RIGOROUS ENCLOSURES OF ELLIPSOIDS AND DIRECTED CHOLESKY FACTORIZATIONS

FERENC DOMES, ARNOLD NEUMAIER *

Abstract.
This paper discusses the rigorous enclosure of an ellipsoid by a rectangular box, its interval hull, providing a convenient preprocessing step for constrained optimization problems.

A quadratic inequality constraint with a positive definite Hessian defines an ellipsoid. The Cholesky factorization can be used to transform a strictly convex quadratic constraint into a norm inequality, for which the interval hull is easy to compute analytically. In exact arithmetic, the Cholesky factorization of a nonsingular symmetric matrix exists iff the matrix is positive definite. However, to cope efficiently with rounding errors in inexact arithmetic is nontrivial. Numerical tests show that even nearly singular problems can be handled successfully by our techniques.

To rigorously account for the rounding errors involved in the computation of the interval hull and to handle quadratic inequality constraints having uncertain coefficients, we define the concept of a directed Cholesky factorization, and give two algorithms for computing one. We also discuss how a directed Cholesky factorization can be used for testing positive definiteness. Some numerical test are given in order to exploit the features and boundaries of the directed Cholesky factorization methods.

Key words. quadratic constraints, interval analysis, constraint satisfaction problems, bounding ellipsoids, interval hull, directed Cholesky factorization, verification of positive definiteness, rounding error control, preprocessing, verified computing

AMS subject classifications. 90C20 Quadratic programming, 15A23, Factorization of matrices, 49M27 Decomposition methods

Several state-of-the-art global optimization solvers, such as BARON (by SAHINIDIS & TAWARMALANI [22]) or COCOS (by SCHICHL et al. [23]), combine a number of methods and strategies to find one or more global solutions of a constrained optimization problem. Most of the techniques (e.g branch and bound, heuristics) require explicit bounds for each variable from below and from above. If a problem lacks these explicit prior bounds, the usual remedy is to set default upper and lower bounds on the variables, thereby changing the problem. If the global minimum lies outside the default bounds, the solver cannot find the solution.

A rigorous enclosure technique for strictly convex quadratic constraints presented in this paper gives the possibility to obtain rigorous bounds on variables that are consequences of the constraints, without the need of giving explicit bounds on them. This makes the method a convenient preprocessing step for constrained optimization problems. On the other hand since the enclosures obtained by the method are rigorous, the method is also applicable in verified global optimization (e.g., KEARFOTT [10], LEBBAH [11], DOMES [7]) and in computer-assisted proofs (see, e.g., NEUMAIER [14]). Since it reduces the search space, it may also be important for stochastic, sampling-based optimization methods.

The paper is logically divided into two parts. The first part (Sections 2 - 4) is about computing rigorous enclosures for strictly convex quadratic constraints. In the second part (Sections 5 - 7) the theory of the directed Cholesky factorization is developed as an essential tool for making the results of the first part rigorous.

In the first section we find an optimal box enclosure of an ellipsoid defined by a simple Euclidean norm inequality constraint. In Section 2 we extend these results and

*Faculty of Mathematics, University of Vienna. Nordbergstrasse 15, A-1090 Vienna, Austria

generate optimal enclosures for strictly convex quadratic constraints. We also consider the case of inexact arithmetic, where the error of the factorization of the coefficient matrix has to be controlled. The need of scaling when confronted with ill-conditioned coefficient matrices is discussed in Section 3. In Section 4 we develop the method into a useful tool for preprocessing constrained optimization problems to get finite bounds on the variables or to improve the existing ones. Since the method is rigorous, our preprocessing step finds finite bounds for all variables if each unbounded variable occurs in some strictly convex quadratic constraint, without losing any feasible point. These bounds on all n variables are obtainable with $O(n^3)$ operations. The method is implemented in the GLOPTLAB optimization environment (see DOMES [7]).

To rigorously account for the rounding errors involved in the computation of the interval hull and to handle quadratic inequality constraint having uncertain coefficients, we define the concept of a *directed* Cholesky factorization.

In the second part of the paper, we give algorithms which compute, if possible, for a real, symmetric matrix A a nonsingular triangular matrix R (a *directed* Cholesky factor) such that the error matrix $A - R^T R$ of the factorization is *small* compared to the entries of $|A|$, and guaranteed to be positive semidefinite. Clearly, this implies that A is positive definite; conversely (in the absence of overflow), any 'sufficiently' positive definite symmetric matrix has such a factorization with R representable in floating point arithmetic. The challenge is to find such a representation which makes the error as small as possible and works even for nearly singular matrices.

Two different versions of the directed Cholesky factorization for real symmetric matrices are discussed in Section 5. Both of them check positive definiteness and, when successful, compute a directed Cholesky factor with positive semidefinite error matrix containing small entries. The first approach uses an a priori error estimate, an approximate Cholesky factorization, and the so-called *Gerschgorin test* (explained later). The second one uses directed rounding and diagonal pivoting to obtain a directed Cholesky factor. Section 6 contains some tests and comparison of the two directed Cholesky factorization methods.

In some applications, it is necessary to safeguard the computations in order to ensure the mathematical correctness of the assertions in spite of rounding errors. This applies to computer-assisted proofs in which positive definiteness must be verified rigorously (a potential application to Lie group representations is described in ADAMS [1]). This also applies to box reduction methods for global optimization (see, e.g., [13, 21]) to guarantee that no feasible point is lost.

The last section is concerned with applications of the directed Cholesky factorization for verifying positive definiteness rigorously. Previous work includes ADJIMAN et al. [2, 4, 5], NEUMAIER [13], RUMP [16, 17, 18, 19]. We show that a directed Cholesky factorization can be employed for the same task, and that the positive definiteness of a complex Hermitian matrix can be checked in real arithmetic by factorizing a related real matrix of twice the size.

Notation. We shall use the following notation. $\mathbb{N}_0$ denotes the set of natural numbers including zero, and $\mathbb{R}_+$ the set of nonnegative reals. The n-dimensional identity matrix is denoted by I_n and the n-dimensional zero matrix is denoted by 0_n. The jth row of a matrix A is denoted by $A_{j:}$, the kth column by $A_{:k}$ and the number of nonzero entries by $\mathrm{nnz}(A)$. For an $n \times n$ matrix A, $\mathrm{diag}(A)$ denotes the n-dimensional vector with $\mathrm{diag}(A)_i = A_{ii}$. For an n-dimensional vector x, $\mathrm{Diag}(x)$

denotes the $n \times n$, diagonal matrix A with

$$\mathrm{Diag}(A)_{ij} := \begin{cases} x_i & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

The expression $\|A\|_2 := \sup\{\|Ax\|_2 \mid \|x\|_2 = 1\}$ denotes the spectral norm of a matrix A. The expression $\lambda(A)$ denotes the set of all eigenvalues of a square matrix A. Furthermore, $\lambda_{max}(A)$ denotes the largest and $\lambda_{min}(A)$ the smallest eigenvalue of A. The comparison operators $=, \neq, <, >, \leq, \geq$ for vectors and matrices, and the absolute value $|A|$ of a matrix A are interpreted componentwise. We denote $(A^T)^{-1}$ by A^{-T}. We classify a matrix A as *very small with respect to a matrix* B if $|A| < |B|$ and for a fixed, given tolerance $0 < \kappa \ll 1$ (chosen by the implementation of the method)

$$\max_{i,j}(D_{ij}) \leq \kappa \text{ where } D_{ij} := \begin{cases} |A_{ij}|/|B_{ij}| & \text{if } |A_{ij}| \geq 1 \\ |A_{ij}| & \text{otherwise.} \end{cases} \tag{0.1}$$

For example if

$$A = \begin{pmatrix} 0.001 & 10 \\ 0.002 & 0.004 \end{pmatrix},\ B = \begin{pmatrix} 0.1 & 10000 \\ 0.003 & 0.02 \end{pmatrix},\ \text{then } D = \begin{pmatrix} 0.001 & 0.001 \\ 0.002 & 0.004 \end{pmatrix}$$

and $|A| < |B|$, the matrix A is very small with respect to B if the tolerance κ satisfies $\kappa \geq \max_{i,j}(D_{ij}) = 10^{-3}$.

An interval vector $\mathbf{x} = [\underline{x}, \overline{x}] \in \overline{\mathbb{IR}}^d$ is the Cartesian product of the closed real intervals $\mathbf{x}_i := [\underline{x}_i, \overline{x}_i]$, representing a (bounded or unbounded) axiparallel box in $\mathbb{R}^d$. The values $-\infty$ and ∞ are allowed as lower and upper bounds, respectively, to take care of one-sided bounds on variables. $\overline{\mathbb{IR}}^d$ denotes the set of interval vectors with d components.

$$\langle \mathbf{x} \rangle := \min(|\underline{x}|, |\overline{x}|)$$

defines the *mignitude*,

$$|\,\mathbf{x}| := \max(|\underline{x}|, |\overline{x}|)$$

defines the *magnitude* and

$$\mathrm{mid}(\mathbf{x}) := (\underline{x} + \overline{x})/2$$

defines the *midpoint* of an interval $\mathbf{x}$. We also use the notation $\mathrm{mid}(\mathbf{x})$ component-wise for a bounded box $\mathbf{x}$.

To account for inaccuracies in computed entries of a matrix, consider the interval matrices, standing for uncertain real matrices whose coefficients are between given lower and upper bounds. The expression $\mathbf{A} := [\underline{A}, \overline{A}] \in \overline{\mathbb{IR}}^{m \times n}$ denotes an $m \times n$ interval matrix with lower bound $\underline{A}$ and upper bound $\overline{A}$. $\mathbf{A} \in \overline{\mathbb{IR}}^{n \times n}$ is symmetric if $\mathbf{A}_{ik} = \mathbf{A}_{ki}$ for all $i, k \in \{1, \ldots, n\}$. The *comparison matrix* $\langle \mathbf{A} \rangle$ of a square interval matrix $\mathbf{A}$ is defined by

$$\langle \mathbf{A} \rangle_{ij} := \begin{cases} -|\,\mathbf{A}_{ij}| & \text{for } i \neq j, \\ \langle \mathbf{A}_{ij} \rangle & \text{for } i = j. \end{cases}$$

A real matrix A is identified with the thin interval matrix with $\underline{A} = \overline{A} = A$; in particular, its comparison matrix is

$$\langle A \rangle_{ij} := \begin{cases} -|A_{ij}| & \text{for } i \neq j, \\ |A_{ij}| & \text{for } i = j. \end{cases}$$

The *width and the radius of an interval matrix* A are real matrices and are defined as

$$\mathrm{wid}(\mathbf{A}) := \overline{A} - \underline{A}, \text{ and } \mathrm{rad}(\mathbf{A}) := \mathrm{wid}(\mathbf{A})/2,$$

respectively. A symmetric interval matrix $\mathbf{A} \in \overline{\mathbb{IR}}^{n\times n}$ is called *positive definite* if all symmetric $A \in \mathbf{A}$ are positive definite:

$$x^T A x > 0 \quad \text{for all } x \in \mathbb{R}^n,\ x \neq 0,\ A = A^T \in \mathbf{A}.$$

An interval matrix $\mathbf{A} \in \overline{\mathbb{IR}}^{n\times n}$ is called an *H-Matrix* iff a vector $u > 0$ with $\langle \mathbf{A} \rangle u > 0$ exists (see, e.g. Neumaier [12]).

The well-known theorem of Gerschgorin, (see, e.g. Stoer & Bulirsch [25]) implies that every symmetric H-matrix with non-negative diagonal entries is positive definite; we call this the *Gerschgorin test for positive definiteness*. Other sufficient conditions for positive definiteness based on scaled Gerschgorin theorems and semidefinite programming, form the basis of the αBB method Adjiman et al. [3] and Androulakis [6] and are given in Adjiman et al. [2, 4]. For further tests see the discussion in Section 7.

1. Bounding strictly convex norm constraints. In this section we construct an optimal box enclosure of an ellipsoid defined by the Euclidean norm constraint

$$\|Rx\|_2^2 + 2a^T x \leq \alpha, \tag{1.1}$$

for a given $R \in \mathbb{R}^{n\times n}$, $a \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$. The following result — for ellipsoids centered on the origin ($a = 0$ in (1.1)) — is the basic tool for finding an optimal enclosure of (1.1).

Proposition 1.1. *Suppose that* $R, C \in \mathbb{R}^{n\times n}$, $0 < \beta \in \mathbb{R}$ *and* $0 < d \in \mathbb{R}^n$ *satisfy*

$$d_i \geq \sqrt{(CC^T)_{ii}} \text{ for all } i = 1, \ldots, n, \tag{1.2}$$

as well as

$$\beta d \leq \langle CR \rangle d. \tag{1.3}$$

Then R *is invertible, and for* $x \in \mathbb{R}^n$

$$\|Rx\|_2^2 \leq \delta^2 \quad \Rightarrow \quad |x| \leq \frac{\delta}{\beta} d. \tag{1.4}$$

Proof. We first note that (1.2) implies $(CC^T)_{ii} \leq d_i^2$. Since CC^T is positive semidefinite, $|(CC^T)_{ik}|^2 \leq (CC^T)_{ii}(CC^T)_{kk} \leq d_i^2 d_k^2$, so that

$$|CC^T| \leq dd^T.$$

Since (1.3) implies that CR is an H-matrix (NEUMAIER [12, Section 3.7]), the matrix CR is invertible (hence also R), and $|(CR)^{-1}| \le \langle CR\rangle^{-1}$. Moreover, multiplying (1.3) by $\langle CR\rangle^{-1}\beta^{-1} \ge 0$, we find $\langle CR\rangle^{-1}d \le \beta^{-1}d$. Now let $z := R^{-T}e^i$ with the ith unit vector $e^i = I_{:i}$. Then $e^i = R^T z$, hence

$$\begin{array}{rcl} x_i^2 & = & (e^{iT}x)^2 = (z^T Rx)^2 \le \|z\|_2^2\|Rx\|_2^2 \le \delta^2\|z\|_2^2 = \delta^2 z^T z \\ & = & \delta^2 e^{iT}R^{-1}R^{-T}e^i = \delta^2 e^{iT}(CR)^{-1}CC^T(CR)^{-T}e^i \\ & \le & \delta^2 e^{iT}\langle CR\rangle^{-1}dd^T\langle CR\rangle^{-T}e^i = (\delta e^{iT}\langle CR\rangle^{-1}d)^2 \\ & \le & (\delta\beta^{-1}e^{iT}d)^2 = (\delta\beta^{-1}d_i)^2, \end{array} \tag{1.5}$$

proving (1.4). □

EXAMPLE. 1.2. *For the ellipsoid*

$$\|Rx\|_2^2 \le \delta^2 \text{ with } R = \begin{pmatrix} 2 & -1 \\ 0 & 1 \end{pmatrix} \text{ and } \delta^2 = 10,$$

choosing $C = R^{-1}$, $d_i = \sqrt{(CC^T)_{ii}}$, $\beta = 1$ *and applying the results of Proposition 1.1 we find the enclosure*

$$|x| \le \frac{\delta}{\beta}d = \begin{pmatrix} \sqrt{5} \\ \sqrt{10} \end{pmatrix} \approx \begin{pmatrix} 2.24 \\ 3.17 \end{pmatrix} \text{ (see Figure 1.1).}$$

MATLAB CODE FOR TESTING EXAMPLE 1.2.

```
R=[2 -1;0 1], delta=sqrt(10), C=inv(R),
d=sqrt(diag(C*C')), beta=1, bound=delta/beta*d
```

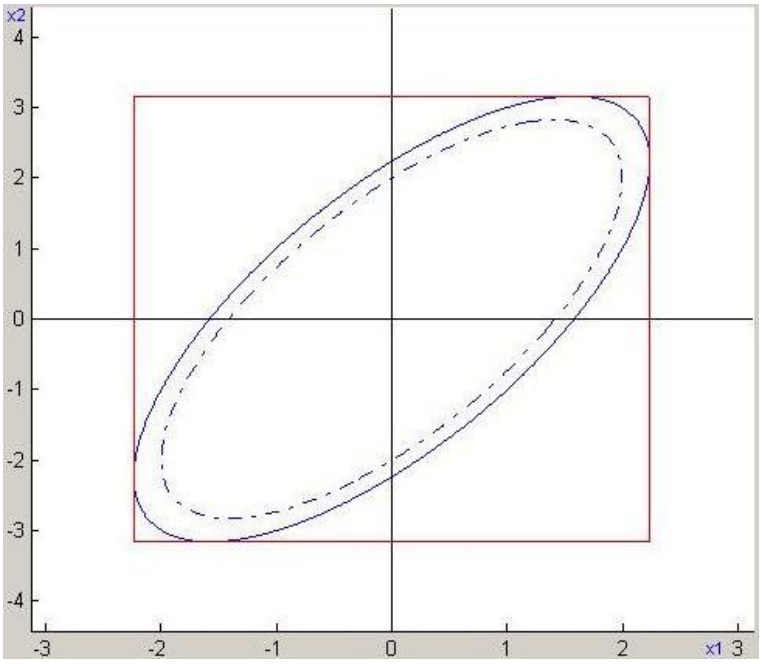

FIG. 1.1. *Box enclosure found for the ellipsoid from Example 1.2*

We now show that the choice for the parameters C, d and β we made in the above example was the optimal one.

PROPOSITION 1.3. *Under the assumptions of Proposition 1.1, the bound on x is optimal if $C = R^{-1}$, $d_i = \sqrt{(CC^T)_{ii}}$ and $\beta = 1$.*

Proof. From $\beta d \leq \langle CR \rangle d$ with $C = R^{-1}$ follows that $\beta \leq 1$. Therefore β is maximal if $\beta = 1$ and d_i is minimal if $d_i = \sqrt{(CC^T)_{ii}}$. The assertion that the bound is optimal follows if we show that for all $i = 1, \ldots, n$ the points

$$\hat{x}^i := \pm \frac{\delta}{d_i} R^{-1} R^{-T} e^i$$

satisfy $\|R\hat{x}^i\|_2 = \delta$ and the ith component of $\hat{x}^i$ matches the boundary of the box $[-\delta d_i, \delta d_i]$ enclosing the ellipsoid. Since

$$d_i = \sqrt{(CC^T)_{ii}} = \sqrt{e^{iT} R^{-1} R^{-T} e^i} = \|R^{-T} e^i\|_2 > 0$$

holds, the first claim follows from

$$\|R\hat{x}^i\|_2 = \| \pm \frac{\delta}{d_i} R R^{-1} R^{-T} e^i\|_2 = \frac{\delta}{d_i} \| \pm R^{-T} e^i\|_2 = \delta,$$

and the second claim follows from

$$d_i \hat{x}^i_i = d_i (e^{iT} \hat{x}^i) = \pm \delta e^{iT} R^{-1} R^{-T} e^i = \pm \delta \|R^{-T} e^i\|_2^2 = \pm \delta d_i^2, \tag{1.6}$$

after division by d_i. □

If we shift the center of the ellipsoid by replacing x in Propositions 1.1 and 1.3 by $x - \tilde{x}$, we find:

COROLLARY 1.4. *Suppose that $R \in \mathbb{R}^{n \times n}$ is invertible, $\tilde{x} \in \mathbb{R}^n$, $d_i \geq \sqrt{(R^{-1}R^{-T})_{ii}}$, and $\beta d \leq \langle R^{-1} R \rangle d$ then*

$$x \in \mathbb{R}^n, \quad \|R(x - \tilde{x})\|_2 \leq \delta \quad \Rightarrow \quad |x - \tilde{x}| \leq \frac{\delta}{\beta} d. \tag{1.7}$$

The bound on $x - \tilde{x}$ is optimal if $d_i = \sqrt{(R^{-1}R^{-T})_{ii}}$ and $\beta = 1$. □

We use Propositions 1.1 and 1.3 to achieve the main result of this section given by the following theorem; we derive cheap and in inexact arithmetic only slightly non optimal bounds on x for the general norm inequality (1.1). The theorem, which is valid for arbitrary $\tilde{z}$, $\tilde{x} \in \mathbb{R}^n$, will be used with

$$\tilde{z} = R^{-T} a, \quad \tilde{x} = -R^{-1} \tilde{z} = -R^{-1} R^{-T} a, \tag{1.8}$$

to make γ small. We know that if the choice is exact then $\gamma = 0$ and the bounds would be optimal.

THEOREM 1.5 (**Ellipsoid Hull**). *For given $R \in \mathbb{R}^{n \times n}$, $a \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$, under the assumptions of Proposition 1.1, for arbitrary $\tilde{z}, \tilde{x} \in \mathbb{R}^n$ and $\gamma, \Delta \in \mathbb{R}$ satisfying*

$$\gamma \geq \|\tilde{z} + R\tilde{x}\|_2 + \beta^{-1} d^T |a - R^T \tilde{z}|, \tag{1.9}$$

and

$$\Delta \geq \gamma^2 + \alpha - 2a^T \tilde{x} - \|R\tilde{x}\|_2^2, \tag{1.10}$$

the following statements hold:

(i) If $\Delta < 0$ *then*

$$\|Rx\|_2^2 + 2a^T x \leq \alpha \tag{1.11}$$

has no solution $x \in \mathbb{R}$*.*

(ii) If $\Delta \geq 0$ *then* (1.11) *implies that*

$$\|R(x-\tilde{x})\|_2 \leq \delta := \gamma + \sqrt{\Delta}, \quad |x-\tilde{x}| \leq \frac{\delta}{\beta} d. \tag{1.12}$$

Proof. For any $x \in \mathbb{R}^n$, Proposition 1.1 implies

$$|x-\tilde{x}| \leq \frac{\varepsilon}{\beta} d, \quad \text{where } \varepsilon = \|R(x-\tilde{x})\|_2. \tag{1.13}$$

If (1.11) holds then

$$\begin{array}{rcl} \|Rx\|_2^2 & \leq & \alpha - 2a^T x \leq \Delta - \gamma^2 + 2a^T\tilde{x} + \|R\tilde{x}\|_2^2 - 2a^T x \\ & = & \Delta - \gamma^2 + \|R\tilde{x}\|_2^2 - 2a^T(x-\tilde{x}). \end{array}$$

Therefore

$$\begin{array}{rcl} \varepsilon^2 & = & \|R(x-\tilde{x})\|_2^2 = (x-\tilde{x})^T R^T R(x-\tilde{x}) \\ & = & x^T R^T Rx - 2\tilde{x}^T R^T Rx + \tilde{x}^T R^T R\tilde{x} \\ & = & \|Rx\|_2^2 - 2\tilde{x}^T R^T Rx + \|R\tilde{x}\|_2^2 \\ & \leq & \Delta - \gamma^2 - 2a^T(x-\tilde{x}) - 2\tilde{x}^T R^T Rx + 2\|R\tilde{x}\|_2^2 \\ & = & \Delta - \gamma^2 - 2(a + R^T R\tilde{x})^T(x-\tilde{x}). \end{array}$$

By (1.9), the inequality

$$\begin{array}{rcl} \Big|(a + R^T R\tilde{x})^T(x-\tilde{x})\Big| & = & \Big|(\tilde{z} + R\tilde{x})^T R(x-\tilde{x}) + (a - R^T\tilde{z})^T(x-\tilde{x})\Big| \\ & \leq & \|\tilde{z} + R\tilde{x}\|_2 \, \|R(x-\tilde{x})\|_2 + |a - R^T\tilde{z}|^T |x-\tilde{x}| \\ & \leq & \|\tilde{z} + R\tilde{x}\|_2 \varepsilon + |a - R^T\tilde{z}|^T \frac{\varepsilon}{\beta} d \leq \varepsilon\gamma, \end{array}$$

holds. We therefore conclude

$$(\varepsilon - \gamma)^2 \leq \Delta - 2\varepsilon\gamma + 2\Big|(a + R^T R\tilde{x})^T(x-\tilde{x})\Big| \leq \Delta.$$

If $\Delta < 0$, we get a contradiction, proving (i). And if $\Delta \geq 0$, we find $\varepsilon \leq \gamma + \sqrt{\Delta} = \delta$, and (ii) follows from (1.13). □

For the choice (1.8) and assuming that the computations are exact, we get the optimal bounds

$$\tilde{x} - \frac{\delta}{\beta} d \leq x \leq \tilde{x} + \frac{\delta}{\beta} d,$$

by using Theorem 1.5. This is shown by the following corollary.

COROLLARY 1.6. *If we chose*

$$\begin{array}{rcl} C & = & R^{-1}, \\ \beta & = & 1, \\ d_i & = & \sqrt{(R^T R^{-T})_{ii}}, \\ \gamma & = & \|\tilde{z} + R\tilde{x}\|_2 + \beta^{-1} d^T |a - R^T \tilde{z}|, \\ \Delta & = & \gamma^2 + \alpha - 2a^T \tilde{x} - \|R\tilde{x}\|_2^2, \\ \tilde{z} & = & R^{-T} a, \\ \tilde{x} & = & -R^{-1} R^{-T} a, \end{array}$$

then Theorem 1.5 holds, and in the case of $\Delta \geq 0$ *the bound on* $x - \tilde{x}$ *in* (1.12) *is optimal.*

Proof. By the choice of C, β, d, γ and Δ we have

$$\gamma = \|\tilde{z} + R\tilde{x}\|_2 + \beta^{-1} d^T |a - R^T \tilde{z}| = 0, \tag{1.14}$$

and

$$\begin{array}{rcl} \Delta & = & \alpha - 2a^T \tilde{x} - \|R\tilde{x}\|_2^2 \\ & = & \alpha + 2a^T R^{-1} R^{-T} a - \| - RR^{-1} R^{-T} a\|_2^2 = \alpha + \|R^{-T} a\|_2^2. \end{array} \tag{1.15}$$

By the choice of $\tilde{z}$ and $\tilde{x}$ we have

$$\begin{array}{rcl} \|R(x - \tilde{x})\|_2^2 & = & \|Rx\|_2^2 + \|R\tilde{x}\|_2^2 - x^T R^T R\tilde{x} - \tilde{x}^T R^T Rx \\ & = & \|Rx\|_2^2 + \|R^{-T} a\|_2^2 + 2a^T x, \end{array}$$

Therefore, $\|R(x - \tilde{x})\|_2^2 \leq \delta^2 = \Delta$ implies $\|Rx\|_2^2 + \|R^{-T} a\|_2^2 + 2a^T x \leq \alpha + \|R^{-T} a\|_2^2$ by (1.15), hence $\|Rx\|_2^2 + 2a^T x \leq \alpha$. This gives the forward direction of

$$\|R(x - \tilde{x})\|_2 \leq \delta \quad \Leftrightarrow \quad \|Rx\|_2^2 + 2a^T x \leq \alpha$$

and the reverse direction follows from (1.12). By the choice of d_i and β, we can apply the second part of Corollary 1.4, proving that the bound on $x - \tilde{x}$ is optimal. □

In practice, one cannot make the required choices in Corollary 1.6 exact, since rounding errors affect the results of the defining formulas. However, using approximations for $\tilde{x}$ and $\tilde{z}$ computed by ordinary floating point arithmetic, tight bounds which take account of the rounding errors are easy to get with directed, upward rounding. In this way we get nearly optimal enclosure. In 2 dimensions, the results are visually indistinguishable from the optimal enclosures. In exact arithmetics however, by Theorem 1.5 and Corollary 1.6 we can summarize:

THEOREM 1.7. *Let* $\|Rx\|_2^2 + 2a^T x \leq \alpha$ *be an ellipsoid. Suppose that* R *is invertible, then*

$$\mathbf{x} = [\tilde{x} - \delta d, \tilde{x} + \delta d]$$

with $d_i = \sqrt{(R^{-1} R^{-T})_{ii}}$, $\tilde{x} = -R^{-1} R^{-T} a$ *and* $\delta = \sqrt{\alpha + \|R^{-T} a\|_2^2}$ *defines the interval hull for the given ellipsoid.* □

EXAMPLE. 1.8. *For the ellipsoid*

$$\|Rx\|_2^2 + 2a^T x \leq \alpha \text{ with } R = \begin{pmatrix} 2 & -1 \\ 0 & 1 \end{pmatrix}, \; 2a^T = (2 \;\; 3), \; \alpha = 10,$$

which is shown Figure 1.2, we apply Theorem 1.7 and obtain the outward rounded (to three significant digits) interval hull

$$\mathbf{x} = \begin{pmatrix} [-3.92, 1.42] \\ [-5.78, 1.78] \end{pmatrix}.$$

MATLAB CODE FOR TESTING EXAMPLE 1.8.

```
R=[2 -1;0 1], a=[2;3]./2, alpha=10, C=inv(R),
d=sqrt(diag(C*C')), delta=sqrt(alpha+norm(C'*a)^2,
xhat=-C*C'*a, xl=xhat-delta*d, xu=xhat+delta*d
```

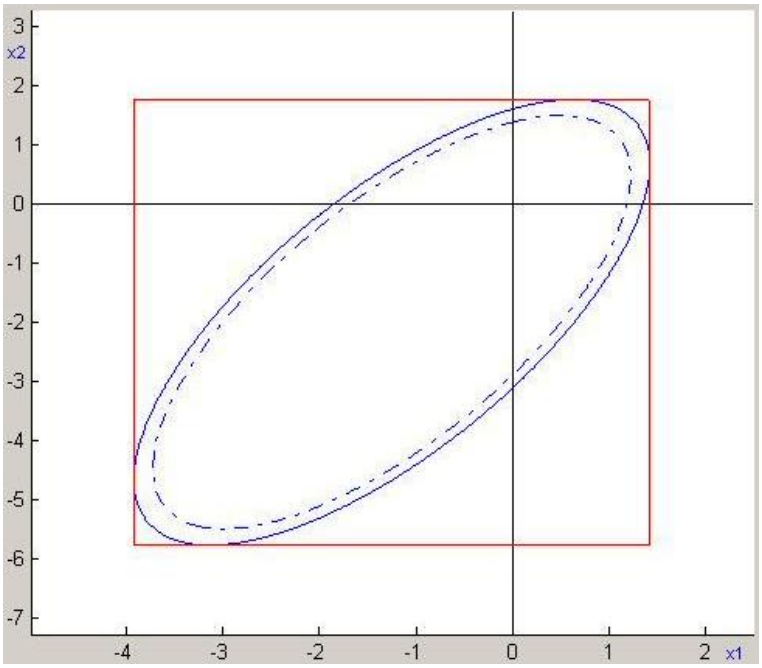

FIG. 1.2. *Optimal box enclosure of the ellipsoid defined in Example 1.8*

2. Enclosing strictly convex quadratic constraints. In this section we apply results of the previous section to enclose strictly convex quadratic constraints in inexact arithmetic. To efficiently cope with the rounding errors we use a method called the directed Cholesky factorization to transform a strictly convex quadratic constraint into a Euclidean norm constraint (1.1). The directed Cholesky factorization takes the rounding errors involved in the transformation into account and is discussed in detail in Sections 5.

Let A be a symmetric, positive definite matrix. The strictly convex quadratic constraint

$$x^T Ax + 2a^T x \leq \alpha \tag{2.1}$$

describes an ellipsoid. We derive a nearly optimal enclosure $\mathbf{x}$ for this ellipsoid such that each x satisfying (2.1) is contained in the box $\mathbf{x}$ (hence the method is rigorous).

We compute a permutation matrix P and an upper triangular matrix R such that the residual matrix $\hat{E} := PAP^T - \hat{R}^T\hat{R}$ is positive semidefinite and is very small

with respect to PAP^T (details in Section 5). If the factorization fails, the positive definiteness of A cannot be verified and the enclosure cannot be computed. (This case only happens when A is an indefinite or a nearly indefinite matrix.) If the factorization is successful the constraint is strictly convex and we have

$$A = P^T \hat{R}^T \hat{R} P + P^T \hat{E} P, \tag{2.2}$$

where the residual matrix $\hat{E}$ (and also $P^T \hat{E} P$) is positive semidefinite and very small with respect to A. Substituting in (2.1) we have

$$x^T (P^T \hat{R}^T \hat{R} P + P^T \hat{E} P) x + 2a^T x = \|\hat{R} P x\|_2^2 + x^T P^T \hat{E} P x + 2a^T x \leq \alpha,$$

and if we define $R := \hat{R} P$, we end up in

$$\|Rx\|_2^2 + 2a^T x \leq \alpha - x^T P^T \hat{E} P x \leq \alpha. \tag{2.3}$$

This proves that the ellipsoid defined by (2.1) is fully contained in the ellipsoid given by the norm constraint

$$\|Rx\|_2^2 + 2a^T x \leq \alpha. \tag{2.4}$$

Note that (2.3) is only then a valid inequality if the residual matrix $\hat{E}$ is positive semidefinite. Since $\hat{E}$ is very small with respect to PAP^T, the relative approximation error

$$\delta(x) := \frac{x^T \hat{E} x}{\|Rx\|_2^2},$$

is also small, for all $x \in \mathbf{x}$.

We apply the main result of Section 1, Theorem 1.5 and Corollary 1.6, to (2.4), choosing $\tilde{z} \approx R^{-T} a$ and $\tilde{x} \approx -R^{-1} \tilde{z}$ by ordinary floating point calculations, and the remaining variables optimally, by computing the corresponding expressions with directed rounding or interval arithmetic. The details are given in the following algorithm.

ALGORITHM 2.1 (Ellipsoid Hull).
Compute a box enclosure of strictly convex quadratic constraint $x^T Ax + 2a^T x \leq \alpha$:

1. *Find a directed Cholesky factorization of the matrix A:*
   (a) *if the factorization fails, the positive definiteness of A cannot be verified and the enclosure cannot be computed,*
   (b) *otherwise a directed Cholesky factor R is obtained.*
2. *Compute the approximative inverse C of the matrix R.*
3. *Compute d with $d_i = \inf(\sqrt{(CC^T)_{ii}})$ by using directed rounding.*
4. *Use upward rounding to compute $h = \langle CR \rangle d$ and obtain $\beta = \max\{h_i/d_i | i = 1 \dots n\}$ which must be approximately one.*
5. *Set $\tilde{z} = C^T a$ and $\tilde{x} = -C\tilde{z}$ and compute an enclosure $[\underline{\gamma}, \overline{\gamma}]$ of the expression*

$$\|\tilde{z} + R\tilde{x}\|_2 + \beta^{-1} d^T |a - R^T \tilde{z}|$$

and an enclosure $[\underline{\Delta}, \overline{\Delta}]$ of the expression

$$\gamma^2 + \alpha - 2a^T \tilde{x} - \|R\tilde{x}\|_2^2,$$

by using interval arithmetic.

6. *Finally, use outward rounding to compute the interval*

$$[\underline{\delta}, \overline{\delta}] := [\underline{\gamma} + \sqrt{\underline{\Delta}}, \overline{\gamma} + \sqrt{\overline{\Delta}}].$$

7. *The result is an approximate but rigorous enclosing ellipsoid for* (2.1)*, given by the norm constraint* $\|R(x - \tilde{x})\|_2 \leq \overline{\delta}$*, as well as the rigorous box enclosure*

$$x \in \left[(\delta/\overline{\beta})d - \tilde{x}, (\delta/\underline{\beta})d + \tilde{x}\right].$$

The algorithm applies with trivial modifications if A and a are uncertain (their components vary in intervals). This form is implemented in GloptLab (see [7]).

3. Scaling. The ellipsoid hull approximation which was presented in the previous section may have difficulties when used on badly scaled systems. Scaling the constraints before applying the Cholesky factorization increases the range of matrices which can be successfully factorized[1].

To demonstrate this behavior we discuss a four dimensional problem, first presented in Domes & Neumaier [8], consisting of the single constraint

$$\begin{array}{rcl} 4x_1^2 + 4Nx_1x_2 + 12x_1x_3 - 28x_1x_4 + (1+N^2)x_2^2 + (6N-2)x_2x_3 & & \\ -(14N+10)x_2x_4 + 11x_3^2 - 32x_3x_4 + 75x_4^2 + 2x_2 + 2Nx_3 + 26 & \leq & 0. \end{array} \tag{3.1}$$

Writing (3.1) in the form $x^T Ax + 2ax \leq -26$ with $x = (x_1, x_2, x_3, x_4)^T$, $2a = (0,\ 2,\ 2N, 0)$ and

$$A = B^T B, \text{ where } B := \begin{pmatrix} R & S \\ 0 & I \end{pmatrix}, \quad R = \begin{pmatrix} 2 & N \\ 0 & 1 \end{pmatrix} \text{ and } S = \begin{pmatrix} 3 & -7 \\ -1 & -5 \end{pmatrix},$$

we see that the symmetric matrix A is manifestly positive definite. Thus (3.1) describes an ellipsoid. If N is chosen large enough, A is very ill-conditioned.

For example if we choose $N = 5 \cdot 10^6$ then the 2-norm condition number of A is approximately $5 \cdot 10^{21}$, therefore A is nearly singular; the lowest eigenvalue is approximately $5 \cdot 10^{-15}$. It is no surprise that for this matrix both the directed Cholesky factorization using the Gerschgorin test and the directed Cholesky factorization with pivoting fails (the reasons for this and both methods are explained in Section 5).

If we use the scaling algorithm ScaleLP from Domes & Neumaier [8] on the problem, we obtain

$$D = \mathrm{Diag}(10^6 \quad 1 \quad 10^6 \quad 10^5)$$

as scaling matrix for the variables. Here, scaling makes an essential difference since the scaled problem

$$x^T DADx + bDx \leq -26$$

has a 10^6 times lower condition number (approximately $6 \cdot 10^{15}$) which — however it is still high — it is small enough for the directed Cholesky factorization with pivoting to be successful. Therefore we obtain the factorization

$$PDADP^T = R^T R + E$$

[1]It is interesting that Rump [19] also mentions the importance of scaling of the matrix A in his algorithm for checking definiteness. He uses a minimum degree ordering and a scaling technique based on the results of the van der Sluis method given in Higham [9, Chapter 7])

with

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}, R = \begin{pmatrix} 5.0 \cdot 10^6 & 3.0 \cdot 10^6 & -7.0 \cdot 10^5 & 2.0 \cdot 10^6 \\ 0 & 1.4 \cdot 10^6 & 3.5 \cdot 10^5 & 2.8 \cdot 10^{-1} \\ 0 & 0 & 3.7 \cdot 10^5 & 2.7 \cdot 10^{-1} \\ 0 & 0 & 0 & 1.1 \cdot 10^{-2} \end{pmatrix}.$$

Finally, computing the interval hull by the method from Section 2 we find the bounds

$$x \in ([-146, 1442],\ [-5 \cdot 10^8, 5 \cdot 10^7],\ [-3 \cdot 10^{-5}, 10^{-12}],\ [-10^{-2}, 10^{-3}])^T$$

for the constraint (3.1).

If we compute an upper bound for the residual E by using interval arithmetic we obtain the positive definite matrix:

$$\overline{E} = \sup(PD\mathbf{A}DP^T - R^T R) = \begin{pmatrix} 3.52 \cdot 10^{-2} & 0 & 0 & 0 \\ 0 & 2.58 \cdot 10^4 & 2.44 \cdot 10^{-4} & 9.77 \cdot 10^{-4} \\ 0 & 2.44 \cdot 10^{-4} & 2.42 \cdot 10^3 & 2.44 \cdot 10^{-4} \\ 0 & 9.77 \cdot 10^{-4} & 2.44 \cdot 10^{-4} & 146 \cdot 10^{-3} \end{pmatrix}.$$

As one can notice the second and the third diagonal entry of $\overline{E}$ is very large. This seem to be a contradiction since E is supposed to be very small. However the component wise relative error

$$\delta := \max_{i,j} \left| \frac{\overline{E}_{ij}}{A_{ij}} \right| = 1.0317 \cdot 10^{-9},$$

indicates that E must be a very small perturbation of A. Thus E is indeed *very small with respect to* A in the sense as defined in (0.1) in notation part of the first section.

4. Preprocessing constrained optimization problems. Optimization is a constantly developing, complex and important field of the numerical mathematics. The goal of solving an optimization problem is to find a local or a global minimum of the objective function $f(x)$, subject to the general constraints $G(x) \in \mathbf{w}$ (including equality and inequality constraints) and to the bound constraints $x \in \mathbf{x}$:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{s.t.} & G(x) \in \mathbf{w}, \\ & x \in \mathbf{x}. \end{array} \tag{4.1}$$

If we search for a global minimum of the problem, it is called a *global optimization problem*.

If an $\hat{x} \in \mathbf{x}$ satisfies $G(\hat{x}) \in \mathbf{w}$, $\hat{x}$ is called a feasible point. If there is no objective function given or it is constant, the goal is to find a good enclosure of the set of all feasible points. In this case the problem is called a *constraint satisfaction problem*. Also in the case of a global search the first step is often to solve the constraint satisfaction problem in order to bound and reduce the search space as much as possible. In this chapter we show how the ellipsoid hull enclosure technique presented in the previous sections can be used for this purpose.

Several state of the art global optimization solvers (e.g., [23] or [22]) combine a number of methods and strategies to find one or more global solution of (4.1). Most of the techniques (e.g branch and bound, heuristics) require that the bound constraints $x \in \mathbf{x}$ are finite. If a problem lacks the desired bounds, the usual remedy is to set

default upper and lower bounds on the variables, thus changing the problem. If the global minimum lies outside the default bounds, the solver cannot find the solution.

The main advantage of the enclosure techniques from sections 1 and 2 is that they give us the possibility to obtain rigorous bounds on some variables which are consequences of the constraints without the need of having explicit bounds on the variables.

Since the enclosures are rigorous, the method is also applicable in verified global optimization (e.g., [7], [10], [11]) and in computer-assisted proofs (see e.g., [14]).

If we already have found a good and feasible point $\hat{x} \in \mathbf{x}$ we can apply Algorithm 2.1 to the constraint satisfaction problem

$$\begin{array}{ll} \text{minimize} & 1 \\ \text{s.t.} & f(x) \le f(\hat{x}), \\ & G(x) \in \mathbf{w}, \\ & x \in \mathbf{x}, \end{array} \tag{4.2}$$

and may obtain new bounds on the variables as well as on the objective function $f(x)$. This makes our method a valuable tool not only for preprocessing and solving constraint satisfaction problems but also for global optimization.

To enhance the results of sections 1 and 2, we discuss the application to optimization problems given in the form of (4.1). The m general constraints are interpreted as componentwise enclosures $G_i(x) \in \mathbf{w}_i$ $(i = 1 \dots m)$. This includes equality constraints if $\mathbf{w}_i = [\underline{w}_i, \overline{w}_i]$ is a degenerate interval with $\underline{w}_i = \overline{w}_i$, inequality constraints if one bound of $\mathbf{w}_i$ is infinite, and two sided inequalities $\underline{w}_i \le G_i(x) \le \overline{w}_i$ if both bounds are finite. If we have bounds on the objective function, it should be treated like an ordinary general constraint. Similarly, the n bound constraints are interpreted as enclosures $x_j \in \mathbf{x}_j$ with $j = 1 \dots n$. Again, fixed variables and one-sided bounds on the variables are included as special cases.

We may apply Algorithm 2.1 for each quadratic constraint $G_i(x) \in \mathbf{w}_i$ separately. Thereby only the finite bounds of $\mathbf{w}_j$ are taken into account resulting in one or two inequality constraints in the form of

$$x^T A x + 2a^T x \le \alpha \tag{4.3}$$

with coefficients obtained by the Taylor expansion of $G_j(x)$ around $x = 0$,

$$\begin{aligned} A_{jk} &= \frac{1}{2}\frac{\partial^2 G_i}{\partial x_j \partial x_k}(0), \quad 2a_j = \frac{\partial G_i}{\partial x_j}(0), \quad \alpha = \overline{w}_j - G_j(0) \quad \text{if} \quad \overline{w}_j \text{ is finite} \\ A_{jk} &= -\frac{1}{2}\frac{\partial^2 G_i}{\partial x_j \partial x_k}(0), \quad 2a_j = -\frac{\partial G_i}{\partial x_j}(0), \quad \alpha = -\underline{w}_j - G_j(0) \quad \text{if} \quad \underline{w}_j \text{ is finite} \end{aligned}$$

The size of the problem can of course be restricted to those variables on which $G_j(x)$ actually depends.

If the constraint (4.3) is strictly convex we obtain new bounding box $\mathbf{u}$ on the variables. If we cut the original bound constraints $x \in \mathbf{x}$ with $\mathbf{u}$ we obtain the new bound constrains

$$x \in \hat{\mathbf{x}} \text{ with } \hat{\mathbf{x}}_i := \mathbf{u}_i \cap \mathbf{x}_i = [\max(\underline{x}_i, \underline{u}_i), \min(\overline{x}_i, \overline{u}_i)]. \tag{4.4}$$

Because $\mathbf{u}$ is bounded, the box $\hat{\mathbf{x}}$ is also bounded. If we process all quadratic $G_j(x) \in \mathbf{w}_j$, we obtain an interval enclosure of the intersection of all strictly convex quadratic constraint.

The method can be greatly enhanced by removing the linear variables. This is crucial in the presence of slack variables which are only linear in the constraints. If the linear variables are not removed the matrix A has a zero row and column, hence it is singular, therefore the directed Cholesky factorization will fail and we cannot compute new bounds. To remove the linear terms from the constraint

$$x^T Ax + 2a^T x \leq \beta$$

we write

$$x_I^T \hat{A} x_I + 2a^T x_I + c^T x_J \leq \beta \tag{4.5}$$

with J being the index set of the variables which are only linear and I being the index set of the variables which have nonlinear terms in the constraint. The dimension of $\hat{A}$ and a is reduced to $n' := |I|$, and the dimension of c is $n'' := |J|$. We modify (4.5) by bounding and removing the linear variables and obtain

$$x_I^T \hat{A} x_I + 2a^T x_I \leq \beta + \sum_{j \in J} (-c_j \underline{x}_j). \tag{4.6}$$

Here bracketing the right hand side of the above expression yields a correct bound when evaluating it using floating-point arithmetic with upward rounding. We can now write the new inequality (4.6) in the form of (4.3), with

$$x := x_I, \quad A := \hat{A}, \quad \alpha := \beta + \sum_{j \in J} (-c_j \underline{x}_j)$$

and factorizing A using a directed Cholesky factorization method. From this point on, all steps are the same as above, except for the fact that we compute new bounds only on the remaining n' variables. This should be accounted for when cutting the resulting $|x| \leq u$ with the original box. Thus we compute $\hat{\mathbf{x}}_i$ as in (4.4) only for $i \in I$ and set the remaining $\hat{\mathbf{x}}_j := \mathbf{x}_j$ for all $j \in J$. Proceeding in this way allows us to handle a bigger class of problems, by avoiding unnecessary singularity of the matrix A.

In practice, many problems have nonquadratic constraints. These relaxations can be handled as above if convex quadratic relaxations of such constraints can be computed (see Skutella [24] and Rendl [15] for possible techniques). This further extends the scope of our methods.

5. Directed Cholesky factorization. Let A be a symmetric matrix. A *directed Cholesky factorization* of the matrix A is an approximate factorization $A \approx R^T R$ with nonsingular upper triangular R such that the error matrix $A - R^T R$ of the factorization is positive semidefinite. The matrix R is called a directed Cholesky factor of A.

Proposition 5.1. *A directed Cholesky factorization exists iff A is positive definite.*

Proof. If R is a directed Cholesky factor of A then $E := A - R^T R$ is positive semidefinite, therefore $A = E + R^T R$ is positive definite. Conversely, if A is positive definite, we may take for R the Cholesky factor of A then $A - R^T R = 0$ is positive semidefinite; hence R is a directed Cholesky factor of A. □

In finite precision arithmetic, however, R is usually not representable exactly, and simply rounding it is often not sufficient to make $A - R^T R$ positive semidefinite. Thus

finding a directed Cholesky factorization needs additional considerations. To represent the general setting we factor a symmetric interval matrix $\mathbf{A} := [\underline{A}, \overline{A}] \in \overline{\mathbb{IR}}^{n\times n}$. This form also represents the case when a matrix is not exactly known, as it is the result of inaccurate measurements or computations. We present the following methods to compute a directed Cholesky factorization such that the residual matrix $\mathbf{A} - R^T R$ is very small with respect to the matrix $\overline{A}$.

5.1. Directed Cholesky factorization using the Gerschgorin test. Our first method for computing a directed Cholesky factorization for an interval matrix $\mathbf{A}$ is based on the Gerschgorin test[2]. If $\underline{A}_{ii} \leq 0$ for any $i \in \{1, \ldots, n\}$ then not all symmetric $A \in \mathbf{A}$ are positive definite and a factorization with a nonsingular R is not possible. If the lower bounds of the diagonal entries of $\mathbf{A}$ are positive, we choose a matrix $\tilde{A} \in \mathbf{A}$ and slightly perturb its diagonal entries by using a suitable chosen a priori error estimation constant σ. Then we apply the approximate Cholesky factorization $R^T R \approx \tilde{A}$ to the perturbed matrix. If the error estimation constant σ was chosen correctly, even positive but nearly indefinite matrices (where the approximate Cholesky factorization would fail for the unperturbed matrix) can be factorized. If the Cholesky factorization succeeds the error matrix $\mathbf{E} := \mathbf{A} - R^T R$ is computed by using interval arithmetic. Finally we test $\mathbf{E}$ for positive definiteness with the Gerschgorin test. Again the right choice of σ is crucial, since if it was chosen unnecessary large the increased width of the interval error matrix has a negative effect on the outcome of the Gerschgorin test. If the Gerschgorin test is positive then R is a directed Cholesky factor of the matrix $\mathbf{A}$.

The following algorithm summarizes the above consideration:

ALGORITHM 5.2 (DirCholG).

Compute a directed Cholesky factorization of a symmetric interval matrix, using the Gerschgorin test:

1. *Let $\mathbf{A} = [\underline{A}, \overline{A}]$ be a symmetric n–dimensional interval matrix.*
2. *If $\underline{A}_{ii} \leq 0$ for some $i \in \{1, \ldots, n\}$ the factorization is not possible. Stop.*
3. *We define the matrix*

$$\tilde{A}_{ij} = \begin{cases} \overline{A}_{ij} & \text{if } \ \overline{A}_{ij} \geq -\underline{A}_{ij} \text{ and } i \neq j \\ \underline{A}_{ij} & \text{otherwise}\,. \end{cases} \tag{5.1}$$

4. *Perturb the diagonal entries of the matrix $\tilde{A}$:*
   (a) *Generate a diagonal perturbation matrix D ($D_{ij} = 0$ for $i \neq j$) which depends on the diagonal entries of $\tilde{A}$ and the width of the interval matrix $\mathbf{A}$:*

$$D_{ii} := \tilde{A}_{ii} - \frac{\sum_{j=1}^{n} (\overline{A}_{ij} - \underline{A}_{ij}) u_j}{u_i}, \text{ where } u_i = 1/\tilde{A}_{ii} \text{ for all } 1 \leq i \leq n.$$

   (b) *Choose an approximate a priori error estimation constant σ such that the Cholesky factorization of $A' := \tilde{A} - \sigma D$ is positive definite enough even in a nearly indefinite case (suitable selections for σ are discussed later).*
5. *Compute $A' \approx R^T R$ approximately, and $\mathbf{E} := \mathbf{A} - R^T R$ by using interval arithmetic.*

[2]the Gerschgorin test is described in the notation of the first section of this paper.

6. *If* $\underline{E}_{ii} \geq 0$ *for all* i *and* $\mathbf{E}$ *is an H-matrix then* $\mathbf{E}$ *is positive definite (Gerschgorin test), the factorization is successful and the directed Cholesky factor* R *is returned.*

PROPOSITION 5.3. *If Algorithm 5.2 is successful we obtain a directed Cholesky factor* R *such that for all symmetric* $A \in \mathbf{A}$, $A - R^T R$ *is positive definite.*

Proof. Let be R the matrix returned by the algorithm. Since we use interval arithmetic in Step 5 of Algorithm 5.2 the bound $\mathbf{E}$ on $\mathbf{A} - R^T R$ is rigorous. Since by Step 6 of Algorithm 5.2, $\mathbf{E}$ is a H-matrix, the Gerschgorin test implies the assertion. □

Comments on Algorithm 5.2:

(ad 3.) The algorithm would be also correct if $\tilde{A}$ in (5.1) is replaced by an arbitrary $\tilde{A} \in \mathbf{A}$ with $\tilde{A}_{ii} = \underline{A}_{ii}$.

(ad 4.) The perturbation applied to the diagonal entries of $\tilde{A}$ is needed for nearly indefinite matrices. By using this approach we may obtain the approximate Cholesky factor of the perturbed matrix, even if we would fail for the unperturbed one.

(ad 6.) The test whether or not the matrix $\mathbf{E}$ is an H-matrix can be done by choosing a suitable $u > 0$ and test whether or not $\langle \mathbf{E} \rangle u > 0$ holds. Different choices of u are

- $u = (1, \ldots, 1)^T$ is the simplest, proving diagonal dominance, but not scaling invariant,
- $u \approx 1/\operatorname{diag}(E)$, is a generally good and cheap choice,
- $u \approx \langle E \rangle^{-1} (1, \ldots, 1)^T$, is the best choice (see NEUMAIER [12]), but requires $O(n^3)$ operation for solving the linear system.

The selection of an approximate a priori error estimation constant σ is critical for nearly indefinite matrices. If we choose σ too small, the approximate Cholesky factorization will possibly fail; if we choose it to large, the error matrix $\mathbf{E}$ will be too large and it will not pass the H-matrix test.

The following theorem which can be found in HIGHAM [9, pp. 203–224] gives information about the feasibility of a numerical Cholesky factorization when all arithmetic operations are executed with a relative error of at most ε (when no overflow or underflow occurs).

THEOREM 5.4 (Demmel). *Let* $A \in \mathbb{R}^{n \times n}$ *be a symmetric matrix with positive diagonal elements, and a diagonal matrix* D *with* $D_{ii} = A_{ii}^{-1/2}$. *If*

$$\lambda_{min}(DAD) > \sigma := \frac{n(n+1)\varepsilon}{1 - 2(n+1)\varepsilon} \tag{5.2}$$

then the Cholesky factorization applied to A *succeeds and produces a nonsingular* R. *If* $\lambda_{min}(A) \leq -\sigma$ *then the computation is certain to fail.*

Theorem 5.4 seems to give a good choice for σ, but in reality it is significantly larger than would be needed to successfully factor nearly indefinite matrices by the approximate Cholesky factorization. This makes it harder to pass Gerschgorin test. By our heuristic experiments we found a more suitable choice for σ; we try $\sigma = \varepsilon(0.015\,\mathrm{nnz}(\mathbf{A}) + 0.5n)$ in the first run, then in the case of failure $\sigma = \varepsilon(0.015\,\mathrm{nnz}(\mathbf{A}) + n)$ in the second one. The results of this strategy are satisfactory.

5.2. Directed Cholesky factorization with pivoting. For a symmetric, positive definite interval matrix $\mathbf{A}$, the *directed Cholesky factorization with diagonal pivoting* computes a permutation matrix P and an upper triangular matrix R such that for every $A \in \mathbf{A}$, the residual matrix $E := PAP^T - R^T R$ is positive semidefinite and is very small with respect to A. We first state the algorithm, then discuss the conditions under which the residual matrix is positive semidefinite and is very small.

The following algorithm either computes a directed Cholesky factor R and a permutation matrix P such that the residual matrix E is positive semidefinite and is very small with respect to A, or it terminates with an error message and returns an incomplete factorization:

ALGORITHM 5.5 (DirCholP).
Directed Cholesky factorization of symmetric interval matrix, using directed rounding and diagonal pivoting.

1. *Let $\mathbf{A} = [\underline{A}, \overline{A}]$ be an n–dimensional symmetric interval matrix. Set $\mathbf{A}_1 = \mathbf{A}$, $R = 0_n$, $P = I_n$, and the rounding mode to upward rounding.*
2. *For $k = 1, \ldots, n$ do the following steps:*
   (a) *Find the pivot element $\alpha = \max(\mathrm{diag}(\underline{A}_k))$ on the diagonal of the matrix $\underline{A}_k \in \mathbb{R}^{n-k+1}$. Let j denote the index of the pivot element; interchange row j with the first row and column j with the first column, in the interval matrix $\mathbf{A}_k$. $\mathbf{A}_k$ remains symmetric. Exchange the same rows and columns in the matrix P.*
   (b) *Partition the permuted interval matrix $\mathbf{A}_k$ as:*

$$\mathbf{A}_k = \begin{pmatrix} \boldsymbol{\alpha}_k & \mathbf{a}^T \\ \mathbf{a}_k & \mathbf{B}_k \end{pmatrix}.$$

   (c) *If $\underline{\alpha}_k \leq 0$ terminate Step 2. and return an error message.*
   (d) *Choose γ_k with $0 < \gamma_k < 1$, $\rho_k = \gamma_k \sqrt{\underline{\alpha}_k}$ and $r_k = (\overline{a}_k + \underline{a}_k)/(2\rho_k)$.*
   (e) *Set $R_{kk} = \rho_k$ and $R_{k,k:n} = r_k^T$.*
   (f) *Compute $\delta_k := -(-\underline{\alpha}_k + \rho_k^2)$ and $d_k := \max(\overline{a}_k + \rho_k(-r_k),\ \rho_k r_k - \underline{a}_k)$.*
   (g) *If the residual pivot $\delta_k \leq 0$ terminate Step 2. and return an error message.*
   (h) *Set $\mathbf{A}_{k+1} := [\underline{B}_k - r_k r_k^T - d_k d_k^T/\delta_k,\ \overline{B}_k + (-r_k) r_k^T + d_k d_k^T/\delta_k]$.*
3. *If Step 2. is finished without an error message we obtain the upper triangular matrix R and the permutation matrix P, if an error message was produced the incomplete factorization is returned.*

THEOREM 5.6. *Suppose that Algorithm 5.5 used for the symmetric interval matrix $\mathbf{A}$ terminates without an error message and returns the matrix R. Then $PAP^T - R^T R$ is positive semidefinite for all symmetric $A \in \mathbf{A}$.*

To prove the proposition we need some preparations:

PROPOSITION 5.7. *Let*

$$A := \begin{pmatrix} \alpha & a^T \\ a & B \end{pmatrix} \in \mathbf{A} := \begin{pmatrix} \boldsymbol{\alpha} & \mathbf{a}^T \\ \mathbf{a} & \mathbf{B} \end{pmatrix} \in \mathbb{IR}^{n \times n}, \tag{5.3}$$

with $\underline{\alpha} > 0$, then:
(i) For arbitrary $\rho \in \mathbb{R}$, $|\rho| < \sqrt{\underline{\alpha}}$, $r \in \mathbb{R}^{n-1}$ and

$$\varepsilon := \alpha - \rho^2 > 0, \quad e := a - \rho r, \quad A_0 := B - rr^T - \frac{ee^T}{\varepsilon}, \tag{5.4}$$

we have

$$A = \begin{pmatrix} \rho \\ r \end{pmatrix} \begin{pmatrix} \rho \\ r \end{pmatrix}^T + \begin{pmatrix} \varepsilon & e^T \\ e & ee^T/\varepsilon \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & A_0 \end{pmatrix}. \tag{5.5}$$

(ii) The bounds $\varepsilon \geq \delta$ and $|e| \leq d$ and $A_0 \in \mathbf{A}_0$ are satisfied if

$$\begin{array}{l} 0 < \delta \leq -(-\underline{\alpha} + \rho^2), \\ d \geq \max(\overline{a} + \rho(-r), \rho r - \underline{a}), \\ \mathbf{A}_0 \supseteq [\underline{B} - rr^T - dd^T/\delta,\ \overline{B} + (-r)r^T + dd^T/\delta]. \end{array} \tag{5.6}$$

Proof. (i) Since $|\rho| < \sqrt{\underline{\alpha}}$ we have $\varepsilon = \alpha - \rho^2 > \alpha - \underline{\alpha} \geq 0$ for all $\alpha \in \boldsymbol{\alpha}$, so that $\varepsilon > 0$. Thus (5.5) is well defined. Substituting (5.4) into (5.5) gives

$$A = \begin{pmatrix} \rho^2 + \varepsilon & \rho r^T + e^T \\ \rho r + e & rr^T + ee^T/\varepsilon + A_0 \end{pmatrix} = \begin{pmatrix} \alpha & a^T \\ a & B \end{pmatrix}.$$

(ii) By (5.4) we have

$$\varepsilon = \alpha - \rho^2 \geq -(-\underline{\alpha} + \rho^2) \geq \delta > 0, \tag{5.7}$$

$$\begin{array}{l} e = a - \rho r \leq \overline{a} + \rho(-r), \quad -e = -a + \rho r \leq \rho r - \underline{a}, \\ |e| \leq \max(\overline{a} + \rho(-r), \rho r - \underline{a}) \leq d, \end{array} \tag{5.8}$$

and

$$A_0 = B - rr^T + ee^T/\varepsilon.$$

Since $|ee^T/\varepsilon| \leq dd^T/\delta$ by (5.7) and (5.8) we find that

$$A_0 \geq \underline{B} - rr^T - dd^T/\delta \geq \underline{A}_0, \quad A_0 \leq \overline{B} + (-r)r^T + dd^T/\delta \leq \overline{A}_0,$$

resulting in $A_0 \in \mathbf{A}_0$. □

In finite precision arithmetic, the right hand side of (5.6) must be evaluated by using upward rounding and the priorities given by the parentheses. Then (5.6) is satisfied, only the inequality $0 < \delta$ may be violated, but this case is handled by the termination condition (g) of Algorithm 5.5.

We now use the Proposition 5.7 to prove the following proposition, which is then used in the induction proof of Theorem 5.6.

PROPOSITION 5.8. *Suppose that for some real constants δ, ε and ρ, for some $(n-1)$-dimensional vectors d, r and e, for*

$$\mathbf{A} := \begin{pmatrix} \boldsymbol{\alpha} & \mathbf{a}^T \\ \mathbf{a} & \mathbf{B} \end{pmatrix} \in \mathbb{IR}^{n \times n}, \tag{5.9}$$

and for some symmetric interval matrix $\mathbf{A}_0 \in \mathbb{IR}^{(n-1)\times(n-1)}$ the inequalities

$$\begin{array}{l} |\rho| < \sqrt{\underline{\alpha}}, \\ 0 < \delta \leq -(-\underline{\alpha} + \rho^2), \\ d \geq \max(\overline{a} + \rho(-r), \rho r - \underline{a}), \\ \mathbf{A}_0 \supseteq [\underline{B} - rr^T - dd^T/\delta,\ \overline{B} + (-r)r^T + dd^T/\delta], \end{array} \tag{5.10}$$

are satisfied. If for all symmetric matrices $A_0 \in \mathbf{A}_0$ an $R_0 \in \mathbb{R}^{(n-1)\times(n-1)}$ exists such that $A_0 - R_0^T R_0$ is positive semidefinite, then for every symmetric matrix $A \in \mathbf{A}$ an $R \in \mathbb{R}^{n\times n}$ exists such that $A - R^T R$ is positive semidefinite.

Proof. By assumption, the Cholesky factorization

$$LL^T = A_0 - R_0^T R_0 \tag{5.11}$$

exists, with a lower triangular matrix $L \in \mathbb{R}^{n\times n}$.

Since by (5.9) every symmetric $A \in \mathbf{A}$ can be written as (5.3), the representation (5.5) holds by Proposition 5.7 for arbitrary $\rho \in \mathbb{R}$, $|\rho| < \sqrt{\underline{\alpha}}$, $r \in \mathbb{R}^{n-1}$ and

$$\varepsilon := \alpha - \rho^2, \quad e := a - \rho r, \quad A_0 := B - rr^T - \frac{ee^T}{\varepsilon}.$$

By (5.10) and the same proposition, the bounds $\varepsilon \geq \delta > 0$ and $A_0 \in \mathbf{A}_0$ are satisfied.

If we substitute (5.11) into (5.5), we get

$$\begin{aligned}
A &= \begin{pmatrix} \rho^2 + \varepsilon & \rho r^T + e^T \\ \rho r + e & rr^T + ee^T/\varepsilon + R_0^T R_0 + LL^T \end{pmatrix} = \\
&= \begin{pmatrix} \rho^2 & \rho r^T \\ \rho r & rr^T + R_0^T R_0 \end{pmatrix} + \begin{pmatrix} \varepsilon & e^T \\ e & ee^T/\varepsilon + LL^T \end{pmatrix} = \\
&= \begin{pmatrix} \rho & 0 \\ r & R_0 \end{pmatrix} \begin{pmatrix} \rho & r^T \\ 0 & R_0 \end{pmatrix} + \begin{pmatrix} \varepsilon & 0 \\ e & L \end{pmatrix} \begin{pmatrix} 1/\varepsilon & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \varepsilon & e^T \\ 0 & L^T \end{pmatrix} = \\
&= R^T R + S^T D S,
\end{aligned}$$

with

$$R = \begin{pmatrix} \rho & r^T \\ 0 & R_0 \end{pmatrix}, \quad S = \begin{pmatrix} \varepsilon & e^T \\ 0 & L^T \end{pmatrix}, \quad D = \begin{pmatrix} 1/\varepsilon & 0 \\ 0 & I \end{pmatrix}.$$

Since $\varepsilon \geq \delta > 0$ the matrix D is positive semidefinite and

$$x^T(A - R^T R)x = x^T(S^T D S)x = (Sx)^T D S x = (Sx)^T D S x \geq 0.$$

holds, proving the assertion. □

We are now prepared to prove that for all symmetric $A \in \mathbf{A}$ the residual matrix of the directed Cholesky factorization computed by Algorithm 5.5 is positive semidefinite:

Proof of Theorem 5.6: First we show by induction that the interval matrices $\mathbf{A}_k$, $k = 1, \ldots, n$ constructed by Algorithm 5.5 are symmetric. Without loss of generality we may assume that A_k is already permuted, such that no further pivoting is required ($P = I_n$).

$\mathbf{A}_1 = \mathbf{A}$ is symmetric by definition. Assuming that the interval matrix $\mathbf{A}_k$ is symmetric, $\mathbf{B}_k$ is also symmetric as an $(n-k) \times (n-k)$ submatrix of $\mathbf{A}_k$. For arbitrary vectors r and d, the matrices rr^T and dd^T are symmetric by construction. Therefore by (h) of Algorithm 5.5, the interval matrix $\mathbf{A}_{k+1} \in \mathbb{IR}^{(n-k)\times(n-k)}$ is symmetric. From this follows that each $\mathbf{A}_k$ can be factorized as in (5.9).

First we assume that every computation is exact and prove by induction on $m := n - k + 1$ that $A_k - R_k^T R_k$ is positive semidefinite for each symmetric $A_k \in \mathbf{A}_k$.

For $m = 1$, $k = n$, $R_n^T R_n = \rho_n^2 = \gamma_n^2 \underline{\alpha}_n$ and $A_n = \alpha_n \in \boldsymbol{\alpha}_n$ and since $\alpha_n \geq \underline{\alpha}_n > 0$ and $0 < \gamma_n < 1$,

$$A_n - R_n^T R_n = \alpha_n - \gamma_n^2 \underline{\alpha}_n \geq \underline{\alpha}_n (1 - \gamma_n^2) > 0$$

is positive semidefinite for all $A_n \in \mathbf{A}_n$.

We now assume for all symmetric matrices $A_{k+1} \in \mathbf{A}_{k+1}$ an $R_{k+1} \in \mathbb{R}^{n-k \times n-k}$ exists such that $A_{k+1} - R_{k+1}^T R_{k+1}$ is positive semidefinite. Since (d), (f) and (h) of Algorithm 5.5 imply (5.10) with $\delta = \delta_k$, $\varepsilon = \varepsilon_k$, $\rho = \rho_k$, $r = r_k$, $e = e_k$, $\mathbf{A} = \mathbf{A}_k$ and $\mathbf{A}_0 = \mathbf{A}_{k+1}$, we can find for every symmetric matrix $A_k \in \mathbf{A}_k$ an $R_k \in \mathbb{R}^{n-k+1 \times n-k+1}$ such that $A_k - R_k^T R_k$ is positive semidefinite.

By induction, this holds for $m = n$, $k = 1$ when $A = A_1$ and $R = R_1$ proving that $PAP^T - R^T R$ is positive semidefinite for $A \in \mathbf{A}$ $\square$.

In finite precision arithmetic, the results satisfy the required inequalities in Proposition 5.8 for d_k, δ_k and $\mathbf{A}_{k+1}$ if the right hand sides of the inequalities in (5.10) are computed with directed rounding.

By successfully factoring a symmetric positive semidefinite interval matrix $\mathbf{A}$ by Algorithm 5.5 we obtain a matrix R such that for all symmetric $A \in \mathbf{A}$ the residual matrix $S := PAP^T - R^T R$ is positive semidefinite. In addition to this we also expect (and our numerical experiments show that it is typically true) that S is very small with respect to A (for a suitable tolerance, e.g. $\kappa = 10^{-6}$). The choices of ρ_k, r_k and γ_k in Algorithm 5.5 were made to satisfy this criteria:

- To make S positive semidefinite, we had to ensure that $\varepsilon > 0$. Therefore we needed $\delta_k > 0$ which is the case when, $|\rho_k| < \sqrt{\underline{\alpha}_k}$. If we additionally want δ_k to be very small and assume that $\underline{\alpha}_k > 0$ (which is true if $\underline{A}$ is positive definite), we can set $\rho_k = \gamma_k \sqrt{\underline{\alpha}_k}$ with $\gamma_k < 1$. If in addition to this we choose $\gamma_k \approx 1$, the condition that $\delta_k \approx 0$ is also satisfied.
- The entries of $d_k = a_k - \rho_k r_k$ can be made to vanish by setting $r_k := a_k / \rho_k$. Even when r_k and ρ_k are computed inaccurately we can set $r_k = (\overline{a}_k + \underline{a}_k)/(2\rho_k)$ using the midpoint of the interval $\mathbf{a}_k$ to get a very small d_k.
- To make $d_k^T d_k / \delta_k$ very small, we also have to guarantee that $d_k^T d_k \ll \delta_k$. In the case of rounding errors $d_k^T d_k = (a_k - (\rho_k a_k)/\rho_k)^T (a_k - (\rho_k a_k)/\rho_k) = O(\epsilon)$ with ϵ representing the machine precision, so $1 \gg \delta_k = \alpha_k - \gamma_k^2 \alpha_k \gg \epsilon$ should be satisfied. Since $a_k \in \mathbf{a}_k$ the width of $\mathbf{a}_k$ should be accounted for, too. The heuristic choice for the regularization term γ_k is

$$\begin{aligned} \gamma_k &= 1 - \min((\sqrt{n-k}+1)\zeta_k, 0.01), \\ \zeta_k &= \epsilon + \max(\mathrm{mid}(\mathbf{a}_k)/w), \\ w_i &= \sqrt{\underline{\alpha}_k (\underline{A}_k)_{ii}}, \end{aligned}$$

  where n is the dimension of A. This choice will usually produce good results. *Edit: the heuristic of γ_k is an improved version of the one published one.*

Using these choices in Algorithm 5.5 makes the residual matrix not only positive semidefinite but also very small with respect to A for all $A \in \mathbf{A}$.

While the Cholesky factorization is numerically stable without pivoting, it is of advantage to use a permuted version. To enhance the robustness of the factorization we use *diagonal pivoting*. Thus in each step we permute two rows and the corresponding two columns of the matrix A_k in order to have the maximum of all diagonal entries as the pivot element α, while retaining the symmetric structure of the matrix. In our implementation, the diagonal pivoting can be turned off in order to reduce the time

needed for the factorization in the case of very big matrices. Testing has shown that when we turn off the pivoting, the factorization will fail more often, and the reduction of the computation time is not really significant unless the dimension is huge.

6. Testing the directed Cholesky factorization. We tested the new methods on random real interval matrices of different dimension (column `dim` in the tables below), width (column `width` in the tables below) and density (column `density` in the tables below). These matrices are constructed to be positive definite but nearly singular, with a very small inverse condition number (column `icond` in the tables below). For the inverse condition number we take the median of the quotients of the absolute value of the smallest eigenvalues and the absolute value of the largest ones of all k test matrices, where the eigenvalues are approximately computed by MATLAB, formally:

$$\texttt{icond} := \operatorname*{med}_i \Big(\frac{|\lambda_{min}(\underline{A}_i)|}{|\lambda_{max}(\underline{A}_i)|} \Big),\ i \in \{1, \ldots, k\}.$$

The following algorithm shows how the test matrices are created:

ALGORITHM 6.1 (Nearly singular positive definite interval matrix generator). *Given is the dimension n, a tiny singularity factor η (e.g. $\eta = 10^{-12}$) and the required relative width $\delta \geq 0$ of the interval matrix $\mathbf{A}$ to be created.*

1. *Generate a random matrix $B \in \mathbb{R}^{n-1 \times n}$ with $B_{ij} \in [-1, 1]$ for all $i = 1, \ldots, n-1$ and $j = 1, \ldots, n$.*
2. *Generate a random vector $u \in \mathbb{R}^n$ with $u_j \in [-1, 1]$ for all $j = 1, \ldots, n$.*
3. *Divide u by $\max(u)$ such that $\|u\|_2 = 1$ holds.*
4. *Compute $C = B^T B \in \mathbb{R}^{n \times n}$ and $d = max(C_{ii})$.*
5. *If $d = 0$ start again with step 1.*
6. *Else set $\underline{A} = C/d + \eta u u^T$ and $\overline{A} = \underline{A} + \delta|\underline{A}|$ and return the interval matrix $\mathbf{A} := [\underline{A}, \overline{A}] \in \mathbb{IR}^{n \times n}$.*

The tests show that both methods can be used to verify the positive definiteness and to decompose ill-conditioned matrices into their directed Cholesky factors. We

first show that the approximative method factors all the matrices and the directed methods factor nearly all of them. The comparison of the approximate Cholesky factorization of MATLAB (row CHOL in the tables below), the directed Cholesky factorization based on the Gerschgorin test (computed by Algorithm 5.2, row DIRCHOLG in the tables below) and the directed Cholesky factorization based with diagonal pivoting (computed by Algorithm 5.5, row DIRCHOLP in the tables below) on 500, 20–dimensional real matrices with a small inverse condition number:

method	dim	density	width	sfact	iters	icond	solved
CHOL	20	100%	0	1e-012	500	1.3e-016	100%
DIRCHOLH	20	100%	0	1e-012	500	1.4e-016	94%
DIRCHOLP	20	100%	0	1e-012	500	1.5e-016	88%

The next few tests of the directed Cholesky factorization based on the Gerschgorin test show increasing the dimension or the width makes the factorization more difficult, while more sparsity makes it easier. A test of the directed Cholesky factorization based using the Gerschgorin test on 500 real matrices of different dimensions (50,100,200) with a small inverse condition number:

method	dim	density	width	sfact	iters	icond	solved
DirCholH	10	100%	0	1e-012	500	1.7e-016	95%
DirCholH	40	100%	0	1e-012	500	1.3e-016	90%
DirCholH	100	100%	0	1e-012	500	1e-016	74%

The test of the directed Cholesky factorization using the Gerschgorin test on 500 real interval matrices of different dimensions (50,100,200), different average density and with a small inverse condition number:

method	dim	density	width	sfact	iters	icond	solved
DirCholH	10	46%	0	1e-012	500	1.3e-016	100%
DirCholH	40	39%	0	1e-012	500	5.6e-017	100%
DirCholH	100	38%	0	1e-012	500	3.7e-017	70%

A test of the directed Cholesky factorization using the Gerschgorin test on 500 real interval matrices of width $1e-014$ of different dimensions (50,100,200) with a small inverse condition number:

method	dim	density	width	sfact	iters	icond	solved
DirCholH	10	100%	1e-014	1e-012	500	1.7e-016	82%
DirCholH	40	100%	1e-014	1e-012	500	1.3e-016	67%
DirCholH	100	100%	1e-014	1e-012	500	9.1e-017	55%

The last five tests Cholesky factorization with diagonal pivoting show similar results with respect to the increasing dimension and more sparsity. We can also see that the results of this factorization method are not as good as the results of the directed Cholesky factorization using the Gerschgorin test. Since most applications are not as ill-conditioned as the problems in our test set and this method also returns an incomplete factorization it is still interesting. The fourth test is done in order to show the correlation between the dimension and the singularity factor, while the last one shows the effect if the pivoting is turned off. A test of the directed Cholesky factorization with diagonal pivoting on 500 real matrices of different dimensions (50,100,200) with a small inverse condition number:

method	dim	density	width	sfact	iters	icond	solved
DirCholP	10	100%	0	1e-012	500	1.8e-016	93%
DirCholP	40	100%	0	1e-012	500	1.3e-016	78%
DirCholP	100	100%	0	1e-012	500	1.1e-016	65%

A test of the directed Cholesky factorization with diagonal pivoting on 500 real interval matrices of different dimensions (50,100,200), different average density and with a small inverse condition number:

method	dim	density	width	sfact	iters	icond	solved
DirCholP	10	46%	0	1e-012	500	1.3e-016	100%
DirCholP	40	40%	0	1e-012	500	5.5e-017	93%
DirCholP	100	36%	0	1e-012	500	3.9e-017	81%

A test of the directed Cholesky factorization with diagonal pivoting on 500 real interval matrices of with $1e-014$ of different dimensions (50,100,200) with a small inverse condition number:

method	dim	density	width	sfact	iters	icond	solved
DirCholP	10	100%	1e-014	1e-012	500	1.7e-016	78%
DirCholP	40	100%	1e-014	1e-012	500	1e-016	65%
DirCholP	100	100%	1e-014	1e-012	500	1.1e-016	54%

A test of the correlation between the inverse condition number and the dimension for the directed Cholesky factorization with diagonal pivoting on 500 real matrices:

method	dim	density	width	sfact	iters	icond	solved
DirCholP	10	100%	0	1e-013	500	1.9e-017	79%
DirCholP	40	100%	0	1e-012	500	1.1e-016	76%
DirCholP	100	100%	0	3e-011	500	2.9e-015	92%

A test of the directed Cholesky factorization with diagonal pivoting on 500 real interval matrices of with $1e-014$ of different dimensions (50,100,200) with a small inverse condition number (pivoting turned off):

method	dim	density	width	sfact	iters	icond	solved
DirCholP(0)	10	100%	0	1e-012	500	1.7e-016	92%
DirCholP(0)	40	100%	0	1e-012	500	1.1e-016	77%
DirCholP(0)	100	100%	0	1e-012	500	1.1e-016	65%

7. Verification of positive definiteness. Proposition 5.1 shows that the existence of a directed Cholesky factorization of a symmetric matrix A implies that A is positive definite, and that of a symmetric interval matrix $\mathbf{A}$ implies that all symmetric matrices $A \in \mathbf{A}$ are positive definite. On the other hand, if the directed factorization fails, $\mathbf{A}$ either contains a singular or indefinite, or a very ill-conditioned matrix. Many of the latter cases can still be verified when we apply an appropriate scaling before verifying positive definiteness, see Section 3.

Such definiteness test are useful independent of the goal of this paper, for several applications ranging from the solution of linear interval equations (see below) over semidefinite programing problems Vandenbergh & Boyd [26] to the representation theory of Lie groups (Adams [1]).

Any test for the positive definiteness of real symmetric matrices can easily be extended to a test for complex Hermitian matrices, using the following result; no complex arithmetic is required.

Theorem 7.1. *A matrix $H = A + iB$ with $A, B \in \mathbb{R}^{n\times n}$ is Hermitian and positive definite iff the real matrix*

$$C := \begin{pmatrix} A & -B \\ B & A \end{pmatrix} \tag{7.1}$$

is symmetric and positive definite.

Proof. The matrix C is symmetric iff $A^T = A$ and $B^T = -B$, and this holds iff H is Hermitian. H is positive definite iff

$$(x+iy)^*(A+iB)(x+iy) > 0 \text{ whenever } \begin{pmatrix} x \\ y \end{pmatrix} \neq 0. \tag{7.2}$$

Now

$$\begin{aligned}(x+iy)^*(A+iB)(x+iy) &= (x-iy)^T(A+iB)(x+iy)\\ &= x^TAx + y^TAy + y^TBx - x^TBy + i(x^TAy - y^TAx + x^TBx + y^TBy)\\ &= x^TAx + y^TAy - 2x^TBy = \begin{pmatrix} x\\ y\end{pmatrix}^T C \begin{pmatrix} x\\ y\end{pmatrix}\end{aligned}$$

since

$$\begin{array}{lll} x^TAy = (x^TAy)^T = y^TA^Tx = y^TAx, & & \\ y^TBx = (y^TBx)^T = x^TB^Ty = -x^TBy, & & \\ x^TBx = (x^TBx)^T = x^TB^Tx = -x^TBx & \Rightarrow & x^TBx = 0,\\ y^TBy = (y^TBy)^T = y^TB^Ty = -y^TBy & \Rightarrow & y^TBy = 0.\end{array}$$

Thus (7.2) holds iff C is positive definite. □

We also note that all the results from Section 5 could be developed for the complex case.

Rump [16, 17, 18, 19] gave criteria for the definiteness of interval matrices in the context of solving linear interval equations. Here we discuss only his most recent work [19]. His method is based on a single floating-point Cholesky factorization; all possible computational and rounding errors, including underflow, are taken into account via a floating-point error analysis. To find an error estimation of the Cholesky factorization, Rump presents three different selection methods. These error estimations are worst case bounds; so when they are used to perturb the diagonal entries of the matrix A and the approximative Cholesky factorization is successful, the positive definiteness of A is guaranteed. Uncertainties in the matrix are accounted for only coarsely by bounding them in the Frobenius norm. Rump & Ogita [20] reduce the computational overhead in Rump's method, but only for exactly given floating-point matrices A.

In contrast, in our directed Cholesky factorization using the Gerschgorin test, the perturbation terms are based on heuristics that account for the typical case rather than a worst case floating-point analysis. To justify the heuristic choice, the actual verification is done by the additional Gerschgorin test. The directed Cholesky factorization with diagonal pivoting is based on different principles and is not directly comparable with Rump's approach. It is likely that the ideas of Rump can be combined with directed Cholesky factorizations to get improved enclosures for linear systems with positive definite interval coefficient matrices.

The following alternative test for positive definiteness of symmetric interval matrices is given in Neumaier [13, p. 32].

Theorem 7.2. *Let* $\mathbf{A}$ *be a symmetric interval matrix.*

(i) If some symmetric matrix $A \in \mathbf{A}$ *is positive definite and all symmetric matrices in* $\mathbf{A}$ *are nonsingular then they are all positive definite.*

(ii) In particular, this holds if the midpoint matrix

$$\hat{A} = (\overline{A} + \underline{A})/2$$

is positive definite with inverse C*, and the preconditioned radius matrix*

$$\Delta = |C|\,\mathrm{rad}(\mathbf{A});$$

satisfies (in an arbitrary norm) the condition

$$\|\Delta\| < 1.$$

Since verifying definiteness is not the focus of this paper we refrain from giving numerical comparison of the various definiteness tests.

Acknowledgment. This research was supported through the research grant FSP 506/003 of the University of Vienna. Numerous suggestions by the referees, which markedly improved the presentation of the paper, are gratefully acknowledged. We also thank Prof. Günter Mayer for his corrections.

REFERENCES

[1] J. Adams, *Computer computations in representation theory III: Unitary representation of real Lie groups*, tech. report, University of Maryland, 2002. Available from: `www.math.umd.edu/~jda/minicourse`.

[2] C. S. Adjiman, I. P. Androulakis, and C. A. Floudas, *A global optimization method, αBB, for general twice-differentiable constrained NLPs - II. implementation and computational results*, Computers and Chemical Engineering, 22 (1998), pp. 1159–1179.

[3] C. S. Adjiman, I. P. Androulakis, C. D. Maranas, and C. A. Floudas, *A global optimization method αBB for process design*, Computers and Chemical Engineering, 20 (1996), pp. 419–424.

[4] C. S. Adjiman, S. Dallwig, C. A. Floudas, and A. Neumaier, *A global optimization method, αBB, for general twice-differentiable constrained NLPs - I. theoretical advances,*, Computers and Chemical Engineering, 22 (1998), pp. 1137–1158.

[5] C. S. Adjiman and C. A. Floudas, *Rigorous convex underestimators for general twice-differentiable problems*, Journal of Global Optimization, 9 (1996), pp. 23–40.

[6] I. P. Androulakis, C. D. Maranas, and C. A. Floudas, *αBB: a global optimization method for general constrained nonconvex problems*, Journal of Global Optimization, 7 (1995), pp. 337–363.

[7] F. Domes, *GloptLab – a configurable framework for the rigorous global solution of quadratic constraint satisfaction problems*, Optimization Methods and Software, 24 (2009), pp. 727–747. Available from: `http://www.mat.univie.ac.at/~dferi/publ/Gloptlab.pdf`.

[8] F. Domes and A. Neumaier, *A scaling algorithm for polynomial constraint satisfaction problems*, Journal of Global Optimization, 43 (2008), pp. 327–345. Available from: `http://www.mat.univie.ac.at/~dferi/publ/Scaling.pdf`.

[9] N. J. Higham, *Accuracy and stability of numerical algorithms*, Siam, Philadelphia, 1996, ch. 10.

[10] R. B. Kearfott, *GlobSol user guide*, Optimization Methods and Software, 24 (2009), pp. 687–708.

[11] Y. Lebbah, *iCOs – Interval COnstraints Solver*, 2003. Available from: `http://ylebbah.googlepages.com/icos`.

[12] A. Neumaier, *Interval methods for systems of equations*, vol. 37 of Encyclopedia of Mathematics and its Applications, Cambridge Univ. Press, Cambridge, 1990.

[13] ———, *Complete search in continuous global optimization and constraint satisfaction*, Acta Numerica, 1004 (2004), pp. 271–369.

[14] ———, *Computer-assisted proofs*, in Proc. 12th GAMM-IMACS (SCAN 2006), IEEE Computer Society, 2007. Available from: `http://www.mat.univie.ac.at/~neum/ms/caps.pdf`.

[15] F. Rendl, G. Rinaldi, and A. Wiegele, *Solving Max-Cut to Optimality by Intersecting Semidefinite and Polyhedral Relaxations*, 2007. Available from: `http://biqmac.uni-klu.ac.at/rrw.pdf`.

[16] S. M. Rump, *Solving algebraic systems with high accuracy*, in A New Approach to Scientific Computation, Academic Press, 1993, pp. 51–120.

[17] ———, *Validated solution of large linear systems*, in Validation Numerics, Springer, 1993.

[18] ———, *Verification methods for dense and sparse systems of equations*, in Topics in validated computations, North Holland, 1994.

[19] ———, *Verification of positive definiteness*, BIT Numerical Mathematics, 46 (2006), pp. 433–452.

[20] S. M. Rump and T. Ogita, *Super-fast validated solution of linear systems*, Journal Comput. Appl. Math., 199 (2007).
[21] N. Sahinidis and M. Tawarmalani, *Convexification and global optimization in continuous and mixed–integer nonlinear programming: theory, algorithms, software, and applications*, Kluwer Academic Pub., 2003.
[22] N. V. Sahinidis and M. Tawarmalani, *BARON 7.2.5: global optimization of mixed-integer nonlinear programs,* User's Manual, 2005. Available from: `http://www.gams.com/dd/docs/solvers/baron.pdf`.
[23] H. Schichl, M. C. Markót, A. Neumaier, Xuan-Ha Vu, and C. Keil, *The COCONUT Environment*, 2000-2010. Software. Available from: `http://www.mat.univie.ac.at/coconut-environment`.
[24] M. Skutella, *Convex quadratic and semidefinite programming relaxations in scheduling*, Journal ACM, 48 (2001), pp. 206–242.
[25] J. Stoer and R. Bulirsch, *Introduction to numerical analysis*, Springer, 2002.
[26] L. Vandenbergh and S. Boyd, *Semidefinite programming*, SIAM Review, 38 (1996), pp. 49–95.