

# Numerical Linear Algebra Homework Project 1: Solution of Linear Systems

Catalano Giuseppe, Cerrato Nunzia

## 1 Problem 1

In the first problem it is required to write a function that computes the LU factorization of a non-singular matrix  $A$  without pivoting. The full code has been written using Python as programming language and it is available at the following link on GitHub [https://github.com/nunziacerrato/Numerical\\_Analysis\\_Optimization](https://github.com/nunziacerrato/Numerical_Analysis_Optimization), where there is also a link to the online documentation of the library <sup>1</sup>. In the following it is reported the function `lufact(A)` that can be found in the library named `Project_1.py`

```
1 def lufact(A):
2     r''' This function computes the LU factorization of a square matrix A
        ↳ without pivoting, giving as output the matrices L and U and the growth
        ↳ factor g, here defined as :math:\frac{\max_{ij} (|L|+|U|)_{ij}}{\max_{ij} (|A|)_{ij}}.
3
4     Paramters:
5     -----
6     A : ndarray
7         Input matrix of dimension :math:(n \times n)
8
9     Returns
10    -----
11    L : ndarray
12        Unit lower triangular matrix
13    U : ndarray
14        Upper triangular matrix
15    g : float
16        Growth factor
17    '''
18
19    # Compute the dimension of the input square matrix
20    dim = A.shape
21    n = dim[0]
22
23    # Define the chosen precision
24    precision = np.finfo(float).eps/2
25
26    # Check that the input matrix is a square matrix
27    assert (dim[0] == dim[1]), "The input matrix is not a square matrix"
```

---

<sup>1</sup>The syntax used in the docstrings of the Python functions is necessary to obtain clear latex formatting of the formulas in the online documentation.

```

28
29 # Check if the determinant of the input matrix is less than the chosen
    ↪ precision
30 if np.abs(np.linalg.det(A)) < precision:
31     logging.warning("The determinant of the input matrix is less than the
        ↪ chosen precisionr")
32 # Check if the hypothesis of the LU factorization theorem hold
33 for k in range(n):
34     if np.abs(np.linalg.det(A[:k+1,:k+1])) < precision:
35         logging.warning(f'The {k}-th principal minor is less than the chosen
            ↪ precision')
36
37 # Create a copy of the input matrix to be modified in order to obtain the
    ↪ matrices L and U
38 B = np.copy(A)
39 for k in range(0,n-1):
40     for i in range(k+1,n):
41         B_kk = B[k,k]
42         # Check if there is a division by a quantity smaller than the chosen
            ↪ precision
43         if np.abs(B_kk) < precision:
44             raise ValueError('Division by a quantity smaller than the chosen
                ↪ precision - B_kk = {B_kk}')
45         B[i,k] = B[i,k]/B_kk
46     for j in range(k+1,n):
47         for l in range(k+1,n):
48             B[l,j] = B[l,j] - B[l,k]*B[k,j]
49
50 # Extract the matrices L and U from B using, respectively, a strictly lower
    ↪ triangular mask and an upper triangular mask.
51 L = np.tril(B,k=-1) + np.eye(n) # Add the Id matrix in order for L to be
    ↪ unit lower triangular
52 U = np.triu(B,k=0)
53
54 # Compute the growth factor
55 LU_abs = np.abs(L)@ np.abs(U)
56 g = np.amax(LU_abs)/np.amax(np.abs(A))
57
58 return L, U, g

```

We included tests in this function in order to ensure that the input matrix has the correct characteristics to perform the LU factorization. The rationale used to insert such tests is explained in the following. We added *warnings*, using the `logging` Python library, for all the checks that, if not satisfied, do not necessarily break the LU factorization, so the flow of the code does not interrupt. For what concern most significant errors, we inserted an *assertion* if the input matrix is not square, so that the flow of the code is interrupted, and a *ValueError* if a division by a quantity smaller than the chosen precision occurs. In this last case, if the LU factorization fails, the exception is handled in the main program and this information is stored in a counter (see subsection 1.1), that is saved in an Excel file for all the types of matrices and the dimensions considered. The counter will be updated each time a failure occurs.

In analyzing the results regarding the correctness of the computational LU factorization for

the chosen input matrices, it is required to study the trend of the growth factor and the relative backward error with respect to the dimension of the considered matrices. While the growth factor is computed within the function `lufact(A)`, the relative backward error is computed separately. The code that computes this value is reported below.

```

1 def relative_backward_error(A,L,U):
2     r''' This function computes the relative backward error of the LU
3     ↪ factorization, defined as :math:\frac{\|A - LU\|}{\|A\| \|L\| \|U\|}
4     ↪ \frac{\|A - LU\|}{\|A\| \|L\| \|U\|}
5
6     Parameters
7     -----
8     A : ndarray
9     Input matrix
10    L : ndarray
11    Unit lower triangular matrix, obtained from the LU factorization of the
12    ↪ input matrix A.
13    U : ndarray
14    Upper triangular matrix, obtained from the LU factorization of the
15    ↪ input matrix A.
16
17    Returns
18    -----
19    out : float
20    Relative backward error
21    '''
22
23    return np.linalg.norm(A - L @ U, ord=np.inf)/np.linalg.norm(A, ord=np.inf)

```

Note that the symbol `@` stands for the operator that performs the matrix multiplication.

Once having constructed these functions, we built the dataset of matrices, on which apply the LU factorization, by creating a function that takes two values as input, namely the number of matrices of each type and their dimension, and gives as output a dictionary containing all the sampled matrices of each type. In particular, we considered random matrices, unitary matrices, Hermitian matrices, positive definite matrices, and diagonally dominant matrices. For the first type of matrices, we distinguished matrices whose entries are real and uniformly sampled in the range  $[0, 1)$  and matrices whose entries are independent complex and normally distributed. To sample unitary and Hermitian matrices we considered, instead, ensembles of random matrices using the `tenpy` library in Python. Positive definite matrices are sampled using the `qutip` library in Python, by considering trace-one matrices of the form  $A^*A$ , where  $A^*$  stands for the Hermitian conjugate of the matrix  $A$ , sampled as a matrix whose entries are independent complex and normally distributed. Finally, we define a function to sample diagonally dominant matrices, which is reported below.

In this last case the idea is to generate random matrices whose entries are normally distributed and substitute each diagonal element with a new one obtained by summing the absolute values of all the elements in the corresponding row (including itself). The sign of each diagonal element is chosen at random, by generating  $n$  numbers in the interval  $[0, 1)$  and applying the sign function to these numbers, shifted by 0.5.

```

1 def diagonally_dominant_matrix(n):
2     ''' This function returns a diagonally dominant matrix of dimension
3     → :math:`(n \times n)`, whose non-diagonal entries are normally
4     → distributed.
5
6     Parameters
7     -----
8     n : int
9     Dimension of the output matrix
10
11     Returns
12     -----
13     out : ndarray
14     Diagonally dominant matrix
15
16     '''
17     # The following steps are made to decide the sign of the diagonal element
18     → of the output matrix
19     # Obtain n random numbers in [0,1) and apply the sign function to this
20     → values, shifted by 0.5
21     diag_sign = np.random.rand(n)
22     diag_sign = np.sign(diag_sign - 0.5)
23     diag_sign[diag_sign == 0] = 1 # Set to 1 the (very unlikely) values equal
24     → to 0
25
26     # Obtain a matrix of dimension (n × n) whose entries are normally
27     → distributed
28     M = np.random.normal(loc=0.0, scale=1.0, size=(n,n))
29     # Substitute all the diagonal elements in this matrix with the sum of the
30     → absolute values of all the elements in the corresponding row (including
31     → itself)
32     for i in range(n):
33         M[i,i] = sum(np.abs(M[i,:])) * diag_sign[i]
34
35     return M

```

The function that creates the dataset contains all the types of matrices listed before. Since that part of the code is trivial, we have decided to not report it here for brevity (it can be found [here](#) in the online repository on GitHub). We do a brief description of how we organize this function. First and foremost, the function `create_dataset` takes as input the number of matrices of each type and their relative dimension and gives as output a dictionary whose keys represent the different types of matrices considered and whose values are 3-dimensional arrays, where the first index cycles on the number of matrices considered. Moreover, we fixed the seed in order to have reproducibility of the results (`tenpy` and `qutip` libraries work using `numpy` in the background, so this choice for the seed holds for all the sampled matrices).

## 1.1 Main program

In the main program, after having imported all the necessary Python libraries, we created the dataset and computed the LU factorization for all the types of matrices considered, while also taking into account different dimensions. All the data are saved in Excel files. We also created a DataFrame to store all the failures of the algorithm, where column names represent

the different types of matrices considered, row indices represent the progressive dimension of the matrices, while elements of the DataFrame represent the total number of failures of the LU factorization for a certain matrix type of a given dimension. In this way, it is possible to clearly see for which types of matrices the algorithm fails and what is the number of failures associated with each matrix dimension.

```

1  # Define global parameters
2  num_matr = 500
3  dim_matr_max = 50
4  common_path = "Project_1"
5
6  keys = create_dataset(1,2).keys()
7
8  # Define a DataFrame to store all the failures of the LU factorization
   → divided by matrix types.
9  df_fails = pd.DataFrame(0, columns = keys, index = range(2,dim_matr_max+1))
10
11 # Cycle on the different dimensions considered
12 for dim_matr in range(2,dim_matr_max+1):
13     logging.info(f'Dimension = {dim_matr}')
14
15 # Create the dataset
16 dataset = create_dataset(num_matr, dim_matr)
17
18 # Create DataFrames in which the growth factor and the relative backward
   → error are stored
19 df_g = pd.DataFrame(columns = keys)
20 df_rel_back_err = pd.DataFrame(columns = keys)
21
22 # Cycle on the different types of matrices considered
23 for matrix_type in keys:
24
25     # Cycle on the number of matrices of each type
26     for i in range(num_matr):
27         # Select the matrix and compute the LU factorization, the growth factor
           → and the relative backward error
28         A = dataset[matrix_type][i,:,:]
29         try:
30             L, U, df_g.at[i,matrix_type] = lufact(A)
31             df_rel_back_err.at[i,matrix_type] = relative_backward_error(A, L, U)
32         except ValueError:
33             df_fails.at[dim_matr,matrix_type] = df_fails.at[dim_matr,matrix_type] +
               → 1
34
35 # Save the growth factor and the relative backward error in Excel files
36 writer = pd.ExcelWriter(f'{common_path}\\Data\\'
37 f'Statistics_for_{num_matr}_matrices_of_dim_{dim_matr}.xlsx')
38 df_g.to_excel(writer, 'growth_factor', index = False)
39 df_rel_back_err.to_excel(writer, 'rel_back_err', index = False)
40 writer.save()
41
42 # Save the failures of the LU factorization in an Excel file

```

```

43 writer = pd.ExcelWriter(f'{common_path}\\Data\\'
44 f'Failures_LUfact_for_{num_matr}_matrices.xlsx')
45 df_fails.to_excel(writer, 'Fails', index = False)
46 writer.save()

```

Choosing this approach, once having created the dataset and computed the quantities of interest, we have all the data stored. Different scripts are dedicated to statistical analysis. In particular, a program is dedicated to computing the minimum and the maximum value of the growth factor and of the relative backward error, as well as the mean value and the standard deviation of such quantities, and plotting them with respect to the dimension of the input matrix (see program `Project_1_Statistics.py`, where there is also a part of the code to obtain summary boxplots for the growth factor and the relative backward error). Moreover, in the file `Project_1_Distributions.py` we obtain distributions of the two latter quantities for a fixed dimension of the input matrix, while in the program `Project_1_Hilbert.py` we analyze such quantities for the Hilbert matrices. For the sake of simplicity, we do not report here such part of the code, as it is of no particular relevance. However, it can be found [here](#) in the online repository on GitHub.

## 1.2 Results

We considered 500 samplings for all the types of square matrices investigated, whose dimension varies from 2 to 50.

**Random Matrices** We generated two different types of random matrices, namely real matrices with entries sampled uniformly in the interval  $[0, 1)$ , which we simply call *random matrices*, and complex matrices with entries of the form  $a+ib$ , where  $a$  and  $b$  are independently sampled from the normal distribution. The second ones are called *Ginibre matrices*. We made this choice to construct a dataset as representative as possible, trying to take into consideration different types of distributions for the entries of the matrices. We report in the following the plot of the characteristic values of the growth factor  $\gamma$  and the relative backward error  $\delta$  as a function of the dimension  $N$  of the matrix.

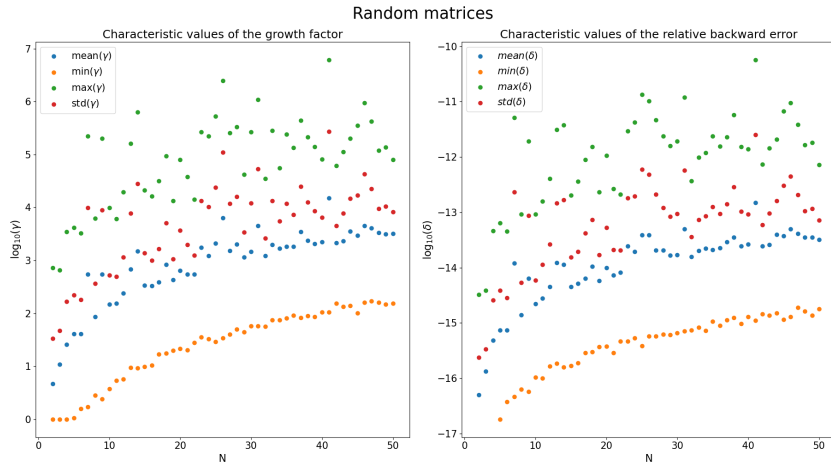


Figure 1: Scatterplot of the characteristic values of  $\gamma$  and  $\delta$  as a function of  $N$  for real random matrices. Logarithmic scale on the ordinate axis.

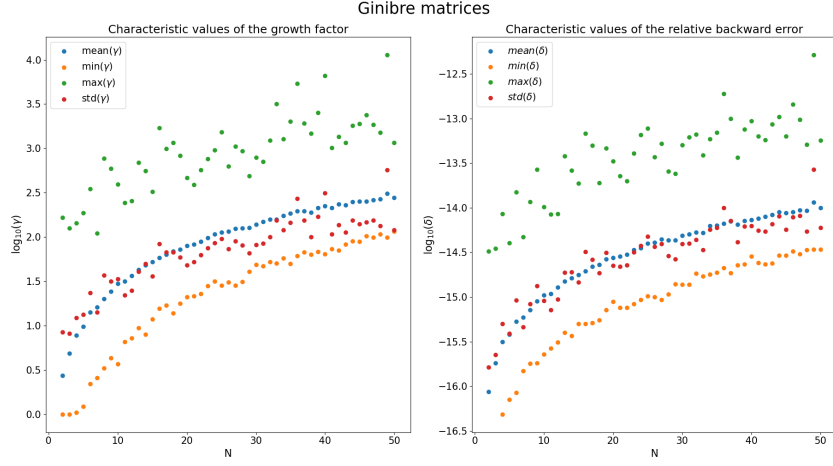


Figure 2: Scatterplot of the characteristic values of  $\gamma$  and  $\delta$  as a function of  $N$  for Ginibre matrices. Logarithmic scale on the ordinate axis.

As can be seen from these plots, the characteristic values of the two considered quantities increase with the dimension of the input matrix for both types of matrices taken into account. This might be due to the fact that, when the dimension of the input matrix increases, the algorithm that performs the LU factorization requires more steps, resulting in a propagation of numerical errors. It is interesting to also consider, for a fixed dimension, the distributions of the growth factor and the relative backward error for both cases. We have chosen to report here the histograms obtained for  $N = 25$ , together with the corresponding boxplots, where the outliers have been removed for clarity of visualization. We have chosen this point as an intermediate representative dimension; similar considerations also hold in the other cases.

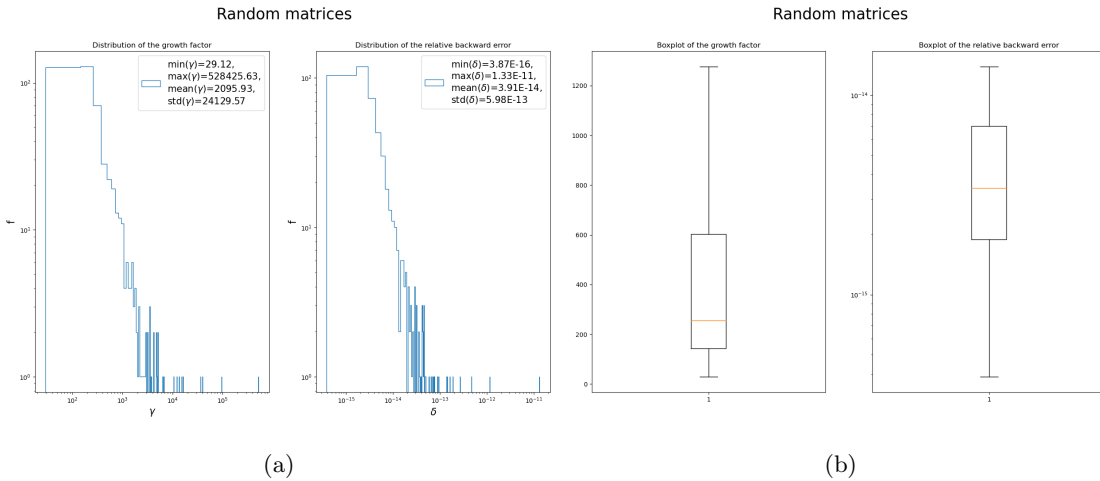


Figure 3: Histograms (panel (a)) and boxplots (panel (b)) of  $\gamma$  and  $\delta$  considering real random matrices of dimension  $25 \times 25$ . Logarithmic scale is reported on both axis of the histograms as well as on the ordinate axes of the relative backward error boxplot.

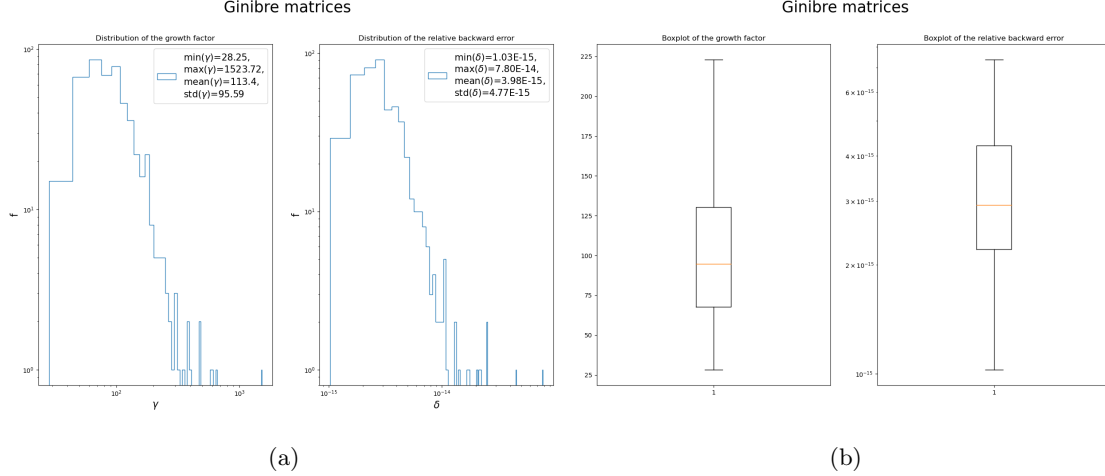


Figure 4: Histograms (panel (a)) and boxplots (panel (b)) of  $\gamma$  and  $\delta$  considering Ginibre matrices of dimension  $25 \times 25$ . Logarithmic scale is reported on both axis of the histograms as well as on the ordinate axes of the relative backward error boxplot.

It can be observed that both the growth factor and the relative backward error have a wide range of variation, if one also considers outliers, in the case of real random matrices, while these ranges are smaller in the case of Ginibre matrices. This is, in general, an indication of the fact that both quantities strongly depend on the input matrix, even if more careful considerations can be made to distinguish the two types of matrices. In fact, taking into account, for example, the relative backward error and considering also the outliers, one can note that for real random matrices, whose entries are uniformly distributed in  $[0, 1]$ , this quantity ranges approximatively from  $10^{-16}$  to  $10^{-11}$ . For what concern the Ginibre matrices, the same quantity ranges approximatively from  $10^{-15}$  to  $10^{-13}$ . It is worth noting that the presence of outliers is more marked in the first case. This difference in the variation range might be due to the fact that when considering the normal distribution to sample the entries of a random matrix, it is more likely to have matrix elements that are centered in the neighborhood of the mean value, resulting in an ensemble of more homogeneous matrices. This, as a consequence, might reflect in the obtained values of relative backward errors, when computing the LU factorization, which may be similar and with a smaller range of variation. In the other case, when considering the uniform distribution to sample the entries of a random matrix, the resulting ensemble of matrices may be less homogeneous when compared to the previous one, and this might result in a wider range of variation of the quantity under examination. However, even if in both cases the relative backward errors grow with the dimension of the matrix, their values are small enough and, at least for the mean value, not that far from the machine epsilon, that is of order  $10^{-16}$ . For what concern the growth factor, even if it grows with the dimension of the input matrix, its growth is reasonable as not exponential. Given these considerations, the algorithm can be considered backward stable for these types of matrices.

**Unitary matrices** We considered in our dataset the ensemble of unitary matrices, namely matrices  $U$  such that  $U^* = U^{-1}$ , where  $*$  stands for the conjugate transpose. These matrices have been sampled using the `tenpy` Python library and they are also called *CUE matrices*, as they are sampled from the circular ensemble. As in the previous case, we report in the following the scatterplot of the growth factor  $\gamma$  and the relative backward error  $\delta$  as a function of the dimension  $N$  of the matrix.



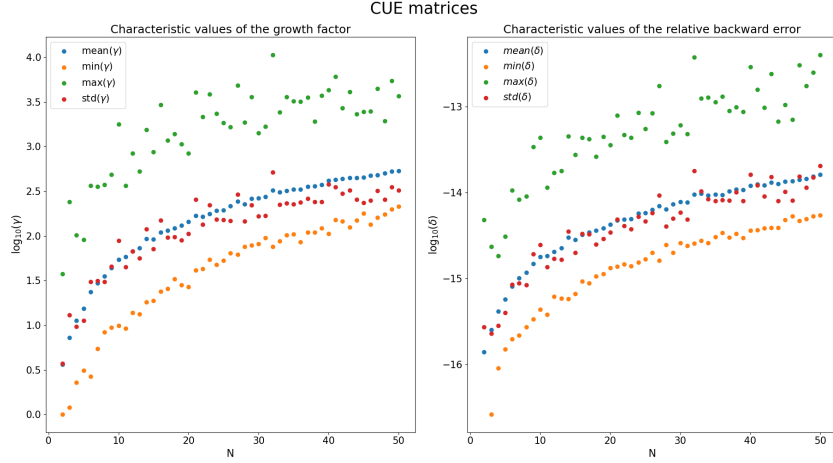


Figure 5: Scatterplot of the characteristic values of  $\gamma$  and  $\delta$  as a function of  $N$  for unitary matrices. Logarithmic scale on the ordinate axis.

The range of variation of both these quantities is comparable with the one obtained in the case of Ginibre matrices. For this reason, the same considerations regarding the overall stability of the LU factorization algorithm can be made. In fact, as can be seen from the values obtained for  $\gamma$  and  $\delta$ , the LU factorization algorithm can be considered backward stable for unitary matrices.

**Hermitian matrices** The next class of matrices that we considered is that of Hermitian matrices, namely matrices  $H$  such that  $H = H^*$ . In order to sample these matrices we used the `tenpy` Python library. These matrices are also called *GUE matrices* as they are sampled from the Gaussian unitary ensemble.

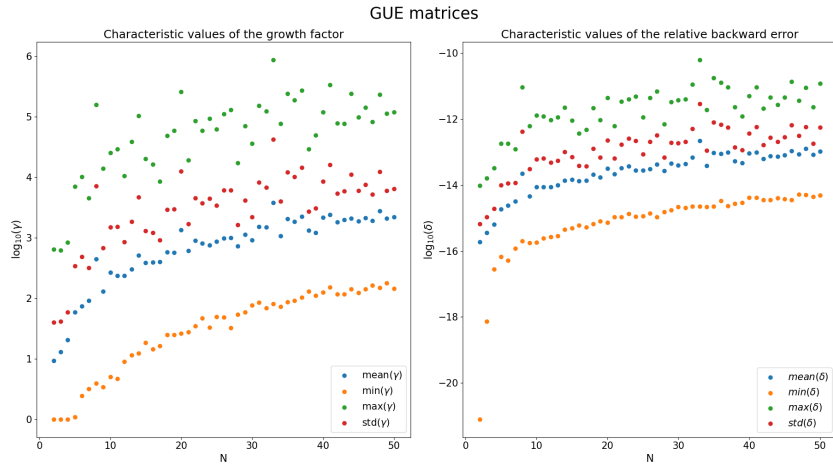


Figure 6: Scatterplot of the characteristic values of  $\gamma$  and  $\delta$  as a function of  $N$  for Hermitian matrices. Logarithmic scale on the ordinate axis.

A characteristic of these scatterplots, as well as of all the others analyzed until now, is that the range of variation of both the growth factor and the relative backward error is quite large, approximatively of three or four orders of magnitude. The choice input matrix has, therefore, an impact in determining these values. However, as in the previous cases, since the trend of the growth factor with the dimension of the input matrix is not exponential and the relative backward error is of the order of the machine precision, these data suggest that the algorithm is backward stable for this class of matrices.

**Positive definite matrices** The next class of matrices that we considered is that of positive definite matrices. Positive definite matrices  $W$  can be obtained by considering matrices of the form  $W = A^*A$ , where  $*$  stands for the conjugate transpose, discarding singular matrices. If the matrix  $A$  is sampled from the Ginibre ensemble, that is it is a complex matrix whose entries are independent and normally distributed, the matrix  $W$  is called *Wishart matrix*. In order to sample these matrices we used the `qutip` Python library, considering, in particular, unit trace matrices. Note that matrices of this kind are also Hermitian. We expect the LU factorization for these matrices to be backward stable.

We report in the following the scatterplot of the growth factor  $\gamma$  and the relative backward error  $\delta$  as a function of the dimension  $N$  of the input matrix.

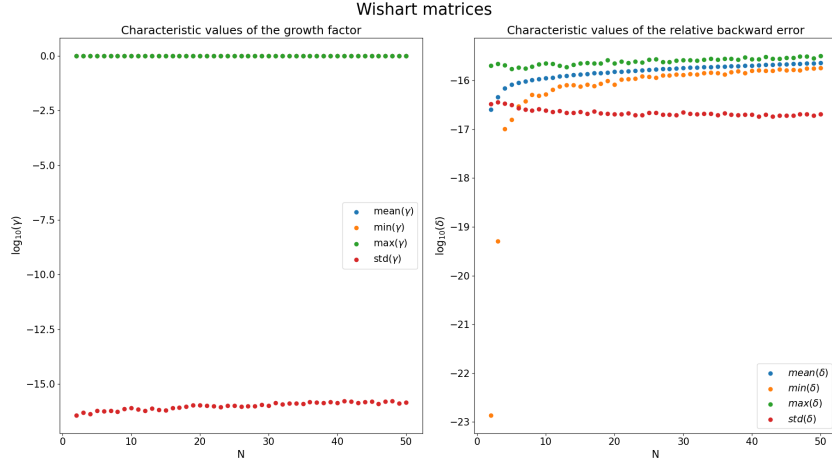


Figure 7: Scatterplot of the characteristic values of  $\gamma$  and  $\delta$  as a function of  $N$  for positive definite matrices. Logarithmic scale on the ordinate axis.

It turns out that this case is quite different from the previous ones. In fact, it can be seen that the minimum, maximum, and mean values of the growth factor are all equal to one, while the standard deviation is of the order of the machine epsilon. Moreover, the relative backward error is significantly small and it is of the order of the machine epsilon. It is interesting to note that, in this latter case, the standard deviation is smaller if compared to the other values, suggesting that the distribution of the relative backward error is peaked in correspondence with the mean value. This means that, differently from the other cases, whatever the input matrix, the values obtained for the relative backward error will be approximately the same, or, in other words, they will be included in a very small range. These considerations imply the backward stability of the LU factorization, as expected, for this class of matrices.

**Diagonally dominant matrices** Another class of matrices that we considered in our dataset is given by the diagonally dominant matrices. These are matrices whose diagonal entries are, in absolute value, greater or equal to the sum of the absolute values of the other elements in the corresponding row, and such that the inequality is strict for at least a row. We have described how we sampled these matrices in the description of the function at the beginning of the chapter. In the following, we report the scatterplot of the growth factor  $\gamma$  and the relative backward error  $\delta$  as a function of the dimension  $N$  of the matrix.

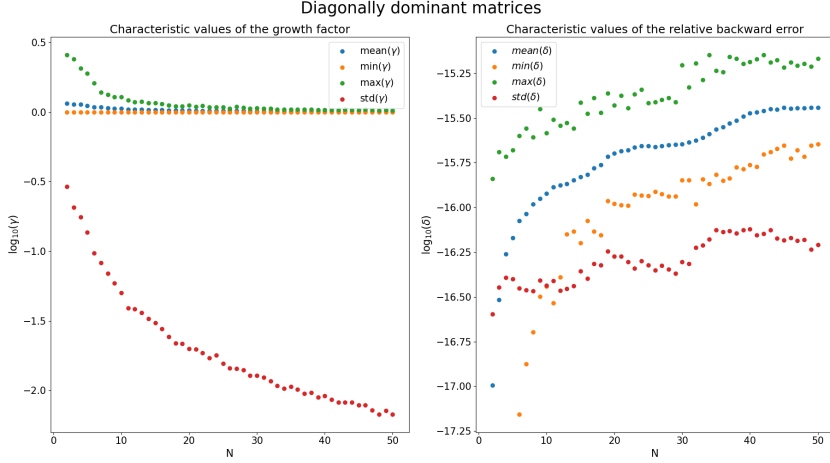


Figure 8: Scatterplot of the characteristic values of  $\gamma$  and  $\delta$  as a function of  $N$  for diagonally dominant matrices. Logarithmic scale on the ordinate axis.

In this case, considerations similar to the case of positive definite matrices can be made. In fact, it can be observed that the growth factor is of order  $\mathcal{O}(1)$ , while the relative backward error is of the order of the machine epsilon, with the standard deviation that suggests, as in the previous case, that the distribution of this quantity, for a fixed dimension of the input matrix, is peaked at the mean value. Therefore, the value obtained for this quantity does not strongly depend on the input matrix. It can be concluded that the LU factorization is backward stable for this class of matrices.

**Hilbert matrices** In addition to the matrices considered in the dataset, we also analyzed the trend of the growth factor and the relative backward error for the Hilbert matrices  $H_N$ , as a function of their dimension. In this case, it emerges that, when considering a precision given by the unit roundoff  $u$ , the algorithm that performs the LU factorization breaks for  $N \geq 26$ . In other words, the growth factor is identically equal to one and the relative backward error is of the order of the machine epsilon for  $N < 26$ , while, as  $N$  increases, a division by a quantity smaller than the machine epsilon occurs, causing the algorithm to break. This means that the problem of solving a linear system  $H_N \mathbf{x} = \mathbf{b}$  is ill-conditioned for  $N \geq 26$  as its condition number increases. In fact, we know from theory that the condition number of the Hilbert matrix increases exponentially with the dimension of the matrix.

### 1.3 Conclusions

Given the classes of matrices considered in the dataset (which does not include the Hilbert matrices, analyzed separately), we found that the LU factorization is backward stable in all the cases considered, with differences that emerge when comparing the last two types of matrices, namely positive definite matrices and diagonally dominant matrices, to the other ones. In fact, in the last two cases, the values obtained for the relative backward error do not depend on the input matrix, differently from the other cases, where this happens. Moreover, we observed that the LU factorization algorithm never failed, as the DataFrame where we stored all the failures of the algorithm, divided by types of matrix considered and their relative dimension, is identically zero. Different considerations hold for the Hilbert matrices, where the LU factorization algorithm breaks for dimensions of the input matrix greater or equal to 26 (given the chosen precision), as a consequence of the increasing condition number.

In the following we report a final plot where we compare the boxplots of all the types of matrices taken into account in the dataset when considering a fixed dimension of the input matrix, that is  $N = 25$ . For greater clarity of visualization, we discard outliers and consider,

in this way, the most likely range of possible values.

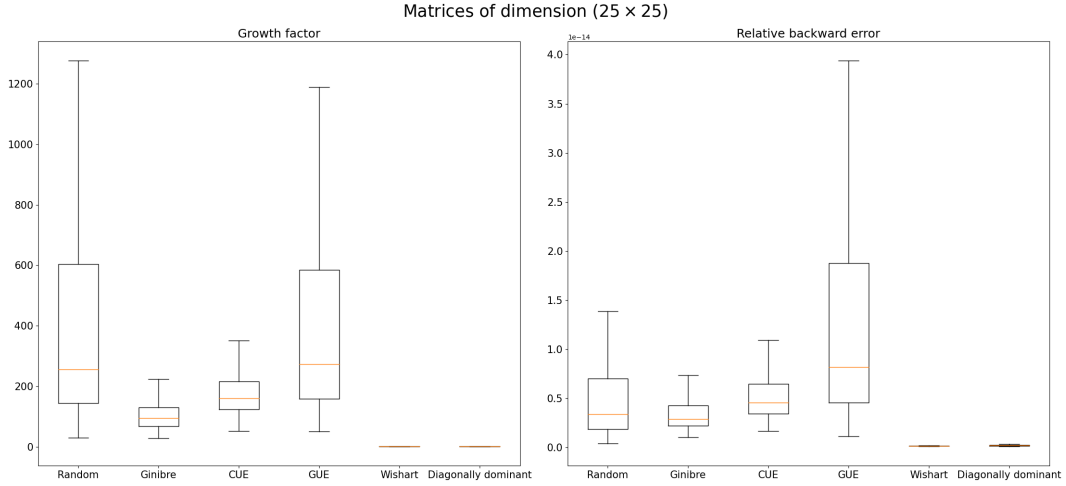


Figure 9: Boxplot of  $\gamma$  and  $\delta$  for matrices of dimension  $25 \times 25$ .

## 2 Problem 2

Consider the  $n \times n$  Wilkinson matrix

$$W_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 1 \\ -1 & 1 & 0 & \cdots & 1 \\ -1 & -1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & -1 & 1 \end{bmatrix} \quad (1)$$

(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  $W_5$ :

$$W_5 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix}, \quad (2)$$

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

$$\mathbf{m}_1 = \begin{bmatrix} 0 \\ w_{21}/w_{11} \\ w_{31}/w_{11} \\ w_{41}/w_{11} \\ w_{51}/w_{11} \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix}. \quad (3)$$

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

$$M_1 = \mathcal{I}_5 - \mathbf{m}_1 \mathbf{e}_1^T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad (4)$$

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

$$W_5^{(1)} = M_1 W_5 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & -1 & 1 & 0 & 2 \\ 0 & -1 & -1 & 1 & 2 \\ 0 & -1 & -1 & -1 & 2 \end{bmatrix}. \quad (5)$$

The second iteration proceeds in a similar way:

$$\mathbf{m}_2 = \begin{bmatrix} 0 \\ 0 \\ w_{32}^{(1)}/w_{22}^{(1)} \\ w_{42}^{(1)}/w_{22}^{(1)} \\ w_{52}^{(1)}/w_{22}^{(1)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -1 \\ -1 \\ -1 \end{bmatrix}, \quad (6)$$

$$M_2 = \mathcal{I}_5 - \mathbf{m}_2 \mathbf{e}_2^T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix}. \quad (7)$$

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

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

Now we start the third iteration:

$$\mathbf{m}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ w_{43}^{(2)}/w_{33}^{(2)} \\ w_{53}^{(2)}/w_{33}^{(2)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1 \\ -1 \end{bmatrix}, \quad (9)$$

$$M_3 = \mathcal{I}_5 - \mathbf{m}_3 \mathbf{e}_3^T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix}. \quad (10)$$

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

$$W_5^{(3)} = M_3 W_5^{(2)} = M_3 M_2 M_1 W_5 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 & 8 \\ 0 & 0 & 0 & -1 & 8 \end{bmatrix}. \quad (11)$$

Similarly, we can perform the last iteration:

$$\mathbf{m}_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ w_{54}^{(2)}/w_{44}^{(2)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -1 \end{bmatrix}, \quad (12)$$

$$M_4 = \mathcal{I}_5 - \mathbf{m}_4 \mathbf{e}_4^T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}. \quad (13)$$

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

$$W_5^{(4)} = M_4 W_5^{(3)} = M_4 M_3 M_2 M_1 W_5 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 & 8 \\ 0 & 0 & 0 & 0 & 16 \end{bmatrix} = U_5. \quad (14)$$

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  $M_i^{-1} = \mathcal{I}_5 + \mathbf{m}_i \mathbf{e}_i^T$ :

$$L_5 = M_1^{-1} M_2^{-1} M_3^{-1} M_4^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ -1 & -1 & -1 & 1 & 0 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix}. \quad (15)$$

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

$$L_n = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -1 & 1 & \ddots & & \vdots \\ \vdots & \ddots & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ -1 & \cdots & \cdots & -1 & 1 \end{bmatrix}, \quad U_n = \begin{bmatrix} 1 & 0 & \cdots & 0 & 2^0 \\ 0 & 1 & \ddots & \vdots & 2^1 \\ \vdots & \ddots & \ddots & 0 & 2^2 \\ \vdots & & \ddots & 1 & \vdots \\ 0 & \cdots & \cdots & 0 & 2^{n-1} \end{bmatrix} \quad (16)$$

(3) In the following, we report the function that generates the  $n \times n$  Wilkinson matrix.

```

1 def wilkin(n):
2     r''' This function computes the Wilkinson matrix of dimension :math:`(n`
3     ↪ \times n)`.
4     Parameters
5     -----
6     n : int
7     Dimension of the Wilkinson matrix
8
9     Returns
10    -----

```

```

11 W : ndarray
12 Wilkinson matrix
13 '''
14 W = np.tril(-np.ones((n,n)),k=-1) + np.eye(n)
15 W[:,n-1] = 1
16 return W

```

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

```

1 def check_when_lufact_W_fails(n_max = 60, threshold = np.finfo(float).eps):
2     r''' This function checks the failures of GEPP for a Wilkinson matrix W_n
3         ↪ with dimension less or equal to n_max. We define a failure as a case in
4         ↪ which the error between the solution found by the algorithm and the
5         ↪ expected solution is higher than a chosen threshold. The default
6         ↪ threshold is set equal to the machine epsilon. When an error is found,
7         ↪ a warning message is printed. The function returns the list of the
8         ↪ dimensions less than or equal to n_max for which the GEPP algorithm
9         ↪ fails.
10
11 Parameters
12 -----
13 n_max : int
14     Maximum dimension of the Wilkinson matrix
15
16 Returns
17 -----
18 fails : list
19     List of dimensions for which the algorithm fails
20 '''
21
22 fails = []
23
24 # Cycle on the dimension of the input
25 for n in range(2,n_max+1):
26     W = wilkin(n)
27     logging.debug(f'W = {W}')
28
29 # Define the vector b
30 vect = np.ones(n)
31 b = W @ vect
32
33 # Solve the system with GEPP
34 x = np.linalg.solve(W,b)
35
36 # Compute the error in 1-norm between the computed solution and the
37     ↪ exact solution, and print a warning message when the error exceeds
38     ↪ the chosen precision
39 error = sum(np.abs(x - vect))
40 if error <= threshold:
41     logging.info(f'n = {n}, ||x - e||_1 = {error}')
42 elif error > threshold:

```

```

34     logging.warning(f'n = {n}, ||x - e||_1 = {error}')
35     fails.append(n)
36     # Cycle on the elements of the computed solution and print a warning
    ↪ message with the wrong elements of the solution.
37     for i in range(n):
38         if abs(x[i] - 1) > threshold:
39             logging.warning(f'x[{i}] = {x[i]}')
40     return fails
41

```

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:

$$\tilde{\mathbf{x}}_{54} = 0 \neq \mathbf{e}_{54} = 1. \quad (17)$$

In order to understand why this happens, we have verified that the matrices  $L_{55}$  and  $U_{55}$  are computed accurately. This is done in the function `compute_error_lufact_W(n)`, which can be found in the library [Project\\_1](#). In this way, we know that the problem must be in the calculation of the forward and backward substitution. Note 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) the following two linear systems with forward and backward substitutions:

$$\begin{cases} L\mathbf{y} = \mathbf{b} \\ U\mathbf{x} = \mathbf{y} \end{cases}, \quad (18)$$

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, an 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, in double precision,  $\text{fl}(2^{53} + 1) = 2^{53}$  or, in other words, 1 is smaller than the numerical distance between  $2^{53}$  and the next representable number, when working in double precision. This can be justified by estimating the absolute error associated with the representation of  $2^{53} + 1$ :

$$|x - \text{fl}(x)| < u|x| = 2^{-53}(2^{53} + 1) = 1 + 2^{-53}. \quad (19)$$

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

$$x_{55} = y_{55}/U_{55,55} = \frac{2^{54}}{2^{54}} = 1. \quad (20)$$

Now we update the vector  $\mathbf{y}$  as follows:

$$\mathbf{y}^{(1)} = \mathbf{y} - x_{55} \mathbf{u}_{55}, \quad (21)$$

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

$$[\mathbf{y}^{(1)}]_{54} = [\mathbf{y}]_{54} - x_{55} U_{54,55} = 2^{53} + 1 - 2^{53}, \quad (22)$$

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:

$$x_n = y_n / U_{n,n} = \frac{2^{n-1}}{2^{n-1}}, \quad (23)$$



that does not lead to catastrophic 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  $\gamma$  assumes the maximum possible value, i.e.  $\gamma = 2^{n-1}$ . However, in most cases, the GEPP algorithm is backward stable.

For the sake of completeness, we have developed a function, named `step_by_step_GEPP_W(n)` that calculates step by step the forward and backward substitution using the GEPP algorithm for Wilkinson matrices, commented in such a way as to make each step explicit, with warnings that signal errors due to catastrophic cancellations.

### 3 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  $\mathbf{u}\mathbf{v}^T = \alpha_1 \beta_1 \mathbf{e}_1 \mathbf{e}_1^T + \alpha_1 \beta_2 \mathbf{e}_1 \mathbf{e}_2^T$ . It is known that we can compute the determinant of a  $n \times n$  matrix  $M$  using the formula

$$\det(M) = \sum_{j=1}^n m_{ij} C_{ij}(M), \quad (24)$$

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

$$\det(\tilde{A}) = \det(A) + \alpha_1 \beta_1 C_{11}(A) + \alpha_1 \beta_2 C_{12}(A). \quad (25)$$

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

$$\det(A) + \alpha_1 \beta_1 C_{11}(A) + \alpha_1 \beta_2 C_{12}(A) \neq 0, \quad (26)$$

and, therefore:

$$\alpha_1 \beta_1 \frac{C_{11}(A)}{\det(A)} + \alpha_1 \beta_2 \frac{C_{12}(A)}{\det(A)} \neq -1, \quad (27)$$

where we divided for  $\det(A)$  both sides of the equation, knowing that  $\det(A) \neq 0$ . At this point, it can be verified that

$$\mathbf{v}^T A^{-1} \mathbf{u} = \alpha_1 \beta_1 \frac{C_{11}(A)}{\det(A)} + \alpha_1 \beta_2 \frac{C_{12}(A)}{\det(A)}, \quad (28)$$

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

$$\mathbf{v}^T A^{-1} \mathbf{u} \neq -1. \quad (29)$$

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

$$\det(\tilde{A}) = \det(A) + \alpha_1 \beta_1 C_{11}(A) + \alpha_1 \beta_2 C_{12}(A) = \det(A) \left( 1 + \alpha_1 \beta_1 \frac{C_{11}(A)}{\det(A)} + \alpha_1 \beta_2 \frac{C_{12}(A)}{\det(A)} \right). \quad (30)$$

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

$$\det(\tilde{A}) = \det(A)(1 + \mathbf{v}^T A^{-1} \mathbf{u}). \quad (31)$$

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

$$\det(\tilde{A}) \neq 0 \quad (32)$$

and this concludes the proof.  $\square$

**(1b)** Show that:

$$\tilde{A}^{-1} = A^{-1} - \alpha A^{-1} \mathbf{u} \mathbf{v}^T A^{-1}, \quad \text{where } \alpha = \frac{1}{\mathbf{v}^T A^{-1} \mathbf{u} + 1}. \quad (33)$$

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 (33) multiplying both sides to the left by  $A$  and to the right by  $\tilde{A}$ , obtaining:

$$\begin{aligned} A &= \tilde{A} - \alpha \mathbf{u} \mathbf{v}^T A^{-1} \tilde{A} \\ &= A + \mathbf{u} \mathbf{v}^T - \alpha \mathbf{u} \mathbf{v}^T A^{-1} (A + \mathbf{u} \mathbf{v}^T) \\ &= A + (1 - \alpha) \mathbf{u} \mathbf{v}^T - \alpha \mathbf{u} \mathbf{v}^T A^{-1} \mathbf{u} \mathbf{v}^T. \end{aligned} \quad (34)$$

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

$$\mathbf{u} \mathbf{v}^T A^{-1} \mathbf{u} \mathbf{v}^T = \mathbf{u} \mathbf{v}^T (\alpha^{-1} - 1). \quad (35)$$

Finally, since  $\mathbf{v}^T A^{-1} \mathbf{u} = \alpha^{-1} - 1$ , the identity (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  $\tilde{A}$ :

$$\tilde{\mathbf{x}} = \tilde{A}^{-1} \tilde{\mathbf{b}} = A^{-1} \tilde{\mathbf{b}} - \frac{A^{-1} \mathbf{u} \mathbf{v}^T A^{-1} \tilde{\mathbf{b}}}{\mathbf{v}^T A^{-1} \mathbf{u} + 1}. \quad (36)$$

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  $A$  exists and has been computed, describe an efficient algorithm for solving the *bordered system*

$$\begin{bmatrix} A & \mathbf{u} \\ \mathbf{v}^T & \beta \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ z \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ c \end{bmatrix}, \quad (37)$$

where  $z$  is unknown and  $\beta$  and  $c$  are given scalars. When does this system have a unique solution?

Solution: Defining

$$A' = \begin{bmatrix} A & \mathbf{u} \\ \mathbf{v}^T & \beta \end{bmatrix}, \mathbf{x}' = \begin{bmatrix} \mathbf{x} \\ z \end{bmatrix}, \mathbf{b}' = \begin{bmatrix} \mathbf{b} \\ c \end{bmatrix}, \quad (38)$$

the system above rewrites as:

$$A'\mathbf{x}' = \mathbf{b}'. \quad (39)$$

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

$$L' = \begin{bmatrix} L & \mathbf{0} \\ \mathbf{f}^T & 1 \end{bmatrix}, U' = \begin{bmatrix} U & \mathbf{g} \\ \mathbf{0} & \gamma \end{bmatrix}. \quad (40)$$

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:

$$\begin{cases} L\mathbf{g} = \mathbf{u} \\ U^T\mathbf{f} = \mathbf{v} \\ \mathbf{f}^T\mathbf{g} + \gamma = \beta \end{cases}. \quad (41)$$

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:

$$\gamma = \beta - \mathbf{v}^T A^{-1} \mathbf{u}. \quad (42)$$

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 uniqueness of the solution becomes:

$$\gamma = \beta - \mathbf{v}^T A^{-1} \mathbf{u} \neq 0 \Rightarrow \mathbf{v}^T A^{-1} \mathbf{u} \neq \beta. \quad (43)$$