International Journal of Pure and Applied Mathematics
Volume 76 No. 1 2012, 113-122
ISSN: 1311-8080 (printed version)
url: http://www.ijpam.eu

ijpam.eu

# ON THE NUMERICAL SOLVING OF COMPLEX LINEAR SYSTEMS

Romulus Militaru[1] §, Ioan Popa[2]

[1]Faculty of Exact Sciences
Department of Applied Mathematics
University of Craiova
13, A.I. Cuza, 200585, Craiova, ROMANIA

[2]Faculty of Electrical Engineering
Electrical Apparatus and Technologies Department
University of Craiova
107, Blvd. Decebal, 200440, Craiova, ROMANIA

**Abstract:** The present paper concerns the numerical solving of complex linear systems. The basic idea is to transform the given complex system into a real one and then to select an appropriate numerical method, direct or iterative, to solve the problem. Also, we focus on the cases of a band and sparse matrix, giving a practical storage scheme in order to be able to solve numerically efficient a linear system having a large sparse coefficient matrix.

**AMS Subject Classification:** 65F10, 65F50
**Key Words:** complex linear system, band matrix, sparse matrix, direct methods, iterative methods, storage scheme

## 1. Introduction

Despite the nonlinearity of many actual dependencies, linear assumptions are widely used throughout engineering and science, many investigations starting

Received: December 12, 2011



§Correspondence address: Cart. G. Enescu, No. 55, Bl. D2, Sc. 1, Apt. 5, Craiova, Dolj, ROMANIA

with linear models. A lot of exact or approximate methods for analyzing and solving mathematical problems are more appropriate for this kind of models, see [10]. Also, most of the more advanced problems in scientific calculations require the solution of linear systems, having real or complex coefficients, often of very large dimension. Since linear systems of equations are so common, their efficient solution is one of the central topics of numerical computation, see [1], [6]. Linear systems with complex coefficients arise from various physical problems. For instance, the Helmholtz equation and Maxwell equations approximated by finite difference or finite element methods lead to large sparse complex linear systems.

## 2. Problem Statement

We consider the following linear system of equations with $n$ unknowns, written in explicit form:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = b_2 \\ \qquad\qquad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n = b_n \end{cases} \tag{1}$$

In matrix form, the above system becomes:

$$A \cdot x = b \tag{2}$$

where $A$ is a $n \times n$ matrix and $b$, $x$ are $n \times 1$ vectors:

$$A = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & ... & a_{nn} \end{pmatrix}, \; b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}, \; x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

All the coefficients are assumed to be complex:

$$a_{ij} \in \mathbf{C}, \; 1 \leq i, j \leq n, \quad b_i \in \mathbf{C}, \; 1 \leq i \leq n.$$

We suppose that each complex element $\alpha$ can be written in the form:

$$\alpha = \alpha^r + i\alpha^c \tag{3}$$

where $\alpha^r, \alpha^c \in \mathbf{R}$.

We consider the following linear system of equations with $2n$ unknowns, written in matrix form:

$$\tilde{A} \cdot \tilde{x} = \tilde{b} \tag{4}$$

where

$$\tilde{A} = \begin{pmatrix} A^r & -A^c \\ A^c & A^r \end{pmatrix} \in \mathbf{R}^{2n\times 2n},\ \tilde{b} = \begin{pmatrix} b^r \\ b^c \end{pmatrix} \in \mathbf{R}^{2n},\ \tilde{x} = \begin{pmatrix} x^r \\ x^c \end{pmatrix} \in \mathbf{R}^{2n} \tag{5}$$

the notations being

$$A^r = \begin{pmatrix} a^r_{11} & a^r_{12} & ... & a^r_{1n} \\ a^r_{21} & a^r_{22} & ... & a^r_{2n} \\ \vdots & \vdots & & \vdots \\ a^r_{n1} & a^r_{n2} & ... & a^r_{nn} \end{pmatrix} \in \mathbf{R}^{n\times n},\ b^r = \begin{pmatrix} b^r_1 \\ b^r_2 \\ \vdots \\ b^r_n \end{pmatrix} \in \mathbf{R}^n,\ x^r = \begin{pmatrix} x^r_1 \\ x^r_2 \\ \vdots \\ x^r_n \end{pmatrix} \in \mathbf{R}^n$$

$$A^c = \begin{pmatrix} a^c_{11} & a^c_{12} & ... & a^c_{1n} \\ a^c_{21} & a^c_{22} & ... & a^c_{2n} \\ \vdots & \vdots & & \vdots \\ a^c_{n1} & a^c_{n2} & ... & a^c_{nn} \end{pmatrix} \in \mathbf{R}^{n\times n},\ b^c = \begin{pmatrix} b^c_1 \\ b^c_2 \\ \vdots \\ b^c_n \end{pmatrix} \in \mathbf{R}^n,\ x^c = \begin{pmatrix} x^c_1 \\ x^c_2 \\ \vdots \\ x^c_n \end{pmatrix} \in \mathbf{R}^n$$

The following statement holds:

**Proposition 1.** *The linear system of equations (1), (or (2)), has an unique solution if and only if the linear system of equations (4) admits an unique solution.*

*Proof.* "⇒" Let $x \in \mathbf{C}^n, x = x^r + ix^c,\ x^r, x^c \in \mathbf{R}^n$, the unique solution for (1). Then $A \cdot x = b$. Using (3) it follows

$$A^r \cdot x^r - A^c \cdot x^c + i(A^r \cdot x^c + A^c \cdot x^r) = b^r + ib^c$$

Thus $\begin{cases} A^r \cdot x^r - A^c \cdot x^c = b^r \\ A^c \cdot x^r + A^r \cdot x^c = b^c \end{cases}$ and taking into account the notations (5) one gets $\tilde{A} \cdot \tilde{x} = \tilde{b}$.

"⇐" Supposing $\tilde{x} = \begin{pmatrix} x^r \\ x^c \end{pmatrix} \in \mathbf{R}^{2n}$ represents the unique solution of (4). Then using (5) it follows $\begin{cases} A^r \cdot x^r - A^c \cdot x^c = b^r \\ A^c \cdot x^r + A^r \cdot x^c = b^c \end{cases}$ which is equivalent with $A \cdot x = b$, where $x \in \mathbf{C}^n$, $x = x^r + ix^c$.

**Consequence 1.** The coefficient matrix $A \in \mathbf{C}^{n\times n}$ of the linear system (2) is nonsingular if and only if the coefficient matrix $\tilde{A} \in \mathbf{R}^{2n\times 2n}$ of the linear system (4) is nonsingular.

**Observation 1.** Taking into account the above results, we can conclude that the computation of the solution of a linear system of $n$ equations with complex coefficients, is equivalent to the computation of the solution of a linear system of $2n$ equations with real coefficients.

## 3. Numerical Solution

As we presented in the first section, a linear system of equations with complex coefficients is equivalent with a linear one, having a double number of unknowns and real coefficients.

There are two large classes of methods for the numerical solution of linear systems of equations with real coefficients (see [2], [4], [7]):

— *direct methods*: they produce the exact solution (supposing the absence of rounding errors), by performing a finite number of arithmetic operations in a finite and "a priori" known number of stages. These methods are preferred when most of unknowns coefficients are nonzero and the dimension $n$ of the system is not too large. Examples of direct methods: the Gaussian elimination (with/without pivoting technique), the LU decomposition etc.

— *iterative methods*: they construct successive approximations $x^{(k)}$, $k = 0, 1, ...$ which converge to the exact solution. These methods compute an approximate solution, whose accuracy is imposed by the user. The number $k$ of accomplished iterations depends directly on the given precision. These methods are preferred when a majority of the coefficients are equal to zero and the dimension $n$ of the system is large.

Examples of iterative methods: the methods of Jacobi and Seidel-Gauss, the Conjugate gradient method, the Successive overrelaxation technique etc.

### 3.1. Large Sparse Linear Systems of Equations

Most of the linear systems of equations encountered in engineering and scientific applications are very large if high accuracy is required and the solution of these systems is an important and costly step. For example, in electromagnetic scattering problems the mesh size (when using finite difference or finite element methods) must be related to the wave length of the incoming wave. The higher

the frequency of the incoming wave, the smaller the mesh size must be. Thus for solving real 3D-problems, it is no longer practical to use direct method solvers, because of the huge memory they need.

Also the resulting matrix coefficients is *sparse*: it has a small percentage of nonzero terms. Different types of sparse matrices: symmetric positive definite matrix, diagonally dominant matrix, structurally symmetric matrix, banded matrix. If the equations and the unknowns of such a system are ordered properly, these sparse matrices can generally be set into banded form, with all nonzero elements confined to a relatively narrow band around the main diagonal.

**Definition 1.** A matrix $A$ is called *banded* if there are natural numbers $p$ and $q$ such that

$$a_{ij} = 0 \text{ when } j - i > p \text{ or } i - j > q$$

**Observation 2.** (i) The natural number $w = p+q+1$ is called bandwidth of the matrix.

(ii) If $p = q = 1$ then such a matrix is called tridiagonal: the nonzero elements lay on the main diagonal as well as on the first diagonal below and above the main one;

If $p = q = 2$ the matrix is called pentadiagonal: the only nonzero entries are on the main diagonal, and the first two diagonals above and below it.

**Proposition 2.** *Let $A \in \mathbf{R}^{n\times n}$ be a band matrix having the bandwidth equal to $w = p + q + 1$. Then the number of nonzero elements, $NZE(A) \in \mathbf{N}$, is characterized by*

$$0 < NZE(A) \leq nw - \frac{1}{2}w(w-1) + pq \tag{6}$$

*Proof.* Let $A \in \mathbf{R}^{n\times n}$, having $p$ diagonals above the main diagonal and $q$ diagonals below it. Thus, the number of the elements within the diagonals above the main diagonal is equal to $np - \dfrac{p(p+1)}{2}$, and the number of the elements within the diagonals below it is equal to $nq - \dfrac{q(q+1)}{2}$. Knowing that the main diagonal of $A$ has $n$ elements, it follows that the nonzero elements $NZE(A)$ is characterized by

$$0 < NZE(A) \leq n(1+p+q) - \frac{1}{2}(p^2+q^2+p+q)$$

Using the expression of the bandwidth $w$ of $A$ one gets the conclusion.

**Proposition 3.** *Let $A \in \mathbf{C}^{n\times n}$ a banded matrix having the bandwidth $w = p + q + 1$. Supposing that $A = A^r + iA^c$, $A^r$, $A^c \in \mathbf{R}^{n\times n}$, then for the matrix $\tilde{A} = \begin{pmatrix} A^r & -A^c \\ A^c & A^r \end{pmatrix} \in \mathbf{R}^{2n\times 2n}$, the following statement holds:*

$$0 < NZE(\tilde{A}) \leq 4nw - 2w(w-1) + 4pq \tag{7}$$

**Observation 3.** If $A \in \mathbf{C}^{n\times n}$ is a band matrix then the corresponding matrix $\tilde{A} \in \mathbf{R}^{n\times n}$ will have a small percentage of non-zero elements or will be a sparse one.

### 3.2. Practical Storage Scheme

The direct solution of a linear system having a sparse matrix can be obtained by the mean of direct methods. Taking into account that the amount of computational work for the solution rises with the third power of the number of equations, in the case of a large linear system this computational cost and the corresponding total time cannot be ignored. In order to significantly improve the efficiency, one has to better exploit the sparse structure of the matrix coefficients, taking advantage of the fact that most of the matrix elements are zero and using special storage schemes, see [3], [5]. Thus, the conclusion that there is a lot of work involved in the sparse direct methods.

The iterative methods represent a better alternative for numerical resolution of the large linear systems of equations. They offer an improvement of the computational time and allow us to conserve the initial structure of the matrix, and thus the use of a minimal memory storage.

The efficiency of each iterative method depends indirectly on the method chosen for storing the coefficient matrix, see [8].

Obviously it is desirable to store only the non-zero coefficient values of a sparse matrix. At the same time, however, information about the position of the stored coefficient values has to be recorded.

In the sequel we will present a storage scheme for sparse matrices, named VRC format (value, row, column). It consists of three linear arrays:

— a real array $v = v_i$, $1 \leq i \leq NZE(A)$ containing the floating-point values of the non-zero elements of the coefficient matrix $A$;

— an integer array $r = r_i$, $1 \leq i \leq n$ containing the number of non-zero elements of the row $i$;

— an integer array $c = c_i$, $1 \leq i \leq NZE(A)$ containing the column index of each non-zero element, starting from the first row until the last one. In terms of

memory requirements the VRC format requires storage only for $2 \cdot NZE(A)+n$ data elements.

Thus the above scheme decreases the storage saving in comparison with other known practical storage schemes as Coordinate (COO) Format, requiring $3 \cdot NZE(A)+n$, Compressed Row Storage (CRS) Format, requiring $2 \cdot NZE(A)+n+1$, see [9].

**Example.** Let $A \in \mathbf{R}^{6\times 6}$,

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 3 & -3 & 1 & 0 & 0 \\ 0 & -3 & 5 & -2 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 2 & 3 \\ 0 & 0 & 0 & 0 & 3 & 6 \end{pmatrix}.$$

Using the VRC storage scheme $A$ can be specified by the following arrays:

| value | 1 | 1 | 3 | -3 | 1 | -3 | 5 | -2 | 1 | -2 | 1 | 1 | 2 | 3 | 3 | 6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| column index | 1 | 5 | 2 | 3 | 4 | 2 | 3 | 4 | 2 | 3 | 4 | 1 | 5 | 6 | 5 | 6 |

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| non-zero elements of row $i$ | 2 | 3 | 3 | 3 | 3 | 2 |

## 4. Numerical Example

Let the following linear complex system $A \cdot x = b$, $A \in \mathbf{C}^{5\times 5}$

$$A = \begin{pmatrix} 19.73 & 12.11 - i & 5i & 0 & 0 \\ -0.51i & 32.3 + 7i & 23.07 & i & 0 \\ 0 & -0.51i & 70 + 7.3i & 3.95 & 19 + 31.83i \\ 0 & 0 & 1 + 1.1i & 50.17 & 45.51 \\ 0 & 0 & 0 & -9.351i & 55 \end{pmatrix}$$

a band matrix having the bandwidth $w = 4$. and $b \in \mathbf{C}^5$, given by

$$b = \begin{pmatrix} 77.38 + 8.82i \\ 157.48 + 19.8i \\ 1175.62 + 20.69i \\ 912.12 - 801.75i \\ 550 - 1060.4i \end{pmatrix}$$

Thus, using the above analysis, the corresponding real matrix $\tilde{A}$ given by (5) is:

$$\begin{pmatrix} 19.73 & 12.11 & 0 & 0 & 0 & 0 & 1 & -5 & 0 & 0 \\ 0 & 32.3 & 23.07 & 0 & 0 & 0.51 & -7 & 0 & -1 & 0 \\ 0 & 0 & 70 & 3.95 & 19 & 0 & 0.51 & -7.3 & 0 & -31.83 \\ 0 & 0 & 1 & 50.17 & 45.51 & 0 & 0 & -1.1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 55 & 0 & 0 & 0 & 9.351 & 0 \\ 0 & -1 & 5 & 0 & 0 & 19.73 & 12.11 & 0 & 0 & 0 \\ -0.51 & 7 & 0 & 1 & 0 & 0 & 32.3 & 23.07 & 0 & 0 \\ 0 & -0.51 & 7.3 & 0 & 31.83 & 0 & 0 & 70 & 3.95 & 19 \\ 0 & 0 & 1.1 & 0 & 0 & 0 & 0 & 1 & 50.17 & 45.51 \\ 0 & 0 & 0 & -9.351 & 0 & 0 & 0 & 0 & 0 & 55 \end{pmatrix}.$$

It is a non-symmetric $10 \times 10$ matrix with 42% non-zero elements.

Based on VRC storage scheme, $\tilde{A}$ is specified by the three arrays $v$, $r$ and $c$, where $v$ is a real 38 elements array, whose components are given by the non-zero elements of $\tilde{A}$, $r$ is an integer 10 elements array containing the non-zero elements of each row $i, 1 \leq i \leq 10$ and $c$ an integer 38 elements array containing the the column index of each non-zero element from the first row until the last one.

Also, the real vector $\tilde{b}$ given by (5) has the following components

$$\tilde{b} = \begin{pmatrix} 77.38 \\ 157.48 \\ 1175.62 \\ 912.12 \\ 550 \\ 8.82 \\ 19.8 \\ 20.69 \\ -801.75 \\ -1060.4 \end{pmatrix}$$

Using the Seidel-Gauss iterative method, with the above storage scheme, for the numerical solving of the given complex linear system we get:

| accuracy $\varepsilon$ | number of accomplished iterations | approximate solution $x$ |
|---|---|---|
| $1e-2$ | 7 | $3.3000256 - 1.001811i$<br>$1.000607 + 0.169971i$<br>$5.499587 - 0.000011i$<br>$8.998377 - 0.000007i$<br>$10.000131 - 17.750112i$ |
| $1e-5$ | 11 | $3.299693 - 1.000375$<br>$0.999761 + 0.169839i$<br>$5.500074 - 0.000046i$<br>$8.999787 - 0.000067i$<br>$10.000012 - 17.749872i$ |

**Observation 4.** The exact solution is:

$$x = \begin{pmatrix} 3.3 - i \\ 1 + 0.17i \\ 5.5 \\ 9 \\ 10 - 17.75i \end{pmatrix}.$$

## References


[1] R.L. Burden, J.D. Faires, *Numerical Analysis*, Pws-Kent (2004).

[2] C. Carasso, *Analyse Numérique*, Lidec, Canada (1970).

[3] J.J. Dongarra, H.A. Van der Vorst, Performance of various computers using standard sparse linear equations solving techniques, In: *Computer Benchmarks*, Elsevier, New-York (1993).

[4] F. Jedrzejewski, *Introduction aux Méthodes Numériques*, Springer (2005).

[5] J. Mellor-Crummey, J. Garvin, Optimizing sparse matrix vector product computations using unroll and jam, *International Journal of High Performance Computing Applications*, **18**, No. 2 (2004), 225-236.

[6] J. Stoer, R. Bulirsch, *Introduction to Numerical Analysis*, Second Edition, Springer-Verlag (2002).

[7] P. Taylor, G. Phillips, *Theory and Applications of Numerical Analysis*, Second Edition, Academic Press (1996).

[8] P. Tvrdik P, I. Simecek, A new approach for accelerating the sparse matrix-vector multiplication, In: *Proceedings of the 8-th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, IEEE Computer Society (2006), 156-163.

[9] C.W. Ueberhuber, *Numerical Computation*, Springer (1997).

[10] S.M. Wong, *Computational Methods in Physics and Engineering*, Second Edition, World Scientific (2003).