DIRECTED MODIFIED CHOLESKY FACTORIZATIONS AND CONVEX QUADRATIC RELAXATIONS

FERENC DOMES, ARNOLD NEUMAIER *

Abstract. A directed Cholesky factorization of a symmetric interval matrix $\mathbf{A}$ consists of a permuted upper triangular matrix R such that for all symmetric $A \in \mathbf{A}$, the residual matrix $A - R^T R$ is positive semidefinite with tiny entries. This must hold with full mathematical rigor, although the computations are done in floating-point arithmetic.

Similarly, a directed modified Cholesky factorization of a symmetric interval matrix $\mathbf{A}$ consists of a nonsingular permuted upper triangular matrix R and a non-negative diagonal matrix D such that for all $A \in \mathbf{A}$ the residual matrix $A + D - R^T R$ is positive semidefinite with tiny entries.

The paper shows how to construct a directed modified Cholesky factorization with the additional property that the entries of D are tiny, too, if $\mathbf{A}$ is nearly positive definite, and they are zero for numerically positive definite matrices. The construction is based on an incomplete version of the directed Cholesky factorization Domes & Neumaier (SIAM J. Matrix Anal. Appl. 32 (2011), 262–285), which performs better on nearly singular positive definite matrices. The incomplete method also allows one to select a set of columns which are eliminated before the other columns are processed. If the factorization fails, but the selected part was successfully processed an incomplete factorization is returned, needed for the new modified factorization. For the new factorization and relaxation methods detailed algorithms are given. Directed rounding or interval computations are used to make sure that the methods are rigorous in spite of the use of floating point arithmetic.

As application, new techniques are given for pruning boxes in the presence of an additional quadratic constraint, a problem relevant for branch and bound methods for global optimization and constraint satisfaction. Using either the incomplete directed Cholesky or the directed modified Cholesky factorization, a convex quadratic relaxation is created and an improved box enclosing the set of points in the original box satisfying the relaxed constraint. If the quadratic constraint is strictly convex and the box is sufficiently big, the relaxation and the enclosure are optimal up to rounding errors.

Numerical test show the usefulness of the new factorization methods in the context of pruning.

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

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

1. Introduction. A directed Cholesky factorization of a symmetric interval matrix $\mathbf{A}$ consists of a permuted upper triangular matrix R such that for all symmetric $A \in \mathbf{A}$, the residual matrix $A - R^T R$ is positive semidefinite with tiny entries. This must holds with full mathematical rigor, although the computations are done in floating-point arithmetic.

Domes & Neumaier [1] presented two methods for obtaining a directed Cholesky factorization. The method proved to be useful for other researchers (e.g., [4, 6, 7, 8]). In this paper we discuss two new methods: the incomplete directed Cholesky factorization and the directed modified Cholesky factorization. The development of these methods was motivated by our research in global optimization (Domes & Neumaier [2]) where we had to factorize the reduced Hessian for which theory requires, that a certain submatrix is positive definite, but the remainder of the matrix can be indefinite.

In general the incomplete directed Cholesky factorization (discussed in Section 4) performs better on nearly singular positive definite matrices as the methods from [1]. The method also allows us to select a set of columns which are eliminated before

*Faculty of Mathematics, University of Vienna. Oskar-Morgenstern-Platz 1, A-1090 Vienna, Austria

the other columns are processed. If the factorization fails, but the selected part was successfully processed the method returns an incomplete factorization which can be useful in many applications.

A directed modified Cholesky factorization of a symmetric interval matrix $\mathbf{A}$ consists of a nonsingular matrix R and a non-negative diagonal matrix D such that the residual matrix $A + D - R^T R$ is positive semidefinite and its entries are tiny for all $A \in \mathbf{A}$. In addition to this if $\mathbf{A}$ is nearly positive definite the entries of D are expected to be tiny and zero for numerically positive definite matrices. This method works for severely ill-conditioned and even for indefinite matrices. In addition to this like in the incomplete directed Cholesky factorization a certain set of columns can be selected which are eliminated before the other columns are processed. An algorithm for constructing a directed modified Cholesky factorization is presented in Section 5. It is a correctly rounding interval version of the (approximate) modified Cholesky factorization discussed, e.g., in Schnabel & Eskow [10, 11].

In the second part of this paper we discuss enhancements to the indefinite case of the `ehull` enclosure technique from Domes & Neumaier [1] for strictly convex quadratic constraints. Two new methods are presented in Section 3. One of them is based on the incomplete directed Cholesky factorization from Section 4 and the other one is based on the directed modified Cholesky factorization from Section 5. The new methods widen the application scope of the old `ehull` since they can compute convex quadratic relaxations for general quadratic constraints. They also perform better on problems where the quadratic coefficient matrix is positive definite but nearly singular.

For the new factorizations and relaxation methods detailed algorithms are given. Directed rounding or interval computations are used to make sure that the methods are rigorous in floating point arithmetic. Numerical test of the new directed Cholesky factorization methods are presented in Section 6.

The techniques presented give the possibility to obtain rigorous bounds on variables that are consequences of the constraints, without the need of giving explicit bounds on them. As already the original `ehull` enclosure, this makes the method a convenient step in branch and bound methods for solving constrained optimization problems (e.g., [3, 6, 7, 9]).

Acknowledgments This research was supported by the Austrian Science Fund (FWF) under the contract numbers P23554-N13 and P22239-N13.

2. Notation.

2.1. Matrices. $\mathbb{R}^{m\times n}$ denotes the vector space of all $m \times n$ matrices A with real entries A_{ik} $(i = 1,\dots,m,\ k = 1,\dots,n)$, and $\mathbb{R}^n = \mathbb{R}^{n\times 1}$ denotes the vector space of all column vectors of length n. For vectors and matrices, the relations $=$, $\neq$, $<$, $>$, $\leq$, $\geq$ and the absolute value $|A|$ of a matrix A are interpreted component-wise.

The n-dimensional identity matrix is denoted by I and the n-dimensional zero matrix is denoted by $\mathit{0}$. The transpose of a matrix A is denoted by A^T, and A^{-T} is short for $(A^T)^{-1}$. The ith row vector of a matrix A is denoted by $A_{i:}$ and the jth column vector by $A_{:j}$. For an $n \times n$ matrix A, $\operatorname{diag}(A)$ denotes the n-dimensional vector with $\operatorname{diag}(A)_i = A_{ii}$.

The number of elements of an index set N is denoted by $|N|$. The set $\neg N$ denotes the complement of N. Let $I \subseteq \{1,\dots,m\}$ and $J \subseteq \{1,\dots,n\}$ be index sets and let $n_I := |I|$, $n_J := |J|$. For an n-dimensional vector x, x_J denotes the n_J-dimensional vector built from the components of x selected by the index set J. For an $m \times n$ matrix A, the expression $A_{I:}$ denotes the $n_I \times n$ matrix built from the rows of A

selected by the index sets I. Similarly, $A_{:J}$ denotes the $m \times n_J$ matrix built from the columns of A selected by the index sets J. Instead of using the index sets I and J we also write $A_{i:k,j:l}$ if $I = \{i, i+1, \dots, k\}$ and $J = \{j, j+1, \dots, l\}$.

2.2. Boxes. A **box** $\mathbf{x} = [\underline{x}, \overline{x}]$, i.e., 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 $\mathbb{R}^n$. $\overline{\mathbb{IR}}^n$ denotes the set of all n-dimensional boxes. To take care of one-sided bounds on variables, the values $-\infty$ and ∞ are allowed as lower and upper bounds of a box, respectively. The condition $x \in \mathbf{x}$ is equivalent to the collection of simple bounds

$$\underline{x}_i \le x_i \le \overline{x}_i \quad (i = 1, \dots, n),$$

or, with inequalities on vectors and matrices interpreted component-wise, to the two-sided vector inequality $\underline{x} \le x \le \overline{x}$. Apart from two-sided constraints, this includes with $\mathbf{x}_i = [a, a]$ variables x_i fixed at a particular value $x_i = a$, with $\mathbf{x}_i = [a, \infty]$ lower bounds $x_i \ge a$, with $\mathbf{x}_i = [-\infty, a]$ upper bounds $x_i \le a$, and with $\mathbf{x}_i = [-\infty, \infty]$ free variables. For the notation in interval analysis we mostly follow [5].

2.3. Uncertain vectors and matrices. To rigorously account for inaccuracies in computed entries of a matrix, we use interval matrices, standing for uncertain real matrices whose coefficients are between given lower and upper bounds. Note that all boxes may be considered as interval vectors, i.,e., column vectors ($n \times 1$ matrices) with uncertain components, whose values are known only to lie withing given intervals. The midpoint, width and the radius of an interval matrix $\mathbf{A}$ are the noninterval matrices defined by

$$\mathrm{mid}(\mathbf{A}) := (\overline{A} + \underline{A})/2, \quad \mathrm{wid}(\mathbf{A}) := \overline{A} - \underline{A}, \quad \mathrm{rad}(\mathbf{A}) := \mathrm{wid}(\mathbf{A})/2,$$

respectively. An interval, interval vector, or interval matrix is called **thin** or **degenerate** if its width is zero, and **thick** if its width is positive. A real matrix A is identified with the thin interval matrix with $\underline{A} = \overline{A} = A$.

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 $i, k \in \{1, \dots, n\}$. The comparison matrix $\langle \mathbf{A} \rangle$ of a square interval matrix $\mathbf{A}$ is defined by

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

3. Convex quadratic relaxations. Domes & Neumaier [1, Section 2] describe a method for computing nearly optimal, rigorous enclosure of a strictly convex quadratic constraint. The method makes use of a directed Cholesky factorization. In this section we present two improved versions of the original method: one of them is based on the incomplete directed Cholesky factorization from Section 4 and the other one is based on the directed modified Cholesky factorization from Section 5. The new methods widen the application scope of the old one. They also perform better for problems where the quadratic coefficient matrix is positive definite but nearly singular.

For a symmetric interval matrix $\mathbf{A} \in \mathbb{IR}^{n \times n}$, an interval vector $\mathbf{a} \in \mathbb{IR}^n$, a constant α and a box $\mathbf{x} \in \overline{\mathbb{IR}}^n$ we define the uncertain quadratic constraint

$$x^T A x + 2a^T x \le \alpha, \quad x \in \mathbf{x}, \quad A \in \mathbf{A}, \quad a \in \mathbf{a}. \tag{3.1}$$

For this constraint we compute a strictly convex quadratic relaxation (which is nearly optimal in case $\mathbf{A}$ is positive definite) as well as a box $\mathbf{x}' \subseteq \mathbf{x}$ such that each x satisfying (3.1) is contained in the box $\mathbf{x}'$ (hence the method is rigorous). We first assume that $\mathbf{A}$ does not contain zero rows. This means that each variables occurs non-linearly in (3.1); the case where some variables enter the constraint only linearly will be discussed separately in Subsection 3.4.

For later use we also define the index sets of the bounded and unbounded variables by

$$N := \{i \mid \mathbf{x}_i \text{ is bounded}\}, \quad M := \neg N, \tag{3.2}$$

and denote the subspace of $\mathbb{R}^n$ defined by the directions x_i, $i \in M$ by $\mathbb{R}^M$.

3.1. Enclosing the solution set of norm constraints. In this subsection we summarize the main result of [1, Section 1] applied to the norm constraint

$$\|Rx\|_2^2 + 2a^T x \leq \widehat{\alpha}. \tag{3.3}$$

Let C be the inverse of the matrix R, $d \in \mathbb{R}^n$ with $d_i = \inf(\sqrt{(CC^T)_{ii}})$, $h = \langle CR \rangle d$, $\beta = \max\{h_i/d_i \mid i = 1 \dots n\} \approx 1$, $\widetilde{z} = C^T a$ and $\widetilde{x} = -C\widetilde{z}$. Denote the enclosure of the expression

$$\|\widetilde{z} + R\widetilde{x}\|_2 + \beta^{-1} d^T |a - R^T \widetilde{z}| \tag{3.4}$$

by $[\underline{\gamma}, \overline{\gamma}]$, and denote the enclosure of the expression

$$\gamma^2 + \widehat{\alpha} - 2a^T \widetilde{x} - \|R\widetilde{x}\|_2^2, \tag{3.5}$$

by $[\underline{\Delta}, \overline{\Delta}]$. If $\Delta \geq 0$ then by [1, Theorem 1.5 and Corollary 1.6], (3.3) implies that

$$\|R(x - \widetilde{x})\|_2 \leq \overline{\delta}, \tag{3.6}$$

must be satisfied with

$$[\underline{\delta}, \overline{\delta}] := [\underline{\gamma} + \sqrt{\underline{\Delta}}, \overline{\gamma} + \sqrt{\overline{\Delta}}]. \tag{3.7}$$

Therefore the ellipsoid defined by (3.6) is an enclosure of (3.3). By the same theorem we also obtain the box

$$\widehat{\mathbf{x}} := \Big[(\delta/\overline{\beta})d - \widetilde{x}, (\delta/\underline{\beta})d + \widetilde{x}\Big], \tag{3.8}$$

enclosing the solution set of (3.3).

In floating point arithmetic we compute $\widetilde{z} \approx R^{-T} a$ and $\widetilde{x} \approx -R^{-1}\widetilde{z}$ by floating point calculations, and the remaining variables optimally, by computing the corresponding expressions with directed rounding or interval arithmetic.

3.2. Convex quadratic relaxation by incomplete Cholesky factorization. The first method is very similar to the one described in DOMES & NEUMAIER [1, Section 2], but instead of using the directed Cholesky factorization on $\mathbf{A}$ we use the incomplete directed Cholesky factorization (Algorithm 3) on $\mathbf{A}$ and the index set M. In the original method, no enclosure was computed if the factorization failed, the constraint was not strictly convex. Now if the incomplete directed Cholesky factorization fails but $\mathbf{A}_{MM}$ was successfully factored, the results of the incomplete factorization, namely the permutation matrix P and the matrix R^m are obtained. If we put $R := R^m P_{MM}$ by Section 4 we know that the residual matrix

$$\Gamma := A_{MM} - RR, \tag{3.9}$$

is positive semidefinite and tiny in respect to the entries of $\mathbf{A}_{MM}$.

PROPOSITION 3.1. *For $x \in \mathbf{x} \in \overline{\mathbb{IR}}^n$ consider an uncertain quadratic constraint as in* (3.1)*. Let $N, M \in \{1, \ldots, n\}$, be defined as in* (3.2) *and let $R := R^m P_{MM} \in \mathbb{R}^{|M| \times |M|}$ be the result of the incomplete directed Cholesky factorization given in Algorithm 3 applied to $\mathbf{A}$ and M. Let for the norm inequality*

$$\|Rx_M\|_2^2 + 2a_M^T x_M \leq \widehat{\alpha},\ x_M \in \mathbf{x}_M,\ A_{MM} \in \mathbf{A}_{MM},\ a_M \in \mathbf{a}_M, \tag{3.10}$$

with

$$\widehat{\alpha} := \alpha - \inf(\mathbf{x}_N^T \mathbf{A}_{NN} \mathbf{x}_N + 2\mathbf{a}_N^T \mathbf{x}_N). \tag{3.11}$$

the box $\widehat{\mathbf{x}}$ be computed by (3.8)*, then all $x \in \mathbf{x}$ satisfying* (3.1) *are also contained in the box*

$$\mathbf{x}' \text{ with } \mathbf{x}'_M := \widehat{\mathbf{x}},\ \mathbf{x}'_N := \mathbf{x}_N. \tag{3.12}$$

Proof. Since $R := R^m P_{MM}$ is the result of the incomplete directed Cholesky factorization, the residual matrix given in (3.9) is positive semidefinite. Therefore for all $x \in \mathbf{x}$, $A \in \mathbf{A}$ and $a \in \mathbf{a}$ we obtain (3.10) and (3.11) by

$$\begin{aligned}
&x^T Ax + 2a^T x \leq \alpha \Rightarrow \\
&\|Rx_M\|_2^2 + x_N^T A_{NN} x_N + 2a_M^T x_M + 2a_N^T x_N \leq x^T Ax + 2a^T x \leq \alpha \Rightarrow \\
&\|Rx_M\|_2^2 + 2a_M^T x_M \leq \widehat{\alpha} := \alpha - \inf(\mathbf{x}_N^T \mathbf{A}_{NN} \mathbf{x}_N + 2\mathbf{a}_N^T \mathbf{x}_N).
\end{aligned}$$

Since by the definition (3.2) of N the bound $\widehat{\alpha}$ is finite, the results of Subsection 3.1 can be applied to the m-dimensional norm constraint (3.10) to obtain a convex quadratic relaxation (3.6) of (3.1) in the subspace $\mathbb{R}^M$. In addition to this since the box $\widehat{\mathbf{x}}$ from (3.8) encloses the solution set of (3.10) and (3.10) is a relaxation of (3.1) (in the subspace $\mathbb{R}^M$), all $x \in \mathbf{x}$ satisfying (3.1) are contained in the box $\mathbf{x}'$. □

Proposition 3.1 forms the basis of the rigorous method given by Algorithm 1.

3.3. Relaxation by directed modified Cholesky factorization. The second method uses the directed modified Cholesky factorization (presented in Section 5) applied to $\mathbf{A}$, M and $\zeta = 0$. If the factorization is successful, a nonsingular matrix R and a diagonal matrix $D \geq 0$ is obtained such that

$$A \leq R^T R - D + \Gamma,\quad \forall A \in \mathbf{A} \text{ and } D_{MM} = 0, \tag{3.13}$$

and the residual matrix Γ is positive semidefinite and very small with respect to $\mathbf{A} - D$ (details in Section 5).

PROPOSITION 3.2. *For $x \in \mathbf{x} \in \overline{\mathbb{IR}}^n$ consider an uncertain quadratic constraint as in* (3.1)*. Let $M \in \{1, \ldots, n\}$ be defined as in* (3.2) *and let R and D be the result of the directed modified Cholesky factorization given in Algorithm 4 applied to $\mathbf{A}$ and M. Let for the norm inequality*

$$\|Rx\|_2^2 + 2a^T x \leq \widehat{\alpha}, x \in \mathbf{x},\ A \in \mathbf{A},\ a \in \mathbf{a} \tag{3.14}$$

Algorithm 1: Convex quadratic relaxation using incomplete directed Cholesky factorization (QRelIDChol)

Input: The constraint $x^TAx + 2a^Tx \le \alpha$, $x \in \mathbf{x}$ with $A \in \mathbf{A}$ and $a \in \mathbf{a}$.
Output: A convex quadratic relaxation and the rigorous box enclosure $x \in \mathbf{x}'$.

Compute M by (3.2) and use the incomplete directed Cholesky factorization (Algorithm 3) on $\mathbf{A}$, M;
if *the factorization failed but* $R^m \neq \emptyset$ **then**
    We have obtained P and R^m; put $R \leftarrow R^m P_{MM}$, $\mathbf{A} \leftarrow \mathbf{A}_{MM}$ and compute $\widehat{\alpha} \leftarrow \alpha'$ by (3.11), using interval arithmetic;
else if *the factorization was successful* **then**
    We have obtained P and R; put $R \leftarrow RP$ and $\widehat{\alpha} \leftarrow \alpha$;
else return *signaling failure*;
Compute the approximative inverse C of the matrix R;
Compute d with $d_i = \inf(\sqrt{(CC^T)_{ii}})$ by using directed rounding;
Use upward rounding to compute the values $h = \langle CR \rangle d$ and $\beta = \max\{h_i/d_i \mid i = 1 \dots n\} \approx 1$;
Set $\widetilde{z} = C^Ta$ and $\widetilde{x} = -C\widetilde{z}$ and compute an enclosure $[\underline{\gamma}, \overline{\gamma}]$ for (3.4), an enclosure $[\underline{\Delta}, \overline{\Delta}]$ for (3.5) using interval arithmetic;
if $\underline{\Delta} < 0$ **then return** *signaling failure*;
else
    Compute the interval $[\underline{\delta}, \overline{\delta}]$ from (3.6) by using outward rounding;
    return *The convex quadratic relaxation, given by the norm constraint* (3.7) *and the rigorous box enclosure* (3.12)
end

with

$$\widehat{\alpha} := \alpha + \sup \sum_{i \in \neg M} D_{ii} \mathbf{x}_i^2. \tag{3.15}$$

the box $\widehat{\mathbf{x}}$ *be computed by* (3.8)*, then all* $x \in \mathbf{x}$ *satisfying* (3.1) *are also contained in the box*

$$\mathbf{x}' := \mathbf{x} \cap \widehat{\mathbf{x}}. \tag{3.16}$$

Proof. Since R and D the result of the directed modified Cholesky factorization given in Algorithm 4 applied to $\mathbf{A}$ and M, they must satisfy (3.13). Substituting (3.13) into (3.1) results in

$$\begin{array}{rcl} x^TAx + 2a^Tx & \le & x^T(R^TR - D + \Gamma)x + 2a^Tx \\ & = & \|Rx\|_2^2 - x^TDx + x^T\Gamma x + 2a^Tx \le \alpha, \end{array}$$

and using that $D \ge 0$ and the residual matrix Γ is positive semidefinite, we end up in

$$\|Rx\|_2^2 + 2a^Tx \le \alpha + x^TDx - x^T\Gamma x \le \alpha + x^TDx \le \widehat{\alpha}, \tag{3.17}$$

with

$$\widehat{\alpha} := \alpha + \sup_{x \in \mathbf{x}} x^TDx = \alpha + \sup \sum_{i \in \neg M} D_{ii} \mathbf{x}_i^2, \tag{3.18}$$

for all $x \in \mathbf{x}$, $A \in \mathbf{A}$ and $a \in \mathbf{a}$, which is exactly (3.3). This proves that the solution set of (3.1) is fully contained in the ellipsoid given by the norm constraint (3.3). If we apply Subsection 3.1 to the norm constraint (3.14) we obtain the convex quadratic relaxation (3.6) of (3.1) as well as the box $\widehat{\mathbf{x}}$. Therefore for all $x \in \mathbf{x}$ satisfying (3.1) $x \in \mathbf{x}' := \mathbf{x} \cap \widehat{\mathbf{x}}$. □

Note that if $D = 0$, A is positive definite, $\widehat{\alpha} = \alpha$ does not depend of the box $\mathbf{x}$ and since Γ is very small with respect to A, the relative approximation error

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

is also small. Therefore if (3.1) is strictly convex, (3.15) is a nearly optimal approximation. On the other hand if $D \neq 0$, by (3.2) the bound (3.15) has to be finite; so (3.3) is a nontrivial inequality.

Proposition 3.2 forms the basis of the rigorous method described by Algorithm 2.

Algorithm 2: Convex quadratic relaxation using directed modified Cholesky factorization (QRelMDChol)

Input: The constraint $x^T Ax + 2a^T x \le \alpha$, $x \in \mathbf{x}$ with $A \in \mathbf{A}$ and $a \in \mathbf{a}$.
Output: A convex quadratic relaxation and the rigorous box enclosure $x \in \mathbf{x}'$.

Compute M by (3.2) and use the directed modified Cholesky factorization (Algorithm 4) on $\mathbf{A}$, M and $\zeta = 0$ to obtain D and R;
if *the factorization failed* **then return** *signaling failure*;
else
  Compute $\widehat{\alpha}$ by (3.15), using interval arithmetic;
  Compute the approximative inverse C of the matrix R;
  Compute d with $d_i = \inf(\sqrt{(CC^T)_{ii}})$ by using directed rounding;
  Use upward rounding to compute the values $h = \langle CR \rangle d$ and $\beta = \max\{h_i/d_i \mid i = 1 \dots n\} \approx 1$;
  Set $\widetilde{z} = C^T a$ and $\widetilde{x} = -C\widetilde{z}$ and compute an enclosure $[\underline{\gamma}, \overline{\gamma}]$ for (3.4), an enclosure $[\underline{\Delta}, \overline{\Delta}]$ for (3.5) using interval arithmetic;
  if $\underline{\Delta} < 0$ **then return** *signaling failure*;
  else
    Compute the interval $[\underline{\delta}, \overline{\delta}]$ from (3.6) by using outward rounding;
    return *The convex quadratic relaxation, given by the norm constraint* (3.7) *and the rigorous box enclosure* (3.16)
  end
end

3.4. Removing purely linear terms. If some variables occur only linearly in (3.1) the corresponding rows and columns of $\mathbf{A}$ have only zero entries, therefore $\mathbf{A}$ is singular, and depending on the method computing the relaxation becomes impossible or at least inefficient. Therefore we define the index sets

$$J := \{j \mid \mathbf{A}_j k = 0,\ \forall k = 1, \dots, n\}, \quad K := \neg J,$$

and eliminate the variables x_J from (3.1) obtaining

$$x_K^T A_{KK} x_K + 2a_K^T x_K \le \alpha',\ x_K \in \mathbf{x}_K,\ A_{KK} \in \mathbf{A}_{KK},\ a_K \in \mathbf{a}_K, \tag{3.19}$$

where

$$\alpha' := \alpha - \inf(2\mathbf{a}_J^T \mathbf{x}_J). \tag{3.20}$$

If α' is finite we can apply the methods discussed in the previous subsections without modifications to the lower dimensional system (3.19) instead of (3.1) to obtain a convex quadratic relaxation of (3.1) in the subspace spanned by x_K as well as the box enclosure

$$x \in \mathbf{x}', \quad \mathbf{x}'_K := \mathbf{x}_K \cap \Big[(\delta/\overline{\beta})d - \widetilde{x}, (\delta/\underline{\beta})d + \widetilde{x}\Big]_i, \quad \mathbf{x}'_J := \mathbf{x}_J. \tag{3.21}$$

3.5. Diagonal test for indefiniteness. Since the new methods require that at least $\mathbf{A}_{MM}$ is positive definite, it is useful to perform a simple diagonal test for positive definiteness and immediately signaling failure if $\mathbf{A}_{ii} < 0$ for any $i \in M$. This saves computation time and should be done every time before a factorization of $\mathbf{A}$ is computed.

4. Incomplete directed Cholesky factorization. A directed Cholesky factorization of a symmetric interval matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ constructs an upper triangular matrix $R' \in \mathbb{R}^{n\times n}$ and a permutation matrix P such that the residual matrix $\Gamma := A - P^T R'^T R' P$ is positive semidefinite for all $A \in \mathbf{A}$, and all Γ_{ij} are tiny. If we combine the matrices P and R to $R := R'P$ (which is in general no longer upper triangular), we obtain $\Gamma = A - R^T R$.

If the directed Cholesky factorization fails since A is not (numerically) positive definite, we want the resulting partial factorization to satisfy at least some of the properties of the full factorization. We achieve this by appropriately modifying Algorithm `DirCholP` by Domes & Neumaier [1, Algorithm 5.5], to either compute 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 terminates with an error message and returns an incomplete factorization. In addition to this we also want to make sure that specific diagonal elements are selected first as pivots. For this reason allow the input of an index set M and ensure for the first $m := |M|$ steps only pivots α_k with $k \in M$ are chosen.

Theorem 5.6 of Domes & Neumaier [1] (which details the properties of the factorization resulting from the old algorithm) still holds for the resulting Algorithm 3, with trivial modifications. In case the factorization failed after k steps, the incomplete factorization computed by the new algorithm still has the property that $\Gamma^k := (PAP^T)_{KK} - R_{KK}^T R_{KK}$ is positive semidefinite for $K := \{1, \dots, k\}$. This is particularly useful in case both $M \neq \emptyset$, $R^m \neq \emptyset$, and we need the fact that

$$\Gamma^m := (PAP^T)_{MM} - R^{mT} R^m \tag{4.1}$$

is positive semidefinite.

The method given in Domes & Neumaier [1] for selecting the parameter γ_k in line 8 of Algorithm 3 proved to have some limitations when a submatrix is nearly singular. We therefore present an improved method for selecting the parameter γ_k in Algorithm 3.

The improvement is based on the expectation (checked in numerical experiments to be typically valid for a suitable tolerance, e.g., $\kappa = 10^{-6}$) that each Γ^k is very small with respect to A. Therefore we choose ρ_k, r_k and γ_k in Algorithm 3 as follows:

Algorithm 3: Incomplete directed Cholesky factorization (IDirChol)

Input: A symmetric interval matrix $\mathbf{A} \in \mathbb{IR}^{n\times n}$ and the index set M (which can be empty) with $M \subseteq \{1, \ldots, n\}$ and $|M| = m$.

Success: The upper triangular matrix $R \in \mathbb{R}^{n\times n}$ and a permutation matrix $P \in \mathbb{R}^{n\times n}$ such that $PAP^T - R^T R$ is positive semidefinite for all symmetric $A \in \mathbf{A}$.

Incomplete: In case $M \neq \emptyset$, the complete factorization failed but the first m steps were successful, return the matrix $\mathbf{A}^m \in \mathbb{IR}^{n-m\times n-m}$, P and $R^m \in \mathbb{R}^{m\times m}$

if $\underline{A}_{ii} < 0$ *for any* $i \in M$ **then return** *signaling failure*;

Put $\mathbf{A}_1 = \mathbf{A}$, $N = M$, $\mathbf{A}^m = \emptyset$, $R = 0_n$, $R^m = \emptyset$, $P = I_n$ and change to upward rounding mode;

for $k = 1, \ldots, n$ **do**

  Find the pivot element

$$\alpha = \max(\operatorname{diag}(\widehat{A})), \quad \widehat{A} := \begin{cases} \underline{A}_{NN} & \text{if } N \neq \emptyset, \\ \underline{A}_k \in \mathbb{R}^{n-k+1} & \text{otherwise.} \end{cases}$$

  Let j denote the index of the pivot element in $\mathbf{A}_k$; exchange row j with the first row and column j with the first column of $\mathbf{A}_k$. Exchange the same rows and columns in the matrix P;

  If the pivot was selected from N remove its index from N;

  Partition the permuted interval matrix $\mathbf{A}_k$ as:

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

  if $\underline{\alpha}_k \leq 0$ **then return** $\mathbf{A}^m$*,* P*,* R^m *and an error message*;

  else

    Choose $0 < \gamma_k < 1$, $\rho_k = \gamma_k\sqrt{\underline{\alpha}_k}$ and $r_k = (\overline{a}_k + \underline{a}_k)/(2\rho_k)$;

    Set $R_{kk} = \rho_k$, $R_{k,k:n} = r_k^T$ and compute $\delta_k := -(-\underline{\alpha}_k + \rho_k^2)$, $d_k := \max(\overline{a}_k + \rho_k(-r_k), \rho_k r_k - \underline{a}_k)$;

    if *the residual pivot* $\delta_k \leq 0$ **then**

      return $\mathbf{A}^m$*,* P*,* R^m *and an error message*;

    else Set $\mathbf{A}_{k+1} := [\underline{B}_k - r_k r_k^T - d_k d_k^T/\delta_k,\ \overline{B}_k + (-r_k)r_k^T + d_k d_k^T/\delta_k]$;

  end

  if $M \neq \emptyset$ *and* $k = m$ **then** put $\mathbf{A}^m = \mathbf{A}_k$ and $R^m := R_{MM}$;

end

return *The matrix* R *and the permutation matrix* P

- To make Γ positive semidefinite, we have to ensure that $\varepsilon > 0$. Therefore we need $\delta_k > 0$, which is the case when, $|\rho_k| < \sqrt{\underline{\alpha}_k}$. If we also 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 $\delta_k \approx 0$ 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 get a very small d_k by setting $r_k = \widetilde{a}_k/(2\rho_k)$ where $\widetilde{a}_k := \overline{a}_k + \underline{a}_k$.
- To make $d_k^T d_k/\delta_k$ very small, we also have to guarantee that $d_k^T d_k \ll \delta_k$. Due

to rounding errors, $d_k^T d_k$ is of order

$$\widetilde{d}_k := |\overline{a}_k - \underline{a}_k| + \varepsilon|\widetilde{a}_k|,$$

where ϵ is the machine precision, so we want $1 \gg \delta_k = \alpha_k - \gamma_k^2 \alpha_k \gg \widetilde{a}_k$.

In order to achieve these goals in each step we need choose a suitable γ_k such that the diagonal elements of $\underline{A}_k$ are likely to remain positive. Writing $\mu_k := \gamma_k^{-2}$, we must ensure that the diagonal of

$$\underline{A}_k = \underline{B}_k - r_k r_k^T - d_k d_k^t/\delta_k \approx \underline{B}_k - \frac{\mu_k}{4\underline{\alpha}_k}\Big(\widetilde{a}_k \widetilde{a}_k^T + \frac{\widetilde{d}_k \widetilde{d}_k^T}{\mu_k - 1}\Big)$$

is not so small. If $\widetilde{a}_k^T \widetilde{a}_k = 0$ then $\widetilde{a}_k = \widetilde{d}_k = 0$ and we may choose $\gamma_k = 1$. Otherwise we note that the trace is maximal for

$$\mu_k = 1 + \sqrt{\frac{\widetilde{d}_k^T \widetilde{d}_k}{\widetilde{a}_k \widetilde{a}_k^T}}.$$

Thus we might choose $\gamma_k = 1/\sqrt{\mu_k}$; but in order to avoid that γ_k gets small, we safeguard it by

$$\gamma_k := \begin{cases} 1/\min(2, \sqrt{\mu_k}) & \text{if } \widetilde{a}_k^T \widetilde{a}_k \neq 0, \\ 1 & \text{otherwise.} \end{cases}$$

Using these choices in Algorithm 3 makes the residual matrix not only positive semidefinite but also very small with respect to A for all $A \in \mathbf{A}$.

5. Directed modified Cholesky factorization. We now use the directed Cholesky factorization discussed in Section 4 to define a modified directed Cholesky factorization that also works for indefinite matrices.

The directed modified Cholesky factorization of a symmetric interval matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, consists of a nonsingular matrix $R \in \mathbb{R}^{n\times n}$ and a non-negative diagonal matrix $D \in \mathbb{R}^{n\times n}$ such that the residual matrix

$$\Gamma := A + D - R^T R \tag{5.1}$$

is positive semidefinite for all $A \in \mathbf{A}$. We compute this factorization by trying to form the directed Cholesky factorization of $A + D$ for different diagonal matrices $D \geq 0$ until we succeed, starting with $D = 0$ and using the partial results of a failed directed Cholesky factorization to select an improved D.

In certain applications (see, e.g., Domes & Neumaier [2]) it is useful to have a preferred index set M where theory predicts that, in all nondegenerate cases, A_{MM} is positive definite. In this case, we want to ensure if possible that

$$D_{ii} = 0 \quad \text{for all} \quad i \in M.$$

Therefore we put $m := |M|$, and write $\mathbf{A}^m \in \mathbb{IR}^{(n-m)\times(n-m)}$ for the interval matrix to be factored after the mth pivoting step. If k pivot steps were successfully performed in a failed directed Cholesky factorization, we define the matrix

$$A' = \begin{cases} \underline{A} & \text{if } k < m, \\ \underline{A}^m & \text{otherwise,} \end{cases} \tag{5.2}$$

Algorithm 4: Modified directed Cholesky factorization (ModDirChol)

Input: A symmetric interval matrix $\mathbf{A} \in \mathbb{IR}^{n\times n}$, the index set $M \subseteq \{1, \dots n\}$ indicating that $\mathbf{A}_{MM}$ is required to be positive definite and the corresponding positive definiteness violation tolerance $\zeta \ll 1$ (e.g, $\zeta = 10^{-6}$).

Output: A nonsingular matrix $R \in \mathbb{R}^{n\times n}$ and a diagonal matrix $D \in \mathbb{R}^{n\times n}$ with $D \geq 0$ and $D_{MM} = 0$ such that Γ defined by (5.1) is positive semidefinite for all $A \in \mathbf{A}$.

Use the incomplete directed Cholesky factorization (Algorithm 3) with $\mathbf{A}$ and M to obtain either $\widehat{R}$ and P or $\mathbf{A}^m$;
if *Algorithm 3 was successful* **then return** $R := \widehat{R}P$ and $D = 0$;
else
    Put $m = |M|$ and denote the number of successfully performed pivot steps by k;
    Find A' and J as given by (5.2) and (5.3);
    Compute the minimum and the maximum eigenvalues $\underline{\lambda}$, $\overline{\lambda}$ of the matrix A' and compute $\gamma := 1 + |\overline{\lambda}| + |\underline{\lambda}|$;
    for $\epsilon = 10^{-12}, 10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1$ **do**
        if $\epsilon > \zeta$ *and* $k < m$ **then**
            the claim that A_{MM} is positive definite is significantly violated therefore **return** signaling failure;
        else
            Compute $\sigma = \epsilon\gamma + \max(-\underline{\lambda}, 0)$ and put $D := \sigma J \in \mathbb{R}^{n\times n}$;
            Use the directed Cholesky factorization Algorithm 3 with $\mathbf{A} + D$ and M to obtain either $\widehat{R}$ and P;
            if *Algorithm 3 was successful* **then return** $R{:=}\widehat{R}P$ *and* D;
        end
    end
end
return *signaling failure*

and the diagonal matrix

$$J \in \mathbb{R}^{n\times n}, \quad J_{ij} := \begin{cases} 1 & \text{if } i = j \text{ and } (k < m \text{ or } i \notin M), \\ 0 & \text{otherwise.} \end{cases} \tag{5.3}$$

The next D is then chosen as a multiple σJ of J, trying increasing values of σ until we succeed.

Algorithm 4 gives a precise description of our modified directed Cholesky factorization algorithm

6. Testing the directed Cholesky factorizations. We compared the new incomplete directed Cholesky factorization and the modified directed Cholesky factorization methods with the old directed Cholesky factorization method on random real interval matrices of different dimension (column `dim` in the tables below) and width (column `width` in the tables below). These matrices can be constructed to be positive definite or indefinite and are always 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 (numerical) eigenvalues and the absolute value of the largest ones of all k test matrices $A^{(i)}$, formally:

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

The following algorithm shows how the test matrices are created:

Algorithm 5: Nearly singular interval matrix generator

Input: Given is the dimension n, a tiny singularity factor $\eta \neq 0$ with $|\eta| \ll 1$ (e.g. $|\eta| = 10^{-12}$) and the required relative width $\omega \geq 0$ of the interval matrix $\mathbf{A}$ to be created.

Output: Nearly singular positive definite (if $\eta > 0$) or indefinite (if $\eta < 0$) interval matrix $\mathbf{A} \in \mathbb{IR}^{n\times n}$, $\underline{A}_{ij} \in [0,1]$ of relative width ω

1 Generate a random matrix $B \in \mathbb{R}^{n-1\times n}$ with $B_{ij} \in [-1,1]$ for all $i = 1,\dots,n-1$ and $j = 1,\dots,n$;
2 Compute $C = B^TB \in \mathbb{R}^{n\times n}$ and $d = max(C_{ii})$;
3 **if** $d = 0$ **then** start again with step 1;
4 **else**
5   Generate a random vector $u \in \mathbb{R}^n$ with $u_j \in [-1,1]$ for all $j = 1,\dots,n$;
6   Divide u by $\max(|u|)$ such that $\|u\|_\infty = 1$ holds;
7   Set $\underline{A} = C/d + \eta uu^T$ and $\overline{A} = \underline{A} + \omega|\underline{A}|$;
8   **return** *the interval matrix* $\mathbf{A} := [\underline{A}, \overline{A}]$;
9 **end**

We first compare the approximate Cholesky factorization (computed with LAPACK, row `Chol` in the tables below), the directed Cholesky factorization with diagonal pivoting (computed by Domes & Neumaier [1, Algorithm 5.5], row `DirChol` in the tables below), the incomplete directed Cholesky factorization (computed by Algorithm 3, row `IDirChol` in the tables below) and the modified directed Cholesky factorization (computed by Algorithm 4, row `MDirChol` in the tables below) on 200 real positive definite matrices with dimensional 20 and a very small inverse condition number.

method	dim	width	iters	icond	diagpert	solved
`Chol`	20	0	200	$1.36{\cdot}10^{-13}$	0	100%
`DirChol`	20	0	200	$1.36{\cdot}10^{-13}$	0	2%
`IDirChol`	20	0	200	$1.36{\cdot}10^{-13}$	0	86%
`MDirChol`	20	0	200	$1.36{\cdot}10^{-13}$	$5.09{\cdot}10^{-13}$	100%

The next table shows how the efficiency of incomplete directed Cholesky factorization scales with the problem dimension.

method	dim	width	iters	icond	diagpert	solved
`IDirChol`	10	0	200	$1.95{\cdot}10^{-13}$	0	97%
`IDirChol`	40	0	200	$1.04{\cdot}10^{-13}$	0	53%
`IDirChol`	100	0	200	$9.94{\cdot}10^{-14}$	0	4%

The next table shows how the efficiency of incomplete directed Cholesky factorization scales with the problem dimension in case the interval entries of the matrices are not thin.

method	dim	width	iters	icond	diagpert	solved
`IDirChol`	10	10^{-14}	200	$2.03{\cdot}10^{-13}$	0	89%
`IDirChol`	40	10^{-14}	200	$1.25{\cdot}10^{-13}$	0	28%
`IDirChol`	100	10^{-14}	200	$1.09{\cdot}10^{-13}$	0	2%

The next table shows how the efficiency of directed modified Cholesky factorization scales with the problem dimension in case the interval entries of the matrices are not thin.

method	dim	width	iters	icond	diagpert	solved
`MDirChol`	10	0	200	$1.99{\cdot}10^{-13}$	$1.58{\cdot}10^{-13}$	100%
`MDirChol`	40	0	200	$1.26{\cdot}10^{-13}$	$1.75{\cdot}10^{-12}$	100%
`MDirChol`	100	0	200	$1.1{\cdot}10^{-13}$	$4.11{\cdot}10^{-10}$	100%

The next table shows how the efficiency of directed modified Cholesky factorization scales with the problem dimension in case the interval entries of the matrices are not thin.

method	dim	width	iters	icond	diagpert	solved
`MDirChol`	10	10^{-14}	200	$2.01{\cdot}10^{-13}$	$2.34{\cdot}10^{-13}$	100%
`MDirChol`	40	10^{-14}	200	$1.24{\cdot}10^{-13}$	$2.76{\cdot}10^{-12}$	100%
`MDirChol`	100	10^{-14}	200	$9.48{\cdot}10^{-14}$	$4.11{\cdot}10^{-10}$	100%

Discussion. The tests show that the new methods both improve the quality of the old one in case we want to factor ill-conditioned matrices. In particular the column `diagpert` shows the average diagonal perturbation

$$\operatorname*{med}_i(\max_k(D^{(i)}_{kk}))$$

(which can be nonzero only for `MDirChol`) and the percentage of successfully factored matrices.

REFERENCES

[1] F. Domes and A. Neumaier. Rigorous enclosures of ellipsoids and directed Cholesky factorizations. *SIAM Journal on Matrix Analysis and Applications*, 32:262–285, 2011.

[2] F. Domes and A. Neumaier. Constraint aggregation in global optimization. *Mathematical Programming*, pages 1–27, 2014. Online First.

[3] F. Domes and A. Neumaier. JGloptLab – a rigorous global optimization software. in preparation, 2014.

[4] A. Frommer and B. Hashemi. Verified stability analysis using the lyapunov matrix equation. *Electronic Transactions on Numerical Analysis*, 40:187–203, 2013.

[5] R.B. Kearfott, M.T. Nakao, A. Neumaier, S.M. Rump, S.P. Shary, and P. van Hentenryck. Standardized notation in interval analysis. In *Proc. XIII Baikal International School-seminar "Optimization methods and their applications"*, volume 4, pages 106–113, Irkutsk: Institute of Energy Systems, Baikal, 2005.

[6] R. Misener and C. A. Floudas. GloMIQO: Global mixed-integer quadratic optimizer. *Journal of Global Optimization*, 57(1):3–50, 2013.

[7] R. Misener and C. A. Floudas. ANTIGONE: Algorithms for coNTinuous / Integer Global Optimization of Nonlinear Equations. *Journal of Global Optimization*, 59(2-3):503–526, 2014.

[8] H. Schichl and M. C. Markót. Algorithmic differentiation techniques for global optimization in the COCONUT environment. *Optimization Methods and Software*, 27(2):359–372, 2012.

[9] H. Schichl, M. C. Markót, A. Neumaier, Xuan-Ha Vu, and C. Keil. The COCONUT Environment, 2000-2010. Software.

[10] R. B. Schnabel and E. Eskow. A new modified Cholesky factorization. *SIAM Journal on Scientific and Statistical Computing*, 11(6):1136–1158, 1990.

[11] R. B. Schnabel and E. Eskow. A revised modified Cholesky factorization algorithm. *SIAM Journal on Optimization*, 9(4):1135–1148, 1999.