Report

Nunzia Cerrato, Giuseppe Catalano

Problem 2

Consider the $n \times n$ Wilkinson matrix

$$W_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 1 \\ -1 & 1 & 0 & \cdots & 1 \\ -1 & -1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & -1 & 1 \end{bmatrix} \tag{0.1}$$

(1) We are interested to compute (by hand) the LU factorization of W_5, therefore to compute two $n \times n$ matrices, L_5 and U_5, that are, respectively, a unit lower triangular matrix and an upper triangular matrix that satisfy the identity $W_5 = L_5 U_5$. We start writing the expression of W_5:

$$W_5 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix}, \tag{0.2}$$

now we compute $\mathbf{m}_1$:

$$\mathbf{m}_1 = \begin{bmatrix} 0 \\ w_{21}/w_{11} \\ w_{31}/w_{11} \\ w_{41}/w_{11} \\ w_{51}/w_{11} \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix}. \tag{0.3}$$

Defining $\mathbf{e}_i$ as the vectors with 1 in the $i-th$ element and 0 otherwise, we can compute:

$$M_1 = \mathcal{I}_5 - \mathbf{m}_1 \mathbf{e}_1^T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{0.4}$$

Using the expression of M_1, we can compute $W_5^{(1)}$:

$$W_5^{(1)} = M_1 W_5 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & -1 & 1 & 0 & 2 \\ 0 & -1 & -1 & 1 & 2 \\ 0 & -1 & -1 & -1 & 2 \end{bmatrix}. \tag{0.5}$$

The second iteration proceeds in a similar way:

$$\mathbf{m}_2 = \begin{bmatrix} 0 \\ 0 \\ w_{32}^{(1)}/w_{22}^{(1)} \\ w_{42}^{(1)}/w_{22}^{(1)} \\ w_{52}^{(1)}/w_{22}^{(1)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -1 \\ -1 \\ -1 \end{bmatrix}, \tag{0.6}$$

$$M_2 = \mathcal{I}_5 - \mathbf{m}_2 \mathbf{e}_2^T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix}. \tag{0.7}$$

Using the expression of M_2, we can compute $W_5^{(2)}$:

$$W_5^{(2)} = M_2 W_5^{(1)} = M_2 M_1 W_5 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 4 \\ 0 & 0 & -1 & 1 & 4 \\ 0 & 0 & -1 & -1 & 4 \end{bmatrix}. \tag{0.8}$$

Now we start the third iteration:

$$\mathbf{m}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ w_{43}^{(2)}/w_{33}^{(2)} \\ w_{53}^{(2)}/w_{33}^{(2)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1 \\ -1 \end{bmatrix}, \tag{0.9}$$

$$M_3 = \mathcal{I}_5 - \mathbf{m}_3\mathbf{e}_3^T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix}. \tag{0.10}$$

Using the expression of M_3, we can compute $W_5^{(3)}$:

$$W_5^{(3)} = M_3 W_5^{(2)} = M_3 M_2 M_1 W_5 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 & 8 \\ 0 & 0 & 0 & -1 & 8 \end{bmatrix}. \tag{0.11}$$

Similarly, we can perform the last iteration:

$$\mathbf{m}_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ w_{54}^{(2)}/w_{44}^{(2)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -1 \end{bmatrix}, \tag{0.12}$$

$$M_4 = \mathcal{I}_5 - \mathbf{m}_4\mathbf{e}_4^T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}. \tag{0.13}$$

Using the expression of M_4, we can compute $W_5^{(4)}$:

$$W_5^{(4)} = M_4 W_5^{(3)} = M_4 M_3 M_2 M_1 W_5 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 & 8 \\ 0 & 0 & 0 & 0 & 16 \end{bmatrix} = U_5. \tag{0.14}$$

Since this last matrix is upper triangular, we call it U_5 and define $L_5^{-1} := M_4 M_3 M_2 M_1$, such that $L_5^{-1} W_5 = U_5$. We can get L_5 from $L_5 = M_1^{-1} M_2^{-1} M_3^{-1} M_4^{-1}$ and knowing that $M_i^{-1} = \mathcal{I}_5 + \mathbf{m}_i \mathbf{e}_i^T$:

$$L_5 = M_1^{-1} M_2^{-1} M_3^{-1} M_4^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ -1 & -1 & -1 & 1 & 0 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix}. \tag{0.15}$$

(2) It is possible to guess the LU factorization of $W_n = L_n U_n$:

$$L_n = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -1 & 1 & \ddots & & \vdots \\ \vdots & \ddots & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ -1 & \cdots & \cdots & -1 & 1 \end{bmatrix}, \quad U_n = \begin{bmatrix} 1 & 0 & \cdots & 0 & 2^0 \\ 0 & 1 & \ddots & \vdots & 2^1 \\ \vdots & \ddots & \ddots & 0 & 2^2 \\ \vdots & & \ddots & 1 & \vdots \\ 0 & \cdots & \cdots & 0 & 2^{n-1} \end{bmatrix} \tag{0.16}$$

(3) In the following, we report the function that generates the $n \times n$ Wilkinson matrix.
*** INSERIRE CODICE (funzione wilkin(n)) ***

(4-5-6) In the following, we report the code that performs the numerical experiment for each $n = 2, \dots, 60$.

*** INSERIRE CODICE (funzione check_when_lufact_W_fails) ***

We have seen that the largest value of n for which $W_n\mathbf{x} = \mathbf{b}$ can be solved accurately is 54, that means that for $n = 55$ the program returns an inaccurate value for the solution $\mathbf{x}$. In particular, instead of computing the value $\mathbf{x} = \mathbf{e} = [1, \dots, 1]^T$, it computes $\tilde{\mathbf{x}} = [1, \dots, 1, 0, 1]^T$. In other words we have:

$$\tilde{\mathbf{x}}_{54} = 0 \neq \mathbf{e}_{54} = 1. \tag{0.17}$$

In order to understand the motivation uder this behavior, we verified that the matrices L_{55} and U_{55} were computed accurately. In the following is reported the code that verifies this computation.

*** INSERIRE CODICE (funzioni expected_LU_wilkin, compute_error_lufact_W) ***

Once verified that the matrices L_{55} and U_{55} are correct, we know that the problem must be in the calculation of the forward and backward substitution. We recall that having performed the LU factorization of the matrix W_{55} allows us to solve the system $W_{55}\mathbf{x} = \mathbf{b}_{55}$ by solving (in order) two linear systems with forward and backward substitutions:

$$\begin{cases} L\mathbf{y} = \mathbf{b} \\ U\mathbf{x} = \mathbf{y} \end{cases} \tag{0.18}$$

where we named $L_{55} = L$, $U_{55} = U$ and $\mathbf{b}_{55} = \mathbf{b}$ for clarity. It can be verified that the analytical solution to the first system is $\mathbf{y} = [2^0 + 1, 2^1 + 1, \dots, 2^{n-2} + 1, 2^{n-1}]$. At this point, a very important observation arises: if we consider $\mathbf{y}_{54} = 2^{53} + 1$, we may notice that, when the sum $2^{53} + 1$ is performed in double precision, the result is 2^{53}. This happens because 1 is less than the machine precision associated to the number 2^{53} in double precision:

$$1 < 2^{53} \cdot \varepsilon \simeq 2^{53} \cdot 2.22 \cdot 10^{-16} \simeq 0.9 \cdot 10^{16} \cdot 2.22 \cdot 10^{-16} \simeq 2. \tag{0.19}$$

After having computed $\mathbf{y}$, we can compute the solution of the second system starting from the bottom:

$$\mathbf{x}_{55} = \mathbf{y}_{55}/U_{55,55} = \frac{2^{54}}{2^{54}} = 1, \tag{0.20}$$

so far so good. Now we update the vector $\mathbf{y}$ as follows:

$$\mathbf{y}^{(1)} = \mathbf{y} - \mathbf{x}_{55}\mathbf{u}_{55}, \tag{0.21}$$

where $\mathbf{u}_{55}$ is the $55 - th$ and last column of U. The $54 - th$ component of $\mathbf{y}^{(1)}$ will be:

$$[\mathbf{y}^{(1)}]_{54} = [\mathbf{y}]_{54} - \mathbf{x}_{55}U_{54,55} = 2^{53} + 1 - 2^{53}, \tag{0.22}$$

but, since in double precision we have $2^{53} + 1 = 2^{53}$, the returned value of $[\mathbf{y}^{(1)}]_{54}$ will be 0 and not 1. If we execute the same algorithm for bigger values of n, the number of elements of the solution vector that will be miscalculated will grow, starting from the penultimate element of the vector $\mathbf{x}$. It is worth noting that the last element of the solution is not affected by this kind of error because it is computed via the operation:

$$\mathbf{x}_n = \mathbf{y}_n/U_{n,n} = \frac{2^{n-1}}{2^{n-1}} \tag{0.23}$$

that does not lead to catastrophical cancellations. These considerations imply that the GEPP algorithm is not backward stable when the input is a Wilkinson matrix. However, it is important to say that this is a very artificial example where the growth factor γ assumes the maximum possible value, i.e. $\gamma = 2^{n-1}$. However, in most cases, the GEPP algorithm is backward stable.
The just described behavior can be observed from the execution of the function reported here.

*** INSERIRE CODICE ***

Problem 3

Suppose that $A \in \mathbb{R}^{n\times n}$ is a nonsingular matrix and the LU factorization of A exists and has been computed. Consider two given vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, we can define the matrix $\tilde{A} = A + \mathbf{u}\mathbf{v}^T$

(1a) Prove that $\tilde{A}$ is nonsingular if and only if $\mathbf{v}^T A^{-1}\mathbf{u} \neq 1$.

Proof: We start proving that $\det(\tilde{A}) \neq 0$ implies that $\mathbf{v}^T A^{-1}\mathbf{u} \neq 1$. We can choose an orthonormal basis $\mathcal{B} = \{\mathbf{e}_i\}_{i=1,\dots,n}$ of $\mathbb{R}^n$ such that $\mathbf{u} = \alpha_1\mathbf{e}_1$ and $\mathbf{v} = \beta_1\mathbf{e}_1 + \beta_2\mathbf{e}_2$ with $\alpha_1, \beta_1, \beta_2 \in \mathbb{R}$. We can represent the matrix A with respect to the basis $\mathcal{B}$, denoting with $a_{ij} = \mathbf{e}_i^T A\mathbf{e}_j$ the element of the $i-th$ row and $j-th$ column of the matrix A written in the basis $\mathcal{B}$. We can represent the matrix $\mathbf{u}\mathbf{v}^T$ with respect to the basis $\mathcal{B}$, obtaining $\mathbf{u}\mathbf{v}^T = \alpha_1\beta_1\mathbf{e}_1\mathbf{e}_1^T + \alpha_1\beta_2\mathbf{e}_1\mathbf{e}_2^T$. Knowing that we can compute the determinant of a matrix M $n\times n$ using the formula:

$$\det(M) = \sum_{j=1}^{n} m_{ij}C_{ij}(M), \tag{0.24}$$

where C_{ij} is the cofactor of the element (i,j) of the matrix M, the determinant of $\tilde{A}$ is:

$$\det(\tilde{A}) = \det(A) + \alpha_1\beta_1 C_{11}(A) + \alpha_1\beta_2 C_{12}(A). \tag{0.25}$$

Since we know that $\det(\tilde{A}) \neq 0$, we can write:

$$\det(A) + \alpha_1\beta_1 C_{11}(A) + \alpha_1\beta_2 C_{12}(A) \neq 0, \tag{0.26}$$

and, therefore:

$$\alpha_1\beta_1\frac{C_{11}(A)}{\det(A)} + \alpha_1\beta_2\frac{C_{12}(A)}{\det(A)} \neq -1, \tag{0.27}$$

where we divided for $\det(A)$ both sides of the equation, knowing that $\det(A) \neq 0$. At this point, it is straightforward to verify that

$$\mathbf{v}^T A^{-1}\mathbf{u} = \alpha_1\beta_1\frac{C_{11}(A)}{\det(A)} + \alpha_1\beta_2\frac{C_{12}(A)}{\det(A)} \tag{0.28}$$

writing $\mathbf{u}$ and $\mathbf{v}$ in terms of $\mathbf{e}_1$ and $\mathbf{e}_2$ and $A^{-1} = \frac{1}{\det(A)}(cof(A))^T$, where $cof(A)$ is the matrix of cofactors of A. This proves that:

$$\mathbf{v}^T A^{-1}\mathbf{u} \neq -1. \tag{0.29}$$

In order to prove the converse implication, we consider again the expression for the determinant of $\tilde{A}$:

$$\det(\tilde{A}) = \det(A) + \alpha_1\beta_1 C_{11}(A) + \alpha_1\beta_2 C_{12}(A) = \det(A)\left(1 + \alpha_1\beta_1\frac{C_{11}(A)}{\det(A)} + \alpha_1\beta_2\frac{C_{12}(A)}{\det(A)}\right). \tag{0.30}$$

Here we can recognize the expression of $\mathbf{v}^T A^{-1}\mathbf{u}$, obtaining:

$$\det(\tilde{A}) = \det(A)(1 + \mathbf{v}^T A^{-1}\mathbf{u}). \tag{0.31}$$

Now, knowing that $\det(A) \neq 0$ and $\mathbf{v}^T A^{-1}\mathbf{u} \neq -1$, we obtain

$$\det(\tilde{A}) \neq 0 \tag{0.32}$$

and this concludes the proof.

(1b) Show that:

$$\tilde{A}^{-1} = A^{-1} - \alpha A^{-1}\mathbf{u}\mathbf{v}^T A^{-1}, \quad \text{where } \alpha = \frac{1}{\mathbf{v}^T A^{-1}\mathbf{u} + 1}. \tag{0.33}$$

Proof: We start noticing that the last expression is well defined since $\tilde{A}$ invertible implies $\mathbf{v}^T A^{-1}\mathbf{u} + 1 \neq 0$. Now we can manipulate the (0.33) multiplying both sides to the left for A and to the right for $\tilde{A}$, obtaining:

$$\begin{aligned} A &= \tilde{A} - \alpha\mathbf{u}\mathbf{v}^T A^{-1}\tilde{A} \\ &= A + \mathbf{u}\mathbf{v}^T - \alpha\mathbf{u}\mathbf{v}^T A^{-1}(A + \mathbf{u}\mathbf{v}^T) \\ &= A + (1-\alpha)\mathbf{u}\mathbf{v}^T - \alpha\mathbf{u}\mathbf{v}^T A^{-1}\mathbf{u}\mathbf{v}^T. \end{aligned} \tag{0.34}$$

Subtracting A from each side and dividing both sides for α (that is nonzero $\forall \mathbf{v}^T A^{-1}\mathbf{u} \in \mathbb{R}$) we obtain:

$$\mathbf{u}\mathbf{v}^T A^{-1}\mathbf{u}\mathbf{v}^T = \mathbf{u}\mathbf{v}^T(\alpha^{-1} - 1). \tag{0.35}$$

Finally, since $\mathbf{v}^T A^{-1}\mathbf{u} = \alpha^{-1} - 1$, the identity (0.35) is verified and this concludes the proof.

(1c) Assuming that LU factorization of A is already available, describe an $\mathcal{O}(n^2)$ algorithm to solve $\tilde{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$ for any right-hand side $\tilde{\mathbf{b}}$.
Supposing that $\tilde{A}$ is invertible, we can write the solution $\tilde{\mathbf{x}}$ using the Sherman-Morrison formula for $\tilde{A}$:

$$\tilde{\mathbf{x}} = \tilde{A}^{-1}\tilde{\mathbf{b}} = A^{-1}\tilde{\mathbf{b}} - \frac{A^{-1}\mathbf{u}\mathbf{v}^T A^{-1}\tilde{\mathbf{b}}}{\mathbf{v}^T A^{-1}\mathbf{u} + 1} \tag{0.36}$$

Algorithm:

- Compute $\mathbf{x}$ s.t. $A\mathbf{x} = \tilde{\mathbf{b}}$ and $\mathbf{y}$ s.t. $A\mathbf{y} = \mathbf{u}$ using backward and forward substitutions. This requires $\mathcal{O}(n^2)$ operations.
- Compute $\gamma = \frac{\mathbf{v}^T\mathbf{x}}{\mathbf{v}^T\mathbf{y}+1}$. This requires $\mathcal{O}(n)$ operations.
- Compute $\tilde{\mathbf{x}} = \mathbf{x} - \gamma\mathbf{y}$. This requires $\mathcal{O}(n)$ operations.

(2) Assuming again that the LU factorization of A existes and has been computed, describe an efficient algorithm for solving the *bordered system*

$$\begin{bmatrix} A & \mathbf{u} \\ \mathbf{v}^T & \beta \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ z \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ c \end{bmatrix}, \tag{0.37}$$

where z is unknown and β and c are given scalars. When does this system have a unique solution?
Solution:
Putting

$$A' = \begin{bmatrix} A & \mathbf{u} \\ \mathbf{v}^T & \beta \end{bmatrix}, \mathbf{x}' = \begin{bmatrix} \mathbf{x} \\ z \end{bmatrix}, \mathbf{b}' = \begin{bmatrix} \mathbf{b} \\ c \end{bmatrix}, \tag{0.38}$$

the system above rewrites as:

$$A'\mathbf{x}' = \mathbf{b}'. \tag{0.39}$$

The LU factorization of the matrix $A' = L'U'$ exists and the matrices L' and U' take the following form:

$$L' = \begin{bmatrix} L & \mathbf{0} \\ \mathbf{f}^T & 1 \end{bmatrix}, U' = \begin{bmatrix} U & \mathbf{g} \\ \mathbf{0} & \gamma \end{bmatrix}. \tag{0.40}$$

In order to get the values of $\mathbf{f}, \mathbf{g} \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}$, we impose $A' = L'U'$, obtaining the following system:

$$\begin{cases} L\mathbf{g} = \mathbf{u} \\ U^T\mathbf{f} = \mathbf{v} \\ \mathbf{f}^T\mathbf{g} + \gamma = \beta \end{cases}. \tag{0.41}$$

Here we can find $\mathbf{f}$ and $\mathbf{g}$ with forward substitutions with $\mathcal{O}(n^2)$ operations. Therefore, we can rewrite the last equation as:

$$\gamma = \beta - \mathbf{v}^T A^{-1}\mathbf{u}. \tag{0.42}$$

In order to impose that the bordered system has a unique solution, we have to require that $\det(A') \neq 0$, that is true if and only if all the diagonal elements of U' are nonzero and, given that $\det(A) \neq 0$, this means requiring that $\gamma \neq 0$. Therefore, the condition for the uniquenes of the solution becomes:

$$\gamma = \beta - \mathbf{v}^T A^{-1}\mathbf{u} \neq 0 \Rightarrow \quad \mathbf{v}^T A^{-1}\mathbf{u} \neq \beta. \tag{0.43}$$