Algorithmen mit Ergebnisverifikation – einige Bemerkungen zu neueren Entwicklungen

Christian Jansson und Siegfried M. Rump

Die Numerische Mathematik beschäftigt sich vorwiegend mit Methoden und Algorithmen zur Lösung mathematischer Grundaufgaben aus Anwendungsgebieten der Mathematik und aus der Praxis sowie deren mathematischer Analyse. Numerische Algorithmen, ausgeführt auf digitalen Rechenanlagen (mit endlicher Mantissenlänge), geben allerdings in den meisten Fällen keine verifizierte Fehlerabschätzung für den Fehler zwischen der exakten Lösung und der berechneten Näherungslösung aus. Algorithmen mit Ergebnisverifikation und Intervallmethoden sind ein Teilgebiet der Numerischen Mathematik mit dem Hauptziel, solche verifizierten Fehlerabschätzungen mitzuberechnen. Verifiziert heißt, daß es sich um bewiesenermaßen richtige Schranken handelt unter Einbeziehung aller aufgetretenen Rundungs- und Verfahrensfehler. In diesem Zusammenhang kann auch von computerunterstützten Beweisen gesprochen werden. Dies gilt sowohl für Punktprobleme als auch für Probleme mit toleranzbehafteten Eingabedaten.

Fügt man die Abschätzungen für den maximalen Fehler jeder einzelnen Gleitpunktoperation (Rundungs-ε) zusammen, so gewinnt man im Prinzip eine Fehlerabschätzung für die durch einen Algorithmus berechnete Näherung; es ist allerdings wohlbekannt, daß solche durch einfache Vorwärtsrechnung gewonnenen Abschätzungen den tatsächlichen Fehler in vielen Fällen beträchtlich überschätzen und für praktische Belange in dieser Form dann wenig brauchbar sind. Im folgenden soll exemplarisch gezeigt werden, wie erstens Überschätzungen in angesprochenem Sinne durch geeignete Verfahren vermieden werden können, und wie zweitens auch aus deutlich überschätzten Fehlern doch noch brauchbare Informationen gewonnen werden können.

Ziel dieser Arbeit ist es, einige neuere Entwicklungen dieses jüngeren Forschungsgebietes vorzustellen. Dabei zeigt es sich, daß insbesondere das Zusammenwirken klassischer numerischer Methoden mit Intervallmethoden sowie mit Fehlerabschätzungen der Vorwärts- und Rückwärtsanalyse zu guten Resultaten führen kann. Es geht also nicht darum, numerische Verfahren durch Intervallalgorithmen zu ersetzen, vielmehr soll gezeigt werden, daß in einigen Anwendungsgebieten Verfahren mit Ergebnisverifikation erfolgreich eingesetzt werden können.

Die vorliegende Arbeit wendet sich nicht an Spezialisten auf diesem Gebiet; vielmehr soll eine Einführung in die Problemstellung und einige weiterführende Hinweise gegeben werden, wobei im gegebenen Rahmen auf die Darstellung vieler interessanter Aspekte und Details bewußt verzichtet wird. Der geneigte Leser sei hier auf die Literatur verwiesen.

Verfahren mit Ergebnisverifikation sowie Möglichkeiten zur Vermeidung der Aufblähung von Fehlerschranken werden anhand linearer Gleichungssysteme in zwei Anwendungen erläutert; zum einen vollbesetzte Gleichungssysteme kleinerer und mittlerer Dimension (z.B. 1000 Unbekannte) und zum anderen Gleichungssysteme mit bandartiger oder spärlich besetzter Systemmatrix deutlich größerer Dimension. Im Falle spärlicher Strukturen wurden in jüngster Zeit Systeme bis zu 1 Million Unbekannten verifiziert gerechnet. Es wird ebenfalls der Fall toleranzbehafteter Eingabedaten betrachtet. Die zu skizzierenden Verfahren liefern dann komponentenweise recht scharfe Fehlerschranken für die maximale und die minimale Variation jeder einzelnen Lösungskomponente. Anschließend wird angedeutet, wie diese Verfahren auf den nichtlinearen Fall erweitert werden können.

Ein weiteres Fallbeispiel sind Verfahren zur globalen Optimierung. Hier läßt sich der Wertebereich einer nichtlinearen Funktion in n Unbekannten über einer Eingabebox rigoros für viele praktische Zwecke hinreichend genau abschätzen, obwohl in den einzelnen Schritten der wahre Wertebereich auf Teilboxen erheblich überschätzt werden kann.

Es werden Ergebnisse eines neuartigen Branch-and-Bound Verfahrens diskutiert, das für viele bekannte Testprobleme Approximationen der Lösungen liefert. Darüber hinaus werden Abschätzungen für das globale Optimum berechnet, die in jedem Falle rigoros sind.

Für einige, auch höherdimensionale, aus der Literatur der globalen Optimierung bekannte Testbeispiele ist der Einschließungsalgorithmus schneller als bekannte Gleitpunktverfahren.

1 Einleitung

Die oft zitierte, stürmische Entwicklung auf dem Rechnermarkt stellt heute Rechenleistungen zu erschwinglichem Preis bereit, die vor wenigen Jahren noch als eher utopisch galten. Damit können bei konstanter investierter Rechenzeit verfeinerte theoretische Modelle berechnet werden, wobei die Anzahl der Rechenoperationen entsprechend zunimmt. Dies kann auch neue Fragen nach der Zuverlässigkeit der Rechenergebnisse aufwerfen, die im folgenden diskutiert werden sollen.

Historisch betrachtet gibt es zwei wesentliche und verschiedenartige Ansätze der Ergebnisverifikation, die kurz beschrieben werden sollen.

Der naheliegende und chronologisch zuerst betrachtete Zugang stellt folgende Frage: Gegeben ein Problem P, eine exakte, gesuchte Lösung $\widehat{x} = \widehat{x}(P)$ und eine (berechnete) Näherungslösung $\widetilde{x}$, wie groß ist der Fehler von $\widetilde{x}$ in bezug auf $\widehat{x}$? Für die Abschätzung dieses Fehlers gibt es verschiedene Zugänge. Ist ein endlicher Algorithmus zur Lösung des gestellten Problems P gegeben und führt der Algorithmus bei exakter Ausführung (d.h. im Körper der reellen oder komplexen Zahlen) zur exakten Lösung $\widehat{x}$, kann man untersuchen, inwieweit das Ersetzen jeder exakten reellen oder komplexen Operation durch die entsprechende Gleitkomma-Operation das Ergebnis verfälscht. Dazu schätzt man den Fehler jeder einzelnen Gleitkomma-Operation ab. Die Zusammensetzung dieser Abschätzungen (wobei durch geschicktes Zusammenfügen Verbesserungen erzielt werden können) liefert den Gesamtfehler.

Diese Vorgehensweise ist eine Möglichkeit, eine „Vorwärtsanalyse" durchzuführen. In einer grundlegenden Arbeit haben J. v. Neumann und H. H. Goldstine [25] bereits in den 40er Jahren diese Vorwärtsanalyse für Systeme linearer Gleichungen rigoros durchgeführt. Leider sind die erzielten Fehlerabschätzungen derart pessimistisch, daß sie für praktische Zwecke in dieser Form wenig brauchbar sind.

Ein ganz anderer Zugang ist die „Rückwärtsanalyse". Hier wird die Frage gestellt, inwieweit die Parameter des gegebenen Problems geändert werden müssen, damit die *exakte* Lösung des derart perturbierten Problems $\widetilde{P}$ *gleich* der errechneten Näherungslösung $\widetilde{x}$ ist. Zieht man in Betracht, daß die Daten eines Problems häufig ohnehin mit nicht allzu hoher Genauigkeit bekannt sind, ist dies eine wesentliche Fragestellung. Mit der Rückwärtsanalyse ist unverrückbar der Name Wilkinson [37], [38] verknüpft, der diese für sehr viele Standardprobleme der numerischen Analysis ausführte.

Unabhängig davon, daß die Rückwärtsanalyse in vielen Fällen ein geeignetes Mittel zur Fehlerabschätzung ist, wird öfter auch die Kenntnis des tatsächlichen Fehlers der Näherungslösung gegenüber der exakten Lösung, also eine Vorwärtsanalyse, von Interesse sein. Mathematisch gesehen ist diese rigorose Abschätzung des Fehlers jeder einzelnen Rechenoperation nichts anderes als der Einsatz von Intervalloperationen, falls auch Terme höherer Ordnung mitgenommen werden. Im folgenden Abschnitt soll der dabei auftretende Effekt der Fehlerüberschätzung zunächst erläutert werden, um dann im weiteren darzustellen, wie diese Überschätzung entweder vermieden werden oder bzw. und aus überschätzten Wertebereichen von Funktionen sinnvolle Information gewonnen werden kann.

2 Vorwärtsanalyse und Abschätzung des Wertebereichs einer Funktion

Treten während einer Rechnung fehlerbehaftete Terme auf, sagen wir $A := a \pm \Delta a$ (die eine Größe $\widetilde{a}$ darstellen, von der nur $a - \Delta a \le \widetilde{a} \le a + \Delta a$ bekannt ist), können diese miteinander verknüpft werden. Für Addition und Multiplikation ergibt sich zum Beispiel

$$A + B := (a + b) \pm (\Delta a + \Delta b) \quad \text{und}$$
$$A \cdot B := (a \cdot b) \pm (a \cdot \Delta b + b \cdot \Delta a + \Delta a \cdot \Delta b).$$

Fährt man in diesem Sinne für die vier Grundrechnungsarten fort (bei Division dürfen keine Nullintervalle vorkommen; ein Nullintervall ist ein Intervall, das die Null enthält), ist klar, daß das Ergebnis einer Folge von Operationen immer innerhalb der Fehlerschranken liegen muß. Arbeitet man mit den Grenzen $a - \Delta a$ und $a + \Delta a$, bekommt man schärfere Schranken für Multiplikation und Division, da der Mittelpunkt eines Produkts nicht gleich dem Produkt der Mittelpunkte sein muß. Mit $A := [a_1, a_2]$ und $B := [b_1, b_2]$ kann man einfach definieren

$$A \circ B := [\min_{i,j \in [1,2]} a_i \circ b_j, \max_{i,j \in [1,2]} a_i \circ b_j] \quad \text{für } \circ \in \{+, -, \cdot, /\}.$$

Man überlegt sich leicht, daß man, möglicherweise mit ein paar Fallunterscheidungen, immer mit genau zwei Verknüpfungen, eine für die untere und eine für die obere Grenze, auskommt, außer bei der Multiplikation zweier Nullintervalle. So lassen sich Intervalloperationen für die vier Grundrechnungsarten leicht definieren.

Für die Addition und Subtraktion folgt unmittelbar, daß der Durchmesser des Ergebnisses gleich der *Summe* der Durchmesser der Operanden ist, also

$$\operatorname{diam}(A + B) = \operatorname{diam}(A) + \operatorname{diam}(B) \quad \text{und}$$
$$\operatorname{diam}(A - B) = \operatorname{diam}(A) + \operatorname{diam}(B).$$

Die einzige Möglichkeit, den Durchmesser von Intervallen wieder zu verkleinern, ist also die Multiplikation mit einer kleinen Zahl bzw. Division durch eine große.

Darüber hinaus lassen sich auch die Wertebereiche einfacher transzendenter Funktionen rigoros abschätzen. Im Falle monotoner Funktion ist das sofort klar, etwa

$$\exp(A) := [\exp(a_1), \exp(a_2)].$$

Bei periodischen Funktionen kommen noch einige Fallunterscheidungen hinzu um festzustellen, in welchem Bereich der Funktion man sich befindet.

Festzuhalten bleibt, daß für die Grundrechnungsarten, Exponentialfunktion und Logarithmus, trigonometrische und deren inverse, hyperbolische trigonometrische und deren inverse Funktionen, aber auch Gamma- oder Fehlerfunktion rigorose Schranken für den Wertebereich über einem Intervall A berechenbar sind. Das gilt auch für die Ausführung auf dem Rechner, dann sind alle Rundungs- und Restgliedfehler mit abzuschätzen. Für alle diese Funktionen gibt es Rechner-Implementationen, die den gewünschten Wertebereich scharf und rigoros berechnen.

Kernpunkt bleibt in jedem Falle, daß für dyadische Operationen $\circ$ immer gilt

$$\forall a \in A \quad \forall b \in B : a \circ b \in A \circ B \tag{2.1}$$

und für monadische Operationen σ

$$\forall a \in A : \sigma(a) \in \sigma(A). \tag{2.2}$$

Diese grundlegende Eigenschaft aller Intervalloperationen ist die *Isotonie* der Intervalloperationen. Da diese „Einschließungseigenschaft" sich fortsetzt, kann damit für jede Funktion f, die durch endlich viele Hintereinanderausführungen der oben besprochenen Funktionen berechnet werden können, der Wertebereich über einem Intervall abgeschätzt werden. Das geschieht dann *vollständig automatisch* ohne eine Kenntnis der Ableitung oder Lipschitzkonstanten von f. Zum Beispiel ergibt sich für die Funktion

$$f(x) = \log(3 \cdot \sin x + \sqrt{x} + 1) + e^x \quad \text{für } X = [0,1]$$

die Einschließung des Wertebereiches

$$\begin{aligned} \{ f(x) \mid x \in X \} = f(X) \\ &\subseteq \log(3 \cdot \sin X + \sqrt{X} + 1) + e^X \\ &\subseteq \log(3 \cdot [0,\ 0.84148] + [0,1] + 1) + [1,\ 2.7183] \\ &\subseteq \log([1,\ 4.5245]) + [1,\ 2.7183] \\ &\subseteq [0,\ 0.65558] + [1,\ 2.7183] \\ &\subseteq [1,\ 3.3739]. \end{aligned}$$

Hierbei wurde fünfstellige Dezimalarithmetik verwendet. In diesem günstig gewählten Beispiel tritt kaum Überschätzung auf. Nehmen wir statt dessen

$$f(x) := (x-2)^2 - 4 = x^2 - 4x \quad \text{auf } X = [1,4].$$

Offensichtlich nimmt f bei $x = 2$ sein Minimum an und wegen der Symmetrie bezüglich $x = 2$ folgt $f(X) = [f(2), f(4)] = [-4, 0]$. In diesem trivialen Beispiel ist eine Kurvendiskussion ebenso trivial; es soll zunächst nur der Effekt der Überschätzung durch Datenabhängigkeit demonstriert werden. Es ist

$$\begin{aligned} f(X) &\subseteq X^2 - 4 \cdot X \subseteq [1,4]^2 - 2 \cdot [1,4] \\ &\subseteq [1,16] - [2,8] = [-7,14]. \end{aligned}$$

Ein zwar richtiges, aber zunächst nicht sonderlich befriedigendes Ergebnis. Die Überschätzung resultiert einzig aus der Nichtbeachtung der Abhängigkeit der Variablen:

statt $\{ x^2 - 4x \mid x \in X \}$ wird $\{ x_1^2 - 4x_2 \mid x_1, x_2 \in X \}$

abgeschätzt, letzteres scharf. Verwendet man das Horner-Schema ergibt sich

$$f(X) \subseteq X(X-4) \subseteq [1,4] \cdot [-3,0] \subseteq [-12,0].$$

Ein Teil der Abhängigkeiten wurde bereits eliminiert, die rechte Grenze ist bereits scharf. Entwickelt man f um $x = 2$, so ergibt sich schließlich

$$f(X) \subseteq (X-2)^2 - 4 = [-1,2]^2 - 4 = [0,4] - 4 = [-4,0]$$

der exakte Wertebereich.

Aus diesem Beispiel können zwei Schlußfolgerungen gezogen werden:

1) Eine Vorwärtsfehlerrechnung, in der einzelne Größen mehrfach vorkommen, führt in der Regel zu Überschätzungen.

2) Durch geeignete algebraische Umformung lassen sich Überschätzungen reduzieren.

Für die praktische Anwendung bedeutet dies insbesondere, daß eine Vorwärtsfehlerberechnung, deren Zwischenergebnisse immer wieder aufeinander aufbauen, ungünstig ausfallen kann. Ein Ausweg ist:

I) Näherungsverfahren einzusetzen, die intervallmäßig auszuwertenden Teile auf ein Minimum zu beschränken und statt der Lösung den Fehler gegenüber einer Näherungslösung einzuschließen,

II) Intervalle möglichst immer mit einem kleinen Faktor zu versehen, und

III) sich immer wieder auf die Ausgangsdaten zu beziehen, da diese originär und nicht fehlerbehaftet sind.

Das hört sich zunächst recht theoretisch an; im folgenden werden anhand von Fallbeispielen solche Algorithmen diskutiert. Regel I) ist übrigens ganz im Sinne von Wilkinson [39], der bereits 1971 schrieb:

> Wilkinson: „In general it is the best in algebraic computations to leave the use of interval arithmetic as late as possible so that it effectively becomes an a posteriori weapon."

Im folgenden werden zunächst Verfahren ganz in diesem Sinne vorgestellt, die den Fehler von gleitkommamäßig berechneten Näherungen abschätzen. In Abschnitt 6 wird auch ein Verfahren angedeutet, daß sich intervallmäßig gewonnene Informationen a priori zu Nutze macht.

Vorher soll noch bemerkt werden, daß sich das Konzept der Fehlerabschätzung ebenso auf Vektoren und Matrizen ausdehnt. Die Komponenten sind dann Intervalle und die Operationen werden nach den üblichen Formeln ausgeführt. Die Menge der Intervallvektoren sei mit $\mathbb{IR}^n$, die der Intervallmatrizen mit $\mathbb{IR}^{m \times n}$ bezeichnet. Was hier in kürzester Form angedeutet wird, bedarf einer theoretischen Untermauerung, wie sie in Standardtextbüchern gegeben wird [1], [2], [18], [22], [24], [28].

Etwas später benötigen wir dabei noch folgende Beobachtung. Seien eine Intervallmatrix $[A]$ und ein reeller Vektor x gegeben. Dann überschätzt die Multiplikation $[A] \cdot x$ das Ergebnis *nicht*. Das sieht man daran, daß alle Intervallkomponenten $[A]_{ij}$ in dem Ausdruck nicht mehr als einmal vorkommen. Bei der Multiplikation einer reellen Matrix A mit einem Intervallvektor $[x]$ hingegen gibt es eine Überschätzung, den „wrapping effect". Es kann zwar keine Komponente verkleinert werden, doch nicht zu jedem Punkt y in $A \cdot [x]$ gibt es ein $x \in [x]$ mit $y = A \cdot x$.

$$A = \begin{pmatrix} -0.75 & 0.50 \\ 0.50 & 0.25 \end{pmatrix}$$
$$[x] = \begin{pmatrix} [1,2] \\ [1,3] \end{pmatrix}$$

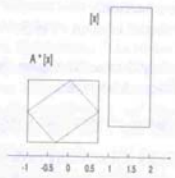


Abbildung 2.1 Der „wrapping effect"

Die gestrichelte Linie in $A \cdot [x]$ hüllt die wahre Wertemenge $\{ A \cdot x \mid x \in [x] \}$ ein, die durchgezogene Linie ist der engste Intervallvektor, der $A \cdot [x]$ einschließt. Letzterer wird durch die intervallmäßige Matrix-Vektor-Multiplikation auch wirklich berechnet. Für Matrix-Matrix-Multiplikation gelten entsprechende Überlegungen.

3 Vollbesetzte, lineare Gleichungssysteme

In diesem Abschnitt beschreiben wir der Einfachheit halber ein Verfahren mit einer Näherungsinversen R, etwas später wird diese dann durch eine Zerlegung ersetzt. Vorab sei bemerkt, daß sich die folgenden Konzepte auf Systeme nichtlinearer Gleichungen ausdehnen lassen. Dies wird später auch behandelt. Wir bevorzugen die Darstellung anhand linearer Gleichungssysteme, um die interessanten Punkte nicht durch den bei nichtlinearen Gleichungssystemen notwendigen Formalismus unnötig zu stören.

Gegeben sei ein lineares Gleichungssystem $Ax = b$, eine Näherungsinverse $R \approx A^{-1}$ und es sei $\widetilde{x} := R \cdot b$ eine Näherungslösung des Gleichungssystems. Gesucht ist die Nullstelle $\widehat{x}$ von $f(x) = b - Ax$. Offenbar ist die Fixpunktgleichung

$$g(x) = x \qquad \text{mit} \qquad g(x) := x + R(b - Ax)$$

äquivalent zur Gleichung $f(x) = 0$, falls R regulär ist. Die vorhin diskutierte Möglichkeit, den Wertebereich von Funktionen rigoros abschätzen zu können, legt die Anwendung eines Fixpunktsatzes nahe. Kann für einen Intervallvektor X nachgewiesen werden, daß

$$g(X) \subseteq X \tag{3.1}$$

gilt, folgt aus der Nichtsingularität von R und A die Existenz und Eindeutigkeit der Nullstelle $\widehat{x}$ von f und es ist $\widehat{x} \in X$.

Ersetzt man in der Funktion g die reelle Variable x durch eine Intervallvariable X und testet

$$X + R \cdot (b - A \cdot X) \subseteq X, \tag{3.2}$$

so kommt der Intervallvektor X zweimal vor, was den Wertebereich überschätzt. Im obigen Falle ist dieser Sachverhalt besonders unglücklich, da

$$\operatorname{diam}\big(X + R \cdot (b - A \cdot X)\big) = \operatorname{diam}(X) + \operatorname{diam}\big(R \cdot (b - A \cdot X)\big),$$

was den Nachweis von (3.1) in dieser Form unmöglich macht. Eine leichte Umformung ergibt aber $g(x) = Rb + (I - RA)x$, wobei I die Einheitsmatrix bezeichnet. Der Test

$$R \cdot b + (I - R \cdot A) \cdot X \subseteq X \tag{3.3}$$

ermöglicht in vielen Fällen leicht den Nachweis von (3.1) . Hier werden die Grundregeln I), II) und III) beherzigt. Es kommt der einzige Intervallterm, nämlich X, in (3.3) genau einmal vor, es gibt also *keine* Überschätzung, dieser Term ist mit dem kleinen Faktor $I - R \cdot A$ versehen, was nochmals dämpft, und mit A und b gehen immer wieder die Originaldaten in die Iteration ein. Für $R \approx A^{-1}$ gilt $I - R \cdot A \approx 0$.

Die Behauptung, es gäbe keine Überschätzung bei der Auswertung von (3.3) bedarf noch einer Erläuterung. Sie ist richtig in dem Sinne, daß die rechte Seite von (3.3) den *engsten* einschließenden Intervallvektor liefert, andererseits ist, wie im vorigen Abschnitt besprochen, die rechte Seite leicht vergrößert durch den wrapping effect.

Es bleibt zu verifizieren, daß R und A nicht singulär sind. Das kann z.B. durch den Nachweis von $\|I - RA\| < 1$ erfolgen (vgl. [15], [16], [21]) oder auch durch folgenden Satz [30].

Satz 3.1 Seien $R, A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$ und $X \in \mathbb{IR}^n$ [1] gegeben. Dann folgt aus

$$R \cdot b + (I - R \cdot A) \cdot X \subseteq \operatorname{int}(X) \tag{3.4}$$

die Regularität von R und A sowie $A^{-1}b \in X$.

Hierbei bezeichnet $\operatorname{int}(X)$ das topologisch Innere von X; in einer Implementation ist das einfach das Ersetzen von $\leq$ durch $<$.

In dieser Form des Satzes liegt bereits ein hinreichendes Kriterium für die Einschließung der Lösung von $Ax = b$ vor. Es fehlt noch ein *konstruktives* Verfahren zur Berechnung eines einschließenden Intervallvektors X. Man kann die linke Seite von (3.4) als Iteration schreiben in der Form

$$X^{k+1} := R \cdot b + (I - RA) \cdot X^k \tag{3.5}$$

für gegebenes $X^0 \in \mathbb{IR}^n$ und $X^{k+1} \subseteq \operatorname{int}(X^k)$ prüfen. Damit wären die Bedingungen von Satz 3.1 erfüllt. Betrachten wir dazu das wohl einfachste lineare Gleichungssystem mit nicht-singulärer Matrix

$$1 \cdot x = 1.$$

Ist die Näherungsinverse $R \approx A^{-1} = 1$ mäßig genau berechnet, etwa $R = 0.8$, und legt man um $\widetilde{x}$ ein kleines Intervall (was dann hoffentlich die Lösung enthält), etwa $X^0 := [0.7, 0.9]$, so ergibt sich aus $I - RA = 0.2$ mit (3.5)

$$X^1 = 0.8 + 0.2 \cdot [0.7, 0.9] = [0.94, 0.98]$$
$$X^2 = [0.988,\ 0.996]$$
$$X^3 = [0.9976,\ 0.9992] \ldots.$$

Das heißt, X^k konvergiert offensichtlich gegen 1 (was auch aus $|I - RA| = 0.2 < 1$ folgt), aber $X^{k+1} \subseteq \operatorname{int}(X^k)$ ist niemals erfüllt. Man muß also die Intervalle während der Iteration künstlich erweitern, und das geschieht durch die ε-Aufblähung. Für gegebenes $X^0 \in \mathbb{IR}^n$ ist eine Form der ε-Aufblähung gegeben durch

[1] $\mathbb{IR}^n$ bezeichnet die Menge der Intervallvektoren mit n Komponenten, $\mathbb{IR}^{n \times n}$ entsprechend Matrizen.

$$Y^k := X^k + [-\varepsilon, +\varepsilon]; \quad X^{k+1} := R \cdot b + (I - RA) \cdot Y^k.$$

Im n-dimensionalen Fall ist die Addition von $[-\varepsilon, \varepsilon]$ komponentenweise zu verstehen. Was zunächst wie ein einfacher Behelf aussieht, führt zu einem vollständigen Überblick der Konvergenz der obigen Einschließungsiteration [32].

Satz 3.2 Seien $R, A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$ und $X^0 \in \mathbb{IR}^n$ gegeben. Für $0 < \varepsilon \in \mathbb{R}$ sei

$$Y^k := X^k + [-\varepsilon, +\varepsilon]; \quad X^{k+1} := R \cdot b + (I - RA) \cdot Y^k. \tag{3.6}$$

Werden in (3.6) alle Operationen intervallmäßig ausgeführt, so sind folgende Aussagen äquivalent:

a) $\exists\, k \in \mathbb{N} : X^{k+1} \subseteq \mathrm{int}(Y^k)$ und somit Nachweis der Regularität von R und A sowie $A^{-1}b \in X^{k+1}$.

b) $\rho(|I - RA|) < 1$.

ρ bezeichnet den Spektralradius einer Matrix. Gegenüber den üblichen Sätzen über die Konvergenz der Residueniteration muß also der Betrag der Iterationsmatrix kontrahierend, also vom Spektralradius kleiner 1 sein. Man beachte, daß *genau dann* eine Einschließung der Lösung durch die Iteration (3.6) gefunden wird.

Für eine praktische Implementation können obige Sätze natürlich nur eine sehr grobe Richtschnur sein. Man würde etwa statt $Ax = b$ das Gleichungssystem $Ay = b - A\tilde{x}$ lösen, um zu scharfen Einschließungen zu gelangen, in (3.4) kann die Überprüfung der Einschließung durch ein Einzelschrittverfahren ersetzt werden, es kann ein genaues Skalarprodukt [18] eingesetzt werden und anderes mehr.

Im Hinblick auf nichtlineare Gleichungssysteme ist eine Anwendung interessant, nämlich die auf toleranzbehaftete Eingabedaten. Sind die Daten des linearen Gleichungssystems nicht exakt, sondern nur innerhalb gewisser Grenzen $[A] \in \mathbb{IR}^{n \times n}$, $[b] \in \mathbb{IR}^n$ gegeben, dann wird man sich für die Menge aller Lösungen

$$\Sigma([A], [b]) := \{\, x \in \mathbb{R}^n \mid \exists\, A \in [A]\ \exists\, b \in [b] : Ax = b \,\}$$

interessieren. Die grundlegende Eigenschaft der Intervalloperationen, die Isotonie, erlaubt es, Satz 3.1 unmittelbar auf toleranzbehaftete Daten auszudehnen. Auch die vollständige Beschreibung der Konvergenzeigenschaften (3.6) einer entsprechenden Iteration läßt sich übertragen:

Satz 3.3 Sei $[A] \in \mathbb{IR}^{n \times n}$, $[b] \in \mathbb{IR}^n$, $R \in \mathbb{R}^{n \times n}$ und $X^0 \in \mathbb{IR}^n$ gegeben. Für $0 < \varepsilon \in \mathbb{R}$ sei

$$Y^k := X^k + [-\varepsilon, \varepsilon]; \quad X^{k+1} := R \cdot [b] + (I - R \cdot [A]) \cdot Y^k. \tag{3.7}$$

Werden in (3.7) alle Operationen intervallmäßig ausgeführt, so folgt für $k \geq 0$ aus

$$X^{k+1} \subseteq \mathrm{int}(Y^k) \tag{3.8}$$

daß R und jedes $A \in [A]$ regulär sind und daß $\Sigma([A], [b]) \subseteq X^{k+1}$ gilt.

Außerdem sind folgende Aussagen äquivalent:

a) $\exists\, k \in \mathbb{N} : X^{k+1} \subseteq \mathrm{int}(Y^k)$.

b) $\rho(|I - R \cdot [A]|) < 1$.

Der Betrag einer Intervallmatrix ist dabei definiert als die komponentenweise betragsmäßig größte Matrix innerhalb dieser Intervallmatrix. Man beachte, daß hier wie in den vorhergehenden Sätzen keinerlei Voraussetzungen an das Gleichungssystem, an R oder an X^0 gestellt wurden, allein die Einschließungsbedingung erlaubt die gezogenen Schlußfolgerungen. Insbesondere ist das Ausgangsintervall X^0 beliebig.

Der Durchmesser des Lösungskomplexes $\Sigma([A], [b])$ wächst mit der Kondition des Gleichungssystems. Für schlecht konditionierte Gleichungssysteme wird man auch für kleine Durchmesser von $[A]$ und $[b]$ mit weitem $\Sigma([A], [b])$ rechnen müssen. Satz 3.3 erlaubt in Form von (3.8) allerdings nur eine (äußere) *Einschließung* X^{k+1} von $\Sigma([A], [b])$. Nach den Vorbemerkungen aus Absatz 2 könnte der wahre Bereich durch X^{k+1} erheblich überschätzt werden. Oder umgekehrt: Läßt eine grobe Einschließung für $\Sigma([A], [b])$ auf schlechte Kondition schließen?

Interessanterweise läßt sich die Genauigkeit, mit der X^{k+1} den Lösungskomplex einschließt, rigoros abschätzen. Betrachten wir den Wertebereich $\Sigma([A], [b])$ komponentenweise. Bezeichne $\Sigma([A], [b])_i$ das engste Intervall, das die i-te Komponente der Lösungsmenge einschließt. Jeder Punkt in diesem Intervall ist tatsächlich i-te Komponente einer Lösung $A^{-1}b$, $A \in [A]$, $b \in [b]$, da alle $A \in [A]$ als regulär nachgewiesen werden und die Lösung eines Gleichungssystems stetig von den Eingabedaten $A \in [A]$, $b \in [b]$ abhängt.

Für eine Einschließung $X \supseteq \Sigma([A], [b])$ gilt dann natürlich $\Sigma([A], [b])_i \subseteq X_i$. Neben der äußeren Einschließung X_i kann fast ohne Mehraufwand auch eine *innere Einschließung* Y_i jeder Komponente mit $Y_i \subseteq \Sigma([A], [b])_i$ berechnet werden [23], [31].

Satz 3.4 Sei $[A] \in \mathbb{IR}^{n \times n}$, $[b] \in \mathbb{IR}^n$, $R \in \mathbb{R}^{n \times n}$, $\tilde{x} \in \mathbb{R}^n$ und $X \in \mathbb{IR}^n$ gegeben. Sei

$$Z := R \cdot ([b] - [A] \cdot \tilde{x}), \quad \Delta := (I - R \cdot [A]) \cdot X$$

und

$$Z + \Delta \subseteq \mathrm{int}(X).$$

Dann ist R und jedes $A \in [A]$ regulär und mit $Z_i := [\underline{Z}_i, \overline{Z}_i]$, $\Delta_i := [\underline{\Delta}_i, \overline{\Delta}_i]$ gilt

$$[\underline{Z}_i + \overline{\Delta}_i, \overline{Z}_i + \underline{\Delta}_i] \subseteq \tilde{x} + \Sigma([A], [b])_i \subseteq [\underline{Z}_i + \underline{\Delta}_i, \overline{Z}_i + \overline{\Delta}_i].$$

Damit ist eine komponentenweise Mindest- und Höchstvariation für den Lösungskomplex gegeben. Die Güte der Abschätzung, d.h. der Unterschied zwischen unterer und oberer Abschätzung für den Durchmesser der Lösungsmenge ist gerade $\mathrm{diam}(\Delta)$. Ist R eine Näherungsinverse der Mittelpunktsmatrix von $[A]$, so ist bei nicht zu großem Durchmesser von $[A]$ die Matrix $I - R \cdot [A]$ klein. Außerdem sind wir zur Lösung des Defektgleichungssystems übergegangen, d.h., X schließt jetzt den *Fehler* gegenüber der Näherungslösung $\tilde{x}$ ein. Demzufolge ist X ebenfalls klein. Das bedeutet, daß das Produkt Δ sehr klein ausfällt und damit die Außen- und Inneneinschließung sehr scharf wird.

Betrachten wir ein Beispiel. Gegeben sei die Matrix $A \in \mathbb{R}^{n \times n}$ mit

$$A_{ij} := \left(\frac{i+j}{p}\right) \quad \text{für } p = n + 1,$$

wobei $\left(\frac{k}{p}\right)$ das Legende-Symbol bezeichnet:

$$\left(\frac{k}{p}\right) := \begin{cases} 0 & \text{falls } k \text{ Teiler von } p \\ 1 & \text{falls } k \equiv c^2 \ mod p \text{ für ein } c \\ -1 & \text{sonst} \end{cases}$$

(siehe [5], Beispiel 3.14). Es ist für $n = 4$

$$A = \begin{pmatrix} -1 & -1 & 1 & 0 \\ -1 & 1 & 0 & 1 \\ 1 & 0 & 1 & -1 \\ 0 & 1 & -1 & -1 \end{pmatrix}.$$

Die rechte Seite b wird aus der Lösung $x \in \mathbb{R}^n$ mit

$$x_i := \frac{(-1)^{i+1}}{i} \quad 1 \le i \le n$$

berechnet. Jetzt wird die Matrix und rechte Seite gestört, und zwar

$$[A] := A \cdot (1 \pm e), \ [b] := b \cdot (1 \pm e) \quad \text{mit } e = 10^{-5}.$$

Betrachten wir den Fall $n = 1008$ ($n + 1 = p$ muß prim sein). Die Rechnung ist in einfacher Genauigkeit entsprechend etwa 7 Dezimalen durchgeführt. Dann ergibt sich für die innere und äußere Einschließung

Innere und äußere Einschließungen für einige Komponenten								$\frac{\text{diam}(X)}{\text{diam}(Y)}$
[0.99973,	1.000127]	$\subseteq$	$\Sigma([A],[b])_1$	$\subseteq$	[0.99969,	1.000131]		0.96980
[−0.500127,	−0.49973]	$\subseteq$	$\Sigma([A],[b])_2$	$\subseteq$	[−0.500131,	−0.49969]		0.96975
[0.33306,	0.33360]	$\subseteq$	$\Sigma([A],[b])_3$	$\subseteq$	[0.33303,	0.33364]		0.96978
			...					
[−0.001121,	−0.00067]	$\subseteq$	$\Sigma([A],[b])_{1006}$	$\subseteq$	[−0.001125,	−0.00063]		0.96979
[0.00066,	0.001120]	$\subseteq$	$\Sigma([A],[b])_{1007}$	$\subseteq$	[0.00062,	0.001124]		0.96981
[−0.001119,	−0.00065]	$\subseteq$	$\Sigma([A],[b])_{1008}$	$\subseteq$	[−0.001123,	−0.00061]		0.96977

Um die Qualität der inneren Einschließung gegenüber der äußeren zu beurteilen, ist das Verhältnis der Durchmesser interessant, also $\text{diam}(X)/\text{diam}(Y)$, wenn X bzw. Y die innere bzw. äußere Einschließung bezeichnet. Diese Verhältnisse sind in der letzten Spalte obiger Tabelle angegeben. Das kleinste und damit ungünstigste Verhältnis für $1 \le i \le 1008$ ist hierbei

Innere und äußere Einschließungen für einige Komponenten						$\frac{\text{diam}(X)}{\text{diam}(Y)}$
[−0.00841,	−0.00894]	$\subseteq$	$\Sigma([A],[b])_{116}$	$\subseteq$	[−0.00851,	−0.00890]	0.96967

M.a.W. bis auf rund 3 % genau kennt man die exakte Ausdehnung des Lösungskomplexes in jeder Komponente von 1 bis 1008. Ersetzt man die relative Störung 10^{-5} durch eine absolute Störung 10^{-5}, d.h. die Nullkomponenten werden auch gestört, ergeben sich fast genau die gleichen Ergebnisse. Das heißt, das lineare Gleichungssystem reagiert wenig auf zusätzliche Störungen der Nullkomponenten.

Für eine Darstellung verschiedener Aspekte und vieler Details der Behandlung linearer und nichtlinearer Gleichungssysteme mit Intervallmethoden verweisen wir auf [1], [22], [24].

Bisher wurde vorausgesetzt, daß die Eingabedaten $A_{ij} \in [A_{ij}]$, $b_i \in [b_i]$ unabhängig voneinander variieren. In praktischen Anwendungen ist diese Voraussetzung häufig nicht erfüllt,

beispielsweise wenn symmetrische Matrizen vorliegen. In diesem Fall ist man an der Menge aller Lösungen

$$\Sigma^{sym}([A],[b]) := \{ x \in \mathbb{R}^n \mid \exists\, A \in [A], b \in [b], A \text{ symmetrisch} : Ax = b \}$$

interessiert. Wir bemerken hier, daß mit einem ähnlichen Verfahrenstyp sehr scharfe komponentenweise Innen- und Außeneinschließungen für die Lösungskomponenten von $\Sigma^{sym}([A],[b])$ berechnet werden können [11].

Abschließend noch eine Bemerkung zur Rechenzeit. Selbst in dieser hier vorgestellten, sehr einfachen Form benötigt ein Einschließungslogarithmus den sechsfachen Aufwand gegenüber einem direkten Eliminationsalgorithmus. Letzterer benötigt $n^3/3$ Operationen im Vergleich zu n^3 Operationen für R und n^3 Operationen für $R \cdot A$ mit Einschließung, da $R \cdot A$ nur nach oben gerundet benötigt wird. Es sei bemerkt, daß durch eine Modifikation des Algorithmus mit Hilfe der Wilkinsonschen Rückwärtsanalyse auf die Berechnung des Produktes $R \cdot A$ ganz verzichtet werden kann, was die Rechenzeit nochmals beträchtlich veringert.

Der Faktor ist unabhängig von n, es werden verifizierte, richtige Ergebnisse geliefert, und im Falle von toleranzbehafteten Daten wird auch eine verifizierte Inneneinschließung des Lösungskomplexes berechnet. Im nächsten Abschnitt werden wir noch effizientere, insbesondere für spärlich besetzte Gleichungssysteme geeignete Algorithmen vorstellen.

4 Große, spärlich besetzte Gleichungssysteme

Für größere, insbesondere für bandartige oder spärlich besetzte Gleichungssysteme kommt die Berechnung einer Näherungsinversen auf keinen Fall in Frage. Diese ist in der Regel vollbesetzt, und ein mit Einschließungsformeln der Art (3.4) aufgebauter Algorithmus würde im Vergleich zu bekannten Näherungsmethoden einen enormen Rechen- und Speicheraufwand benötigen und wäre auf den meisten Rechenanlagen schon für $n \ge 10000$ nicht durchführbar. Gehen wir im folgenden von einer Bandmatrix A mit unterer bzw. oberer Bandbreite p bzw. q oder von einer spärlich besetzten Matrix aus. Die folgenden Verfahren benutzen die Tatsache, daß ein Eliminationsalgorithmus wie LU, LDL^T, LDM^T o.ä. die Bandbreite bzw. das Profil nicht verändert.

Man könnte in (3.4) die Näherungsinverse R durch eine näherungsweise berechnete Zerlegung, z.B. LU, ersetzen. Im folgenden bezeichnen wir symbolisch mit U^{-1} bzw. L^{-1} irgendeine Vorwärts- bzw. Rückswärtsauflösung für eine obere bzw. untere Dreiecksmatrix. Dann ergibt sich für g mit $R := (L \cdot U)^{-1} = U^{-1} \cdot L^{-1}$

$$\begin{aligned} g(x) &= U^{-1} \cdot L^{-1} \cdot b + (I - U^{-1} \cdot L^{-1} \cdot A)x \\ &= U^{-1} \cdot L^{-1} \cdot \{b + (L \cdot U - A)x\}. \end{aligned} \tag{4.1}$$

Die Größe $LU - A$ kann während der LU-Zerlegung von A günstig mit berechnet werden; mit der Crout-Variante sind das gerade zusätzliche n^2 Operationen, für Bandmatrizen nur $n \cdot (p+q+1)$. Aus der Fehlertheorie des Gauß-Algorithmus ist bekannt, daß $LU - A$ in der Regel sehr klein sein wird. Geht man zum Defektgleichungssystem $Ay = b - A\tilde{x}$ für eine Näherungslösung $\tilde{x}$ über, dann wird aus (4.1)

$$g(x) = U^{-1}L^{-1}\{b - A\tilde{x} + (LU - A) \cdot x\}. \tag{4.2}$$

Der Nachweis von $g(X) \subseteq \mathrm{int}(X)$ für ein X genügt wieder für alle Schlußfolgerungen, die im letzten Abschnitt gezogen wurden. Außerdem wird der Ausdruck in der geschweiften Klammer in (4.2) jetzt i.d.R. sehr klein. Er müßte allerdings zur Verifikation eingeschlossen werden, was einen Intervallvektor etwa mit Komponenten $[-\varepsilon, +\varepsilon]$ ergäbe. Läßt man zunächst noch der Einfachheit halber die Rückwärtsauflösung mit U weg und skaliert die geschweifte Klammer zu einem Intervallvektor mit Komponenten $[-1, 1]$, so ist es zum Nachweis von $g(X) \subseteq \mathrm{int}(X)$ mittels (4.2) notwendig, mindestens

$$L^{-1}[e] \quad \text{mit} \quad [e] \in \mathbb{IR}^n,\ [e]_i \equiv [-1, 1] \tag{4.3}$$

einzuschließen. Dies könnte mittels intervallmäßiger Vorwärtssubstitution erreicht werden. Betrachten wir als Beispiel

$$L = \begin{pmatrix} 1 & & & & \\ 1 & 1 & & & \\ 1 & 1 & 1 & & \\ & 1 & 1 & 1 & \\ & & 1 & 1 & 1 \\ & & & & \ddots \end{pmatrix}. \tag{4.4}$$

Dann ergibt sich der Reihe nach für $[X] := L^{-1}[e]$ intervallmäßig ausgeführt:

$$\begin{aligned} [X]_1 &= [e]_1 \\ [X]_2 &= [e]_2 - [X]_1 = [-1, 1] - [-1, 1] = [-2, 2] \\ [X]_3 &= [e]_3 - [X]_1 - [X]_2 = [-1, 1] - [-1, 1] - [-2, 2] = [-4, 4] \\ [X]_4 &= [e]_4 - [X]_2 - [X]_3 = [-1, 1] - [-2, 2] - [-4.4] = [-7, 7] \\ &\ldots \end{aligned}$$

Offensichtlich ist das Ergebnis X ein zum Nullpunkt symmetrisches Intervall $[-x, x]$ und x kann aus

$$\begin{pmatrix} 1 & & & & \\ -1 & 1 & & & \\ -1 & -1 & 1 & & \\ & -1 & -1 & 1 & \\ & & -1 & -1 & 1 \\ & & & & \ldots \end{pmatrix} \cdot x = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ \vdots \end{pmatrix} \tag{4.5}$$

berechnet werden. Die Matrix in (4.5) ist die Ostrowskische Vergleichsmatrix $\langle L \rangle$ von L. Die durch die intervallmäßig ausgeführte Vorwärtssubstitution eingeschleppte Überschätzung ist offenbar $\|\langle L \rangle^{-1}\|_\infty \,/\, \|L^{-1}\|_\infty$. Für einige Werte von n ergibt sich

n	10	20	30	40	50
$\|\langle L \rangle^{-1}\|_\infty \,/\, \|L^{-1}\|_\infty$	20.4	1265	1.0e5	9.9e6	9.7e8

Das Verhältnis wächst exponentiell mit der Dimension n. Die Multiplikation mit der Inversen von L gibt zwar scharfe Ergebnisse, scheidet jedoch aus, da die Inverse des (spärlichen) L i.d.R. vollbesetzt ist.

Bei der intervallmäßig ausgeführten Vorwärtssubstitution ist Regel III) aus Abschnitt 2 verletzt. Ein intervallmäßig durchgeführter Algorithmus, dessen Zwischenergebnisse laufend *aufeinander* aufbauen, neigt häufig zu Überschätzungen. Dies muß allerdings nicht so sein, z.B. für den Spezialfall von M-Matrizen ist $\langle L \rangle = L$, und obiges Verfahren berechnet den exakten Lösungskomplex (siehe [1], [24], [34]).

Statt (4.3) benötigen wir ja in Wirklichkeit die Intervallhülle von

$$\{ y \in \mathbb{R}^n \mid Ly = e, e \in [e] \}; \tag{4.6}$$

es ist also ein Verfahren gesucht, (4.6) einzuschließen, ohne die Inverse von L explizit berechnen zu müssen. Seien $\sigma_1 \ldots \sigma_n$ die Singulärwerte von L. Dann gilt

$$x \in L^{-1} \cdot b \quad \Rightarrow \quad \|x\|_2 \leq \sigma_n^{-1} \cdot \|b\|. \tag{4.7}$$

Mit einer *unteren* Abschätzung des kleinsten Singulärwerts von L hätte man also eine Einschließung von (4.6). Das bedeutet aber die Berechnung eines *verifizierten* Konditionschätzers. Letzterer liefert aber nur obere Schranken für σ_n, während in der Ungleichung (4.7) eine verifizierte untere Schranke für σ_n benötigt wird.

σ_n^2 ist kleinster Eigenwert von LL^T. Von einem $\widetilde{\lambda} \in \mathbb{R}$ könnte $\widetilde{\lambda} \leq \sigma_n^2$ durch den Nachweis der positiven Semidefinitheit von $LL^T - \widetilde{\lambda} I$ gezeigt werden. Letzteres bedeutet, daß die Cholesky-Zerlegung von $LL^T - \widetilde{\lambda} I$ durchführbar ist. Eine intervallmäßige Durchführung des Cholesky-Verfahrens führte zwar den Nachweis der positiven Semidefinitheit von $LL^T - \widetilde{\lambda} I$, falls alle Diagonalelemente nichtnegativ bleiben, kommt wegen der Verletzung von Regel II) jedoch nicht in Betracht. Ab einigen hundert Unbekannten blähen sich die Intervalle zu sehr auf.

Unter Berücksichtigung von Regel I) und mit Perturbationssätzen für Eigenwerte symmetrischer Matrizen rechnet man in *Gleitkommaarithmetik* einen genäherten Cholesky-Faktor $\widetilde{G}$ mit $\widetilde{G}\widetilde{G}^T \approx LL^T - \widetilde{\lambda} I$ aus. Ist der Algorithmus durchführbar, d.h. bleiben alle Diagonalelemente von $\widetilde{G}$ nichtnegativ, so kann

$$\sigma_n \geq (\widetilde{\lambda} - \|LL^T - \widetilde{\lambda} I - \widetilde{G}\widetilde{G}^T\|)^{1/2}$$

gezeigt werden für jede konsistente Norm $\|\cdot\|$, also etwa $\|\cdot\|_1$ oder $\|\cdot\|_\infty$. Die Abschätzung von $\|LL^T - \widetilde{\lambda} I - \widetilde{G}\widetilde{G}^T\|$ kann recht günstig durchgeführt werden ohne die Matrix LL^T explizit speichern zu müssen. Die Näherung $\widetilde{\lambda}$ ermittelt man mit inverser Potenziteration für L und L^T oder mit einem Konditionsschätzer.

Um auf die Form (4.2) zurückzukommen, wird eine untere Abschätzung des kleinsten Singulärwerts von LU benötigt. Diese ergibt sich aus $\sigma_n(LU) \geq \sigma_n(L) \cdot \sigma_n(U)$. Ein Einschließungsalgorithmus kann dann auf folgendem Satz aufgebaut werden.

Satz 4.1 Sei $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$, $\widetilde{x} \in \mathbb{R}^n$, $\widetilde{L}, \widetilde{U} \in \mathbb{R}^{n \times n}$ regulär und $\widetilde{A} := \widetilde{L}\widetilde{U}$. Sei $\Delta A := \widetilde{A} - A$ und $\sigma_n(\widetilde{A}) > (\|\Delta A\|_1 \cdot \|\Delta A\|_\infty)^{1/2}$. Dann ist A nicht singulär und für $\widehat{x} := A^{-1}b$ gilt

$$\|\widehat{x} - \widetilde{x}\|_\infty \leq \|\widehat{x} - \widetilde{x}\|_2 \leq \frac{\|b - A\widetilde{x}\|_2}{\sigma_n(\widetilde{A}) - (\|\Delta A\|_1 \cdot \|\Delta a\|_\infty)^{1/2}}.$$

Die Regularität von $\widetilde{L}$ und $\widetilde{U}$ erkennt man unmittelbar an den Diagonalelementen; ΔA kann wie gesagt zusammen mit $\widetilde{L}$ und $\widetilde{U}$ praktisch ohne Mehraufwand gewonnen werden. Der Gesamtaufwand dieses Verfahrens *mit* Ergebnisverifikation beträgt für Bandmatrizen A mit unterer bzw. oberer Bandbreite p bzw. q

A	M-Matrix:	$n \cdot pq$	Operationen
A	symmetrisch positiv definit:	$n \cdot p^2$	Operationen
A	symmetrisch indefinit:	$\frac{3}{2}n \cdot p^2$	Operationen
A	allgemeine Matrix:	$n \cdot (pq + p^2 + q^2)$	Operationen

Der Aufwand der Einschließung entspricht also dem der Zerlegung, der Gesamtaufwand steigt um den Faktor 2, unabhängig von n.

Betrachten wir etwa als Beispiel (4.16) aus [5], gerechnet in doppelter Genauigkeit entsprechend etwa 17 Dezimalen und rechter Seite so berechnet, daß die i-te Komponente der Lösung gleich $(-1)^{i+1}/i$ ist.

n	cond(A)	$\sigma_{\min}$	$\|\widehat{x} - \widetilde{x}\|_\infty / \|\widetilde{x}\|_\infty$
2 000	2.63E+12	2.46E-06	7.01E-13
5 000	1.03E+14	3.95E-07	2.53E-11
10 000	1.64E+15	9.87E-08	5.38E-10
20 000	2.63E+16	2.47E-08	1.83E-08
50 000	1.03E+18	4.05E-09	Abbruch

Abbildung 4.1. Matrix (4.16) aus [5]

$\sigma_{\min}$ bezeichnet den kleinsten Singulärwert von L. Als weiteres Beispiel betrachten wir die Matrix $A = 0.1 \cdot LL^T$, wobei L durch (4.4) definiert ist.

n	cond	$\sigma_{\min}(A)$	$\|\widehat{x} - \widetilde{x}\|_\infty / \|\widetilde{x}\|_\infty$
10 000	1.22E+08	2.72E-04	3.39E-17
20 000	4.87E+08	1.36E-04	1.35E-16
50 000	3.04E+09	5.44E-05	8.47E-16
100 000	1.22E+10	2.72E-05	3.39E-15
500 000	3.04E+11	5.44E-06	8.47E-14
1 000 000	1.22E+12	2.72E-06	3.39E-13

Abbildung 4.2. $A = 0.1 \cdot LL^T$, L definiert durch (4.4)

Der Faktor 0.1 ist eingeführt, damit der Gleitpunkt-Zerlegungsalgorithmus nicht das exakte L berechnet. Wie man sieht, werden auch für größere Dimensionen ordentliche maximale relative Fehler erzielt.

Als letztes Beispiel noch einige spärlich besetzte Gleichungssysteme aus den Harwell Testcases. Die rechte Seite ist wiederum so berechnet, daß die i-te Komponente der Lösung $(-1)^{i+1}/i$ wird.

Matrix	n	p	q	profile	cond	$\|A - \widetilde{L}\widetilde{U}\|_2$	$\|\widehat{x} - \widetilde{x}\|_\infty / \|\widetilde{x}\|_\infty$
gre_216	216	14	36	876	2.7e+02	3.1e-15	7.9e-27
gre_343	343	18	49	1435	2.5e+02	5.6e-15	2.4e-26
gre_512	512	24	64	2192	3.8e+02	7.4e-15	6.8e-26
west0167	167	158	20	507	2.8e+06	1.6e-16	4.6e-22
west0381	381	363	153	2157	2.0e+06	1.1e-15	8.8e-25
bcsstk08	1074	590	590	7017	6.1e+06	1.6e-16	6.6e-23
bcsstk14	1806	161	161	32630	4.3e+04	1.8e-15	1.8e-25

Abbildung 4.3. Die „Harwell test cases"

Die im vorigen Abschnitt entwickelten Konzepte für toleranzbehaftete Daten und für Inneneinschließungen lassen sich auch auf den Fall großer, spärlich besetzter Gleichungssysteme übertragen. Es sei jedoch betont, daß die in diesem Abschnitt vorgestellten Verfahren noch sehr jung sind und die Forschung gerade am Anfang steht. Verfahren für spärliche Systeme, die auf intervallmäßige Rückwärtsauflösung verzichten und somit exponentielle Überschätzungen vermeiden, wurden unseres Wissens nach erstmals in [33] behandelt.

5 Nichtlineare Gleichungssysteme

Es wurde bereits angedeutet, daß die Konzepte für lineare Gleichungssysteme, insbesondere auch für solche mit großer, spärlicher Struktur, sich auf Systeme nichtlinearer Gleichungen übertragen. Gegeben sei ein parametrisiertes Gleichungssystem

$$f(x, p) = 0 \quad \text{mit } f : \mathbb{R}^{n \times k} \to \mathbb{R}^n.$$

Wir setzen voraus, daß für f eine intervallmäßige Auswertung F gegeben ist, d.h.

$$X \in \mathbb{IR}^n, P \in \mathbb{IR}^k : (x, p) \in (X, P) \Rightarrow f(x, p) \in F(X, P). \tag{5.1}$$

Ist f beispielsweise in Form eines Programms gegeben, das aus arithmetischen Operationen, Standardfunktionen, Schleifen usw. besteht, kann F gemäß den Bemerkungen aus Abschnitt 2 einfach hergestellt werden, indem jede reelle (oder komplexe) Operation durch die entsprechende Intervalloperation ersetzt wird. Aus der Isotonie der Einzeloperationen folgt dann Eigenschaft (5.1).

f kann allerdings auch implizit gegeben sein, die Konstruktion eines geeigneten F wird dann schwieriger. Eine Funktion, die durch Funktionswerte an endlich vielen Stützstellen gegeben ist, scheidet naturgemäß aus, solange keine weiteren Informationen wie eine Lipschitz-Konstante o.ä. bekannt sind.

Die Funktion f kann gemäß $g(x, p) := x - R \cdot f(x, p)$ in eine Fixpunktform umgeschrieben werden. Für ein geeignetes R entspricht das einem vereinfachten Newton-Verfahren. In Abschnitt 3 hatten wir bereits bemerkt, daß die Einschließung des Fehlers einer Näherungslösung $\widetilde{x}$ zu wesentlich schärferen Einschließungen führt. Wir definieren daher

$$g : \mathbb{R}^{n \times k} \to \mathbb{R}^n \quad \text{mit} \quad g(x, p) := x - R \cdot f(\widetilde{x} + x, p)$$

für festes $\widetilde{x} \in \mathbb{R}^n$, $R \in \mathbb{R}^{n \times n}$. Es werden keine weiteren Anforderungen an $\widetilde{x}$ oder an R gestellt. Gilt dann $g(X, \widetilde{p}) \subseteq X$ für festes $\widetilde{p} \in \mathbb{R}^k$, so besitzt g einen Fixpunkt $\widetilde{x} = \widetilde{x}(\widetilde{p})$, und es folgt für reguläres R sofort $f(\widetilde{x} + \widetilde{x}, \widetilde{p}) = 0$. Für den Nachweis von $g(X, p) \subseteq X$ wird $f(\widetilde{x} + x, \widetilde{p})$ um die Stelle $\widetilde{x}$ entwickelt. Für festes $\widetilde{p}$ sei $M_S(x)$ so definiert, daß

$$f(\widetilde{x}+x,\widetilde{p}) = f(\widetilde{x},\widetilde{p}) + M_{\widetilde{x}}(x)\cdot((\widetilde{x}+x)-\widetilde{x}) = f(\widetilde{x},\widetilde{p}) + M_{\widetilde{x}}(x)\cdot x$$

gilt. Dann ist

$$\begin{aligned} g(x,\widetilde{p}) &= x - R\cdot\{f(\widetilde{x},\widetilde{p}) + M_{\widetilde{x}}(x)\cdot x\} \\ &= -R\cdot f(\widetilde{x},\widetilde{p}) + \{I - R\cdot M_{\widetilde{x}}(x)\}\cdot x. \end{aligned} \tag{5.2}$$

Ist die Funktion f stetig differenzierbar, kann $M_{\widetilde{x}}(x)$ durch eine Jacobi-Matrix definiert werden:

$$M_{\widetilde{x}}(x)_i = \frac{\partial f_i}{\partial x}(\widetilde{x}+\xi_i(x-\widetilde{x})) \quad \text{für ein } \xi_i \in (0,1), \tag{5.3}$$

wobei $M_{\widetilde{x}}(x)_i$ die i-te Zeile der Matrix $M_{\widetilde{x}}(x)$ bezeichnet. Ist f in Form eines Programms gegeben, läßt sich die Jacobi-Matrix und damit M mittels automatischer Differentiation berechnen (vgl. [6], [27], [35]). Programmtechnisch gesehen ist das zumindest für die Vorwärts-Differentiation nahezu trivial und liefert gute Approximationen für $M_{\widetilde{x}}(x)$ ohne die üblichen Schwierigkeiten bei numerischer Differentiation. Bei der Rückwärts-Differentiation ist der Aufwand für die Berechnung des *gesamten* Gradienten einer Funktion höchstens der 5-fache gegenüber einer Auswertung der Funktion; das gilt unabhängig von n.

Um den gesamten Wertebereich $g(X,\widetilde{p})$ abschätzen zu können, benötigt man $M_{\widetilde{x}}(x)$ für alle $x \in X$. Dies kann mittels Intervalloperationen und Ausnutzung der Isotonie durch Auswertung von $M_{\widetilde{x}}(X)$ erreicht werden. Das bedeutet, daß X statt x in der zeilenweisen Auswertung der Jacobi-Matrix eingesetzt wird und damit alle $\xi_i \in (0,1)$ erfaßt werden. Liegt $\widetilde{x}$ nicht in X, so wird X durch die konvexe Hülle von $\widetilde{x}$ und X ersetzt.

Satz 5.1 Sei $f : D_n \times D_k \to \mathbb{R}^n$, $D_n \subseteq \mathbb{R}^n$, $D_k \in \mathbb{R}^k$ abgeschlossen, $\widetilde{x} \in D_n$, $R \in \mathbb{R}^{n\times n}$, $X \in \mathbb{IR}^n$ mit $X \subseteq D_n$ und $\widetilde{p} \in D_k$ gegeben. Für eine Funktion $M_{\widetilde{x}}(x) : \mathbb{R}^n \to \mathbb{IR}^{n\times n}$ sei

$$f(\widetilde{x}+x,\widetilde{p}) \in f(\widetilde{x},\widetilde{p}) + M_{\widetilde{x}}(x)\cdot x \quad \forall\, x \in X$$

und es gelte

$$-R\cdot f(\widetilde{x},\widetilde{p}) + \{I - R\cdot M_{\widetilde{x}}(X)\}\cdot X \subseteq \mathrm{int}(X). \tag{5.4}$$

Dann gibt es eine Nullstelle $\widehat{x} = \widehat{x}(\widetilde{p})$ von $f(x,\widetilde{p})$ mit $\widehat{x} \in \widetilde{x}+X$. Ist $M_{\widetilde{x}}(x)$ nach (5.3) berechnet, so ist die Nullstelle $\widehat{x}$ in $\widetilde{x}+X$ eindeutig.

Bedingung (5.4) impliziert die Lösbarkeit von $f(x,\widetilde{p}) = 0$, die Regularität von R und die Existenz und Eindeutigkeit einer Nullstelle von $f(x,\widetilde{p})$ in $\widetilde{x}+X$. Der Nachweis wird programmtechnisch gesehen *automatisch* geführt, d.h. ohne a priori Information über $f, \widetilde{x}, X$ oder R seitens des Benutzers, ohne Lipschitz-Konstanten oder ähnliches.

Satz 5.1 liefert eine algorithmisch verifizierte Vorwärtsabschätzung des Faktors. In einem Algorithmus legt man um eine gegebene Näherungslösung $\widetilde{x}$ eine Fehlerschranke X und mittels (5.4) kann verifiziert werden, ob in $\widetilde{x}+X$ genau eine Nullstelle der Funktion $f(x,\widetilde{p})$ liegt. Versagt der Test, kann sehr ähnlich dem linearen Fall iteriert werden wie in Abschnitt 3 beschrieben.

Ist die Jacobi-Matrix in $\widetilde{x}$ nicht zu schlecht konditioniert und ist $\widetilde{x}$ eine hinreichend gute Näherung, werden in der Regel nur ein oder zwei Iterationsschritte notwendig sein. In jedem Fall setzt der Einschließungsalgorithmus eine ordentliche Näherung $\widetilde{x}$ voraus. Der Punkt ist eben, daß über die Güte von $\widetilde{x}$ keine a priori-Information notwendig ist. Es sei einschränkend hinzugefügt, daß die Konvergenztheorie im nichtlinearen Fall zusätzliche Annahmen über f benötigt, da die Iterationsmatrix nicht mehr konstant ist.

Auch im nichtlinearen Fall können toleranzbehaftete Eingabedaten betrachtet werden. Ist der Parameter p nicht exakt, sondern nur innerhalb gewisser Toleranzen $P \in \mathbb{IR}^k$ bekannt, bleiben die Nullstellen der Funktionenschar $f(x,p)$, $p \in P$ unter gewissen Bedingungen zusammenhängend. Man kann dann nach der Einschließung eines gesamten Nullstellenclusters von $f(x,p)$, $p \in P$ fragen. Außerdem wird man an der wahren Ausdehnung dieses Clusters interessiert sein, d.h. nicht nur an Außen-, sondern auch an *Innen*einschließungen [31].

Satz 5.2 Sei $f : D_n \times D_k \to \mathbb{R}^n$, $D_n \subseteq \mathbb{R}^n$, $D_k \in \mathbb{R}^k$ abgeschlossen, $\widetilde{x} \in D_n$, $\widetilde{p} \in D_k$, $R \in \mathbb{R}^{n\times n}$, $X \subseteq \mathbb{IR}^n$ mit $X \subseteq D_n$ und $P \subseteq D_k$ gegeben. Für eine Funktion $M(x,p) : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{IR}^n$ sei

$$f(\widetilde{x}+y,p) \in f(\widetilde{x}+x,\widetilde{p}) + M(x,p)\cdot(y-x) \quad \forall\, x,y \in X\ \forall\, p \in P \tag{5.5}$$

und mit

$$Z := -R\cdot f(\widetilde{x},P), \quad \Delta := \{I - R\cdot M(X,P)\}\cdot X \tag{5.6}$$

gelte

$$Z + \Delta \subseteq \mathrm{int}(X). \tag{5.7}$$

Dann besitzt für jedes $p \in P$ das Gleichungssystem $f(x,p) = 0$ eine Lösung $\widehat{x} \in \widetilde{x}+X$. Diese Lösung ist in $\widetilde{x}+X$ eindeutig. Darüber hinaus gilt für die i-te Komponente des so eindeutig definierten Lösungskomplexes

$$\Sigma(f,\widetilde{x}+X,P)_i := \{\,\widehat{x}_i \mid \widehat{x} \in \widetilde{x}+X \text{ und } \exists\, p \in P : f(\widehat{x},p) = 0\,\}$$

und $Z := [\underline{Z},\overline{Z}]$, $\Delta := [\underline{\Delta},\overline{\Delta}]$ folgende Innen- und Außeneinschließung:

$$[\underline{Z}+\overline{\Delta},\overline{Z}+\underline{\Delta}]_i \subseteq \Sigma(f,\widetilde{x}+X,P)_i - \widetilde{x}_i \subseteq [\underline{Z}+\underline{\Delta},\overline{Z}+\overline{\Delta}]_i. \tag{5.8}$$

Zur rigorosen Berechnung von (5.6) und (5.7) braucht in der oben beschriebenen Berechnung von $M_{\widetilde{x}}(x)$ nur jeweils $\widetilde{p}$ durch P ersetzt zu werden und (5.5) ist wegen der Isotonie erfüllt. Die Bemerkungen zur Güte der Abschätzungen (5.8) aus Abschnitt 3 gelten sinngemäß für nichtlineare Gleichungssysteme. Auch hier ist der Unterschied zwischen Innen- und Außeneinschließung gerade Δ.

Der Vollständigkeit halber sei bemerkt, daß zur rigorosen Auswertung von (5.8) eine Inneneinschließung von Z notwendig ist. Nach den Schlußbemerkungen des ersten Abschnitts ist das im linearen Fall mühelos erzielbar; im nichtlinearen Fall ist etwas mehr Aufwand notwendig. Weiterhin sei angemerkt, daß zur Einschließung einer Nullstelle M nicht notwendig eine Art Jacobi-Matrix sein muß. Bedingung an M ist (5.5), und das kann auch durch gewisse Steigungs- oder Slope-Matrizen erreicht werden [17]. Es geht dann u.U. die Eindeutigkeit der Nullstelle in $\widetilde{x}+X$ verloren.

Abschließend geben wir noch ein Beispiel. Sei f eine Diskretisierung von $3\ddot{x}x - \dot{x}^2 = 0$ für $x(0) = 0, x(1) = 20$, also

$$f_1 = 3x_1(x_2 - 2x_1) + x_2^2/4$$
$$f_i = 3x_i(x_{i+1} - 2x_i + x_{i-1}) + (x_{i+1} - x_{i-1})^2/4 \quad \text{für } 2 \le i \le n-1$$
$$f_n = 3x_n(20 - 2x_n + x_{n-1}) + (20 - x_{n-1})^2/4.$$

Die exakte Lösung der Differentialgleichung ist $x(t) = 20 \cdot t^{3/4}$. Für

$$n = 400, \ \widetilde{x}_i \equiv 10.0 \quad \text{für } 1 \le i \le 400,$$

also einer sehr schlechten Startnäherung ergibt sich folgende Einschließung

$$\begin{aligned}
X_1 &= [\ 0.206611908273, \quad 0.206611908274] \\
X_2 &= [\ 0.360737510102, \quad 0.360737510104] \\
X_3 &= [\ 0.495574119032, \quad 0.495574119035] \\
&\ \ \vdots \\
X_{398} &= [19.9249135121276, \quad 19.9249135121282] \\
X_{399} &= [19.9624596920554, \quad 19.9624596920557] \\
X_{400} &= [19.9999823472852, \quad 19.9999823472853]
\end{aligned}$$

Die Rechnung wurde in doppelter Genauigkeit (entsprechend etwa 17 Dezimalen) ausgeführt; in allen Lösungskomponenten stimmten linke und rechte Grenzen in mindestens 11 Dezimalstellen überein. Eingeschlossen wurde die Lösung des diskretisierten, nichtlinearen Gleichungssystems, nicht des kontinuierlichen Problems.

6 Globale Optimierung

Als Anwendung der in Abschnitt 2 besprochenen Vorgehensweise zur Berechnung von Einschließungsfunktionen möchten wir hier noch kurz auf Verfahren für globale Optimierungsaufgaben, basierend auf Intervallmethoden, sowie auf einige Beispiele eingehen.

Ist für das globale Optimierungsproblem

$$Min\{\, f(x) \mid x \in X \,\},\ X \in \mathbb{IR}^n \tag{6.1}$$

eine Einschließungsfunktion F gegeben, d.h.

$$Y \in \mathbb{IR}^n,\ Y \subseteq X \Rightarrow \{\, f(y) \mid y \in Y \,\} \subseteq F(Y) := [\underline{F}(Y), \overline{F}(Y)], \tag{6.2}$$

so können Branch-and-Bound-Strategien zur Berechnung verifizierter Schranken des globalen Optimalwertes $f^* := Min\{\, f(x) \mid x \in X \,\}$ und der globalen Optimalpunkte $X^* := \{\, x^* \in X \mid f(x^*) = f^* \,\}$ angewandt werden. In [7], [8], [9], [20], [29] sind Verfahren dieses Typs unter Benutzung von Intervallmethoden angegeben; in [9] findet man zahlreiche Beispiele. Zu weiteren Branch-and-Bound-Methoden sei insbesondere auf [10], [26] verwiesen.

Nach unserer Erfahrung erfordert die Berechnung verifizierter Schranken für X^* in vielen Fällen die Benutzung von Einschließungsfunktionen für den Gradienten und die Hesse-Matrix von f und ist daher häufig recht aufwendig. Deshalb wurde ein Verfahren entwickelt, das verifizierte Schranken $\underline{F}^*$ und $\overline{F}^*$ für den Optimalwert f^* berechnet mit

$$\underline{F}^* \le f^* \le \overline{F}^*. \tag{6.3}$$

Darüber hinaus liefert dieses Verfahren eine Approximation $\widetilde{x} \in X$ mit

$$\underline{F}^* \le f(\widetilde{x}) \le \overline{F}^*. \tag{6.4}$$

Das Verfahren arbeitet *ableitungsfrei*. Es macht sich die Vorteile lokaler Optimierungsverfahren und intervallmäßiger Auswertungen wechselseitig in einer Branch-and-Bound-Strategie zu Nutze:

a) Lokale Optimierungsverfahren benötigen einen hinreichend guten Startpunkt; daher wird mit intervallmäßiger Auswertung der Startpunkt verbessert.

b) Ist für eine Box $Y \subseteq X$ der intervallmäßig abgeschätzte Wertebereich $F(Y) = [\underline{F}(Y), \overline{F}(Y)]$ und ist $\underline{F}(Y)$ größer als ein bereits berechneter Funktionswert $f(\widetilde{x}), \widetilde{x} \in X$, so kann Y keinen globalen Optimalpunkt mehr enthalten und kann somit ausgeschlossen werden.

Der Ausschluß von Boxen mittels Punkt b) geht um so schneller, je besser $f(\widetilde{x})$ den globalen Optimalwert f^* approximiert, was wiederum durch Punkt a) begünstigt wird. Der wechselseitige Einsatz von Gleitkomma- und Intervallrechnung erzielt beachtliche Geschwindigkeitserfolge. Diese Strategie ist wieder ganz im Sinne der in Abschnitt 2 besprochenen Regeln.

Auf eine genauere Darstellung des Verfahrens soll im Rahmen dieser Arbeit verzichtet werden, da eine Vielzahl von Details zu beachten sind. Der eindimensionale Fall wurde in [12] behandelt; eine erste mehrdimensionale Version mit ca. 50 aus der Literatur bekannten Beispielen findet man in [14]. Eine detaillierte Darstellung des mehrdimensionalen Falls mit Konvergenzbetrachtung wird in [13] gegeben.

Die Forschung steht hier ebenfalls noch am Anfang. Wir geben im folgenden numerische Resultate für einige Testbeispiele an und beginnen mit einem Satz von Testfunktionen, der von Dixon und Szegö [4] für den Vergleich globaler Optimierungsverfahren vorgeschlagen wurde. Für rechnerunabhängige Vergleiche geben wir die Rechenzeit in Einheiten der Standardzeit STU an [Standard Unit Time, 1 Einheit entspricht 1000 Aufrufen der reellen Shekelfunktion S5 in (4,4,4,4)]. Eine Einheit entspricht auf einer SUN-4 etwa 0.2 Sekunden.

Wir vergleichen mit den Zeiten für eine Reihe bekannter Algorithmen. Diese Zeiten (der obere Teil der Tabelle) wurden [3] entnommen.

Methode	GP	BR	H3	H6	S5	S7	S10
Törn	4	4	8	16	10	13	15
De Biase	15	14	16	21	23	20	30
Price	3	4	8	46	14	20	20
Branin	-	-	-	-	9	8.5	9.5
Boender et al.	1.5	1	1.7	4.3	3.5	4.5	7
unser Verfahren	0.45	0.45	5.65	6.45	0.70	0.80	0.90
$f^*, \overline{F}^*$	3	0.397887	-3.86278	-3.32237	-10.1532	-10.4049	-10.5364
$\underline{F}^*$	3	0.397887	-4.34853	-4.17324	-10.2008	-10.6772	-10.8517

Tabelle 6.1. Standardzeiten und Resultate mit Ergebnisverifikation

In den letzten beiden Zeilen ist der globale Optimalwert f^* angegeben sowie die untere Schranke $\underline{F}^*$. *Die berechnete Approximation $f(\widetilde{x})$ stimmt in allen Fällen auf mindestens 6 Stellen mit dem Optimalwert f^* und der verifizierten oberen Schranke $\overline{F}^*$ überein.*

Durch Erhöhung des Aufwandes läßt sich die Genauigkeit der berechneten Schranken verbessern. Eine genauere Analyse des Verfahrens zeigt die monotone Konvergenz der unteren und oberen Schranken $\underline{F}^*$, $\overline{F}^*$ gegen f^*.

Im Anhang ist der Plot von fünf Funktionen dargestellt. Bei den ersten drei Beispielen handelt es sich um die Maximierung des kleinsten Singulärwertes einer parametrisierten Matrix $M(x)$, also die Maximierung des $\|\cdot\|_2$-Abstandes zur nächsten singulären Matrix:

$$Max\, f(x), \quad f(x) := \sigma_{\min}(M(x)).$$

Das vierte Beispiel ist eine bekannte Funktion von Levy, und im fünften Beispiel wird der maximale Realteil der Eigenwerte einer parametrisierten Matrix $M(x)$ minimiert. Die Einschließungsfunktionen für die Beispiele 1, 2, 3, 5 sind mit Hilfe einer leicht modifizierten Variante des Verfahrens von Lohner [19] zur Einschließung von Eigenwerten berechnet worden. Im vierten Beispiel wurden lediglich die auftretenden Variablen und Operationen durch Intervallvariablen und Intervalloperationen ersetzt.

In allen Bildern ist der Plot von $-f(x)$ gezeigt. Das erste Beispiel ist durch die parametrisierte Matrix

$$M(x) = \begin{pmatrix} 2\sin \pi x_1 & \sin \pi x_1 & \sin \pi x_2 & \sin \pi x_1 x_2 & \cos \pi x_1 x_2 \\ \sin \pi x_1 & 2\sin 4\pi x_2 & \cos \pi(1-x_1) & \cos \pi(1-x_2) & \cos \pi x_1 \\ \sin \pi x_2 & \cos \pi(1-x_1) & 2\cos 5\pi x_1 x_2 & \cos \pi x_1 & \cos \pi x_2 \\ \sin \pi x_1 x_2 & \cos \pi(1-x_2) & \cos \pi x_1 & 2\sin \pi x_2 & \sin \pi(1-x_1) \\ \cos \pi x_1 x_2 & \cos \pi x_1 & \cos \pi x_2 & \sin \pi(1-x_1) & 2\sin 4\pi x_1 \end{pmatrix}$$

$$0 \le x_i \le 1, \quad \text{für } i = 1, 2.$$

definiert. Das globale Optimum befindet sich in diesem Beispiel auf einer nicht differenzierbaren Kante der Funktion. Im folgenden geben wir neben der Rechenzeit in STU und Schranken für f^* jeweils die Anzahl der reellen Funktionsauswertungen NRF von f und die Anzahl der intervallmäßigen Auswertungen NIF der Einschließungsfunktion an. Es ergibt sich

NRF	NIF	STU	$\underline{F}^*$	$f^*, \overline{F}^*$
207	2687	278	−2.00159	−1.67555
216	3891	397	−1.71291	−1.67555

Im zweiten Beispiel ist

$$M(x) = \begin{pmatrix} d_1(x_1,x_2) & k\sin x_1 & k\sin x_2 & k\cos x_2 & k\cos x_2 \\ k\sin x_1 & d_2(x_1,x_2) & kx_1 & kx_2 & kx_1x_2 \\ k\sin x_2 & kx_1 & p(10,-10) & kx_1^2 & kx_2^2 \\ k\cos x_1 & kx_2 & kx_1^2 & -5 & k\sin x_1x_2 \\ k\cos x_2 & kx_1x_2 & kx_2^2 & k\sin x_1x_2 & p(10,-10) \end{pmatrix}$$

$$\begin{aligned} \text{mit} \quad d_1(x_1,x_2) &= 2e^{-500p(-4,-4)} + 2.5 + \frac{x_1+x_2}{20} \\ d_2(x_1,x_2) &= \frac{x_1^2+x_2^2}{7} + (x_2+5)\cos\frac{\pi}{2}(x_1^2+x_2^2) \\ p(a,b) &= (x_1-a)^2 + (x_2-b)^2 \\ k &= 1.0\cdot 10^{-3} \end{aligned}$$

$$-5 \le x_i \le 5, \quad i = 1, 2$$

NRF	NIF	STU	$\underline{F}^*$	$f^*, \overline{F}^*$
135	33	8.9	−4.10190	−4.10000
136	49	11.2	−4.10010	−4.10000

Das dritte Beispiel ist gleich dem zweiten bis auf die letzten drei Diagonalelemente von $M(x)$: $M(x)_{33} = 6\cos 2\pi x_1$, $M(x)_{44} = d_3(x_1,x_2)$, $M(x)_{55} = 6\cos 2\pi x_2$ mit $d_3(x_1,x_2) = (x^4+y^4)/128 + 2 + 0.5\cos 6\pi x_1 x_2$.

Alle anderen Daten sind gleich. Man beachte im Plot des zweiten und dritten Beispiels die kleine Spitze des globalen Optimums am linken Rand. Es ergibt sich für das dritte Beispiel

NRF	NIF	STU	$\underline{F}^*$	$f^*, \overline{F}^*$
135	33	6.4	−4.10245	−4.10000
136	49	11.9	−4.10010	−4.10000

Das vierte Beispiel ist die von Levy angegebene Funktion

$$\begin{aligned} f(x) \;=\; & \sum_{i=1}^{5} i\cos\big((i+1)x_1+i\big)\sum_{j=1}^{5} j\cos\big((j+1)x_2+j\big) \\ & +(x_1+1.42513)^2+(x_2+0.80032)^2 \end{aligned}$$

$$-10 \le x_i \le 10, \quad i = 1, 2$$

NRF	NIF	STU	$\underline{F}^*$	$f^*, \overline{F}^*$
199	664	5.1	−189.796	−176.138
202	1236	13.6	−177.200	−176.138

Das fünfte Beispiel ist ein MinMax-Problem zur Auffindung einer Matrix, deren maximaler Realteil der Eigenwerte minimiert wird, also eine Maximierung im Sinne der Stabilität

$$f(x) := \min_{x\in X} \max_{1\le i\le m} \Re\{\lambda_i(M(x))\}.$$

Dabei bezeichnet $\Re\{\lambda_i(M(x))\}$ den Realteil der Eigenwerte, wobei

$$M(x) = \begin{pmatrix} d_1(x_1,x_2) & k\sin x_1 & k\sin x_2 & k\cos x_1 & k\cos x_2 \\ k\sin 2x_1 & d_2(x_1,x_2) & kx_1 & kx_2 & kx_1x_2 \\ k\sin 2x_2 & k(x_1+x_2) & d_3(x_1,x_2) & kx_1^2 & kx_2^2 \\ k\cos 2x_1 & k(x_1-x_2) & k(x_1+x_2)^2 & d_4(x_1,x_2) & k\sin x_1x_2 \\ k\cos 2x_2 & kx_1x_2^2 & 4kx_2^2 & k\sin(x_1+x_2) & d_5(x_1,x_2) \end{pmatrix}$$

$$\begin{aligned} \text{mit} \quad d_1(x_1,x_2) &= 17.5 - 2e^{-500\left((x_1+4)^2+(x_2+4)^2\right)} - \frac{x_1+x_2}{20} \\ d_2(x_1,x_2) &= 20 - \frac{x_1^2+x_2^2}{7} - (x_2+5)\cos\frac{\pi}{2}(x_1^2+x_2^2) \\ d_3(x_1,x_2) &= 20 - 6\cos 2\pi x_1 \\ d_4(x_1,x_2) &= 18 - \frac{x_1^4+x_2^4}{128} + \frac{1}{2}\cos 6\pi x_1 x_2 \\ d_5(x_1,x_2) &= 20 - 6\cos 2\pi x_2 \\ k &= 10^{-3} \\ x_1, x_2 &\in [-5,5] \end{aligned}$$

In den Bildern ist wieder $-f(x)$ gezeichnet. Das erste Bild zeigt f über X, das zweite das Detail aus X mit globalem Optimum. Es ergibt sich für obiges Beispiel:

NRF	NIF	STU	$\underline{F}^*$	$f^*, \overline{F}^*$
133	33	21.5	15.8974	15.9000
134	49	28.5	15.8999	15.9000

Als letztes, höherdimensionales Beispiel schließlich betrachten wir die Griewank-Funktion

$$f_G(x) := \sum_{i=1}^{n} \frac{x_i^2}{d} - \prod_{i=1}^{n} \cos \frac{x_i}{\sqrt{i}} + 1$$

mit $X = [-600,\ 600]^n$, $d = 4000$ und als Optimum $f^* = 0$. Im zweidimensionalen sieht die Funktion aus wie ein leicht nach oben gewölbter Eierkarton. Im angegebenen Bereich X befinden sich etwa 1000 lokale Minima. Die uns bekannten Resultate (vgl. [36]) sind:

	NRF	STU
Griewank 1981	6600*	-
Snyman, Fatti (1987)	23399	90

* globales Minimum nicht gefunden

Tabelle 6.2. Bekannte Resultate für Griewank-Funktion ($n = 10$)

n	NRF	NIF	STU	$\underline{F}^*$	$\overline{F}^*$
10	417	421	4.3	0	$1.31 \cdot 10^{-14}$
50	743	1601	48.1	0	$2.25 \cdot 10^{-14}$

Tabelle 6.3. Resultate Griewank-Funktion für unser Verfahren mit Ergebnisverifikation

Literaturverzeichnis

[1] G. Alefeld and J. Herzberger. *Introduction to Interval Computations*. Academic Press, New York, 1983.

[2] H. Bauch, K.-U. Jahn, D. Oelschlägel, H. Süsse, and V. Wiebigke. *Intervallmathematik, Theorie und Anwendungen*, volume Bd. 72 of *Mathematisch-naturwissenschaftliche Bibliothek*. B.G. Teubner, Leipzig, 1987.

[3] C. Boender, A.R. Kan, G. Timmer, and L. Stongie. A Stocastic Method for Global Optimization. *Mathematical Programming 22*, pages 125–140, 1982.

[4] L.C.W. Dixon and G.P. Szegö (eds.). *Towards Global Optimization*. North-Holland, Amsterdam, 1975.

[5] R.T. Gregory and D.L. Karney. *A Collection of Matrices for Testing Computional Algorithms*. John Wiley & Sons, New York, 1969.

[6] A. Griewank. On Automatic Differentiation. In *Mathematical Programming 88*. Kluwer Academic Publishers, Boston, 1989.

[7] E.R. Hansen. Global Optimization Using Interval Analysis — the One Dimensional Case. *J. Optim. Theory Appl. 29*, pages 331–344, 1979.

[8] E.R. Hansen. Global Optimization Using Interval Analysis — the Multidimensional Case. *Numer. Math. 34*, pages 247–270, 1980.

[9] E.R. Hansen. *Global Optimization using Interval Analysis*. Marcel Dekker, New York, Basel, Hong Kong, 1992.

[10] R. Horst and H. Tuy. *Global Optimization*. Springer-Verlag, Berlin, 1990.

[11] C. Jansson. Interval Linear Systems with Symmetric Matrices, Skew-Symmetric Matrices, and Dependencies in the Right Hand Side. *Computing 46*, pages 265–274, 1991.

[12] C. Jansson. A Global Optimization Method Using Interval Arithmetic. In L. Atanassova and J. Herzberger, editors, *Computer Arithmetic and Enclosure Methods*, IMACS, pages 259–267. Elsevier Science Publishers B.V., 1992.

[13] C. Jansson and O. Knüppel. A Branch-and-Bound Algorithm for Bound Constrained Optimization Problems without Derivatives. zur Veröffentlichung eingereicht.

[14] C. Jansson and O. Knüppel. A Global Minimization Method: The Multi-dimensional case. Technical Report 92.1, Forschungsschwerpunkt Informations- und Kommunikationstechnik, TU Hamburg-Harburg, 1992.

[15] W.M. Kahan. A More Complete Interval Arithmetic. *Lecture notes for a summer course at the University of Michigan*, 1968.

[16] R. Krawczyk. Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken. *Computing 4*, pages 187–201, 1969.

[17] R. Krawczyk and A. Neumaier. Interval Slopes for Rational Functions and Associated Centered Forms. *SIAM J. Numer. Anal.*, 22(3):604–616, 1985.

[18] U. Kulisch and W.L. Miranker. *Computer Arithmetic in Theory and Practice*. Academic Press, New York, 1981.

[19] R. Lohner. Enclosing all Eigenvalues for Symmetric Matrices. In Ch. Ullrich and J. Wolff von Gudenberg, editors, *Accurate Numerical Algorithms*, New York, 1989.

[20] R.E. Moore. On Computing the Range of Values of a Rational Function of n Variables over a Bounded Region. *Computing 16*, pages 1–15, 1976.

[21] R.E. Moore. A Test for Existence of Solutions for Non-Linear Systems. *SIAM J. Numer. Anal. 4*, 1977.

[22] R.E. Moore. *Methods and Applications of Interval Analysis*. SIAM, Philadelphia, 1979.

[23] A. Neumaier. Rigorous sensitivity analysis for parameter-dependent systems of equations. *J. Math. Anal. Appl. 144*, pages 16–25, 1989.

[24] A. Neumaier. *Interval Methods for Systems of Equations, Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 1990.

[25] J.v. Neumann and H.H. Goldstine. Numerical Inverting of Matrices of High Order. *Bull. Amer. Math. Soc. 53*, pages 1021–1099, 1947.

[26] P.M. Pardalos and J.B. Rosen. Constrained Global Optimization: Algorithms and Applications. *Springer Lecture Notes Comp. Sci. 268, Berlin*, 1987.

[27] L.B. Rall. Automatic Differentiation: Techniques and Applications. In *Lecture Notes in Computer Science 120*. Springer Verlag, Berlin-Heidelberg-New York, 1981.

[28] H. Ratschek and J. Rokne. *Computer Methods for the Range of Functions*. Halsted Press (Ellis Horwood Limited), New York (Chichester), 1984.

[29] H. Ratschek and J. Rokne. *New Computer Methods for Global Optimization*. John Wiley & Sons (Ellis Horwood Limited), New York (Chichester), 1988.

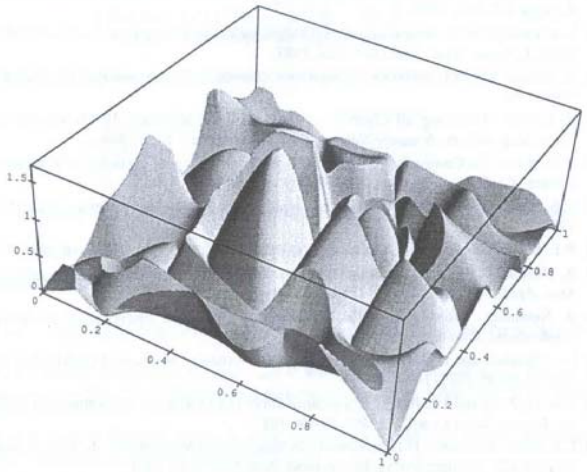
[30] S.M. Rump. Solving Algebraic Problems with High Accuracy. In U.W. Kulisch and W.L. Miranker, editors, *A New Approach to Scientific Computation*, pages 51–120. Academic Press, New York, 1983.

[31] S.M. Rump. Rigorous Sensitivity Analysis for Systems of Linear and Nonlinear Equations. *Math. of Comp.*, 54(10):721–736, 1990.

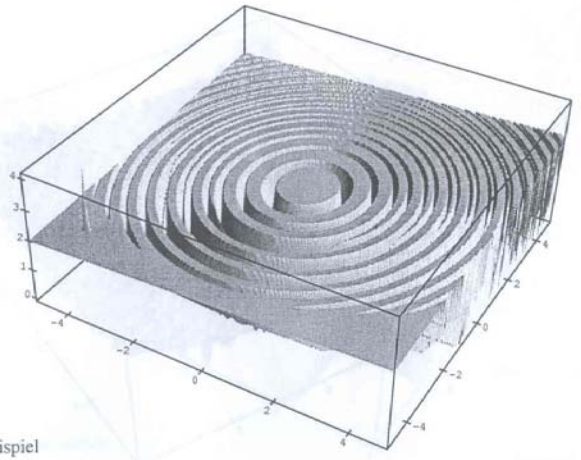
[32] S.M. Rump. On the Solution of Interval Linear Systems. *Computing 47*, pages 337–353, 1992.

[33] S.M. Rump. Validated Solution of Large Linear Systems. *Computing Supplementum*, to appear.
[34] H. Schwandt. An Interval Arithmetic Approach for the Construction of an almost Globally Convergent Method for the Solution of the Nonlinear Poisson Equation on the Unit Square. *SIAM J. Sci. Stat. Comp.*, 5(2), 1984.
[35] B. Speelpennig. *Compiling Fast Partial Derivatives of Functions given by Algorithms*. Urbana, Illinois, 1980.
[36] A. Törn and A. Žilinskas. *Global Optimization*. Springer-Verlag, Berlin, Heidelberg, New York, 1989.
[37] J.H. Wilkinson. *Rounding Errors in Algebraic Processes*. Prentice-Hall Inc., 1963.
[38] J.H. Wilkinson. *The Algebraic Eigenvalue Problem*. Oxford Univerity Press, Oxford, 1969.
[39] J.H. Wilkinson. Modern Error Analysis. *SIAM Rev. 13*, pages 548–568, 1971.

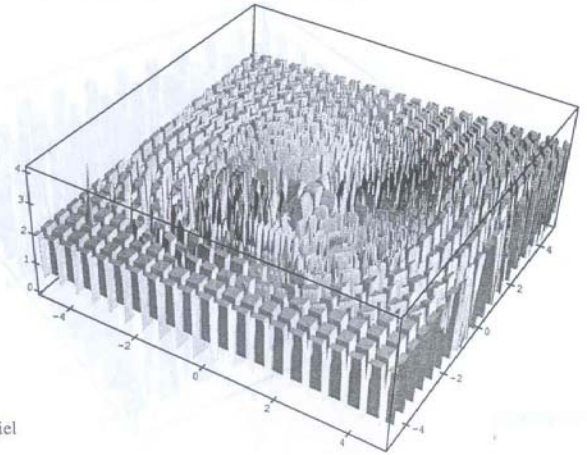
Anhang



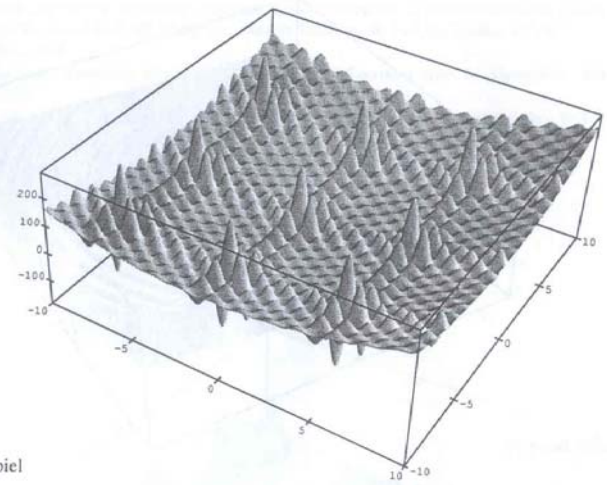
Erstes Beispiel



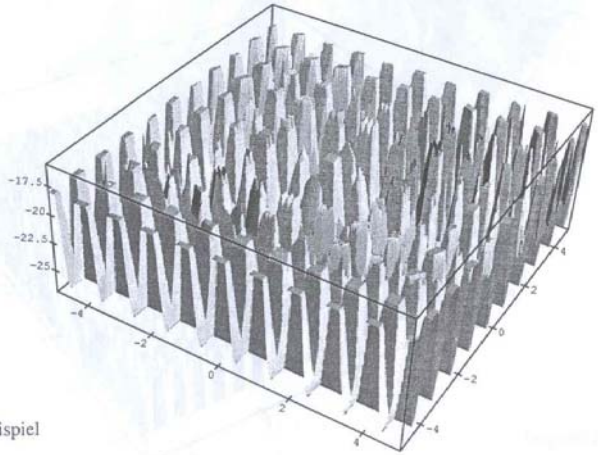
Zweites Beispiel



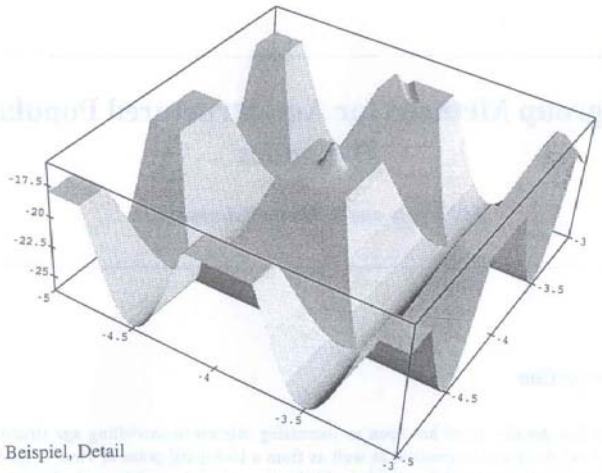
Drittes Beispiel



Viertes Beispiel



Fünftes Beispiel



Fünftes Beispiel, Detail