Validated Solutions of Linear Equations[*]

Jiří Rohn[†]

Abstract

It is shown that Rump's method for computing validated 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.

Key words. Linear equations, validated enclosure, refinement, interval-free, finite termination

1 Introduction

This report was made as a transcript of transparencies. This is the reason for its terse style and division into short sections. It consists of two parts: an overview of Rump's method (sections 2 to 6) and its reformulation avoiding the use of Brouwer's fixed-point theorem and interval arithmetic (sections 7 to 18). The principal result is Theorem 1.

2 Problem

Solve

$$A\tilde{x} = b$$

(A square $n \times n$) in finite precision arithmetic and estimate the accuracy of the solution obtained by means of the finite precision arithmetic.

[*] This work was supported in part by the Czech Republic Grant Agency under grant GAČR 201/95/1484

[†] Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague (rohn@uivt.cas.cz), and Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (rohn@kam.ms.mff.cuni.cz)

3 Recasting

For arbitrary nonsingular R and x_0,

$$A\tilde{x} = b$$

is equivalent to

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

Hence

$$\tilde{x} = x_0 + x^*$$

where x^* solves

$$x^* = Gx^* + g \tag{1}$$

with

$$\begin{aligned} G &= I - RA \\ g &= R(b - Ax_0) \end{aligned}$$

(in practice: $R \approx A^{-1}$, $x_0 = Rb$, so that G and g are of small norms and x^* is close to 0). In the sequel we consider the equation (1).

4 Rump's inclusion idea

Let an interval vector (”box”) X satisfy

$$G \cdot X + g \subset X^0 \tag{2}$$

(where $G \cdot X + g = \{Gx + g;\ x \in X\}$, and X^0 is the interior of X). Then, in view of Brouwer's fixed-point theorem,

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

has a unique solution $x^* \in G \cdot X + g$ (Rump [1]).
How to verify (2): if

$$G \odot X \oplus g \subset X^0$$

holds in interval arithmetic, then (2) holds since

$$G \cdot X + g \subset G \odot X \oplus g$$

due to definitions of interval operations.

5 Interlude: interval arithmetic

Operations over intervals are defined by the general rule

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

explicitly:

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

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

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

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 $0 \notin [\underline{b},\overline{b}]$. A real number a is identified with $[a,a]$.

6 Rump's algorithm for solving $x^* = Gx^* + g$

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

The algorithm (Rump [1, p. 62]) proved to perform excellently:

- small number of iterations (usually ≤ 10),
- high accuracy achieved,
- relative independence on the "inflation parameter" ε.

Now, *what is behind it ?*

7 Enclosure theorem

Theorem 1 *Let x and $d > 0$ satisfy*

$$|(I-G)x - g| < (I - |G|)d. \tag{3}$$

Then the equation

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

has a unique solution x^* *and*

$$x - d < x^* < x + d$$

holds.

Comment The most important part of the assumption is the existence of a positive solution d of the inequality

$$|G|d < d.$$

If d possesses this property, then for *each* x there exists a positive real number α such that x and $d := \alpha d$ satisfy (3).

8 Proof

(3) implies $|G|d < d$, hence $\varrho(|G|) < 1$, $(I - |G|)^{-1} \geq 0$ and $I - G$ is nonsingular, so that (1) has a unique solution x^*. From

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

we obtain

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

which implies

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

and

$$(I - |G|)|x^* - x| \leq |(I - G)x - g|.$$

Premultiplying this inequality by $(I - |G|)^{-1}$ yields

$$|x^* - x| \leq (I - |G|)^{-1}|(I - G)x - g|$$

and from (3), also by premultiplying by the same matrix, we have

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

which together gives

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

□

9 Refinement

Theorem 2 *Let all rows of G be nonzero and let x and $d > 0$ satisfy (3). Then*

$$\begin{aligned} x' &:= Gx + g \\ d' &:= |G|d \end{aligned}$$

also satisfy (3) and

$$0 < d' < d$$

holds.

Comment Hence, once a solution to (3) has been found, a *nested* sequence of enclosures can be constructed whose radii tend to 0 provided $\varrho(|G|) < 1$ (since then $|G|^k \to 0$).

10 Proof

Under the assumptions we have

$$|(I-G)x'-g| = |G((I-G)x-g)| \le |G|\cdot|(I-G)x-g| < |G|\cdot(I-|G|)d = (I-|G|)d'.$$

Since no row of $|G|$ is a zero vector,

$$0 < d' = |G|d < d$$

follows from (3). □

11 Relationship to Rump's inclusion result

Theorem 3 *Let $X = [x-d, x+d]$. Then Rump's inclusion*

$$G \odot X \oplus g \subset X^0 \tag{4}$$

is equivalent to (3).

12 Proof

It follows from the formulae in section 5 that

$$G \odot X \oplus g = [Gx - |G|d + g, Gx + |G|d + g],$$

hence $G \odot X \oplus g \subset X^0$ if and only if

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

and

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

hold, which is equivalent to

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

□

Hence, Rump's inclusion result can be proved by elementary means without using Brouwer's fixed-point theorem. Also, the inclusion (4) can be verified as (3) without using interval arithmetic.

Now, the problem reduces to solving (3).

13 Solvability

Theorem 4 *The inequality (3) has a solution $x,\ d > 0$ if and only if*

$$\varrho(|G|) < 1. \tag{5}$$

14 Proof

If (3) holds, then $|G|d < d$, implying (5). Conversely, if (5) holds, then there exists a $d > 0$ satisfying $|G|d < d$, and the inequality (3) holds e.g. with x^* and d. □

15 Algorithm

Theorem 5 *If (5) holds, then a solution to (3) can be computed by the following finite algorithm:*

```
$f$ := a small positive vector;
$x' := 0; d' := 0;$
**repeat**
    $x := x'; d := d';$
    $x' := Gx + g;$
    $d' := |G|d + |x' - x| + f$
**until** $|d' - d| < f.$
```

Then x and d solve (3) and satisfy

$$(I - |G|)^{-1}|(I - G)x - g| < d < (I - |G|)^{-1}(|(I - G)x - g| + 2f). \tag{6}$$

Comment 1 Both downwardly and upwardly directed rounding must be used to guarantee that the key inequality $d - f < d' < d + f$ is satisfied.

Comment 2 The algorithm was formulated under the condition (5). In practice it is usually not known beforehand whether it is satisfied. In this case, to ensure the finiteness of the algorithm, we may change the stopping rule to

$$\textbf{until } (|d' - d| < f \textit{ or } |G|d' > d' \textit{ or } k = k_0)$$

where k is an iteration counter (which should be added into the main loop) and k_0 is a prescribed maximum number of iterations. If $|G|d' > d'$ holds, then $\varrho(|G|) > 1$ and (3) does not have a solution.

16 Proof

The algorithm generates sequences

$$\begin{aligned} x_{k+1} &= Gx_k + g \\ d_{k+1} &= |G|d_k + |x_{k+1} - x_k| + f \end{aligned}$$

that can be shown to be cauchian, hence convergent. Thus $|d_{k+1} - d_k| \to 0$ and

$$d_{j+1} - d_j \leq |d_{j+1} - d_j| < f$$

after a finite number of steps. Then

$$|G|d_j + |(I - G)x_j - g| < d_j$$

hence x_j, d_j solve (3). The first inequality in (6) follows from (3) as in the proof of Theorem 1. The second one follows from the identity

$$(I - |G|)d_j = |x_{j+1} - x_j| + f + d_j - d_{j+1}$$

and the facts that $(I - |G|)^{-1} \geq 0$ and $d_j - d_{j+1} < f$. □

Note The algorithm is not identical with that of Rump, which, due to the use of interval arithmetic, generates another sequence of boxes $[x_k - d_k, x_k + d_k]$.

17 Refinement procedure

The enclosure produced by the algorithm can be further refined by this procedure (based on Theorem 2):

$h :=$ a positive vector of accuracy wanted;
$x' := x; d' := d;$

```
$x'' := Gx' + g; d'' := |G|d';$
**repeat**
    $x := x'; d := d';$
    $x' := x''; d' := d'';$
    $x'' := Gx' + g;$
    $d'' := |G|d'$
**until** $(\mathit{not}\ |x' - x''| < d' - d''\ \mathit{or}\ d' < h);$
**if** $\mathit{not}\ |x' - x''| < d' - d''$ **then** $\{x - d < x^* < x + d\}$
**else** $\{x' - d' < x^* < x' + d'$ and $d' < h\}.$
```

The procedure generates a strictly nested sequence, i.e

$$[x' - d', x' + d'] \subset [x - d, x + d]^0$$

at each iteration. It either finds an enclosure with prescribed accuracy, or stops when the condition $|x' - x''| < d' - d''$ cannot be verified more. If (5) holds, then the radii tend to 0 and the condition $d' < h$ guarantees finite termination. As in the main algorithm, downwardly and upwardly directed rounding must be used.

18 Appendix: Interval–free version of Rump's iterations and a finite termination condition

The results of this appendix were found when the previous part had already been completed. Let us denote the interval vectors Y and X appearing in Rump's algorithm (section 6) by $Y = [\underline{y}, \overline{y}]$ and $X = [\underline{x}, \overline{x}]$. Since for $\varepsilon \in (0, 1)$ we have

$$[1 - \varepsilon, 1 + \varepsilon] \odot [\underline{y}, \overline{y}] = [\underline{y} - \varepsilon|\underline{y}|, \overline{y} + \varepsilon|\overline{y}|]$$

and

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

(section 12), the original Rump's algorithm can be equivalently rewritten in the following interval-free form:

```
select $\varepsilon \in (0, 1);\ \underline{y} := g;\ \overline{y} := g;$
**repeat**
    $\underline{x} := \underline{y} - \varepsilon|\underline{y}|;$
    $\overline{x} := \overline{y} + \varepsilon|\overline{y}|;$
    $\underline{y} := \frac{1}{2}G(\underline{x} + \overline{x}) - \frac{1}{2}|G|(\overline{x} - \underline{x}) + g;$
    $\overline{y} := \frac{1}{2}G(\underline{x} + \overline{x}) + \frac{1}{2}|G|(\overline{x} - \underline{x}) + g$
**until** $(\underline{x} < \underline{y}\ \mathit{and}\ \overline{y} < \overline{x});$
$\{$then $\underline{y} \leq x^* \leq \overline{y}\}.$
```

It is worth emphasizing that this algorithm generates *the same* sequence of boxes $Y = [\underline{y}, \overline{y}],\ X = [\underline{x}, \overline{x}]$ as the original Rump's algorithm, but the interval arithmetic is not used here.

The explicit form of iterations enables us to formulate a sufficient condition for finite termination, which is different from that one by Rump [1]:

Theorem 6 *Rump's algorithm for solving (1) terminates in a finite number of steps for each ε satisfying*

$$0 < \varepsilon < \frac{1}{2} \tag{7}$$

$$(1+\varepsilon)\varrho(|G|) < \frac{1}{2} \tag{8}$$

$$4\varepsilon(I - |G|)^{-1}|G| \cdot |x^*| < |x^*|. \tag{9}$$

Comment The assumptions imply that $\varrho(|G|) < \frac{1}{2}$ and $|x^*| > 0$. Conversely, if this is true, then ε satisfying (7)–(9) exists.

Proof. Denote the iterated boxes by $Y_k = [\underline{y}_k, \overline{y}_k],\ X_k = [\underline{x}_k, \overline{x}_k]$. From the explicit formulae

$$\begin{aligned}
\underline{x}_{k+1} &= \underline{y}_k - \varepsilon|\underline{y}_k| \\
\overline{x}_{k+1} &= \overline{y}_k + \varepsilon|\overline{y}_k| \\
\underline{y}_{k+1} &= \frac{1}{2}G(\underline{x}_{k+1} + \overline{x}_{k+1}) - \frac{1}{2}|G|(\overline{x}_{k+1} - \underline{x}_{k+1}) + g \\
\overline{y}_{k+1} &= \frac{1}{2}G(\underline{x}_{k+1} + \overline{x}_{k+1}) + \frac{1}{2}|G|(\overline{x}_{k+1} - \underline{x}_{k+1}) + g
\end{aligned}$$

we have

$$\begin{pmatrix} |\underline{x}_{k+1} - \underline{x}_k| \\ |\overline{x}_{k+1} - \overline{x}_k| \end{pmatrix} \le (1+\varepsilon) \begin{pmatrix} |G| & |G| \\ |G| & |G| \end{pmatrix} \begin{pmatrix} |\underline{x}_k - \underline{x}_{k-1}| \\ |\overline{x}_k - \overline{x}_{k-1}| \end{pmatrix}$$

for each k and since the spectral radius of the matrix on the right-hand side is equal to $2\varrho(|G|)$, from (8) we see that the sequences $\{\underline{x}_k\}$ and $\{\overline{x}_k\}$ are cauchian, hence $\underline{x}_k \to \underline{x},\ \overline{x}_k \to \overline{x},\ \underline{y}_k \to \underline{y},\ \overline{y}_k \to \overline{y}$. Taking the limits, we obtain

$$\begin{aligned}
\underline{x} &= \underline{y} - \varepsilon|\underline{y}| \\
\overline{x} &= \overline{y} + \varepsilon|\overline{y}| \\
\underline{y} &= \frac{1}{2}G(\underline{x} + \overline{x}) - \frac{1}{2}|G|(\overline{x} - \underline{x}) + g \\
\overline{y} &= \frac{1}{2}G(\underline{x} + \overline{x}) + \frac{1}{2}|G|(\overline{x} - \underline{x}) + g
\end{aligned}$$

which after some rearrangements leads to

$$\hat{y} = \varepsilon M|\hat{y}| + \hat{x} \tag{10}$$

where

$$\hat{y} = \begin{pmatrix} \underline{y} \\ \overline{y} \end{pmatrix}$$

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

and

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

Since

$$|M| \le \begin{pmatrix} (I-|G|)^{-1}|G|, & (I-|G|)^{-1}|G| \\ (I-|G|)^{-1}|G|, & (I-|G|)^{-1}|G| \end{pmatrix},$$

for ε satisfying (7) and (8) we have

$$\varrho(\varepsilon|M|) \le 2\varepsilon\varrho((I-|G|)^{-1}|G|) = \frac{2\varepsilon\varrho(|G|)}{1-\varrho(|G|)} < 2\varepsilon < 1,$$

hence from (10) we obtain

$$|\hat{y}| \le \varepsilon|M| \cdot |\hat{y}| + |\hat{x}|$$

and consequently (since $(I-\varepsilon|M|)^{-1} \ge 0$)

$$|\hat{y}| \le (I-\varepsilon|M|)^{-1}|\hat{x}|,$$

hence

$$|\varepsilon M|\hat{y}|| \le \varepsilon|M| \cdot |\hat{y}| \le (I-\varepsilon|M|)^{-1}\varepsilon|M| \cdot |\hat{x}|. \qquad (11)$$

But from (9) we have

$$2\varepsilon|M| \cdot |\hat{x}| < |\hat{x}|,$$

hence

$$\varepsilon|M| \cdot |\hat{x}| < (I-\varepsilon|M|)|\hat{x}|$$

and (by premultiplying)

$$(I-\varepsilon|M|)^{-1}\varepsilon|M| \cdot |\hat{x}| < |\hat{x}|$$

which combined with (11) gives

$$|\varepsilon M|\hat{y}|| < |\hat{x}|,$$

hence from (10) we have that $|\hat{y}| > 0$ (i.e., all entries of $\hat{y}$ are nonzero), so that $|\underline{y}| > 0$ and $|\overline{y}| > 0$. Then from the limit expressions above we obtain

$$\underline{x} = \underline{y} - \varepsilon|\underline{y}| < \underline{y}$$

and

$$\overline{x} = \overline{y} + \varepsilon|\overline{y}| > \overline{y},$$

hence $\underline{x}_k < \underline{y}_k$ and $\overline{y}_k < \overline{x}_k$ for some k, so that the algorithm is finite. □

Note Consider a system (1) for which there exists an i such that $g_i = 0$ and $G_{ij} = 0$ for each j (i.e., $x_i^* = 0$). Then $(\underline{x}_k)_i = (\overline{x}_k)_i = (\underline{y}_k)_i = (\overline{y}_k)_i = 0$ for each k, so that the algorithm will not terminate in a finite number of steps. (This is a theoretical result; in practical computations finite termination may occur due to roundoff errors.)

19 Final remark

As we have seen, Theorem 1 forms the common basis for Rump's algorithm and for the algorithm given in section 15. Another alternative algorithms may be formulated for solving the inequality (3); thus the area is open for further research.

20 Acknowledgments

I highly appreciate discussions on the subject of this report with Prof. G. Heindl in Wuppertal, July 1994, with Prof. S. M. Rump in Hamburg, October 1994, and with Prof. H.-G. Rex in Leipzig, December 1994.

During my visit in Leipzig I learned that Prof. Rex had arrived independently, earlier and in another way, at results almost identical with those of Theorems 1 and 3 here. Hence his work (unpublished so far) has the priority.

References

[1] S. M. Rump, *Solving algebraic problems with high accuracy*, in: A New Approach to Scientific Computation (U. Kulisch and W. Miranker, eds.), Academic Press, New York 1983, pp. 51-120