Numerische Mathematik 3, 226—240 (1961)

Jordan-Elimination und Ausgleichung nach kleinsten Quadraten*

Von

PETER LÄUCHLI

1. Einleitung

Die klassische Ausgleichsrechnung nach der Methode der kleinsten Quadrate hat bis heute ihre Bedeutung beibehalten, auch wenn sie in neuerer Zeit durch das Studium anderer Ausgleichsprinzipien, so vor allem des Tschebyscheffschen, teilweise in den Hintergrund gedrängt wurde. Die Stärke der Gaußschen Methode liegt unter anderem darin, daß eine geschlossene elementare Theorie vorliegt, welche die Lösungen explizit als lineare Funktionen der Meßgrößen gibt.

Ist nun auch die Theorie einfach, so treten doch sofort ernsthafte Probleme auf, wenn es darum geht, umfangreichere Ausgleichungen tatsächlich durchzuführen. Solche Aufgaben können sich — etwas salopp ausgedrückt — bei der Behandlung gut- oder bösartig verhalten.

Genauer: Die algebraische Lösung der Ausgleichsaufgabe — n Gleichungen für m Unbekannte, $n > m$ — führt über die Forderung, daß die Quadratsumme (v, v) der Verbesserungen minimal werden soll, auf ein System von m linearen Gleichungen (Gaußsche Normalgleichungen). Bekanntlich sind die Schwierigkeiten der numerischen Gleichungsauflösung größer bei Systemen schlechter Kondition, das heißt dann, wenn das Verhältnis zwischen absolut größtem und absolut kleinstem Eigenwert der Koeffizientenmatrix groß ist.

In diesem Falle sind dann auch die Unbekannten bezüglich gewisser Achsenrichtungen des „Fehlerellipsoides" $(v, v) = \text{konst.}$ viel schlechter bestimmt als für andere Richtungen. Und schließlich ist ein schlecht konditioniertes System auch dadurch charakterisiert, daß die Ebene F', welche von den Spaltenvektoren der Fehlergleichungsmatrix aufgespannt wird, schlecht definiert ist, da diese Vektoren fast abhängig sind.

In der vorliegenden Arbeit wird nun ein Verfahren beschrieben, welches speziell auf schlecht konditionierte Systeme zugeschnitten ist. Man hat natürlich zu beachten, daß bei vielen Fällen, welche numerisch sehr schlecht liegen, eben schon die Aufgabe nicht gut gestellt ist. Dennoch kann die Diskussion dieser Zusammenhänge auch von theoretischem Interesse sein.

Die vorgeschlagene Methode verlangt zu Beginn eine gewisse Transformation des Fehlergleichungssystems (Jordan-Elimination) mit einem Rechenaufwand von der selben Größenordnung wie beim Aufstellen der Normalgleichungen. Anschließend werden mit irgend einem iterativen Verfahren die transformierten

* Bei der vorliegenden Arbeit handelt es sich um die leicht gekürzte Fassung eines Manuskriptes, welches auf der Hauptbibliothek der ETH, Zürich, eingesehen werden kann.

Normalgleichungen behandelt, ohne daß diese explizit gebildet werden müssen. Dabei gewinnt man folgende Vorteile:

1. Das transformierte Normalgleichungssystem kann theoretisch nicht beliebig schlecht konditioniert sein, gleichgültig wie schlecht auch das ursprüngliche System war. Genauer: Es wird eine nur von n und m abhängige universelle obere Schranke für die Kondition des transformierten Systems angegeben werden.

2. Bei der iterativen Behandlung der transformierten Normalgleichungen ergeben sich in jedem Schritt ohne wesentlichen Mehraufwand beidseitige Schranken für die zu minimalisierende Größe (v, v). Diese Schranken werden nun aus dem oben angeführten Grunde, gerade bei schlechtkonditionierten Systemen, meistens schon nahe zusammenrücken, bevor die Iteration an den Unbekannten steht. Man wird den Prozeß dementsprechend früher abbrechen können, da offensichtlich ein Weiterrechnen gar nicht sinnvoll wäre.

Natürlich werden auch bei unserer Methode die prinzipiellen numerischen Schwierigkeiten nicht weggezaubert. Es zeigt sich aber doch, daß in manchen Fällen noch brauchbare Resultate gewonnen werden können, wo dies auf normalem Wege beim Rechnen mit der üblichen Stellenzahl nicht mehr möglich ist. Dabei kann es vorkommen, daß wenigstens ein Teil der nachstehend beschriebenen Jordan-Transformation stabil verläuft, und die Verbesserungen (d.h. der Lösungspunkt im n-dimensionalen Raum) gut herauskommen, nicht aber die eigentlichen Unbekannten der vermittelnden Ausgleichung (Punkt im m-dimensionalen Raum).

Es sei hier gleich als Resultat einiger Versuche vorweggenommen, daß sich die Methode der konjugierten Gradienten für die iterative Behandlung der transformierten Gleichungen ganz besonders bewährt hat. Die (v, v)-Schranken rücken im allgemeinen schon nach wesentlich weniger als m Schritten (für die exakte Lösung sind theoretisch m Schritte erforderlich) so nahe zusammen, daß der Prozeß abgebrochen werden kann.

2. Die Transformation der Fehlergleichungen

Es möge ein überbestimmtes lineares Gleichungssystem vorliegen, bestehend aus n Gleichungen in den m Unbekannten $y_1, \ldots, y_m$ $(n > m)$, vektoriell geschrieben

$$C\,y - l = 0.$$

Von der rechteckigen Matrix C wollen wir ein für allemal voraussetzen, daß sie den Rang m habe. Die obigen Gleichungen lassen sich im allgemeinen nicht erfüllen, sondern es werden auf der rechten Seite gewisse Residuen v_i übrigbleiben:

$$C\,y - l = v, \tag{2.1}$$

und die Ausgleichung nach kleinsten Quadraten besteht bekanntlich darin, die y_i so zu bestimmen, daß (v, v) minimal wird. Die v_i haben den Charakter von Verbesserungen, welche an den Meßwerten l_i anzubringen sind. (2.1) nennt man die Fehlergleichungen.

Zur geometrischen Interpretation im R^n betrachten wir die m-dimensionale Ebene F', welche von den m Spaltenvektoren von C aufgespannt wird. (Unter einer „Ebene", welche parallel zu einem Unterraum U liegt, ist natürlich eine Restklasse mod U zu verstehen.) Die Aufgabe der Ausgleichsrechnung ist es dann, denjenigen Punkt $x = C\,y$ von F' zu bestimmen, welcher den kleinsten Abstand $|v| = |x - l|$ von einem gegebenen Punkt l außerhalb F' hat.

Für die algebraische Bestimmung der Unbekannten y_i wird die Tatsache benützt, daß v senkrecht auf F' stehen muß, also $C^T v = 0$ (die Transponierte einer Matrix A bezeichnen wir mit A^T), oder:

$$C^T C\, y - C^T l = 0 \qquad \text{(Gaußsche Normalgleichungen).} \tag{2.2}$$

F' ist also der Bildbereich der regulären linearen Abbildung

$$x = C\, y \tag{2.3}$$

des R^m in den R^n. Betrachten wir umgekehrt die Abbildung $y = C^T x$ von R^n auf R^m, so ist deren Kern das orthogonale Komplement zu F' und der Kern F'' von

$$y = C^T x - C^T l \tag{2.4}$$

die $(n-m)$-dimensionale Parallelebene durch den Punkt l. Diese beiden total senkrechten Ebenen F' und F'', welche einzig den Lösungspunkt s der Ausgleichungsaufgabe gemeinsam haben, spielen eine wichtige Rolle bei den Abschätzungen für (v, v) (s. Abschnitt 4).

Es sollen nun in der Abbildung (2.3), welche ausgeschrieben

$$x_i = \sum_1^m c_{ik}\, y_k \quad \text{für} \quad i = 1, \ldots, n \tag{2.5}$$

lautet, die y_k gegen m der Variablen x_i ausgetauscht werden. Einen einzelnen solchen Austausch, z.B. von y_q mit x_p, nennen wir einen *Jordan-Schritt* (s. [5] oder [7])*. Dieser Schritt besteht darin, daß die p-te Gleichung (2.5) nach y_q aufgelöst und der erhaltene Ausdruck in die übrigen Gleichungen eingesetzt wird.

Dabei entsteht aus der Matrix C eine neue Matrix C', deren Elemente nach folgenden Rechenregeln bestimmt sind:

$$c'_{ik} = \begin{cases} 1/c_{pq} & \text{für} \quad i = p, \quad k = q \\ c_{iq}/c_{pq} & \text{für} \quad i \neq p, \quad k = q \\ -c_{pk}/c_{pq} & \text{für} \quad i = p, \quad k \neq q \\ (c_{ik} - c_{iq}\, c_{pk}/c_{pq}) & \text{für} \quad i \neq p, \quad k \neq q. \end{cases} \tag{2.6}$$

Das Element c_{pq}, welches am Kreuzungspunkt der p-ten Zeile und der q-ten Kolonne von C steht, heißt *Pivot* des betreffenden Austauschschrittes und muß natürlich $\neq 0$ sein.

Aus unserer Voraussetzung, wonach die Matrix C den Rang m haben soll, folgt unter Benützung des Steinitzschen Austauschsatzes, daß es möglich ist, alle y_k als unabhängige Variable zu eliminieren. Dabei ist zunächst die Frage noch offen, gegen welche der x_i ausgetauscht werden soll. Man hat lediglich die Pivots so zu wählen, daß sie nicht verschwinden und in einer Zeile und Kolonne stehen, welche bis dahin noch an keinem Austausch beteiligt waren. Das heißt, daß die Pivots zueinander in „Turmstellung" (Schachspiel) stehen müssen. Ein vernünftiges Auswahlprinzip scheint dieses zu sein, daß unter allen noch zur Konkurrenz zugelassenen Elementen das absolut größte genommen wird, da durch das Pivot dividiert werden muß. Auf diese Frage wird in Abschnitt 3 noch näher eingetreten.

Diese Jordan-Elimination steckt in irgendeiner Form in sehr vielen Prozessen der numerischen linearen Algebra, sei es bei der Auflösung von linearen Gleichungs-

* Literaturverzeichnis am Ende der Arbeit.

systemen nach dem Gaußschen Algorithmus, bei der Matrizeninversion, oder im Simplex-Algorithmus bei der linearen Programmierung.

In den folgenden Überlegungen wollen wir zur Vereinfachung immer annehmen daß gerade die ersten m von den x_i ausgetauscht werden, so daß als Pivots der Reihe nach $c_{11}, \ldots, c_{mm}$ auftreten. Dieser Fall wird natürlich in Wirklichkeit selten vorliegen, jedoch ist es vorteilhaft, bei der praktischen Durchführung durch entsprechende Vertauschungen von Zeilen und Kolonnen diese natürliche Anordnung herzustellen, da nachher mit der transformierten Matrix weitergerechnet wird. Man hat sich dabei lediglich die Permutationen der Indizes zu merken um am Schluß die Komponenten der Lösungsvektoren identifizieren zu können.

Unter dieser Voraussetzung sollen die ersten m Komponenten von x zu einem Vektor z zusammengefaßt werden, die restlichen zu einem Vektor w. Aus der Abbildung (2.3) wird dann, nach Ausführung der m Jordan-Schritte (im folgenden kurz Jordan-Transformation genannt), die folgende:

$$\begin{bmatrix} y \\ \hline w \end{bmatrix} = \begin{bmatrix} A \\ \hline B \end{bmatrix} \cdot [z] = H \cdot z. \tag{2.7}$$

Wir brauchen vor allem den unteren Teil mit der $(n-m)\times m$-Matrix B. Man erhält also die Punkte von F', indem man die ersten m Komponenten, eben die z_k, beliebig wählt und die restlichen, nämlich die w_i mit Hilfe von (2.7) durch jene ausdrückt:

$$w = B z. \tag{2.8}$$

Nun üben wir unsere Jordan-Transformation in genau derselben Weise auf die Abbildung (2.4) aus. Da aber deren Matrix gerade die transponierte von (2.3) ist (duale Abbildung), besteht auch zwischen den Jordan-transformierten ein sehr enger Zusammenhang: Infolge der Unsymmetrie in der Behandlung der Pivotkolonne und -zeile in (2.6) sind A und B durch ihre Transponierte bzw. Negativ-Transponierte zu ersetzen. Der Konstantenvektor $-C^T l$ von (2.4) ist wohl in die Austauschschritte, nicht aber in die Pivotauswahl einzubeziehen. Den Vektor, der nach der Transformation an der Stelle von $-C^T l$ steht, nennen wir $-d$. So entsteht die folgende Abbildung:

$$[z] = \left[\begin{array}{c|c|c} A^T & -B^T & -d \end{array}\right] \cdot \begin{bmatrix} y \\ \hline w \\ \hline 1 \end{bmatrix} \tag{2.9}$$

Um einen Punkt aus F'' zu erhalten, hat man $y=0$ zu setzen, das heißt aber nach (2.9), daß die letzten $(n-m)$ Komponenten, nämlich die w_i, frei gewählt werden dürfen, und daß sich dann die restlichen m, die z_k, aus dem rechten Teil von (2.9)

$$z = -B^T w - d \tag{2.10}$$

ergeben.

Der Lösungspunkt s der Ausgleichsaufgabe muß schließlich als Schnittpunkt von F' und F'' die Eigenschaft haben, daß die beiden Teilvektoren z und w gleichzeitig die Gleichungen (2.8) und (2.10) erfüllen. Durch Elimination eines der beiden erhält man sofort ein Gleichungssystem für die Komponenten des anderen:

$$z = -B^T B z - d \qquad (2.11)$$

oder:

$$N z + d = 0 \qquad (2.12)$$

mit

$$N = E_m + B^T B$$

$$w = -B B^T w - B d \qquad (2.13)$$

oder:

$$M w + B d = 0 \qquad (2.14)$$

mit

$$M = E_{n-m} + B B^T$$

(E_j ist die j-reihige Einheitsmatrix).

Die Gleichungen für z aufstellen heißt nun: die ursprüngliche Fehlergleichungsmatrix C durch

$$C_1 = \begin{array}{|c|}\hline E_m \\ \hline B \\ \hline \end{array}$$

ersetzen. Die Unbekannten fallen dann eben mit dem oberen Teil z des Vektors x der Meßvariablen zusammen. Die andere Variante hingegen ist äquivalent mit der Formulierung desselben Problems als solches der bedingten Ausgleichung. Die Transponierte der Matrix der Bedingungsgleichungen ist dann:

$$C_2 = \begin{array}{|c|}\hline B^T \\ \hline -E_{n-m} \\ \hline \end{array}$$

und ihre Spalten sind orthogonal zu denen von C_1. Der Vektor der Korrelaten fällt, bis auf eine konstante Verschiebung, mit $-w$ zusammen.

Bei dieser Betrachtungsweise kommt die in [4] angetönte Dualität zwischen vermittelnder und bedingter Ausgleichung besonders deutlich zum Ausdruck. Wir werden uns allerdings im folgenden auf die vermittelnde Ausgleichung beschränken, das heißt nur das Gleichungssystem für z diskutieren, da doch in den meisten praktischen Fällen sehr viele überschüssige Messungen vorliegen ($n > 2m$) und die andere Variante nichts wesentlich neues liefert. Es ist übrigens klar, daß die entsprechenden Transformationen auch gemacht werden können, wenn das Problem ursprünglich als bedingtes formuliert ist.

Die Form (2.11) des transformierten Normalgleichungssystems legt es nahe, zur Auflösung ein Iterationsverfahren zu wählen. Dabei kommt weniger die primitive Iteration mit der Matrix $-B^T B$ in Frage, als vielmehr eine der bekannten feineren Relaxationsmethoden. Im Abschnitt 5 wird ein solches Verfahren genau beschrieben.

Des weiteren ist zu beachten, daß die sicher in irgendeiner Form auftretende Iteration zweckmäßigerweise in die beiden Schritte (2.8) und (2.10) aufgespalten

wird, damit das Matrizenprodukt $B^T B$ nicht berechnet zu werden braucht. Das heißt aber, daß in jedem Schritt des Rechenprozesses je ein Punkt aus F' und aus F'' auftritt. Diesen Umstand werden wir in Abschnitt 4 zur Gewinnung von beidseitigen Schranken für (v, v) ausnützen.

3. Die Kondition des transformierten Systems

Es ist im allgemeinen nicht so einfach, Angaben über die Kondition einer Matrix zu machen, da vor allem eine positive untere Schranke für die Beträge der Eigenwerte nicht ohne weiteres angegeben werden kann. Diese Schwierigkeit fällt hier dahin, denn trivialerweise sind alle Eigenwerte von $N = E + B^T B$ mindestens gleich 1. Für $n < 2m$ ist diese Schranke sicher auch das Minimum, da dann $B^T B$ immer singulär ist.

Eine obere Schranke für die Eigenwerte ist bei Normalgleichungsmatrizen leicht anzugeben, da diese symmetrisch und positiv definit sind. Wir werden sogar eine universelle Schranke finden, welche nur von n und m, nicht aber von der Matrix C abhängt (immer vorausgesetzt, daß diese den Rang m hat).

Zunächst soll folgende wichtige Eigenschaft der Jordan-Transformation bewiesen werden, welche zwar nicht weiter benützt wird, aber doch von Interesse ist:

Satz. *Die Matrix B ist dem Unterraum F', welcher durch die Spalten von C aufgespannt wird, eineindeutig zugeordnet.*

Wenn man zur Vereinfachung wieder voraussetzt, daß die Pivots in natürlicher Reihenfolge in der Diagonale des oberen Quadrates P von C stehen, kann die Jordan-Transformation folgendermaßen dargestellt werden (s. (2.7)):

$$C = \begin{array}{|c|}\hline P \\ \hline Q \\ \hline \end{array} \xrightarrow{J} H = \begin{array}{|c|}\hline A = P^{-1} \\ \hline B = Q P^{-1} \\ \hline \end{array}$$

Das obere Quadrat A der Transformierten H wird erst wieder bei der Berechnung von y gebraucht und soll vorläufig außer acht gelassen werden.

Nun wird der Unterraum F' genau dann festgehalten, wenn man C mit einer regulären $(m \times m)$-Matrix R transformiert: $C_1 = CR$. Dann ist aber

$$B_1 = Q_1 P_1^{-1} = Q R (P R)^{-1} = Q P^{-1} = B.$$

Umgekehrt folgt aus

$$B_1 = B$$

$$Q_1 P_1^{-1} = Q P^{-1}$$

und mit

$$Q_1 = Q P^{-1} P_1,$$

$$P_1 = P P^{-1} P_1$$

$$C_1 = CR \quad \text{mit} \quad R = P^{-1} P_1,$$

womit die Behauptung bewiesen ist.

Wenn nun schon die Matrix B nur von der Ebene F' abhängt, dann gilt dies vielmehr auch für die Eigenwerte von $B^T B$. So liegt es nahe, dem Sachverhalt die folgende geometrische Deutung zu geben:

Für ein x aus F', das wir wieder aus den beiden Teilvektoren z und w zusammensetzen, gilt, wie in Abschnitt 2 gezeigt, $w = Bz$. Damit ist

$$\frac{(x,x)}{(z,z)} = 1 + \frac{(w,w)}{(z,z)} = 1 + \frac{(z, B^T B z)}{(z,z)} = \frac{(z, Nz)}{(z,z)}. \tag{3.1}$$

Diese Gleichung sagt, daß der größte Eigenwert von N gleich $1/c^2$ ist, wobei $c = |z|/|x|$ gleich dem Kosinus des größten Winkels ist, den ein Vektor x aus F' mit seiner Projektion z auf die erste m-dimensionale Koordinatenebene bilden kann.

Bei der Jordan-Transformation eine gute Pivotauswahl treffen heißt somit: diejenige m-dimensionale Koordinatenebene suchen, welche mit F' einen möglichst kleinen Winkel bildet. Und die Behauptung, daß die Eigenwerte von N beschränkt seien, bedeutet, daß F' nicht gleichzeitig auf allen m-dimensionalen Koordinatenebenen „beinahe" senkrecht stehen kann.

Für den Beweis der erwähnten Behauptung hat man die Beträge der Matrixelemente in jedem Schritt der Jordan-Transformation abzuschätzen:

Nach dem j-ten Schritt seien die Matrixelemente $c_{ik}^{(j)}$, mit $c_{ik}^{(0)} = c_{ik}$, $c_{ik}^{(m)} = b_{ik}$. Man hat sich bei der Diskussion auf die Elemente von der $(j+1)$-ten Zeile an zu beschränken.

Es soll nun durch Induktion nach j bewiesen werden, daß

$$|c_{ik}^{(j)}| \leqq 2^{j-k} \quad \text{für} \quad \begin{cases} i = j+1, \ldots, n \\ k = 1, \ldots, j. \end{cases} \tag{3.2}$$

Die Rechenvorschrift für einen Jordan-Schritt lautet:

Das Element $\underset{\substack{i=j+1,\ldots,n \\ k=j+1,\ldots,m}}{\operatorname{Max}} |c_{ik}^{(j)}|$ ist durch Zeilen- und Spaltenvertauschungen an die Stelle $(j+1, j+1)$ zu bringen.

Dann bilde gemäß (2.6):

$$c_{i,j+1}^{(j+1)} = -\, c_{i,j+1}^{(j)} / c_{j+1,j+1}^{(j)} \quad \text{für} \quad i = j+2, \ldots, n$$

$$c_{ik}^{(j+1)} = c_{ik}^{(j)} + c_{j+1,k}^{(j)} \cdot c_{i,j+1}^{(j+1)} \quad \text{für} \quad \begin{cases} i = j+2, \ldots, n \\ k = 1, \ldots, j, j+2, \ldots, m. \end{cases}$$

Da nun wegen der Pivotauswahl

$$|c_{i,j+1}^{(j+1)}| \leqq 1 \quad \text{für} \quad i = j+2, \ldots, n,$$

folgt aus Induktionsvoraussetzung und Rechenvorschrift sofort

$$|c_{ik}^{(j+1)}| \leqq 2^{j-k} + 2^{j-k} = 2^{j+1-k} \quad \text{für} \quad \begin{cases} i = j+2, \ldots, n \\ k = 1, \ldots, j+1. \end{cases}$$

Die Verankerung für $j = 0$ ist trivial.

Es gilt somit insbesondere $|b_{ik}| \leqq 2^{m-k}$.

Eine obere Schranke für die Eigenwerte λ der positiv definiten Matrix $B^T B$ wird durch deren Spur gegeben. Diese Abschätzung ist zwar im allgemeinen recht grob. Sie führt jedoch auf einfach gebaute Ausdrücke, welche für unsere theoretischen Überlegungen gut geeignet sind. Zudem erfüllt sie die bei einer

der in Abschnitt 5 angetönten numerischen Methoden unerläßliche Bedingung, daß die Elemente von $B^T B$ nicht berechnet werden müssen.

So erhält man:

$$\lambda \leqq \mathrm{Sp}\,(B^T B) = \sum_{i,k} b_{ik}^2.$$

Diese Größe wird schließlich nach der obigen Ungleichung eingeschränkt:

$$\sum_{i,k} b_{ik}^2 \leqq \sum_{k=1}^{m} (n-m) \cdot 2^{2(m-k)} = \tfrac{1}{3}(n-m)\,(4^m - 1). \tag{3.3}$$

Damit haben wir den folgenden

Satz. *Der größte Eigenwert von $N = E + B^T B$, und damit auch die Kondition dieser Matrix ist höchstens gleich*

$$1 + \tfrac{1}{3}(n-m)\,(4^m - 1) \approx \tfrac{1}{3}(n-m)\,4^m.$$

Man beachte, daß der angegebene Ausdruck nur mit der Anzahl der Unbekannten, nicht aber mit der Anzahl der Fehlergleichungen stark anwächst.

Diese Schranke ist übrigens in Bezug auf alle $(n \times m)$-Fehlergleichungsmatrizen C wirklich ein Maximum, das heißt sie wird für gewisse leicht zu konstruierende Matrizen angenommen. Allerdings wäre in diesen Fällen durch eine andere Pivotauswahl eine bedeutend günstigere Kondition zu erreichen. Man kann somit sicher nicht behaupten, daß die Auswahl nach dem absolut größten Element bezüglich der Kondition von $B^T B$ immer die beste sei, sondern nur daß sie sicher nicht beliebig schlecht, und im allgemeinen wohl auch vernünftig ist.

Zur Illustration sei auf das in Abschnitt 6.3 näher ausgeführte Beispiel einer Matrix C hingewiesen, welche einen Parameter ε enthält, derart daß die Spaltenvektoren von C für $\varepsilon \to 0$ zusammenfallen. Die schlechte Kondition von C ist also „echt", das heißt sie läßt sich nicht durch Normierung der Spalten beheben. Dennoch erhält man für ein beliebig kleines $\varepsilon > 0$ durch die Jordan-Transformation eine Matrix B mit guter Kondition:

$$\begin{aligned} &(\text{Kondition von } C^T C) && \approx \frac{m}{\varepsilon^2}. \\ &(\text{Kondition von } N = E + B^T B) && \approx m. \end{aligned}$$

4. Schranken für die Quadratsumme der Verbesserungen

Wie in Abschnitt 1 auseinandergesetzt wurde, hat man bei einer iterativen Behandlung der Ausgleichungsaufgabe ein Interesse daran, während der Rechnung beidseitige Schranken für (v, v) zu kennen, da ein nahes Zusammenrücken derselben bedeutet, daß der Prozeß abgebrochen werden soll.

Eine obere Schranke für (v, v) wird trivialerweise von jedem Punkt x' aus F' geliefert, da ja der Lösungspunkt diese Größe zu einem Minimum macht. Das Abstandsquadrat eines Punktes x'' aus F'' von l gibt zwar im allgemeinen keine untere Schranke; eine solche kann jedoch bei gleichzeitiger Kenntnis eines x' *und* eines x'' in sehr einfacher Weise berechnet werden.

Wir wollen für das folgende voraussetzen, daß ein Rechenverfahren benützt werde, welches in jedem Schritt je ein x' und ein x'' liefert. Nach den Ausführungen von Abschnitt 2 ergeben sich diese Punkte mit unseren transformierten Normalgleichungen in zwangloser Weise, und zwar gleichgültig, ob das System für z oder dasjenige für w aufgelöst werde.

Es erübrigt sich somit, von einem Näherungspunkt der einen Ebene den in [4] beschriebenen, ziemlich künstlichen Übergang zu einem Punkt der anderen Ebene zu machen.

Zur Gewinnung einer unteren (v, v)-Schranke fällen wir von x' aus das Lot auf die durch l und x'' gehende Gerade. Mit den Bezeichnungen $v' = x' - l$, $v'' = x'' - l$ erhält man für den Fußpunkt u'' des Lotes:

$$u'' = l + \frac{(v', v'')}{(v'', v'')} v''. \tag{4.1}$$

Nun ist aber wegen

$$(x' - u'', u'' - l) = 0$$

und

$$(s - x', u'' - l) = 0 \quad (s = \text{Lösungspunkt})$$

auch

$$(s - u'', u'' - l) = 0,$$

das heißt u'' ist auch die Projektion von s auf dieselbe Gerade. Somit ist $(u'' - l, u'' - l)$ eine untere Schranke für (v, v), und zwar die beste, welche auf Grund von x'' gewonnen werden kann; ganz unabhängig von x'.

Somit haben wir für (v, v) folgende Ungleichungen:

$$\frac{(v', v'')^2}{(v'', v'')} \leqq (v, v) \leqq (v', v'). \tag{4.2}$$

Bei diesen Abschätzungen handelt es sich um einen Spezialfall von solchen, welche die Hyperkreismethode liefert, wenn in jeder der Ebenen F' und F'' je ein System von Vektoren gegeben ist. Dieser Fragenkomplex ist in [8] mit vielen Beispielen beschrieben; für die Anwendung auf die Ausgleichsrechnung s. [4].

5. Iterative Behandlung der transformierten Gleichungen

Wir beschränken uns im Rahmen dieser Arbeit auf die Diskussion eines speziellen Relaxationsverfahrens, welches sich in einigen Versuchen gut bewährt hat, nämlich der *Methode der konjugierten Gradienten* (cg-Verfahren). Es werden die expliziten Rechenvorschriften für unseren Fall der transformierten Normalgleichungen angegeben werden.

Der Algorithmus wurde in der ursprünglichen Form von [3] verwendet; in jener Arbeit findet man auch eine ausführliche Theorie. Für Berichte über numerische Experimente, auch in Kombination mit anderen Methoden, s. [1]; für die Anwendung auf die Ausgleichsrechnung s. [4].

Einige Versuche wurden auch mit dem Verfahren von FLANDERS und SHORTLEY [2] durchgeführt, einem Gradientenverfahren, welches als Residuenpolynome die Tschebyscheffschen Polynome verwendet (s. [1]). Es lag deshalb nahe, diese Methode anzuwenden, weil sie Schranken für die Eigenwerte der Koeffizientenmatrix benötigt, und diese sind im vorliegenden Falle leicht anzugeben. Das cg-Verfahren hat sich jedoch als überlegen erwiesen (s. erste Fußnote).

Es sei allgemein der unbekannte Vektor z aus dem Gleichungssystem $Az = k$ zu bestimmen, wo A für eine symmetrische positiv definite Matrix steht. Wenn z_j (durch den Index werden in diesem Zusammenhang nicht Komponenten,

sondern verschiedene Vektoren unterschieden) die Näherung des j-ten Iterationsschrittes ist, dann nennen wir

$$r_j = k - A z_j \tag{5.1}$$

den zugehörigen Residuenvektor.

Das cg-Verfahren gehört zu einer bestimmten Klasse von Gradientenverfahren, bei denen die Korrektur Δz_j des Unbekanntenvektors einer dreigliedrigen Rekursion

$$\Delta z_j = \frac{1}{\varkappa_j}(r_j + \pi_{j-1} \cdot \Delta z_{j-1}) \tag{5.2}$$

gehorcht. Durch die Wahl der Skalare $\varkappa_j$, π_j wird das spezielle Verfahren festgelegt (s. [1]).

Der Rechenprozeß soll nun auf unser spezielles System mit $A = N = E + B^T B$, $k = -d$ angewendet und so umgeschrieben werden, daß A nicht explizit gebildet wird. Dafür kommen in jedem Schritt zwei Multiplikationen der Rechtecksmatrix B, bzw. von B^T mit einem Vektor vor. Diese beiden Multiplikationen würden auch in einem primitiveren Iterationsverfahren stehen bleiben, so daß der Mehraufwand des cg-Verfahrens bescheiden ist.

Die Formeln für den allgemeinen Fall werden hier nicht reproduziert; wir legen für das folgende die Bezeichnungen von [4] zugrunde.

Wenn man den Vektor $q = B p$ einführt, wird

$$(p, A p) = (p, (E + B^T B) p) = (p, p) + (q, q).$$

Um in jedem Schritt die für die (v, v)-Schranken notwendigen Punkte x' und x'' zu erhalten, welche in den Ebenen F' und F'' liegen, berechnet man, ausgehend von der Näherung z des j-ten Schrittes (der Index j ist im folgenden weggelassen):

$$w = B z; \qquad \hat{z} = -B^T w - d.$$

Allerdings kann w rekursiv berechnet werden, wegen $\Delta w = B \Delta z = \lambda \cdot q$.

Dann besteht x' aus den beiden Teilvektoren z und w, x'' aus $\hat{z}$ und w, und der Residuenvektor ergibt sich in einfacher Weise aus

$$r = k - A z = -d - z - B^T B z = \hat{z} - z.$$

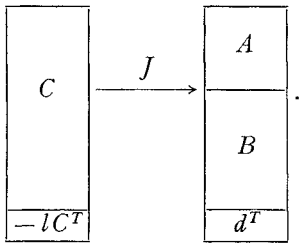
Zweckmäßigerweise zerlegt man auch $v = x' - l$ und l entsprechend x je in die beiden Teilvektoren v_z und v_w bzw. l_z und l_w.

Für den Start eines Relaxationsverfahrens wird man im allgemeinen $z_0 = 0$ setzen müssen, wenn über die Lösung weiter nichts bekannt ist. Im vorliegenden Falle kann man sich jedoch einen Teil der in die Jordan-Transformation gesteckten Arbeit zunutze machen, da die Unbekannten (Komponenten von z) nach der Transformation gerade m von den Meßvariablen sind, und deshalb die entsprechenden Komponenten des Meßvektors l als Startwerte benützen. Das heißt geometrisch, daß man, wenn die Fehlergleichungen als Gleichungen von Hyperebenen im R^m aufgefaßt werden, durch geeignete Pivotauswahl einen möglichst gut definierten Schnittpunkt von m solchen Hyperebenen bestimmt und diesen, statt des Nullpunktes, als Ausgangspunkt für die Iteration nimmt.

Damit kann nun die Rechenvorschrift für das cg-Verfahren formuliert werden:

1. Bilde $-C^T l$ und füge diesen Vektor als zusätzliche Zeile zur Matrix C der Fehlergleichungen.

2. Jordan-Transformation (s. Abschnitt 2)



Die auf Grund der Pivotauswahl durchgeführten Zeilen- und Spaltenvertauschungen hat man sich zu merken. Die letzte Zeile darf nicht Pivotzeile werden. Die Zeilenvertauschungen sind auch auf die entsprechenden Komponenten von l auszuüben (wegen der Berechnung von $v = x - l$).

3. Relaxationsprozeß. (Wenn die Formeln in der richtigen Reihenfolge geschrieben werden, erübrigt sich deren Belastung mit dem Index j des Zyklus. Bei der Durchführung der Rechnung auf einem Automaten werden ja auch tatsächlich die entsprechenden Größen immer wieder am selben Ort gespeichert. Das Zeichen „:=" soll andeuten, daß es sich nicht um Gleichungen im üblichen Sinne, sondern um Rechenvorgänge handelt. Genauer: der links vom Zeichen stehenden Variablen ist der momentane Wert des Ausdrucks der rechten Seite zuzuweisen. Die Zuweisung des Wertes „∞" ist im Sinne der Gleitkommamaschinen aufzufassen).

Vorbereitung:

$$p := l_z; \qquad z := 0; \qquad w := 0;$$
$$\varrho := \infty; \qquad \varrho' := \infty; \qquad \tau := \infty.$$

Beginn an der Stelle * des *allgemeinen Rechenzyklus:*

$r := \hat{z} - z; \qquad \varrho := (r, r);$
$\varepsilon := \varrho/\varrho';$
$\varrho' := \varrho;$
$p := r + \varepsilon \cdot p; \qquad \tau := (p, p);$
*
$q := Bp; \qquad \sigma := (q, q);$
$\lambda := \varrho/(\sigma + \tau);$
$z := z + \lambda \cdot p;$
$w := w + \lambda \cdot q;$
$\hat{z} := -B^T w - d;$
$v_z := z - l_z; \qquad \omega_1 := (v_z, v_z);$
$v_w := w - l_w; \qquad \omega_2 := (v_w, v_w);$
$\hat{v}_z := \hat{z} - l_z; \qquad \omega_3 := (\hat{v}_z, \hat{v}_z);$
$\omega_4 := (v_z, \hat{v}_z);$
$\eta_1 := (\omega_2 + \omega_4)^2/(\omega_2 + \omega_3);$
$\eta_2 := \omega_1 + \omega_2$

(v, v)-Schranken nach (4.2): $\eta_1 \leqq (v, v) \leqq \eta_2$. Dieser Teil der Rechnung wird für den Relaxationsprozeß nicht benützt und braucht daher nicht unbedingt in jedem Schritt durchgeführt zu werden.

4. Um die Komponenten des Lösungsvektors x in der richtigen Reihenfolge zu bekommen, sind die bei der Jordan-Transformation ausgeführten *Zeilen*vertauschungen an den Komponenten des aus z und w zusammengesetzten Vektors wieder rückgängig zu machen.

5. Falls schließlich die eigentlichen Unbekannten der vermittelnden Ausgleichung verlangt werden, berechne man $y := A z$, worauf die Komponenten von y gemäß den *Spalten*vertauschungen der Jordan-Transformation wieder in die ursprüngliche Ordnung gebracht werden müssen.

Bemerkung zu den (v, v)-Schranken: Die oberen Schranken $\eta_2 = (v', v')$ nehmen beim cg-Verfahren monoton ab (s. [4]), wogegen die unteren Schranken η_1 nicht notwendigerweise monoton wachsen. Würde man hingegen die Gleichung (2.14) für w lösen, das heißt bedingte Ausgleichung treiben, dann hätte man nach [4] Monotonie bei η_1 und nicht bei η_2, und die unteren Schranken η_1 würden schon durch die cg-Näherungspunkte x'' selbst geliefert, so daß der, allerdings bescheidene Rechenaufwand für die Auswertung der linken Seite von (4.2) entfiele.

Der Vollständigkeit halber seien noch folgende Identitäten angeführt, welche beim cg-Verfahren gelten:

Die Residuenvektoren von verschiedenen Schritten sind orthogonal:

$$(r_j, r_k) = 0 \quad \text{für} \quad j \neq k. \tag{5.3}$$

Die Gewichtsvektoren von verschiedenen Schritten sind konjugiert bezüglich der Matrix N des Gleichungssystems:

$$(p_j, N p_k) = (p_j, p_k) + (q_j, q_k) = 0 \quad \text{für} \quad j \neq k. \tag{5.4}$$

Ferner gelten auf Grund der in [4] angegebene Beziehungen für jeden Rechenschritt:

$$(v_z, z) + (v_w, w) = 0 \tag{5.5}$$

und

$$(l_z, p) + (l_w, q) = \varrho. \tag{5.6}$$

Der Wert solcher Identitäten als Rechenproben ist immerhin fragwürdig, da nicht leicht entschieden werden kann, wie weit Unstimmigkeiten von Rundungsfehlern herrühren.

Um schließlich den Rechenaufwand bei den verschiedenen Methoden grob abschätzen zu können, wird noch angegeben, wie sich die Anzahl Multiplikationen asymptotisch verhält:

a) Bildung der Normalgleichungen: $n\, m^2/2$; Auflösung der Normalgleichungen durch Elimination: $m^3/6$. (Bei beiden Ausnützung der Symmetrie.)

b) Jordan-Transformation: $n\, m^2$; Ein Relaxationsschritt: $2n\, m$.

6. Beispiele

6.1. Allgemeine Bemerkungen zur Polynom-Approximation

Die Aufgabe, bei gegebenen n Stützwerten $(n > m)$ das Polynom $(m-1)$-ten Grades bester Approximation im Sinne der kleinsten Quadrate zu bestimmen, kann als Problem der vermittelnden Ausgleichung formuliert werden. Die Matrix C der Fehlergleichungen enthält in ihrer k-ten Spalte die $(k-1)$-ten Potenzen der Stützabszissen und führt bekanntlich auf sehr schlecht konditionierte Normal-

gleichungen. Man hat deshalb andere Methoden entwickelt, welche sich besser für die numerische Rechnung eignen (z. B. Bildung der Orthogonalpolynome unter Benützung der Rekursionsformel).

Mit dem nachstehend beschriebenen Versuch sollte immerhin festgestellt werden, ob nicht auch in Fällen, wo die Auflösung der gewöhnlichen Normalgleichungen versagt, eventuell mit Hilfe der vorgängigen Jordan-Transformation wenigstens die Berechnung der ausgeglichenen Polynomwerte zu retten sei. Das Resultat war durchaus positiv. Hingegen wird in vielen Fällen die Bestimmung der Polynomkoeffizienten auch mit unserer Methode hoffnungslos sein.

Man kann übrigens im Falle der Polynomausgleichung der Methode noch eine sehr einfache Deutung geben: Durch die Pivotauswahl werden m Stützstellen von allen n ausgezeichnet. Bei der Bestimmung eines Punktes x' aus F' gibt man an den m ausgezeichneten Stellen die Polynomwerte, nämlich die Komponenten von z vor, und erhält mit $w = Bz$ die Werte an den übrigen Stellen. Die Spalten der Matrix C_1 (s. Abschnitt 2), welche ja Linearkombinationen der Spalten von C sind, enthalten also gerade die Werte der Grundpolynome für Lagrangesche Interpolation bezüglich der ausgezeichneten Stützstellen. Und wenn wir die Iteration dann mit $z_0 = l_z$ beginnen, so heißt dies, daß als Ausgangsnäherung nicht etwa das identisch verschwindende Polynom, sondern das Interpolationspolynom genommen wird, welches an den ausgezeichneten Stützstellen die zu approximierenden Werte exakt annimt. Nur führen wir eben nicht die Interpolation aus, sondern üben auf die Fehlergleichungsmatrix die Jordan-Transformation aus.

Wenn man nun aus einer großen Anzahl n von Stützstellen deren m für diese Interpolation auswählen müßte, würde man sie wohl aus Gründen der Stabilität möglichst wie die Extremalstellen des Tschebyscheff-Polynoms $(m-1)$-ten Grades anordnen. Die angestellten Überlegungen erfahren somit nochmals eine gewisse Rechtfertigung dadurch, daß bei dem durchgerechneten Fall die Pivotauswahl nach dem absolut größten Element von selbst auf eine derartige Stützstellenverteilung führt.

Es sei in diesem Zusammenhang auch auf [6] verwiesen, wo vor allem im 4. Abschnitt (Theorie der S-Funktionen) das Problem der optimalen Abszissen von einer anderen Stelle her beleuchtet wird.

6.2. Polynom-Beispiel

Das folgende Beispiel wurde auf einer 11stelligen dezimalen Gleitkommamaschine (ERMETH) durchgerechnet:

An den Stellen $\xi = 0(0{,}05)1$ gegebene Stützwerte (Komponenten von l) sind durch ein Polynom 9. Grades zu approximieren (Komponenten von x), also $n = 21$, $m = 10$.

Zuerst wurden die gewöhnlichen Normalgleichungen gebildet und diese durch Elimination aufgelöst. Durch Einsetzen der Lösungen y_i (nun wieder Komponenten) in die Fehlergleichungen ergaben sich für die ausgeglichenen x_i total unbrauchbare Werte, bei welchen zum Teil sogar das Vorzeichen falsch war. Der damit errechnete Wert von (v, v) war 10mal zu groß.

Andererseits lieferte die Jordan-Transformation mit anschließendem cg-Verfahren ganz befriedigende Resultate.

In den ersten cg-Schritten ergaben sich folgende (v, v)-Schranken:

Schritt	η_2	η_1
Start	0,30230	0,07420
1	0,13051	0,09288
2	0,10758	0,10735
3	0,10749	0,10748
4	0,10748	0,10748

Die 6. Stelle von η_1 und η_2 war nicht mehr korrekt. Auf Grund der obigen Tabelle würde man in einem praktischen Fall sicher nach dem 2. oder 3. Schritt aufhören.

Die x_i wiesen im 8. Schritt noch absolute Fehler von $5 \cdot 10^{-6}$ auf. Im selben Schritt wurden ferner nach $y = A z$ die Polynomkoeffizienten y_i berechnet, und daraus wiederum $x = C y$. Diese x-Werte hatten dann Fehler von $5 \cdot 10^{-5}$.

6.3. Theoretisches Beispiel mit beliebig schlechter Kondition

Es soll hier kurz auf das Beispiel eingegangen werden , welches am Schluß von Abschnitt 3 angedeutet wurde.

Um nochmals an die geometrische Interpretation der Ausgleichsaufgabe anzuknüpfen: Der Vektor l soll so gut wie möglich durch eine Linearkombination der Spaltenvektoren von C, d.h. also durch einen Vektor $C y$ aus dem Teilraum F' angenähert werden. Die schlimmste Situation für die numerische Lösung liegt nun wohl dann vor, wenn diese Spaltenvektoren beinahe zusammenfallen, und die Projektion von l auf F' ungefähr senkrecht auf ihnen steht.

Ein solcher Fall ergibt sich (z.B. bei $m = 5$, $n = 6$) mit den folgenden Daten:

$$C = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ \varepsilon & & & & \\ & \varepsilon & & 0 & \\ & & \varepsilon & & \\ & 0 & & \varepsilon & \\ & & & & \varepsilon \end{bmatrix}; \quad l = \begin{bmatrix} \varepsilon \\ 0 \\ -5 \\ 5 \\ -5 \\ 0 \end{bmatrix}.$$

Die Koeffizienten der Normalgleichungen wären:

$$C^T C = \begin{bmatrix} 1+\varepsilon^2 & 1 & & 1 & 1 \\ 1 & 1+\varepsilon^2 & & & \vdots \\ & & \ddots & & 1 \\ 1 & \cdots & & 1 & 1+\varepsilon^2 \end{bmatrix}; \quad C^T l = \varepsilon \cdot \begin{bmatrix} 1 \\ -4 \\ 6 \\ -4 \\ 1 \end{bmatrix}.$$

Die Eigenwerte von $C^T C$ sind $\lambda_1 = 5 + \varepsilon^2$, $\lambda_2 = \cdots = \lambda_5 = \varepsilon^2$; die Kondition somit $\approx \frac{5}{\varepsilon^2}$. Falls man ε genügend klein wählt, (z.B. $\varepsilon = 10^{-6}$, bei 11stelligen Mantissen), wird $C^T C$ sogar numerisch exakt singulär.

Hingegen führt die Jordan-Transformation auf die (einzeilige) Matrix

$$B = [\varepsilon \; -1 \; -1 \; -1 \; -1].$$

Die Eigenwerte von $N = E + B^T B$ sind $\lambda_1 = 5 + \varepsilon^2$, $\lambda_2 = \cdots = \lambda_5 = 1$; die Kondition somit ≈ 5.

Ein cg-Schritt, oder natürlich auch direkte Auflösung ergeben die x-Lösung

$$s = \begin{bmatrix} 0 \\ 1 \\ -4 \\ 6 \\ -4 \\ 1 \end{bmatrix} \text{ und die eigentlichen Unbekannten } y = A\,x = \frac{1}{\varepsilon} \cdot \begin{bmatrix} 1 \\ -4 \\ 6 \\ -4 \\ 1 \end{bmatrix}.$$

Bei der Jordan-Transformation passiert in der Rechnung für B keine Stellenauslöschung. Das gilt jedoch nicht für A und d. Die Tatsache, daß hier mit runden Zahlen gerechnet wird und so alle Resultate doch exakt herauskommen, ist deshalb nicht sehr interessant.

Um eine realistischere Lage zu schaffen, wurde das theoretisch äquivalente Problem mit (im x-Raum) gedrehten Vektoren, welche jetzt allgemein im Koordinatensystem lagen, durchgerechnet. C und l wurden zuvor noch mit Nullen derart ergänzt, daß die Dimensionen $n = 10$, $m = 5$ entstanden. Erwartungsgemäß wurde die Situation infolge der nun auftretenden Rundungen gegenüber dem vorherigen Spezialfall etwas verwischt.

Das Resultat war dennoch eindeutig: Der normale Weg über $C^T C$ lieferte völlig unbrauchbare Werte, wogegen mit Jordan-Transformation und 5 cg-Schritten Resultate mit 4 bis 5 korrekten Stellen herauskamen, und zwar für *x und y*.

Literatur

[1] ENGELI, M., TH. GINSBURG, H. RUTISHAUSER and E. STIEFEL: Refined Iterative Methods for Computation of the Solution and the Eigenvalues of Self-Adjoint Boundary Value Problems. Mitt. Nr. 8, Inst. f. angew. Math., ETH (Basel und Stuttgart: Birkhäuser 1959).

[2] FLANDERS, D. A., and G. SHORTLEY: Numerical Determination of Fundamental Modes. J. appl. Phys. **21**, 1326–1332 (1950).

[3] HESTENES, M. R., and E. STIEFEL: Methods of Conjugate Gradients for Solving Linear Systems. J. Res. NBS **49**, No. 6, 409 (1952).

[4] LÄUCHLI, P.: Iterative Lösung und Fehlerabschätzung in der Ausgleichsrechnung. ZAMP **10**, 245–280 (1959).

[5] RUTISHAUSER, H.: Zur Matrizeninversion nach GAUSS-JORDAN. ZAMP **10**, 281–291 (1959).

[6] STIEFEL, E.: Über diskrete und lineare Tschebyscheff-Approximationen. Numerische Mathematik **1**, 1–28 (1959).

[7] — Note on Jordan elimination, linear programming and Tchebycheff approximation. Numerische Mathematik **2**, 1–17 (1960).

[8] SYNGE, J. L.: The Hypercircle in Mathematical Physics. Cambridge: Cambridge University Press 1957.

Institut für angewandte Mathematik
der Eidgenössischen Technischen Hochschule
Zürich 6 (Schweiz)

(Eingegangen am 27. April 1961)