Technische Hochschule Leipzig
Wissenschaftliche Zeitschrift
Jahrgang 15/1991 · Heft 6 · Seite 427–431

S. M. Rump

Convergence Properties of Iterations Using Sets

Abstract
In the literature efficient algorithms have been described for calculating guaranteed inclusions for the solution of a number of standard numerical problems [3, 4, 7, 10, 11, 12]. The inclusions are given by means of a set containing the solution. In [11, 12] this set is calculated using an affine iteration which stops after a nonempty and compact set has been mapped into itself.
In this paper different types of such sets are investigated, namely general sets, hyperrectangles and standard simplices. For affine iterations using those types of sets global convergence properties are given. Here, global convergence means that the iteration stops for every starting set with a set being mapped into itself.

0. Introduction

Let T denote one of the sets $\mathbb{R}$, $\mathbb{C}$, $\mathbb{R}^n$ (real vectors with n components), $\mathbb{C}^n$ (complex vectors with n components), $R^{n\times n}$ (real square matrices with n rows and columns) or $\mathbb{C}^{n\times n}$ (complex square matrices with n rows and columns). Throughout this paper the letter "n" is reserved in the prescribed way; only square matrices (which are $n \times n$) will occur. $\mathbb{P}T$ denotes the power set over T.
In the following $* \in\{+, -, \cdot, /\}$ denotes the binary real or complex operations, resp. These operations extend to power set operations in the usual way. If $x * y \in T_3$ is defined for $x \in X \in \mathbb{P}T_1$, $y \in Y \in \mathbb{P}T_2$, then

$$X * Y := \{x * y \mid x \in X, y \in Y\} \in \mathbb{P}T_3 .$$

The set of all n-dimensional resp. n^2-dimensional hyperrectangles parallel to the axis over real or complex numbers is denoted by $\mathbb{IRR}^n$, $\mathbb{IC}^n$, $\mathbb{IR}^{n\times n}$ or $\mathbb{IC}^{n\times n}$, resp. This is one way to represent interval vectors and interval matrices.
Intervals are always supposed to be nonempty.
The rounding of an arbitrary set X into the smallest hyperrectangle containing X is denoted by $\diamondsuit: \mathbb{P}T \to \mathbb{I}T$

$$X \in \mathbb{P}T \Rightarrow \diamondsuit(X) := \bigcap \{Y \in \mathbb{I}T \mid X \subseteq Y\} \in \mathbb{I}T .$$

The set $\diamondsuit(X)$ is well-defined. We define operations $\diamondplus$, $\diamondminus$, $\diamondcdot$, $\diamondslash$ over $\mathbb{I}T$ by

$$[X], [Y] \in \mathbb{I}T \Rightarrow [X] \diamond [Y] := \diamondsuit([X] * [Y]) \quad \text{for} \quad * \in\{+, -, \cdot, /\} .$$

This is the smallest hyperrectangle containing the result of the power set operation between the operands. It is uniquely defined and effectively computable (cf. [2, 8, 9, 10]).
For a set X, $Y \subseteq T$, $\operatorname{int}(X)$ denotes the interior of X, $Re(X)$ denotes the real part, $Im(X)$ the imaginary part of X. For a real matrix A we define $|A|$ to be the matrix of absolute values of the components of A, for a complex matrix is $|Re(A)| + |Im(A)|$ (cf. [2]). For an interval $[X] \in \mathbb{I}S$ we define $|[X]| = \max\{|x| \mid x \in [X]\}$ extending componentwise to interval vectors and

matrices. For $A \in S^{n\times n}$, $S \in \{\mathbf{R}, \mathbf{C}\}$ the spectral radius of A is denoted by $\rho(A)$, for $[A] \in \mathbf{I}S^{n\times n}$ we define $\varrho([A]) := \max\{\varrho(A) \mid A \in [A]\}$. A_i denotes the ith row of A.

1. Iterations using sets

Let $S \in \{\mathbf{R}, \mathbf{C}\}$ and $f: S^n \to S^n$ with $f \in C^1$. Furthermore, let $\tilde{x} \in S^n$, $R \in S^{n\times n}$ and for $X \in \mathbf{P}S^n$ define

$$f': \mathbf{P}S^n \to \mathbf{P}S^{n\times n} \quad \text{by} \quad (f'(X))_i := \bigcup \left\{ \frac{\partial f_i}{\partial x}(\zeta) \mid \zeta \in X \right\}$$

and

$$K(X) := \tilde{x} - R \cdot f(\tilde{x}) + \{I - R \cdot f'(\tilde{x} \cup X)\} \cdot (X - \tilde{x}). \tag{1.1}$$

$K(X)$ denotes the Krawczyk-operator [7]. Then it can be shown [11, 12, 15] that for non-empty, compact and convex X

$$K(X) \subseteq \operatorname{int}(X) \quad \text{implies} \quad \exists^{1-1} \hat{x} \in X: f(\hat{x}) = 0. \tag{1.2}$$

Note that there are no a priori assumptions on $\tilde{x}$ or R. For a practical application we use some $\tilde{x}$, R with $f(\tilde{x}) \approx 0$ and $R \approx f'(\tilde{x})^{-1}$.
The problem of finding a suitable X satisfying (1.2) can be attacked by means of an iteration. By simplifying (1.1) we obtain

$$K(X) = Z + A \cdot X \quad \text{where} \quad A = \{I - R \cdot f'(\tilde{x} \cup X)\} \in \mathbf{P}S^{n\times n} \quad \text{and}$$
$$Z = \tilde{x} - R \cdot f(\tilde{x}) - A \cdot \tilde{x} \in \mathbf{P}S^n.$$

Then, a first approach is the iteration scheme

$$X^{k+1} := Z + A \cdot X^k \tag{1.3}$$

for given $X^0 \in \mathbf{P}S^n$.
Our aim is to investigate the global convergence of such an iteration, i.e. to guarantee $X^{k+1} \subseteq \operatorname{int}(X^k)$ for some $k \in \mathbf{N}$ für every $\emptyset \neq X^0 \in \mathbf{P}S^n$. However, (1.3) is not suitable to achieve global convergence because of two reasons. First, the interior of X^0 must be non-empty because $\operatorname{int}(X^0) = \emptyset$ implies $\operatorname{int}(X^k) = \emptyset$ for every $k \in \mathbf{N}$. Second, (1.3) implies $\hat{x} \in X^0$. In other words only those sets X^0 already containing $\hat{x}$ are suitable in order to achieve (1.3). For practical applications this is hardly acceptable.
To overcome those three difficulties in [11] the so-called ε-inflation has been introduced. One possible definition for general sets is the following.
Definition 1. For a set $X \subseteq S^n$, $S \in \{\mathbf{R}, \mathbf{C}\}$ the ε-inflation $X \circ \varepsilon$ is definied by

$$X \circ \varepsilon := X + U_\varepsilon(0) \quad \text{for} \quad 0 < \varepsilon \in \mathbf{R},$$

where $U_\varepsilon(0)$ is some closed and bounded set containing the origin as an interior point. Obviously $X \subseteq \operatorname{int}(X \circ \varepsilon)$. An example for $U_\varepsilon(0)$ is the closed ball of radius ε around the origin. Using the ε-inflation we can define an iteration scheme allowing a complete analysis of the conditions under which the iteration stops for some $k \in \mathbf{N}$.

Theorem 2. Let $A \in S^{n\times n}$ be an arbitrary matrix, $\emptyset \neq Z \subseteq S^n$ be a compact set of vectors, $S \in \{\mathbf{R}, \mathbf{C}\}$. Let

$$X^{k+1} := (Z + A \cdot X^k) \circ \varepsilon_k \quad \text{for} \quad 0 \leq k \in \mathbf{N}, \tag{1.4}$$

where $U_{\varepsilon_{k+1}} \subseteq U_{\varepsilon_k}$ and $U \subseteq U_{\varepsilon_k}$ for every $k \in \mathbf{N}$ and some compact $\emptyset \neq U \subseteq S^n$ with $0 \in \operatorname{int}(U)$. Then the following two conditions are equivalent:

a) $\forall\, \emptyset \neq X^0 \in S^n$ compact $\exists\, k \in \mathbf{N}$: $Z + A \cdot X^k \subseteq \operatorname{int}(X^k)$

b) $\varrho(A) < 1$.

Proof. See [12].

Note that in theorem 2 the matrix A is a single matrix, not a set of matrices. In practical applications the diameter of A is small. However, a direct generalization of theorem 2 to $A \in \mathbb{P}S^{n\times n}$ by replacing part b) by

$$\varrho(a) < 1 \quad \text{for all} \quad a \in A$$

is not true. By the definition of the power set operations part a) implies $\varrho\left(\prod_{\nu=1}^{m} A_\nu\right) < 1$ for all $A_\nu \in A$. Consider (see [12])

$$a_1 := \begin{pmatrix} 0.5 & 0.27 \\ 0.92 & 0.5 \end{pmatrix} \quad \text{and} \quad a_2 := \begin{pmatrix} 0.17 & 0.6 \\ 0.94 & 0.25 \end{pmatrix}$$

and $A := \{a = a_1 + \sigma(a_2 - a_1) \mid 0 \le \sigma \le 1\}$. Then $\varrho(a_1) < 0.998\,5$, $\varrho(a_2) < 0.962\,2$ and $\max_{0\le\sigma\le 1} \varrho(a_1 + \sigma(a_2 - a_1)) < 0.999\,618 < 1$. Hence $\varrho(a) < 1$ for all $a \in A$, but $\varrho(a_1 \cdot a_2) > 1.016\,5 > 1$. It is an open problem to find a general criterion for global convergence of (1.4) for $A \in \mathbb{P}S^{n\times n}$.

2. Interval iterations

Interval vectors and interval matrices offer the great advantage that operations between those are very easily and effectively computable [2, 9]. This advantage has to be paid by some overestimation which can be severe (data dependencies, "wrapping effect"). However, the overestimation in (1.4) is diminished because those terms where intervals occur are kept small.
In case of hyperrectangles an ε-inflation should consist of an absolute and a relative part in order to maintain (1.4) for a small value of k. A possible definition which turned out to be very suitable for practical applications is

$$[X] \in \mathbb{I}S: \quad [X] \circ \varepsilon := [J] \diamondsuit [X] \oplus [E]$$

with a diagonal matrix $[J] \in \mathbb{I}S^{n\times n}$ and $[E] \in \mathbb{I}S^n$ and $1 \in [J_{ii}]$, $0 \in \operatorname{int}([E_i])$ for $1 \le i \le n$. In the following we state a theorem similar to theorem 4 for intervals (hyperrectangles) and the corresponding operations $\diamondsuit$. In practical applications it turned out to be useful to adapt E to the iteration process. Therefore we use a more general definition of the ε-inflation in the following theorem.

Theorem 3. Let $[A] \in \mathbb{I}S^{n\times n}$ be an interval matrix, $[Z] \in \mathbb{I}S^n$ be an interval vector, $S \in \{\mathbb{R}, \mathbb{C}\}$. For

$$f: \mathbb{I}S^n \to \mathbb{I}S^n \text{ with } [Y] \in \mathbb{I}S^n: \quad f([Y]) := [Z] \oplus [A] \diamondsuit [Y]$$

and $[X^0] \in \mathbb{I}S^n$ we define the iteration

$$[X]^{k+1} := J \diamondsuit f([X^k]) \oplus E^k \tag{2.1}$$

with diagonal matrix $[J] \in \mathbb{I}S^{n\times n}$, $[E^k] \in \mathbb{I}S^n$, for $0 \le k \in \mathbb{N}$. Let $[E^k] \to [E] \in \mathbb{I}S^n$, $0 \in \operatorname{int}([E])$, $1 \in [J_{ii}]$ for $1 \le i \le n$ and $\varrho(|[J]| \cdot |[A]|) < 1$. Then the following two conditions are equivalent:

a) $\forall [X^0] \in \mathbb{I}S^n \quad \exists k \in \mathbb{N}: \quad f([X^k]) \subseteq \operatorname{int}([X^k])$

b) $\varrho(|[A]|) < 1$.

Proof. See [16].

The simplicity of interval operations has to be paid by the special structure, especially the midpoint-symmetry of the hyperrectangles implying the contractivity of the absolute value of the iteration matrix. Nevertheless theorem 3 gives a necessary and sufficient criterion for global convergence.

The assumption $\varrho(|[J]| \cdot |[A]|) < 1$ in Theorem 3 is necessary. Consider

$$[A] := \begin{pmatrix} 0 & 2 \\ 1/8 & 0 \end{pmatrix}, \qquad Z := 0, \qquad [X^0] = \begin{pmatrix} [-1,1] \\ [-1,1] \end{pmatrix},$$

$$[J] := \begin{pmatrix} [-4,4] & 0 \\ 0 & [-4,4] \end{pmatrix} \quad \text{and}$$

$$[E^k] = [E] := \begin{pmatrix} [-1/4,\ 1/4] \\ [-1/4,\ 1/4] \end{pmatrix} \quad \text{for} \quad k \in \mathbb{N}.$$

It can be shown that $X^{k+1} \subseteq \operatorname{int}(X^k)$ is not true for any $k \in \mathbb{N}$. In this counterexample $\varrho(|[J]| \cdot |[A]|) = 2$. The matrix A is not primitive, and this seems to be crucial to construct a counterexample.

3. Standard simplices

The special structure of hyperrectangles requires $|A|$ or $|Re(A)| + |Im(A)|$ to be convergent in order to allow $f(X^k) \subseteq \operatorname{int}(X^k)$ for some $k \in \mathbb{N}$ (see theorem 3). This is a necessary and sufficient condition. For general sets or general simplices $f(X^k) \subseteq \operatorname{int}(X^k)$ is equivalent to $\varrho(A) < 1$, $A \in S^{n \times n}$, $S \in \{\mathbb{R}, \mathbb{C}\}$. One might try to use other representations of sets to omit the assumption $\varrho(|A|) < 1$ or $\varrho(|Re(A)| + |Im(A)|) < 1$. The representation should be simple enough to allow fast computation of $f(X^k)$ but "general" enough to cover as many matrices as possible.
One such representation are standard simplices:

$$S = \{s_0, \sigma_1, \ldots, \sigma_n\}$$
$$= \left\{ x \in \mathbb{R}^n \mid x = s_0 + \sum_{\nu=0}^{n} \lambda_\nu \sigma_\nu e_\nu, \qquad 0 \leq \lambda_\nu \in \mathbf{R}, \qquad \sum_{\nu=1}^{n} \lambda_\nu \leq 1 \right\}$$

Standard simplices still allow inexpensive set operations. A problem, however, is the rounding. For example, the product of a matrix and a standard simplex yields a general simplex where its rounding into the "smallest" standard simplex is not unique.
With respect to our problem of stating criterions for the iteration (1.4) to converge standard simplices play a special role. On the one hand, $\varrho(|A|) < 1$ is not necessary, on the other hand it is not sufficient either. Consider

$$A = \begin{pmatrix} 0.9 & -0.05 \\ -0.9 & -0.8 \end{pmatrix}. \tag{3.1}$$

The eigenvalues of A are $0.05 \pm \sqrt{0.7675}$, those of $|A|$ are $0.85 \pm \sqrt{0.0475}$ implying $\varrho(A) < 1 < \varrho(|A|)$. However, short computation yields that the standard simplex $S = \{(-2.7, -1.15), 4.9, 5\}$ satisfies $A \cdot S \subseteq \operatorname{int}(S)$.
On th other hand consider

$$A = \begin{pmatrix} 0.5 & -0.5 \\ 0{:}25 & 0.5 \end{pmatrix}. \tag{3.2}$$

The eigenvalues of $|A|$ are $0.5 \pm \sqrt{0.125}$ implying $\varrho(A) \leq \varrho(|A|) < 1$. It can be shown that for every standard simplex $S = \{(a, b)^T, c, d\}$ we have $A \cdot S \not\subseteq \operatorname{int}(S)$ despite $\varrho(|A|) < 1$.
Therefore, in the first example (3.1) $A \cdot X \subseteq \operatorname{int}(X)$ would be impossible for any $X \in \mathbf{I}S^n$ whereas we found a standard simplex S with $A \cdot S \subseteq \operatorname{int}(S)$, in the second example the interval iteration (2.1) is globally convergent whereas $A \cdot S \not\subseteq \operatorname{int}(S)$ for every standard simplex S.

REFERENCES

[1] ACRITH High-Accuracy Arithmetic Subroutine Library: General Information Manual, IBM Publications, GC33-6133 (1985).

[2] *Alefeld, G.; Herzberger, J.:* Introduction to Interval Computations. Academic Press (1983).

[3] *Hansen, E.:* Interval Arithmetic in Matrix Computations. Part I. SIAM J. Numer. Anal. 2, S. 308–320 (1965).

[4] *Hansen, E.:* Interval Arithmetic in Matrix Computations. Part II. SIAM J. Numer. Anal. 4, S. 1–9 (1967).

[5] IEEE 754 Standard for Floating-Point Arithmetic (1986).

[6] *Jansson, C.:* A Posteriori Error Bounds for the Solution of Linear Systems with Simplex Techniques. Submitted to publication.

[7] *Krawczyk, R.:* Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken. Computing 4, S. 187–220 (1969).

[8] *Kulisch, U.; Miranker, W. L.:* Computer Arithmetic in Theory and Practice. Academic Press, New York (1981).

[9] *Moore, R. E.:* Interval Analysis. Prentice Hall, Englewood Cliffs, New Jersey (1966).

[10] *Neumaier, A.:* Interval Methods for Systems of Equations. Cambridge University Press, to be published.

[11] *Rump, S. M.:* Kleine Fehlerschranken bei Matrixproblemen. Dissertation. Universität Karlsruhe (1980).

[12] *Rump, S. M.:* New Results on Verified Inclusions. In: *Miranker, W. L.; R. Toupin* (eds.): Accurate Scientific Computations. Springer Lecture Notes in Computer Science 235, 39 Seiten (1986).

[13] Siemens AG: Arithmos (BS2000). Benutzerhandbuch, (1986).

[14] *Varga, R. S.:* Matrix Iterative Analysis. Prentice Hall, Englewood Cliffs, New Jersey (1962).

[15] *Rump, S. M.:* Rigorous Sensitivity Analysis for Systems of Linear and Nonlinear Equations. MATH. of Comp., Vol. 54, No. 190, S. 721–736 (1990).

[16] *Rump, S. M.:* On the Solution of Interval Linear Systems. To appear in Computing.

Verfasser:

Prof. Dr. *S. M. Rump*,
Technische Informatik III,
Technische Universität Hamburg–Harburg,
Eißendorfer Straße 38,
W-2100 Hamburg 90.