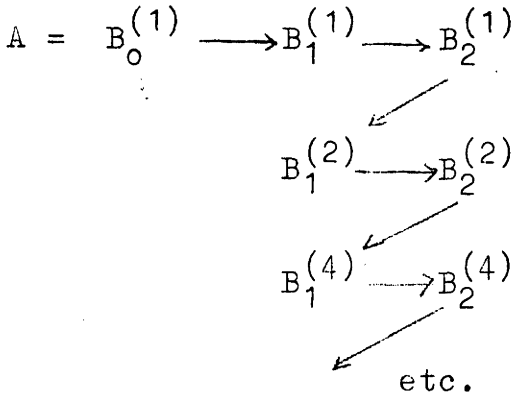
QR-Transformation und Eigenwertquadrierung.
von Herrn Rutishauser

Es sei die Matrix A mit $B_o^{(1)}$ bezeichnet, und die Matrix, welche aus $B_o^{(1)}$ nach k QR-Schritten erhalten ist, sei $B_k^{(1)}$. Es sei weiter

$B_o^{(p)}$ die p-te Potenz von A und das Resultat von k QR-Schritten auf $B_o^{(p)}$ sei $B_k^{(p)}$, dann gilt

$$B_k^{(p)} = \left(B_{k\cdot p}^{(1)}\right)^p .$$

Jeder QR-Schritt, angewendet auf A^p entspricht p QR-Schritten angewendet auf A, welches folgenden Rechenprozeß nahe legt (ein Pfeil nach rechts bedeutet einen QR-Schritt, ein Pfeil nach links unten hingegen eine Matrixquadrierung):



Jede QR-Transformation im letzten Schema stellt eine orthogonale Transformation dar

$$B_2^{(2^k)} = U_k^T \, B_1^{(2^k)} \, U_k \qquad (k = o,1,2,\ldots)$$

Bezeichnet man $A^{(o)} = B_1^{(1)}$ und führt die gleichen Transformationen U_k auf die $A^{(k)}$ aus,

$$A^{(k+1)} = U_k^T \, A^{(k)} \, U_k ,$$

dann sind die $A^{(k)}$ theoretisch die Matrizen, welche sich nach 2^k QR-Schritten aus A ergeben. Daraus folgt die quadratische Konvergenz des Verfahrens. Das Auftreten von großen bzw. kleinen Zahlen in diesem Graeffe-artigen Prozeß kann durch eine geeignete Skalierung umgangen werden. Es zeigt sich, daß nicht die Skaliergrößen selbst, sondern nur deren Quotienten mitgeführt zu werden brauchen.

Diskussion: Herr Bauer bemerkt, daß es vorteilhaft sein kann, die Transformationen U_k zu akkumulieren, beim Stehen der Skalierfaktoren die akkumulierte Transformationsmatrix eventuell noch nachzuorthogonalisieren und dann auf die Diagonalmatrix A anzuwenden.

Treppeniteration an Funktionswerten.

von Herrn Rutishauser

Es wird darauf hingewiesen, daß die durch die Formeln

$$P_k^{(v+1)}(x) := \frac{xP_k^{(v)}(x) - P_{k-1}^{(v)}(x)}{q_k^{(v)}} \qquad (k = 1,2,\ldots n; v = 0,1,\ldots\infty)$$

(wobei $P_k^{(v)}(x) = x^{n-k}+\ldots$ - Polynome vom Grade $n-k$, $P_o^{(v)}(x) \equiv P(x)$ ein festes Polynom vom Grade n, und $P_k^{(o)}(x) = x^{n-k}+\ldots$ willkürlich gewählte Anfangspolynome sind) festgelegte Iterationsvorschrift (sog. freie Treppeniteration von F.L. Bauer) nicht nur in der üblichen Potenzreihendarstellung der beteiligten Polynome $P_k^{(v)}(x)$ durchgeführt werden kann, sondern auch, wenn die $P_k^{(v)}(x)$ durch ihre Werte an $n+1$ fest gewählten Stützstellen $x_o, x_1, \ldots x_n$ dargestellt sind. Man kann nämlich die obgenannte Formel auch mit diesen Stützwerten durchführen; nur die Normierungsvorschrift, daß die höchste Potenz aller beteiligten Polynome den Koeffizienten 1 haben soll, muß durch die Bedingung "$(n-k)$-te dividierte Differenz von $P_k(x) = 1$" ersetzt werden.

Dies gilt insbesondere auch für die "gebundene Treppeniteration", bei der die Polynome $P_k^{(v)}(x)$ untereinander, außer durch die obige Formel auch noch durch

$$P_{k+1}^{(v)} = \frac{P_k^{(v+1)} - P_k^{(v)}}{e_k^{(v)}}$$

verbunden sind, wobei die $e_k^{(v)}$ wieder die Normierungsfaktoren sind, die den höchsten Koeffizienten von $P_{k+1}^{(v)}$ zu 1 machen. Es läßt sich damit offenbar der Euklidische Algorithmus:

$$P_1^{(1)} = \frac{xP_1^{(o)} - P_o^{(1)}}{q_1^{(o)}} \qquad P_2^{(o)} = \frac{P_1^{(1)} - P_1^{(o)}}{e_1^{(o)}}$$

$$P_2^{(1)} = \frac{xP_2^{(o)} - P_1^{(1)}}{q_2^{(o)}}, \qquad \text{etc.}$$

direkt an den Funktionswerten durchführen, wenn nur $P_1^{(o)} = x^{n-1}+\ldots$ und $P_o^{(1)} = x^n+\ldots$ durch die Stützwerte an den Stellen $x_o, x_1, \ldots, x_n$ gegeben sind und auf diese Weise die Koeffizienten $q_o^{(o)}$, $e_o^{(o)}$ des S-Kettenbruches der rationalen Funktion $\frac{P_1^{(o)}}{P_o^{(1)}}$ auf numerisch stabilere Weise berechnen als dies unter Verwendung der Polynom-Koeffizienten möglich wäre.

Donnerstag, den 11. Juni

Nachiteration der Matrixinversen. (Nachtrag zu Problemkreis A)
Der Einfluß von Rundungsfehlern bei der Iteration nach Schulz
von Herrn Veltkamp

Sei A eine nicht singuläre Matrix und sei X_0 eine Näherung von A^{-1} in dem Sinne, daß für $R_0 := I - AX_0$ gilt $\|R_0\| < 1$. Dann ist der Prozeß

$$R_j := I - AX_j$$
$$X_{j+1} := X_j + RX_j$$

quadratisch convergent und $\lim X_j = A^{-1}$. Es wird untersucht, wie der Einfluß von Rundungsfehlern bei Gleitkommarechnung ist.

Angenommen wird, daß die skalaren Produkte mit doppelter Genauigkeit akkumuliert werden. Dann gilt

$$|fl(a + \sum_i^n b_i c_i) - (a + \sum_i^n b_i c_i)| \leq$$

$$\leq \varepsilon_1 \, |a + \sum_i^n b_i c_i| + n\varepsilon_2(|a| + \sum_i^n |b_i c_i|) ,$$

wo ε_1 die Relativgenauigkeit der Gleitkommazahlen ist (2^{-t} bei t Binärstellen) und ε_2 die Relativgenauigkeit des doppelt langen Akkumulators ist (ungefähr 2^{-2t}).

Wird nun gefordert, daß

$$\|A\| \, \|A^{-1}\| \leq 1/(2\varepsilon_1)$$

(mit Schurscher Norm), dann stellt sich heraus, daß der Prozeß erst zu stagnieren anfängt, wenn

$$\|R_j\| \leq 2\varepsilon_1 \|A\| \, \|A^{-1}\|$$

und

$$\frac{\|X_j - A^{-1}\|}{\|A^{-1}\|} \leq 2(\varepsilon_1 + n\varepsilon_2(\sqrt{n} + \|A\| \, \|A^{-1}\|))$$

geworden ist.

Dieses Resultat wird mit einigen Beispielen belegt.

Diskussion: Die Abschätzung über den relativen Fehler stimmt mit $\varepsilon_1 = \varepsilon_2$ (einfach genaue Rechnung) mit dem klassischen Resultat überein. Ferner sieht man, daß man im skalaren Produkt zur Berechnung von I - AX mindestens so viel Stellen mehr mitnehmen muß, wie die Kondition angibt. Vor dem letzten Schritt sollte eine a priori - Abschätzung angewendet werden, damit das Residuum nachträglich nicht nochmals berechnet werden muß.

Ostrowski verweist auf eine seiner Publikationen, unabhängig von Schulz, welche die Verbesserung einer Näherung von A^{-1} auf die Neumannsche Reihe zurückführt.

Herr Bauer: Ein Rechenautomat mit beweglicher Stellenzahl kann so programmiert werden, daß er selbsttätig auf Grund der Rechnung die größere Stellenzahl mitführt.

Zum Problemkreis D folgt der Beitrag

Charakterisierung der Operatornormen im Raum der n-reihigen Matrizen. von Herrn Stoer

Eine Matrixnorm $\nu(A)$ wird Operatornorm genannt, wenn ν durch eine Vektornorm $\|x\|$ im C^n erzeugt wird: $\nu(A) = \mathrm{lub}(A) = \sup\{\|Ax\| \mid \|x\| \leq 1\}$. Bezeichnet man mit $\mathrm{lub}^D(A) := \sup\{\mathrm{Re}\, tr(AB) \mid \mathrm{lub}(B) \leq 1$ die zur Operatornorm lub duale Norm, so hat jede Operatornorm folgende Eigenschaften:

1) $\{A \mid \mathrm{lub}^D(A) \leq 1\} = \mathcal{H}\{A = xy^H \mid \mathrm{lub}^D(A) \leq 1\}$ ($\mathcal{H}(\ldots)$ = konvexe Hülle von (...))
2) $\mathrm{lub}(A) \leq \mathrm{lub}^D(A)$ für alle Matrizen A.
3) $\mathrm{lub}(A) = \mathrm{lub}^D(A)$ für alle Matrizen $A = xy^H$ vom Höchstrang 1.
4) $\mathrm{lub}^D(A) = \inf\{\sum \lambda_i \mid A = \sum \lambda_i x_i y_i^H, \lambda_i \geq 0, \mathrm{lub}(x_i y_i^H) \leq 1\}$
5) $\mathrm{lub}^D(A)$ ist multiplikativ.

Als Konsequenzen ergeben sich die Sätze:

1) $\mathrm{lub}_1(A) \leq \mathrm{lub}_2(A)$ für alle A impliziert $\mathrm{lub}_1(A) = \mathrm{lub}_2(A)$ für alle A.
2) Jede multiplikative Matrixnorm ν, die minimal unter allen multiplikativen Matrixnormen ist, ist eine Operatornorm und umgekehrt.
3) Hat die Matrixnorm ν die Eigenschaften:
   a) ν ist multiplikativ,
   b) $\nu(A) \leq \nu^D(A)$ für alle A,
   c) $\{A \mid \nu^D(A) \leq 1\} = \mathcal{H}\{A = xy^H \mid \nu^D(A) \leq 1\}$,

so ist . Operatornorm.

4) Besitzt die Matrixnorm die Eigenschaften:

a) Es gibt einen Vektor $a \neq 0$, $a \in C^n$, so daß

$$\nu(xy^H u a) \leq (xy^H)\nu(u a^H)$$

für alle $x,y,u \in C^n$ ("schwache Multiplikavität"),

b) $\nu(A) = \nu^D(A)$ für alle $A = xy^H$,

c) $\{A \mid \nu^D(A) \leq 1\} = \mathcal{H}\{A = xy^H \mid \nu^D(A) \leq 1\}$,

so ist ν eine Operatornorm.

Die Nachmittagssitzung ist dem Problemkreis B über Fragen der Pivotwahl bei Eliminationsprozessen gewidmet.

Fragen der Pivotwahl bei Eliminationsprozessen.

von Herrn Bauer

Für positiv definite Matrizen ist die Situation weitgehend geklärt: Nur Diagonalelemente sinnvollerweise als Pivots, a priori Fehler-Analysis ergibt Schranken, die von der Pivotwahl unabhängig sind und erfahrungsgemäß kaum praktisch unterschritten werden, Pivotwahl längs der Diagonale bringt dementsprechend kaum einen Vorteil - hat aber den Nachteil umständlicherer Organisation - und ergibt sich automatisch, wenn man nach dem betragsgrößten Element der ganzen Matrix als Pivot sucht (sog. Maximal-Pivot-Strategie). Bei Gleitkommarechnung tritt in der Schranke für die a priori - Fehleranalysis die Minimalconditionszahl $\inf_{D \neq 0} \|DAD\| \, \|D^{-1}A^{-1}D^{-1}\|$ auf, entsprechend dem Umstand, daß eine optimale Skalierung der Ausgangsmatrix, d.h. Ersetzung von A durch DAD, wo D eine Diagonalmatrix mit 2er bzw. 10er-Potenzen ist, den Gang der numerischen Berechnung nicht beeinflußt.

Bei allgemeinen Matrizen ist die Situation unbefriedigender. Die richtige Pivotwahl bekommt entscheidende Bedeutung für die numerische Stabilität der Rechnung. Die Maximal-Pivot-Strategie hat sich praktisch bewährt, während eine nur zeilen- bzw. spaltenweise Maximal-Pivot-Suche zu Schwierigkeiten führen kann. WILKINSON konnte für die Maximal-Pivot-Strategie Schranken herleiten, die gegenüber dem positiv-definiten Fall einen von n abhängigen Verschlechterungsfaktor zeigen, die Abschätzung läßt erkennen, daß sie zu pessimistisch ist, die Praxis zeigt, daß in aller Regel keine nennenswerte Verschlechterung eintritt.

Unbefriedigend ist, daß die Maximal-Pivot-Strategie von der Skalierung D_1AD_2 der Ausgangsmatrix abhängt und daß durch geeignete

extreme Skalierung jedes von 0 verschiedene Matrixelement zum Pivot gemacht werden kann. Die Ergebnisse der Elimination mit Maximal-Pivot-Strategie sind dementsprechend auch nur befriedigend, wenn extreme Skalierung vermieden wird. Praktisch wird zu diesem Zweck oft eine "Equilibrierung" durch geeignete Skalierung vorgenommen. Es liegt nahe zu vermuten, daß die Maximal-Pivot-Strategie optimal arbeitet, wenn die Matrix so skaliert ist, daß die Minimalkonditionszahl $\inf_{D_1 \neq 0, D_2 \neq 0} \|D_1 A D_2\| \, \|D_2^{-1} A^{-1} D_1^{-1}\|$ erreicht oder gut approximiert wird. Praktische Schwierigkeiten, diese Equilibrierung durchzuführen, die eigentlich vor jedem Eliminationsschritt vorgenommen werden müßte. Eine andere Möglichkeit bestünde darin, eine Pivotwahl vorzunehmen, die von vornherein gegen Skalierung invariant ist. Die Quotienten $a_{ik}a_{lj}/a_{ij}a_{lk}$ von Matrixelementen sind solche Invarianten. Daß solche Quotienten entscheidend sind, zeigt die Fehleranalysis der Gauß-Jordan-Invertierung für den simplen Fall der reellen zweireihigen Matrix $\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$; der Fehler in der berechneten Inversen, hervorgerufen durch den Einfluß von Rundungsfehlern, beträgt bei Gleitkommarechnung, beginnend mit a_{11} als Pivot

$$\begin{pmatrix} 1+3\eta & 1 \\ 1 & 1 \end{pmatrix}\varepsilon, \text{ wenn } |a_{11}a_{22}-a_{12}a_{21}| = |a_{11}a_{22}| + |a_{12}a_{21}|,$$

$$\begin{pmatrix} \eta & 1 \\ 1 & 1 \end{pmatrix}\frac{1+\eta}{1-\eta}\varepsilon, \text{ wenn } |a_{11}a_{22}-a_{12}a_{21}| = ||a_{11}a_{22}|-|a_{12}a_{21}||,$$

wobei $\eta = |a_{12}a_{21}|/|a_{11}a_{22}|$.

Dies zeigt klar, daß die Pivotwahl so erfolgen muß (das heißt eine solche Umstellung der Matrix vorgenommen werden muß), daß η^{-1} maximal wird.

Eine sichere Möglichkeit, im Eliminationsverfahren Schwierigkeiten zu vermeiden, besteht in der Verwendung einer geeigneten linearen Kombination der Matrixzeilen als Eliminationszeilen. Werden als Gewichte der Linearkombination die Elemente der Arbeitsspalte verwendet, so ergibt sich in der Fehleranalysis dieselbe günstige Abschätzung, die für positiv-definite Matrizen gilt. Die Methode stellt sich als Variante einer Reduktion auf Dreiecksgestalt durch orthogonale Transformationen heraus (und liefert auch ein Verfahren zur Ausgleichung ohne Aufstellung der Normalgleichungen). Das Problem taucht hier auf, ob derselbe Effekt für die Fehleranalysis auch durch Verwendung einiger weniger geeigneter Matrixzeilen erreicht werden kann.

Freitag, den 12. Juni

Behandlung des Problemkreises E:

Die Reduktions einer symmetrischen Bandmatrix auf tridiagonale Form. von Herrn Schwarz

Eine geeignete Anordnung von zweidimensionalen Jacobi-Drehungen vermag eine symmetrische Bandmatrix auf tridiagonale Form zu reduzieren, wobei die Bandgestalt ständig bewahrt wird. Das geschilderte Verfahren besteht im wesentlichen darin, daß die Bandbreite systematisch in einem Durchlauf um 1 reduziert wird. Die dabei notwendigerweise auftretenden von 0 verschiedenen Elemente außerhalb des Bandes werden durch zusätzliche geeignete Drehungen nach unten geschoben und schließlich über den Rand hinausgewischt. Eine Analyse des Rechenaufwandes zeigt, daß höchstens

$$n^2\left[4(m-1)+\frac{13}{2}\sum_{k=2}^{m}\left(\frac{1}{k}\right)\right]$$

Multiplikationen für die vollständige Reduktion notwendig sind, wobei n die Ordnung und m(<n) die Bandbreite bedeuten. Die Methode wird als Bindeglied zur Eigenwertberechnung symmetrischer Bandmatrizen zu den bekannten Methoden für tridiagonale Matrizen gedacht. Bei Bestimmung von sämtlichen Eigenwerten spricht der Rechenaufwand mit zunehmender Bandbreite immer deutlicher für die Bandreduktion und nachfolgendem QD-Algorithmus im Vergleich zur LR-Transformation.

Eine andere Variante der Reduktion gestattet, die Verbindung zur Methode von Householder herzustellen.

Diskussion: In der Diskussion schlägt Herr Stoer eine andere Variante vor, welche einen kleineren Rechenaufwand verspricht.

Ferner wird die Frage nach einer analogen Anpassung des klassischen Jacobi-Verfahrens für tridiagonale Matrizen aufgeworfen. Das entsprechende Verfahren dürfte instabil sein.

Weiter wird das Problem gestellt, ob das Verfahren auf symmetrische Matrizen mit systematischen Nullen ausgedehnt werden kann (drei Bänder). Dies scheint nach gemachten Versuchen nicht möglich zu sein.

Eigenwerte normaler Matrizen.
von Herrn Rutishauser

Eine normale Matrix A werde in ihren hermiteschen und schiefsymme-

trischen Teil zerlegt. $A = H + S$. H werde auf Diagonalform transformiert $U^T H U = D$ und mit demselben U die Matrix S transformiert: $U^T S U = S_1$. Die Matrix $A_1 = D + S_1 = U^T A U$ kann nicht beliebig sein, vielmehr muß sie in Kästchen zerfallen, welche sich längs der Diagonale aufreihen und welche von folgender Bauart sind:

$$\begin{pmatrix} \lambda & & T \\ & \ddots & \\ -T^T & & \lambda \end{pmatrix}$$

In der Diagonale stehen lauter gleiche Werte. Nach Subtraktion von λI bleibt eine schiefsymmetrische Matrix, welche nach Giveus auf eine Jacobi-Matrix transformiert werde:

$$\begin{pmatrix} 0 & a_1 & & & & \\ -a_1 & 0 & a_2 & & & \\ & -a_2 & 0 & \ddots & & \\ & & \ddots & \ddots & a_{n-1} & \\ & & & -a_{n-1} & 0 & \end{pmatrix}$$

Die QR-Transformation auf diese Matrix läuft auf den QD-Algorithmus hinaus, welcher in seiner stabilen Form durchführbar ist, da sämtliche q-und e-Werte positiv sind.

Mit der Householder-Transformation an Stelle von Giveus, angepaßt an den schiefsymmetrischen Fall, gewinnt man einiges bezüglich Rechenaufwand.

<u>Die LR-Transformation für unendliche, positiv-definite, symmetrische Matrizen.</u> von Herrn Rutishauser

Sei A eine unendliche hermitesche Matrix mit der Eigenschaft $0 < (A\eta, \bar{\eta}) < M$ für alle η mit $|\eta| = 1$; außerdem sei A vollstetig. Unter diesen Bedingungen ist

1) A nach Choleski in $R^H R$ zerlegbar,
2) $A_1 = RR^H$ existent, hermitesch, zu A ähnlich und wieder vollstetig.
3) Bei Iteration dieses Prozesses nach $A_k \to R_k^H R_k$, $R_k R_k^H \to A_{k+1}$ entsteht eine Folge von Matrizen A_0, A_1, A_2, A_3... mit der Eigenschaft

$$\lim_{k\to\infty} A_k = \text{Diagonalmatrix} = \Lambda .$$

Es ist nicht geklärt, inwieweit die Diagonalelemente von Λ alle Eigenwerte von A sind, oder ob allenfalls einige "ausgelassen" werden. Falls jedoch $\sum_i \sum_k |a_{ik}|^2 < \infty$, sind solche Auslassungen natürlich unmöglich.

Diskussion: Koordinatenverschiebung ist nicht möglich, da dadurch entweder die positive Definitheit oder die Vollstetigkeit zerstört wird. Der Grund liegt darin, daß sich die Eigenwerte der vollstetigen Matrix bei Null häufen.

Falls man nur den größten Eigenwert zu berechnen wünscht, so hat dies mit Abschnitten der unendlichen Matrix zu erfolgen. Zur Verfeinerung ist der Abschnitt zu rändern. Die Methode von Jacobi ist für die geränderte Matrix möglich.

QR-Transformation mit Nullpunktsverschiebung.
von Herrn Bauer

Sei A eine Matrix vom Grad n, A_{11} der (n - 1)-reihige Minor zum Element a_{nn}. Sei $f(\lambda) = \det(A - \lambda I)$ und $g(\lambda) = \det(A_{11} - \lambda I)$. Während die Newton-Korrektur mit $f(\lambda)$, $-f(\lambda)/f'(\lambda)$ die unerwünschte Eigenschaft hat, nicht einmal dann sofort einen Eigenwert zu liefern, wenn A diagonal ist, zeigt die Newton-Korrektur mit $f(\lambda)/g(\lambda)$, $-[f(\lambda)/g(\lambda)]/[f(\lambda)/g(\lambda)]'$ eine bessere Wirkung je näher A bereits der Diagonalgestalt ist und bietet sich deshalb für die Nullpunktsverschiebung im Zusammenhang mit der LR-Transformation an. Es zeigt sich nun, daß man diese Newton-Korrektur gerade dann erhält, wenn man an $A - \lambda I$ zwei LR-Schritte (für symmetrisches A äquivalenterweise einen QR-Schritt) durchführt und zwar als a_{nn}-Element der sich ergebenden Transformierten. Dies mag die beobachtete gute Wirkung der auf dem a_{nn}-Element beruhenden Nullpunktsverschiebung, insbesondere bei der QR-Transformierten, erklären.

Diskussion: Das Verfahren ist sogar kubisch konvergent. Ein Einwand gegen die Wahl der Nullpunktsverschiebung mit a_{nn} besteht darin, daß bei reeller Matrix die Verschiebungen so lange reell bleiben, als nicht komplex gerechnet wird.

Herr Rutishauser diskutiert die Nullpunktsverschiebung an einem numerischen Beispiel für die LR-Transformation. Insbesondere betrachtet er den Fall, wie kleine Diagonalelemente gegen das untere Ende durchgedrückt werden, wobei naturgemäß Auslöschung führender

Stellen stattfinden muß. Es konnte jedoch noch nie eine numerische Instabilität festgestellt werden.

Herr Bauer zweifelt die Stabilität der Akkumulierung der Transformationsmatrix in diesem Falle an. Bereits die Stabilität für die Eigenwerte muß als besonderer Glücksfall angesehen werden.

Der letzte Nachmittag wird nochmals dem Problem der optimalen Organisation eines Rechenzentrums gewidmet. Die Diskussion (Herr Ostrowski) geht vom Standpunkt aus, daß das Recheninstitut eine Haftung für die gelieferten Resultate zu übernehmen hat und somit für eine entsprechende Fehleranalyse verpflichtet ist, was Konsequenzen mit sich bringt. Als Modell dient wieder die Auflösung von linearen Gleichungssystemen.

I) Standardprobleme:

Für solche Probleme erhält der Benutzer der Rechenanlage ein Merkblatt, auf dem er sämtliche Angaben findet, wie Angaben über die Ordnung, eventuelle Beschränkungen bezüglich des Speicherbedarfs, Zeitaufwand in Funktion der Ordnung, Angaben über die Art der Datenlieferung (nur von Null verschiedene Elemente der Matrix). Solche Probleme sind mechanisch zu erledigen, und zwar ohne hochqualifizierte Leute. Eine genaue Schätzung des Umfanges von Standardproblemen hat selbstverständlich zu erfolgen. Insbesondere ist neben der Ordnung n der Matrix noch die Anzahl N der verschiedenen Matrixelemente wichtig. Der Benutzer hat selbst zu entscheiden, welche Art der Lösung er wünscht:

Fall a) $|A\xi_0 - \eta| \ll 1$, d.h. das Residuum soll möglichst klein sein. In diesem Fall wird folgendes Vorgehen vorgeschlagen: Die Dreieckszerlegung der Matrix A und die Berechnung der ersten Näherung sollen mit einfacher Genauigkeit erfolgen. Die Berechnung des Residuums hat auf jeden Fall mit doppelter Genauigkeit zu erfolgen. Die Nachiteration geschieht dann wieder mit einfacher Genauigkeit. Die Nachiteration soll maximal viermal wiederholt werden. Dem Kunden soll das Residuum mitgeteilt werden, und die Maschine soll dem Kunden neben den Resultaten gedruckte Information in Form von Text mitliefern. Es könnte ja sein, daß die Nachiteration nicht genügend gut konvergiert.

Andererseits sollte nach Herrn Rutishauser das maximale Residuum nicht durch den Kunden umschreibbar sein, sondern das Rechenzentrum sollte in jedem Fall stereotyp die Lösung bis zur Übereinstimmung $A\xi = \eta$ innerhalb Maschinengenauigkeit treiben ($A\xi$ exakt, bzw. doppelt genau), um für eine gleichmäßige Behandlung aller primiti-

ven Standardprobleme zu ermöglichen. Kunden, die Fehleranalysen ausführen können und den Rechenprozeß damit steuern wollen, sind auf eine "höhere Ebene" zu verweisen.

Fall b) $|\xi_o - \xi| \ll 1$. Die berechnete Lösung soll mit vorgeschriebener Genauigkeit mit der exakten Lösung übereinstimmen. Falls diese Forderung vorliegt, ist A^{-1} unbedingt zu berechnen, damit die Einflußzahlen greifbar sind. Im Merkblatt ist darauf hinzuweisen, daß die Abschätzungen pessimistisch sind. Ist die vom Kunden gestellte Forderung auf Grund der Kondition nicht zu erfüllen, so ist ihm mitzuteilen, daß er die Lösung mit einfacher Rechengenauigkeit nicht haben kann. Kostenfrage, ob er bereit ist, ein Vielfaches auszulegen.

Falls in einem der beiden Fälle mehr verlangt wird, als das Standardprogramm zu leisten im Stande ist, muß das Problem an die nächst höhere Stufe weitergeleitet werden.

Zwischendurch wird die Methode der konjugierten Gradienten als Muster zur Lösung von Gleichungen mit einer Matrix A mit vielen Nullen diskutiert. Auch hier ist eine Nachiteration nötig, indem der Residuenvektor doppelt genau bestimmt wird. Aus den Skalarprodukten mit den früheren Residuenvektoren lassen sich Korrekturen bestimmen. Es wird weiter die Feststellung gemacht, daß eine Nachiteration mit derselben Methode rationeller und billiger ist, als der Übergang zu einem anderen Verfahren.

II) Spezialisierte Halbstandardprobleme:

Auf dieser Stufe ist ein Mathematiker nötig. Diese Probleme sind etwas spezieller Natur, doch lassen sie sich meistens durch Standardprogramme lösen. Beispiele: Inversion mit Nachiteration; Zerlegung eines Gleichungssystems in Blöcke. Jedenfalls funktioniert ein normales Standardprogramm nicht, der Kunde kann nicht sofort bedient werden, vielmehr hat er einen neuen Auftrag zu erteilen. Dazu hat der zuständige Mathematiker das Problem richtig einzuschätzen und eventuell auf Grund der negativen Erfahrungen beim Versuch, das Problem mit Standardmethoden zu lösen, abzuschätzen, welche finanziellen Konsequenzen für den Kunden entstehen.

Die Frage, nach der Erhebung von Pauschalpreisen für solche Probleme ruft ihrerseits dem Problem, ob sich das Recheninstitut bei einer Versicherung gegen Kostenüberschreitungen versichern lassen kann. Zur Feststellung der zu erwartenden Zahl von Kostenüberschreitungen dürften jedoch nicht stochastische Überlegungen über die Art der pathologischen Fälle zugrunde gelegt werden, sondern

statistische Untersuchungen. Anhand der Matrixinversion wird festzustellen versucht, ob die Information $|XA - I| = \eta$ mit $0,1 \leq \eta \leq 10$ hinreichend sei, die Schwierigkeit des Problems und die Kondition richtig zu beurteilen. Es kristallisiert sich die Ansicht heraus, daß solche Information keine hundertprozentige Sicherheit bietet, daß Wiederholung der Rechnung mit doppelter Genauigkeit erfolgreich ist.

Die Tagungsleiter danken Herrn H.R. Schwarz für seine sorgfältige und hingebungsvolle Führung des Protokolls.