March 26, 2018

Lecture 2 – Decompositions, perturbations

A Triangular systems

Exercise 2.1. Let $\mathbf{L} = (L_{ij})$ be an $n \times n$ lower triangular matrix ($L_{ij} = 0$ if $i > j$).

(a) Prove that $\mathbf{L}$ is non-singular if and only if $L_{ii} \neq 0$ for all i.

Assume $\mathbf{L}$ is non-singular. Let $\mathbf{B} \equiv \mathbf{L}^{-1}$.

(b) Prove that $\mathbf{B} = (B_{ij})$ is lower triangular as well.

For a given n-vector $\mathbf{b}$, we are interested in solving $\mathbf{Lx} = \mathbf{b}$ for $\mathbf{x}$. Note $\mathbf{x} = \mathbf{Bb}$.

(c) Suppose all B_{ij}, $i \leq j$, are non-zero as well as all coordinates b_i.
Compute the number of floating-point operations (flop) that are required to perform the matrix-vector multiplication (MV) $\mathbf{Bb}$.

(d) How many flop are needed to solve $\mathbf{Lx} = \mathbf{b}$ for $\mathbf{x}$ by forward substitution?

(e) Determine $\mathbf{B} = \mathbf{L}^{-1}$ in case $\mathbf{L} = \mathbf{I} + \mathbf{DS}$, where $\mathbf{S} = (S_{ij})$ is the shift matrix (all entries S_{ij} are zero, except for S_{ii+1} which equals 1 for all i) and $\mathbf{D}$ a diagonal matrix.

(f) Suppose $\mathbf{L}$ is banded (i.e., there is a k such that $L_{ij} = 0$ if $i - k < j$). Will $\mathbf{B}$ have a specific sparsity structure?

(g) Discuss the efficiency of computing $\mathbf{x}$ given $\mathbf{L}$ and $\mathbf{b}$.

Exercise 2.2. Let $\mathbf{L}$ be a $k \times k$ lower triangular matrix with non-zero diagonal elements. Put $\ell_j \equiv \mathbf{L}\mathbf{e}_j - \mathbf{e}_j$. Note that $\mathbf{e}_i^* \ell_j = 0$ if $i < j$.

(a) Prove that $\mathbf{L} = (\mathbf{I} + \ell_1 \mathbf{e}_1^*) \cdot \ldots \cdot (\mathbf{I} + \ell_k \mathbf{e}_k^*)$.

(b) Prove that, if $\mathrm{diag}(\mathbf{L}) = \mathbf{I}$ (i.e., all ones on the diagonal), then $\mathbf{e}_j^* \ell_j = 0$ and

$$(\mathbf{I} + \ell_j \mathbf{e}_j^*)^{-1} = \mathbf{I} - \ell_j \mathbf{e}_j^*.$$

Let $\mathbf{A}$ be an $k \times n$ matrix. Consider the sequence $(\mathbf{A}_j)$ of $k \times n$ matrices for which

$$\mathbf{A}_0 \equiv \mathbf{A}, \quad (\mathbf{I} + \ell_j \mathbf{e}_j^*)\mathbf{A}_j = \mathbf{A}_{j-1} \quad (j = 1, \ldots, k) \tag{2.1}$$

($\mathbf{A}_j$ is to be solved). Put $\mathbf{U} \equiv \mathbf{A}_k$.

(c) Prove that $\mathbf{A} = \mathbf{LU}$.
Note that with $n = 1$, this describes the solution process of lower triangular system $\mathbf{Ly} = \mathbf{b}$ with 'forward elimination' (i.e., take $\mathbf{A}_0 = \mathbf{b}$ and $\mathbf{y} = \mathbf{A}_k$).

B LU-decompositions

Exercise 2.3. **Gauss elimination.** Let $\mathbf{A}$ be an $k \times n$ matrix and let $\mathbf{L}$ be such that $\mathbf{e}_i^* \mathbf{L} \mathbf{e}_j = \mathbf{e}_i^* \mathbf{A}_{j-1} \mathbf{e}_j$ if $i \geq j$ (i.e., the jth column of $\mathbf{L}$ is the jth column of $\mathbf{A}_{j-1}$ below the diagonal), with $\mathbf{A}_{j-1}$ as in (2.1) ($j = 1, \ldots, k$).

(a) Prove that $\mathbf{U}$ is upper triangular (and $\mathrm{diag}(\mathbf{U}) = \mathbf{I}$).

(b) Check that (2.1) with $\mathbf{L}$ with columns scaled ($\mathbf{e}_i^* \mathbf{L} \mathbf{e}_j = \frac{1}{\nu_j} \mathbf{e}_i^* \mathbf{A}_{j-1} \mathbf{e}_j$ if $i \geq j$ with $\nu_j \equiv \mathbf{e}_j^* \mathbf{A}_{j-1} \mathbf{e}_j$, assuming $\nu_j \neq 0$) such that $\mathrm{diag}(\mathbf{L}) = \mathbf{I}$ represents the standard **Gauss elimination process** for computing an LU-decomposition (see Alg. 2.1).

Exercise 2.4. Let $\mathbf{A}$ be an $n \times n$ matrix.

(a) Show that if $\mathbf{A}$ has an LU-decomposition, $\mathbf{A} = \mathbf{LU}$, with $\mathrm{diag}(\mathbf{L}) = \mathbf{I}$, and is non-singular, then $\mathbf{L}$ and $\mathbf{U}$ are unique.

LU-FACTORISATION (STANDARD)

Put $\mathbf{U} = \mathbf{A} = (U_{ik})$.
For $k = 1, \ldots, n-1$ do
  For $i = k+1, \ldots, n$ do
    $L_{ik} = U_{ik}/U_{kk}$
    For $j = k+1, \ldots, n$ do
      $U_{ij} \leftarrow U_{ij} - L_{ik}U_{kj}$

LU-FACTORISATION (IKJ-VARIANT)

Put $\mathbf{U} = \mathbf{A} = (U_{ik})$.
For $i = 2, \ldots, n$ do
  For $k = 1, \ldots, i-1$ do
    $L_{ik} = U_{ik}/U_{kk}$
    For $j = k+1, \ldots, n$ do
      $U_{ij} \leftarrow U_{ij} - L_{ik}U_{kj}$

ALGORITHM 2.1. Computing the LU-factorisation with Gauss elimination of an $n \times n$ matrix $\mathbf{A}$: $\mathbf{A} = \mathbf{LU}$, where, with $\mathbf{L} = (L_{ik})$ and $\mathbf{U} = (U_{ik})$, the L_{ik} values for $k < i$ and the U_{ik} values for $k \geq i$ are as computed in the algorithm, $L_{kk} = 1$ all k and all other L_{ik} and U_{ik} values are 0. The left panel displays the standard variant, also called kij-variant. The right panel shows the so-called ikj-variant. In practise, the resulting U and the L factors are stored in the same location as the original matrix (i.e, if $\mathbf{U}$ is the original matrix, then the value L_{ik} in the above algorithms is stored in the location for U_{ik}: replace L_{ik} by U_{ik} in the algorithms. Then, upon termination, $L_{ik} = U_{ik}$ for $k < i$).

LU-FACTORISATION WITH PARTIAL PIVOTING

Put $\mathbf{U} = \mathbf{A} = (U_{ik})$, $\pi = (1, 2, \ldots, n)$
For $k = 1, \ldots, n-1$ do
  $m = \mathtt{argmax}\{|U_{\pi(j),k}| \,\big|\, j, j \geq k\}$
  $\ell = \pi(k)$, $\pi(k) = \pi(m)$, $\pi(m) = \ell$
  For $i = k+1, \ldots, n$ do
    $L_{\pi(i)\pi(k)} = U_{\pi(i)k}/U_{\pi(m)k}$
    For $j = k+1, \ldots, n$ do
      $U_{\pi(i)j} \leftarrow U_{\pi(i)j} - L_{\pi(i)\pi(k)}\, U_{\pi(k)j}$

ALGORITHM 2.2. Computing the LU-factorisation with Gauss elimination of an $n \times n$ matrix $\mathbf{A}$ using partial pivoting. Here, for a function f on $\mathbf{k} \equiv \{1, \ldots, k\}$ with values ≥ 0, argmax is the smallest index at which f takes its maximum ($j_0 = \mathtt{argmax}\{f(j) \,\big|\, j \in \mathbf{k}\}$ is such that $f(j) \leq f(j_0)$ for all $j \in \mathbf{k}$ and $j_0 \leq j$ if $f(j) = f(j_0)$). With $U_{\pi(i)k} = 0$ for $k < \pi(i)$, $L_{\pi(i)\pi(k)} = 0$ for $\pi(k) > \pi(i)$ and $L_{\pi(k)\pi(k)} = 1$ we have that $\mathbf{A} = \mathbf{LU}$, and, following the MATLAB conventions, $\mathbf{A}(\pi, :) = \mathbf{L}(\pi, \pi)\mathbf{U}(\pi, :)$ with $\mathbf{L}(\pi, \pi)$ lower triangular with ones on the diagonal and $\mathbf{U}(\pi, :)$ upper triangular; π defines a permutation $\mathbf{P}$: $\mathbf{PA} = \mathbf{A}(\pi, :)$. Note that this algorithm avoids swapping rows: it simply only swaps indices. Note that $\mathbf{x}$ solves $\mathbf{Ax} = \mathbf{b}$ if and only if $\mathbf{x}$ solves $\mathbf{A}(\pi, :)\mathbf{x} = \mathbf{b}(\pi)$.

(b) Show that if $\mathbf{A}$ has an LDU-decomposition, $\mathbf{A} = \mathbf{LDU}$, with $\mathrm{diag}(\mathbf{L}) = \mathrm{diag}(\mathbf{U}) = \mathbf{I}$, and is non-singular, then $\mathbf{L}$, $\mathbf{D}$ and $\mathbf{U}$ are unique.

Exercise 2.5. Let $\mathbf{A}$ be an $n \times n$ matrix. Below we neglect terms that are lower order in n. Show the following:

(a) It requires $2n^3/3$ flop (floating point operations) to compute an LU-decomposition.

(b) It requires $2n^2$ flop to solve both the systems $\mathbf{Ly} = \mathbf{b}$ and $\mathbf{Ux} = \mathbf{y}$.

(c) If $\mathbf{A}$ is banded with **bandwith** p, i.e., $A_{ij} = 0$ if $|i - j| > p$, then it requires $2p^2n$ flop to compute an LU-decomposition.

(d) Both $\mathbf{L}$ and $\mathbf{U}$ have also bandwidth p and it requires $2pn$ flop to solve both the systems $\mathbf{Ly} = \mathbf{b}$ and $\mathbf{Ux} = \mathbf{y}$.

Exercise 2.6. In Gauss elimination with **partial pivoting** (GEPP), rows of the partially eliminated matrices may be switch in each step of the process.

(a) Show that this can be represented by one permutation matrix $\mathbf{P}$ (a matrix that arises by permuting rows of the identity matrix $\mathbf{I}$), i.e., $\mathbf{PA} = \mathbf{LU}$, with $\mathbf{U}$ the upper triangular matrix as resulting from GEPP and $|L_{ij}| \leq 1$ for all matrix entries L_{ij} of $\mathbf{L}$: $\|\mathbf{L}\|_{\mathrm{M}} = 1$.

(b) How is $\mathbf{L}$ related to the L_{ij}-terms as computed in the steps of GEPP?

(c) Give an adaption of ALG. 2.1 (standard variant) that incorporates partial pivotting. Include also the updating of the permutation π of $(1, 2, \ldots, n)$ that represents the permutation matrix $\mathbf{P}$ (i.e., the jth row of $\mathbf{P}$ equals $\mathbf{e}^{\mathrm{T}}_{\pi(j)}$. Hint: start with $\pi = (1, 2, \ldots, n)$.)

(d) Give another variant of ALG. 2.1 (standard variant) where the partial pivotting is incorporated not by swapping rows, but by going through the rows in a customized order (swap the 'pointers' to the rows rather than the rows themselves), i.e., derive Alg. 2.2. Can you guess what strategy MATLAB is following?

Exercise 2.7. Suppose $\mathbf{A} = \mathbf{LU}$ with $\mathbf{L} = (L_{ij})$, $\mathbf{U} = (U_{ij})$ and $L_{ii} = 1$ all i. Derive an algorithm to compute L_{ij} and U_{ij} by comparing the product $\mathbf{LU}$ with $\mathbf{A}$.

Exercise 2.8. Let $\mathbf{A} = (a_{ij})$ be an $n \times n$ matrix.
Assume $\mathbf{A}$ is upper Hesenberg, that is, $a_{ij} = 0$ if $i > j + 1$.

(a) Assume LU-factorisation does not require pivoting (cf., Exercise 2.6). Show that $\mathbf{L}$ is bidiagonal. Give an (efficient) algorithm to compute the LU-factorisation, $\mathbf{A} = \mathbf{LU}$, of $\mathbf{A}$. Show that the factorisation can be computed with n^2 flop.

Tridiagonal matrices $\mathbf{T} = (t_{ij})$ are special Hessenberg matrices: $t_{ij} = 0$ also if $i < j - 1$. Give an efficient algorithm to compute the LU-factorisation for such a matrix (also assuming pivotting is not required).

(b) Assume partial pivoting is required. Does that affect the Hessenberg structure in the elimination steps? If the resulting factorisation is represented as $\mathbf{PA} = \mathbf{LU}$ with $\mathbf{P}$ an appropriate permutation, is $\mathbf{PA}$ still Hessenberg? Is a tridiagonal structure affected by partial pivotting?

(c) Suppose, for a sequence $\sigma_1, \ldots, \sigma_n$ of scalars and a sequence $\mathbf{b}_1, \ldots, \mathbf{b}_n$ of n-vectors, we want to solve all systems

$$(\mathbf{A} - \sigma_j \mathbf{I})\mathbf{x}_j = \mathbf{b}_j, \qquad (j = 1, \ldots, n). \tag{2.2}$$

Show that this can be done with $2n^3$ flop (neglecting terms of order n^2, i.e., modest multiples of n^2).

Now, suppose $\mathbf{A}$ is a general $n \times n$ matrix (no specific algebraic structure, no specific sparsity structure).

(d) In Exercise 3.19 in Lecture 3, we will see that with $\frac{8}{3}n^3$ flop we can compute an $n \times n$ unitary matrix $\mathbf{Q}$ and an upper Hessenberg matrix $\mathbf{H}$ such that $\mathbf{AQ} = \mathbf{QH}$ (a Hessenberg factorisation). Use this and explain how the n shifted systems (2.2) can be solved in $8\frac{2}{3}n^3$ flop. Analyse the costs if the n shifted systems are solved without a Hessenberg factorisation.

Variants for computing the LU-decomposition. Matrices are often stored row-wise. The variant of the Gaussian elimination process in the right panel of ALG. 2.1, the so called **ikj-variant**, fits better this row-wise storage format. The standard variant in Exercise 2.3 is the **kij-variant**. Note that the ikj-variant does not allow pivoting of rows. Column pivoting is possible.

Exercise 2.9. Show that $\mathbf{A} = \mathbf{LU}$, with $\mathbf{U}$ the upper triangular part of the matrix as constructed in the ikj-variant and $\mathbf{L} = \mathbf{I} + \mathbf{L}'$, where $\mathbf{L}' = (L_{ij})$ also as in the ikj-variant: both variants give the same $\mathbf{L}$ and $\mathbf{U}$ factors (also in floating-point arithmetic).

Theorem 2.1 *An $n \times n$ matrix $\mathbf{A}$ has a non-singular LU-decomposition (without pivoting), $\mathbf{A} = \mathbf{LU}$, i.e., $\mathbf{L}$ has all 1 on its diagonal $\mathbf{U}$ is non-singular, if and only if each left upper-block of $\mathbf{A}$ is non-singular. The LU-decomposition is unique in this case.*

Exercise 2.10. ***Proof of Theorem*** 2.1**.**

(a) Show that both $\mathbf{L} = (L_{ij})$ and $\mathbf{U} = (U_{ij})$ are non-singular if $\mathbf{A} = \mathbf{LU}$ and $\mathbf{A}$ is non-singular. Note that $\det(\mathbf{A}) = \det(\mathbf{L})\det(\mathbf{U}) = \det(\mathbf{U}) = \prod_{i=1}^{n} U_{ii}$. Conclude that $\mathbf{A}$ is non-singular if and only if all diagonal entries U_{ii} of $\mathbf{U}$ are non-zero.

(b) Let $\mathbf{A}_j$, $\mathbf{L}_j$, and $\mathbf{U}_j$ be the $j \times j$ left upper block of $\mathbf{A}$, $\mathbf{L}$, and $\mathbf{U}$, respectively. Prove that $\mathbf{A}_j = \mathbf{L}_j\mathbf{U}_j$ is the LU-decomposition of $\mathbf{A}_j$ if $\mathbf{A} = \mathbf{LU}$ is the LU-decomposition of $\mathbf{A}$ $(j = 1, \ldots, n)$.

(c) Assume $\mathbf{A}$ is non-singular. Use an induction argument and the above results to show that $\mathbf{A}$ has an LU-decomposition if and only if each square left upper-block of $\mathbf{A}$ is non-singular.

(d) Prove that the LU-decomposition is unique if the diagonal elements of $\mathbf{L}$ are fixed to 1 and all square left upper blocks of $\mathbf{A}$ are non-singular.

For an $n \times n$ matrix $\mathbf{A}$ it is often convenient to consider a some block partitioning

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix},$$

where for some m, $1 \leq m \leq n$, $\mathbf{A}_{11}$ is the $m \times m$ left upper-block of $\mathbf{A}$, with $k \equiv n - m$, $\mathbf{A}_{22}$ the $k \times k$ right lower-block, etc..

In case $\mathbf{A}_{11}$ is non-singular, consider the **block LU-factorisation**

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} = \mathbf{A} = \mathbf{LU} = \begin{bmatrix} \mathbf{I}_1 & \mathbf{0} \\ \mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{I}_2 \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{21} \\ \mathbf{0} & \mathbf{U}_{22} \end{bmatrix}$$

with $\mathbf{I}_1$ and $\mathbf{I}_2$ identity matrices of appropriate size and $\mathbf{U}_{22} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{21}$. $\mathbf{U}_{22}$ is the **Schur complement** (of $\mathbf{A}_{11}$ in $\mathbf{A}$).

C Cholesky decomposition

A complex $n \times n$ matrix $\mathbf{A}$ is **positive definite** if
(i) $\mathbf{A}$ is Hermitian, i.e, $\mathbf{A}^* = \mathbf{A}$, and (ii) $\mathbf{x}^*\mathbf{A}\mathbf{x} > 0 \quad (\mathbf{x} \in \mathbb{C}^n,\ \mathbf{x} \neq \mathbf{0})$.
$\mathbf{A}$ is said to be **positive semi-definite** if (i) $\mathbf{A}$ is Hermitian and (ii)' $\mathbf{x}^*\mathbf{A}\mathbf{x} \geq 0 \quad (\mathbf{x} \in \mathbb{C}^n)$.
A real matrix $\mathbf{A}$ is **positive definite** if (i)' $\mathbf{A}$ is symmetric, i.e., $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$ and (ii)" $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} > 0$ $(\mathbf{x} \in \mathbb{R}^n,\ \mathbf{x} \neq \mathbf{0})$.

As observed in Exercise 0.29 of Lecture 0, property (i) follows from property (ii), in the complex case. If the matrix is real and $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{0}$, then (i) does not follow and the additional restriction $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$ is required for positive definiteness.

In the literature, 'positive definite' and 'positive semi-definite' are also called 'strictly positive definite' and 'positive definite', respectively.

Exercise 2.11. Prove that a positive definite matrix is non-singular.

Theorem 2.2 *Let* $\mathbf{A}$ *be a complex or real positive (semi-)definite* $n \times n$ *matrix. Then there is a* **Cholesky factorisation***, i.e., there is an upper triangular* $n \times n$ *matrix* $\mathbf{C}$ *with all diagonal entries* ≥ 0*, a so called* **Cholesky factor***, such that*

$$\mathbf{A} = \mathbf{C}^*\mathbf{C}.$$

In particular, $\|\mathbf{C}\|_2^2 = \|\mathbf{A}\|_2$*. If* $\mathbf{A}$ *is real, then* $\mathbf{C}$ *is real.*
$\mathbf{A}$ *is positive definite if and only if all diagonal entries of* $\mathbf{C}$ *are* > 0*. The Cholesky factor is unique if* $\mathbf{A}$ *is positive definite. Then, we also have that* $\mathcal{C}_2^2(\mathbf{C}) = \mathcal{C}_2(\mathbf{A})$*.*

The existence of a Cholesky factorisation follows from Theorem 2.1 and the fact that each left upper block of a complex positive definite matrix is positive definite as well (why?) and therefore non-singular. The proof in the exercise below essentially folows this argument, but is a but more straight-forward.

Exercise 2.12. ***Proof of Theorem*** 2.2**.** First assume that $\mathbf{A}$ is positive definite.

(a) If the Cholesky decomposition exists, then prove that the Cholesky factors are non-singular and $\mathcal{C}_2^2(\mathbf{C}) = \mathcal{C}_2(\mathbf{A})$. Note that the Cholesky factor is non-singular if and only if all its diagonal entries are non-zero.

To prove existence and uniqueness of the Cholesky factorisation, partition $\mathbf{C}$ and $\mathbf{A}$ as

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}' & \mathbf{c} \\ \mathbf{0}^* & \gamma \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \mathbf{A}' & \mathbf{a} \\ \mathbf{a}^* & \alpha \end{bmatrix},$$

where $\mathbf{C}'$ is the left $(n-1) \times (n-1)$ of $\mathbf{C}$, $(\mathbf{c}^{\mathrm{T}}, \gamma)^{\mathrm{T}}$ is the last column of $\mathbf{C}$, and we have a similar partitioning of $\mathbf{A}$.

Assume $\mathbf{A}'$ has a Cholesky factorisation with unique Cholesky factor $\mathbf{C}'$. We will prove that the Cholesky decomposition of $\mathbf{A}$ (uniquely) exists.

(b) Show that $\mathbf{c}$ is the (unique) solution of the lower triangular system $(\mathbf{C}')^*\mathbf{c} = \mathbf{a}$, which can be solved by forward substitution.

Put $\beta \equiv \alpha - \|\mathbf{c}\|_2^2$.

(c) Prove that, with $\gamma = \sqrt{\beta} \geq 0$ if $\beta \geq 0$, $\mathbf{C}$ is the Cholesky factor of $\mathbf{A}$.

(d) Derive a contradiction if $\beta = 0$.

(e) Assume that $\beta < 0$. Consider the Cholesky decomposition for the matrix $\mathbf{A} + |\beta| e_n e_n^*$. Prove that this matrix is positive definite and derive a contradiction.

(f) Prove Theorem 2.2.

(g) Show that $\mathbf{C}$ is real if $\mathbf{A}$ is real.

(h) If $\mathbf{A}$ has a Cholesky factorisation, then $\mathbf{A}$ is positive definite.

Assume that $\mathbf{A}$ is positive semi-definite.

(i) Prove the existence of a Cholesky factorisation (Hint: Note that $\mathbf{A} + \varepsilon\mathbf{I}$ is positive definite for all $\varepsilon > 0$ and recall that $\|\mathbf{C}_\varepsilon\|_2^2 = \|\mathbf{A} + \varepsilon\mathbf{I}\|_2$ for the appropriate Cholesky factor $\mathbf{C}_\varepsilon$).

(j) Give an example to illustrate the fact that the Cholesky factors need not to be unique if $\mathbf{A}$ is positive semi-definite.

Exercise 2.13. Suppose that $\mathbf{A}$ is a symmetric and positive definite tridiagonal matrix. Give an (efficient) algorithm to compute the $\mathrm{LDL}^{\mathrm{T}}$-decomposition of $\mathbf{A}$, i.e., compute the factors $\mathbf{D}$ and $\mathbf{L}'$ such that $\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^{\mathrm{T}}$ with $\mathbf{L} = \mathbf{I} + \mathbf{L}'$, $\mathbf{L}'$ strict lower triangular, and $\mathbf{D}$ diagonal.

Exercise 2.14. Let $\mathbf{A}$ be a complex positive semi-definite $n \times n$ matrix.

(a) For $\mathbf{x} \in \mathbb{C}^n$, show that $\mathbf{x} \in \mathcal{N}(\mathbf{A})$ if and only if $\mathbf{x}^*\mathbf{A}\mathbf{x} = 0$. Does this equivalence hold for any Hermitian matrix $\mathbf{A}$?

(b) Assume that $\mathbf{A} = \mathbf{L}\mathbf{L}^*$ for some $n \times n$ matrix $\mathbf{L}$. Show that

$$\mathcal{N}(\mathbf{L}^*) = \mathcal{N}(\mathbf{A}) \quad \text{and} \quad \mathcal{R}(\mathbf{L}) = \mathcal{R}(\mathbf{A}).$$

D Iterative refinement

Exercise 2.15. Let $\mathbf{A}$ be a non-singular $n \times n$ matrix. Let $\mathbf{b}$ be an n-vector. We are interested in solving $\mathbf{Ax} = \mathbf{b}$ for $\mathbf{x}$. Suppose we have a numerical procedure that, for some $\delta \in (0, \frac{1}{2})$, produces for any n-vector $\mathbf{r}$ a computed solution $\widehat{\mathbf{u}}$ of the system $\mathbf{Au} = \mathbf{r}$ such that $(\mathbf{A}+\Delta)\widehat{\mathbf{u}} = \mathbf{r}$ for some $n \times n$ matrix Δ; Δ may depend on $\mathbf{r}$, but $\|\mathbf{A}^{-1}\Delta\| \leq \delta$.

```
Solve Ax = b for x.
Put x_0 ≡ x̂, r_0 ≡ b, u_0 = x_0.
For i = 1,... do
    Compute r_i ≡ r_{i-1} − Au_{i-1}
    Stop if ‖r_i‖ is sufficiently small
    Solve Au = r_i for u
    With u_i ≡ û, update x_i = x_{i-1} + u_i
```

We assume that the errors in the updates or $\mathbf{r}_{i-1}$ and $\mathbf{x}_{i-1}$ and in the matrix vector multiplication $\mathbf{Au}_{i-1}$ are much smaller than δ, that is, we neglect these errors.

Prove the following claims.

(a) $\mathbf{r}_i = \mathbf{b} - \mathbf{Ax}_{i-1}$.

(b) $\mathbf{e}_i \equiv \mathbf{x} - \mathbf{x}_i = \mathbf{e}_{i-1} - (\mathbf{A} + \Delta_i)^{-1}\mathbf{Ae}_{i-1} = (\mathbf{I} + \mathbf{E}_i)^{-1}\mathbf{E}_i e_{i-1}$,
where Δ_i is such that $(\mathbf{A} + \Delta_i)\mathbf{u}_i = \mathbf{r}_i$ and $\mathbf{E}_i \equiv \mathbf{A}^{-1}\Delta_i$.

(c) $\|\mathbf{e}_i\| \leq \frac{\delta}{1-\delta}\|\mathbf{e}_{i-1}\| \leq (\frac{\delta}{1-\delta})^{i+1}$.

(d) Argue that the assumption on the size of the errors in the updates is a reasonable one as long as $\|\mathbf{e}_i\|$ is much larger (some factors n) than $\mathbf{u}\,\|\mathbf{x}\|$ with $\mathbf{u}$ unit round off.

(e) Assume that $\delta \leq 10^{-2}$. Conclude that the error in $\mathbf{x}_5$ as a solution of $\mathbf{Ax} = \mathbf{b}$ is smaller than 10^{-12}.

(f) Suppose the numerical procedure is the LU-decomposition (with partial pivoting). Compare the costs (in flop) for computing $\mathbf{x}_0$ and $\mathbf{x}_5$.

A practical stopping criterion would be 'Stop if $\|\mathbf{r}_i\|/(\|\mathbf{A}\|\,\|\mathbf{x}_0\|) \leq 10^{-12}$'; see (1.26) in Section C of Lecture 1.

E Rounding errors

We follow the conventions as introduced in Section E of Lecture 1.

LU-decomposition.

The following theorem gives the backward error in the solution process based on LU-decomposition (from Gauss elimination).

An $n \times n$ matrix $\mathbf{A} = (A_{ij})$ has **bandwidth** p if $|A_{ij}| = 0$ whenever $|i - j| \geq p$.

Theorem 2.3 *Let $\mathbf{A}$ be an $n \times n$ matrix with bandwidth p. Let $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{U}}$ be the L and U factors of the LU-decomposition of $\mathbf{A}$ as computed by Gauss elimination (without pivoting). Then*[1]

$$\mathbf{A} + \Delta_{LU} = \widehat{\mathbf{L}}\widehat{\mathbf{U}} \quad \textit{for some } \Delta_{LU} \textit{ for which} \quad |\Delta_{LU}| \leq p\,\mathbf{u}\,|\widehat{\mathbf{L}}|\,|\widehat{\mathbf{U}}|.$$

Let $\widehat{\mathbf{x}}$ be the solution of the system $\mathbf{Ax} = \mathbf{b}$ as computed by solving $\widehat{\mathbf{L}}\mathbf{y} = \mathbf{b}$ and $\widehat{\mathbf{U}}\mathbf{x} = \widehat{\mathbf{y}}$. Then

$$(\mathbf{A} + \Delta)\widehat{\mathbf{x}} = \mathbf{b} \quad \textit{for some } \Delta \textit{ for which} \quad |\Delta| \leq 3\,p\,\mathbf{u}\,|\widehat{\mathbf{L}}|\,|\widehat{\mathbf{U}}|.$$

[1] As in Section E in Lecture 1, inequalities and absolute values are matrix entry wise.

Exercise 2.16. ***Proof of Theorem*** 2.3**.** Let $\mathbf{L}$ be an $n \times n$ lower non-singular triangular matrix with bandwith p.

(a) Let $\mathbf{b}$ be an n-vector. Let $\widehat{\mathbf{x}}$ be the solution of $\mathbf{L}\mathbf{x} = \mathbf{b}$ as computed by forward substitution. Note that the j coordinate x_j of $\mathbf{x}$ is computed as a difference of the scalar b_j (the jth coordinate of $\mathbf{b}$) and an inner product. Use this fact and the results in (1.29) of Section E in Lecture 1 to show that, for some $n \times n$ matrix Δ_L we have that

$$(\mathbf{L} + \Delta_L)\widehat{\mathbf{x}} = \mathbf{b} \quad \text{with} \quad |\Delta_L| \leq p\,\mathbf{u}\,|\mathbf{L}| \qquad \text{and} \qquad |\mathbf{b} - \mathbf{L}\widehat{\mathbf{x}}| \leq p\,\mathbf{u}\,|\mathbf{L}|\,|\widehat{\mathbf{x}}|.$$

Let $\mathbf{A}$ be an $n \times n$ matrix with bandwidth p.

(b) Let $\widehat{\mathbf{U}}$ be the $n \times n$ matrix as computed by forward substitution by solving $\mathbf{L}\mathbf{U} = \mathbf{A}$ column-wise, i.e., $\mathbf{L}\mathbf{u}_i = \mathbf{A}\mathbf{e}_i$ is solved for $\mathbf{u}_i$, for $i = 1, 2, \ldots, n$ and we put $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$. Show that, for some $n \times n$ matrix Δ

$$(\mathbf{A} + \Delta) = \mathbf{L}\widehat{\mathbf{U}} \quad \text{with} \quad |\Delta| = |\mathbf{A} - \mathbf{L}\,\widehat{\mathbf{U}}| \leq p\,\mathbf{u}|\mathbf{L}|\,|\widehat{\mathbf{U}}|.$$

Do we also have that $\mathbf{A} = (\mathbf{L} + \Delta'_L)\widehat{\mathbf{U}}$ for some $n \times n$ matrix Δ'_L for which $|\Delta'_L| \leq p\,\mathbf{u}\,|\mathbf{L}|$?

(c) Suppose $\widehat{\mathbf{L}}$ is the lower triangular matrix as *computed* by the Gauss elimination process. Show that the factor $\widehat{\mathbf{U}}$ as computed by Gauss elimination is equal (in rounded arithmetic) to the solution of $\widehat{\mathbf{L}}\mathbf{U} = \mathbf{A}$ as computed as indicated in (b) (do we have to worry about the zeros in the lower triangular part of $\widehat{\mathbf{U}}$?). Note that for the claim in (b) it is irrelevant where the lower triangular matrix $\mathbf{L}$ comes from and whether it is perturbed by an error. Conclude that the computed factors in the LU-decomposition satisfy

$$\mathbf{A} + \Delta_{LU} = \widehat{\mathbf{L}}\,\widehat{\mathbf{U}} \quad \text{with} \quad |\Delta_{LU}| \leq p\,\mathbf{u}\,|\widehat{\mathbf{L}}|\,|\widehat{\mathbf{U}}|$$

(d) Let $\widehat{\mathbf{x}}$ be the solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ as computed by solving $\widehat{\mathbf{L}}\mathbf{y} = \mathbf{b}$ with computed solution $\widehat{\mathbf{y}}$ and $\widehat{\mathbf{U}}\mathbf{x} = \widehat{\mathbf{y}}$. Show that

$$(\mathbf{A} + \Delta)\widehat{\mathbf{x}} = \mathbf{b} \quad \text{with} \quad |\Delta| = |\Delta_{LU} + \widehat{\mathbf{L}}\Delta_U + \Delta_L\widehat{\mathbf{U}}| \leq 3\,p\,\mathbf{u}\,|\widehat{\mathbf{L}}|\,|\widehat{\mathbf{U}}|.$$

Here, Δ_L, Δ_U, is the perturbation term that represents the error in solving $\widehat{\mathbf{L}}\mathbf{y} = \mathbf{b}$ for $\mathbf{y}$, and solving $\widehat{\mathbf{U}}\mathbf{x} = \widehat{\mathbf{y}}$ for $\mathbf{x}$, respectively.

Theorem 2.3 leads to the following (sharp) estimates for the backward errors in Gauss elimination. For a proof, we refer to Exercise 2.17. Note that with partial pivoting, the bandstructure will be (partly) destroyed. With pivotting we may have to replace p by n.

Corollary 2.4 *For the backward error in the solution of a system obtained with and LU-decomposition we have that*

$$\frac{\|\Delta\|_\infty}{\|\mathbf{A}\|_\infty} \leq 3\,p\,\mathbf{u}\,\frac{\|\,|\widehat{\mathbf{L}}|\,|\widehat{\mathbf{U}}|\,\|_\infty}{\|\mathbf{A}\|_\infty}. \tag{2.3}$$

With partial pivoting (cf., Exercise 2.6), this leads to

$$\frac{\|\Delta\|_\infty}{\|\mathbf{A}\|_\infty} \leq 3\,n^3\,\mathbf{u}\,\rho_{\text{pv}}(\mathbf{A}), \quad \textit{where} \quad \rho_{\text{pv}}(\mathbf{A}) \equiv \frac{\|\widehat{\mathbf{U}}\|_M}{\|\mathbf{A}\|_M}. \tag{2.4}$$

The result indicates that the *stability of the Gauss elimination process* is a factor of order $\|\,|\widehat{\mathbf{L}}|\,|\widehat{\mathbf{U}}|\,\|/\|\,|\mathbf{A}|\,\|$ worse than the stability of the problem "solve $\mathbf{A}\mathbf{x} = \mathbf{b}$ for $\mathbf{x}$", cf., Section F in Lecture 1. This factor can be viewed as the conditioning of Gauss elimination.

Once the LU factors are available, the factor $\|\,|\widehat{\mathbf{L}}|\,|\widehat{\mathbf{U}}|\,\|/\|\,|\mathbf{A}|\,\|$ can be computed. Since, with $\mathbf{1} \equiv (1, 1, \ldots, 1)^{\mathrm{T}}$, $\|\,|\widehat{\mathbf{L}}|\,|\widehat{\mathbf{U}}|\,\|_\infty = \|\,|\widehat{\mathbf{L}}|\,(|\widehat{\mathbf{U}}|\mathbf{1})\,\|_\infty$ (why?) the computation of $\|\,|\widehat{\mathbf{L}}|\,|\widehat{\mathbf{U}}|\,\|_\infty$ requires n^2 flop; much less than the computation of the LU-factors. Nevertheless, the so-called **growth factor** $\rho_{\text{pv}}(\mathbf{A})$ (cf., (2.4)) of the Gauss elimination process with partial pivoting, is available for free as a side product of the elimination process and can conveniently be used saving

computational costs. Note, however, that $\| \, |\widehat{\mathbf{L}}| \, |\widehat{\mathbf{U}}| \, \|_\infty / \|\mathbf{A}\|_\infty$ can be smaller than $n^2 \, \rho_{\rm pv}(\mathbf{A})$ by a factor n^2.

It can be proved that the quantity $\rho_{\rm pv}(\mathbf{A})$ in (2.4) is at most 2^{n-1}. In Exercise 2.18 below, we will see that this bound is sharp. Already for $n \leq 50$, this upper bound is larger than 10^{15}. However, in practice, it appears that $\rho_{\rm pv}(\mathbf{A})$ rarely is larger than 16 (**Wilkinson's miracle**, 1965). Nevertheless, the Gauss elimination process with pivoting can be unstable also since the factor n^3 can be large. Full systems of size up to $n = 10000$ are standardly solved with Gauss elimination with partial pivoting. With $n = 10000$, the factor $n^3 = 10^{12}$ allows a loss of 12 digits of accuracy (or more depending on $\mathcal{C}(\mathbf{A})$ even if $\rho_{\rm pv}(\mathbf{A})$ is modest). For large n, the matrix should have some sparsity structure to make Gauss elimination feasible. In such a case, pivoting has to be limited in order to avoid destruction of the sparsity.

Exercise 2.17. **The stability of Gauss elimination.** Consider Theorem 2.3.
The norm $\| \cdot \|$ in this exercise is a p-norm for $p \in [1, \infty]$. (Otherwise p is the bandwidth of $\mathbf{A}$.)
(a) Prove that

$$\|\mathbf{A}^{-1} \Delta_{LU}\| \leq p \, \mathbf{u} \, \mathcal{C}(\mathbf{A}) \, \frac{\| \, |\widehat{\mathbf{L}}| \, |\widehat{\mathbf{U}}| \, \|}{\|\mathbf{A}\|}$$

Use the results of Theorem 1.11 in Lecture 1 to discuss the non-singularity of $\widehat{\mathbf{L}}\widehat{\mathbf{U}}$ and to estimate $\|\mathbf{A} - \widehat{\mathbf{L}}\widehat{\mathbf{U}}\|$ and $\|\mathbf{A}^{-1} - (\widehat{\mathbf{L}}\widehat{\mathbf{U}})^{-1}\|$.
(b) Prove that (the backward error [cf., Theorem 1.10] can be bounded by

$$\frac{\|\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}\|}{\|\mathbf{A}\| \, \|\widehat{\mathbf{x}}\|} \leq 3 \, p \, \mathbf{u} \, \frac{\| \, |\widehat{\mathbf{L}}| \, |\widehat{\mathbf{U}}| \, \|}{\|\mathbf{A}\|}. \tag{2.5}$$

(c) Prove that (the forward can be bounded by)

$$\frac{\|\mathbf{x} - \widehat{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{\delta \mathcal{C}(\mathbf{A})}{1 - \delta \mathcal{C}(\mathbf{A})}, \quad \text{where} \quad \delta \equiv 3p \, \mathbf{u} \, \frac{\| \, |\widehat{\mathbf{L}}| \, |\widehat{\mathbf{U}}| \, \|}{\|\mathbf{A}\|}$$

and assuming that $\delta \mathcal{C}(\mathbf{A}) < 1$.
(d) Assume partial pivoting has been applied (cf., Exercise 2.6). Now, consider $\mathbf{PA}$ instead of $\mathbf{A}$ (note that partial pivoting applied to $\mathbf{PA}$ does not lead to a permutation of the rows. We take $p = n$). Prove that then

$$\|\widehat{\mathbf{L}}\|_M \leq 1 \quad \text{and} \quad \frac{\| \, |\widehat{\mathbf{L}}| \, |\widehat{\mathbf{U}}| \, \|_\infty}{\|\mathbf{A}\|_\infty} \leq n^2 \, \rho(\mathbf{A}),$$

with $\rho(\mathbf{A})$ as in (2.4). Hence, δ can be estimated by $3 \, \mathbf{u} \, n^3 \, \rho_{\rm pv}(\mathbf{A})$ (in case $\| \cdot \| = \| \cdot \|_\infty$).

Partial pivoting limits the growth factor in Gauss elimination. Nevertheless, there are (exceptional) situations where the growth is exponential in the dimension as we will see in the next exercise. Although complete pivoting provably limits the growth even more, the bound might still be huge for high dimensions: $\rho_{\rm comp}(\mathbf{A}) \leq \kappa n^{\frac{3}{2} + \frac{1}{4} \log(n)}$ for some modest constant κ (it is conjectured that $\rho_{\rm comp}(\mathbf{A}) \leq \kappa n$). With **complete pivoting**, the absolute largest matrix entry of the 'active' matrix (rather than of the first column of this matrix) is brought to 'pivot position': after j sweeps of Gauss elimination, the active matrix is the $(n-j) \times (n-j)$ right lower block of the matrix obtained after elimination of the first j columns (with the pivot strategy). The absolute largest matrix entry of this matrix is brought to the top row of this matrix by switching two rows. Then it is brought to the left top position (the pivot position) by switching two columns. Eventually, we have an LU-decomposition of $\mathbf{P}_1 \mathbf{A} \mathbf{P}_2$ for some permutations $\mathbf{P}_1$ and $\mathbf{P}_2$.

Stability of Gauss elimination requires pivoting. Complete pivoting is hardly ever applied: in practise partial pivoting appears to be sufficiently stable. For high dimensional, sparse systems, pivot strategies on rows as well as on columns are frequently applied. However, in these cases, the pivot strategies aim for efficiency (to keep fill-in as low as possible) and to

maintain some symmetry structure rather than for stability. At best they try to keep the lost of stability limited: often more stability leeds to more fill, that is to a computationally more costly process.

Exercise 2.18. **Stability LU-decomposition.** For $\varepsilon \in \mathbb{R}, \varepsilon \neq 0$, let

$$\mathbf{A} \equiv \begin{bmatrix} \varepsilon & 1 \\ 1 & 0 \end{bmatrix}.$$

(a) Determine the LU-decomposition $\mathbf{A} = \mathbf{LU}$ of $\mathbf{A}$ (without pivoting).

(b) Compute (in exact arithmetic) $\frac{\| \, |\mathbf{L}| \, |\mathbf{U}| \, \|_\infty}{\|\mathbf{A}\|_\infty}$ and $\frac{\|\mathbf{L}\|_\infty \, \|\mathbf{U}\|_\infty}{\|\mathbf{A}\|_\infty}$.

Exercise 2.19. Here, for an $n \times n$ matrix $\mathbf{A}$, we estimate $\rho_{\text{pv}}(\mathbf{A})$, where

$$\rho_{\text{pv}}(\mathbf{A}) \equiv \frac{\|\mathbf{U}\|_M}{\|\mathbf{A}\|_M} \tag{2.6}$$

with the LU-factors $\mathbf{L}$ and $\mathbf{U}$ as obtained by Gauss elimination with partial pivoting.

(a) Let $\mathbf{A}_0 \equiv \mathbf{A}$, $\mathbf{A}_1, \ldots,$ $\mathbf{U} \equiv \mathbf{A}_{n-1}$ the intermediate matrices as computed in the Gauss elimination process with partial pivoting (cf., (2.1)): $\mathbf{A}_k$ is obtained after elimination of the first k-columns. Recall that with partial pivoting the multiplication factors are ≤ 1 in absolute value and prove that $\|\mathbf{A}_1\|_M \leq 2\|\mathbf{A}_0\|_M$. Conclude that $\|\mathbf{U}\|_M \leq 2^{n-1}\|\mathbf{A}\|_M$, whence

$$\rho_{\text{pv}}(\mathbf{A}) \leq 2^{n-1}. \tag{2.7}$$

To prove that (2.7) is sharp, consider the $n \times n$ matrix $\mathbf{A}$ given by

$$\mathbf{A} \equiv \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 & 1 \\ -1 & 1 & 0 & \ldots & 0 & 1 \\ -1 & -1 & 1 & & 0 & 1 \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ -1 & -1 & -1 & \ldots & 1 & 1 \\ -1 & -1 & -1 & \ldots & -1 & 1 \end{bmatrix}. \tag{2.8}$$

(b) Determine the LU-decomposition $\mathbf{A} = \mathbf{LU}$ of $\mathbf{A}$ (without pivoting). Note that partial pivoting leads to the same result (no rows have to be switched).

(c) Show that $\rho_{\text{pv}}(\mathbf{A}) = 2^{n-1}$.

(d) Compute (in exact arithmetic) $\frac{\| \, |\mathbf{L}| \, |\mathbf{U}| \, \|_\infty}{\|\mathbf{A}\|_\infty}$ and $\frac{\|\mathbf{L}\|_\infty \, \|\mathbf{U}\|_\infty}{\|\mathbf{A}\|_\infty}$.

(e) Solve $\mathbf{Ly} = \mathbf{e}_1$. Show that $\|\mathbf{L}^{-1}\|_1 = \|\mathbf{L}^{-1}\|_\infty = 2^{n-1}$ and $\mathcal{C}_\infty(\mathbf{L}) \geq n\, 2^{n-1}$.

(f) Compute, now with complete pivoting, the L and U factors of the matrix $\mathbf{A}$ from (2.8) for $n = 4$: $\mathbf{P}_1\mathbf{A}\mathbf{P}_2 = \mathbf{LU}$. Determine $\rho_{\text{comp}}(\mathbf{A})$.

Exercise 2.20. Here we discuss the estimate

$$\frac{\| \, |\mathbf{L}| \, |\mathbf{U}| \, \|_\infty}{\|\mathbf{A}\|_\infty} \leq n^2 \rho_{\text{pv}}(\mathbf{A}) \quad \text{where} \quad \rho_{\text{pv}}(\mathbf{A}) = \frac{\|\mathbf{U}\|_M}{\|\mathbf{A}\|_M} \tag{2.9}$$

with the LU-factorisation as obtained by Gauss elimination with partial pivoting.

(a) Derive (2.9).

To show that this estimate is sharp (except for some modest factor), consider the $n \times n$ matrix $\mathbf{S}$ defined by $\mathbf{S}\mathbf{e}_j \equiv \mathbf{e}_{j+1}$ for $j = 1, \ldots, n-1$ and $\mathbf{S}\mathbf{e}_n = \mathbf{0}$: $\mathbf{S}$ shifts the basisvectors. Let $\mathbf{a}$ be the n-vector $\mathbf{a} \equiv (1, -1, 1, -1, \ldots)^{\text{T}}$.

(b) Show that $(\mathbf{I}+\mathbf{S})\mathbf{a} = \mathbf{e}_1$.
Prove that $(\mathbf{I}-\mathbf{S})^{-1}$ is the $n \times n$ lower triangular matrix with all ones in the lower triangle.
With $\mathbf{b} \equiv (\mathbf{I}-\mathbf{S}^*)^{-1}\mathbf{a}$, show that $\mathbf{b} = (0,-1,0,-1,\ldots)^{\mathrm{T}}$.

(c) Consider the $(n+1)\times(n+1)$ lower triangular matrix $\mathbf{L}_1$ defined (as block matrices) by

$$\mathbf{L}_1 \equiv \begin{bmatrix} \mathbf{I}-\mathbf{S} & \mathbf{0} \\ \mathbf{a}^* & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{U}_1 \equiv \begin{bmatrix} \mathbf{I}+\mathbf{S}^* & \mathbf{0} \\ \mathbf{0}^* & 1 \end{bmatrix}.$$

Compute $\mathbf{L}_1^{-1}$. Define the $(2n+2)\times(2n+2)$ matrices $\mathbf{L}$, $\mathbf{U}$ and $\mathbf{A}$ (as block matrices) by

$$\mathbf{L} \equiv \begin{bmatrix} \mathbf{L}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \quad \mathbf{U} \equiv \begin{bmatrix} \mathbf{U}_1 & \mathbf{L}_1^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \quad \text{and} \quad \mathbf{A} \equiv \mathbf{L}\mathbf{U}.$$

Here $\mathbf{I}$ is the identity matrix and $\mathbf{0}$ is the matrix of all zeros. Both are $(n+1)\times(n+1)$.
Show that $\mathbf{L}$ is lower triangular, $\mathbf{U}$ is upper triangular, and $\mathbf{L}$ and $\mathbf{U}$ are the L and U factors of $\mathbf{A}$ as obtained by Gauss elimination with partial pivoting.

(d) Show $\|\mathbf{A}\|_M = 1$, $\|\mathbf{A}\|_\infty = 3$, $\|\mathbf{L}\|_\infty = n$, $\|\mathbf{U}\|_\infty = \frac{1}{2}n$, $\|\mathbf{U}\|_M = 1$, and $\| \, |\mathbf{L}| \, |\mathbf{U}| \, \|_\infty = \frac{1}{2}n^2$.

Cholesky factorisation.

The following theorem describes estimates for the rounding errors in the Cholesky factorisation. The proof of this theorem is similar to the proof of Theorem 2.3 in Exercise 2.16: we will not give details here.

Theorem 2.5 *Let $\mathbf{A}$ be a complex $n \times n$ positive definite matrix.*
Let $\widehat{\mathbf{C}}$ be the Cholesky factor as computed by Cholesky factorisation. Then

$$\mathbf{A} + \Delta_{CC} = \widehat{\mathbf{C}}^*\widehat{\mathbf{C}} \quad \textit{for some } \Delta_{CC} \textit{ for which} \quad |\Delta_{CC}| \le n\mathbf{u}\,|\widehat{\mathbf{C}}|^*\,|\widehat{\mathbf{C}}|.$$

Let $\widehat{\mathbf{x}}$ be the solution of the system $\mathbf{A}\mathbf{x} = \mathbf{b}$ as computed by solving $\widehat{\mathbf{C}}^*\mathbf{y} = \mathbf{b}$ and $\widehat{\mathbf{C}}\mathbf{x} = \widehat{\mathbf{y}}$. Then

$$(\mathbf{A}+\Delta)\widehat{\mathbf{x}} = \mathbf{x} \quad \textit{for some } \Delta \textit{ for which} \quad |\Delta| \le 3n\mathbf{u}\,|\widehat{\mathbf{C}}|^*\,|\widehat{\mathbf{C}}|.$$

Although the estimates are similar to the ones in Theorem 2.3, they allow to prove that the Cholesky factorisation allows a stable computation of the solution of systems involving positive definite systems, as we will see in Exercise 2.21.

Corollary 2.6 *For the backward error in the solution of a complex positive definite system obtained with Cholesky's decomposition we have that*

$$\frac{\|\Delta\|_2}{\|\mathbf{A}\|_2} \le 3\,\mathbf{u}\,n^2. \tag{2.10}$$

Exercise 2.21.

(a) Prove that $\| \, |\mathbf{C}|^* \, |\mathbf{C}| \, \|_2 \le n \, \min(\|\mathbf{A}\|_2, \| \, |\mathbf{A}| \, \|_2) = n \, \| \, |\mathbf{A}| \, \|_2$.

(b) Conclude that solving positive definite systems with Cholesky factorisation is stable (with respect to the 2-norm, with condition number n, cf., Exercise 2.17).

The Estimate (2.10) can, with some effort, be improved to $\|\Delta\|_2/\|\mathbf{A}\|_2 \le 2.5\,\mathbf{u}\,n\sqrt{n}$.