ZAMM 61, T 313 – T 315 (1981)

S. M. Rump

Kleine, exakte Fehlerschranken für die Lösung linearer Gleichungssysteme

1. Einleitung

Bekanntlich können auf Rechenanlagen durch Rundungsfehler große Fehler entstehen. Dies ist um so mehr der Fall, wenn die Arithmetik nicht sauber implementiert ist. So sind etwa die Operanden der Differenz

$$134217728.0 - 134217727.0$$

exakt (d. h. ohne Konvertierungs- oder Rundungsfehler) auf einer weit verbreiteten Großrechenanlage darstellbar, ebenso wie das Ergebnis 1.0; der Computer errechnet jedoch als Ergebnis 2.0. Die relative Rundungsfehlereinheit müßte nach der Akkumulatorlänge auf diesem Rechner 2^{-27} sein, in diesem Beispiel ist sie jedoch 1. Für viele bekannte Fehlerabschätzungen sind damit die Voraussetzungen für deren Gültigkeit nicht erfüllt.

In Gleitkommaalgorithmen müssen zur Vermeidung von schwerwiegenden Fehlern Kontrollen im Algorithmus und am Ergebnis angebracht werden. Gleichwohl wird kein Beweis für die maximale Ungenauigkeit des Ergebnisses gegeben und die laienhafte Anwendung kann gefährlich werden. Es werden Algorithmen entwickelt, die bewiesene Fehlerschranken berechnen, und zwar zunächst für die Lösung linearer Gleichungssysteme. Der Zeitaufwand liegt in der Größenordnung des Gleitkomma-Gauss-Algorithmus, es entfällt jedoch jeglicher Aufwand des Benutzers für die Kontrolle, da die Ergebnisse als richtig bewiesen sind.

2. Theoretischer Hintergrund

Gegeben sei das lineare Gleichungssystem

$$Ax = b \quad \text{mit} \quad A \in M_n\mathbb{R} \quad \text{und} \quad b \in V_n\mathbb{R}\,. \tag{1}$$

Ein Fixpunkt der Funktion

$$f(x) = x + R(b - Ax)\,, \qquad R \in M_n\mathbb{R}\,, \tag{2}$$

ist für nicht-singuläres R offenbar Lösung des Gleichungssystems (1). Bereits in [1] wurde der folgende Operator

$$g(x) = f(\tilde{x}) + (E - RA)(x - \tilde{x})\,, \qquad \tilde{x} \in V_n\mathbb{R}\,, \tag{3}$$

betrachtet. Für beliebiges $\tilde{x}$ ist offenbar

$$f(x) = g(x) \quad \text{für alle} \quad x \in V_n\mathbb{R}\,. \tag{4}$$

Gilt für einen Intervallvektor $X \in \mathbb{I}V_n\mathbb{R}$

$$f(X) \subseteq X\,, \tag{5}$$

so besitzt f nach dem Fixpunktsatz von Brouwer mindestens einen Fixpunkt. Das Erfülltsein der Bedingung (5) kann wegen (4) durch einfaches Ersetzen der Gleitkommaoperationen durch die entsprechenden Intervalloperationen in (3) nachgeprüft werden. Ein Problem ist nachzuprüfen, ob die Matrix R nicht-singulär ist und, damit die Lösung eindeutig ist, ob auch A nicht-singulär ist. Der folgende Satz zeigt, daß bereits eine schwache Verschärfung von (5) ausreicht um beide Fragen positiv zu entscheiden. Zuvor eine Definition.

Definition 1: Seien $I, J \in \mathbb{I}\mathbb{R}$ reelle Intervalle. I heißt *echt enthalten in* J, in Zeichen $I \subsetneqq J$ wenn gilt:

$$I \subseteq J \quad \text{und} \quad I \cap \partial J = \emptyset\,.$$

D. h. I besteht nur aus inneren Punkten von J. Die entsprechende Definition für Intervallvektoren gilt komponentenweise.

Satz 2: *Gegeben sei das lineare Gleichungssystem* (1) *und f wie in* (2) *für beliebiges $R \in M_n\mathbb{R}$. Aus*

$$f(X) \subsetneqq X \quad \text{für} \quad X \in \mathbb{I}V_n\mathbb{R} \tag{6}$$

folgt dann, daß

die Matrizen R und A nicht-singulär sind, (7)

das Gleichungssystem $Ax = b$ eindeutig lösbar ist und (8)

für $A\hat{x} = b$ mit $\hat{x} \in V_n\mathbb{R}$ ist $\hat{x} \in f^k(X)$ für $0 \leqq k \in \mathbb{N}$. (9)

Beweis: Nach dem Fixpunktsatz von Brouwer besitzt f einen Fixpunkt $\hat{x} \in V_n\mathbb{R}$, für den nach (2) gilt

$$b - A\hat{x} \in \mathrm{Ker}\,(R)\,.$$

Für $y \in \mathrm{Ker}\,(A)$ und $\lambda \in \mathbb{R}$ ist

$$f(\hat{x} + \lambda y) = \hat{x} + \lambda y + R(b - A\hat{x} - Ay) = \hat{x} + \lambda y + R(b - A\hat{x}) = \hat{x} + \lambda y\,.$$

Da $\hat{x} \in X$ und $\hat{x} + \lambda y$ für jedes $\lambda \in \mathbb{R}$ ebenfalls Fixpunkt von f ist, gäbe es für $y \neq 0$ einen Fixpunkt $\hat{x} + \mu y$ auf dem Rand von X im Widerspruch zu (6). Die Matrix A ist also nicht-singulär. Angenommen $0 \neq y \in \mathrm{Ker}\,(R)$. Wegen der Nicht-Singularität von A folgt $A^{-1}y \neq 0$ und wie oben die Existenz eines $\mu \in \mathbb{R}$ mit $\hat{x} + \mu(A^{-1}y) \in \partial X$. Andererseits ist aber

$$f\big(\hat{x} + \mu(A^{-1}y)\big) = \hat{x} + \mu(A^{-1}y) + R(b - A\hat{x} - \mu y) = \hat{x} + \mu(A^{-1}y)$$

wieder im Widerspruch zu (6). Damit ist (7) gezeigt und (8) folgt sofort. Aus $\hat{x} \in X = f^0(X)$ und

$$\hat{x} \in f^k(X) \Rightarrow \hat{x} = f(\hat{x}) \in f\big(f^k(X)\big) = f^{k+1}(X)$$

folgt mittels vollständiger Induktion (9) und der Satz ist vollständig bewiesen.

3. Algorithmus

Aus diesem Satz läßt sich sofort ein Algorithmus zur Einschließung der Lösung von (1) ableiten: ($\circledast$ bedeutet Intervalloperation für $* \in \{+, -, \cdot, /\}$)

$$\begin{aligned} &X := \tilde{x} \subset R \odot (b \ominus A \odot \tilde{x}):\\ &\text{repeat} \quad Y := X; \qquad X := f(X)\\ &\text{until} \quad X \subseteq Y\,. \end{aligned} \tag{10}$$

Hierbei sind $\tilde{x}$ und R eine Näherung für $\hat{x}$ und die Inverse von A, an deren Güte nach Satz 2 keinerlei Voraussetzungen geknüpft sind. Mit dem Eintreten von (6) ist die Voraussetzung von Satz 2 erfüllt und es gelten seine Folgerungen. Der Algorithmus kann in vielerlei Hinsicht wesentlich verbessert werden. Insbesondere folgt mit

$$h(x) := R(b - A\tilde{x}) + (E - RA)\,x$$

aus

$$h(X) \subsetneqq X$$

sofort (7), (8) und

$$\hat{x} \in \tilde{x} \oplus X \quad \text{mit} \quad A\hat{x} = b .$$

Das so gewonnene Ergebnisintervall ist bereits wesentlich schärfer. So ist es möglich bei einfacher Grundgenauigkeit, wobei nur der Defekt $b - Ax$ doppelt-genau berechnet wird, die Lösung auf weit mehr als einfache Genauigkeit einzuschließen. Bei Verwendung höherer Defekte und eines langen Akkumulators ist es so möglich, bei im wesentlichen einfacher Rechengenauigkeit die Lösung beliebig genau einzuschließen. Für eine ausführliche Beschreibung dieser und weiterer Verbesserungen siehe [2].

4. Zusammenfassung

Der wesentliche Fortschritt der beschriebenen Methode ist, daß ausgehend von einer Gleitkommanäherung der Lösung ohne Voraussetzungen an deren Güte eine Einschließung der Lösung von (1) numerisch konstruiert und gleichzeitig deren Richtigkeit bewiesen wird. Das bisher notwendige Beschaffen einer Ersteinschließung entfällt ebenso wie der bisher notwendige große Mehraufwand des Anwenders für diverse Kontrollen im Algorithmus und am Ergebnis. Die Rechenzeit des Algorithmus ist die 6-fache gegenüber dem Gleitkomma-GAUSS-Algorithmus unabhängig von der Größe des Systems. Kommt die Schleife (10) nicht zum Stillstand, ist die Rechengenauigkeit gemessen an der Kondition des Problems unzureichend und das Problem ist mit höherer Genauigkeit zu rechnen (oder die Problemstellung ist zu überprüfen). In fast allen gerechneten Beispielen trat die Einschließung nach einem Schritt ein. Der Algorithmus wurde bis $n = 200$ gerechnet, wobei die Zahl 200 durch die begrenzte Speicherkapazität bedingt ist und keine Grenze für den Algorithmus darstellt.

Die Algorithmen sind in FORTRAN implementiert und verfügbar und sind leicht erweiterbar auf komplexe Gleichungssysteme. An der Verallgemeinerung auf schwach besetzte Matrizen wird gearbeitet.

Literatur

1 KRAWCZYK, R., NEWTON-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken, Computing **4** (1969), 187–201.
2 RUMP, S. M., Kleine Fehlerschranken bei Matrixproblemen, Dr.-Dissertation, Universität Karlsruhe, Februar 1980.
3 RUMP, S. M.; KAUCHER, E., Small Bounds for the Solution of Systems of Linear Equations, Supplementum **2** Computing.

Anschrift: Dr. SIEGFRIED M. RUMP, Universität Karlsruhe, Institut für Angewandte Mathematik, Kaiserstraße 12, D-7500 Karlsruhe, BRD

ZAMM 61, T 315 – T 317 (1981)

R. SCHERER / K. ZELLER

Rundungsfehler bei linearen Gleichungen

1. Einleitung

Zahlreiche Veröffentlichungen über Rundungsfehler in neuerer Zeit verdeutlichen das Interesse an diesem Gebiet. Wir knüpfen an einige dieser Untersuchungen an und behandeln eine Kurzschrift für Rundungsfehler, die vergleichbar ist mit früheren Darstellungsweisen (z. B. WILKINSON [18], VAN DER SLUIS [16], OLVER [8]), aber in ihrer ausgeprägten Form manche Vorteile bietet. Sie liefert zentrale Abschätzungen in ziemlich einfacher und übersichtlicher Weise. Wir behandeln zunächst Dreieckszerlegungen (Lemma 1, vgl. SAUTTER [11] [12], VELDHUIZEN [17]). Dann betrachten wir einige wichtige Ergänzungen (DE BOOR, PINKUS [1], REID [10], Lemma 2). Weiter gehen wir auf die OETTLI-PRAGER-Schranke ein (Lemma 3, vgl. SCHABACK [13], SKEEL [15]). Zum Schluß weisen wir auf Ausbaumöglichkeiten und auf weitere Literatur hin.

Wir nehmen an, daß die Gleitpunkt-Arithmetik der Regel

$$fl(a \circ b) = (a \circ b)\varrho \quad \text{mit} \quad |\log \varrho| \leqq \log \bar{\varrho} =: r^*$$

genügt. Natürlich hängt der Faktor ϱ von den Parametern ab, wir verzichten aber auf eine Kennzeichnung dieser Abhängigkeit. Bei mehrmaligem Auftreten wird also ϱ im allgemeinen verschiedene Zahlen bedeuten. Konsequent schreiben wir ϱ^2, ϱ^3, ... für ein Produkt von zwei oder mehr Zahlen dieses Typs.

2. Dreieckszerlegung

Ein lineares Gleichungssystem behandeln wir numerisch mittels Dreieckszerlegung und zweistufiger Auflösung. Die Ungenauigkeiten beschreiben wir in der Form $A = LR - G$, $(LR + H)x = b$, also $(A + F)x = b$ mit $F = G + H$ (siehe Lemma 1). Wir nehmen an, daß die numerischen Operationen (Anordnung wie bei GAUSS-Elimination; $l_{ii} \neq 0$ gewählt) durchführbar sind.

Unsere Grundformel lautet

$$\left((\ldots (a - l_1 r_1 \varrho)\varrho - \ldots - l_{m-1} r_{m-1}\varrho)\varrho / l_m\right)\varrho =: r_m .$$