On the Applicability of the Interval Gaussian Algorithm

Günter Mayer * and Jiří Rohn†

Dedicated to Professor Dr. U. Kulisch, Karlsruhe, on the occasion of his 65^{th} *birthday.*

Abstract

We consider a linear interval system with a regular $n \times n$ interval matrix $[A]$ which has the form $[A] = I + [-R, R]$. For such a system we prove necessary and sufficient conditions for the applicability of the interval Gaussian algorithm where applicability means that the algorithm does not break down by dividing by an interval which contains zero. If this applicability is guaranteed we compare the output vector $[x]^G$ with the interval hull of the solution set $S = \{\tilde{x} \mid \exists \tilde{A} \in [A],\ \tilde{b} \in [b] \ :\ \tilde{A}\tilde{x} = \tilde{b}\}$. In particular, we show that in each entry of $[x]^G$ at least one of the two bounds is optimal. Linear interval systems of the above–mentioned form arise when a given general system is preconditioned with the midpoint inverse of the underlying coefficient matrix.

AMS Subject Classifications: 65F05, 65G10

Key words: Systems of linear equations, linear interval equations, preconditioned linear interval equations, Gaussian algorithm, interval Gaussian algorithm, feasibility of the interval Gaussian algorithm, interval hull, optimal enclosure.

1 Introduction

The interval Gaussian algorithm is an interval arithmetic counterpart of the well–known Gaussian algorithm for solving systems of linear equations. Starting with an $n \times n$ interval matrix $[A]$ and an interval vector $[b]$, it produces an interval vector $[x]^G = \mathrm{IGA}([A], [b])$ which contains all solutions $\tilde{x}$ of linear systems $\tilde{A}\tilde{x} = \tilde{b}$ with

*Fachbereich Mathematik, Universität Rostock, D–18051 Rostock, Germany, guenter.mayer@mathematik.uni-rostock.de

†Faculty of Mathematics and Physics, Charles University, Malostranské nám. 25, 11800 Prague, Czech Republic, and Institute of Computer Science, Academy of Sciences, Pod vodárenskou věží 2, 18207 Prague, Czech Republic, rohn@uivt.cas.cz

arbitrary $\tilde{A} \in [A]$ and $\tilde{b} \in [b]$. The set of all these solutions is called *solution set* (cf., e.g., [4], [9], [13] – [18]). We denote it by S. For a linear system with real data it is known that – from a theoretical point of view – the Gaussian algorithm is applicable without interchanging rows and columns if and only if all leading principal submatrices as defined in Section 2 are nonsingular. For the interval Gaussian algorithm such a necessary and sufficient condition is still missing. There are various sufficient conditions and also necessary and sufficient ones if one restricts the class of admissible matrices – cf. [2], [11] or [13] for an overview on such criteria known up to 1991, and [6], [7], [12] for some newer ones. In the present paper we derive necessary and sufficient conditions for the interval Gaussian algorithm to be applicable when the underlying interval matrix $[A]$ has the form

$$[A] = I + [-R, R]\,, \tag{1}$$

where I is the identity matrix. If one of these conditions holds we compare $[x]^G$ with the interval hull $[x]^S$ of the solution set S, i.e., the tightest interval enclosure of S. We show that the last components of these two vectors always coincide while the other ones generally do not. We state a necessary and sufficient condition for this situation. At the end of our paper we show a way how to generalize our results onto matrices of the form $[A] := D + [-R, R]$ where D is an arbitrary real regular diagonal matrix.

Matrices of the form (1) occur, e.g., when a general system is preconditioned by the midpoint inverse of the coefficient matrix. Note that in this case S refers to the preconditioned system and in general does not coincide with the solution set of the initial unpreconditioned one. In fact, S encloses the original solution set.

In Section 2 we start with some notations and with formulae for the interval Gaussian algorithm which we wrote down mainly for notational reasons. In Section 3 we present our results.

2 Preliminaries

By $\boldsymbol{R}^n, \boldsymbol{R}^{n\times n}, \boldsymbol{IR}, \boldsymbol{IR}^n, \boldsymbol{IR}^{n\times n}$ we denote the set of real vectors with n components, the set of real $n \times n$ matrices, the set of intervals, the set of interval vectors with n components and the set of $n \times n$ interval matrices, respectively. By *interval* we always mean a real compact interval. We write interval quantities in brackets with the exception of *point quantities* (i.e., degenerate interval quantities) which we identify with the element which they contain. Examples are the null matrix O and the identity matrix I. We use the notation $[A] = [\underline{A}, \overline{A}] = ([a]_{ij}) = ([\underline{a}_{ij}, \overline{a}_{ij}]) \in \boldsymbol{IR}^{n\times n}$ simultaneously without further reference, and we proceed similarly for the elements of $\boldsymbol{R}^n, \boldsymbol{R}^{n\times n}, \boldsymbol{IR}$ and $\boldsymbol{IR}^n$. We equip the interval spaces with the usual interval arithmetic which the reader can find, e.g., in [2] or [13]. We assume that the reader is familiar with this arithmetic. We only recall the formula $r([a] + [b]) = r[a] + r[b]$ where $r \in \boldsymbol{R}$, $[a]$, $[b] \in \boldsymbol{IR}$. It is well–known that this formula does not hold, in general, if $r \in \boldsymbol{R}$ is replaced by $[r] \in \boldsymbol{IR}$.

By $A \geq 0$ we denote a *non–negative* $n \times n$ matrix, i.e., $a_{ij} \geq 0$ for $i, j = 1, \ldots, n$. We call $x \in \boldsymbol{R}^n$ *positive* and write $x > 0$ if $x_i > 0$ for all $i = 1, \ldots, n$.

We also mention the standard notation from interval analysis ([2], [13])

$$
\begin{aligned}
\check{a} &:= \operatorname{mid}([a]) := \frac{\underline{a}+\overline{a}}{2} && (\textit{midpoint})\\
\operatorname{rad}([a]) &:= \frac{\overline{a}-\underline{a}}{2} && (\textit{radius})\\
|[a]| &:= \max\{|\tilde{a}| \,|\, \tilde{a}\in[a]\} = \max\{|\underline{a}|,|\overline{a}|\} && (\textit{absolute value})\\
\langle[a]\rangle &:= \min\{|\tilde{a}| \,|\, \tilde{a}\in[a]\} = \begin{cases} \min\{|\underline{a}|,|\overline{a}|\} & \text{if } 0\notin[a] \\ 0 & \text{otherwise}\end{cases} && (\textit{minimal absolute value})\\
q([a],[b]) &:= \max\{\,|\underline{a}-\underline{b}|, |\overline{a}-\overline{b}|\,\} && (\textit{Hausdorff distance})
\end{aligned}
$$

for intervals $[a]$, $[b]$. For $[A]\in \boldsymbol{IR}^{n\times n}$ we obtain $\check{A}=\operatorname{mid}([A])$, $\operatorname{rad}([A])$, $|[A]|\in \boldsymbol{R}^{n\times n}$ by applying $\operatorname{mid}(\,\cdot\,)$, $\operatorname{rad}(\,\cdot\,)$ and $|\,\cdot\,|$ entrywise, and we define the *comparison matrix* $\langle[A]\rangle=(c_{ij})\in \boldsymbol{IR}^{n\times n}$ by setting

$$
c_{ij} := \begin{cases} -|[a]_{ij}| & \text{if } i\neq j \\ \langle[a_{ii}]\rangle & \text{if } i=j \end{cases} .
$$

Since real quantities can be viewed as degenerate interval ones, $|\,\cdot\,|$ and $\langle\,\cdot\,\rangle$ can also be used for them.

As in [13] we call $[A]\in \boldsymbol{IR}^{n\times n}$ *regular* if $[A]$ contains only nonsingular $n\times n$ matrices, and *strongly regular* if its midpoint matrix $\check{A}$ is regular together with $\check{A}^{-1}[A]$. It is easy to prove that every strongly regular interval matrix is regular.

We call $[a]\in \boldsymbol{IR}$ *symmetric* (with respect to zero) if $[a]=-[a]$, i.e., if $[a]=[-|[a]|,|[a]|]$, and we denote by $\boldsymbol{IR}_{sym}$ the set of all symmetric intervals. The following lemma sums up some known properties of $\boldsymbol{IR}_{sym}$.

Lemma 1 *(Cf., e.g., [13])*
Let $[a]$, $[b]$, $[c]$, $[g]\in \boldsymbol{IR}$.

a) $[a]\in \boldsymbol{IR}_{sym}$ *if and only if* $\check{a}=0$.

b) For $[a]$, $[b]$, $[c]\in \boldsymbol{IR}_{sym}$ *and* $0\notin[g]$ *we get*

$$
\begin{aligned}
[a]+[b]=[a]-[b] &= [-(|[a]|+|[b]|),|[a]|+|[b]|];\\
[a]\cdot[b] &= [-|[a]|\cdot|[b]|,|[a]|\cdot|[b]|];\\
[a]/[g] &= [-|[a]|/\langle[g]\rangle,|[a]|/\langle[g]\rangle];\\
[g]\cdot[a] &= |[g]|\cdot[a];\\
[g]\cdot([a]+[b]) &= [g]\cdot[a]+[g]\cdot[b] .
\end{aligned}
$$

c) For $[a]\in \boldsymbol{IR}_{sym}$ *we get*

$$
\operatorname{mid}([g]+[a])=\operatorname{mid}([g]-[a])=\check{g} .
$$

□

This means, in particular, that $\boldsymbol{IR}_{sym}$ is closed under the arithmetic operations $+, -, *$ and that the distributive law holds in $\boldsymbol{IR}_{sym}$. It is easily seen that many of the properties in Lemma 1 transfer directly to vectors and matrices with symmetric entries. Thus $[A]$, $[G] \in \boldsymbol{IR}^{n\times n}$ with $[A] = -[A]$ implies $[G][A] = |[G]|[A]$ and $\mathrm{mid}([G] + [A]) = \check{G}$.

As in [21], pp. 19 – 20, we introduce the concept of a *directed graph* $G(A) := (\boldsymbol{X}_A, \boldsymbol{E}_A)$ associated with a matrix $A \in \boldsymbol{R}^{n\times n}$. This graph consists of the set $\boldsymbol{X}_A := \{\, i \mid i = 1, \ldots, n \,\}$ of *nodes* i and of the set $\boldsymbol{E}_A := \{\, (i,j) \mid a_{ij} \neq 0 \,\}$ of *edges* (i,j) which are ordered pairs of nodes. A sequence of edges is called a *path of length* r if it has the form $\{(i_l, i_{l+1})\}_{l=0}^{r-1}$. We will write $i_0 \to i_1 \to \cdots \to i_{r-1} \to i_r$ for this path. If there is a path $i =: i_0 \to i_1 \to \cdots \to i_r := j$ we say i is *connected* to j. If in $G(A)$ any node i can be connected to any node j then A is defined to be *irreducible*; otherwise it is called *reducible*. It is well–known (cf. [21], e.g.) that in the case $n > 1$ reducibility is equivalent to finding a permutation matrix P such that

$$PAP^T = \begin{pmatrix} A_{11} & O \\ A_{21} & A_{22} \end{pmatrix},$$

where A_{11}, A_{22} are square submatrices. Investigating the reducibility for these submatrices finally yields the so–called *reducible normal form* (cf. [21], p. 46)

$$PAP^T = \begin{pmatrix} A_{11} & & & O \\ A_{21} & A_{22} & & \\ \vdots & \vdots & \ddots & \\ A_{s1} & A_{s2} & \ldots & A_{ss} \end{pmatrix}, \tag{2}$$

where each diagonal submatrix A_{ii} is either square and irreducible, or a 1×1 null matrix.

We term $A \in \boldsymbol{R}^{n\times n}$ an *M–matrix* if $a_{ij} \leq 0$ for $i \neq j$ and if A is regular with $A^{-1} \geq 0$. An interval matrix $[A] \in \boldsymbol{IR}^{n\times n}$ is an M–matrix if it contains only M–matrices as elements. It is called an *H–matrix* if $\langle [A] \rangle$ is an M–matrix.

For $A = (a_{ij}) \in \boldsymbol{R}^{n\times n}$ the *k–th leading principal submatrix* A_k is defined by

$$A_k := \begin{pmatrix} a_{11} & \ldots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \ldots & a_{kk} \end{pmatrix} \in \boldsymbol{R}^{k\times k},$$

the *spectral radius* $\rho(A)$ is given by $\rho(A) := \max\{\, |\lambda| \mid \lambda \text{ eigenvalue of } A \,\}$ and the *row sum norm* $\|A\|_\infty$ is defined by $\|A\|_\infty := \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|$.

For a given interval matrix $[A] \in \boldsymbol{IR}^{n\times n}$ and a given interval vector $[b] \in \boldsymbol{IR}^n$ the *interval Gaussian algorithm* as described in [2] reads as follows:

First, we compute, consequently, $[A]^{(k)} \in \boldsymbol{IR}^{n\times n}$ and $[b]^{(k)} \in \boldsymbol{IR}^n$, $k = 1, \ldots, n$, by using the formulae

$$[A]^{(1)} := [A],\ [b]^{(1)} := [b],$$

$$[a]_{ij}^{(k+1)} := \begin{cases} [a]_{ij}^{(k)} & i = 1, \ldots, k;\ j = 1, \ldots, n, \\ [a]_{ij}^{(k)} - \dfrac{[a]_{ik}^{(k)} \cdot [a]_{kj}^{(k)}}{[a]_{kk}^{(k)}} & i = k+1, \ldots, n;\ j = k+1, \ldots, n, \\ 0 & \text{for all other pairs } (i,j), \end{cases}$$

$$[b]_i^{(k+1)} := \begin{cases} [b]_i^{(k)} & i = 1, \ldots, k, \\ [b]_i^{(k)} - \dfrac{[a]_{ik}^{(k)}}{[a]_{kk}^{(k)}} \cdot [b]_k^{(k)} & i = k+1, \ldots, n, \end{cases}$$

$$k = 1, \ldots, n-1\,.$$

When we know $[A]^{(n)}$, $[b]^{(n)}$, we compute, consequently, the components $[x]_i^G$, $i = n, n-1, \ldots, 1$, of the interval vector $[x]^G = \mathrm{IGA}([A],[b])$ by using the formulae

$$\begin{aligned} [x]_n^G &:= [b]_n^{(n)} / [a]_{nn}^{(n)}, \\ [x]_i^G &:= \left([b]_i^{(n)} - \sum_{j=i+1}^{n} [a]_{ij}^{(n)} [x]_j^G \right) / [a]_{ii}^{(n)}, \quad i = n-1, n-2, \ldots, 1. \end{aligned}$$

Note that $[x]^G$ is defined without permuting rows or columns. The construction of $[x]^G = \mathrm{IGA}([A],[b])$ is called the interval Gaussian algorithm or, shortly, IGA. It is applicable (for $[A] \in \boldsymbol{IR}^{n\times n}$ and for any $[b] \in \boldsymbol{IR}^n$) if and only if $0 \notin [a]_{kk}^{(k)}$, $k = 1, \ldots, n$, which, by the definition of the matrices $A^{(k)}$, is equivalent to $0 \notin [a]_{kk}^{(n)}$, $k = 1, \ldots, n$. It is easy to see that the IGA reduces to the ordinary Gaussian algorithm if $[A]$ and $[b]$ are point quantities. As usual, we speak of the IGA *with partial pivoting* if either for $k = 1, \ldots, n-1$ two of the rows $k, k+1, \ldots, n$ are interchanged in $[A]^{(k)}$ such that $|[a]_{kk}^{(k)}| = \max\{\, |[a]_{ik}^{(k)}| \mid k \le i \le n,\ 0 \notin [a]_{ik}^{(k)} \,\}$ or for $k = 1, \ldots, n-1$ the corresponding columns are permuted such that $|[a]_{kk}^{(k)}| = \max\{\, |[a]_{kj}^{(k)}| \mid k \le j \le n,\ 0 \notin [a]_{kj}^{(k)} \,\}$.

In Section 3 we will apply the IGA also to the preconditioned interval matrix $\check{A}^{-1}[A]$ and to the corresponding right–hand side $\check{A}^{-1}[b]$ instead of $[A]$ and $[b]$, respectively; i.e., we will compute $[x]_{prec}^G := \mathrm{IGA}(\check{A}^{-1}[A], \check{A}^{-1}[b])$. We will call the construction of $[x]_{prec}^G$ *preconditioned interval Gaussian algorithm* or, shortly, PIGA. Analogously, we define PIGA *with partial pivoting.*

3 Results

We start this section with the announced necessary and sufficient conditions for the applicability of the IGA for matrices of the form (1).

Theorem 1

Let $[A] := I + [-R, R] \in \boldsymbol{IR}^{n\times n}$, $[b] \in \boldsymbol{IR}^n$. *Then the following conditions are equivalent.*

a) IGA *is applicable, i.e.,* $[x]^G = IGA([A],[b])$ *exists,*

b) IGA *with partial pivoting is applicable,*

c) $I - R$ *is an* M*–matrix,*

d) $\rho(R) < 1$,

e) $[A]$ *is regular,*

f) $[A]$ *is strongly regular,*

g) $[A]$ *is an* H*–matrix.*

Proof

By Proposition 4.1.1 in [13], d), e), f) and g) are equivalent.

c) $\Longleftrightarrow$ d) follows from Proposition 3.6.3 (iii) in [13].

a) $\Longrightarrow$ e) is easy to see since none of the diagonal entries $[a]_{kk}^{(n)} = [a]_{kk}^{(k)}$ contains zero. Therefore, the Gaussian algorithm is applicable for any $\tilde{A} \in [A]$ which means that $\tilde{A}$ is nonsingular. Hence $[A]$ is regular.

g) $\Longrightarrow$ a) is a result of Alefeld stated in [1]; cf. also Theorem 4.5.7 in [13].

a) $\Longrightarrow$ b)

Start the IGA with $[A]^{(1)} := [A]$. Then $\mathrm{mid}([A]^{(1)}) = I$, and an inductive argument based on Lemma 1 yields $\mathrm{mid}([A]^{(k)}) = I,\ k = 1, \ldots, n$, i.e., $0 \in [a]_{ij}^{(k)}$ for $i \neq j$. This shows that partial pivoting necessarily results in the original IGA.

b) $\Longrightarrow$ e)

Let P be the permutation matrix which describes column pivoting. Then, by assumption, IGA is applicable for $P[A]$, hence this matrix is regular. But then $[A]$ is regular, too. The proof performs analogously for row pivoting.

□

For the important case of the preconditioned IGA we will restate Theorem 1 as the subsequent corollary. To this end we mention the representation

$$\check{A}^{-1}[A] = \check{A}^{-1}(\check{A} + [-\mathrm{rad}([A]), \mathrm{rad}([A])]) = I + [-|\check{A}^{-1}|\mathrm{rad}([A]), |\check{A}^{-1}|\mathrm{rad}([A])] \ .$$

Corollary 1

Let $[A] \in \boldsymbol{IR}^{n\times n}$, $[b] \in \boldsymbol{IR}^n$ *and let* $\check{A}^{-1}$ *exist. Then the following conditions are equivalent.*

a) PIGA is applicable, i.e., $[x]_{prec}^{G} = IGA(\check{A}^{-1}[A], \check{A}^{-1}[b])$ *exists,*

b) PIGA with partial pivoting is applicable,

c) $I - |\check{A}^{-1}|\mathrm{rad}([A])$ *is an* M*–matrix,*

d) $\rho(|\check{A}^{-1}|\mathrm{rad}([A])) < 1$,

e) $\check{A}^{-1}[A]$ *is regular,*

f) $[A]$ *is strongly regular,*

g) $\check{A}^{-1}[A]$ *is an* H*–matrix.*

□

Here we have slightly modified condition f) of Theorem 1 by applying e) and the definition of strong regularity. Note that a similar condition as in Corollary 1 d) was stated in [13], p. 166, using the so–called Gauss inverse. This condition is sufficient, but not necessary for the applicability of the PIGA.

In order to prove parts of our subsequent theorems we need some auxiliary results which we summarize in Lemma 2.

Lemma 2

Let $C := I - R \in \boldsymbol{R}^{n\times n}$, $R \geq 0$, $\rho(R) < 1$, *and let* $M := C^{-1}$. *Then the following assertions hold.*

a) The Gaussian algorithm is applicable for C*; all matrices* $C^{(k)}$, $k = 1, \ldots, n$, *are* M*–matrices.*

b) For arbitrary $i, j \in \{1, \ldots, n\}$ *we have* $m_{ij} \neq 0$ *if and only if* i *is connected to* j *in the graph* $G(C)$.

c) Let $C = LU$, L *lower triangular with ones in the diagonal,* U *upper triangular. Then* $U = C^{(n)}$, *and* $c_{ij}^{(k)} \neq 0$ *for* $i \neq j$ *and* $k \leq m := \min\{i, j\}$ *if and only if* i *is connected to* j *in the graph* $G(C)$ *such that all intermediate nodes* i_s *in the corresponding path satisfy* $i_s < \min\{i, j, k\}$.

In particular, $l_{ij} \neq 0$ *for* $i > j$ *if and only if* i *is connected to* j *in* $G(C)$ *such that all intermediate nodes* i_s *in the corresponding path satisfy* $i_s < j$.

Proof

a) The first part of the assertion follows from Theorem 1; in particular, C is an M–matrix. As in the proof of Theorem 3 in [2], pp. 186 f, or with Lemma 4.5.6 in [13] one can see that $C^{(k)}$, $k = 1, \ldots, n$, are M–matrices, too.

b) ¿From

$$M = (I - R)^{-1} = \sum_{p=0}^{\infty} R^p$$

and from the nonnegativity of R we get $m_{ii} \geq 1$. Since C is an M–matrix it also has positive diagonal entries. Therefore, the assertion is true for $i = j$. Assume now $i \neq j$. Then $m_{ij} \neq 0$ if and only if $(R^p)_{ij} > 0$ for at least one $p \in \boldsymbol{N}$. For $p > 1$ this holds if and only if in the representation

$$(R^p)_{ij} = \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \ldots \sum_{i_{p-1}=1}^{n} r_{i,i_1} \cdot r_{i_1,i_2} \cdot \ldots \cdot r_{i_{p-1},j}$$

at least one summand differs from zero. The corresponding indices determine the path required in the assertion, and vice versa. For $p = 1$ the path is $i \to j$.

c) The equality $U = C^{(n)}$ can be found, e.g., in [8], § 3.2.5.

$\Longrightarrow$:

Let $c_{ij}^{(k)} \neq 0$ hold with $k \leq m$. Then in $G(C^{(k)})$ there exists the path $i \to j$. If $m = 1$, this implies the assertion. If $m > 1$, we see from the formula

$$c_{ij}^{(k)} = c_{ij}^{(k-1)} - \frac{c_{i,k-1}^{(k-1)} c_{k-1,j}^{(k-1)}}{c_{k-1,k-1}^{(k-1)}} \tag{3}$$

and from the signs of the entries of the M–matrix $C^{(k-1)}$ that $c_{ij}^{(k)} \neq 0$ if and only if $c_{ij}^{(k-1)} \neq 0$ or both $c_{i,k-1}^{(k-1)},\ c_{k-1,j}^{(k-1)} \neq 0$. Therefore, in $G(C^{(k-1)})$ we obtain the path $i \to j$ or $i \to (k-1) \to j$. Repeat the arguments while the upper index goes down to 1. This yields the path in $G(C^{(1)}) = G(C)$ as it was asserted.

$\Longleftarrow$:

Let the path of the assertion exist. Without loss of generality assume that it contains none of the nodes twice. (Otherwise cut off the piece between the two equal nodes including one of them.) We proceed by induction on the length p of the path. If $p = 1$ then the path reads $i \to j$, hence $c_{ij} = c_{ij}^{(1)} \neq 0$. Using the arguments following (3) we obtain $c_{ij}^{(k)} \neq 0$ for all $k \leq m$. Assume now that the assertion is true for all pairs of indices for which the corresponding path in $G(C)$ has a length which is less than or equal to some p. Let i, j be connected by a path of length $p+1$ and let i_s be the largest node among all intermediate nodes of this path. Then by the hypothesis of the induction $c_{i,i_s}^{(i_s)} \neq 0$ and $c_{i_s,j}^{(i_s)} \neq 0$ whence $c_{ij}^{(i_s+1)} \neq 0$ by (3). As above we get $c_{ij}^{(k)} \neq 0$ for all k with $i_{s+1} \leq k \leq m$.

The assertion for l_{ij} follows immediately, since $l_{ij} = \dfrac{c_{ij}^{(j)}}{c_{jj}^{(j)}} \neq 0$ if and only if $c_{ij}^{(j)} \neq 0$.

□

Our next lemma provides explicit formulae for the quantities arising in the IGA. To this end we will define a diagonal matrix $D = \mathrm{diag}(d_1, \ldots, d_n)$ with $|D| = I$ such that $|\check{b}| = D\check{b}$ whence

$$\max(d_i[b]_i + [a]) = |d_i[b]_i + [a]| = |[b]_i| + |[a]|$$

for any symmetric interval $[a]$.

Lemma 3

Let the IGA be applicable for $[A] = I + [-R, R] \in \boldsymbol{IR}^{n\times n}$ and $[b] \in \boldsymbol{IR}^n$, and apply it also to the real matrix $C := I - R$ and the right–hand side $w := |\check{b}| + \mathrm{rad}([b])$. Define the diagonal matrix $D = \mathrm{diag}(d_1, \ldots, d_n) \in \boldsymbol{R}^{n\times n}$ by

$$d_i := \begin{cases} 1 & \text{if } \check{b}_i \geq 0\,, \\ -1 & \text{if } \check{b}_i < 0\,. \end{cases}$$

Then the quantities of the IGA *can be represented as follows:*

$$
\begin{aligned}
[A]^{(k)} &= [C^{(k)}, 2I - C^{(k)}] = I + [-(I - C^{(k)}), I - C^{(k)}] && (4)\\
\langle [A]^{(k)} \rangle &= C^{(k)} && (5)\\
[b]^{(k)} &= D[2|\check{b}| - w^{(k)}, w^{(k)}] = D\left(|\check{b}| + [-(w^{(k)} - |\check{b}|), w^{(k)} - |\check{b}|]\right)\\
&= \check{b} + [-(w^{(k)} - |\check{b}|), w^{(k)} - |\check{b}|] && (6)\\
\max(D[b]^{(k)}) &= |[b]^{(k)}| = w^{(k)} \geq 0 && (7)\\
\max(D[x]^G) &= |[x]^G| = C^{-1}w =: x^* \geq 0 && (8)
\end{aligned}
$$

Proof

Lemma 2 guarantees that $C^{(k)}$, $k = 1, \ldots, n$, are M–matrices. By inspecting the formulae for the IGA one easily sees that all the vectors $w^{(k)}$ are nonnegative.

We first prove (4) and (5). For $k = 1$ these formulae are trivial. Let them hold for some $k < n$ and choose i, $j > k$. Then $\underline{A}^{(k)} = C^{(k)}$ and $\langle [A]^{(k)} \rangle = C^{(k)}$ by assumption, and

$$
\begin{aligned}
\underline{a}_{ij}^{(k+1)} &= \underline{a}_{ij}^{(k)} - \max\left(\frac{[a]_{ik}^{(k)}[a]_{kj}^{(k)}}{[a]_{kk}^{(k)}}\right) = c_{ij}^{(k)} - \frac{|[a]_{ik}^{(k)}||[a]_{kj}^{(k)}|}{\langle [a]_{kk}^{(k)} \rangle}\\
&= c_{ij}^{(k)} - \frac{c_{ik}^{(k)} c_{kj}^{(k)}}{c_{kk}^{(k)}} = c_{ij}^{(k+1)}\,.
\end{aligned}
$$

Hence $\underline{A}^{(k+1)} = C^{(k+1)}$. Since $\check{A}^{(k)} = I$ the same holds for $[A]^{(k+1)}$ as can be seen from Lemma 1. Therefore, $\mathrm{rad}([A]^{(k+1)}) = I - C^{(k+1)}$ which yields (4) for $k + 1$. From $[a]_{ij}^{(k+1)} = [c_{ij}^{(k+1)}, -c_{ij}^{(k+1)}]$ for $i \neq j$, and from $\langle [a]_{ii}^{(k+1)} \rangle = c_{ii}^{(k+1)}$ we get (5).

We now prove (6) and (7) which certainly are trivial if $k = 1$. Let them hold for some $k < n$. Then $\max(D[b]^{(k)}) = w^{(k)} = |D[b]^{(k)}| = |[b]^{(k)}|$ by assumption and by $|D| = I$. This implies

$$
\begin{aligned}
\max(d_i[b]_i^{(k+1)}) &= \max(d_i[b]_i^{(k)}) - \min\left(d_i \frac{[a]_{ik}^{(k)}}{[a]_{kk}^{(k)}}[b]_k^{(k)}\right) = w_i^{(k)} - \min\left(\frac{[a]_{ik}^{(k)}}{\langle [a]_{kk}^{(k)} \rangle}|[b]_k^{(k)}|\right)\\
&= w_i^{(k)} - \min\left(\frac{[a]_{ik}^{(k)}}{\langle [a]_{kk}^{(k)} \rangle} w_k^{(k)}\right) = w_i^{(k)} + \frac{|[a]_{ik}^{(k)}|}{\langle [a]_{kk}^{(k)} \rangle} w_k^{(k)}\\
&= w_i^{(k)} - \frac{c_{ik}^{(k)}}{c_{kk}^{(k)}} w_k^{(k)} = w_i^{(k+1)},
\end{aligned}
$$

where we exploited the symmetry of $[a]_{ik}^{(k)}$ and (4). With Lemma 1 one sees that $\mathrm{mid}(d_i[b]_i^{(k+1)}) = \mathrm{mid}(d_i[b]_i^{(k)}) = |\check{b}_i|$ whence $\mathrm{rad}([b]_i^{(k+1)}) = \mathrm{rad}(d_i[b]_i^{(k+1)}) = w_i^{(k+1)} - |\check{b}_i|$. This yields (6) and (7) for $k + 1$.

Now we address ourselves to (8). For $k = n$ we get

$$
\max(d_n[x]_n^G) = |d_n[b]_n^{(n)}| / \langle [a]_{nn}^{(n)} \rangle = |[b]_n^{(n)}| / \langle [a]_{nn}^{(n)} \rangle
$$

which equals $|[x]_n^G|$ as well as $w_n^{(n)}/c_{nn}^{(n)} = x_n^*$. Here we used (5) and (7). Assume now that $\max(d_j[x]_j^G) = |[x]_j^G| = x_j^*$ holds for $j = n, n-1, \ldots, i+1$. Then

$$\begin{aligned}
\max(d_i[x]_i^G) &= \max\left(\left(d_i[b]_i^{(n)} - \sum_{j=i+1}^{n} d_i[a]_{ij}^{(n)}[x]_j^G\right)/[a]_{ii}^{(n)}\right) \\
&= \max\left(\left(d_i[b]_i^{(n)} - \sum_{j=i+1}^{n} [a]_{ij}^{(n)}|[x]_j^G|\right)/[a]_{ii}^{(n)}\right) \\
&= \left(|d_i[b]_i^{(n)}| + \sum_{j=i+1}^{n} |[a]_{ij}^{(n)}||[x]_j^G|\right)/\langle[a]_{ii}^{(n)}\rangle \\
&= \left(w_i^{(n)} - \sum_{j=i+1}^{n} c_{ij}^{(n)} x_j^*\right)/c_{ii}^{(n)} = x_i^* \ ,
\end{aligned}$$

where we exploited the symmetry of $[a]_{ij}^{(n)}$. Note that the next to the last formula equals $|d_i[x]_i^G| = |[x]_i^G|$.

□

Based on Lemma 3 the question arises quite naturally whether $[x]^G$ also has a simple representation. The answer is contained in our next theorem which shows that for this purpose it is sufficient to solve the single real system $Cx = w$ by means of the Gaussian algorithm.

Theorem 2

Let the IGA be applicable for $[A] := I + [-R, R] \in \boldsymbol{IR}^{n\times n}$ *and* $[b] \in \boldsymbol{IR}^n$, *and let* $C := I - R$, $w := |\check{b}| + \mathrm{rad}([b])$, $x^* := C^{-1}w$. *With* $C^{(n)}$ *from the IGA define* $f_i := 1/c_{ii}^{(n)}$, $i = 1, \ldots, n$. *Then for each* $i \in \{1, \ldots, n\}$ *we have*

$$\underline{x}_i^G = \min\{\underset{\sim}{x}_i, \mu_i \underset{\sim}{x}_i\}, \qquad \overline{x}_i^G = \max\{\tilde{x}_i, \mu_i \tilde{x}_i\}, \tag{9}$$

where

$$\underset{\sim}{x}_i := -x_i^* + f_i(\check{b} + |\check{b}|)_i, \quad \tilde{x}_i := x_i^* + f_i(\check{b} - |\check{b}|)_i, \tag{10}$$

and

$$\mu_i := \frac{1}{2f_i - 1} \in (0, 1] \ .$$

Proof

¿From (4) we get $c_{ii}^{(n)} \le 2 - c_{ii}^{(n)}$. Hence $c_{ii}^{(n)} \le 1$, $2f_i - 1 \ge 1 > 0$ and $0 < \mu_i \le 1$.

Let

$$[z]_i := [b]_i^{(n)} - \sum_{j=i+1}^{n} [a]_{ij}^{(n)}[x]_j^G \ .$$

With Lemma 3 we obtain

$$[z]_i = [b]_i^{(n)} - \sum_{j=i+1}^{n} [a]_{ij}^{(n)} |[x]_j^G| = [b]_i^{(n)} - \sum_{j=i+1}^{n} [a]_{ij}^{(n)} x_j^*$$

and

$$\overline{z}_i = \overline{b}_i^{(n)} - \sum_{j=i+1}^{n} c_{ij}^{(n)} x_j^* = \check{b}_i - |\check{b}_i| + w_i^{(n)} - \sum_{j=i+1}^{n} c_{ij}^{(n)} x_j^* = \check{b}_i - |\check{b}_i| + c_{ii}^{(n)} x_i^* = c_{ii}^{(n)} \tilde{x}_i \, .$$

In particular, $\mathrm{sign}(\overline{z}_i) = \mathrm{sign}(\tilde{x}_i)$.

If $\overline{z}_i \geq 0$ then $\tilde{x}_i \geq 0$ and

$$\overline{x}_i^G = \frac{\overline{z}_i}{\underline{a}_{ii}^{(n)}} = \frac{\overline{z}_i}{c_{ii}^{(n)}} = f_i(\check{b}_i - |\check{b}_i|) + x_i^* = \tilde{x}_i \geq \mu_i \tilde{x}_i \, .$$

If $\overline{z}_i < 0$ then $\tilde{x}_i < 0$ and by (4) we get

$$\overline{x}_i^G = \frac{\overline{z}_i}{\overline{a}_{ii}^{(n)}} = \frac{\overline{z}_i}{2 - c_{ii}^{(n)}} = \mu_i \frac{\overline{z}_i}{c_{ii}^{(n)}} = \mu_i \tilde{x}_i \geq \tilde{x}_i \, .$$

This proves the second equality in (9).

In order to prove the first one we replace $[b]$ by $-[b]$. Then $[y]^G := \mathrm{IGA}([A], -[b]) = -\mathrm{IGA}([A], [b]) = -[x]^G$. Applying the second equality of (9) to $\overline{y}_i^G$ yields $\underline{x}_i^G = -\overline{y}_i^G = -\max\{\tilde{y}_i, \mu_i \tilde{y}_i\} = \min\{-\tilde{y}_i, -\mu_i \tilde{y}_i\}$ with $\tilde{y}_i := x_i^* + f_i(-\check{b}_i - |\check{b}_i|) = -\underset{\sim}{x}_i$.

□

In combination with Lemma 3 Theorem 2 shows that for matrices of the form (1) and arbitrary right–hand sides $[b]$ IGA and PIGA can be performed without using interval arithmetic.

In addition, this theorem exhibits a remarkable analogy to the following theorem of Hansen and Rohn.

Theorem 3 *([9], [18])*

Let $[A] := I + [-R, R] \in \boldsymbol{IR}^{n \times n}$, $[b] \in \boldsymbol{IR}^n$, $\rho(R) < 1$, $M := (I - R)^{-1}$. *Denote by* $[x]^S$ *the interval hull of the solution set* $S := \{x \mid \exists \tilde{A} \in [A],\ \tilde{b} \in [b] :\ \tilde{A}x = \tilde{b}\}$. *Then for each* $i \in \{1, \ldots, n\}$ *we have*

$$\underline{x}_i^S := \min\{\underset{\sim}{x}_i^S, \nu_i \underset{\sim}{x}_i^S\}, \qquad \overline{x}_i^S := \max\{\tilde{x}_i^S, \nu_i \tilde{x}_i^S\}, \tag{11}$$

where

$$\underset{\sim}{x}_i^S := -x_i^* + m_{ii}(\check{b} + |\check{b}|)_i, \quad \tilde{x}_i^S := x_i^* + m_{ii}(\check{b} - |\check{b}|)_i, \quad x_i^* := \Big(M(|\check{b}| + \mathrm{rad}([b])) \Big)_i \tag{12}$$

and

$$\nu_i := \frac{1}{2m_{ii} - 1} \in (0, 1] \, .$$

□

Note that x^* is the same vector in both theorems since $M = C^{-1}$. By these theorems it is easily seen that $f_i = m_{ii}$ implies $[x]_i^G = [x]_i^S$. The following result is essentially based on this fact. It shows that at least one bound of each component $[x]_i^G$ coincides with the corresponding bound of $[x]_i^S$.

Theorem 4

The assumptions of Theorem 2 imply

$$[x]_n^G = [x]_n^S \ . \tag{13}$$

Moreover, for any $i \in \{1, \ldots, n\}$ we get

$$\overline{x}_i^G = \overline{x}_i^S = x_i^* \ \textit{if} \ \check{b}_i \geq 0, \ \textit{and} \ \underline{x}_i^G = \underline{x}_i^S = -x_i^* \ \textit{if} \ \check{b}_i \leq 0 \ .$$

In particular, with x^* from Theorem 2 the equality $[x]_i^G = [x]_i^S = [-x_i^*, x_i^*]$ holds if $\check{b}_i = 0$.

Proof

With the notation of the two preceding theorems the last column of M can be written as $y := C^{-1}e^{(n)}$ where $e^{(n)}$ denotes the n^{th} column of I. By Cramer's rule $y_n = \det(C')/\det(C)$ with C' being the $(n-1) \times (n-1)$ leading principal submatrix of C. Since $\det(C) = c_{11}^{(n)} \cdot c_{22}^{(n)} \cdot \ldots \cdot c_{nn}^{(n)} = \det(C') \cdot c_{nn}^{(n)}$ one gets $f_{nn} = m_{nn}$. Therefore, (13) holds. The remaining part of the theorem follows directly from the Theorems 2 and 3 with $\tilde{x}_i^S = x_i^* = \tilde{x}_i \geq 0$ if $\check{b}_i \geq 0$ and $\underset{\sim}{x}_i^S = -x_i^* = \underset{\sim}{x}_i \leq 0$ if $\check{b}_i \leq 0$ which implies $\overline{x}_i^G = \overline{x}_i^S = x_i^*$ if $\check{b}_i \geq 0$ and $\underline{x}_i^G = \underline{x}_i^S = -x_i^*$ if $\check{b}_i \leq 0$.

□

A. Neumaier [15] remarked that the 'Moreover'–part of the preceding theorem can also be proved using the Theorems 4.4.8, 4.4.10 and 4.5.11 in [13] if one applies the first of these theorems to the M–matrix $\underline{A} = C$ and if one takes into account the inclusion $\mathrm{IGA}(\underline{A}, [b]) \subseteq \mathrm{IGA}([A], [b])$ and $\underline{A}^F[b] \subseteq [A]^F[b]$. Here, $[A]^F[b]$ denotes the limit of the Jacobi iteration for $[A]$ and $[b]$. We leave the details to the reader.

By Theorem 4 it is easily seen that

$$[x]_i^S = \Big(\mathrm{IGA}(P[A]P^T, P[b])\Big)_n = \Big(P^T \, \mathrm{IGA}(P[A]P^T, P[b])\Big)_i$$

if P is a permutation matrix which effects an exchange of the rows/columns i and n. Note that the assumptions of the Theorems 1 – 4 remain true when rows and corresponding columns are permuted.

Unfortunately, $m_{ii} = f_i, \ i = 1, \ldots, n-1$, does not hold, in general. Hence $[x]^G \neq [x]^S$, i.e., the enclosure of S by $[x]^G$ is not optimal. This is even true in the 2×2 case and, therefore, for tridiagonal matrices, as the following example shows.

Example 1

Let

$$[A] := \begin{pmatrix} 1 & [-1,1] \\ [-\frac{1}{2},\frac{1}{2}] & 1 \end{pmatrix} , \qquad [b] := \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \check{b} .$$

Then

$$C = \begin{pmatrix} 1 & -1 \\ -\frac{1}{2} & 1 \end{pmatrix} \qquad C^{(n)} = \begin{pmatrix} 1 & -1 \\ 0 & \frac{1}{2} \end{pmatrix}$$

$$M = C^{-1} = \begin{pmatrix} 2 & 2 \\ 1 & 2 \end{pmatrix} \qquad x^* = M|\check{b}| = \begin{pmatrix} 4 \\ 3 \end{pmatrix}$$

$$[x]^G = \begin{pmatrix} [-4,2] \\ [\frac{1}{3},3] \end{pmatrix} \neq [x]^S = \begin{pmatrix} [-4,0] \\ [\frac{1}{3},3] \end{pmatrix} .$$

With $P := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ we get

$$P^T \,\mathrm{IGA}\left(P[A]P^T, P[b]\right) = \begin{pmatrix} [-4,0] \\ [-1,3] \end{pmatrix} ,$$

where the first component is now optimal.

□

Our next theorem completes Theorem 4. Reformulating it appropriately it fits into the class of optimality results for the interval hull of the solution set S given in [3], [4], [9], [10], [16] and [18].

Theorem 5

The assumptions of Theorem 2 imply

$$[x]_i^G \neq [x]_i^S \tag{14}$$

if and only if the following two properties (i), (ii) or, equivalently (i), (iii) hold:

(i) $\check{b}_i \neq 0$*;*

(ii) using the matrices L from the LU decomposition of C and $M := (I - R)^{-1}$ there is an index $k > i$ such that $m_{ik} \neq 0$ and $l_{ki} \neq 0$;

(iii) there is an index $k > i$ such that the node i is connected to k in the graph $G(C)$, and, vice versa, the node k is connected to i, where in this latter case the intermediate nodes i_j of the path have to satisfy $i_j < i$.

Proof

We use again the notations from the Theorems 2 and 3. By these theorems the inequality (14) is equivalent to (i) and $m_{ii} \neq f_i$. In order to derive an equivalent

condition for $m_{ii} - f_i = (M - (C^{(n)})^{-1})_{ii} \neq 0$ let $C = LU = LC^{(n)}$ be the LU decomposition of C. Then $(C^{(n)})^{-1} = C^{-1}L = ML$ implies

$$M - (C^{(n)})^{-1} = M(I - L) \geq 0 \ . \tag{15}$$

Hence $(M - (C^{(n)})^{-1})_{ii} \neq 0$ if and only if there is an index $k > i$ such that $m_{ik} \neq 0$ and $l_{ki} \neq 0$. Here, we exploited $M \geq 0,\ I - L \geq 0$; $k > i$ is required since $I - L$ is a strictly lower triangular matrix. This proves the equivalence of (14) with (i), (ii). Since by Lemma 2 (ii) is equivalent to (iii) Theorem 5 is proved.

□

Note that property (ii) can certainly never be fulfilled if $i = n$. This confirms (13).

Example 1 can be used to illustrate the preceding theorem. For $i = 1$ the paths in (iii) necessarily are $1 \to 2$ and $2 \to 1$, where no intermediate nodes occur.

Theorem 5 shows that $[x]^S = [x]^G$ is true for 2×2 matrices of the form (1) if and only if $\check{b}_1 = 0$ or $[a]_{12} = 0$ or $[a]_{21} = 0$. This is equivalent to '$\check{b}_1 = 0$ or C is reducible'. In particular, if $\check{b}_1 \neq 0$ and if C is irreducible then $[x]_1^G \neq [x]_1^S$ holds. This can be generalized to $n \times n$ matrices of the form (1).

Corollary 2

Let the assumptions of Theorem 2 hold with $n > 1$, and let C be irreducible. Then for an arbitrary $i < n$

$$[x]_i^G = [x]_i^S \ \textit{if and only if}\ \check{b}_i = 0 \ . \tag{16}$$

Proof

Since C is irreducible and $i < n$ there are two paths in the graph $G(C)$ which connect i to $i+1$ and $i+1$ to i, respectively. Concatenate these paths to end up with $i \to \ldots \to i+1 \to \ldots \to i$. Trace back this path starting with the right node i until you find the first node, say i_0, which is larger than i. (At latest, $i+1$ is such a node.) Then (iii) of Theorem 5 is fulfilled with $k = i_0$, i.e., (iii) always holds if C is irreducible. Thus (14) is equivalent to (i) which proves (16).

□

Corollary 2 shows that for matrices of the form (1) the IGA often overestimates the interval hull of S, a phenomenon which is well–known in the literature. (Cf. [13], p. 160 ff, [14] or [20], e.g.)

For reducible matrices the following corollary clarifies the situation. It generalizes (13).

Corollary 3

a) Let the assumptions of Theorem 2 hold with $n > 1$, and let P be the permutation matrix such that $\hat{C} = (\hat{C}_{ij}) := PCP^T$ is the reducible normal form of C where $\hat{C}_{ij}$ denote the submatrices of $\hat{C}$ according to (2). Let π be the permutation associated with P and let i_k be the largest row number in C such that $\pi(i_k)$ belongs to the row numbers of $\hat{C}_{kk}$ (in $\hat{C}$). Then $[x]_{i_k}^G = [x]_{i_k}^S$.

b) Let $[A]$ be a triangular matrix of the form (1) with $0 \notin [a]_{ii},\ i = 1, \ldots, n$. Then the IGA *is applicable and $[x]^G = [x]^S$.*

Proof

a) If $[x]^G_{i_k} \neq [x]^S_{i_k}$ then according to Theorem 5 there is a node $s > i_k$ such that i_k is connected to s and vice versa. Hence $\pi(s)$ belongs to the row numbers of $\hat{C}_{kk}$ (in $\hat{C}$) which implies the contradiction $s \leq i_k$.

b) The applicability of the IGA follows from Theorem 1 d) since R is triangular and $r_{ii} < 1$ for each i due to $0 \notin [a]_{ii}$. Since C is already in reducible normal form (eventually after reflection on the counterdiagonal) and since the corresponding diagonal submatrices are 1×1, the assertion follows directly from a).

□

The result in Corollary 3 b) shows that $[x]^G$ is optimal for triangular systems of the form (1). It is not always optimal for interval systems which first have to be brought to the form (1) by preconditioning when $[x]^G$ is compared with the solution set of the *initial* system. An example which illustrates this situation by a 4×4 matrix can be found in [19].

The matrix of the subsequent example satisfies the assumptions of Corollary 3 .

Example 2

Let

$$[\hat{A}] := \begin{pmatrix} [A] & O \\ [B] & [A] \end{pmatrix}$$

with

$$[A] := \begin{pmatrix} 1 & [-1,1] \\ [-\frac{1}{2}, \frac{1}{2}] & 1 \end{pmatrix} \text{ as in Example 1 and with } [B] := \begin{pmatrix} [-1,1] & [-1,1] \\ [-1,1] & [-1,1] \end{pmatrix}.$$

Furthermore, let $[b] := (-1, 1, -1, 1)^T = \check{b}$. Then $[\hat{A}]$ has the form (1) and the matrix

$$\hat{C} := \langle [\hat{A}] \rangle = \begin{pmatrix} C & O \\ \underline{B} & C \end{pmatrix}$$

is already in reducible normal form where

$$C := \begin{pmatrix} 1 & -1 \\ -\frac{1}{2} & 1 \end{pmatrix} \qquad \text{(cf. the previous example)}.$$

Moreover,

$$\hat{M} := \hat{C}^{-1} = \begin{pmatrix} M & O \\ N & M \end{pmatrix} \text{ where } M := \begin{pmatrix} 2 & 2 \\ 1 & 2 \end{pmatrix} \text{ and } N := \begin{pmatrix} 12 & 16 \\ 9 & 12 \end{pmatrix}.$$

Obviously, $\hat{M}$ is nonnegative which shows that $\hat{C}$ is an M–matrix, i.e., due to Theorem 1 the assumptions of the Theorems 2 and 3 are fulfilled. Now $x^* := \hat{M}|\check{b}| = (4,3,32,24)^T$, and from Example 1 we easily deduce

$$\hat{c}_{ii}^{(n)} = c_{ii}^{(n)} = \begin{cases} 1 & \text{if } i \in \{1,3\} \\ \frac{1}{2} & \text{if } i \in \{2,4\} \end{cases} \qquad \text{whence} \quad f_i := \frac{1}{\hat{c}_{ii}^{(n)}} = \begin{cases} 1 & \text{if } i \in \{1,3\} \\ 2 & \text{if } i \in \{2,4\} \end{cases} .$$

With the notation of the Theorems 2 and 3 and with $m_{ii} = 2$ we get for $i = 1, \ldots, 4$

$$\mu_i = \begin{cases} 1 & \text{if } i \in \{1,3\} \\ \frac{1}{3} & \text{if } i \in \{2,4\} \end{cases} , \qquad \nu_i = \frac{1}{3} ,$$

and

$$\underset{\sim}{x}_i = (-4,1,-32,-20)^T, \qquad \tilde{x}_i = (2,3,30,24)^T,$$

$$\underset{\sim}{x}_i^S = (-4,1,-32,-20)^T, \qquad \tilde{x}_i^S = (0,3,28,24)^T,$$

whence

$$\begin{aligned} [x]^G &= \left([-4,2]\,,\,[\tfrac{1}{3},3]\,,\,[-32,30]\,,\,[-20,24]\right)^T , \\ [x]^S &= \left([-4,0]\,,\,[\tfrac{1}{3},3]\,,\,[-32,28]\,,\,[-20,24]\right)^T . \end{aligned}$$

In particular, $[x]_2^G = [x]_2^S$, $[x]_4^G = [x]_4^S$ as predicted by Corollary 3 a) .

□

Now we address ourselves to the overestimation among $[x]^G$ and $[x]^S$. By the Theorems 2 and 3 this overestimation can be given explicitly when distinguishing six cases. (See the proof of the subsequent theorem.) However, we prefer to state an optimal bound which does not depend on the various cases.

Theorem 6

Let the assumptions of Theorem 2 hold. With the notation of the preceding theorems we get for any $i \in \{1, \ldots, n\}$

$$m_{ii} \geq f_i, \qquad \nu_i \leq \mu_i, \tag{17}$$

$$q([x]_i^G, [x]_i^S) \leq 2(m_{ii} - f_i)|\check{b}_i| \,, \tag{18}$$

where equality can hold in (17) and in (18).

Proof

¿From (15) the inequalities in (17) follow immediately. Choosing $[A] = I$ yields equality there.

In order to prove (18) use the Theorems 2 and 3 to derive the following expressions for $\underline{x}_i^S - \underline{x}_i^G$ and $\overline{x}_i^S - \overline{x}_i^G$, respectively.

Case $\check{b}_i \geq 0$:

$$\begin{aligned}
\underline{x}_i^G \leq 0,\ \underline{x}_i^S \leq 0 \quad &\Rightarrow \quad \underline{x}_i^S - \underline{x}_i^G = 2(m_{ii} - f_i)\check{b}_i; \\
\underline{x}_i^G \leq 0,\ \underline{x}_i^S > 0 \quad &\Rightarrow \quad \underline{x}_i^S - \underline{x}_i^G = (1-\nu_i)x_i^* + 2(\nu_i m_{ii} - f_i)\check{b}_i; \\
\underline{x}_i^G > 0,\ \underline{x}_i^S > 0 \quad &\Rightarrow \quad \underline{x}_i^S - \underline{x}_i^G = (\mu_i - \nu_i)x_i^* + 2(\nu_i m_{ii} - \mu_i f_i)\check{b}_i.
\end{aligned}$$

Since $\underline{x}_i^S > 0$ implies $x_i^* < 2m_{ii}\check{b}_i$ we get $\underline{x}_i^S - \underline{x}_i^G \leq 2(1-\nu_i)m_{ii}\check{b}_i + 2(\nu_i m_{ii} - f_i)\check{b}_i = 2(m_{ii} - f_i)\check{b}_i$ in the second subcase. In the third $\underline{x}_i^G > 0$ implies $x_i^* < 2f_i\check{b}_i$ hence $\underline{x}_i^S - \underline{x}_i^G \leq 2(\mu_i - \nu_i)f_i\check{b}_i + 2(\nu_i m_{ii} - \mu_i f_i)\check{b}_i = 2\nu_i(m_{ii} - f_i)\check{b}_i \leq 2(m_{ii} - f_i)\check{b}_i$. Together with Theorem 4 this proves (18) when $\check{b}_i \geq 0$. Note that in the estimates we used $\mu_i,\ \nu_i \in (0, 1]$ and (17).

Case $\check{b}_i < 0$:

$$\begin{aligned}
\overline{x}_i^G \geq 0,\ \overline{x}_i^S \geq 0 \quad &\Rightarrow \quad \overline{x}_i^G - \overline{x}_i^S = 2(f_i - m_{ii})\check{b}_i = 2(m_{ii} - f_i)|\check{b}_i|; \\
\overline{x}_i^G \geq 0,\ \overline{x}_i^S < 0 \quad &\Rightarrow \quad \overline{x}_i^G - \overline{x}_i^S = (1-\nu_i)x_i^* + 2(f_i - \nu_i m_{ii})\check{b}_i; \\
\overline{x}_i^G < 0,\ \overline{x}_i^S < 0 \quad &\Rightarrow \quad \overline{x}_i^G - \overline{x}_i^S = (\mu_i - \nu_i)x_i^* + 2(\mu_i f_i - \nu_i m_{ii})\check{b}_i.
\end{aligned}$$

Since $\overline{x}_i^S < 0$ implies $x_i^* < -2m_{ii}\check{b}_i$ we get $\overline{x}_i^G - \overline{x}_i^S \leq -2(1-\nu_i)m_{ii}\check{b}_i + 2(f_i - \nu_i m_{ii})\check{b}_i = 2(f_i - m_{ii})\check{b}_i = 2(m_{ii} - f_i)|\check{b}_i|$ in the second subcase. In the third $\overline{x}_i^G < 0$ implies $x_i^* < -2f_i\check{b}_i$ hence $\overline{x}_i^G - \overline{x}_i^S \leq -2(\mu_i - \nu_i)f_i\check{b}_i + 2(\mu_i f_i - \nu_i m_{ii})\check{b}_i = 2\nu_i(f_i - m_{ii})\check{b}_i = 2\nu_i(m_{ii} - f_i)|\check{b}_i| \leq 2(m_{ii} - f_i)|\check{b}_i|$. Together with Theorem 4 this proves (18) including the possibility of equality.

□

By means of the next lemma we are able to generalize our results to matrices of the form $[A] = D + [-R, R]$ where D is now an arbitrary regular real diagonal matrix. Such matrices were first considered in [16]. Let $[\hat{A}] := D^{-1}[A] = I + |D^{-1}|[-R, R]$, $[\hat{b}] := D^{-1}[b]$, and denote by $[\hat{x}]^G$, $[\hat{x}]^S$ the associated quantities, corresponding to $[x]^G$, $[x]^S$ of the original data. Then $[\hat{x}]^S = [x]^S$, and Lemma 4 a) shows $[\hat{x}]^G = [x]^G$. Therefore, we can apply our preceding results to $[\hat{A}]$, $[\hat{b}]$ in order to get properties for $[x]^G$, $[x]^S$. We leave the details to the reader.

Lemma 4

Let $[A] \in \boldsymbol{IR}^{n\times n}$, $[b] \in \boldsymbol{IR}^n$, and let $D = \mathrm{diag}(d_1, \ldots, d_n) \in \boldsymbol{R}^{n\times n}$ be a regular diagonal matrix. Then the following assertions hold.

a) $\mathrm{IGA}([A], [b])$ *exists if and only if* $\mathrm{IGA}(D[A], D[b])$ *exists. In this case both vectors are equal, and for the intermediate quantities of the IGA one gets*

$$(D[A])^{(k)} = D[A]^{(k)}, \quad (D[b])^{(k)} = D[b]^{(k)} \ . \tag{19}$$

b) $[A]$ is an H–matrix if and only if $D[A]$ is an H–matrix.

c) $[A]$ is regular if and only if $D^{-1}[A]$ is regular.

Proof

a) We first prove (19) by induction. For $k = 1$ this is trivial. Let (3) hold for some $k < n$. Then

$$\begin{aligned}(D[A])_{ij}^{(k+1)} &= d_i[a]_{ij}^{(k)} - d_i[a]_{ik}^{(k)} d_k[a]_{kj}^{(k)} / (d_k[a]_{kk}^{(k)}) \\ &= d_i \left([a]_{ij}^{(k)} - [a]_{ik}^{(k)} [a]_{kj}^{(k)} / [a]_{kk}^{(k)} \right) = d_i[a]_{ij}^{(k+1)} ,\end{aligned}$$

and, similarly,

$$\begin{aligned}(D[b])_{i}^{(k+1)} &= d_i[b]_{i}^{(k)} - d_i[a]_{ik}^{(k)} d_k[b]_{k}^{(k)} / (d_k[a]_{kk}^{(k)}) \\ &= d_i \left([b]_{i}^{(k)} - [a]_{ik}^{(k)} [b]_{k}^{(k)} / [a]_{kk}^{(k)} \right) = d_i[b]_{i}^{(k+1)} .\end{aligned}$$

Let $[x]^G = \mathrm{IGA}([A], [b])$ and $[y]^G = \mathrm{IGA}(D[A], D[b])$. Then

$$[y]_n^G = d_n[b]_n^{(n)} / (d_n[a]_{nn}^{(n)}) = [x]_n^G ,$$

and $[y]_i^G = [x]_i^G$ for each $i \geq k+1$ implies

$$[y]_k^G = \left(d_k[b]_k^{(n)} - \sum_{j=k+1}^{n} d_k[a]_{kj}^{(n)} [y]_j^G \right) / (d_k[a]_{kk}^{(n)}) = [x]_k^G .$$

b) Let $[A]$ be an H–matrix, i.e., $\langle [A] \rangle$ is an M–matrix. By a well–known result of Fan [5] there is a vector $u > 0$ such that $\langle [A] \rangle u > 0$. Hence $\langle D[A] \rangle u = |D| \langle [A] \rangle u > 0$, and $D[A]$ is an H–matrix, too. Applying this result to D^{-1} and $D[A]$ instead of D and $[A]$, respectively, proves the converse assertion since $D^{-1}(D[A]) = (D^{-1}D)[A]$ for diagonal matrices D. (Note that the interval matrix multiplication is not associative, in general.)

c) is trivial.

□

Acknowledgements

We thank Prof. Arnold Neumaier, University of Vienna, and the referees for their valuable comments.
The second author's work was supported by the Czech Republic Grant Agency under grant 201/98/0222.

References

[1] Alefeld G.: Über die Durchführbarkeit des Gaußschen Algorithmus bei Gleichungen mit Intervallen als Koeffizienten. Computing Suppl. **1**, 15 – 19 (1977).

[2] Alefeld, G., Herzberger, J.: Introduction to Interval Computations. Academic Press, New York, 1983.

[3] Barth, W., Nuding, E.: Optimale Lösung von Intervallgleichungssystemen. Computing **12**, 117 – 125 (1974).

[4] Beeck, H.: Zur scharfen Außenabschätzung der Lösungsmenge bei linearen Intervallgleichungssystemen. Z. Angew. Math. Mech. **54**, T208 – T209 (1974).

[5] Fan, K.: Topological proof for certain theorems on matrices with non–negative elements. Monatsh. Math. **62**, 219 – 237 (1958).

[6] Frommer, A., Mayer, G.: A new criterion to guarantee the feasibility of the interval Gaussian algorithm. SIAM J. Matrix Anal. Appl. **14**, 408 – 419 (1993).

[7] Frommer, A., Mayer, G.: Linear Systems with Ω–diagonally Dominant Matrices and Related Ones. Linear Algebra Appl. **186**, 165 – 181 (1993).

[8] Golub, H. G., van Loan, C. F.: Matrix Computations. Third Edition. The Johns Hopkins University Press, Baltimore, 1996.

[9] Hansen, E. R.: Bounding the solution of linear interval equations. SIAM J. Numer. Anal. **29**, 1493 – 1503 (1992).

[10] Hansen, E. R.: Gaussian Elimination in Interval Systems. Preprint (1997).

[11] Mayer, G.: Old and new aspects of the interval Gaussian algorithm. In Kaucher, E., Markov, S. M., Mayer, G., eds.: Computer Arithmetic, Scientific Computation and Mathematical Modelling. IMACS Annals on Computing and Applied Mathematics **12**, Baltzer, Basel, 1991, 329 – 349.

[12] Mayer, G., Pieper, L.: A necessary and sufficient criterion to guarantee the feasibility of the interval Gaussian algorithm for some class of matrices. Appl. Math. **38**, 205 – 220 (1993).

[13] Neumaier, A.: Interval Methods for Systems of Equations. Cambridge University Press, Cambridge, 1990.

[14] Neumaier, A.: The Wrapping Effect, Ellipsoid Arithmetic, Stability and Confidence Regions. In Albrecht, R., Alefeld, G., Stetter, H. J. (eds.): Validation Numerics. Theory and Applications. Computing Supplementum **9**, Springer, Wien, 1993, 175 – 190.

[15] Neumaier, A.: Personal communication, 1997.

[16] Ning, S., Kearfott, R. B.: A Comparison of Some Methods for Solving Linear Interval Equations. SIAM J. Numer. Anal. **34**, 1289 – 1305 (1997).

[17] Rohn, J.: Systems of linear interval equations. Linear Algebra Appl. **126**, 39 – 78 (1989).

[18] Rohn, J.: Cheap and Tight Bounds: The Recent Result by E. Hansen Can Be Made More Efficient. Interval Computations **4**, 13 – 21 (1993).

[19] Rohn, J.: On Overestimations Produced by the Interval Gaussian Algorithm. Reliable Computing **4**, 363 – 368 (1997).

[20] Rump, S. M.: Validated Solution of Large Linear Systems. In Albrecht, R., Alefeld, G., Stetter, H. J. (eds.): Validation Numerics. Theory and Applications. Computing Supplementum **9**, Springer, Wien, 1993, 191 – 212.

[21] Varga, R. S.: Matrix Iterative Analysis. Prentice–Hall, Englewood Cliffs, N. J., 1962.