https://doi.org/10.18453/rosdok_id00005822

A Survey on Properties and Algorithms for the Symmetric Solution Set

Günter Mayer*

Preprint, Universität Rostock, December 2011

Dedicated to Professor Dr. Götz Alefeld on the occasion of his 70^{th} birthday.

Abstract

We give a survey on the symmetric solution set of linear systems of equations with perturbed input data starting with a description of its form and presenting several algorithms whose resulting interval vector encloses it. In particular, we derive the describing inequalities in a new way and represent them in a modified form as compared with a recent result by Hladík. This form enables us to improve a known upper bound for their number essentially. We thus end up with an Oettli–Prager like theorem for the symmetric solution set which we can prove by elementary means and completely different from Hladík's paper.

Key words: Interval analysis, linear system of equations with symmetric coefficient matrix, solution set, symmetric solution set, Oettli–Prager theorem, Oettli–Prager like theorem, Fourier–Motzkin elimination, interval Cholesky algorithm, incomplete interval Cholesky method, Jansson's method, Rohn's method, historical remarks.

AMS Subject Classification: 65G20, 65G30

*Institut für Mathematik, Universität Rostock, Ulmenstr. 69, Haus 3, D–18057 Rostock, Germany (guenter.mayer@uni-rostock.de)

Contents

1 Introduction **3**

2 Preliminaries **5**

3 Characterizations of Σ_{sym} **9**

4 Interval Cholesky method and modifications **28**

5 Incomplete Cholesky decomposition **43**

6 Jansson's method **49**

7 Rohn's method **50**

8 Some historical remarks **54**

References **55**

1 Introduction

We give a survey on equivalent descriptions, on properties and on enclosures of the symmetric solution set

$$\Sigma_{\rm sym} = \{x \in \mathbb{R}^n |\ Ax = b,\ A = A^T \in [A] = [A]^T,\ b \in [b]\}, \tag{1.1}$$

where $[A] = [\underline{A}, \overline{A}]$ is a given $n \times n$ interval matrix satisfying $[A] = [A]^T$, and $[b]$ is a corresponding interval vector. This set is apparently a subset of the more general and more frequently considered general solution set

$$\Sigma = \{x \in \mathbb{R}^n |\ Ax = b,\ A \in [A],\ b \in [b]\}, \tag{1.2}$$

which also contains the solutions of linear systems with unsymmetric matrices from $[A]$. In [39] Neumaier illustrates both sets for a simple 2×2 example and shows that the tightest enclosure of $\Sigma_{\rm sym}$, the so–called interval hull $\Box\Sigma_{\rm sym}$ of this set, can be smaller than the corresponding interval hull $\Box\Sigma$ of Σ, which in his terminology is denoted as hull inverse. Therefore he remarks [39], p. 95:

Thus the hull inverse is not adapted to the optimal treatment of symmetric matrices. However, at present, no special methods have been devised for this case, and we shall content ourselves with the unsymmetric treatment of symmetric matrices.

To show that things changed since 1990 – at least slightly – is one of the purposes of this paper. There is another statement which fits into this subject. It was made by Babuška when talking to Jiří Rohn in 1992 and is quoted according to an email sent by Rohn to a group of scientists working on the field of interval analysis [49].

Unless you are able to handle dependent data, you will never gain interest of the engineers.

Note that by the symmetry of the matrices $\Sigma_{\rm sym}$ certainly contains a data dependence as mentioned in the quotation. Moreover, the symmetric solution set can equivalently be introduced as a solution set of particular parameter dependent linear systems

$$(\check{A} + T \circ \mathrm{rad}([A]))\, x = \check{b} + \tau \circ \mathrm{rad}([b]), \tag{1.3}$$

where the entries t_{ij} of the symmetric matrix $T \in \mathbb{R}^{n\times n}$ and τ_i of the vector $\tau \in \mathbb{R}^n$ are the parameters which are allowed to vary in the interval $[-1, 1]$; cf. [50]. The symbol $\circ$ denotes the Hadamard product which is defined as the entrywise multiplication of matrices and vectors, respectively; cf. [25]. The midpoints $\check{A}$, $\check{b}$ and the radii $\mathrm{rad}([A])$, $\mathrm{rad}([b])$ of $[A]$ and $[b]$ are introduced in Section 2. By virtue of (1.3) it is obvious that enclosures for parameter dependent linear systems can also be adapted for enclosures of $\Sigma_{\rm sym}$; cf. [27], [41], [42], [43], [44], [50], [51], [59], [61]. They supplement enclosure methods which originate from interval analysis like the interval Cholesky method [8] or a modification of the Krawczyk method [26]. We shall recall these methods later–on together with their properties.

In [30] we mentioned three equivalent descriptions for the general solution set Σ: The Oettli–Prager criterion [40] which is a vector inequality using the midpoints and radii of $[A] = ([a]_{ij})$ and $[b] = ([b]_i)$, Beeck's criterion [13] based on interval

arithmetic and intersection of sets, and Hartfiel's characterization which consists of a variety of inequalities depending on the orthants and on the endpoints of $[a]_{ij}$ and $[b]_i$. For each of these characterizations of Σ there are analogua for describing $\Sigma_{\rm sym}$; cf. [7], [24], [30], [33], and Section 3 of the present survey. It turns out that for regular matrices $[A]$ the general solution set Σ is the union of at most 2^n convex polyhedra which, together, form a starlike, connected but not necessarily convex polyhedron themselves. In contrast to the piecewise plain surface of Σ the boundary of $\Sigma_{\rm sym}$ is formed by parts of quadrics and hyperplanes in $\mathbb{R}^n$, i.e., of solutions of algebraic equations of degree at most two in n variables. This structure complicates the study of $\Sigma_{\rm sym}$ tremendously. On the other hand enclosures for $\Sigma_{\rm sym}$ are needed in practice. This can be seen for instance from [11] and [29] which deal with Markov chains and truss mechanics, respectively.

Up to now essentially enclosures for the corresponding general solution set Σ are used to bound $\Sigma_{\rm sym}$ – mainly for simplicity – although each of the methods described in the Sections 4 – 6 could be used either.

We finally mention the symmetric tolerance solution set [58]

$$\Sigma_{\rm sym}^{\rm tol} = \{x \in \mathbb{R}^n | \; \forall \, A = A^T \in [A] \; \exists \, b \in [b] \; : \; Ax = b \, \}$$

and the symmetric control solution set

$$\Sigma_{\rm sym}^{\rm contr} = \{x \in \mathbb{R}^n | \; \forall \, b \in [b] \; \exists \, A = A^T \in [A] \; : \; Ax = b \, \}$$

which we do not consider in this paper.

Many of our results can be transferred to skew–symmetric matrices $A = -A^T$, and to persymmetric matrices which are, by definition, symmetric with respect to the counterdiagonal; cf. [5].

We have organized our paper as follows: In Section 2 we list our notation, in Section 3 we describe $\Sigma_{\rm sym}$, in the Sections 4 and 5 we study the interval Cholesky method and the interval version of an iterative process based on an incomplete Cholesky decomposition, and in Section 6 we recall Jansson's variant of the well known Krawczyk method. Section 7 is devoted to a method suggested by Rohn, and the final Section 8 contains some historical remarks. The Sections 3 and 5 contain new material which we intend to publish elsewhere in the nearer future.

This text is a very enlarged version of an invited talk which the author presented at the Conference INVA 2008 on Okinawa, Japan. In particular, it was supplemented by the results in [24] on $\Sigma_{\rm sym}$ which were not available at that time. These results are modified now and deduced and proved in a new and elementary manner. They are represented for the first time in a matrix–vector form which resembles more that form in the Oettli–Prager theorem for Σ. In addition, we improve the upper bound of the number of inequalities in [24] which are necessary to describe $\Sigma_{\rm sym}$. We need now asymptotically at most $3^n/2$ inequalities as compared to $4^n/2$ inequalities in [24]. The exponential increase of this number with increasing dimension n of $[A]$ could, however, not be changed.

2 Preliminaries

By $\mathbb{R}^n, \mathbb{R}^{n\times n}, \mathbb{IR}, \mathbb{IR}^n, \mathbb{IR}^{n\times n}$ 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 their single element. Examples are the zero matrix O, the identity matrix I, its columns $e^{(i)},\ i = 1, \ldots, n$, and the vector $e = (1, 1, \ldots, 1)^T$. We use the notation $[A] = [\underline{A}, \overline{A}] = ([a]_{ij}) = ([\underline{a}_{ij}, \overline{a}_{ij}]) \in \mathbb{IR}^{n\times n}$ simultaneously without further reference, and we proceed similarly for the elements of $\mathbb{R}^n, \mathbb{R}^{n\times n}, \mathbb{IR}$ and $\mathbb{IR}^n$. We also mention the standard notation from interval analysis ([2], [36], [39])

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

for intervals $[a], [b]$. For $[A],\ [B] \in \mathbb{IR}^{n\times n}$ we obtain $\check{A},\ \operatorname{rad}([A]),\ |[A]|$, and $q([A], [B]) \in \mathbb{R}^{n\times n}$ by applying the operators $\operatorname{mid}(\cdot),\ \operatorname{rad}(\cdot),\ |\cdot|$, and $q(\cdot, \cdot)$ entrywise, and we proceed similarly for interval vectors. The comparison matrix $\langle [A] \rangle = (c_{ij}) \in \mathbb{R}^{n\times n}$ is defined by

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

Since real numbers can be viewed as degenerate intervals, $\operatorname{rad}(\cdot),\ |\cdot|,\ q(\cdot, \cdot)$, and $\langle \cdot \rangle$ can be used for real numbers, vectors and matrices, too.

By $A \geq O$ we denote a nonnegative $n \times n$ matrix, i.e., $a_{ij} \geq 0$ for $i, j = 1, \ldots, n$. Analogously, we define $x \geq 0$ for $x \in \mathbb{R}^n$. We call $x \in \mathbb{R}^n$ positive writing $x > 0$ if $x_i > 0$ for $i = 1, \ldots, n$. We use $Z^{n\times n}$ for the set of real $n \times n$ matrices with non–positive off–diagonal entries. Trivially, $Z^{n\times n}$ contains the $n \times n$ matrix $\langle A \rangle$. As usual we call $A \in \mathbb{R}^{n\times n}$ inverse nonnegative if it is regular with $A^{-1} \geq O$. It is an M–matrix if it is inverse nonnegative and in $Z^{n\times n}$. It is a Stieltjes matrix if it is a symmetric M–matrix, and it is an H–matrix if $\langle A \rangle$ is an M–matrix. Moreover, it is totally positive (totally nonnegative) if each minor of A is positive (nonnegative), and it is an oscillatory matrix if it is totally nonnegative and if at least one of its powers A^k is totally positive; cf. [14], [17], [25]. For $p \in \mathbb{R}^n$ we define the matrix $D = \operatorname{diag}(p)$ as the $n \times n$ diagonal matrix with $d_{ii} = p_i,\ i = 1, \ldots, n$. By $\rho(A)$ we denote the spectral radius of a matrix $A \in \mathbb{R}^{n\times n}$.

An interval matrix $[A] \in \mathbb{IR}^{n\times n}$ is regular, if each element $\tilde{A} \in [A]$ is regular; otherwise it is called singular. It is defined to be an M–matrix if each element

$\tilde{A} \in [A]$ is an M–matrix. In the same way the terminology 'Stieltjes matrix', 'H–matrix', 'inverse nonnegative matrix', 'totally nonnegative matrix', etc. can be extended to $\mathbb{IR}^{n\times n}$. It is easy to verify that $[A] \in \mathbb{IR}^{n\times n}$ is an M–matrix if and only if $\underline{A}$ is an M–matrix and $\overline{a}_{ij} \leq 0$ for $i \neq j$, and that $[A] \in \mathbb{IR}^{n\times n}$ is an H–matrix if and only if $\langle [A] \rangle$ is an M–matrix. By Kuttler's theorem (Proposition 3.6.6 in [39]) $[A]$ is inverse nonnegative if and only if $\underline{A}$ and $\overline{A}$ have this property. Specializing Theorem 1 in [18] the interval matrix $[A]$ is totally positive if and only if the same holds for the matrices $A^L = (a_{ij}^L)$, $A^U = (a_{ij}^U) \in \mathbb{R}^{n\times n}$ which are defined by

$$a_{ij}^L = \begin{cases} \overline{a}_{ij}, & \text{if } i+j \text{ is odd,} \\ \underline{a}_{ij}, & \text{if } i+j \text{ is even,} \end{cases} \qquad a_{ij}^U = \begin{cases} \overline{a}_{ij}, & \text{if } i+j \text{ is even,} \\ \underline{a}_{ij}, & \text{if } i+j \text{ is odd.} \end{cases}$$

It is conjectured in [18] that this property also holds if 'totally positive' is replaced by 'regular and totally nonnegative'; see also [19]. Additional sufficient criteria for this latter property can be found in [34], Theorem 6.1 .

We call $[A] \in \mathbb{IR}^{n\times n}$ irreducible if $\langle [A] \rangle$ is irreducible.[1] In the same way we define $[A]$ to be diagonally dominant, strictly diagonally dominant, and irreducibly diagonally dominant, respectively. If there is a positive vector x such that

$$\langle [A] \rangle x \geq 0 \tag{2.1}$$

then we call $[A]$ generalized diagonally dominant. Moreover, we define $[A]$ to be generalized strictly diagonally dominant if strict inequality holds in (2.1). Analogously, a generalized irreducibly diagonally dominant matrix $[A]$ is irreducible and generalized diagonally dominant with $(\langle [A] \rangle x)_i > 0$ in (2.1) for at least one component i. It is well known that generalized strictly diagonally dominant matrices are H–matrices and vice versa. Note that the add–in 'generalized' can be omitted if $x = e$ can be chosen in (2.1).

The smallest interval which encloses a given bounded set $S \subseteq \mathbb{R}$ is called interval hull of S. It is denoted by $\square S$. For $S \subseteq \mathbb{R}^n$ and $S \subseteq \mathbb{R}^{n\times n}$ the interval hull is defined analogously. If $[A] \in \mathbb{IR}^{n\times n}$ is regular we write $[A]^{-1}$ for the interval hull of $S = \{ \tilde{A}^{-1} \mid \tilde{A} \in [A] \}$ and call it inverse of $[A]$.

We equip $\mathbb{IR}, \mathbb{IR}^n, \mathbb{IR}^{n\times n}$ with the usual real interval arithmetic as described in [2], [36], [39]. We assume that the reader is familiar with the basic properties of this arithmetic. Here we only recall

$$[a] \circ [b] = \{ \tilde{a} \circ \tilde{b} \mid \tilde{a} \in [a],\ \tilde{b} \in [b] \} \tag{2.2}$$

with $[a], [b] \in \mathbb{IR}$, $\circ \in \{ +, -, \cdot, / \}$, and $0 \notin [b]$ in case of division. From (2.2) one sees immediately that $[a] \circ [b]$ can be expressed by means of the interval bounds $\underline{a}, \overline{a}, \underline{b}, \overline{b}$ which is interesting for practical computation. The structures $(\mathbb{IR}, +)$ and $(\mathbb{IR}, \cdot)$ are two commutative monoids, i.e., commutative semigroups with neutral element, but neither $(\mathbb{IR}, +)$ nor $(\mathbb{IR}\backslash\{0\}, \cdot)$ is a group. For the addition and multiplication the so–called subdistributive law

$$[a]([b] + [c]) \subseteq [a][b] + [a][c]$$

[1] Note that sometimes $[A]$ is defined to be irreducible if $|[A]|$ has this property. Unless $n = 1$, i.e., $[A] \in \mathbb{IR}^{1\times 1}$, both definitions are equivalent.

holds with '$\subsetneq$' being possible as the example $[-1,1](1-1) = 0 \neq [-2,2] = [-1,1]\cdot 1 + [-1,1]\cdot 1$ shows. Moreover, $[A][A]^{-1} = I$ does not hold in general for interval matrices $[A]$. Despite of these algebraic deficiencies the topological behavior is satisfactory: With respect to the Hausdorff distance q the arithmetic depends continuously on the operands, and $(\mathbb{IR}, q)$ is a complete metric space.

For $[a] \in \mathbb{IR}$ we define

$$\sqrt{[a]} = \{\sqrt{\tilde{a}} \mid \tilde{a} \in [a]\} \quad \text{for} \quad 0 \leq \underline{a} \tag{2.3}$$

and

$$[a]^2 = \{\tilde{a}^2 \mid \tilde{a} \in [a]\}. \tag{2.4}$$

Instead of $\sqrt{[a]}$ we also write $[a]^{1/2}$.

If all symmetric matrices $\tilde{A} \in [A] = [A]^T \in \mathbb{IR}^{n\times n}$ are positive definite we introduce the interval matrix

$$[A]^{\frac{1}{2}} = \sqrt{[A]} = \square\{\sqrt{\tilde{A}} \mid \tilde{A} = \tilde{A}^T \in [A]\}, \tag{2.5}$$

where $\sqrt{\tilde{A}} = \tilde{A}^{\frac{1}{2}} \in \mathbb{R}^{n\times n}$ denotes the unique symmetric positive definite matrix whose square equals $\tilde{A}$ (cf. [25], e.g.). In passing we note that $[A]$ is regular since the symmetric part $(\tilde{A} + \tilde{A}^T)/2$ of *any* matrix $\tilde{A} \in [A]$ (not only the symmetric ones!) is symmetric and positive definite by assumption. Therefore, we have $0 < x^T \frac{1}{2}(\tilde{A} + \tilde{A}^T)x = x^T \tilde{A} x$ for any nonzero vector $x \in \mathbb{R}^n$; hence $\tilde{A}$ cannot be singular.

In the sequel we shall use the following basic facts on intervals.

Lemma 2.1

Let $[a]$, $[b]$, $[c]$, $[a]_i \in \mathbb{IR}$, $i = 1, \ldots, n$, and let $\gamma \in \mathbb{R}$. Then the following properties hold.

a) $\operatorname{mid}(\gamma[a]) = \gamma\check{a}$; $\qquad \operatorname{mid}([a] \pm [b]) = \check{a} \pm \check{b}$;

b) $\operatorname{rad}(\gamma[a]) = |\gamma|\operatorname{rad}([a])$; $\qquad \operatorname{rad}([a] \pm [b]) = \operatorname{rad}([a]) + \operatorname{rad}([b])$;

c) $[a] \cap [b] \neq \emptyset \quad \Leftrightarrow \quad |\check{a} - \check{b}| \leq \operatorname{rad}([a]) + \operatorname{rad}([b]) \quad \Leftrightarrow \quad \underline{a} \leq \overline{b} \ \wedge \ \underline{b} \leq \overline{a}$;

d) If $n \geq 2$ then

$$\bigcap_{i=1}^{n} [a]_i \neq \emptyset \quad \Leftrightarrow \quad \max_{1\leq i\leq n} \underline{a}_i \leq \min_{1\leq i\leq n} \overline{a}_i$$
$$\Leftrightarrow \quad [a]_i \cap [a]_j \neq \emptyset \ \ \textit{for } i < j, \ i, j = 1, \ldots, n;$$

e) $([a] + [b]) \cap [c] \neq \emptyset \quad \Leftrightarrow \quad [a] \cap ([c] - [b]) \neq \emptyset$;

f) $[a] \subseteq [b] \quad \Leftrightarrow \quad |\check{a} - \check{b}| \leq \operatorname{rad}([b]) - \operatorname{rad}([a])$.

Lemma 2.2 *(Theorem 4.4 in* [54]*)*

Let $[A] = [A]^T \in \mathbb{R}^{n\times n}$ *be an M–matrix and define* $[A]^{-\frac{1}{2}}$ *by* $[A]^{-\frac{1}{2}} = ([A]^{\frac{1}{2}})^{-1}$*. Then each symmetric matrix* $\tilde{A} \in [A]$ *is positive definite, and*

$$[A]^{\frac{1}{2}} = [\underline{A}^{\frac{1}{2}}, \overline{A}^{\frac{1}{2}}], \qquad [A]^{-\frac{1}{2}} = [(\overline{A}^{\frac{1}{2}})^{-1}, (\underline{A}^{\frac{1}{2}})^{-1}].$$

In particular, $([A]^{-\frac{1}{2}})^2 = [A]^{-1} = [\overline{A}^{-1}, \underline{A}^{-1}]$.

Theorem 2.1 *(Moore's Theorem; cf.* [2] *or* [39]*)*

Let $f(x)$ *be an expression for the function* $f : x \in D \subseteq \mathbb{R}^n \to \mathbb{R}$*. Let* $f(x)$ *be defined by the basic operations* $+, -, \cdot, /$ *and the 'usual' programmable elementary functions. If each variable* x_i *occurs at most once in* $f(x)$ *then the interval arithmetic evaluation* $f([x])$ *yields the range* $\{f(x) |\ x \in [x]\}$ *for arbitrary interval vectors* $[x] \subseteq D$*.*

3 Characterizations of Σ_{sym}

First we prove the characterizations of the general solution set Σ mentioned in the introduction.

Theorem 3.1

Let $[A] \in \mathbb{IR}^{n\times n}$ be regular and $[b] \in \mathbb{IR}^n$. Then the following statements are equivalent for $x \in \mathbb{R}^n$.

a) $x \in \Sigma$;

b) $[b] \cap ([A]x) \neq \emptyset$; *(Beeck [13])*

c) $|\check{b} - \check{A}x| \leq \operatorname{rad}([A]) \cdot |x| + \operatorname{rad}([b])$; *(Oettli and Prager [40])*

d) $\underline{b}_i \leq \sum_{j=1}^{n} a_{ij}^{+} x_j \quad$ *and* $\quad \sum_{j=1}^{n} a_{ij}^{-} x_j \leq \overline{b}_i \quad$ *for* $i = 1, \ldots, n,$

where a_{ij}^{-}, a_{ij}^{+} *are defined by* $[a]_{ij} = [\underline{a}_{ij}, \overline{a}_{ij}] = \begin{cases} [a_{ij}^{-}, a_{ij}^{+}] & \text{if } x_j \geq 0 \\ [a_{ij}^{+}, a_{ij}^{-}] & \text{if } x_j < 0 \end{cases}$.

(Hartfiel [23])

Proof.

a) $\Rightarrow$ b):
Follows directly from the existence of some $\tilde{A} \in [A]$, $\tilde{b} \in [b]$ such that $\tilde{A}x = \tilde{b}$.

b) $\Rightarrow$ c):
Follows directly by means of Lemma 2.1 a), b), c).

c) $\Rightarrow$ d):
From c) one trivially gets

$$-\operatorname{rad}([A])\,|x| - \operatorname{rad}([b]) \leq \check{b} - \check{A}x \leq \operatorname{rad}([A])\,|x| + \operatorname{rad}([b]),$$

whence

$$\check{A}x - \operatorname{rad}([A])\,|x| \leq \overline{b} \quad \text{and} \quad \underline{b} \leq \check{A}x + \operatorname{rad}([A])\,|x|.$$

Now d) follows immediately by rewriting $|x|$ without absolute value bars.

d) $\Rightarrow$ a) :
For $t \in [0,1]$ define $a_{ij}(t) = a_{ij}^{-} + t(a_{ij}^{+} - a_{ij}^{-})$ and $b_i(t) = \sum_{j=1}^{n} a_{ij}(t)x_j$. Then $a_{ij}(t) \in [a]_{ij}$, and d) implies $b_i(0) \leq \overline{b}_i$, $\underline{b}_i \leq b_i(1)$. If $b_i(1) \leq \overline{b}_i$ choose $t_i = 1$; then $b_i(1) \in [b]_i$. Otherwise choose t_i such that $b_i(t_i) = \overline{b}_i$, whence $b_i(t_i) \in [b]_i$ trivially. For $A = (a_{ij}(t_i))$, $b = (b_i(t_i))$ we finally get $Ax = b$ and $x \in \Sigma$.

□

All properties in Theorem 3.1 can be retrieved for $\Sigma_{\rm sym}$ in a modified way. Unfortunately it lacks shortness, and in the cases b) and d) it is based on a Fourier–Motzkin elimination technique and hence on a finite iterative process.

We first consider an analogue of d); cf. [5]. To this end let O_1 be the closed first orthant and – for simplicity – let $x \in O_1$. (The general case can be handled analogously.)

Trivially, $x \in \Sigma_{\rm sym} \cap O_1$ is equivalent to the existence of $A = A^T \in \mathbb{R}^{n\times n},\ b \in \mathbb{R}^n$ such that

$$x \in O_1 \ \wedge \ \left\{ \begin{array}{c} \underline{b}_i \le \sum_{j=1}^{n} a_{ij}x_j \le \overline{b}_i \\ \underline{a}_{ij} \le a_{ij} \le \overline{a}_{ij} \end{array} \right\}.$$

(Initialization step)

This in turn is equivalent to the existence of $A = A^T \in \mathbb{R}^{n\times n},\ b \in \mathbb{R}^n$ such that

$$x \in \Sigma \cap O_1 \ \wedge \ \left\{ \begin{array}{c} \underline{b}_i x_i \le \sum_{j=1}^{n} a_{ij}x_i x_j \le \overline{b}_i x_i \\ \underline{a}_{ij}x_i x_j \le a_{ij}x_i x_j \le \overline{a}_{ij}x_i x_j \end{array} \right\}.$$

(Multiplication step)

Here Σ comes into the play for the direction '$\Leftarrow$' in case of $x_i = 0$ for some index i; cf. [4] for details. Consider now those inequalities which contain a_{12}.

$$\left\{ \begin{array}{rcccl} \{\underline{b}_1 - \sum_{\substack{j=1\\ j\neq 2}}^{n} a_{1j}x_j\}x_1 & \le & a_{12}x_1x_2 & \le & \{\overline{b}_1 - \sum_{\substack{j=1\\ j\neq 2}}^{n} a_{1j}x_j\}x_1 \\ \{\underline{b}_2 - \sum_{j=2}^{n} a_{2j}x_j\}x_2 & \le & a_{12}x_1x_2 & \le & \{\overline{b}_2 - \sum_{j=2}^{n} a_{2j}x_j\}x_2 \\ \underline{a}_{12}x_1x_2 & \le & a_{12}x_1x_2 & \le & \overline{a}_{12}x_1x_2 \end{array} \right\} \qquad (3.1)$$

(Isolation step)

The basic observation is now that these inequalities hold if and only if the *maximum* of the left–hand sides is less or equal than the *minimum* of the right–hand sides of (3.1). And this is true if and only if *each* left–hand side is less or equal than *each*

right–hand side. Therefore, we finally get the six non–trivial inequalities

$$\left\{\begin{array}{c} \{\underline{b}_1 - \sum\limits_{\substack{j=1\\ j\neq 2}}^{n} a_{1j}x_j\}x_1 \leq \overline{a}_{12}x_1x_2 \\ \{\underline{b}_2 - \sum\limits_{j=2}^{n} a_{2j}x_j\}x_2 \leq \overline{a}_{12}x_1x_2 \\ \{\underline{b}_1 - \sum\limits_{\substack{j=1\\ j\neq 2}}^{n} a_{1j}x_j\}x_1 \leq \{\overline{b}_2 - \sum\limits_{j=2}^{n} a_{2j}x_j\}x_2 \\ \underline{a}_{12}x_1x_2 \leq \{\overline{b}_1 - \sum\limits_{\substack{j=1\\ j\neq 2}}^{n} a_{1j}x_j\}x_1 \\ \underline{a}_{12}x_1x_2 \leq \{\overline{b}_2 - \sum\limits_{j=2}^{n} a_{2j}x_j\}x_2 \\ \{\underline{b}_2 - \sum\limits_{j=2}^{n} a_{2j}x_j\}x_2 \leq \{\overline{b}_1 - \sum\limits_{\substack{j=1\\ j\neq 2}}^{n} a_{1j}x_j\}x_1 \end{array}\right\} \qquad (3.2)$$

(Elimination step)

in which a_{12} apparently is replaced by $\underline{a}_{12}$ and $\overline{a}_{12}$.

Repeating the elimination process (= last two steps) for (3.2) and the remaining a_{12}–free inequalities successively eliminates the entries of A and introduces those of $\underline{A}$ and $\overline{A}$. The degree of the algebraic inequalities does not change. Therefore, we finally get a set of algebraic inequalities of degree at most two as mentioned already in Section 1. The number of inequalities grows exponentially with n.

Example 3.1 [5]

Let

$$[A] = \begin{pmatrix} 1 & [0,1] \\ [0,1] & [-4,-1] \end{pmatrix}, \quad [b] = \begin{pmatrix} [0,2] \\ [0,2] \end{pmatrix}.$$

Then $[A] = [A]^T$ with

$$A = \begin{pmatrix} 1 & \alpha \\ \beta & -\gamma \end{pmatrix} \in [A] \quad \Rightarrow \quad A^{-1} = \frac{1}{\gamma + \alpha\beta} \begin{pmatrix} \gamma & \alpha \\ \beta & -1 \end{pmatrix}$$

for $\alpha, \beta \in [0,1],\ \gamma \in [1,4]$. Since $\underline{b} \geq 0$ the first component of $A^{-1}b$ is nonnegative for all $b \in [b]$. Therefore, the general solution set Σ is completely contained in the union $O_1 \cup O_4$ of the first and the fourth quadrant.

Omitting all redundant inequalities Σ can be characterized as follows

$$\left.\begin{array}{ll} \Sigma \cap O_1 : & x_1 \leq 2, \quad x_1 \geq x_2 \\ \Sigma \cap O_4 : & x_2 \geq -2, \quad x_2 \leq 2 - x_1 \end{array}\right\} \qquad (3.3)$$

For $\Sigma_{\rm sym}$ we need the additional inequalities

$$\left.\begin{array}{ll} \Sigma_{\rm sym} \cap O_1 : & (x_1-1)^2 + x_2^2 \le 1 \\ \Sigma_{\rm sym} \cap O_4 : & (x_1-1)^2 + (x_2+1)^2 \le 2 \end{array}\right\} \tag{3.4}$$

Both solution sets are illustrated in Fig. 3.1.

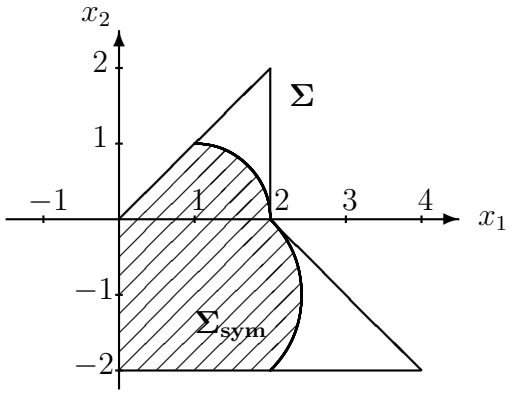


Fig. 3.1 The solution sets Σ and $\Sigma_{\rm sym}$ of Example 3.1

□

The analogue of Theorem 3.1 b) is based on Lemma 2.1; cf. [30]. We start with $x \in \Sigma_{\rm sym}$. Then there is a matrix $A = A^T$ such that

$$\left\{\begin{array}{c} \{\sum_{j=1}^{n} a_{ij}x_j\} \cap [b]_i \ne \emptyset \\ \{a_{ij}\} \cap [a]_{ij} \ne \emptyset \end{array}\right\}$$

(Initialization step)

Note that the first set of the intersections contains only one element.

Similarly as above this is equivalent to the existence of $A = A^T \in \mathbb{R}^{n\times n}$ such that

$$x \in \Sigma \ \wedge \ \left\{\begin{array}{c} \{\sum_{j=1}^{n} a_{ij}x_ix_j\} \cap [b]_i x_i \ne \emptyset \\ \{a_{ij}x_ix_j\} \cap [a]_{ij}x_ix_j \ne \emptyset \end{array}\right\}$$

(Multiplication step)

The isolation step uses Lemma 2.1 e) and results in

$$\left\{\begin{array}{c} \{a_{12}x_1x_2\} \cap ([b]_1 x_1 - \sum_{\substack{j=1 \\ j\ne 2}}^{n} a_{1j}x_1x_j) \ne \emptyset \\ \\ \{a_{12}x_1x_2\} \cap ([b]_2 x_2 - \sum_{j=2}^{n} a_{2j}x_2x_j) \ne \emptyset \\ \\ \{a_{12}x_1x_2\} \cap [a]_{12}x_1x_2 \ne \emptyset \end{array}\right\}, \tag{3.5}$$

where we only listed the sets which contain a_{12}.

The elimination step transforms (3.5) equivalently into

$$[a]_{12}x_1x_2 \cap \left([b]_1x_1 - \sum_{\substack{j=1\\ j\neq 2}}^{n} a_{1j}x_1x_j\right) \cap \left([b]_2x_2 - \sum_{j=2}^{n} a_{2j}x_2x_j\right) \neq \emptyset.$$

Thus the entry a_{12} is replaced by $[a]_{12}$. Transferring the triple intersection via Lemma 2.1 d) equivalently into intersections with two operands allows repeating the last two steps for the remaining entries a_{ij} with $(i,j) \notin \{(1,2), (2,1)\}$. One finally ends up with a variety of intersections which for Example 3.1 read

$$x \in \Sigma_{\text{sym}} \quad \Leftrightarrow \quad [b] \cap [A]x \neq \emptyset \quad \wedge \quad ([0,2]x_1 - x_1^2) \cap ([0,2]x_2 + [1,4]x_2^2) \neq \emptyset. \tag{3.6}$$

The first set theoretical inequality in (3.6) is Beeck's formulation of the Oettli–Prager criterion which yields to the inequalities in (3.3). The final set theoretical inequality in (3.6) combines the inequalities in (3.4) : For $x \in O_1$ it is equivalent to

$$[-x_1^2\,,\, 2x_1 - x_1^2] \;\cap\; [x_2^2\,,\, 2x_2 + 4x_2^2] \neq \emptyset$$

which, by virtue of Lemma 2.1 c), means

$$-x_1^2 \leq 2x_2 + 4x_2^2 \;\wedge\; 2x_1 - x_1^2 \geq x_2^2$$

or

$$x_1^2 + 4x_2^2 + 2x_2 \geq 0 \;\wedge\; (x_1 - 1)^2 + x_2^2 \leq 1.$$

Since the first inequality is obviously true for $x \in O_1$ the second one restricts $\Sigma_{\text{sym}} \cap O_1$ as in (3.4).

For $x \in O_4$ we similarly get

$$-x_1^2 \leq 4x_2^2 \;\wedge\; 2x_1 - x_1^2 \geq 2x_2 + x_2^2.$$

Here the first inequality holds trivially while the second one can be rewritten as $(x_1 - 1)^2 + (x_2 + 1)^2 \leq 2$ as in (3.4).

As an analogue of Theorem 3.1 c) we cite results of [24]. We start with a preparatory lemma.

Lemma 3.1 [24]

Let $a, b, d \in \mathbb{R}^m$, $a', b', d' \in \mathbb{R}^n$, and $C \in \mathbb{R}^{m\times n}$. The function

$$f(u,v) = a^Tu + b^T|u| + (a')^Tv + (b')^T|v| + \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}|d'_ju_i + d_iv_j| \tag{3.7}$$

is nonnegative for all $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$ if and only if it is nonnegative for all u, v satisfying at least one of the following conditions:

(i) $u_i \in \{0, d_i\}$ *for all* $i = 1, \ldots, m$, *and* $v_j \in \{0, -d'_j\}$ *for all* $j = 1, \ldots, n$;

(ii) $u_i \in \{0, -d_i\}$ *for all* $i = 1, \ldots, m$, *and* $v_j \in \{0, d'_j\}$ *for all* $j = 1, \ldots, n$;

(iii) $(u^T, v^T)^T = \pm e^{(k)}$ *for some* $k \in \{1, \ldots, m+n\}$, *where* $e^{(k)}$ *denotes the* k*–th column of the identity matrix* I *in* $\mathbb{R}^{(m+n)\times(m+n)}$.

Lemma 3.1 shows that the global nonnegativity of the particular piecewise linear function f in (3.7) can be proved by restricting the domain $\mathbb{R}^m \times \mathbb{R}^n$ to an appropriate subset.

From linear programming we know (cf. [57], § 7.4, e.g.) that the linear programs

$$\text{maximize } \tilde{b}^T \tilde{y} \text{ such that } \tilde{A}^T \tilde{y} \leq \tilde{c} \tag{3.8}$$

and

$$\text{minimize } \tilde{c}^T \tilde{x} \text{ such that } \tilde{A}\tilde{x} = \tilde{b},\ \tilde{x} \geq 0 \tag{3.9}$$

are dual to each other, i.e., their optimal values exist and are equal provided that both sets in (3.8) and (3.9) are nonempty. By means of this result and by virtue of Lemma 3.1 the following auxiliary result can be proved.

Theorem 3.2 [24]

Let $[A] \in \mathbb{IR}^{n\times n}$, $[b], [d] \in \mathbb{IR}^n$. *Then the vectors* $x, y \in \mathbb{R}^n$ *form a solution of the system*

$$Ax = b, \quad A^T y = d \tag{3.10}$$

for some $A \in [A]$, $b \in [b]$, $d \in [d]$ *if and only if they satisfy the following system of inequalities*

$$\begin{aligned} \operatorname{rad}([A])\,|x| + \operatorname{rad}([b]) &\geq |r|, \\ \operatorname{rad}([A]^T)\,|y| + \operatorname{rad}([d]) &\geq |r'|, \\ \sum_{i,j=1}^{n} \operatorname{rad}([a]_{ij})\,|y_i x_j (p_i - q_j)| + \sum_{i=1}^{n} \left(\operatorname{rad}([b]_i)\,|y_i p_i| + \operatorname{rad}([d]_i)\,|x_i q_i| \right) & \\ \geq \left| \sum_{i=1}^{n} (r_i y_i p_i - r'_i x_i q_i) \right| & \end{aligned} \tag{3.11}$$

for all vectors $p, q \in \{0,1\}^n$, *where* $r = \check{b} - \check{A}x$, $r' = \check{d} - \check{A}^T y$.

With the diagonal matrices $D_p = \operatorname{diag}(p)$, $D_q = \operatorname{diag}(q)$ (p, q as in Theorem 3.2) the inequality (3.11) can be simply rewritten as

$$\begin{aligned} &|y|^T \cdot |D_p \operatorname{rad}([A]) - \operatorname{rad}([A]) D_q| \cdot |x| + |y|^T D_p \operatorname{rad}([b]) + |x|^T D_q \operatorname{rad}([d]) \\ &\geq |y^T D_p r - x^T D_q r'| \ . \end{aligned} \tag{3.12}$$

Note that (3.12) can also be derived directly from (3.6) in the following way:

From (3.10) we get

$$y^T D_p A x = y^T D_p b \quad \text{and} \quad x^T D_q A^T y = x^T D_q d.$$

Subtracting both equalities and introducing the representations

$$A = \check{A} + \Delta, \quad b = \check{b} + \beta, \quad d = \check{d} + \delta$$

yields

$$y^T D_p \Delta x - x^T D_q \Delta^T y - y^T D_p \beta + x^T D_q \delta = y^T D_p r - x^T D_q r' \ .$$

Using $x^T D_q \Delta^T y = y^T \Delta D_q x$ and absolute values results in

$$|y^T D_p r - x^T D_q r'| \leq |y|^T \cdot |D_p \Delta - \Delta D_q| \cdot |x| + |y|^T D_p |\beta| + |x|^T D_q |\delta|.$$

Since

$$\begin{aligned} |D_p \Delta - \Delta D_q|_{ij} &= |(D_p)_{ii} \Delta_{ij} - \Delta_{ij} (D_q)_{jj}| = |\Delta_{ij}||(D_p)_{ii} - (D_q)_{jj}| \\ &\leq \operatorname{rad}([A])_{ij} |(D_p)_{ii} - (D_q)_{jj}| \\ &= |D_p \operatorname{rad}([A]) - \operatorname{rad}([A]) D_q|_{ij} \end{aligned}$$

and $|\beta| \leq \operatorname{rad}([b]),\ |\delta| \leq \operatorname{rad}([d])$ we finally end up with (3.12) .

Since $(A + A^T)/2$ is symmetric and in $[A]$ for any $A \in [A] = [A]^T$, and since $(b^{(1)} + b^{(2)})/2 \in [b]$ for any $b^{(1)},\ b^{(2)} \in [b]$ one sees immediately that

$$x \in \Sigma_{\text{sym}} \quad \text{if and only if} \quad Ax = b^{(1)} \quad \text{and} \quad A^T x = b^{(2)}$$

holds for some matrix $A \in [A]$ and for some vectors $b^{(1)},\ b^{(2)} \in [b]$. Therefore, Theorem 3.2 is the basis for the following equivalent description of Σ_{sym} which can be considered as an analogue of the Oettli–Prager criterion. Here the symbol $\prec_{\text{lex}}$ means strict lexicographic ordering of vectors, i.e., $u \prec_{\text{lex}} v$ if for some k we have $u_i = v_i,\ i < k$, and $u_k < v_k$.

Theorem 3.3 [24]

Let $[A] = [A]^T \in \mathbb{IR}^{n \times n},\ [b] \in \mathbb{IR}^n,\ x \in \mathbb{R}^n,\ r = \check{b} - \check{A}x$. Then $x \in \Sigma_{\text{sym}}$ if and only if the following system of inequalities holds.

$$\operatorname{rad}([A])\,|x| + \operatorname{rad}([b]) \ \geq \ |r| \tag{3.13}$$

$$\sum_{i,j=1}^{n} \operatorname{rad}([a]_{ij})\,|x_i x_j (p_i - q_j)| + \sum_{i=1}^{n} \operatorname{rad}([b]_i)\,|x_i (p_i + q_i)| \ \geq \ \left| \sum_{i=1}^{n} r_i x_i (p_i - q_i) \right| \tag{3.14}$$

for all vectors $p, q \in \{0,1\}^n \backslash \{0, e\}$ such that

$$p \prec_{\text{lex}} q \quad \textit{and} \quad (\, p = e - q \ \vee \ \exists\, i \ :\ p_i = q_i = 0 \,).$$

The system (3.14) consists of $(4^n - 3^n - 2^{n+1} + 3)/2$ inequalities.

Again inequality (3.14) can be rewritten in a dense form

$$|x|^T \cdot |D_p \operatorname{rad}([A]) - \operatorname{rad}([A])\, D_q| \cdot |x| + |x|^T (D_p + D_q)\operatorname{rad}([b]) \geq |x^T (D_p - D_q) r| \tag{3.15}$$

with the notation above.

Analogously to (3.12) an inequality for Σ_{sym} can be derived directly from

$$x^T D_p A x = x^T D_p b \quad \text{and} \quad x^T D_q A x = x^T A D_q x = x^T D_q b$$

for symmetric matrices A. At the end one obtains

$$|x|^T \cdot |D_p \operatorname{rad}([A]) - \operatorname{rad}([A])\, D_q| \cdot |x| + |x|^T |D_p - D_q| \operatorname{rad}([b]) \geq |x^T (D_p - D_q) r| \tag{3.16}$$

with a slightly smaller $\operatorname{rad}([b])$–summand as in (3.15) and – at the moment – without any restrictions on p and q. Although different, both sets of inequalities determine the same solution set when supplemented by (3.13): If x satisfies (3.13) and (3.15) then $x \in \Sigma_{\text{sym}}$ by Theorem 3.3, hence x satisfies (3.16) since we deduced this inequality only for such vectors. The converse is trivial since (3.15) follows immediately from (3.16) by virtue of $|D_p - D_q| \leq D_p + D_q$.

With the complementary vectors $\overline{p} = e - p,\ \overline{q} = e - q$ we can transform the matrix

$$R(p, q) = |D_p \operatorname{rad}([A]) - \operatorname{rad}([A]) D_q|$$

in the following way:

$$\begin{aligned} R(p,q) &= |D_p \operatorname{rad}([A]) - D_p \operatorname{rad}([A]) D_q - D_{\overline{p}} \operatorname{rad}([A]) D_q| \\ &= |D_p \operatorname{rad}([A])(I - D_q) - D_{\overline{p}} \operatorname{rad}([A]) D_q| \\ &= |D_p \operatorname{rad}([A]) D_{\overline{q}} - D_{\overline{p}} \operatorname{rad}([A]) D_q| \\ &= D_p \operatorname{rad}([A]) D_{\overline{q}} + D_{\overline{p}} \operatorname{rad}([A]) D_q, \end{aligned} \tag{3.17}$$

where for the last equality we proceeded entrywise and exploited $p + \overline{p} = e,\ p \in \{0, 1\}$, so that either $p_i = 1,\ \overline{p}_i = 0$ or vice versa. Therefore, at most one summand of $(R(p, q))_{ij}$ in (3.17) differs from zero.

The third equality in (3.17) implies

$$\begin{aligned} R(\overline{p}, \overline{q}) &= |D_{\overline{p}} \operatorname{rad}([A]) D_q - D_p \operatorname{rad}([A]) D_{\overline{q}}| \\ &= |D_p \operatorname{rad}([A]) D_{\overline{q}} - D_{\overline{p}} \operatorname{rad}([A]) D_q| = R(p, q). \end{aligned}$$

Moreover, $D_{\overline{p}} - D_{\overline{q}} = -(D_p - D_q)$, and $q \prec_{\text{lex}} p$ is equivalent to $\overline{p} \prec_{\text{lex}} \overline{q}$. Therefore, the inequality (3.16) also holds for $q \prec_{\text{lex}} p$ if it is true for $p \prec_{\text{lex}} q$.

If $p = 0$ then (3.16) reduces to

$$|x|^T \operatorname{rad}([A]) D_q |x| + |x|^T D_q \operatorname{rad}([b]) \geq |x^T D_q r|$$

which follows also from (3.13) and $|x^T D_q r| \leq |x|^T D_q |r|$ provided that the Oettli–Prager criterion holds, of course. The same is true for the case $q = 0$. Since $e = \overline{p}$ for $p = 0$ we can restrict to $p, q \in \{0, 1\}^n \backslash \{0, e\}$ if (3.13) is assumed.

We show now that all inequalities (3.16) with $p_i = q_i = 1$ for some indices i can be omitted. To this end choose p', q' such that

$$p'_i = \begin{cases} p_i, & \text{if } p_i q_i \neq 1 \\ 0 & \text{otherwise} \end{cases} \qquad q'_i = \begin{cases} q_i, & \text{if } p_i q_i \neq 1 \\ 0 & \text{otherwise} \end{cases}$$

and let $p'' = p - p'$. Then $p''_i = \begin{cases} 1, & \text{if } p_i = q_i = 1 \\ 0 & \text{otherwise} \end{cases}$, whence $p'' = q - q'$, $\overline{p} = \overline{p}' - p''$, $\overline{q} = \overline{q}' - p''$. Hence

$$\begin{aligned} R(p,q) &= D_{p'}\mathrm{rad}([A])D_{\overline{q}} + D_{p''}\mathrm{rad}([A])D_{\overline{q}} + D_{\overline{p}}\mathrm{rad}([A])D_{q'} + D_{\overline{p}}\mathrm{rad}([A])D_{p''} \\ &= D_{p'}\mathrm{rad}([A])D_{\overline{q}'} - D_{p'}\mathrm{rad}([A])D_{p''} + D_{p''}\mathrm{rad}([A])D_{\overline{q}} \\ &\quad + D_{\overline{p}'}\mathrm{rad}([A])D_{q'} - D_{p''}\mathrm{rad}([A])D_{q'} + D_{\overline{p}}\mathrm{rad}([A])D_{p''}. \end{aligned}$$

From this we get

$$\begin{aligned} |x|^T R(p,q)|x| &= |x|^T R(p',q')|x| \\ &\quad + |x|^T D_{p''}\mathrm{rad}([A])(-D_{p'} + D_{\overline{q}} - D_{q'} + D_{\overline{p}})|x| \\ &= |x|^T R(p',q')|x| + |x|^T D_{p''}\mathrm{rad}([A])(I - D_{p'+q} + I - D_{p+q'})|x| \\ &\geq |x|^T R(p',q')|x|, \end{aligned}$$

since $I - D_{p'+q} \geq O$ and $I - D_{p+q'} \geq O$. (Note that $p' + q \leq e$ and $p + q' \leq e$ hold.) Moreover, $D_p - D_q = D_{p'+p''} - D_{q'+p''} = D_{p'} + D_{p''} - (D_{q'} + D_{p''}) = D_{p'} - D_{q'}$. Therefore, the inequality (3.16) for p, q follows from that for p', q' and can be omitted. This shows that in (3.16) only vectors p, $q \in \{0,1\}^n \backslash \{0, e\}$ with $p \prec_{\mathrm{lex}} q$ and $p_i q_i = 0$, $i = 1, \ldots, n$, need to be considered if (3.13) is assumed.

This reduces the number of inequalities (3.16) further over that given in Theorem 3.3 : There are 3^n possibilities for inequalities with $(p_i, q_i) \in \{(0,0), (0,1), (1,0)\}$ for all $i = 1, \ldots, n$. Moreover, there are

$$\binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{n} = 2^n \tag{3.18}$$

possibilities of inequalities with $p = 0$, and the same amount of inequalities occurs for $q = 0$. (The lower number in the binomials counts the exact number of ones in q, and p, respectively.) Both numbers in (3.18) have to subtracted from 3^n. Since hereby the case $p = q = 0$ is subtracted twice we must add again a 1. Notice that we did not allow the case $p_i = q_i = 1$ and excluded the cases $p = 0$ and $q = 0$. This implies that we simultaneously excluded the cases $p = e$ and $q = e$. Taking into account $p \prec_{\mathrm{lex}} q$ we finally end up with

$$\frac{1}{2}(3^n - 2 \cdot 2^n + 1) \tag{3.19}$$

which is less than the number $(4^n - 3^n - 2^{n+1} + 3)/2$ in Theorem 3.3 if $n > 2$. (For $n = 1$ and $n = 2$ both bounds coincide.)

With (3.16) and (3.19) we can reformulate and improve Theorem 3.3 in the following way.

Theorem 3.4 [33]

Let $[A] = [A]^T \in \mathbb{IR}^{n\times n},\ [b] \in \mathbb{IR}^n,\ x \in \mathbb{R}^n,\ r = \check{b} - \check{A}x$. *Then* $x \in \Sigma_{\rm sym}$ *if and only if the following system of inequalities holds.*

$$\operatorname{rad}([A])\,|x| + \operatorname{rad}([b]) \ \geq\ |r| \tag{3.20}$$

$$|x|^T \cdot |D_p \operatorname{rad}([A]) - \operatorname{rad}([A])\, D_q| \cdot |x| + |x|^T |D_p - D_q| \operatorname{rad}([b]) \ \geq\ |x^T (D_p - D_q) r| \tag{3.21}$$

for all vectors $p, q \in \{0,1\}^n \backslash \{0, e\}$ *such that*

$$p \prec_{\rm lex} q \quad \textit{and} \quad p^T q = 0, \tag{3.22}$$

where $D_p = \operatorname{diag}(p)$, $D_q = \operatorname{diag}(q) \in \mathbb{R}^{n\times n}$.

The system (3.21) *consists of* $(3^n - 2^{n+1} + 1)/2$ *inequalities.*

The system (3.20) is just the Oettli–Prager criterion, and the system (3.21) are the additional inequalities caused by symmetry. Thus for $n = 2$ there is only one additional inequality and for $n = 3$ already six.

The proof of one direction of Theorem 3.4 is already contained in the lines preceding the theorem. For the converse direction we need the some preparations:

Without loss of generality we may assume that x is an element of the first orthant O_1. Otherwise replace $[A]$, $[b]$, x by $[B] = D_x[A]D_x = [B]^T$, $[c] = D_x\,[b]$, $z = D_x\,x = |x| \in O_1$ with $D_x = \operatorname{diag}(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{n\times n}$, where $\sigma_i = 1$ if $\operatorname{sign}(x_i) = 0$, and $\sigma_i = \operatorname{sign}(x_i)$ otherwise. Then $\operatorname{rad}([B]) = \operatorname{rad}([A])$, $\operatorname{rad}([c]) = \operatorname{rad}([b])$, $|\check{c} - \check{B}z| = |D_x(\check{b} - \check{A}x)| = |D_x r| = |r|$, $z^T(D_p - D_q)(\check{c} - \check{B}z) = x^T D_x (D_p - D_q) D_x r = x^T (D_p - D_q) r$. Therefore, the inequalities (3.20), (3.21) hold for $[B]$, $[c]$, z if they hold for $[A]$, $[b]$, x, and vice versa.

The Theorems 3.2 – 3.4 are formulated using midpoint and radius of $[A]$ and $[b]$. In the following we will show how an endpoint formulation looks like if $x \in O_1$. With $p_i q_i = 0,\ i = 1, \ldots, n$, we get $|D_p - D_q| = D_p + D_q$ in (3.16) and (3.21). Omitting there the absolute value in the righthand side and using (3.17) results in

$$\begin{aligned} & x^T D_q r + x^T D_q \operatorname{rad}([b]) + x^T D_q \operatorname{rad}([A]) D_{\overline{p}} x \\ & \qquad \geq\ x^T D_p r - x^T D_p \operatorname{rad}([b]) - x^T D_p \operatorname{rad}([A]) D_{\overline{q}} x \end{aligned} \tag{3.23}$$

or, equivalently,

$$x^T D_q (\overline{b} - \underline{A} D_{\overline{p}} x) \geq x^T D_p (\underline{b} - \overline{A} D_{\overline{q}} x). \tag{3.24}$$

Here, we used $\check{A} = \check{A}(D_p + D_{\overline{p}}) = \check{A}(D_q + D_{\overline{q}})$ and dropped the term $x^T D_q \check{A} D_p x = x^T D_p \check{A} D_q x$ which arises on both sides of (3.23). The second inequality hidden behind (3.16) and (3.21) reads

$$x^T D_q (\underline{b} - \overline{A} D_{\overline{p}} x) \leq x^T D_p (\overline{b} - \underline{A} D_{\overline{q}} x) \tag{3.25}$$

which is (3.24) if one interchanges p and q (forbidden if our assumption $p \prec_{\text{lex}} q$ shall hold!).

The inequalities (3.24) and (3.25) will form the starting point for our proof of Theorem 3.4 . Therefore, we will introduce some short notation and list some properties which we will apply without further reference. Note that the meaning of q' below is now different from that which we used previously.

Definition 3.1

Let $p, \tilde{p}, q, q', \tilde{q}, \tilde{q}' \in \{0,1\}^n$ and let ' $\circ$ ' denote the Hadamard product. Then we define

$$\overline{p} = e - p, \quad \overline{q} = e - q, , \quad q' = (0, 0, q_3, \ldots, q_n)^T, \quad \tilde{q}' = (0, 0, \tilde{q}_3, \ldots, \tilde{q}_n)^T,$$

$$\begin{array}{llllllll}
p_c = p \circ \tilde{p}, & p_r = p - p_c, & q_c = q \circ \tilde{q}, & q_r = q - q_c, \\
\tilde{p}_c = \tilde{p} \circ p, & \tilde{p}_r = \tilde{p} - \tilde{p}_c, & \tilde{q}_c = \tilde{q} \circ q, & \tilde{q}_r = \tilde{q} - \tilde{q}_c, \\
 & & q'_c = q' \circ \tilde{q}', & q'_r = q' - q'_c \\
 & & \tilde{q}'_c = \tilde{q}' \circ q', & \tilde{q}'_r = \tilde{q}' - \tilde{q}'_c \\
p_C = p \circ \tilde{q}, & p_R = p - p_C, & q_C = q \circ \tilde{p}, & q_R = q - q_C, \\
\tilde{p}_C = \tilde{p} \circ q, & \tilde{p}_R = \tilde{p} - \tilde{p}_C, & \tilde{q}_C = \tilde{q} \circ p, & \tilde{q}_R = \tilde{q} - \tilde{q}_C,
\end{array}$$

Notice the difference between the subscripts 'c' and 'C', 'r' and 'R'. The subscripts 'c' and 'C' remind of 'common components', i.e., components which are one for both operands simultaneously. The subscripts 'r' and 'R' mean 'remaining components'.

Definition 3.2

Let $[A] = [A]^T \in \mathbb{IR}^{n \times n}$, $[b] \in \mathbb{IR}^n$, $x \in \mathbb{R}^n$. Moreover, let $p, q \in \{0,1\}^n$. Then we define

$$\begin{array}{llll}
\underline{L}_p^q = x^T D_p (\underline{b} - \overline{A} D_{\overline{q}} x), & \dot{\underline{L}}_p^q = -x^T D_p \overline{A} D_q x, \\
\overline{L}_p^q = x^T D_p (\overline{b} - \underline{A} D_{\overline{q}} x), & \dot{\overline{L}}_p^q = -x^T D_p \underline{A} D_q x.
\end{array}$$

The position of the bar in $\underline{L}_p^q$, $\overline{L}_p^q$ is the same as with $\underline{b}$, $\overline{b}$. Notice the missing bar at 'q' when defining $\dot{\underline{L}}_p^q$, $\dot{\overline{L}}_p^q$. If $p_i = q_j = 1$ the entry $\overline{a}_{ij}$ is missing when computing $\underline{L}_p^q$ while it is present when computing $\dot{\underline{L}}_p^q$.

Lemma 3.2

With the assumptions and the notations in the Definitions 3.1 and 3.2 we get the following properties which (mostly) hold for q and $\overline{L}_p^q$ analogously.

a) $p_c = \tilde{p}_c, \qquad p + \tilde{p}_r = \tilde{p} + p_r, \qquad p_C = \tilde{q}_C, \qquad p + \tilde{q}_R = \tilde{q} + p_R.$

b) The inequalities (3.24) and (3.25) can be written as

$$\underline{L}_p^q \leq \overline{L}_q^p, \qquad \underline{L}_q^p \leq \overline{L}_p^q. \tag{3.26}$$

c) If $p = 0$ then $\underline{L}_p^q = \dot{\underline{L}}_p^{\,q} = 0$ for arbitrary q.
If $q = 0$ then $\dot{\underline{L}}_p^{\,q} = 0$ for arbitrary p.
If $p = e^{(i)}$, $q = e$ then $\underline{L}_p^q = x_i \underline{b}_i$.

d) Let $p = 0$, $q = e^{(i)}$, $x \in O_1$ (closed first orthant). Multiplying the i–th Oettli–Prager inequality in (3.20) by x_i results in two particular inequalities (3.26) which read

$$\left.\begin{aligned} \underline{L}_0^{e^{(i)}} = 0 \le \overline{L}_{e^{(i)}}^0 &= x_i \left(\overline{b}_i - \textstyle\sum_{j=1}^n \underline{a}_{ij} x_j \right), \\ \underline{L}_{e^{(i)}}^0 = x_i \left(\underline{b}_i - \textstyle\sum_{j=1}^n \overline{a}_{ij} x_j \right) &\le 0 = \overline{L}_0^{e^{(i)}}. \end{aligned}\right\} \tag{3.27}$$

If $x_i > 0$ the i–th Oettli–Prager inequality is equivalent to (3.27).

e) $\underline{\dot{L}}_p^q = \underline{\dot{L}}_q^p = \underline{\dot{L}}_{p_c}^q + \underline{\dot{L}}_{p_r}^q \quad \left(\le \overline{\dot{L}}_p^{\,q} \ \text{ if } x \in O_1 \right)$.

f) $\underline{L}_p^q = \underline{L}_p^{q+\tilde{q}_r} + \underline{\dot{L}}_{\tilde{q}_r}^p = \underline{L}_{p_c}^q + \underline{L}_{p_r}^q \le \overline{L}_p^q$.

g) $\underline{L}_{p_c}^q + \underline{L}_{\tilde{q}_r}^{\tilde{p}} = \underline{L}_{p_c}^{q+\tilde{q}_r} + \underline{L}_{\tilde{q}_r}^{\tilde{p}_r}, \qquad \underline{L}_{q_C}^p + \underline{L}_{q_C}^{\tilde{q}} = \underline{L}_{q_C}^{p+\tilde{q}_R} + \underline{L}_{q_C}^{\tilde{q}_C}$.

The proof of Lemma 3.2 is based on simple calculations and is therefore omitted.

Proof of Theorem 3.4

In order to prove the converse direction of Theorem 3.4 we assume that x satisfies (3.20) and (3.26) for $0 \ne p \prec_{\text{lex}} q$ with $p^T q = 0$ (which implies $q \ne e$). As various previous remarks show the inequalities in (3.26) then hold for arbitrary $p, q \in \{0,1\}^n$. Moreover, w.l.o.g. we assume $x \in O_1$.

We present now the ideas of our proof which is based on a reversed Fourier–Motzkin elimination technique as used earlier in this section: Successively for each index pair (i,j) with $i < j$ and $x_i x_j > 0$ we will replace the entries $[a]_{ij}$ and $[a]_{ji} = [a]_{ij}$ in $[A]$ simultaneously by some point intervals $[a_{ij}, a_{ij}]$, $[a_{ji}, a_{ji}]$ with $a_{ij} = a_{ji} \in [a]_{ij}$ so that the inequalities (3.20), (3.21), (3.26) still hold. At the end of the complete replacement process we will apply the Oettli–Prager Theorem to the resulting final matrix $[A]^{\text{new}} = ([A]^{\text{new}})^T \subseteq [A]$ and to $[b]$ in order to obtain a (possibly unsymmetric) matrix $\tilde{A} \in [A]^{\text{new}}$ and a vector $\tilde{b} \in [b]$ with $\tilde{A}x = \tilde{b}$. From $\tilde{A}$ we can construct at once a symmetric matrix $\tilde{A}^{\text{sym}} \in [A]$ which satisfies

$$\tilde{A}^{\text{sym}} x = \tilde{A} x = \tilde{b} \tag{3.28}$$

and thus implies $x \in \Sigma_{\text{sym}}$.

For ease of notation we show how

$$\underline{a}_{12}, \ \underline{a}_{21}, \ \overline{a}_{12}, \ \overline{a}_{21} \tag{3.29}$$

and $[a]_{12} = [a]_{21}$ can be replaced assuming $x_1 x_2 > 0$. (If $x_1 x_2 = 0$ then $[a]_{12} = [a]_{21}$ remains unchanged for the moment, and another entry is considered.) By virtue of $x_1 x_2 > 0$ and Lemma 3.2 d) we can replace the first two Oettli–Prager inequalities

equivalently by those in (3.27) for $i = 1$ and $i = 2$. Therefore, we will enlarge the choice of p and q in Theorem 3.4 by the cases $p = 0$, $q = e^{(1)}$ and $p = 0$, $q = e^{(2)}$ without mentioning it furthermore. This extension avoids the study of special subcases. Now we consider only those inequalities (3.26) which contain at least one entry of (3.29) explicitly. Taking into account (3.22) there are only three cases for this :

Case 1: $q = (1, 0, *)^T$, $p = (0, 0, *)^T$
Case 2: $q = (0, 1, *)^T$, $p = (0, 0, *)^T$
Case 3: $q = (1, 1, *)^T$, $p = (0, 0, *)^T$

Here '$*$' replaces the remaining components of p and q. Notice that for $q = (1, 0, *)^T$, $p = (0, 1, *)^T$ the inequalities (3.26) do not contain one of the entries in (3.29) explicitly. Moreover, the entries $\underline{a}_{12} = \underline{a}_{21}$ and $\overline{a}_{12} = \overline{a}_{21}$ cannot appear in one and the same of the two inequalities (3.26) because of $p^T q = 0$.

Our first step consists of isolating the entries (3.29) in (3.26). With the notation of Definition 3.1 and $\underline{a}_{12} = \underline{a}_{21}$, $\overline{a}_{12} = \overline{a}_{21}$ we get

$$\underline{a}_{12} x_1 x_2 \leq \overline{L}_{e^{(1)}}^{e^{(2)}+p} + \overline{L}_{q'}^{p} - \underline{L}_{p}^{q} \quad \text{(Case 1)} \tag{3.30}$$

$$\underline{a}_{12} x_1 x_2 \leq \overline{L}_{e^{(2)}}^{e^{(1)}+p} + \overline{L}_{q'}^{p} - \underline{L}_{p}^{q} \quad \text{(Case 2)} \tag{3.31}$$

$$\underline{a}_{12} x_1 x_2 \leq \left(\overline{L}_{e^{(1)}}^{e^{(2)}+p} + \overline{L}_{e^{(2)}}^{e^{(1)}+p} + \overline{L}_{q'}^{p} - \underline{L}_{p}^{q} \right) /2 \quad \text{(Case 3)} \tag{3.32}$$

and

$$\underline{L}_{e^{(1)}}^{e^{(2)}+p} + \underline{L}_{q'}^{p} - \overline{L}_{p}^{q} \leq \overline{a}_{12} x_1 x_2 \quad \text{(Case 1)} \tag{3.33}$$

$$\underline{L}_{e^{(2)}}^{e^{(1)}+p} + \underline{L}_{q'}^{p} - \overline{L}_{p}^{q} \leq \overline{a}_{12} x_1 x_2 \quad \text{(Case 2)} \tag{3.34}$$

$$\left(\underline{L}_{e^{(1)}}^{e^{(2)}+p} + \underline{L}_{e^{(2)}}^{e^{(1)}+p} + \underline{L}_{q'}^{p} - \overline{L}_{p}^{q} \right) /2 \leq \overline{a}_{12} x_1 x_2 \quad \text{(Case 3)} \tag{3.35}$$

If we can show (which we will do at the end of the proof) that each left–hand side of (3.33) – (3.35) is less or equal than each right–hand side of (3.30) – (3.32) (with another admissible choice of p, q) then the same holds for the maximum M_ℓ of all such left–hand sides as compared with the minimum m_r of all such right–hand sides. If $m_r > \overline{a}_{12} x_1 x_2$ we redefine it by $\overline{a}_{12} x_1 x_2$ knowing by (3.33) – (3.35) that $M_\ell \leq \overline{a}_{12} x_1 x_2$. Analogously, if $M_\ell < \underline{a}_{12} x_1 x_2$ we redefine it by $\underline{a}_{12} x_1 x_2$. Now we choose any number from $[M_\ell, m_r]$. Obviously, it is representable as $a_{12} x_1 x_2$ with some number $a_{12} \in [a]_{12}$. The inequalities (3.30) – (3.35) hold with a_{12} in place of $\underline{a}_{12}$, $\overline{a}_{12}$, and so do the inequalities (3.26) if we define $a_{21} = a_{12}$ and replace the entries (3.29) correspondingly. Replacing the entries $[a]_{12}$, $[a]_{21}$ in $[A]$ by $a_{12} = a_{21}$ results in a matrix $[A]'$ for which the assumptions of Theorem 3.4 are also satisfied. It forms the starting point of our next replacement. Repeating this process for all entries $[a]_{ij}$, $i < j$, with $x_i x_j > 0$ we finally end up with the matrix $[A]^{\text{new}}$ which we already mentioned at the beginning. It has degenerate symmetric entries $a_{ij} = a_{ji} \in [a]_{ij}$ whenever $x_i x_j > 0$ is true. Moreover, it satisfies the inequalities (3.20), (3.21). Therefore, a matrix $\tilde{A} = (\tilde{a}_{ij}) \in [A]^{\text{new}}$ and a vector $\tilde{b} \in [b]$ exist with $\tilde{A}x = \tilde{b}$. Trivially, $\tilde{a}_{ij} = \tilde{a}_{ji} = a_{ij}$ if $x_i x_j > 0$. If $x_i = 0$ the value of $\tilde{a}_{ji}$ does not

matter in the product $\tilde{A}x$. Therefore, we may replace it by $\tilde{a}_{ij}$, a step in view of symmetry. Similarly, if $x_j = 0$ we replace $\tilde{a}_{ij}$ by $\tilde{a}_{ji}$ ending up with a symmetric matrix $\tilde{A} \in [A]$ which satisfies (3.28) and finishes the proof.

It remains to show that each left–hand side of (3.33) – (3.35) is less or equal than each right–hand side of (3.30) – (3.32). To this end we have to combine each right–hand side of (3.30) – (3.32) with each left–hand side of (3.33) – (3.35) which leads to nine combinations.

1) Case 1 vs. Case 1:

Let $\tilde{p},\ \tilde{q}$ and p, q be chosen according to Case 1, independently of each other. We have to show that

$$\underline{L}_{e^{(1)}}^{e^{(2)}+\tilde{p}} + \underline{L}_{\tilde{q}'}^{\tilde{p}} - \overline{L}_{\tilde{p}}^{\tilde{q}} \leq \overline{L}_{e^{(1)}}^{e^{(2)}+p} + \overline{L}_{q'}^{p} - \underline{L}_{p}^{q} \tag{3.36}$$

holds. With the notation of Definition 3.1 and with Lemma 3.2 we get

$$\begin{aligned}
\underline{L}_{e^{(1)}}^{e^{(2)}+\tilde{p}} + \underline{L}_{\tilde{q}'}^{\tilde{p}} + \underline{L}_{p}^{q} &= \underline{L}_{e^{(1)}}^{e^{(2)}+\tilde{p}} + \underline{L}_{\tilde{q}'}^{\tilde{p}} + \underline{L}_{p_c}^{q} + \underline{L}_{p_r}^{e^{(1)}+q'} \\
&= \underline{L}_{e^{(1)}}^{e^{(2)}+\tilde{p}+p_r} + \underline{L}_{\tilde{q}'}^{\tilde{p}} + \underline{L}_{p_c}^{q} + \underline{L}_{p_r}^{q'} \\
&= \underline{L}_{e^{(1)}}^{e^{(2)}+\tilde{p}+p_r} + \underline{L}_{\tilde{q}'_c}^{\tilde{p}} + (\underline{L}_{\tilde{q}'_r}^{\tilde{p}} + \underline{L}_{p_c}^{q}) + \underline{L}_{p_r}^{q'} \\
&= \underline{L}_{e^{(1)}}^{e^{(2)}+\tilde{p}+p_r} + \underline{L}_{\tilde{q}'_c}^{\tilde{p}} + (\underline{L}_{\tilde{q}'_r}^{\tilde{p}_r} + \underline{L}_{p_c}^{q+\tilde{q}'_r}) + \underline{L}_{p_r}^{q'} \\
&= \underline{L}_{e^{(1)}}^{e^{(2)}+\tilde{p}+p_r} + \underline{L}_{\tilde{q}'_c}^{\tilde{p}+p_r} + \underline{L}_{\tilde{q}'_r}^{\tilde{p}_r} + \underline{L}_{p_c}^{q+\tilde{q}'_r} + \underline{L}_{p_r}^{q'_r}.
\end{aligned}$$

Analogously we obtain

$$\begin{aligned}
\overline{L}_{e^{(1)}}^{e^{(2)}+p} + \overline{L}_{q'}^{p} + \overline{L}_{\tilde{p}}^{\tilde{q}} &= \overline{L}_{e^{(1)}}^{e^{(2)}+p+\tilde{p}_r} + \overline{L}_{q'_c}^{p+\tilde{p}_r} + \overline{L}_{q'_r}^{p_r} + \overline{L}_{\tilde{p}_c}^{\tilde{q}+q'_r} + \overline{L}_{\tilde{p}_r}^{\tilde{q}'_r} \\
&= \overline{L}_{e^{(1)}}^{e^{(2)}+\tilde{p}+p_r} + \overline{L}_{q'_c}^{\tilde{p}+p_r} + \overline{L}_{\tilde{p}_r}^{\tilde{q}'_r} + \overline{L}_{\tilde{p}_c}^{q+\tilde{q}'_r} + \overline{L}_{q'_r}^{p_r}.
\end{aligned}$$

For the last formula we used the equality $q'_r = q_r,\ \tilde{q}'_r = \tilde{q}_r$ which holds by virtue of the particular form of q and $\tilde{q}$ in Case 1 .

Comparing both final expressions and using Lemma 3.2 b) and f) proves (3.36).

2) Case 2 vs. Case 2 is proved analogously.

3) Case 1 vs. Case 2:

Let $\tilde{p},\ \tilde{q}$ be chosen according to Case 1, and let p, q be chosen according to Case 2. We have to show that

$$\underline{L}_{e^{(1)}}^{e^{(2)}+\tilde{p}} + \underline{L}_{\tilde{q}'}^{\tilde{p}} - \overline{L}_{\tilde{p}}^{\tilde{q}} \leq \overline{L}_{e^{(2)}}^{e^{(1)}+p} + \overline{L}_{q'}^{p} - \underline{L}_{p}^{q} \tag{3.37}$$

holds. With $q_R = e^{(2)} + q'_R,\ \tilde{q}_R = e^{(1)} + \tilde{q}'_R$ and $p_C = \tilde{q}_C = \tilde{q}'_C$ we obtain

$$
\begin{aligned}
&\underline{L}^{e^{(2)}+\tilde{p}}_{e^{(1)}} + \underline{L}^{\tilde{p}}_{\tilde{q}'} + \underline{L}^{q}_{p} \\
&= \underline{L}^{e^{(2)}+\tilde{p}}_{e^{(1)}} + \underline{L}^{\tilde{p}}_{\tilde{q}'_C} + \underline{L}^{\tilde{p}}_{\tilde{q}'_R} + \underline{L}^{q}_{p_C} + \underline{L}^{q}_{p_R} \\
&= (\underline{L}^{e^{(2)}+\tilde{p}+q'_R}_{e^{(1)}} + \underline{\dot{L}}^{q'_R}_{e^{(1)}}) + (\underline{L}^{\tilde{p}}_{p_C} + \underline{L}^{q}_{p_C}) + (\underline{L}^{\tilde{p}+q'_R}_{\tilde{q}'_R} + \underline{\dot{L}}^{q'_R}_{\tilde{q}'_R}) + (\underline{L}^{q+\tilde{p}_R}_{p_R} + \underline{\dot{L}}^{\tilde{p}_R}_{p_R}) \\
&= \underline{L}^{e^{(2)}+\tilde{p}+q'_R}_{e^{(1)}} + (\underline{L}^{\tilde{p}_C}_{p_C} + \underline{L}^{q+\tilde{p}_R}_{p_C}) + (\underline{L}^{e^{(2)}+\tilde{p}+q'_R}_{\tilde{q}'_R} + \underline{\dot{L}}^{\tilde{q}'_R}_{e^{(2)}} + \underline{L}^{q+\tilde{p}_R}_{p_R}) + \underline{\dot{L}}^{q'_R}_{e^{(1)}} + \underline{\dot{L}}^{q'_R}_{\tilde{q}'_R} + \underline{\dot{L}}^{\tilde{p}_R}_{p_R} \\
&= \underline{L}^{q+\tilde{p}_R}_{\tilde{q}+p_R} + \underline{L}^{\tilde{p}_C}_{p_C} + \underline{\dot{L}}^{q'_R}_{e^{(1)}} + \underline{\dot{L}}^{\tilde{q}'_R}_{e^{(2)}} + \underline{\dot{L}}^{q'_R}_{\tilde{q}'_R} + \underline{\dot{L}}^{\tilde{p}_R}_{p_R}.
\end{aligned}
$$

Analogously we obtain

$$
\begin{aligned}
\overline{L}^{e^{(1)}+p}_{e^{(2)}} + \overline{L}^{p}_{q'} + \overline{L}^{\tilde{q}}_{\tilde{p}} &= \overline{L}^{\tilde{q}+p_R}_{q+\tilde{p}_R} + \overline{L}^{p_C}_{\tilde{p}_C} + \dot{\overline{L}}^{\tilde{q}'_R}_{e^{(2)}} + \dot{\overline{L}}^{q'_R}_{e^{(1)}} + \dot{\overline{L}}^{\tilde{q}'_R}_{q'_R} + \dot{\overline{L}}^{p_R}_{\tilde{p}_R} \\
&= \overline{L}^{\tilde{q}+p_R}_{q+\tilde{p}_R} + \overline{L}^{p_C}_{\tilde{p}_C} + \dot{\overline{L}}^{q'_R}_{e^{(1)}} + \dot{\overline{L}}^{\tilde{q}'_R}_{e^{(2)}} + \dot{\overline{L}}^{q'_R}_{\tilde{q}'_R} + \dot{\overline{L}}^{\tilde{p}_R}_{p_R}
\end{aligned}
$$

which proves (3.37) as above.

4) Case 3 vs. Case 3:

Let $\tilde{p},\ \tilde{q}$ and p, q be chosen according to Case 3. We have to show that

$$
\underline{L}^{e^{(2)}+\tilde{p}}_{e^{(1)}} + \underline{L}^{e^{(1)}+\tilde{p}}_{e^{(2)}} + \underline{L}^{\tilde{p}}_{\tilde{q}'} - \overline{L}^{\tilde{q}}_{\tilde{p}} \leq \overline{L}^{e^{(2)}+p}_{e^{(1)}} + \overline{L}^{e^{(1)}+p}_{e^{(2)}} + \overline{L}^{p}_{q'} - \underline{L}^{q}_{p} \tag{3.38}
$$

holds. With $\tilde{q}_C = \tilde{q}'_C = p_C$ we obtain

$$
\begin{aligned}
&\underline{L}^{e^{(2)}+\tilde{p}}_{e^{(1)}} + \underline{L}^{e^{(1)}+\tilde{p}}_{e^{(2)}} + \underline{L}^{\tilde{p}}_{\tilde{q}'} + \underline{L}^{q}_{p} \\
&= (\underline{L}^{q+\tilde{p}_R}_{e^{(1)}} + \underline{\dot{L}}^{e^{(1)}+q'_R}_{e^{(1)}}) + (\underline{L}^{q+\tilde{p}_R}_{e^{(2)}} + \underline{\dot{L}}^{e^{(2)}+q'_R}_{e^{(2)}}) + (\underline{L}^{\tilde{p}}_{\tilde{q}'_C} + \underline{L}^{\tilde{p}}_{\tilde{q}'_R}) + (\underline{L}^{q}_{p_C} + \underline{L}^{q}_{p_R}) \\
&= \underline{L}^{q+\tilde{p}_R}_{e^{(1)}+e^{(2)}} + (\underline{L}^{q+\tilde{p}_R}_{p_C} + \underline{L}^{\tilde{p}_C}_{p_C}) + (\underline{L}^{\tilde{p}+q_R}_{\tilde{q}'_R} + \underline{\dot{L}}^{q_R}_{\tilde{q}'_R}) + (\underline{L}^{q+\tilde{p}_R}_{p_R} + \underline{\dot{L}}^{\tilde{p}_R}_{p_R}) \\
&\quad + \underline{\dot{L}}^{e^{(1)}+q'_R}_{e^{(1)}} + \underline{\dot{L}}^{e^{(2)}+q'_R}_{e^{(2)}} \\
&= \underline{L}^{q+\tilde{p}_R}_{\tilde{q}+p_R} + \underline{L}^{\tilde{p}_C}_{p_C} + \underline{\dot{L}}^{e^{(1)}+q'_R}_{e^{(1)}} + \underline{\dot{L}}^{e^{(2)}+q'_R}_{e^{(2)}} + \underline{\dot{L}}^{q_R}_{\tilde{q}'_R} + \underline{\dot{L}}^{\tilde{p}_R}_{p_R} \\
&= \underline{L}^{q+\tilde{p}_R}_{\tilde{q}+p_R} + \underline{L}^{\tilde{p}_C}_{p_C} + (\underline{\dot{L}}^{e^{(1)}}_{e^{(1)}} + \underline{\dot{L}}^{e^{(2)}}_{e^{(2)}} + \underline{\dot{L}}^{e^{(1)}+e^{(2)}}_{q'_R}) + (\underline{\dot{L}}^{e^{(1)}+e^{(2)}}_{\tilde{q}'_R} + \underline{\dot{L}}^{q'_R}_{\tilde{q}'_R}) + \underline{\dot{L}}^{\tilde{p}_R}_{p_R}.
\end{aligned}
$$

Analogously we obtain

$$
\begin{aligned}
&\overline{L}^{e^{(2)}+p}_{e^{(1)}} + \overline{L}^{e^{(1)}+p}_{e^{(2)}} + \overline{L}^{p}_{q'} + \overline{L}^{\tilde{q}}_{\tilde{p}} \\
&= \overline{L}^{\tilde{q}+p_R}_{q+\tilde{p}_R} + \overline{L}^{p_C}_{\tilde{p}_C} + (\dot{\overline{L}}^{e^{(1)}}_{e^{(1)}} + \dot{\overline{L}}^{e^{(2)}}_{e^{(2)}} + \dot{\overline{L}}^{e^{(1)}+e^{(2)}}_{\tilde{q}'_R}) + (\dot{\overline{L}}^{e^{(1)}+e^{(2)}}_{q'_R} + \dot{\overline{L}}^{\tilde{q}'_R}_{q'_R}) + \dot{\overline{L}}^{p_R}_{\tilde{p}_R}
\end{aligned}
$$

which proves (3.38) as above.

5) Case 1 vs. Case 3:

Let $\tilde{p},\ \tilde{q}$ be chosen according to Case 1, and let p, q be chosen according to Case 3. We have to show that

$$
2\underline{L}^{e^{(2)}+\tilde{p}}_{e^{(1)}} + 2\underline{L}^{\tilde{p}}_{\tilde{q}'} - 2\overline{L}^{\tilde{q}}_{\tilde{p}} \leq \overline{L}^{e^{(2)}+p}_{e^{(1)}} + \overline{L}^{e^{(1)}+p}_{e^{(2)}} + \overline{L}^{p}_{q'} - \underline{L}^{q}_{p} \tag{3.39}
$$

holds. With $\tilde{q}_r = \tilde{q}'_r,\ q = e^{(1)} + (e^{(2)} + q')$ we obtain

$$
\begin{aligned}
&2\underline{L}_{e^{(1)}}^{e^{(2)}+\tilde{p}} + 2\underline{L}_{\tilde{q}'}^{\tilde{p}} + \underline{L}_{p}^{q} \\
&\quad = (\underline{L}_{e^{(1)}}^{e^{(2)}+\tilde{p}} + \underline{L}_{\tilde{q}'}^{\tilde{p}} + \underline{L}_{p}^{q}) + (\underline{L}_{e^{(1)}}^{e^{(2)}+\tilde{p}} + \underline{L}_{\tilde{q}'}^{\tilde{p}}) \\
&\quad = (\underline{L}_{e^{(1)}}^{e^{(2)}+\tilde{p}+p_r} + \underline{L}_{\tilde{q}'_c}^{\tilde{p}} + \underline{L}_{\tilde{q}'_r}^{\tilde{p}} + \underline{L}_{p_c}^{q} + \underline{L}_{p_r}^{e^{(2)}+q'}) + (\underline{L}_{e^{(1)}}^{e^{(2)}+\tilde{p}} + \underline{L}_{\tilde{q}'}^{\tilde{p}}) \\
&\quad = (\underline{L}_{e^{(1)}}^{e^{(2)}+\tilde{p}+p_r} + \underline{L}_{\tilde{q}'_c}^{\tilde{p}} + \underline{L}_{\tilde{q}'_r}^{\tilde{p}_r} + \underline{L}_{p_c}^{q+\tilde{q}'_r} + \underline{L}_{p_r}^{e^{(2)}+q'}) + (\underline{L}_{e^{(1)}}^{e^{(2)}+\tilde{p}} + \underline{L}_{\tilde{q}'}^{\tilde{p}}) \\
&\quad = (\underline{L}_{e^{(1)}}^{e^{(2)}+\tilde{p}+p_r} + \underline{L}_{\tilde{q}'_c}^{\tilde{p}+p_r} + \underline{L}_{p_c}^{q+\tilde{q}_r} + \underline{L}_{\tilde{q}'_r}^{\tilde{p}_r}) + (\underline{L}_{e^{(1)}}^{e^{(2)}+\tilde{p}} + \underline{L}_{\tilde{q}'}^{\tilde{p}} + \underline{L}_{p_r}^{e^{(2)}+q'_r}).
\end{aligned}
$$

Similarly, with $q_r = e^{(2)} + q'_r$, we get

$$
\begin{aligned}
&\overline{L}_{e^{(1)}}^{e^{(2)}+p} + \overline{L}_{e^{(2)}}^{e^{(1)}+p} + \overline{L}_{q'}^{p} + 2\overline{L}_{\tilde{p}}^{\tilde{q}} \\
&\quad = (\overline{L}_{e^{(1)}}^{e^{(2)}+p} + \overline{L}_{q'}^{p} + \overline{L}_{\tilde{p}}^{\tilde{q}}) + (\overline{L}_{e^{(2)}}^{e^{(1)}+p} + \overline{L}_{\tilde{p}}^{\tilde{q}}) \\
&\quad = (\overline{L}_{e^{(1)}}^{e^{(2)}+p+\tilde{p}_r} + \overline{L}_{q'_c}^{p} + \overline{L}_{q'_r}^{p} + \overline{L}_{\tilde{p}_c}^{\tilde{q}} + \overline{L}_{\tilde{p}_r}^{\tilde{q}'}) + (\overline{L}_{e^{(2)}}^{e^{(1)}+p} + \overline{L}_{\tilde{p}}^{\tilde{q}}) \\
&\quad = (\overline{L}_{e^{(1)}}^{e^{(2)}+p+\tilde{p}_r} + \overline{L}_{q'_c}^{p} + \overline{L}_{q'_r}^{p_r} + \overline{L}_{p_c}^{\tilde{q}+q'_r} + \overline{L}_{\tilde{p}_r}^{\tilde{q}'}) + (\overline{L}_{e^{(2)}}^{e^{(1)}+p} + \overline{L}_{\tilde{p}}^{\tilde{q}}) \\
&\quad = (\overline{L}_{e^{(1)}}^{e^{(2)}+p+\tilde{p}_r} + \overline{L}_{q'_c}^{p} + \overline{L}_{p_c}^{q+\tilde{q}_r} + \dot{\overline{L}}_{p_c}^{e^{(2)}} + \overline{L}_{\tilde{p}_r}^{\tilde{q}'}) + (\overline{L}_{e^{(2)}}^{e^{(1)}+p} + \overline{L}_{q'_r}^{p_r} + \overline{L}_{\tilde{p}}^{\tilde{q}}) \\
&\quad = (\overline{L}_{e^{(1)}}^{e^{(2)}+p+\tilde{p}_r} + \overline{L}_{q'_c}^{p+\tilde{p}_r} + \overline{L}_{p_c}^{q+\tilde{q}_r} + \overline{L}_{\tilde{p}_r}^{\tilde{q}'_r}) + (\overline{L}_{e^{(2)}}^{e^{(1)}+p_r} + \overline{L}_{q'_r}^{p_r} + \overline{L}_{\tilde{p}}^{\tilde{q}}) \\
&\quad = (\overline{L}_{e^{(1)}}^{e^{(2)}+\tilde{p}+p_r} + \overline{L}_{\tilde{q}'_c}^{\tilde{p}+p_r} + \overline{L}_{p_c}^{q+\tilde{q}_r} + \overline{L}_{\tilde{p}_r}^{\tilde{q}'_r}) + (\overline{L}_{e^{(2)}}^{e^{(1)}+p_r} + \overline{L}_{q'_r}^{p_r} + \overline{L}_{\tilde{p}}^{\tilde{q}})
\end{aligned}
$$

Comparing the final expressions one sees that (3.39) certainly holds if

$$
\underline{L}_{e^{(1)}}^{e^{(2)}+\tilde{p}} + \underline{L}_{\tilde{q}'}^{\tilde{p}} + \underline{L}_{p_r}^{e^{(2)}+q'_r} \leq \overline{L}_{e^{(2)}}^{e^{(1)}+p_r} + \overline{L}_{q'_r}^{p_r} + \overline{L}_{\tilde{p}}^{\tilde{q}} \tag{3.40}
$$

is true. But this is just Case 1 vs. Case 2 with $\tilde{p}$, $\tilde{q}$ as above and with p, q there being replaced by the present p_r, $e^{(2)} + q'_r = q_r$. Therefore, (3.40) holds, and (3.39) is proved.

Case 2 vs. Case 1, Case 3 vs. Case 1, Case 2 vs. Case 3, and Case 3 vs. Case 2 are proved using the same ideas. □

As an illustration of the Theorems 3.3 and 3.4 we reconsider Example 3.1. Here we get the equivalence

$$
\begin{aligned}
&x \in \Sigma_{\mathrm{sym}} \\
&\Leftrightarrow \begin{cases}
\mathrm{rad}([A])|x| + \mathrm{rad}([b]) \geq |\check{b} - \check{A}x| \\
\mathrm{rad}([a]_{11})x_1^2 + \mathrm{rad}([a]_{22})x_2^2 + \mathrm{rad}([b]_1)|x_1| + \mathrm{rad}([b]_2)|x_2| \\
\qquad\qquad \geq |-x_1^T(\check{b} - \check{A}x)_1 + x_2^T(\check{b} - \check{A}x)_2|
\end{cases}
\end{aligned} \tag{3.41}
$$

since $p = \begin{pmatrix} 0 \\ 1 \end{pmatrix},\ q = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ are the only admissible vectors for these theorems. As just noticed the first inequality in (3.41) is the Oettli–Prager criterion which –

after some calculations – yields to the linear inequalities (3.3) in Example 3.1. The second inequality in (3.41) reads

$$\frac{3}{2}x_2^2 + |x_1| + |x_2| \geq |x_1^2 - x_1 + \frac{5}{2}x_2^2 + x_2|. \tag{3.42}$$

Here,

$$\begin{aligned} T = x_1^2 - x_1 + \frac{5}{2}x_2^2 + x_2 = 0 \quad &\Leftrightarrow \quad (x_1 - \frac{1}{2})^2 + \frac{5}{2}(x_2 + \frac{1}{5})^2 = \frac{7}{20} \\ &\Leftrightarrow \quad \left(\frac{x_1 - \frac{1}{2}}{\sqrt{\frac{7}{20}}}\right)^2 + \left(\frac{x_2 + \frac{1}{5}}{\sqrt{\frac{7}{50}}}\right)^2 = 1 \end{aligned}$$

describes an ellipse with midpoint $(\frac{1}{2}, -\frac{1}{5})$. It contains the points $(0,0)$, $(1,0)$, $(0,-2/5)$, and therefore points in the interior of the three quadrants O_1, O_3, O_4.

For $x \in O_1$ and $T \geq 0$ (i.e., x in O_1 but on or outside the ellipse) the inequality (3.42) is equivalent to $0 \geq x_1^2 - 2x_1 + x_2^2$, or

$$(x_1 - 1)^2 + x_2^2 \leq 1 \tag{3.43}$$

which is the first inequality in (3.4). For $x \in O_1$ and $T < 0$ (i.e., x in O_1 and inside the ellipse) we get

$$x_1^2 + 4x_2^2 + 2x_2 \geq 0$$

which is always true for $x \in O_1$. Moreover, $T < 0$ means

$$x_1^2 - 2x_1 + x_1 + x_2^2 + \frac{3}{2}x_2^2 + x_2 < 0,$$

i.e.,

$$\left\{(x_1 - 1)^2 + x_2^2\right\} + \left\{\frac{3}{2}x_2^2 + x_2 + x_1\right\} < 1.$$

Since for $x \in O_1$ the expression within the second pair of braces is nonnegative the expression within the first pair of braces must be less than one, i.e., (3.43) holds again. Thus the second inequality in (3.41) and the inequality in (3.43) are equivalent for $x \in O_1$.

For $x \in O_4$ and $T \geq 0$ one similarly obtains $0 \geq x_1^2 - 2x_1 + x_2^2 + 2x_2$ and finally

$$(x_1 - 1)^2 + (x_2 + 1)^2 \leq 2 \tag{3.44}$$

while for $x \in O_4$ and $T < 0$ the inequality implies

$$x_1^2 + 4x_2^2 \geq 0$$

which is always true. Moreover, $T < 0$ can be rewritten as

$$\left\{(x_1 - 1)^2 + (x_2 + 1)^2\right\} + \left\{x_1 - x_2 + \frac{3}{2}x_2^2\right\} < 2.$$

Since $x_2 \leq 0$ for $x \in O_4$ this again implies (3.44) which coincides with the second inequality in (3.4).

We close this section with some general remarks on Σ_{sym}. For regular matrices $[A]$ this set is connected and compact, but not necessarily convex – even if one intersects Σ_{sym} with an orthant, as Fig. 3.3 shows. For singular matrices $[A]$ this result may be false as Jansson's example $[A] = [-1,1] \in \mathbb{IR}^{1\times 1}$, $[b] = 1 \in \mathbb{IR}$ shows. Here $\Sigma = \Sigma_{\text{sym}} = (-\infty, 1] \cup [1, \infty)$ which is neither compact nor connected. Missing compactness and missing convexity can also be seen from the following example whose modification shows (cf. Fig. 3.3) that the interval hull $\square\Sigma$ can overestimate the corresponding hull $\square\Sigma_{\text{sym}}$ of Σ_{sym} arbitrarily large.

Example 3.2 [3]

Let

$$[A] = \begin{pmatrix} 1 & [-1,1] \\ [-1,1] & -1 \end{pmatrix}, \quad [b] = \begin{pmatrix} 2 \\ 2 \end{pmatrix}.$$

Then $[A]$ contains the two singular matrices

$$A_1 = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix} \qquad \text{and} \qquad A_2 = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}.$$

The linear system $A_1 x = b \equiv [b]$ has no solution while the solutions of $A_2 x = b$ are given by $x_1 - x_2 = 2$. No singular *symmetric* matrix is contained in $[A]$, since $\det \begin{pmatrix} 1 & s \\ s & -1 \end{pmatrix} = -1 - s^2 \leq -1$. The solution sets Σ and Σ_{sym} can be represented by

$$\Sigma = \left\{ \frac{2}{1+st} \begin{pmatrix} 1+s \\ -1+t \end{pmatrix} \;\middle|\; -1 \leq s,\, t \leq 1,\; st \neq -1 \right\} \cup \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \;\middle|\; x_1 - x_2 = 2 \right\},$$

and

$$\Sigma_{\text{sym}} = \left\{ \frac{2}{1+s^2} \begin{pmatrix} 1+s \\ -1+s \end{pmatrix} \;\middle|\; -1 \leq s \leq 1 \right\} \subseteq O_4,$$

where O_4 denotes again the fourth quadrant of $\mathbb{R}^2$. The inequalities characterizing $\Sigma \cap O_4$ read

$$2 - x_1 + x_2 \leq 0, \qquad -2 + x_1 + x_2 \leq 0, \qquad -2 - x_1 - x_2 \leq 0,$$

supplemented by the equation

$$(x_1 - 1)^2 + (x_2 + 1)^2 = 2 \tag{3.45}$$

when describing $\Sigma_{\text{sym}} \cap O_4$. Thus, Σ_{sym} is the half–circle which results as the intersection of the circle (3.45) with O_4, while Σ is the union of the half–strip in Fig. 3.2 and the straight line $x_1 - x_2 = 2$.

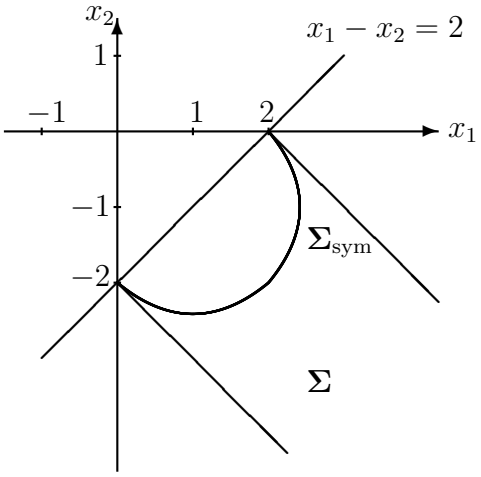


Fig. 3.2 The solution sets Σ and Σ_{sym} of Example 3.2

Replacing the entries $[-1, 1]$ in $[A]$ by $[-1 + \varepsilon, 1 - \varepsilon]$ with a small number $\varepsilon > 0$ results in a regular interval matrix. In the corresponding solution set the two parallel bordering half lines in Fig. 3.2 are now slightly shifted and inclined towards each other so that they intersect at $(2/\varepsilon, -2/\varepsilon)$. Hence Σ ends at their intersection. The straight line $x_1 - x_2 = 2$ is replaced by two finite ones in O_4 which together with the previous ones form a kite in the interior of O_4; cf. Fig. 3.3, where $\varepsilon = 1/2$.

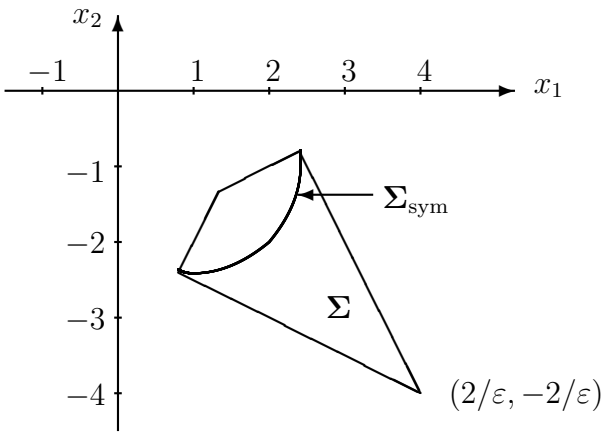


Fig. 3.3 The solution sets Σ and Σ_{sym} of Example 3.2 with $[-1, 1]$ being replaced by $[-1 + \varepsilon, 1 - \varepsilon]$, $\varepsilon = 1/2$.

If $\varepsilon > 0$ tends to zero the trailing end of the kite moves arbitrarily far away from the origin while Σ_{sym} always remains a part of the half circle in Fig. 3.2 . Therefore, depending on the value of ε the interval hull $\square\Sigma$ can overestimate $\square\Sigma_{\text{sym}}$ arbitrarily large (in the sense of Hausdorff distance, e.g.).

□

The same phenomenon as in Example 3.2 can be seen by a detailed example in [24].

4 Interval Cholesky method and modifications

In Section 3 we saw that the symmetric solution set $\Sigma_{\rm sym}$ cannot easily be described. Therefore, it is reasonable to enclose $\Sigma_{\rm sym}$ by simpler sets. Interval vectors undoubtly belong to this class. Such vectors can be constructed using the interval Cholesky method which defines a lower triangular matrix

$$[L] = \mathrm{ICh}([A]) \tag{4.1}$$

via

$$\left.\begin{array}{rcl} [l]_{jj} & = & \left([a]_{jj} - \sum_{k=1}^{j-1} [l]_{jk}^2 \right)^{1/2} , \\ [l]_{ij} & = & \left([a]_{ij} - \sum_{k=1}^{j-1} [l]_{ik}[l]_{jk} \right) / \, [l]_{jj} \, , \quad i = j+1, \ldots, n \, , \end{array}\right\} \quad j = 1, \ldots, n.$$

By virtue of a an analogue of a forward/backward substitution we get

$$[y]_i = \left([b]_i - \sum_{j=1}^{i-1} [l]_{ij}[y]_j \right) / \, [l]_{ii} \, , \qquad i = 1, \ldots, n \; ; \tag{4.2}$$

and

$$[x]_i^C = \left([y]_i - \sum_{j=i+1}^{n} [l]_{ji}[x]_j^C \right) / \, [l]_{ii} \, , \quad i = n, n-1, \ldots, 1 \; . \tag{4.3}$$

Here, sums with an upper index smaller than the lower one are defined to be zero; the squares in the first formula are evaluated by applying the interval square function (2.4). For later reasons we use equivalently the notations

$$[y] = \mathrm{IFS}([L], [b]) \qquad (= \text{ interval forward substitution}) \tag{4.4}$$

for the vector $[y]$ resulting from (4.2), and similarly

$$[x]^C = \mathrm{IBS}([L]^T, [y]) \qquad (= \text{ interval backward substitution}) \tag{4.5}$$

for the vector $[x]^C$ from (4.3).

The enclosure property of interval arithmetic implies

$$\Sigma_{\rm sym} \subseteq [x]^C \quad \text{and} \quad [A] \subseteq [L][L]^T. \tag{4.6}$$

Our next example shows that $[x]^C$ can be a tighter enclosure for $\Sigma_{\rm sym}$ than the interval hull $\square\Sigma$ for Σ.

Example 4.1

Let $[A] = \begin{pmatrix} 4 & [-1,1] \\ [-1,1] & 4 \end{pmatrix}$, $[b] = \begin{pmatrix} 6 \\ 6 \end{pmatrix}$. Setting $A = \begin{pmatrix} 4 & \alpha \\ \beta & 4 \end{pmatrix}$ for $A \in [A]$, we get $A^{-1}b = \dfrac{6}{16-\alpha\beta}\begin{pmatrix} 4-\alpha \\ 4-\beta \end{pmatrix}$ with $\alpha,\ \beta \in [-1,1]$. If $A = A^T \in [A]$ then $\beta = \alpha$ yields $A^{-1}b = \dfrac{6}{4+\alpha}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$. Thus

$$\begin{aligned} \Box\Sigma_{\mathrm{sym}} &= ([\tfrac{18}{15},2]\,,\,[\tfrac{18}{15},2])^T\,, \quad \Box\Sigma &= ([\tfrac{18}{17},2]\,,\,[\tfrac{18}{17},2])^T\,, \\ [x]^C &= ([1,2]\,,\,[\tfrac{18}{16},2])^T\,, \quad [x]^G &= ([1,2]\,,\,[\tfrac{18}{17},2])^T\,, \end{aligned}$$

where $[x]^G$ denotes the vector resulting from the interval Gaussian algorithm [2], [39]. The sets

$$\Sigma_{\mathrm{sym}} = \left\{ \frac{6}{4+\alpha}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \mid -1 \le \alpha \le 1 \right\} = \left\{ \gamma \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} \mid \frac{6}{5} \le \gamma \le 2 \right\}$$

and (see [23])

$$\Sigma = \text{convex hull}\left(\{ (\frac{6}{5},\frac{6}{5})^T,\ (2,2)^T,\ (\frac{18}{17},\frac{30}{17})^T,\ (\frac{30}{17},\frac{18}{17})^T \} \right)$$

can be seen in Fig. 4.1 .

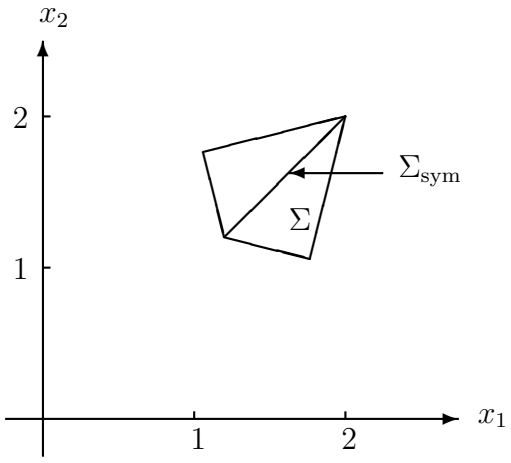


Fig. 4.1 The sets Σ and Σ_{sym}

Example 4.1 illustrates the following properties:

$\Box\Sigma_{\mathrm{sym}} \neq \Box\Sigma$ (cf. also [39]), $\Box\Sigma_{\mathrm{sym}} \neq [x]^C$, $\Box\Sigma \neq [x]^G$, $\Sigma \not\subseteq [x]^C$ (but $\Sigma_{\mathrm{sym}} \subseteq [x]^C$; cf. (4.6)), $[x]^C \subseteq [x]^G$ with $[x]^C \neq [x]^G$.

□

The enclosure $[y]$ and $[x]^C$ in (4.2), (4.3) can alternatively be expressed by the products

$$\begin{array}{rcl} [y] & = & [D^n]([L^{n-1}]([D^{n-1}](\ldots([L^2]([D]^2([L^1]([D^1][b])))))\ldots))), \\ [x]^C & = & [D^1]([L^1]^T([D^2](\ldots([L^{n-2}]^T([D^{n-1}]([L^{n-1}]^T([D^n][y])))))\ldots))). \end{array} \tag{4.7}$$

Here, $[D^s],\ s = 1, \ldots, n$, are diagonal matrices and $[L^s],\ s = 1, \ldots, n-1$, are lower triangular matrices which are defined by

$$\begin{array}{rcl} [d^s_{ij}] & = & \left\{ \begin{array}{cl} 1 & \text{if } i = j \neq s \\ 1/[l_{ss}] & \text{if } i = j = s \\ 0 & \text{otherwise} \end{array} \right. , \\ [l^s_{ij}] & = & \left\{ \begin{array}{cl} 1 & \text{if } i = j \\ -[l_{is}] & \text{if } i > j = s \\ 0 & \text{otherwise} \end{array} \right. . \end{array} \tag{4.8}$$

There is a third equivalent way to define $[x]^C \in \mathbb{IR}^n$ which for $n > 1$ uses the partition $[A] = \begin{pmatrix} [a]_{11} & [c]^T \\ [c] & [A]' \end{pmatrix}$ and the Schur complement $\Sigma^C_{[A]} = [A]' - [c][c]^T/[a]_{11}$ if $n > 1,\ 0 \notin [a]_{11}$, where $[c]_i[c]_i$ is evaluated as $[c]_i^2$.

Definition 4.1

The pair $([L], [L]^T)$ is called the Cholesky decomposition of $[A] = [A]^T \in \mathbb{IR}^{n\times n}$ if $0 < \underline{a}_{11}$ and if either $n = 1$, $[L] = (\sqrt{[a]_{11}})$ or if $n > 1$ and

$$[L] = \begin{pmatrix} \sqrt{[a]_{11}} & 0 \\ [c]/\sqrt{[a]_{11}} & [L]' \end{pmatrix}, \tag{4.9}$$

where $([L]', ([L]')^T)$ is the Cholesky decomposition of $\Sigma^C_{[A]}$. If $0 \in [a]_{11}$ the Cholesky decomposition does not exist.

The matrix $[L]$ in Definition 4.1 is the same as that defined by the interval Cholesky method above; cf. [8]. Therefore, it can be used to define the vector $[x]^C$ as in (4.3). In particular, the existence of the Cholesky decomposition is equivalent to the existence of $[x]^C$ in (4.3). Apparently it depends only on $[A]$ but not on the right–hand side $[b]$ (which influences the value of $[x]^C$, of course). For shortness we will say that $[x]^C$ exists if the Cholesky decomposition of $[A] = [A]^T \in \mathbb{IR}^{n\times n}$ exists skipping the right–hand side $[b]$.

It is a basic fact of matrix analysis that the existence of the Cholesky decomposition of a symmetric point matrix $A \in \mathbb{R}^{n\times n}$ is equivalent to A being positive definite, to A having only positive eigenvalues, and to A having only positive leading principal minors; cf. for instance [25]. This does not hold for interval matrices as the following example shows which is essentially due to Reichmann [45] and is represented here in a form slightly modified by Neumaier [37].

Example 4.2

Let $[A] = \begin{pmatrix} 1 & [a] & [a] \\ [a] & 1 & [a] \\ [a] & [a] & 1 \end{pmatrix}$ with $[a] = [0, \frac{2}{3}]$. Then for the leading principal submatrices $\tilde{A}_k$ of $\tilde{A} = \begin{pmatrix} 1 & a & b \\ a & 1 & c \\ b & c & 1 \end{pmatrix}$ with $a, b, c \in [0, \frac{2}{3}]$ we get $\det \tilde{A}_1 = 1 > 0$, $\det \tilde{A}_2 = 1 - a^2 > 0$, $\det \tilde{A}_3 = 1 - a^2 - b^2 - c^2 + 2abc > 0$; see [8] for a proof of the last inequality. Hence $\tilde{A}$ is symmetric and positive definite, but the interval Cholesky method breaks down since $0 \in [a_{33}] - [l_{31}]^2 - [l_{32}]^2 = [-\frac{11}{45}, 1]$, i.e., $[l]_{33}$ does not exist.

□

By virtue of Example 4.2 it is necessary to find classes of interval matrices for which $[x]^C$ exists. Perhaps the most prominent one is the class of H–matrices with positive diagonal entries.

Theorem 4.1 [8]

Let $[A] = [A]^T \in \mathbb{IR}^{n \times n}$ be an H–matrix with $0 < \underline{a}_{ii}$, $i = 1, \ldots, n$. Then the following statements hold.

a) The vector $[x]^C$ exists, and $[L]$ is again an H–matrix.

b) Each symmetric matrix $\tilde{A} \in [A]$ is positive definite.

Theorem 4.1 implies several corollaries.

Corollary 4.1 [10]

Let $[A] = [A]^T \in \mathbb{IR}^{n \times n}$ be an H–matrix. Then the following statements are equivalent.

(i) The vector $[x]^C$ exists.

(ii) The sign condition $\underline{a}_{ii} > 0$, $i = 1, \ldots, n$, holds.

(iii) The matrix $[A]$ contains at least one symmetric and positive definite element $\tilde{A} \in [A]$.

Corollary 4.2 [8]

Let $[A] = [A]^T \in \mathbb{IR}^{n \times n}$ with $0 < \underline{a}_{ii}$, $i = 1, \ldots, n$. Then in each of the following cases, $[A]$ is an H–matrix and $[x]^C$ exists.

(i) $\langle [A] \rangle$ *is generalized strictly diagonally dominant.*

(ii) $\langle [A] \rangle$ *is generalized irreducibly diagonally dominant.*

(iii) $\langle [A] \rangle$ *is strictly diagonally dominant.*

(iv) $\langle [A] \rangle$ *is irreducibly diagonally dominant.*

(v) $\langle [A] \rangle$ *is regular and diagonally dominant.*

(vi) $\langle [A] \rangle$ *is positive definite.*

Corollary 4.3 [8]

Let $[A] = [A]^T \in \mathbb{IR}^{n \times n}$ *be a tridiagonal matrix and let* $\tilde{A} \in [A]$ *be any symmetric matrix which satisfies* $\langle \tilde{A} \rangle = \langle [A] \rangle$ *and which is positive definite. Then* $[A]$ *is an* H*–matrix; in particular, all symmetric matrices* $A \in [A]$ *are positive definite, and* $[x]^C$ *exists.*

Corollary 4.4 [8]

Let $[A] = [A]^T \in \mathbb{IR}^{n \times n}$ *be a tridiagonal matrix and let* $\tilde{A} \in [A]$ *be any symmetric matrix which satisfies* $\langle \tilde{A} \rangle = \langle [A] \rangle$*. If* $\tilde{A}$ *can be chosen such that it fulfills one of the following three properties*

(i) $\tilde{A}$ *is totally positive,*

(ii) $\tilde{A}$ *is regular and totally nonnegative,*

(iii) $\tilde{A}$ *is oscillatory,*

then $[A]$ *is an* H*–matrix; in particular, all symmetric matrices* $A \in [A]$ *are positive definite, and* $[x]^C$ *exists.*

In some results on the interval Cholesky method the feasibility of the interval Gaussian algorithm comes into the play which ends up with an interval vector $[x]^G$. It is well known that $[x]^G$ encloses the general solution set Σ and therefore, Σ_{sym}, too. Some sort of connection between the existence of $[x]^C$ and $[x]^G$ is not very surprising since for real symmetric positive definite matrices A both vectors exist simultaneously. For interval matrices, however, only the following weaker result can be shown.

Theorem 4.2 [10]

Let $[A] = [A]^T \in \mathbb{IR}^{n \times n}$ *contain a symmetric and positive definite matrix* $\tilde{A}$*. If* $[x]^G$ *exists then* $[x]^C$ *exists, too.*

A first result towards the converse of Theorem 4.2 is the following one.

Theorem 4.3 [10]

Let $[A] = [A]^T \in \mathbb{IR}^{n\times n}$ *contain a symmetric and positive definite matrix* $\tilde{A}$ *and let* $n \leq 3$. *If* $[x]^C$ *exists then* $[x]^G$ *exists, too.*

Unfortunately the restriction $n \leq 3$ cannot be dropped in Theorem 4.3 as can be seen by our next example.

Example 4.3 [10]

Let

$$[A] = \begin{pmatrix} 1 & [-1,1] & 0 & 0 \\ [-1,1] & 2 & 1 & 2 \\ 0 & 1 & 2 & 2 \\ 0 & 2 & 2 & 16/3 \end{pmatrix}$$

Then

$$[L] = \begin{pmatrix} 1 & 0 & 0 & 0 \\ [-1\,,1] & [1\,,\sqrt{2}\,] & 0 & 0 \\ 0 & [1/\sqrt{2}\,,1] & [1\,,\sqrt{3/2}\,] & 0 \\ 0 & [2/\sqrt{2}\,,2] & [0\,,1] & [\,\sqrt{1/3}\,,\sqrt{10/3}\,] \end{pmatrix},$$

i.e., $[x]^C$ exists while the forward substitution of the interval Gaussian algorithm ends up with the entry $[-4/9\,,\,4]$ at position $(4,4)$ which contains zero. Thus $[x]^G$ cannot exist.

Example 4.3 was unexpected since in [10] it was shown that the existence of $[x]^C$ implies the existence of $[x]^G$ for all point matrices $A \in [A]$ (and not only for the symmetric ones).

Despite of the Example 4.3 there are some classes of matrices for which both $[x]^C$ and $[x]^G$ exist. Our first result in this respect is a slight generalization of Theorem 4.1.

Theorem 4.4 [2], [8]

Let $[A] = [A]^T \in \mathbb{IR}^{n\times n}$ *be an H–matrix with* $0 < \underline{a}_{ii}$, $i = 1,\ldots,n$. *Then both* $[x]^C$ *and* $[x]^G$ *exist.*

Based on this theorem the Corollaries 4.1 – 4.4 can be generalized in the same way. Similarly, we get the following result. (Cf. also the Theorems 3.6.7 and 4.5.8 in [39].)

Theorem 4.5 [8], [12]

Let $[A] = [A]^T \in \mathbb{IR}^{n\times n}$ *be an M–matrix and let* $[b] \in \mathbb{IR}^n$. *Then both* $[x]^C$ *and* $[x]^G$ *exist. If, in addition,* $\underline{b} \geq 0$ *or* $\underline{b} \leq 0 \leq \overline{b}$ *or* $\overline{b} \leq 0$ *then*

$$[x]^C = \square\Sigma_{\mathrm{sym}} = [x]^G = \square\Sigma = \begin{cases} [\overline{A}^{-1}\underline{b}, \underline{A}^{-1}\overline{b}], & \text{if } \underline{b} \geq 0, \\ [\underline{A}^{-1}\underline{b}, \underline{A}^{-1}\overline{b}], & \text{if } \underline{b} \leq 0 \leq \overline{b}, \\ [\underline{A}^{-1}\underline{b}, \overline{A}^{-1}\overline{b}], & \text{if } \overline{b} \leq 0. \end{cases}$$

Theorem 4.5 does not generalize to inverse nonnegative matrices as the following example shows which was presented by Neumaier in [39], p. 160, in connection with the (non–) feasibility of the interval Gaussian algorithm.

Example 4.4

Let

$$[A] = \begin{pmatrix} [4,5] & [-3,-2] & 1 \\ [-3,-2] & 4 & [-3,-2] \\ 1 & [-3,-2] & [4,5] \end{pmatrix} = [A]^T.$$

Then

$$\underline{A}^{-1} = \frac{1}{6}\begin{pmatrix} 7 & 9 & 5 \\ 9 & 15 & 9 \\ 5 & 9 & 7 \end{pmatrix} \geq \overline{A}^{-1} = \frac{1}{8}\begin{pmatrix} 2 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 2 \end{pmatrix} \geq O.$$

Hence by Kuttler's theorem mentioned in Section 2 the interval matrix $[A]$ is inverse nonnegative; in particular, it is regular. Therefore, by reasons of continuity, $\det \tilde{A}$ does not change its sign when $\tilde{A}$ varies in $[A]$. It is positive by virtue of $\det \underline{A} = 6 > 0$. One can easily see that the leading principal 1×1 and 2×2 submatrices of $\tilde{A} \in [A]$ have positive determinants, too. Therefore, each symmetric matrix $\tilde{A} \in [A]$ is positive definite. It is mentioned in [39] that $[x]^G$ does not exist, whence $[x]^C$ cannot exist by virtue of Theorem 4.3 .

□

Despite of this negative result parts of Theorem 4.5 remain true for inverse nonnegative interval matrices. Based on the continuity of eigenvalues (which all are positive for a symmetric positive definite point matrix $\tilde{A} \in [A]$, remain real and cannot assume the value zero as long as one perturbs $\tilde{A}$ symmetrically within a *regular* interval matrix $[A]$) or based on Theorem 4.15 at the end of this section one can easily prove the following theorem.

Theorem 4.6 [8], [12]

Let $[A] = [A]^T \in \mathbb{IR}^{n\times n}$ be inverse nonnegative and contain at least one symmetric and positive definite matrix $\tilde{A} \in [A]$. Then all symmetric matrices in $[A]$ are positive definite. If $[b] \in \mathbb{IR}^n$ satisfies $\underline{b} \geq 0$ or $\underline{b} \leq 0 \leq \overline{b}$ or $\overline{b} \leq 0$. Then

$$\square\Sigma_{\mathrm{sym}} = \square\Sigma = \begin{cases} [\overline{A}^{-1}\underline{b}, \underline{A}^{-1}\overline{b}], & \text{if } \underline{b} \geq 0, \\ [\underline{A}^{-1}\underline{b}, \underline{A}^{-1}\overline{b}], & \text{if } \underline{b} \leq 0 \leq \overline{b}, \\ [\underline{A}^{-1}\underline{b}, \overline{A}^{-1}\overline{b}], & \text{if } \overline{b} \leq 0. \end{cases}$$

Theorem 4.7 [10]

Let $[A] = [A]^T \in \mathbb{IR}^{n\times n}$ be tridiagonal. Then the following statements are equivalent.

(i) The vector $[x]^G$ exists and $[A]$ contains at least one symmetric and positive definite matrix.

(ii) The vector $[x]^C$ *exists.*

(iii) Each symmetric matrix $\tilde{A} \in [A]$ *is positive definite.*

Theorem 4.7 can be illustrated by the tridiagonal matrix $[A] = \mathrm{trid}([-1,1], 2, [-1,1]) \in \mathbb{IR}^{n\times n}$ with diagonal entries 2.

Theorem 4.8 [10]

Let $[A] = I + [-R, R]$ *with* $O \leq R = R^T \in \mathbb{R}^{n\times n}$ *and* $0 < \underline{a}_{ii},\ i = 1, \ldots, n$. *Then the following statements are equivalent.*

(i) The vector $[x]^G$ *exists.*

(ii) The vector $[x]^C$ *exists.*

(iii) The spectral radius of R *is less than one.*

(iv) The matrix $[A]$ *is an* H*–matrix.*

Note that the assumption $[A] = I + [-R, R]$ in Theorem 4.8 is fulfilled if one preconditions $[A]$ and $[b]$ by the midpoint inverse $\check{A}^{-1}$ provided that the diagonal entries of $\check{A}^{-1}[A]$ do not contain zero.

Our next two results need some preparation.

The undirected graph of a real matrix $A \in \mathbb{R}^{n\times n}$ consists of the nodes $1, \ldots, n$ and the edges $\{i, j\}, i \neq j$, whenever $|a_{ij}| + |a_{ji}| \neq 0$; cf. for instance [16] and [22]. We call j a neighbor of the node i $(\neq j)$ if $\{i, j\}$ is an edge. The number of neighbors of i are the degree of i in the underlying graph. Denote by $[A]^{(k)} = ([a]_{ij}^{(k)}) \in \mathbb{IR}^{n\times n}$ the matrix just before the k–th elimination step of the interval Gaussian algorithm is executed. (Thus $[A]^{(1)} = [A]$ while $[A]^{(n)}$ is the final matrix of the forward substitution having upper triangular form.) Let G_k denote the k–th elimination graph of $[A]$, i.e., the undirected graph of $|[A]^{(k)}|$ in which the nodes $1, \ldots, k-1$ and the corresponding edges have been removed and for which we assume that $[a]_{ij}^{(k-1)} \neq 0$ implies $[a]_{ij}^{(k)} \neq 0,\ i, j \geq k$ (no accidental zeros!); cf. [16], [22]. If in G_k the node k has the smallest degree and if this holds for all $k = 1, \ldots, n$ then we say that $[A]$ is ordered by minimum degree. If the graph of such a matrix has tree structure (i.e., it is connected and there are no cycles of length ≥ 3; cf. [16] or [22]) the following result holds.

Theorem 4.9 [10]

Let $[A] = [A]^T \in \mathbb{IR}^{n\times n}$ *contain a symmetric and positive definite matrix* $\tilde{A}$. *If the undirected graph of* $\langle [A] \rangle$ *is a tree and if it is ordered by minimum degree then the following statements are equivalent.*

(i) The vector $[x]^G$ *exists.*

(ii) The vector $[x]^C$ *exists.*

(iii) Each symmetric matrix in $[A]$ *is positive definite.*

Theorem 4.9 can be illustrated by symmetric tridiagonal interval matrices and symmetric arrowhead interval matrices [53] (provided that they contain a symmetric and positive definite matrix $\tilde{A}$). A simple example is

$$[A] = \begin{pmatrix} 2 & & & [-1,1] \\ & \ddots & O & \vdots \\ & O & 2 & [-1,1] \\ [-1,1] & \dots & [-1,1] & n \end{pmatrix} .$$

Definition 4.2

Let $[A] \in \mathbb{IR}^{n \times n}$.

a) The matrix $S \in \mathbb{R}^{n \times n}$ *with* $s_{ij} = \operatorname{sign} \check{a}_{ij}$ *is called the sign matrix of* $[A]$.

b) With S *from a) the extended sign matrix* S' *of* $[A]$ *is defined as follows.*

$S' = S$
`for` $k = 1 : (n-1)$
    `for` $i = (k+1) : n$
        `for` $j = (k+1) : n$
            `if` $s'_{ij} == 0$ `then` $s'_{ij} = -s'_{ik} s'_{kk} s'_{kj}$.

Note that the values of s'_{ij} only depend on S. Any other matrix $[\hat{A}]$ with the same sign matrix S as $[A]$ yields the same extended sign matrix S'.

Theorem 4.10 [10]

Let $[A] = [A]^T \in \mathbb{IR}^{n \times n}$ *be irreducible and generalized diagonally dominant with* $0 < \underline{a}_{ii}$, $i = 1, \dots, n$. *Moreover, let* S' *be the extended sign matrix of* $[A]$ *defined in Definition* 4.2 . *Then the following statements are equivalent.*

(i) The vector $[x]^C$ *exists.*

(ii) The vector $[x]^G$ *exists.*

(iii) The matrix $[A]$ *is generalized irreducibly diagonally dominant or the sign condition*

$$s'_{ij}\, s'_{ik}\, s'_{kj} = 1 \tag{4.10}$$

holds for some triple (i, j, k) *with* $k < j < i$.

Corollary 4.5 [10]

Let $[A] = [A]^T \in \mathbb{IR}^{n\times n}, n \geq 3$, *be irreducible and generalized diagonally dominant with* $0 < \underline{a}_{ii}$, $i = 1, \ldots, n$. *Moreover, let* S *be the sign matrix of* $[A]$ *from Definition* 4.2 . *If*

$$s_{ij}\, s_{ik}\, s_{kj} = 1$$

holds for some triple (i, j, k) *with* $k < j < i$ *then* $[x]^C$ *exists.*

The example

$$[A]_\alpha = \begin{pmatrix} 4 & [\alpha, 2] & [\alpha, 2] \\ [\alpha, 2] & 4 & 2 \\ [\alpha, 2] & 2 & 4 \end{pmatrix}, \qquad \alpha \in [-2, 2],$$

illustrates Theorem 4.10. Here $\langle [A]_\alpha \rangle\, e = 0$. For $-2 < \alpha \leq 2$ we have $S = ee^T = S'$, and (4.10) is fulfilled with $(i, j, k) = (3, 2, 1)$, hence $[x]^C$ exists. For $\alpha = -2$ property (4.10) does not hold and $[x]^C$ does not exist.

The example

$$[A] = \begin{pmatrix} 4 & 0 & [0, 2] & [-2, 0] \\ 0 & 4 & [0, 2] & [0, 2] \\ [0, 2] & [0, 2] & [6, 9] & [-2, 2] \\ [-2, 0] & [0, 2] & [-2, 2] & [6, 9] \end{pmatrix}$$

shows that $S \neq S'$ can occur. Here (4.10) holds for $(i, j, k) = (4, 3, 2)$ and $[x]^C$ exists.

We continue with some perturbation results. To this end let $[x]^C$ exist for $[A] \in \mathbb{IR}^{n\times n}$, $[b] \in \mathbb{IR}^n$, and define the nonnegative real matrix $|[A]^C|$ by

$$|\,[A]^C\,| = \langle [L]^T \rangle^{-1} \langle [L] \rangle^{-1}. \tag{4.11}$$

It can be shown that $|[A]^C|$ is the absolute value of the interval matrix whose j–th column is the result of the interval Cholesky method applied to $[A]$ and the vector $[b] = [-e^{(j)}, e^{(j)}]$; cf. [9].

This definition is applied together with the following lemma in order to formulate and prove the subsequent crucial Theorem 4.11 .

Lemma 4.1 [9]

Let $[x]^C$ *exist for* $[A] = [A]^T \in \mathbb{IR}^{n\times n}$ *with the Cholesky decomposition* $([L], [L]^T)$. *Let* $[B] = [B]^T \supseteq [A]$ *be such that for a suitable positive vector* u *we have*

$$q([A], [B])u < \langle [L] \rangle \langle [L]^T \rangle u\ . \tag{4.12}$$

Then the Cholesky decomposition exists for $[B]$, *too.*

Theorem 4.11 [9]

Let $[A],\ [B] \in \mathbb{IR}^{n\times n}$, $[A] = [A]^T$, $[B] = [B]^T$. *Suppose that* $[x]^C$ *exists for* $[A]$. *If*

$$\rho(\,|[A]^C|\, q([A], [B])\,) < 1 \tag{4.13}$$

then $[x]^C$ *exists for* $[B]$, *too.*

Specializing Theorem 4.11 one immediately gets a simple corollary.

Corollary 4.6 [9]

Let the midpoint matrix $\check{A}$ *of* $[A] = [A]^T \in \mathbb{IR}^{n\times n}$ *be positive definite, and assume that*

$$\rho\left(\frac{1}{2}\,|\check{A}^C|\,d([A])\right) < 1\ .$$

Then $[x]^C$ *exists for* $[A]$.

As an illustration consider $[A] = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 4 & [0,2] \\ 2 & [0,2] & 4 \end{pmatrix}$ with

$$\check{A} = LL^T, \quad L = \begin{pmatrix} 2 & 0 & 0 \\ 1 & \sqrt{3} & 0 \\ 1 & 0 & \sqrt{3} \end{pmatrix}, \quad |\check{A}^C| = \frac{1}{12}\begin{pmatrix} 5 & 2 & 2 \\ 2 & 4 & 0 \\ 2 & 0 & 4 \end{pmatrix},$$

and $\rho\left(\frac{1}{2}\,|\check{A}^C|\,d([A])\right) = 1/3 < 1.$

Like the interval Gaussian algorithm the interval Cholesky method often suffers from overestimation caused by interval dependency (cf. [36]). Example 4.2 illustrates that interval pivots $[l]_{ii}^2$ can contain zero in its interior so that the square root necessary to compute $[l]_{ii}$ does not exist. In [20], [21] ways are indicated how to avoid this breakdown of the method. They are based on lower bounds for the smallest eigenvalue $\lambda_1(\tilde{A})$ of point matrices $\tilde{A} = \tilde{A}^T \in [A] = [A]^T \in \mathbb{IR}^{n\times n}$ since $\lambda_1(\tilde{A}) \leq l_{nn}^2(\tilde{A})$. Here $l_{nn}(\tilde{A})$ denotes the final entry of L in the Cholesky decomposition $\tilde{A} = LL^T$ which is assumed to exist. One possibility to tighten $[l]_{nn}^2$ is to use the inequality

$$\lambda_1(\check{A}) - \rho(\mathrm{rad}([A])) \leq \min\{\,\lambda_1(\tilde{A}) \mid \tilde{A} = \tilde{A}^T \in [A]\,\} \leq \min\{\,l_{nn}^2(\tilde{A}) \mid \tilde{A} = \tilde{A}^T \in [A]\,\}.$$

There are other lower bounds for the rightmost minimum which can be computed recursively without computing any eigenvalues. Cf. [20], [21] for details.

A modification of the interval Cholesky method was given by Frommer in [16] and denoted by *symmetrized Gaussian elimination*. It is the well–known interval Gaussian elimination [2], which starts – as usually – with $[A]^{(1)} = [A]$ and computes the intermediate matrices $[A]^{(k)} = ([a]_{ij}^{(k)}) \in \mathbb{IR}^{n\times n},\ k = 2,\ \ldots, n$, by means of three loops and

$$[a]_{ij}^{(k+1)} = [a]_{ij}^{(k)} - [a]_{ik}^{(k)}[a]_{kj}^{(k)}/[a]_{kk}^{(k)}, \quad i,j > k. \tag{4.14}$$

For symmetric interval matrices $[A] = [A]^T$ one can prove by induction that $[a]_{ij}^{(k)} = [a]_{ji}^{(k)}$ holds for $i,j \geq k$. Therefore, for the symmetrized version of the interval

Gaussian elimination one can replace the computation of the diagonal entries in the two innermost loops by

$$[a]_{ii}^{(k+1)} = [a]_{ii}^{(k)} - ([a]_{ik}^{(k)})^2/[a]_{kk}^{(k)},\ i > k,$$

using the square function, while the other entries of $[A]^{(k+1)}$ are computed according to (4.14). After the usual forward/backward substitution the final vector $[x]_{\rm sym}^G$ encloses $\Sigma_{\rm sym}$. In [16] the following result was proved for $[x]_{\rm sym}^G$ with an assumption as in Theorem 4.9.

Theorem 4.12 [16]

Let $[A] = [A]^T \in \mathbb{IR}^{n\times n}$ and assume that the undirected graph of $\langle [A] \rangle$ is a tree which is ordered by minimum degree. If the point vector x^G exists for all symmetric point matrices $\tilde{A} \in [A]$ then the interval vector $[x]_{\rm sym}^G$ exists for $[A]$.

Two other variants of the interval Cholesky method were given by Schäfer in [54] introducing blocks. He starts with block partitions

$$[A] = \begin{pmatrix} [A]_{11} & \dots & [A]_{1p} \\ \vdots & & \vdots \\ [A]_{p1} & \dots & [A]_{pp} \end{pmatrix} = \begin{pmatrix} [A]_{11} & \dots & [A]_{1p} \\ \vdots & & \\ [A]_{p1} & & [A]' \end{pmatrix} \quad \text{and} \quad [b] = \begin{pmatrix} [b]^{(1)} \\ \vdots \\ [b]^{(p)} \end{pmatrix} \tag{4.15}$$

of $[A] = [A]^T \in \mathbb{IR}^{n\times n}$ and $[b] \in \mathbb{IR}^n$, respectively, with $[A]_{ij} \in \mathbb{IR}^{n_i \times n_j}$, $[b]^{(i)} \in \mathbb{IR}^{n_i}$, $\sum_{i=1}^p n_i = n$. With the notation in (4.1) – (4.5) the block interval Cholesky method computes a lower triangular matrix $[L] = ([L]_{ij})$ in block form as follows:

$$\left.\begin{aligned} [L]_{jj} &= \mathrm{ICh}\left([A]_{jj} - \sum_{k=1}^{j-1} [L]_{jk}[L]_{jk}^T \right), \\ [L]_{ij} &= \left(\mathrm{IFS}\Big([L]_{jj}, [A]_{ji} - \sum_{k=1}^{j-1} [L]_{jk}[L]_{ik}^T \Big) \right)^T,\ i = j+1, \dots, p, \end{aligned}\right\} \ j = 1, \dots, p,$$

where IFS (cf. (4.4)) is applied to $[L]_{jj}$ and the individual columns of the matrix in the second argument. Then a forward/backward substitution leads to

$$[y]^{(i)} = \mathrm{IFS}\left([L]_{ii}, [b]^{(i)} - \sum_{j=1}^{i-1} [L]_{ij}[y]^{(j)} \right), \qquad i = 1, \dots, p\ , \tag{4.16}$$

and

$$[x^C]^{(i)} = \mathrm{IBS}\left([L]_{ii}^T, [y]^{(i)} - \sum_{j=i+1}^{p} [L]_{ji}^T [x^C]^{(j)} \right), \quad i = p, p-1, \dots, 1\ , \tag{4.17}$$

with $\Sigma_{\rm sym} \subseteq [x^C]$. (Do not mix up the notation $[x^C]$ in (4.17) with $[x]^C$ in (4.3) !)

There is also a recursive definition of $[L]$ analogously to Definition 4.1.

Definition 4.3

The pair $([L],[L]^T)$ *is called block interval Cholesky decomposition of* $[A]=[A]^T\in\mathbb{IR}^{n\times n}$ *in* (4.15) *if* $[A]_{11}$ *has an interval Cholesky decomposition in the sense of Definition* 4.1 *and if either* $p=1,\ [L]=\mathrm{ICh}([A])=\mathrm{ICh}([A]_{11})$*, or if* $p>1$ *and*

$$[L]=\begin{pmatrix}\mathrm{ICh}([A]_{11}) & O\\ (\mathrm{IFS}([L]_{11},[A]_{12}))^T & \\ \vdots & [L]'\\ (\mathrm{IFS}([L]_{11},[A]_{1p}))^T & \end{pmatrix},$$

where $([L]',([L]')^T)$ *is the block interval Cholesky decomposition of*

$$[A]'-\begin{pmatrix}(\mathrm{IFS}([L]_{11},[A]_{12}))^T\\ \vdots\\ (\mathrm{IFS}([L]_{11},[A]_{1p}))^T\end{pmatrix}(\mathrm{IFS}([L]_{11},[A]_{12}),\ldots,\mathrm{IFS}([L]_{11},[A]_{1p}))$$

with $[A]'$*,* p *as in* (4.15)*. If* $\mathrm{ICh}([A]_{11})$ *does not have an interval Cholesky decomposition in the sense of Definition* 4.1 *then the block interval Cholesky decomposition does not exist.*

Theorem 4.13 [54]

Let $[A]=[A]^T\in\mathbb{IR}^{n\times n}$ *be an* H*–matrix with* $0<\underline{a}_{ii},\ i=1,\ldots,n$*. Then the interval block decomposition exists, and therefore, the same holds for* $[x^C]$ *from* (4.17) *for any partition* (4.15) *of* $[A]$*.*

There is a modification of the block interval Cholesky method which replaces IFS, IBS by a square root (cf. (2.5)) and by multiplications with $[L]_{ii}^{-1}$:

$$\left.\begin{aligned}[L]_{jj} &= \sqrt{\left([A]_{jj}-\sum_{k=1}^{j-1}[L]_{jk}[L]_{jk}^T\right)},\\ [L]_{ij} &= \left([A]_{ij}-\sum_{k=1}^{j-1}[L]_{ik}[L]_{jk}^T\right)[L]_{jj}^{-1},\quad i=j+1,\ldots,p,\end{aligned}\right\}\quad j=1,\ldots,p,$$

$$[y_{\mathrm{mod}}]^{(i)}=[L]_{ii}^{-1}\left([b]^{(i)}-\sum_{j=1}^{i-1}[L]_{ij}[y_{\mathrm{mod}}]^{(j)}\right),\qquad i=1,\ldots,p,$$

$$[x_{\mathrm{mod}}^C]^{(i)}=([L]_{ii}^{-1})^T\left([y]^{(i)}-\sum_{j=i+1}^{p}[L]_{ji}^T[x_{\mathrm{mod}}^C]^{(j)}\right),\quad i=p,p-1,\ldots,1,$$

with $\Sigma_{\mathrm{sym}}\subseteq[x_{\mathrm{mod}}^C]$, as one can easily see. Note that here $[L]=([L]_{ij})$ is a lower *block* triangular matrix but not necessarily triangular as a whole.

As an analogue of Definition 4.3 one obtains the following one.

Definition 4.4

The pair $([L],[L]^T)$ *is called modified block interval Cholesky decomposition of* $[A]=[A]^T\in\mathbb{IR}^{n\times n}$ *in* (4.15) *if* $\sqrt{[A]_{11}}$ *exists and is regular and if either* $p=1$, $[L]=\sqrt{[A]_{11}}$, *or if* $p>1$ *and*

$$[L]=\begin{pmatrix}\sqrt{[A]_{11}} & O\\ [A]_{21}(\sqrt{[A]_{11}})^{-1} & \\ \vdots & [L]'\\ [A]_{p1}(\sqrt{[A]_{11}})^{-1} & \end{pmatrix},$$

where $([L]',([L]')^T)$ *is the modified block interval Cholesky decomposition of*

$$[A]'-\begin{pmatrix}[A]_{21}(\sqrt{[A]_{11}})^{-1}\\ \vdots\\ [A]_{p1}(\sqrt{[A]_{11}})^{-1}\end{pmatrix}\left((\sqrt{[A]_{11}})^{-1}[A]_{12}\,,\,\ldots\,,\,(\sqrt{[A]_{11}})^{-1}[A]_{1p}\right)$$

with $[A]'$, p *as in* (4.15). *If* $\sqrt{[A]_{11}}$ *does not exist then the modified block interval Cholesky decomposition does not exist.*

Theorem 4.14 ([54]; cf. also Lemma 2.2)

Let $[A]=[A]^T\in\mathbb{IR}^{n\times n}$ *be an* M*–matrix. Then the modified interval block decomposition exists for any partition* (4.15) *of* $[A]$.

As was mentioned in [54] one cannot predict whether the interval Cholesky method or one of its block variants yields to better enclosures of $\Sigma_{\rm sym}$. This is demonstrated by the following example.

Example 4.5 [54]

For the symmetric M–matrix

$$[A]=\begin{pmatrix}[4,9] & 0 & -3 & 0\\ 0 & [4,9] & 0 & [-3,-1.5]\\ -3 & 0 & [7.25,10] & [-1,0]\\ 0 & [-3,-1.5] & [-1,0] & [7.25,9.25]\end{pmatrix}$$

and the vector $[b]=(1,[2,3],0,[-1,1])^T$ we consider the 2×2 partition $n_1=n_2=2$. Then we obtain

$$[x]^C=[x^C]=\begin{pmatrix}[\,\frac{40-\sqrt{5}}{324}\,,\,\frac{15}{32}\,]\\ [\,\frac{11}{72}\,,\,\frac{41}{32}\,]\\ [\,\frac{4-\sqrt{5}}{108}\,,\,\frac{7}{24}\,]\\ [\,-\frac{5}{36}\,,\,\frac{17}{24}\,]\end{pmatrix}\approx\begin{pmatrix}[\,0.116555\,,\,0.468750\,]\\ [\,0.152777\,,\,1.281250\,]\\ [\,0.016332\,,\,0.291666\,]\\ [\,-0.138888\,,\,0.708333\,]\end{pmatrix}$$

which is not comparable with

$$[x^C_{\mathrm{mod}}] = \begin{pmatrix} [\,\frac{65}{648} + \frac{1}{27\sqrt{6}}\,,\,\frac{15}{32}\,] \\ [\,\frac{3}{16} - \frac{1}{12\sqrt{6}}\,,\,\frac{41}{32}\,] \\ [\,\frac{1}{9\sqrt{6}} - \frac{7}{216}\,,\,\frac{7}{24}\,] \\ [\,-\frac{5}{72} - \frac{1}{6\sqrt{6}}\,,\,\frac{17}{24}\,] \end{pmatrix} \approx \begin{pmatrix} [\,0.115428\,,\,0.468750\,] \\ [\,0.153479\,,\,1.281250\,] \\ [\,0.012953\,,\,0.291666\,] \\ [\,-0.137486\,,\,0.708333\,] \end{pmatrix}.$$

□

For block variants we also refer to [55] and [56]. A survey on the interval Cholesky method was given in [28]; see also [32].

Recall that if $[x]^C$ exists then all symmetric matrices $\tilde{A} \in [A]$ must be positive definite. Our next theorems provide criteria for this property.

Theorem 4.15 [47]

Let $[A] = [A]^T \in \mathbb{IR}^{n\times n}$. Then each symmetric matrix $\tilde{A} \in [A]$ is positive definite if and only if one of the subsequent equivalent properties holds.

(i) *$[A]$ is regular and contains at least one symmetric positive definite matrix.*

(ii) *The matrices $\check{A} + D\,\mathrm{rad}([A])\,D$ are positive definite for all signature matrices D (i.e., for all real diagonal matrices D with $|D| = I$; see also [15] and [21]).*

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

While the properties in Theorem 4.15 are necessary and sufficient for the positive definiteness of each symmetric matrix $\tilde{A} \in [A]$ our next theorem lists criteria which are only sufficient for this property. (Cf. also Corollary 4.4 .)

Theorem 4.16 (See also Theorem 5.1 in [34])

Let $[A] = [A]^T \in \mathbb{IR}^{n\times n}$. Then each symmetric matrix $\tilde{A} \in [A]$ is positive definite if one of the following properties holds.

(i) *$[A]$ is totally positive.*

(ii) *$[A]$ is regular and totally nonnegative.*

(iii) *$[A]$ is oscillatory.*

Proof.

Each of the criteria in (i), (ii), (iii) guarantees that $\tilde{A} = \tilde{A}^T \in [A]$ is regular and satisfies $\det(\tilde{A}) \geq 0$. Therefore, $\det(\tilde{A}) > 0$. Now Lemma 5 in [17], p. 443, and the totally nonnegativity of $\tilde{A}$ guarantee that all principal minors are positive whence $\tilde{A}$ is positive definite.

□

5 Incomplete Cholesky decomposition

In case of sparse symmetric positive definite matrices A the Cholesky method for A may suffer from an enormous fill–in which sometimes can be avoided by an appropriate renumbering. The arrowhead matrix

$$A = \begin{pmatrix} n & 1 & 1 & \dots & 1 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 1 & 0 & \dots & 0 & 1 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

is a simple example for this fact. It is irreducibly diagonally dominant and cannot have zero as eigenvalue as Gershgorin's theorem for irreducible matrices shows. By the same theorem all eigenvalues are positive, hence A is symmetric and positive definite. It is easily seen that the matrix L from its Cholesky decomposition is dense in its lower triangle. Renumbering $i \to n+1-i,\ i = 1, \dots, n$, results again in an arrowhead matrix. This time the arrow is reflected at the counter–diagonal, and no fill–in occurs. In many cases fill–in cannot be controlled so easily as in the example above. For this situation an iteration based on a so–called incomplete Cholesky decomposition can be created. Such a decomposition is a splitting

$$A = LL^T - R,$$

where the sparsity of L (with $l_{ii} > 0,\ i = 1, \dots, n$) is controlled by a given set J of index pairs (i, j) with $i > j$. In fact, one requires $l_{ij} = 0$ for $(i, j) \in J$. From $Ax = b$ one equivalently gets

$$LL^T x = LL^T x + (b - Ax)$$

which implies the iteration

$$LL^T(x^{k+1} - x^k) = b - Ax^k. \tag{5.1}$$

No R is present here. Collecting in J all index pairs $(i, j),\ i > j$, for which $a_{ij} = 0$ guarantees that L and A have the same zero pattern at the expense of an iterative process. In [35] it is shown that (5.1) is convergent if A is a Stieltjes matrix. The iteration (5.1) cannot be transferred directly to interval analysis: The direct analogue would read

$$[x]^{k+1} = [x]^k + f_J([L], [b] - [A][x]^k),$$

where $f_J([L], [c])$ is defined as result of the backward substitution of $[L]^T$ and the vector obtained by the forward substitution with $[L]$ and $[c] \in \mathbb{IR}^n$. If the sequence $([x]^k)$ is convergent to some vector $[x]^*$ then

$$[x]^* = [x]^* + f_J([L], [b] - [A][x]^*)$$

must hold which, in general, is impossible as one can see by considering the radius of both sides. An alternative starts with the representation

$$LL^T x = b + Rx$$

and ends up with the iteration

$$[x]^{k+1} = f_J([L], [b] + [R][x]^k) \tag{5.2}$$

provided that $[A] \subseteq [L][L]^T - [R]$. The disadvantage of (5.2) is the presence of the matrix $[R]$ which usually destroys the sparsity of $[A] = [A]^T$ and $[L]$. Nevertheless it seems useful to study (5.2) in order to get a feeling what is going on when iterating more practically by

$$\begin{aligned} [x]^{k+1} - \tilde{x} &= f_J\left(\tilde{L}, ([b] - [A]\tilde{x}) + (\tilde{A} - [A])([x]^k - \tilde{x})\right) \\ &= f_J\left(\tilde{L}, [b] - [A]\tilde{x}\right) + f_J\left(\tilde{L}, (\tilde{A} - [A])([x]^k - \tilde{x})\right), \end{aligned} \tag{5.3}$$

following an idea of Tanabe [60]. Here, $\tilde{A}$ is an arbitrarily chosen symmetric matrix from $[A]$, preferably its midpoint $\check{A}$. The matrix $\tilde{L}$ results from the incomplete Cholesky decomposition of $\tilde{A}$ and $\tilde{x}$ is any vector from $\mathbb{R}^n$, preferably an approximate solution of some linear system $Ax = b$ with $A = A^T \in [A]$ and $b \in [b]$. The last equality in (5.3) follows analogously to Theorem 4.5.1 (iv) in [39] since $\tilde{L}$ is a point matrix; cf. also Theorem 4.4.4 (iv) there. The iteration (5.3) is induced by the equivalence

$$Ax = b \;\Leftrightarrow\; x - \tilde{x} = C(b - A\tilde{x}) + C(C^{-1} - A)(x - \tilde{x}) \tag{5.4}$$

with $C = (\tilde{L}\tilde{L}^T)^{-1}$. It exploits the sparsity of $[A]$ if $\tilde{L}$ is chosen appropriately.

We consider here only the iteration (5.2) for which we need the matrices $[L]$ and $[R]$. Given the index set J, the matrix $[A] = [A]^T \in \mathbb{IR}^{n\times n}$, and the vector $[b] \in \mathbb{IR}^n$ they are defined by

$[L] = [R] = O$
`for` $j = 1 : n$
  $[l]_{jj} = ([a]_{jj} - \sum_{k=1}^{j-1} [l]_{jk}^2)^{1/2}$
  `for` $i = (j+1) : n$
    $[h] = ([a]_{ij} - \sum_{k=1}^{j-1} [l]_{ik}[l]_{jk})$
    `if` $(i,j) \in J$ `then` $[r]_{ij} = [r]_{ji} = -[h]$
    `else` $[l]_{ij} = [h] \,/\, [l]_{jj}$

Now one can iterate according to (5.2). For $J = \emptyset$ one gets the interval Cholesky method, for $J = J_{\max} = \{\,(i,j) \mid n \geq i > j)\,\}$ one obtains a modification of the Jacobi–method. As in Section 4 there are three ways of representing the expression $f_J([x])$ which is defined by

$$f_J([x]) = f_J([L], [R][x] + [b]) \tag{5.5}$$

and which will be used in the rest of this section. The first one uses the formulae just described, the second one represents $f_J([x])$ as the product

$$\begin{aligned} f_J([x]) &= [D^1]([L^1]^T(\ldots([L^{n-1}]^T([D^n](\\ & \quad [D^n]([L^{n-1}]([D^{n-1}](\ldots([L^1]([D^1]([R][x]+[b])))\ldots))) \end{aligned}$$

with the matrices $[D^s]$, $[L^s]$ as in (4.8) (with $[L]$ from (5.5)). The third one starts with the following recursive definition of incomplete Cholesky decomposition.

Definition 5.1

The triple $([L],[R],J)$ *is called incomplete Cholesky decomposition of* $[A] = [A]^T = \begin{pmatrix} [a]_{11} & [c]^T \\ [c] & [A]' \end{pmatrix} \in \mathbb{IR}^{n\times n}$ *with* $[c] = ([c]_2,\ldots,[c]_n)^T \in \mathbb{IR}^{n-1}$ *if* $0 < \underline{a}_{11}$ *and if either* $n = 1$, $[L] = (\sqrt{[a]_{11}})$, $[R] = O$, *or* $n > 1$ *and*

$$[L] = \begin{pmatrix} \sqrt{[a]_{11}} & 0 \\ [\hat{c}]/\sqrt{[a]_{11}} & [L]' \end{pmatrix}, \quad [R] = \begin{pmatrix} 0 & [r]^T \\ [r] & [R]' \end{pmatrix},$$

where

$$\begin{cases} [r]_i = -[c]_i \text{ if } (i,1) \in J \text{ and } [r]_i = 0 \text{ otherwise}, \\ [\hat{c}] = ([\hat{c}]_2,\ldots,[\hat{c}]_n)^T \in \mathbb{IR}^{n-1} \text{ with } [\hat{c}]_i = 0 \text{ if } (i,1) \in J \text{ and } [\hat{c}]_i = [c]_i \text{ otherwise}, \\ J' = J\backslash\{\,(i,1) \mid (i,1) \in J\,\}, \\ ([L]',[R]',J') \text{ is the incomplete Cholesky decomposition of } [A]' - [\hat{c}][\hat{c}]^T/[a]_{11}. \\ \text{(Here the numbering of the rows and columns starts with 2.)} \end{cases}$$

If $0 \in [a]_{11}$ *then the incomplete Cholesky decomposition does not exist.*

In the third case $f_J([x])$ from (5.5) is computed by means of the incomplete Cholesky decomposition and by a forward/backward substitution as in in the first case.

Unfortunately, the incomplete Cholesky decomposition of a point matrix A does not necessarily exist if A is symmetric and positive definite. This can be seen from the following example.

Example 5.1

The matrix $A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 8 \end{pmatrix}$ is symmetric and positive definite.

For $J = \{\,(3,1)\,\}$ the incomplete Cholesky decomposition does not exist since by the recursive definition of this decomposition yields to

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 1 & & \\ 0 & & L' \end{pmatrix}, \quad R = \begin{pmatrix} 0 & 0 & -1 \\ 0 & & \\ -1 & & R' \end{pmatrix}, \quad J' = \emptyset$$

with (L', R', J') being the incomplete Cholesky decomposition of

$$A' - \hat{c}\hat{c}^T/a_{11} = \begin{pmatrix} 2 & 3 \\ 3 & 8 \end{pmatrix} - \begin{pmatrix} 1 \\ 0 \end{pmatrix} (1,0) = \begin{pmatrix} 1 & 3 \\ 3 & 8 \end{pmatrix}.$$

But the latter matrix is no longer positive definite. Hence it cannot have a Cholesky decomposition which, by virtue of $J' = \emptyset$, coincides with the incomplete Cholesky decomposition with respect to J'.

□

For symmetric M–matrices, and even for particular symmetric H–matrices the existence of the incomplete Cholesky decomposition can be shown together with a lot of additional theoretical results.

Theorem 5.1

If $[A] = [A]^T \in \mathbb{IR}^{n\times n}$ is an H–matrix with $0 < \underline{a}_{ii}$ for $i = 1, \ldots, n$ then for any index set $J \subseteq J_{\max} = \{ (i,j) \mid n \geq i > j \}$ the following properties hold:

a) The incomplete Cholesky decomposition $([L], [R], J)$ of $[A]$ exists.

b) The incomplete Cholesky decomposition $(\hat{L}, \hat{R}, J)$ of $\langle [A] \rangle$ exists and satisfies $\langle [A] \rangle = \hat{L}\hat{L}^T - \hat{R}$. This splitting is a regular one (i.e., $\hat{R} \geq O$, $(\hat{L}\hat{L}^T)^{-1}\hat{R} \geq O$; cf. [14]) with the spectral radius $\rho_J = \rho(\hat{L}^{-T}\hat{L}^{-1}\hat{R}) < 1$.

c) The function f_J from (5.5) is a P–contraction which satisfies

$$q(\, f_J([x])\,,\, f_J([y])\,) \leq \hat{L}^{-T}\hat{L}^{-1}\hat{R}\, q(\,[x]\,,\,[y]\,)$$

for all $[x]$, $[y] \in \mathbb{IR}^n$. In particular, the iteration

$$[x]^{k+1} = f_J([x]^k),\ k = 0, 1, 2, \ldots \tag{5.6}$$

is convergent to some vector $[x]^*_J \supseteq \Sigma_{\text{sym}}$ independently of the starting vector $[x]^0 \in \mathbb{IR}^n$.

Theorem 5.2

Let $[A] = [A]^T \in \mathbb{IR}^{n\times n}$ be an H–matrix with $0 < \underline{a}_{ii}$ for $i = 1, \ldots, n$, and let $J \subseteq J' \subseteq J_{\max}$ (where the meaning of J' differs from that in Definition 5.1). With the notation of Theorem 5.1 and an analogous one for J' the following properties hold:

a) The number ρ_J is an upper bound of the asymptotic convergence factor (= R_1–factor)

$$\alpha_J = \sup_{[x]^0 \in \mathbb{IR}^n} \left(\limsup_{k\to\infty} \|q([x]^k, [x]^*_J)\|^{1/k} \right), \quad \|\cdot\| \text{ any vector norm},$$

whose value does not depend on the particular norm being chosen.

b) $\rho_J \leq \rho_{J'} < 1$.

c) If $[A]$ *is an* M*–matrix then* ρ_J, $\rho_{J'}$ *coincide with the asymptotic convergence factors.*

d) If $[A]$ *is an* M*–matrix then* $[x]^*_J \supseteq [x]^*_{J'}$. *(Conjecture! Unproven up to now.)*

Example 5.2

Given the M–matrix

$$[A] = \begin{pmatrix} [2,3] & [-1,0] & [-1,0] \\ [-1,0] & [4,6] & [-2,0] \\ [-1,0] & [-2,0] & [6,8] \end{pmatrix}$$

and the vector $[b] = ([1,2],[1,2],[-1,1])^T$. Consider the iteration (5.6) for various index sets J and start it with $[x]^0 = 0$. Stop it whenever

$$|\underline{x}_i^{k+1} - \underline{x}_i^k| \leq \varepsilon \cdot |\underline{x}_i^k| \quad \text{and} \quad |\overline{x}_i^{k+1} - \overline{x}_i^k| \leq \varepsilon \cdot |\overline{x}_i^k|$$

was fulfilled for $i = 1, \ldots, n$ and $\varepsilon = 10^{-6}$ and denote by #it the number of iterates up to this point. The results are computed with INTLAB [52] and listed in the following table.

J	#it	$10 \cdot [x]_1^k$	$10 \cdot [x]_2^k$	$10 \cdot [x]_3^k$
$\emptyset$	1	$[2.234, 23.847]$	$[0.197, 16.539]$	$[-2.693, 11.154]$
$\{(2,1)\}$	19	$[2.426, 23.847]$	$[0.759, 16.539]$	$[-2.223, 11.154]$
$\{(3,2)\}$	25	$[2.591, 23.847]$	$[0.909, 16.539]$	$[-1.819, 11.154]$
$\{(2,1),(3,2)\}$	31	$[2.591, 23.847]$	$[1.060, 16.539]$	$[-1.819, 11.154]$
$\{(2,1),(3,1),(3,2)\}$	38	$[2.777, 23.847]$	$[1.111, 16.539]$	$[-1.667, 11.154]$

□

Example 5.2 illustrates the conjecture d) in Theorem 5.2 and indicates part b) of this theorem.

Example 5.3

Given the H–matrix

$$[A] = \begin{pmatrix} [2,3] & [-1,1] & [-1,1] \\ [-1,1] & [4,6] & [-2,2] \\ [-1,1] & [-2,2] & [6,8] \end{pmatrix}$$

and the vector $[b] = ([1,2],[1,2],[-1,1])^T$). Proceed as in Example 5.2 in order to obtain the following results.

J	#it	$10 \cdot [x]_1^k$	$10 \cdot [x]_2^k$	$10 \cdot [x]_3^k$
$\emptyset$	1	$[-9.764, 23.847]$	$[-7.968, 16.539]$	$[-11.154, 11.154]$
$\{(2,1)\}$	19	$[-8.847, 23.847]$	$[-9.039, 16.539]$	$[-11.154, 11.154]$
$\{(3,2)\}$	25	$[-9.764, 23.847]$	$[-7.968, 16.539]$	$[-11.154, 11.154]$
$\{(2,1),(3,2)\}$	31	$[-8.847, 23.847]$	$[-9.039, 16.539]$	$[-11.154, 11.154]$
$\{(2,1),(3,1),(3,2)\}$	38	$[-8.847, 23.847]$	$[-9.039, 16.539]$	$[-11.154, 11.154]$

□

Example 5.3 shows that Theorem 5.2 d) becomes false if the assumption 'M–matrix' is dropped. Part b) of this theorem is confirmed by the second column of the table.

6 Jansson's method

Jansson's method [26] is a quite general method to enclose the bounds $\min(\Sigma_{\rm sym})_i$, $\max(\Sigma_{\rm sym})_i$, $i = 1, \ldots, n$, where $(\Sigma_{\rm sym})_i$ denotes the projection of $\Sigma_{\rm sym}$ onto the x_i–coordinate axis. The method does only require the symmetry $[A] = [A]^T$ but not any further restriction as positive definiteness of $A \in [A]$. In order to derive the algorithm let $A = A^T \in [A] = [A]^T \in \mathbb{IR}^{n\times n}$, $b \in [b] \in \mathbb{IR}^n$, $C \in \mathbb{R}^{n\times n}$ regular, $\tilde{x} \in \mathbb{R}^n$. We start with the equivalence

$$Ax = b \;\Leftrightarrow\; x - \tilde{x} = C(b - A\tilde{x}) + (I - CA)(x - \tilde{x}),$$

apparently a modification of (5.4). Using Theorem 2.1 for the subsequent first equality we get

$$\begin{aligned}(C(b - A\tilde{x}))_i \;&\in\; \{\, (C(\hat{b} - \hat{A}\tilde{x}))_i \mid \hat{A} = \hat{A}^T \in [A],\ \hat{b} \in [b] \,\} \\ &=\; \sum_{j=1}^{n} c_{ij}([b]_j - [a]_{jj}\tilde{x}_j) - \sum_{j=1}^{n}\sum_{l=1}^{j-1}(c_{ij}\tilde{x}_l + c_{il}\tilde{x}_j)[a]_{jl} = [z]_i^{\rm sym},\end{aligned}$$

where the last equality is to be understood as definition.

Given $[x]_\Delta^0 \in \mathbb{IR}^n$ we now iterate according to

$$[x]_\Delta^{k+1} = [z]^{\rm sym} + (I - C[A])[x]_\Delta^k = \; [z]^{\rm sym} + [\Delta]^k, \quad k = 0, 1, \ldots, \tag{6.1}$$

with $[\Delta]^k = (I - C[A])[x]_\Delta^k$. Then the following theorem holds.

Theorem 6.1 [26]

If

$$([x]_\Delta^{k_0+1})_i \subsetneq ([x]_\Delta^{k_0})_i \tag{6.2}$$

holds for $i = 1, \ldots, n$ and for some iterate of (6.1) *then*

a) C, $[A]$ *are regular,*

b) $\Sigma_{\rm sym} \subseteq \tilde{x} + [x]_\Delta^{k_0}$,

c) $\tilde{x}_i + \underline{z}_i^{\rm sym} + \underline{\Delta}_i^{k_0} \le \min(\Sigma_{\rm sym})_i \le \tilde{x}_i + \underline{z}_i^{\rm sym} + \overline{\Delta}_i^{k_0}$,
$\tilde{x}_i + \overline{z}_i^{\rm sym} + \underline{\Delta}_i^{k_0} \le \max(\Sigma_{\rm sym})_i \le \tilde{x}_i + \overline{z}_i^{\rm sym} + \overline{\Delta}_i^{k_0}$.

Note that (6.2) guarantees regularity of C a posteriori. For practical computations C is usually chosen as an approximation of the midpoint inverse $\check{A}^{-1}$ in order to obtain (at least approximately) $O \in I - C[A]$. If $\tilde{x} \in \Sigma_{\rm sym}$ then $0 \in [x]_\Delta^{k_0}$ by Theorem 6.1 b) so that for matrices $[A]$ with small diameters, one expects (roughly speaking) $[x]_\Delta^{k_0}$ to have a quadratically small radius while $[z]^{\rm sym}$ only has a linearly small one. In view of Theorem 6.1 c) this means that $\min(\Sigma_{\rm sym})_i$ and $\max(\Sigma_{\rm sym})_i$ are tightly enclosed by the bounds given there.

7 Rohn's method

Rohn's method [50] represents another way to construct an interval vector which encloses $\Sigma_{\rm sym}$. Given $[A] = [A]^T \in \mathbb{IR}^{n\times n}$, $[b] \in \mathbb{IR}^n$, it starts with the representation (1.3) which leads to

$$x \in \Sigma_{\rm sym} \;\Leftrightarrow\; (\check{A} + T \circ \mathrm{rad}([A]))\, x = \check{b} + \tau \circ \mathrm{rad}([b]) \tag{7.1}$$

for some $T = T^T = (t_{ij}) \in \mathbb{R}^{n\times n}$, $\tau = (\tau_i) \in \mathbb{R}^n$, t_{ij}, $\tau_i \in [-1, 1]$. In a first step one computes $[B] \in \mathbb{IR}^{n\times n}$, $[z] \in \mathbb{IR}^n$ such that

$$[B] \supseteq \{\, A^{-1} \mid A \in [A] \,\} \;\text{ and }\; [z] \supseteq \Sigma = \Sigma([A], [b])\,,$$

where the last notation is introduced for clarity. Such enclosures can be obtained, e.g., by virtue of the interval Gaussian algorithm, applied to $[A]$, to $[b]$ and the n righthand sides $e^{(i)}$, $i = 1, \ldots, n$. Note that no symmetry is exploited for $[B]$ and $[z]$. Consider now the solution x of the linear system $(\check{A}+T\circ\mathrm{rad}([A]))x = \check{b}+\tau\circ\mathrm{rad}([b])$ as a function of the parameters $T = (t_{ij})$ and $\tau = (\tau_i)$, i.e., $x = x(T, \tau)$. Then the following properties can be shown:

$$\left.\begin{array}{rcl} \dfrac{\partial x_i}{\partial t_{kk}} & \in & -\mathrm{rad}([a]_{kk})[B]_{ik}[z]_k \\ \dfrac{\partial x_i}{\partial t_{kj}} & \in & -\mathrm{rad}([a]_{kj})\,([B]_{ik}[z]_j + [B]_{ij}[z]_k) \;\text{ for } k \neq j \\ \dfrac{\partial x_i}{\partial \tau_\ell} & \in & \mathrm{rad}([b]_\ell)[B]_{i\ell} \end{array}\right\} \tag{7.2}$$

We want to compute $\overline{x}_{i_0} = \max(\Sigma_{\rm sym})_{i_0}$ for fixed $i = i_0$:

If, for instance, $\min(\mathrm{rad}([b]_\ell)[B]_{i_0\ell}) \geq 0$ then $x_{i_0}(T, \tau)$ increases monotonically with respect to τ_ℓ, hence $\overline{x}_{i_0}$ is achieved for $\tau_\ell = 1$ if the other parameters are chosen appropriately. Thus τ_ℓ can be fixed for further procedure if one wants to compute $\overline{x}_{i_0}$.

Generally, if zero is not contained in the interior of some righthand side of (7.2) then x_{i_0} behaves monotonically with respect to the corresponding parameter and assumes its two extreme values (with respect to this parameter) for the parameter values -1 and 1, respectively. Since – with the exception of $t_{ij} = t_{ji}$ – all parameters are independent from each other their values can be fixed as soon as they are known for $\overline{x}_{i_0}$.

Consider now the equation

$$(\check{A} + T \circ \mathrm{rad}([A]))x = \check{b} + \tau \circ \mathrm{rad}([b]), \tag{7.3}$$

for some element $x \in \Sigma_{\rm sym}$ such that $x_{i_0} = \overline{x}_{i_0}$. Then we can divide the parameters into two groups: Those which can be fixed and those which cannot, i.e., those which, at the moment, are still unknown. Therefore, T and τ in (7.3) can be split in two direct sums

$$T = T_f + T_{nf} \;\text{ and }\; \tau = \tau_f + \tau_{nf},$$

where T_f and τ_f contain the fixed parameter values and $T_{nf},\ \tau_{nf}$ contain the non–fixed ones. Obviously, all entries of T_f and τ_f are ± 1 or zero while those of $T_{nf},\ \tau_{nf}$ are zero at positions, where $T_f,\ \tau_f$ are ± 1, or remain undetermined. Equation (7.3) reads now

$$\big(\check{A} + T_f \circ \mathrm{rad}([A]) + T_{nf} \circ \mathrm{rad}([A])\big)\, x = \check{b} + \tau_f \circ \mathrm{rad}([b]) + \tau_{nf} \circ \mathrm{rad}([b]).$$

Define $[A]' = \check{A} + T_f \circ \mathrm{rad}([A]) + [-1,1]S_{nf} \circ \mathrm{rad}([A])$ and $[b]' = \check{b} + \tau_f \circ \mathrm{rad}([b]) + [-1,1]s_{nf} \circ \mathrm{rad}([b])$, where $S_{nf} \in \mathbb{R}^{n\times n}$ is the matrix with ones at places of T which are not fixed and zero otherwise; similarly $s_{nf} \in \mathbb{R}^n$ is the vector with ones at places of τ which are not fixed and zero otherwise. Compute $[B]',\ [z]'$ for $[A]',\ [b]'$ analogously to $[B],\ [z]$. Then

$$[B]' \supseteq \{\, A^{-1} \mid A \in [A]' \,\} \quad \text{and} \quad [z]' \supseteq \Sigma([A]', [b]').$$

Since $[A]' \subseteq [A],\ [b]' \subseteq [b]$, it can be expected that additional parameters can be fixed by virtue of (7.2) applied to $[B]'$ and $[z]'$. If so, the steps can be repeated ending with a vector $[z]^*$ analogously to $[z]$, $[z]'$ whose i_0–th component contains $\overline{x}_{i_0}$. The whole procedure can be applied for $\underline{x}_{i_0}$ and the remaining components of $\overline{x}$, $\underline{x}$ in a similar way.

Example 7.1

We apply Rohn's method to the H–matrix $[A] = \begin{pmatrix} 4 & [-1,1] \\ [-1,1] & 4 \end{pmatrix}$ and to the vector $[b] = \begin{pmatrix} 6 \\ 6 \end{pmatrix}$ as in Example 4.1. From $A = \begin{pmatrix} 4 & \alpha \\ \beta & 4 \end{pmatrix},\ \alpha, \beta \in [-1,1],$ we get $A^{-1} = \dfrac{1}{16 - \alpha\beta} \begin{pmatrix} 4 & -\alpha \\ -\beta & 4 \end{pmatrix}$, whence for the interval inverse $[A]^{-1}$ we obtain

$$[A]^{-1} = \begin{pmatrix} [\frac{4}{17}, \frac{4}{15}] & [-\frac{1}{15}, \frac{1}{15}] \\ [-\frac{1}{15}, \frac{1}{15}] & [-\frac{4}{17}, \frac{4}{15}] \end{pmatrix}.$$

From the interval Gaussian algorithm applied to $[A]$ and $e^{(i)},\ i = 1, 2,$ we obtain the columns of the matrix

$$[B] = \begin{pmatrix} [\frac{7}{30}, \frac{4}{15}] & [-\frac{1}{15}, \frac{1}{15}] \\ [-\frac{1}{15}, \frac{1}{15}] & [-\frac{4}{17}, \frac{4}{15}] \end{pmatrix}$$

which slightly overestimates $[A]^{-1}$. Using the same method for $[A]$ and $[b]$ leads to

$$[z] = [x]^G = \Big([1,2], [\frac{18}{17}\,, 2]\Big)^T$$

as in Example 4.1. With (7.2) we get

$$\left.\begin{array}{rcl} \dfrac{\partial x_i}{\partial t_{kk}} & = & 0 \quad \text{for } i,k \in \{1,2\} \\ \dfrac{\partial x_i}{\partial \tau_\ell} & = & 0 \quad \text{for } i,\ell \in \{1,2\} \end{array}\right\} \tag{7.4}$$

$$\begin{aligned} \frac{\partial x_1}{\partial t_{21}} &\in -([B]_{12}[z]_1 + [B]_{11}[z]_2) = [-\frac{2}{3}, -\frac{29}{255}], \\ \frac{\partial x_2}{\partial t_{21}} &\in -([B]_{22}[z]_1 + [B]_{21}[z]_2) = [-\frac{2}{3}, -\frac{26}{255}]. \end{aligned}$$

Therefore, $x_i(T,\tau)$ behaves monotonically with respect to all parameters, i.e., $T_{nf} = O,\ \tau_{nf} = 0$, and the algorithm stops already after one step. In (7.3) one can choose $T = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ for $\underline{x}$ and $T = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$ for $\overline{x}$ with $\tau = (1,1)^T$ in both cases. (Note that, by virtue of (7.4), the parameters $t_{11},\ t_{22},\ \tau_1,\ \tau_2$ can also be replaced by -1.) Now (7.3) reads

$$\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix} \underline{x} = \begin{pmatrix} 6 \\ 6 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 4 & -1 \\ -1 & 4 \end{pmatrix} \overline{x} = \begin{pmatrix} 6 \\ 6 \end{pmatrix},$$

whence $\underline{x} = \left(\frac{18}{15}, \frac{18}{15}\right)^T$, $\overline{x} = (2\,,\,2)^T$. From Example 4.1 we see that $[\underline{x}, \overline{x}] = \square\Sigma_{\text{sym}} \subsetneq \square\Sigma$ holds.

□

Several important questions remain open up to now:

1. How many steps are needed until the algorithm stops because T_{nf} can no longer be improved?
2. Does the algorithm always end with the optimum $T_{nf} = O$?
3. Are there classes of matrices such that the algorithm ends with $T_{nf} = O$?
4. What can be said on the diameter of the final component $[z]^*_{i_0}$, i.e., on the quality of enclosure for $\overline{x}_{i_0}$?

A first, not very surprising result in this direction is the following one.

Theorem 7.1

Let $[A] = [A]^T \in \mathbb{IR}^{n\times n}$ be an M–matrix and let $[b] \in \mathbb{IR}^n$ satisfy $\underline{b} \geq 0$ or $\overline{b} \leq 0$. Then for each bound $\underline{x}_i,\ \overline{x}_i,\ i \in \{1,\ldots,n\}$, Rohn's method terminates after one step with $T = T_f$ if $[B],\ [z]$ are computed by the interval Gaussian algorithm.

Proof.

According to a result of Barth and Nuding (cf. [2] or [39]) applied to $[A]$ and $e^{(i)} \geq 0$ the interval Gaussian algorithm yields $[B] = [A]^{-1} = [\overline{A}^{-1}, \underline{A}^{-1}]$ whence $\underline{B} \geq O$. Similarly, this algorithm yields $\underline{z} \geq 0$ and $\overline{z} \leq 0$, respectively, according to the two cases of $[b]$ in the theorem.

In the first case $\underline{b} \geq 0$ we get from (7.2) the inequalities

$$\frac{\partial x_i}{\partial t_{kk}} \leq 0, \quad \frac{\partial x_i}{\partial t_{kj}} \geq 0 \ \ (k \neq j), \quad \frac{\partial x_i}{\partial \tau_\ell} \geq 0.$$

Hence $x_i(T, \tau)$ behaves monotonically with respect to each parameter t_{ij} and τ_ℓ. Therefore, the correct parameter values are known for $\underline{x}_i$ and $\overline{x}_i$ after one step of Rohn's algorithm.

The second case $\overline{b} \leq 0$ can be handled analogously.

□

Note that for $\underline{b} < 0 < \overline{b}$ things change: $x_i(T, \tau)$ increases monotonically with respect to τ_ℓ, but nothing can be said about the behavior with respect to t_{ij} since zero is contained in the interior of $[z]$ and therefore in the corresponding right hand sides of (7.2).

By virtue of Theorem 4.5 the present Theorem 7.1 is of purely academic character, of course, since $[z] = [x]^G = \square\Sigma_{\mathrm{sym}}$ has to be computed first in order to compute its bounds anew by applying Rohn's algorithm. In view of Theorem 4.6 the same remark also applies to our next theorem which can be proved analogously to Theorem 7.1 .

Theorem 7.2

Let $[A] = [A]^T \in \mathbb{IR}^{n \times n}$ be invers nonnegative and let $[b] \in \mathbb{IR}^n$ satisfy $\underline{b} \geq 0$ or $\overline{b} \leq 0$. Then for each bound $\underline{x}_i,\ \overline{x}_i,\ i \in \{1, \ldots, n\}$, Rohn's method terminates after one step with $T = T_f$ if $[B] = [\overline{A}^{-1}, \underline{A}^{-1}]$ and $[z] = [\overline{A}^{-1}\underline{b}, \underline{A}^{-1}\overline{b}]$ in the first case and $[z] = [\underline{A}^{-1}\underline{b}, \overline{A}^{-1}\overline{b}]$ in the second one.

8 Some historical remarks

To the author's knowledge the necessity to consider Σ_{sym} was first mentioned in a letter addressed by Neumaier to Rohn in December 23, 1985; cf. [38], [50]. Inspired by this letter Rohn invented his method (cf. Section 7) in June 1986 and presented it at an Oberwolfach conference in February 1990 [46], but did not publish it until 2004; cf. [48], [50]. In 1990 Jansson talked on his own method at the SCAN conference in Albena, Bulgaria; cf. [26], [51] and Section 6. Experimentally he showed on a slide that for a particular example the boundary of Σ_{sym} was slightly curvilinear, but he could not prove this phenomenon theoretically. In 1995 – cf. [9] – Alefeld and the author found a proof for 2×2 matrices. Together with Kreinovich they finally could prove in 1996 the particular curvilinear form of the boundary of Σ_{sym} in the general case using Fourier–Motzkin elimination techniques; cf. [3], [5], and Section 3. They applied this technique also in [6] to general parameter dependent linear systems, and generalized it in [7]. Based on this technique the author published in [30] a description of Σ_{sym} which was similar to Beeck's description of Σ; see again Section 3. Recently [24] Hladík used ideas from linear programming to describe Σ_{sym} in a way which can be thought as an analogue of the Oettli–Prager criterion for Σ and which were also presented in Section 3. We modified his result slightly in the present paper using a matrix–vector form, which we derived in a new, short way and for which we reduced the number of inequalities essentially. Moreover, we reproved the criterion elementarily.

The Cholesky method was first applied to interval data by Alefeld and the author in 1993; cf. [8] and Section 4. In [8] also some results on the feasibility of the method were given. Neumaier's perturbation result on the feasibility of the interval Gaussian algorithm (Theorem 4.5.15 in [39] and his subsequent remark) were transferred to the interval Cholesky method in 1995; cf. [9] and again Section 4. Further criteria of feasibility and a connection to the interval Gaussian algorithm were presented at the conference INVA 2007 in Tokio. These results were extended for the succeeding conference INVA 2008 at Okinawa, Japan. They are published now in [10]. Methods to reduce the overestimation caused by interval dependency when computing the interval pivots of the Cholesky method were first derived 2010 by Garloff in [20]; see also [21]. These methods help to avoid an early breakdown. A variant of the interval Cholesky method was given by Frommer in [16]; cf. also the interval–affine Gaussian algorithm for constraint systems in [1]. Block variants were considered by Schäfer in [54], [55], [56]. The incomplete Cholesky method was first presented at the conference INVA 2008 mentioned above, and written down without proofs in Section 5. It is intended to publish this method together with the proofs separately and in greater detail.

References

[1] Akhmerov, R. R.: Interval–Affine Gaussian Algorithm for Constrained Systems, *Reliab. Comput.* **11** (5) (2005) 323–341.

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

[3] Alefeld, G., Kreinovich, V., and Mayer, G.: The Shape of the Symmetric Solution Set, in Kearfott, R. B. and Kreinovich, V. (Eds.): *Applications of Interval Computations*, Kluwer, Boston, MA (1996) 61–79.

[4] Alefeld, G., Kreinovich, V., and Mayer, G.: Symmetric Linear Systems with Perturbed Input Data, in Alefeld, G., Herzberger, J. (Eds.): *Numerical Methods and Error Bounds*, Akademie Verlag, Berlin (1996) 16–22 .

[5] Alefeld, G., Kreinovich, V., and Mayer, G.: On the Shape of the Symmetric, Persymmetric and Skew–Symmetric Solution Set, *SIAM J. Matrix Anal. Appl.* **18** (1997) 693–705.

[6] Alefeld, G., Kreinovich, V., and Mayer, G.: The Shape of the Solution Set for Systems of Interval Linear Equations with Dependent Coefficients, *Math. Nachr.* **1998** (1998) 23–36.

[7] Alefeld, G., Kreinovich, V., and Mayer, G.: On the Solution Sets of Particular Classes of Linear Interval Systems, *J. Comp. Appl. Math.* **152** (2003) 1–15.

[8] Alefeld, G. and Mayer, G.: The Cholesky Method for Interval Data, *Linear Algebra Appl.* **194** (1993) 161–182.

[9] Alefeld, G. and Mayer, G.: On the Symmetric and Unsymmetric Solution Set of Interval Systems, *SIAM J. Matrix Anal. Appl.* **16** (1995) 1223–1240.

[10] Alefeld, G. and Mayer, G.: New Criteria for the Feasibility of the Cholesky Method with Interval Data, *SIAM J. Matrix Anal. Appl.* **30** (4) (2008) 1392–1405. (IWRMM–Preprint Nr. 07/07, University of Karlsruhe, 2007.)

[11] Araiza, R., Xiang, G., Kosheleva, O., and Škulj, D.: Under Interval and Fuzzy Uncertainty, Symmetric Markov Chains are More Difficult to Predict, in Reformat, M. and Bertold, M. R. (eds.) *Proceedings of the 26th International Conference of the North American Fuzzy Information Processing Society NAFIPS'2007*, San Diego, CA (2007) 526–531.

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

[13] Beeck, H.: Über Struktur und Abschätzungen der Lösungsmenge von linearen Gleichungssystemen mit Intervallkoeffizienten, *Computing* **10** (1972) 231–244. Submitted for publication.

[14] Berman, A. and Plemmons, R. J.: *Nonnegative Matrices in the Mathematical Sciences*, Classics in Applied Mathematics 9, SIAM, Philadelphia, 1994.

[15] Białas, S. and Garloff, J.: Intervals of P–Matrices and Related Matrices, *Linear Algebra Appl.* **58** (1984) 33–41.

[16] Frommer, A.: A Feasibility Result for Interval Gaussian Elimination Relying on Graph Structure, in: Alefeld, G., Rohn, J., Rump, S., and Yamamoto, T. (eds.), *Symbolic Algebraic Methods and Verification Methods*, SpringerMathematics, Springer, Wien, 2001, 79–86.

[17] Gantmacher, F. S.: *Matrizentheorie*, Springer, Berlin, 1986.

[18] Garloff, J.: Criteria for Sign Regularity of Sets of Matrices, *Linear Algebra Appl.* **44** (1982) 153–160.

[19] Garloff, J.: Interval Gaussian Elimination with Pivot Tightening, *SIAM Matrix Anal. Appl.* **30** (4) (2009) 1761–1772.

[20] Garloff, J.: Pivot Tightening for the Interval Cholesky Method, *Proc. Appl. Math. Mech.* **10** (2010) 549–550.

[21] Garloff, J.: Pivot Tightening for Direct Methods for Solving Symmetric Positive Definite Systems of Linear Interval Equations, *Computing*, Online First, October 29, 2011.

[22] George, A. and Liu, J. W. H.: *Computer Solution of Large Sparse Positive Definite Systems*, Prentice Hall, Englewood Cliffs, 1981.

[23] Hartfiel, D. J.: Concerning the Solution Set of $Ax = b$ where $P \leq A \leq Q$ and $p \leq b \leq q$, *Numer. Math.* **35** (1980) 355–359.

[24] Hladík, M.: Description of Symmetric and Skew–Symmetric Solution Set, *SIAM J. Matrix Anal. Appl.* **30** (2) (2008) 509–521.

[25] Horn, R. A. and Johnson, C. R.: *Matrix Analysis*, Cambridge University Press, Cambridge, 1994.

[26] Jansson, C.: Rigorous Sensitivity Analysis for Real Symmetric Matrices with Uncertain Data, in Kaucher, E., Markov, S. M., and Mayer, G. (Eds.), *Computer Arithmetic Scientific Computation and Mathematical Modelling*, IMACS Annals on Computing and Applied Mathematics, **12** (1991) 293–316.

[27] Jansson, C.: Interval Linear Systems with Symmetric Matrices, Skew–Symmetric Matrices and Dependencies in the Right Hand Side, *Computing* **46** (1991) 265–274.

[28] Krakowczyk, J.: *Intervall–Cholesky–Verfahren*, Diploma Thesis, Institut für Mathematik, Universität Rostock, Rostock, 2008.

[29] Kulpa, Z., Pownuk, A., and Skalna, I.: Analysis of Linear Mechanical Structures with Uncertainties by Means of Interval Methods, *Comput. Assist. Mech. Eng. Sci.* **5** (1998) 443–477.

[30] Mayer, G.: A New Way to Describe the Symmetric Solution Set S_{sym} of Linear Interval Systems, in Alefeld, G. and Chen, X. (Eds.), *Topics in Numerical Analysis with Special Emphasis on Nonlinear Problems*, Computing Suppl. **15** (2001) 151–163.

[31] Mayer, G.: On Regular and Singular Interval Systems, *J. Comp. Appl. Math.* **199** (2) (2007) 220–228.

[32] Mayer, G.: Direct Methods for Linear Systems with Inexact Input Data, *Japan J. Indust. Appl. Math.* **26** (2009) 279–296.

[33] Mayer, G.: An Oettli–Prager Like Theorem for the Symmetric Solution Set. To be submitted 2012.

[34] Mayer, J.: An Approach to Overcome Division by Zero in the Interval Gauss Algorithm, *Reliab. Comput.* **8** (3) (2002) 229–237.

[35] Meijerink, J. A. and van der Vorst, H. A.: An Iterative Solution Method for Linear Systems of Which the Coefficient Matrix is a Symmetric M–Matrix, *Math. Comp.* **31** (1977) 148–162.

[36] Moore, R. E., Kearfott, R. B., and Cloud, M. J.: *Introduction to Interval Analysis*, SIAM, Philadelphia, 2009.

[37] Neumaier, A.: New Techniques for the Analysis of Linear Interval Equations, *Linear Algebra Appl.* **58** (1984) 273–325.

[38] Neumaier, A.: Letter to J. Rohn, December 23, 1985. Copy and personal communication.

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

[40] Oettli W. and Prager, W.: Compatibility of Approximate Solution of Linear Equations with Given Error Bounds for Coefficients and Right–Hand Sides, *Numer. Math.* **6** (1964) 405–409.

[41] Popova, E.: On the Solution of Parametrised Linear Systems, in Krämer, W., Wolff von Gudenberg, J. (Eds.): *Scientific Computing, Validated Numerics, Interval Methods*, Kluwer, Dordrecht, 2001, 127–138.

[42] Popova, E.: Quality of the Solution Sets of Parameter–Dependent Interval Linear Systems, *Z. Angew. Math. Mech.* **82** (2002) 723–727.

[43] Popova, E.: Explicit Description of 2D Parametric Solution Sets, *BIT Numerical Analysis*, Online First, June 08, 2011.

[44] Popova, E. and Krämer, W.: Visualizing Parametric Solution Sets, *BIT Numerical Analysis* **48** (2008) 95–115.

[45] Reichmann, K.: Abbruch beim Intervall–Gauss–Algorithmus, *Computing* **22** (1979) 355–361.

[46] Rohn, J.: Linear Interval Equations with Dependent Coefficients, Symposium "Interval Methods for Numerical Computation", Oberwolfach, 1990.

[47] Rohn, J.: Positive definiteness and stability of interval matrices, *SIAM J. Matrix Anal. Appl.* **15** (1994) 175û-184.

[48] Rohn, J.: Letter to the author, June 5, 1999.

[49] Rohn, J.: Email to `reliable_computing@interval.louisiana.edu`, Mon, 25 March 2002.

[50] Rohn, J.: *A Method for Handling Dependent Data in Interval Linear Systems*, Technical Report No. 911, Institute of Computer Science, Academy of Sciences of the Czech Republic, July 7, 2004, 1–7.

[51] Rump, S. M.: Verification Methods for Dense and Sparse Systems of Equations, in Herzberger, J. (Ed.): *Topics in Validated Computations*, Elsevier, Amsterdam, 1994, 63–135.

[52] Rump, S. M.: INTLAB – INTerval LABoratory, in: Csendes, T. (Ed.), *Developements in Reliable Computing*, Kluwer, Dordrecht, 1999, 77–104.

[53] Schäfer, U.: The Feasibility of the Interval Gaussian Algorithm for Arrowhead Matrices, *Reliab. Comput.* **7** (4) (2001) 59–62.

[54] Schäfer, U.: Two Ways to Extend the Cholesky Decomposition to Block Matrices with Interval Entries, *Reliab. Comput.* **8** (1) (2002) 1–20.

[55] Schäfer, U.: Aspects for a Block Version of the Interval Cholesky Algorithm, *J. Comput. Appl. Math.* **152** (1) (2003) 481–491.

[56] Schäfer, U.: Über Blockversionen des Intervall–Cholesky–Verfahrens, *Proc. Appl. Math. Mech.* **2** (1) (2003) 495–496.

[57] Schrijver, A.: *Theory of Linear and Integer Programming*, Wiley, New York, 1986.

[58] Sharaya, I. and Shary, S. P.: Tolerable Solution Set for Interval Linear Systems with Constraints on Coefficients, Reliable Computing **15** (4) (2011) 345–357.

[59] Shary, S. P.: Solving Tied Interval Linear Systems, *Sib. J. Comput. Math.* **7** (4) (2004) 363–376 (in Russian).

[60] Tanabe, K.: Personal communication at the conference INVA 2008, Okinawa, Japan, March 2008.

[61] Zimmer, M., Krämer, W., and Popova, E. D.: Solvers for the Verified Solution of Parametric Linear Systems, *Computing*, Online First, Nov. 16, 2011.