ENCLOSING SOLUTIONS OF LINEAR EQUATIONS

JIRI ROHN* AND GEORG REX†

Abstract. It is shown that Rump's method for enclosing solutions of linear equations can be reformulated in an interval–free form and that the underlying inclusion result can be proved by elementary means without using Brouwer's fixed–point theorem. A sufficient condition on Rump's "inflation parameter" ε is given under which finite termination occurs. Also, a more general modified algorithm is studied for which the number of iterations can be expressed by an explicit formula.

Key words. Linear equations, enclosure, interval–free, finite termination

AMS subject classifications. 15A06, 65G10

1. Introduction. S. M. Rump in his basic paper [11] and in a series of subsequent papers [12], [13], [14], [15] developed a method for enclosing solutions of systems of linear equations. The most attractive feature of the method consists in the fact that it yields a *validated* enclosure (i.e., a narrow hyperrectangle containing the solution) computed by a finite precision arithmetic; hence, the effect of rounding errors arising in finite precision computations is controlled by means of finite precision computations. The method is described in sufficient detail in section 3 below. For the purposes of the Introduction, we shall give only a brief sketch of it here.

In order to solve a system of linear equations

$$A\hat{x} = b \tag{1.1}$$

with $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$, after computing an approximate solution x_0 of (1.1) and an approximate inverse of A, the equation (1.1) is transformed into a residual form

$$x^* = Gx^* + g, \tag{1.2}$$

where $x^* = \hat{x} - x_0$ is the difference between the exact and the approximate solution. The basic idea of the enclosure method consists in applying the Brouwer's fixed–point theorem to (1.2) to guarantee existence of the solution x^* of (1.2) in a given interval vector (hyperrectangle) X. The respective inclusion is theoretically formulated in power set operations, but in practical computations interval arithmetic operations must be used; then the inclusion takes on the form

$$G \odot X \oplus g \subset \mathrm{Int}(X) \tag{1.3}$$

(Rump's Theorem 3.1 below), where $\odot$, $\oplus$ denote the interval arithmetic multiplication and addition (defined in section 2) and $\mathrm{Int}(X)$ is the interior of X. Rump proposed in [11] an algorithm for finding an interval vector X satisfying (1.3). The algorithm is formulated in terms of interval arithmetic operations and contains a parameter $\varepsilon \in (0,1)$ (called the "inflation parameter" by Rump) as a tool for enforcing finite termination of the algorithm. Even so, finite termination is guaranteed only

* Faculty of Mathematics and Physics, Charles University, Prague (rohn@ms.mff.cuni.cz) and Institute of Computer Science, Academy of Sciences, Prague, Czech Republic (rohn@uivt.cas.cz). This author's work was supported by the Czech Republic Grant Agency under grant GAČR 201/95/1484.

† Institute of Mathematics, University of Leipzig, Augustusplatz 10-11, D-04109 Leipzig, Germany (rex@mathematik.uni-leipzig.d400.de).

under some conditions [11], [13], but in general the algorithm proved to perform very well: in most practical cases it terminates in a few iterations [2] and the computed enclosure often exhibits a least significant bit accuracy [11].

The present work was motivated by an attempt to understand what is going on behind Rump's method in terms of classical numerical analysis. This has led to formulation of three questions. First, can the basic inclusion (1.3) be formulated in a more usual way without using interval arithmetic operations? And, as a related issue, is Brouwer's fixed–point theorem really needed in considerations concerning a simple linear problem (1.2)? Second, can Rump's algorithm be formulated in an interval–free form? And third, what is the role of the "inflation parameter" ε and under what conditions on it can finite termination of the algorithm be guaranteed?

After a brief introduction of interval arithmetic in section 2 and a more detailed description of Rump's method in section 3, we address the above three questions in sections 4 to 6. In Theorem 4.1 we show that the inclusion (1.3) can be equivalently written in the form of a simple inequality

$$|(I - G)x - g| < (I - |G|)d \tag{1.4}$$

involving only the usual real arithmetic operations, where x is the center and d is the radius of the interval vector X. Next we prove by elementary means that if (1.4) holds, then the solution x^* of (1.2) satisfies $x - d < x^* < x + d$ (Theorem 4.2); this, in view of Theorem 4.1, gives an elementary proof of Rump's Theorem 3.1, avoiding an explicit use of Brouwer's fixed–point theorem. Based on some simple properties of interval arithmetic operations given in Lemma 2.1 and Lemma 2.2 of section 2, we then give in section 5 an interval–free description of Rump's algorithm, using only the usual real arithmetic operations and absolute values. This interval–free version of the algorithm generates *the same* sequence of interval vectors as the original Rump's algorithm, hence it can be used alternatively. These results yield answers to the first two questions. Next, in Theorem 6.3 of section 6 we give a sufficient condition

$$(1 + 4\varepsilon)|G| \cdot |x^*| < |x^*| \tag{1.5}$$

for a finite termination of Rump's algorithm (both in the original or in the interval–free version) using an inflation parameter ε. The proof of this result is rather complicated and for clarity it is preceded by two auxiliary lemmas. Finally, in section 7 we investigate a modification of the original algorithm (also proposed by Rump [13]) in which an additive constant is employed instead of the multiplicative parameter ε. In Theorem 7.1 we show that this modified algorithm has a finite termination property if and only if

$$\varrho(|G|) < 1$$

holds (independently of the choice of the additive constant), which, compared with the sufficient condition (1.5) for the original algorithm, is a much better result. Also, the number of iterations taken by the modified algorithm can be expressed by an explicit formula (Theorem 7.2).

The proofs of all these results are carried out by linear algebraic means. Our basic tool, frequently employed throughout, is the equivalence of the four assertions

(i) $\varrho(|G|) < 1$,
(ii) $|G|x < x$ for some $x > 0$,
(iii) $(I - |G|)^{-1} \geq 0$,

(iv) $|G|^j \to 0$

for a square matrix G (since $|G|$ is nonnegative), which can be found e.g. in Varga [16] or Neumaier [6].

We hope that these results may contribute to better understanding of the principles of validated computations from a classical (noninterval) point of view. Also, the noninterval inequality (1.4) of Theorem 4.1 may serve as a theoretical basis for deriving other alternative methods for computing validated solutions of linear equations.

2. Interval arithmetic. In this section we briefly survey the basic rules of interval arithmetic (described in detail in Alefeld and Herzberger [1] or Neumaier [6]) and we prove some simple properties to be used later. By an *interval* we always understand a nonempty compact real interval $[\underline{a}, \overline{a}] = \{a;\ \underline{a} \le a \le \overline{a}\}$. Operations over intervals are defined by the general rule

$$[\underline{a}, \overline{a}] \circ [\underline{b}, \overline{b}] = \{\alpha \circ \beta;\ \alpha \in [\underline{a}, \overline{a}],\ \beta \in [\underline{b}, \overline{b}]\}, \tag{2.1}$$

where $\circ$ denotes any of the four arithmetic operations. To make a clear distinction from the usual real arithmetic operations, we denote the interval arithmetic operations by $\oplus$, $\ominus$, $\odot$ and $\oslash$. It is easy to show ([1], [6]) that the general definition (2.1) yields the following explicit formulae:

$$[\underline{a}, \overline{a}] \oplus [\underline{b}, \overline{b}] = [\underline{a} + \underline{b}, \overline{a} + \overline{b}],$$

$$[\underline{a}, \overline{a}] \ominus [\underline{b}, \overline{b}] = [\underline{a} - \overline{b}, \overline{a} - \underline{b}],$$

$$[\underline{a}, \overline{a}] \odot [\underline{b}, \overline{b}] = [\min M, \max M],$$

where

$$M = \{\underline{a}\underline{b}, \underline{a}\overline{b}, \overline{a}\underline{b}, \overline{a}\overline{b}\},$$

and

$$[\underline{a}, \overline{a}] \oslash [\underline{b}, \overline{b}] = [\underline{a}, \overline{a}] \odot \left[\frac{1}{\overline{b}}, \frac{1}{\underline{b}}\right]$$

provided $0 \notin [\underline{b}, \overline{b}]$. In particular, interval arithmetic operations also apply to real numbers if we identify a real number a with the interval $[a, a]$; in this sense interval arithmetic is an extension of real arithmetic.

We shall prove here two very specific properties of interval arithmetic multiplication that will be used later. Let us recall that when speaking of an interval $[\underline{a}, \overline{a}]$, we always understand implicitly that it is nonempty, i.e., $\underline{a} \le \overline{a}$.

LEMMA 2.1. *Let α, β and $\delta \ge 0$, $\varepsilon \in [0, 1]$ be real numbers. Then we have*

$$\alpha \odot [\beta - \delta, \beta + \delta] = [\alpha\beta - |\alpha|\delta, \alpha\beta + |\alpha|\delta] \tag{2.2}$$

and

$$[1 - \varepsilon, 1 + \varepsilon] \odot [\alpha, \beta] = [\alpha - \varepsilon|\alpha|, \beta + \varepsilon|\beta|]. \tag{2.3}$$

Proof. 1) According to the above explicit rules, the lower bound of $\alpha\odot[\beta-\delta,\beta+\delta]$ is equal to $\min\{\alpha\beta-\alpha\delta,\alpha\beta+\alpha\delta\}=\alpha\beta-|\alpha\delta|=\alpha\beta-|\alpha|\delta$, and the upper bound is $\max\{\alpha\beta-\alpha\delta,\alpha\beta+\alpha\delta\}=\alpha\beta+|\alpha\delta|=\alpha\beta+|\alpha|\delta$.

2) In view of $\varepsilon\in[0,1]$, the lower bound of $[1-\varepsilon,1+\varepsilon]\odot[\alpha,\beta]$ is equal to $\min\{(1-\varepsilon)\alpha,(1-\varepsilon)\beta,(1+\varepsilon)\alpha,(1+\varepsilon)\beta\}=\min\{(1-\varepsilon)\alpha,(1+\varepsilon)\alpha\}=\alpha-|\varepsilon\alpha|=\alpha-\varepsilon|\alpha|$; the proof for the upper bound is analogous. □

An n-dimensional *interval vector* is a set of the form

$$[\underline{x},\overline{x}]=\{x;\ \underline{x}\le x\le\overline{x}\} \tag{2.4}$$

(componentwise inequalities), where $\underline{x}$, $\overline{x}\in\mathbb{IR}^n$, $\underline{x}\le\overline{x}$. For the purposes of applicability of interval operations, an interval vector (2.4) is identified with an n-dimensional vector with interval components $[\underline{x}_i,\overline{x}_i]$ $(i=1,\dots,n)$, i.e. we adopt the convention that

$$([\underline{x},\overline{x}])_i=[\underline{x}_i,\overline{x}_i]$$

for each i. This enables us to define a matrix–vector interval multiplication and addition

$$G\odot[\underline{x},\overline{x}]\oplus[\underline{g},\overline{g}], \tag{2.5}$$

where $G=(g_{ij})$ is a real $n\times n$ matrix, as an interval vector with the components

$$(G\odot[\underline{x},\overline{x}]\oplus[\underline{g},\overline{g}])_i=g_{i1}\odot[\underline{x}_1,\overline{x}_1]\oplus\ldots\oplus g_{in}\odot[\underline{x}_n,\overline{x}_n]\oplus[\underline{g}_i,\overline{g}_i] \tag{2.6}$$

$(i=1,\dots,n)$; we can see that (2.6) is a usual matrix multiplication and addition where the real operations are replaced by the interval ones. The following lemma (which turns out to be the basic tool for an interval–free reformulation of Rump's algorithm) shows that the expression (2.5) can be evaluated *without* using interval arithmetic operations. To this end, the interval vector $[\underline{x},\overline{x}]$ must be written in the form $[x-d,x+d]$ (hence, $x=\frac{1}{2}(\underline{x}+\overline{x})$ and $d=\frac{1}{2}(\overline{x}-\underline{x})$), and we employ the absolute value of G defined by $|G|=(|g_{ij}|)$; notice that only real operations appear on the right–hand side:

LEMMA 2.2. *We have*

$$G\odot[x-d,x+d]\oplus[\underline{g},\overline{g}]=[Gx-|G|d+\underline{g},Gx+|G|d+\overline{g}]. \tag{2.7}$$

Proof. From (2.6), using (2.2) we have

$$\begin{aligned}
&(G\odot[x-d,x+d]\oplus[\underline{g},\overline{g}])_i\\
=\ & g_{i1}\odot[x_1-d_1,x_1+d_1]\oplus\ldots\oplus g_{in}\odot[x_n-d_n,x_n+d_n]\oplus[\underline{g}_i,\overline{g}_i]\\
=\ & [g_{i1}x_1-|g_{i1}|d_1,g_{i1}x_1+|g_{i1}|d_1]\oplus\ldots\oplus[g_{in}x_n-|g_{in}|d_n,g_{in}x_n+|g_{in}|d_n]\oplus[\underline{g}_i,\overline{g}_i]\\
=\ & [\sum_j g_{ij}x_j-\sum_j|g_{ij}|d_j+\underline{g}_i,\sum_j g_{ij}x_j+\sum_j|g_{ij}|d_j+\overline{g}_i]\\
=\ & [(Gx-|G|d+\underline{g})_i,(Gx+|G|d+\overline{g})_i]\\
=\ & [Gx-|G|d+\underline{g},Gx+|G|d+\overline{g}]_i,
\end{aligned}$$

which is (2.7). □

3. Rump's method. Consider a system of linear equations

$$A\hat{x} = b \tag{3.1}$$

with an $n \times n$ matrix A. For an arbitrary nonsingular $n \times n$ matrix R and arbitrary $x_0 \in \mathbb{R}^n$, (3.1) is equivalent to

$$\hat{x} - x_0 = (I - RA)(\hat{x} - x_0) + R(b - Ax_0).$$

Hence if we put

$$G = I - RA, \tag{3.2}$$

$$g = R(b - Ax_0), \tag{3.3}$$

then we have

$$\hat{x} = x_0 + x^*, \tag{3.4}$$

where x^* solves

$$x^* = Gx^* + g. \tag{3.5}$$

In practice it is recommended to choose $R \approx A^{-1}$ and $x_0 \approx \hat{x}$, so that G and g are of small norm and x^* is close to 0.

Rump's basic idea on enclosing the solution of (3.5), which goes back to Krawczyk [3], [4] and Moore [5], is contained in the following theorem, where $\mathrm{Int}(X)$ denotes the interior of X (i.e., $\mathrm{Int}(X) = \{x;\ \underline{x} < x < \overline{x}\}$ for $X = [\underline{x}, \overline{x}]$) and g is identified with $[g, g]$:

THEOREM 3.1. (Rump [11]) *Let an interval vector X satisfy*

$$G \odot X \oplus g \subset \mathrm{Int}(X). \tag{3.6}$$

Then the equation (3.5) has a unique solution $x^* \in \mathrm{Int}(X)$.

In fact, in view of the basic property (2.1) of interval arithmetic operations, from (3.6) follows

$$G \cdot X + g := \{Gx + g;\ x \in X\} \subseteq G \odot X \oplus g \subset \mathrm{Int}(X) \subset X,$$

hence, due to the Brouwer fixed–point theorem, the mapping $x \mapsto Gx + g$ has a fixed point x^* in X. Therefore (3.5) holds, which implies that $x^* = Gx^* + g \in \mathrm{Int}(X)$. As explained in [11], the use of $\mathrm{Int}(X)$ instead of X on the right–hand side of (3.6) (which is not necessary for application of Brouwer's theorem) implies nonsingularity of $I - G$ and consequently the uniqueness of the solution of (3.5). The relationship of this result to the original problem of solving (3.1) is provided by the following theorem based on (3.4):

THEOREM 3.2. (Rump [11]) *Let G and g be given by (3.2) and (3.3) and let (3.6) hold for some interval vector X. Then A is nonsingular and the solution $\hat{x}$ of (3.1) satisfies*

$$\hat{x} \in x_0 \oplus \mathrm{Int}(X).$$

In view of this result, we may restrict our attention to enclosing the solution of the "residual equation" (3.5) in the sequel.

In his paper [11, p. 62], Rump proposed the following algorithm for finding an interval vector containing the solution x^* of (3.5):

select $\varepsilon \in (0,1)$;
$Y := [g,g]$;
repeat
  $X := [1-\varepsilon, 1+\varepsilon] \odot Y$;
  $Y := G \odot X \oplus g$
until $Y \subset \mathrm{Int}(X)$;
{then $x^* \in Y$}.

If the stopping rule $Y \subset \mathrm{Int}(X)$ is satisfied, then (3.6) holds, hence by Theorem 3.1, $x^* = Gx^* + g \in G \odot X \oplus g = Y$. When using the interval arithmetic operations, downwardly and upwardly oriented rounding must be used to guarantee that $Y \subset \mathrm{Int}(X)$ holds; then Y is a verified enclosure of the solution x^*.

The algorithm proved to perform very well. Practical experience shows that if it terminates in a finite number of steps, then the number of loops is relatively small ("it is an empirical fact that the inner inclusion is satisfied nearly always after a few steps or never" [2, p. 180]) and the solution is often computed with least significant bit accuracy [11]. However, finite termination is not guaranteed: if $\varrho(|G|) \geq 1$, then the stopping rule is never satisfied (Theorem 4.3 below) and the algorithm constructs an infinite sequence of interval vectors. Although it has been reported that the number of loops is approximately independent of the choice of the "inflation parameter" ε even for values far exceeding the prescribed range $(0,1)$ (see [2]), it seems that the problem of choosing an appropriate value of ε that would guarantee finite termination of the algorithm still remains open.

Summing up, there are three basic questions concerning Rump's method: first, whether Rump's condition (3.6) can be given a more transparent form; second, whether the algorithm can be formulated without using interval arithmetic; and third, what values of the inflation parameter ε (if they exist at all) guarantee finite termination of the algorithm. We shall address these questions in the next three sections.

4. Reformulation of Rump's condition. We start with an equivalent reformulation of the condition (3.6) of Theorem 3.1 in the form of a simple inequality (cf. [7], [10]). Notation: I is the unit matrix, the absolute value of $x = (x_i)$ is given by $|x| = (|x_i|)$, and vector inequality $x < y$ is understood componentwise.

THEOREM 4.1. *Let $X = [x-d, x+d]$. Then Rump's condition*

$$G \odot X \oplus g \subset \mathrm{Int}(X)$$

is equivalent to

$$|(I-G)x - g| < (I - |G|)d. \tag{4.1}$$

Proof. Since $G \odot X \oplus g = [Gx - |G|d + g, Gx + |G|d + g]$ by Lemma 2.2, (3.6) is equivalent to

$$x - d < Gx - |G|d + g,$$

$$Gx + |G|d + g < x + d,$$

which in turn is equivalent to

$$-(I - |G|)d < (I - G)x - g < (I - |G|)d$$

and thus also to (4.1). □

In this way, we have avoided the use of interval arithmetic in the formulation of Rump's condition. Let us note that in terms of the original problem (3.1) the condition (4.1) reads

$$|R(A(x + x_0) - b)| < (I - |I - RA|)d.$$

We shall now prove by elementary means that (4.1) implies $x^* \in \mathrm{Int}(X)$. This, in the light of Theorem 4.1, gives an elementary proof of Rump's Theorem 3.1, avoiding an explicit use of Brouwer's fixed–point theorem (cf. [13, Lemma 10]).

THEOREM 4.2. *If x and $d > 0$ satisfy (4.1), then the equation (3.5) has a unique solution x^* and*

$$x - d < x^* < x + d \tag{4.2}$$

holds.

Proof. From (4.1) we have $0 \leq |G|d < d$, hence $d > 0$, so that the inequality $|G|d < d$ implies $\varrho(G) \leq \varrho(|G|) < 1$ (see [16]), hence $I - G$ is nonsingular and (3.5) has a unique solution x^*. Next, from

$$x^* = Gx^* + g$$

we have

$$x^* - x = G(x^* - x) + g - (I - G)x$$

and taking absolute values we obtain

$$|x^* - x| \leq |G| \cdot |x^* - x| + |(I - G)x - g|,$$

hence in view of (4.1),

$$(I - |G|)|x^* - x| \leq |(I - G)x - g| < (I - |G|)d. \tag{4.3}$$

Since $\varrho(|G|) < 1$ implies $(I - |G|)^{-1} \geq 0$, premultiplying (4.3) by this nonnegative matrix yields

$$|x^* - x| < d,$$

which is (4.2). □

Next we prove a necessary and sufficient condition for solvability of (4.1). A similar result for (3.6) was proved by Rump in [13].

THEOREM 4.3. *The inequality (4.1) has a solution x and $d \geq 0$ if and only if*

$$\varrho(|G|) < 1 \tag{4.4}$$

holds.

Proof. The "only if" part was proved in the proof of Theorem 4.2. If (4.4) holds, then $I - G$ is nonsingular, hence (3.5) has a solution x^*, and there exists a $d > 0$ satisfying $|G|d < d$ [16]. Then (4.1) is satisfied by x^* and d. □

Hence, no interval vector X satisfies the inclusion (3.6) if $\varrho(|G|) \geq 1$; this means that Rump's algorithm will never terminate in this case.

Next we give a description of all solutions of (4.1) which employs a positive parameter vector δ. This result is a generalization of Theorem 2 in [8].

THEOREM 4.4. *Let (4.4) hold. Then $X = [x - d, x + d]$ satisfies (4.1) (or, equivalently, (3.6)) if and only if d is of the form*

$$d = (I - |G|)^{-1}(|(I - G)x - g| + \delta) \tag{4.5}$$

for some $\delta > 0$.

Proof. If x and d satisfy (4.1), then for

$$\delta := (I - |G|)d - |(I - G)x - g|$$

we have $\delta > 0$ and (4.5). Conversely, if (4.5) holds for some $\delta > 0$, then

$$(I - |G|)d = |(I - G)x - g| + \delta > |(I - G)x - g|,$$

so that x and d satisfy (4.1). □

Hence, if (4.4) holds, then for any $x \in \mathbb{R}^n$ we can construct an interval vector $X = [x - d, x + d]$ satisfying Rump's condition when computing d from (4.5) for some (but arbitrary) $\delta > 0$. In view of nonnegativity of $(I - |G|)^{-1}$, (4.5) gives the lower bound

$$d > (I - |G|)^{-1}|(I - G)x - g|$$

which is independent of the choice of δ.

5. Interval–free version of the algorithm. The results of Lemma 2.1 and Lemma 2.2 enable us to formulate an interval–free version of Rump's algorithm. Let us denote the interval vectors X and Y appearing in the algorithm (section 3) by $X = [x - d, x + d]$ and $Y = [y - h, y + h]$. Since from the updating formulae

$$\begin{aligned} X &:= [1 - \varepsilon, 1 + \varepsilon] \odot Y, \\ Y &:= G \odot X \oplus g \end{aligned}$$

we have

$$[x - d, x + d] := [1 - \varepsilon, 1 + \varepsilon] \odot [y - h, y + h] = [y - h - \varepsilon|y - h|, y + h + \varepsilon|y + h|]$$

(Lemma 2.1, (2.3)) and

$$[y - h, y + h] := G \odot [x - d, x + d] \oplus g = [Gx - |G|d + g, Gx + |G|d + g],$$

(Lemma 2.2), which gives

$$\begin{aligned} x &:= y + \frac{\varepsilon}{2}(|y + h| - |y - h|), \\ d &:= h + \frac{\varepsilon}{2}(|y + h| + |y - h|), \\ y &:= Gx + g, \\ h &:= |G|d, \end{aligned}$$

and since $Y \subset \mathrm{Int}(X)$ is equivalent to

$$x - d < y - h,$$

$$y + h < x + d,$$

and thereby also to

$$|x - y| < d - h,$$

the Rump's original algorithm can be equivalently rewritten in the following interval–free form:

select $\varepsilon \in (0, 1)$;
$y := g$; $h := 0$;
repeat
  $x := y + \frac{\varepsilon}{2}(|y + h| - |y - h|)$;
  $d := h + \frac{\varepsilon}{2}(|y + h| + |y - h|)$;
  $y := Gx + g$;
  $h := |G|d$
until $|x - y| < d - h$;
{then $y - h \le x^* \le y + h$}.

It is worth emphasizing that this algorithm generates *the same* sequence of interval vectors $X = [x - d, x + d]$, $Y = [y - h, y + h]$ as the Rump's original algorithm, but interval arithmetic is not used here. As before, downwardly and upwardly oriented rounding must be used to guarantee that the stopping rule holds; then we have a verified enclosure $y - h \le x^* \le y + h$.

Consider any system

$$x^* = Gx^* + g$$

with an $n \times n$ matrix G for which Rump's algorithm terminates in a finite number of steps. Let us construct an $(n + 1) \times (n + 1)$ matrix G' and an $(n + 1)$-dimensional vector g' by

$$G' = \begin{pmatrix} G & 0 \\ 0^T & 0 \end{pmatrix},$$

$$g' = \begin{pmatrix} g \\ 0 \end{pmatrix}.$$

Then a simple computation shows that

$$\varrho(|G'|) = \varrho(|G|) < 1 \tag{5.1}$$

(due to Theorem 4.3), hence the equation

$$x' = G'x' + g' \tag{5.2}$$

has a unique solution

$$x' = \begin{pmatrix} x^* \\ 0 \end{pmatrix}.$$

However, due to the special structure of G and g, Rump's algorithm when applied to (5.2) generates for *each* $\varepsilon > 0$ a sequence of interval vectors $X = [x - d, x + d]$, $Y = [y - h, y + h]$ satisfying

$$x_{n+1} = d_{n+1} = y_{n+1} = h_{n+1} = 0 \tag{5.3}$$

at *each* iteration, as it can be easily seen from the above interval–free description. Hence, the stopping rule $|x - y| < d - h$ is never satisfied for any $\varepsilon > 0$ and finite termination is lost.

Rump formulated in [13, Lemma 21] a very general "inflation" condition under which the algorithm is finite. In our case the condition requires an existence of a vector $s > 0$ such that $\varepsilon|y - h| \geq s$, $\varepsilon|y + h| \geq s$ hold at each iteration. The equation (5.3) shows that in our example the condition is violated, hence Rump's result does not apply. In section 7 we shall describe a modification of Rump's algorithm that will be able to handle even this heavily degenerated example.

6. Finite termination conditions. The explicit form of iterations given in section 5 makes it possible to formulate another sufficient condition for finite termination of the algorithm. In order to make the proof of the main result more clear, we shall precede it by two auxiliary lemmas. The first lemma gives a sufficient condition for the four sequences

$$x_{j+1} = y_j + \frac{\varepsilon}{2}(|y_j + h_j| - |y_j - h_j|), \tag{6.1}$$

$$d_{j+1} = h_j + \frac{\varepsilon}{2}(|y_j + h_j| + |y_j - h_j|), \tag{6.2}$$

$$y_{j+1} = Gx_{j+1} + g, \tag{6.3}$$

$$h_{j+1} = |G|d_{j+1} \tag{6.4}$$

with $y_0 = g$, $h_0 = 0$ (see the description of the algorithm, section 5) to converge.

LEMMA 6.1. *Let* $\varepsilon \in (0, 1)$ *satisfy*

$$(1 + 2\varepsilon)\varrho(|G|) < 1. \tag{6.5}$$

Then the sequences $\{x_j\}$, $\{d_j\}$, $\{y_j\}$, $\{h_j\}$ *given by (6.1)–(6.4) are convergent.*

Proof. From (6.1)–(6.4) we obtain

$$\begin{aligned} |x_{j+1} - x_j| &\leq |y_j - y_{j-1}| + \frac{\varepsilon}{2}\,||y_j + h_j| - |y_{j-1} + h_{j-1}|| + \frac{\varepsilon}{2}\,||y_j - h_j| - |y_{j-1} - h_{j-1}|| \\ &\leq (1 + \varepsilon)|y_j - y_{j-1}| + \varepsilon|h_j - h_{j-1}| \\ &\leq (1 + \varepsilon)|G| \cdot |x_j - x_{j-1}| + \varepsilon|G| \cdot |d_j - d_{j-1}|, \end{aligned}$$

and in a similar way we get

$$|d_{j+1} - d_j| \leq \varepsilon|G| \cdot |x_j - x_{j-1}| + (1 + \varepsilon)|G| \cdot |d_j - d_{j-1}|,$$

which together gives

$$\begin{pmatrix} |x_{j+1} - x_j| \\ |d_{j+1} - d_j| \end{pmatrix} \leq \begin{pmatrix} (1 + \varepsilon)|G| & \varepsilon|G| \\ \varepsilon|G| & (1 + \varepsilon)|G| \end{pmatrix} \begin{pmatrix} |x_j - x_{j-1}| \\ |d_j - d_{j-1}| \end{pmatrix}. \tag{6.6}$$

Now, the condition (6.5) implies existence of a vector $x > 0$ satisfying

$$(1 + 2\varepsilon)|G|x < x$$

(see [16]). Then we have

$$\begin{pmatrix} (1+\varepsilon)|G| & \varepsilon|G| \\ \varepsilon|G| & (1+\varepsilon)|G| \end{pmatrix} \begin{pmatrix} x \\ x \end{pmatrix} = \begin{pmatrix} (1+2\varepsilon)|G|x \\ (1+2\varepsilon)|G|x \end{pmatrix} < \begin{pmatrix} x \\ x \end{pmatrix},$$

hence the spectral radius of the matrix

$$\begin{pmatrix} (1+\varepsilon)|G| & \varepsilon|G| \\ \varepsilon|G| & (1+\varepsilon)|G| \end{pmatrix}$$

is less than one, and the inequality (6.6) implies that the sequence

$$\left\{ \begin{pmatrix} x_j \\ d_j \end{pmatrix} \right\}$$

is a Cauchy sequence, hence it is convergent. This proves that $\{x_j\}$ and $\{d_j\}$ converge; convergence of $\{y_j\}$, $\{h_j\}$ then follows immediately from (6.3), (6.4). □

The second lemma gives a sufficient condition for a special nonlinear equation to have a solution whose all entries are nonzero:

LEMMA 6.2. *Let $Q \in \mathbb{R}^{m\times m}$, $q \in \mathbb{R}^m$ and let*

$$2|Q| \cdot |q| < |q| \tag{6.7}$$

hold. Then the equation

$$y = Q|y| + q \tag{6.8}$$

has a unique solution $\hat{y}$ and all entries of $\hat{y}$ are nonzero and are of the same signs as the respective entries of q.

Proof. As before, from (6.7) we deduce that $\varrho(|Q|) < \frac{1}{2}$, hence $|Q|^j \to 0$ and $(I - |Q|)^{-1} \geq 0$. If we construct the iteration

$$y_{k+1} = Q|y_k| + q \tag{6.9}$$

$(k = 0, 1, \ldots)$, $y_0 = q$, then we have

$$|y_{k+1} - y_k| \leq |Q| \cdot |y_k - y_{k-1}|,$$

which in view of $|Q|^j \to 0$ implies that $\{y_k\}$ is a Cauchy sequence, hence $y_k \to \hat{y}$, so that (6.9) gives

$$\hat{y} = Q|\hat{y}| + q, \tag{6.10}$$

and the solution is unique since from $y = Q|y| + q$ we obtain $|\hat{y} - y| \leq |Q| \cdot |\hat{y} - y|$, hence $(I - |Q|)|\hat{y} - y| \leq 0$ and premultiplying by $(I - |Q|)^{-1} \geq 0$ gives $|\hat{y} - y| \leq 0$, hence $\hat{y} = y$. Now, from (6.10) we obtain

$$|\hat{y}| \leq |Q| \cdot |\hat{y}| + |q|,$$

hence

$$|\hat{y}| \leq (I - |Q|)^{-1}|q|,$$

and again from (6.10),

$$|\hat{y} - q| \leq |Q| \cdot |\hat{y}| \leq |Q|(I - |Q|)^{-1}|q| = (I - |Q|)^{-1}|Q| \cdot |q|. \tag{6.11}$$

But (6.7) implies $|Q| \cdot |q| < (I - |Q|)|q|$, hence

$$(I - |Q|)^{-1}|Q| \cdot |q| < |q|,$$

which combined with (6.11) gives

$$|\hat{y} - q| < |q|.$$

This means that each $\hat{y}_i$ is nonzero and is of the same sign as q_i. □

Now we give a finite termination condition for Rump's algorithm (both in the original version of section 3 or in the interval–free version of section 5). Unfortunately, the condition involves the solution x^* and cannot be a priori verified.

THEOREM 6.3. *Rump's algorithm is finite for each inflation parameter $\varepsilon \in (0,1)$ satisfying*

$$(1 + 4\varepsilon)|G| \cdot |x^*| < |x^*|. \tag{6.12}$$

Proof. First, the inequality (6.12) implies

$$(1 + 2\varepsilon)\varrho(|G|) \leq (1 + 4\varepsilon)\varrho(|G|) < 1, \tag{6.13}$$

hence $x_j \to x$, $d_j \to d$, $y_j \to y$, $h_j \to h$ by Lemma 6.1. Taking the limits in (6.1)–(6.4), we obtain

$$x = y + \frac{\varepsilon}{2}(|y + h| - |y - h|), \tag{6.14}$$

$$d = h + \frac{\varepsilon}{2}(|y + h| + |y - h|), \tag{6.15}$$

$$y = Gx + g, \tag{6.16}$$

$$h = |G|d. \tag{6.17}$$

This implies

$$y = Gy + \frac{\varepsilon}{2}G(|y + h| - |y - h|) + g$$

and

$$h = |G|h + \frac{\varepsilon}{2}|G|(|y + h| + |y - h|),$$

hence

$$y = \frac{\varepsilon}{2}(I - G)^{-1}G(|y + h| - |y - h|) + x^* \tag{6.18}$$

and

$$h = \frac{\varepsilon}{2}(I - |G|)^{-1}|G|(|y + h| + |y - h|). \tag{6.19}$$

Then by subtracting and adding (6.18) and (6.19) we get

$$\hat{y} = Q|\hat{y}| + q,$$

where

$$\hat{y} = \begin{pmatrix} y - h \\ y + h \end{pmatrix},$$

$$q = \begin{pmatrix} x^* \\ x^* \end{pmatrix}$$

and

$$Q = \frac{\varepsilon}{2} \begin{pmatrix} -(I-G)^{-1}G - (I-|G|)^{-1}|G|, & (I-G)^{-1}G - (I-|G|)^{-1}|G| \\ -(I-G)^{-1}G + (I-|G|)^{-1}|G|, & (I-G)^{-1}G + (I-|G|)^{-1}|G| \end{pmatrix}.$$

Since (6.12) implies in the usual way that

$$4\varepsilon(I-|G|)^{-1}|G| \cdot |x^*| < |x^*|$$

holds, we have

$$2|Q| \cdot |q| \leq \begin{pmatrix} 4\varepsilon(I-|G|)^{-1}|G| \cdot |x^*| \\ 4\varepsilon(I-|G|)^{-1}|G| \cdot |x^*| \end{pmatrix} < |q|,$$

and Lemma 6.2 implies that $|\hat{y}| > 0$, hence $|y-h| > 0$ and $|y+h| > 0$. Then from (6.14) and (6.15) we have

$$|x-y| = \frac{\varepsilon}{2}(|y+h| - |y-h|) < \frac{\varepsilon}{2}(|y+h| + |y-h|) = d-h,$$

which means that

$$|x_j - y_j| < d_j - h_j$$

holds from some j on, hence the stopping rule is satisfied at some iteration and the algorithm is finite. □

COROLLARY 6.4. *Let*

$$|G| \cdot |x^*| < |x^*| \tag{6.20}$$

hold. Then there exists an $\varepsilon_0 > 0$ *such that Rump's algorithm is finite for each* $\varepsilon \in (0, \varepsilon_0)$.

Proof. In fact, according to Theorem 6.3 it is sufficient to take ε_0 as the supremum of all ε's satisfying (6.12). This ε_0 is positive due to (6.20). □

Notice that (6.20) implies $\varrho(|G|) < 1$, which is (4.4). The condition (6.20) is not as restrictive as it may seem since in practice the matrix G, computed by $G = I - RA$, where R is an approximation of A^{-1} (section 3), is close to 0.

7. Modified Rump's algorithm. Rump proposed in [13] also another algorithm scheme which employs an additive constant instead of a multiplicative one. This is done by replacing the statement

$$X := [1-\varepsilon, 1+\varepsilon] \odot Y$$

in the original algorithm (section 3) by

$$X := Y \oplus [-f, f],$$

where f is some (sufficiently small) prescribed positive vector. We shall call the resulting algorithm a *modified* Rump algorithm. In this section we show that this algorithm is much easier to analyze and that the number of steps can be given by an explicit formula.

Let us denote the interval vectors appearing in the modified algorithm by $X = [x-d, x+d]$, $Y = [y-h, y+h]$. Then from the updating formulae

$$\begin{aligned} X &:= Y \oplus [-f, f], \\ Y &:= G \odot X \oplus g \end{aligned}$$

we have

$$\begin{aligned}[x-d, x+d] &:= [y-h-f, y+h+f],\\ [y-h, y+h] &:= [Gx-|G|d+g, Gx+|G|d+g],\end{aligned}$$

(Lemma 2.2), which amounts to

$$\begin{aligned}x &:= y,\\ d &:= h+f,\\ y &:= Gx+g,\\ h &:= |G|d,\end{aligned}$$

and $Y \subset \mathrm{Int}(X)$ is equivalent to $|x-y| < d-h$. Hence the modified Rump algorithm can be written in the following interval–free form:

select $f > 0$;
$y := g$; $h := 0$;
repeat
  $x := y$;
  $d := h + f$;
  $y := Gx + g$;
  $h := |G|d$
until $|x - y| < d - h$;
{then $y - h \le x^* \le y + h$}.

It turns out that, in contrast to the original algorithm, finite termination of the modified algorithm can be characterized easily:

THEOREM 7.1. *The modified Rump algorithm terminates in a finite number of steps for each $f > 0$ if and only if*

$$\varrho(|G|) < 1 \tag{7.1}$$

holds.

Proof. Let $\{x_j\}$, $\{d_j\}$, $\{y_j\}$ and $\{h_j\}$ be the sequences generated by the modified algorithm, with $y_0 = g$, $h_0 = 0$. Then from the recurrences

$$\begin{aligned}x_{j+1} &= y_j,\\ d_{j+1} &= h_j + f,\\ y_{j+1} &= Gx_{j+1} + g,\\ h_{j+1} &= |G|d_{j+1}\end{aligned}$$

it follows easily by induction that

$$\begin{aligned}x_j &= \sum_{\ell=0}^{j-1} G^\ell g,\\ d_j &= \sum_{\ell=0}^{j-1} |G|^\ell f,\\ y_j &= \sum_{\ell=0}^{j} G^\ell g,\end{aligned}$$

$$h_j \quad = \quad \sum_{\ell=1}^{j} |G|^\ell f$$

($j = 1, 2, \ldots$). Hence, the stopping rule

$$|x_k - y_k| < d_k - h_k$$

is satisfied for some k if and only if

$$|G^k g| + |G|^k f < f \tag{7.2}$$

holds. Now, if the algorithm terminates in a finite number of steps for some $f > 0$, then (7.2) holds, hence $|G|^k f < f$, implying $\varrho(|G|^k) < 1$ and

$$(\varrho(|G|))^k \le \varrho(|G|^k) < 1,$$

which gives (7.1). Conversely, if (7.1) holds, then $G^j \to 0$ and $|G|^j \to 0$, hence for each $f > 0$ there exists a k such that (7.2) is satisfied, which means that $|x_k - y_k| < d_k - h_k$, and the modified algorithm terminates. □

Since the condition (7.1) is identical with (4.4), this result shows a remarkable property: if Rump's inclusion (3.6) (equivalently, (4.1)) has a solution, then a solution to it can be found by the modified algorithm. Hence, it is more general than the original algorithm of section 3 for which the finite termination condition of Theorem 6.3 is more restrictive. In particular, the example given in section 5 on which Rump's algorithm fails can be solved by the modified algorithm since $\varrho(|G'|) < 1$ (eq. (5.1)).

THEOREM 7.2. *If (7.1) holds, then the modified Rump algorithm terminates at the k-th iteration, where*

$$k = \min\{j;\ |G^j g| + |G|^j f < f\}. \tag{7.3}$$

Proof. Obviously, k is the minimum value of j for which $|x_j - y_j| < d_j - h_j$ holds; this, according to the first part of the previous proof, is equivalent to $|G^j g| + |G|^j f < f$. Hence (7.3) follows. □

As explained in section 3, in practice $G := I - RA$ is small, hence G^j, $|G|^j$ will converge rapidly to 0 and the stopping rule (7.2) can be expected to be satisfied after a few steps. In particular, if $|Gg| + |G|f < f$, then $k = 1$. Let us note that Theorems 7.1 and 7.2 improve the result of Lemma 2.4 in [11].

8. Final remarks. The interval-free versions of Rump's algorithms described in sections 5 and 7 are not only easier to understand, but also advantageous in practice since they require fewer switchings of the rounding mode (which is as costly as multiplication or addition). In Theorem 4.2 we showed that if x and $d > 0$ satisfy (4.1), then $x - d < x^* < x + d$. Hence, other alternative methods for computing validated solutions of linear equations, based on solving directly the inequality (4.1), may be designed. Such methods were proposed by Rex [9] and Rohn [10].

Acknowledgments. A part of this work was done during the first author's stay at the Center of Theoretical Sciences of the University of Leipzig. The authors wish to thank Prof. S. M. Rump and Prof. G. Heindl for valuable discussions on the subject of this paper.

REFERENCES

[1] G. Alefeld and J. Herzberger, *Introduction to Interval Computations*, Academic Press, New York, 1983.
[2] R. Hammer, M. Hocks, U. Kulisch and D. Ratz, *Numerical Toolbox for Verified Computing I*, Springer–Verlag, Berlin, 1993.
[3] R. Krawczyk, *Newton–Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken*, Computing, 4 (1969), pp. 187–201.
[4] ———, *Intervalliterationsverfahren*, Bericht 186, Mathematisch–Statistische Sektion im Forschungszentrum Graz, Graz, 1982.
[5] R. E. Moore, *A test for existence of solutions to nonlinear systems*, SIAM Journal on Numerical Analysis, 14 (1977), pp. 611–615.
[6] A. Neumaier, *Interval Methods for Systems of Equations*, Cambridge University Press, Cambridge, 1990.
[7] H.-G. Rex, *Zur Lösungseinschließung linearer Gleichungssysteme*, Wissenschaftliche Zeitschrift, Technische Hochschule Leipzig, 15 (1991), pp. 441–447.
[8] G. Rex, *Zu a posteriori Fehlerabschätzungen bei linearen Gleichungssystemen*, Zeitschrift für Angewandte Mathematik und Mechanik, 72 (1992), pp. T640–T643.
[9] ———, *Parameterabhängige Lösungseinschließungen linearer Gleichungssysteme*, Zeitschrift für Angewandte Mathematik und Mechanik, 74 (1994), pp. T683–T685.
[10] J. Rohn, *Validated solutions of linear equations*, Technical Report 620, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, January 1995.
[11] S. M. Rump, *Solving algebraic problems with high accuracy*, in A New Approach to Scientific Computation, U. Kulisch and W. Miranker, ed., New York, 1983, Academic Press, pp. 51–120.
[12] ———, *Solution of linear and nonlinear algebraic problems with sharp, guaranteed bounds*, Computing Supplementum, 5 (1984), pp. 147–168.
[13] ———, *New results on verified inclusions*, in Accurate Scientific Computations, W. L. Miranker and R. A. Toupin, ed., Lecture Notes in Computer Science 235, Berlin, 1986, Springer–Verlag, pp. 31–69.
[14] ———, *On the solution of interval linear systems*, Computing, 47 (1992), pp. 337–353.
[15] ———, *Verification methods for dense and sparse systems of equations*, in Topics in Validated Computations, J. Herzberger, ed., Amsterdam, 1994, North–Holland, pp. 63–135.
[16] R. Varga, *Matrix Iterative Analysis*, Prentice–Hall, Englewood Cliffs, 1962.