DEMONSTRATIO MATHEMATICA
Vol. XXX No 4 1997

Andrei Nicolaide

A NEW PROOF OF THE FORMULAE USED IN THE CONJUGATE GRADIENT METHOD WITH PRE-CONDITIONING FOR SOLVING LARGE SYSTEMS OF LINEAR EQUATIONS

Abstract. An original proof, easily accessible and accurate is proposed, for establishing the formulae on which the computing algorithm is based, in full agreement with the purpose aimed at in computing applications. The proof is founded only on the minimization conditions of a corresponding functional.

1. Introduction

In many applications, for instance in the finite element method, it is necessary to solve systems of equations with a large number of unknowns from approximately 200 to 10 000, called large systems of linear equations. For solving these systems one can use direct and iterative methods [1]–[9]. In the present paper, the conjugate gradient methods will be examined, and namely the conjugate gradient methods, which may be called also conjugate directions methods, with pre-conditioning, which have given favourable results in many applications.

It is interesting to be mentioned, that there are different variants of the conjugate gradient method with pre-conditioning. It must be underlined, that various methods, although correct, cannot be utilized, because of the too great influence of the rounding errors, which arise when they are applied on a computer. That is why, it is necessary to examine the properties of these methods when they are applied in practice.

We shall propose a new proof for the relations on which the conjugate gradient methods are based, in an easily accessible manner and useful for their application in practice. The symbols will be the usual ones [1], [7].

2. The conjugate directions method with pre-conditioning

2.1. The variants of the method

Given the system of equations in matrix form

$$\mathbf{Ax} = \mathbf{b}, \tag{I}$$

where $\mathbf{A}$ is a symmetric positive definite matrix. The conjugate directions methods are based on searching, by an iterative procedure, of the solution of a linear system of equations in the form of the vector

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + a_m \mathbf{p}^{(m)}, \tag{II}$$

and at every step of the iterative procedure, it is required that the direction of vector $\mathbf{p}^{(m)}$ should be conjugate with the direction of the previous iteration vector. The condition is imposed, that the two directions should be conjugate versus the matrix $\mathbf{A}$ of the system. One obtains

$$(\mathbf{p}^{(m+1)})^T \mathbf{A} \mathbf{p}^{(m)} = 0. \tag{III}$$

If the matrix $\mathbf{A}$ of the system were equal to the unit matrix, the two vectors would be orthogonal. There are, also, other equivalent manners for expressing the stated condition. There are different variants for deriving the computing formulae [2], [3], [5], [9], analyzed in paper [6]. In these variants, the condition that certain directions should be orthogonal or conjugate (versus certain matrices) is utilized. We shall consider two typical variants:

1^0. In the first variant [9, Part 2, p. 167, 171], the set of utilized formulae is obtained directly, by a congruent transformation containing a dyadic decomposition and an endogenous transformation. Thus, a quadratic matrix is transformed into a diagonal one, which will be inverted.

2^0. In the second variant, the set of formulae given in paper [2, p. 243] is obtained. This set is obtained by utilizing the iterative method with common steps with convergence acceleration coefficient, i.e. the method of Richardson, and putting the condition, that the rests (residuals) of the system of equations, should be conjugate with a certain symmetric matrix. It is proved, that one of the two main coefficients which occur corresponds to a minimization procedure of a functional, whereas the second follows imposing, as previously, the condition that some vectors should be conjugate versus a certain symmetric matrix.

We consider that in establishing the conjugate gradient methods, it would be best to utilize only the minimization of a functional.

Indeed, referring to the minimization procedure, a remark we have made in a previous study [6] can be used, namely, in the case of numerical applications, instead of dealing with the error value (which may oscillate or increase, even if the iterative process is convergent), it is the value of the

functional that must be followed, in order to ensure that this decreases, meaning that the process is convergent. Moreover, according to the performed numerical experiments, we have found, that the vectors which in accordance with the imposed conditions, should be orthogonal, do not satisfy accurately this condition (because of the rounding errors). In order to emphasize that all the computing formulae may be obtained by minimizing a functional, hence without resorting to conditions which are not satisfied in the course of iterations, we shall give a very simple proof, which is not known in literature. Moreover this demonstration does not require special mathematical knowledge. We shall examine two computing variants.

2.2. Establishing of the computing formulae for the first variant

Let us search for the iterative solution of a system of n linear equations with n unknowns:

$$\mathbf{A}\mathbf{x} = \mathbf{b}. \tag{1}$$

Generally the system of equations is assumed a normal one in the known sense [1, p. 306], that is matrix $\mathbf{A}$ is symmetric and positive definite. The case of an unsymmetric matrix can be examined separately.

The solution of the system of equations (1) minimizes the functional

$$F = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{x}^T\mathbf{b}, \tag{2}$$

where $\mathbf{x}$ is any vector, which can be relatively simply established. The quantity $\mathbf{x}$ which minimizes the functional (2), may be searched by an iterative method. The solution for any iteration m will be searched in the form

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + a_m\mathbf{p}^{(m)}, \tag{3}$$

with

$$\mathbf{p}^{(m)} = \mathbf{z}^{(m)} + c_m\mathbf{p}^{(m-1)}, \tag{4}$$

and it will be adopted

$$\mathbf{M}\mathbf{z}^{(m)} = \mathbf{r}^{(m)}, \tag{5}$$

$$\mathbf{p}^{(0)} = \mathbf{z}^{(0)}, \tag{6}$$

$$\mathbf{r}^{(m)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(m)}, \tag{7}$$

where the symbols have the following meanings: $\mathbf{M}$ is a symmetric matrix, invertible, of the same order (rank) as matrix $\mathbf{A}$, the quantities $\mathbf{p}$ and $\mathbf{z}$ represent the vectors along the directions of which the minimization is to be performed, and $\mathbf{r}$ represents the rest or residual. From relation (4) it can be seen that at every iteration, the previous iteration is taken into account.

In the particular case when one takes $\mathbf{M} = \mathbf{A}$ and $\mathbf{M}$ is easily invertible, and $a_0 = 1$ then, from equations (3), (5), (7) it follows immediately that after the first iteration, one obtains the solution.

It is useful to be mentioned, that the previous relations are not of an arbitrary nature. With that end in view, we shall remark that the solution of the system of equations (1) is

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}. \tag{a}$$

With these symbols, relations (7), (5), (4), become:

$$\mathbf{r}^{(m)} = \mathbf{A}(\mathbf{x} - \mathbf{x}^{(m)}), \tag{b}$$

$$\mathbf{z}^{(m)} = \mathbf{M}^{-1}\mathbf{A}(\mathbf{x} - \mathbf{x}^{(m)}), \tag{c}$$

$$\mathbf{p}^{(m)} = \mathbf{M}^{-1}\mathbf{A}(\mathbf{x} - \mathbf{x}^{(m)}) + c_m\mathbf{p}^{(m-1)}. \tag{d}$$

From relations (a)–(d) and (3), it follows that by relation (3), at every iteration, just the deviation from the solution is compensated. From relations (7), (3), one obtains

$$\mathbf{r}^{(m+1)} = \mathbf{r}^{(m)} - a_m\mathbf{A}\mathbf{p}^{(m)}. \tag{8}$$

At the beginning one takes $\mathbf{x}^{(0)}$ and one obtains $\mathbf{r}^{(0)}$. In order to obtain the value of the quantity $\mathbf{x}$ which ensures the minimum value of the functional, we shall determine a_m and c_m so that the value of the functional should be minimum. From the minimum condition

$$\frac{\partial F}{\partial a_m} = 0, \tag{9}$$

and utilizing relation (3), we obtain successively

$$(\mathbf{p}^{(m)})^T[\mathbf{A}\mathbf{x}^{(m+1)} - \mathbf{b}] = 0, \tag{10}$$

$$(\mathbf{p}^{(m)})^T\mathbf{r}^{(m+1)} = 0, \tag{11}$$

$$(\mathbf{p}^{(m)})^T\mathbf{A}[\mathbf{x}^{(m)} + a_m\mathbf{p}^{(m)}] - (\mathbf{p}^{(m)})^T\mathbf{b} = 0. \tag{12}$$

It follows that

$$a_m = \frac{(\mathbf{p}^{(m)})^T\mathbf{r}^{(m)}}{(\mathbf{p}^{(m)})^T\mathbf{A}\mathbf{p}^{(m)}}. \tag{13}$$

Taking into account relation (4), we can write

$$(\mathbf{p}^{(m)})^T\mathbf{r}^{(m)} = [\mathbf{z}^{(m)} + c_m\mathbf{p}^{(m-1)}]^T\mathbf{r}^{(m)}. \tag{14}$$

Taking into account the relation of the form (11), we get

$$(\mathbf{p}^{(m)})^T\mathbf{r}^{(m)} = (\mathbf{z}^{(m)})^T\mathbf{r}^{(m)}. \tag{15}$$

From relations (13), (15), (5), it follows that

$$a_m = \frac{(\mathbf{z}^{(m)})^T \mathbf{M}\mathbf{z}^{(m)}}{(\mathbf{p}^{(m)})^T \mathbf{A}\mathbf{p}^{(m)}}. \tag{16}$$

From the minimum condition

$$\frac{\partial F}{\partial c_m} = 0, \tag{17}$$

we obtain successively

$$(\mathbf{p}^{(m-1)})^T[\mathbf{A}\mathbf{x}^{(m+1)} - \mathbf{b}] = 0, \tag{18}$$

$$(\mathbf{p}^{(m-1)})^T \mathbf{r}^{(m+1)} = 0, \tag{19}$$

$$(\mathbf{p}^{(m-1)})^T[\mathbf{A}\mathbf{x}^{(m)} + a_m \mathbf{A}\mathbf{z}^{(m)} + a_m c_m \mathbf{A}\mathbf{p}^{(m-1)} - \mathbf{b}] = 0, \tag{20}$$

$$(\mathbf{p}^{(m-1)})^T[-\mathbf{r}^{(m)} + a_m \mathbf{A}\mathbf{z}^{(m)} + a_m c_m \mathbf{A}\mathbf{p}^{(m-1)}] = 0. \tag{21}$$

Multiplying relation (4) by $(\mathbf{r}^{(m+1)})^T$ and utilizing relations (11), (19), (5), we get

$$(\mathbf{r}^{(m)})^T \mathbf{z}^{(m+1)} = 0. \tag{22}$$

Taking into account relation (11), the first term of relation (21) is null and it follows that

$$c_m = -\frac{(\mathbf{z}^{(m)})^T \mathbf{A}\mathbf{p}^{(m-1)}}{(\mathbf{p}^{(m-1)})^T \mathbf{A}\mathbf{p}^{(m-1)}}. \tag{23}$$

Taking into account relations of the form (8), (11), (15), the denominator of expression (23) may be expressed in terms of vector $\mathbf{z}$. Considering relations of the form (8), (22), (5), the numerator of expression (23) may be, also expressed in terms of vector $\mathbf{z}$. Finally, we get

$$c_m = \frac{(\mathbf{z}^{(m)})^T \mathbf{M}\mathbf{z}^{(m)}}{(\mathbf{z}^{(m-1)})^T \mathbf{M}\mathbf{z}^{(m-1)}}. \tag{24}$$

If the quantities given by relations (16) and (24) are always positive, it follows that the minimum conditions are satisfied. For this purpose it is necessary and sufficient that besides matrix $\mathbf{A}$, the matrix $\mathbf{M}$ should be, also, positive definite.

Similarly, other relations between the matrices $\mathbf{p}$, $\mathbf{z}$, $\mathbf{r}$, $\mathbf{A}$ may be established. For instance, from relations (8), (11), (19), it follows that

$$(\mathbf{p}^{(m-1)})^T \mathbf{A}\mathbf{p}^{(m)} = 0. \tag{25}$$

It is important to know, if in the course of computations, conditions (11), (19) are satisfied, hence to know if the values of the left-hand side are

near to zero. The usual manner, known in literature, (of considering the values of the left-hand side, so as they are obtained), cannot be accepted as satisfactory, because the results are almost always different from zero. To avoid this disadvantage, we referred to the case of the product of two vectors. In this case, for obtaining the value of the cosine of the angle between the two vectors, it is necessary to divide their scalar product by the product of their moduli. We have replaced the values of the vectors in the left-hand side of the referred relations, by their values divided by moduli of corresponding vectors. The obtained results can be compared regardless of the number of equations or of the values of the system coefficients.

Further on, we shall give the results that we have obtained in the course of iterations for relation (22), with the mentioned divisions, for the system of equations considered in Table 1 given at the end of this paper and obtained from a computation problem of an electromagnetic field.

The results are given below in Table a.

Table a: Numerical results

$$(\mathbf{r}^{(01)})^T\mathbf{z}^{(00)} = -0.1196241\text{E}{-14}$$
$$(\mathbf{r}^{(11)})^T\mathbf{z}^{(10)} = -0.1965751\text{E}{-06}$$
$$(\mathbf{r}^{(21)})^T\mathbf{z}^{(20)} = -0.1128731\text{E}{-06}$$
$$(\mathbf{r}^{(31)})^T\mathbf{z}^{(30)} = -0.2166486\text{E}{-06}$$
$$(\mathbf{r}^{(41)})^T\mathbf{z}^{(40)} = -0.4953121\text{E}{-07}$$
$$(\mathbf{r}^{(51)})^T\mathbf{z}^{(50)} = -0.3179850\text{E}{-07}$$
$$(\mathbf{r}^{(61)})^T\mathbf{z}^{(60)} = -0.6660150\text{E}{-07}$$

If we had not performed the specified divisions, we would have obtained other results, given in Table b, for instance.

Table b: Numerical results

$$(\mathbf{r}^{(01)})^T\mathbf{z}^{(00)} = -0.4862583\text{E}{-12}$$
$$(\mathbf{r}^{(11)})^T\mathbf{z}^{(10)} = -0.6910249\text{E}{-03}$$
$$(\mathbf{r}^{(21)})^T\mathbf{z}^{(20)} = -0.4738768\text{E}{-04}$$
$$(\mathbf{r}^{(31)})^T\mathbf{z}^{(30)} = -0.4370694\text{E}{-05}$$

Also, in the numerical experiments performed relating to formulae (13) and (16), we have found that relation (16) leads to a better convergence than relation (13).

2.3. Establishing of the computing formulae for the second variant

In order to obtain another expression of the solution, it is possible to proceed as follows. One considers the relations obtained from formula (3) for m and $m-1$, one expresses $\mathbf{p}^{(m)}$ and $\mathbf{p}^{(m-1)}$. Then, with the help of formula (4), one eliminates $\mathbf{p}^{(m)}$ and $\mathbf{p}^{(m-1)}$. It results that:

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m-1)} + \omega_{m+1}[\alpha_m \mathbf{z}^{(m)} + \mathbf{x}^{(m)} - \mathbf{x}^{(m-1)}], \tag{26a}$$

$$\alpha_m = \frac{a_m}{1 + \frac{a_m c_m}{a_{m-1}}}; \quad \omega_{m+1} = 1 + \frac{a_m c_m}{a_{m-1}}; \tag{26b–c}$$

$$\alpha_m = \frac{a_m}{\omega_{m+1}}; \quad \omega_{m+1} = \frac{1}{1 - \frac{\alpha_m}{\alpha_{m-1}\omega_m} c_m}; \quad \omega_1 = 1. \tag{26d–f}$$

In order to obtain a more simple expression for α_m we can proceed as follows. We multiply both sides of relation (26 a) first by $\mathbf{A}$ and then by $(\mathbf{z}^{(m)})^T$ and we emphasize the quantities $\mathbf{r}^{(m)}$ and $\mathbf{r}^{(m-1)}$ instead of $\mathbf{x}^{(m)}$ and $\mathbf{x}^{(m-1)}$. Taking into account relations (22) and (5), we obtain immediately the expression

$$\alpha_m = \frac{(\mathbf{z}^{(m)})^T \mathbf{M} \mathbf{z}^{(m)}}{(\mathbf{z}^{(m)})^T \mathbf{A} \mathbf{z}^{(m)}}. \tag{26g}$$

These results are those obtained for the variant of point 2^0 of section 2, and established in literature starting from quite different considerations.

Because relations (3) and (26) have been obtained above starting from the same formulae, we must expect to obtain the same results when applying the two methods, apart from the rounding errors. When applying the two variants, for enough complicated cases, including those of paper [6], we have found that the rounding errors have not had any influence and the results are exactly the same.

Also, it is possible to establish the general relations

$$(\mathbf{p}^{(i)})^T \mathbf{r}^{(j)} = 0; \quad \forall i \in [0, n-1],\ j \in [1, n];\ i < j; \tag{27}$$

$$(\mathbf{p}^{(i)})^T \mathbf{A} \mathbf{p}^{(j)} = 0; \quad \forall i \in [0, n-1],\ j \in [1, n];\ i < j. \tag{28}$$

We have shown [6], that these relations may be obtained directly by utilizing conveniently relations (4), (5), (8), (11), (19). The deduction may be done by an inductive reasoning, considering successively, the relations obtained for the superscript indices of orders: $i = 0, 1, 2, \ldots, k$; $j = k+1$, putting $k = 0, 1, 2, \ldots, n-1$. Also, it is possible to establish, at the same time, the general relation

$$(\mathbf{z}^{(i)})^T \mathbf{M} \mathbf{z}^{(j)} = 0; \quad \forall i \neq j;\ i, j \in [0, n]. \tag{29}$$

However, for computations, only relations (11), (19), (25), derived above in the paper, are of interest.

3. Pre-conditioning of the matrix of the coefficients of the system

There are various manners for proving that the convergence of the conjugate gradient method depends on the conditioning degree (condition number) $K = \lambda_{max}/\lambda_{min}$ of the matrix $(\mathbf{M}^{-1}\mathbf{A})$ of the system of equations, where λ_{max} and λ_{min} represent the greatest and the smallest eigenvalue respectively of the matrix referred to above. For the particular case where $\mathbf{M} = \mathbf{A}$, the condition number is equal to unity and the solution is obtained after the first iteration. The closer to unity the conditioning degree of $(\mathbf{M}^{-1}\mathbf{A})$ is, the more rapidly convergent the iterative process will be.

The principles for obtaining a pre-conditioning matrix are presented in paper [8]. There are different procedures for obtaining these matrices. We have experimented several pre-conditioning matrices. We have obtained the best results with the matrix obtained by incomplete Cholesky factorization. There are several procedures for obtaining a matrix of this type.

For a better explanation of the results, we shall present shortly the utilized procedure [2, p. 207, 211], [6] adding, also, some specifications useful for the procedure. Generally, one can search to obtain the relation

$$\mathbf{A} = \mathbf{LSU}, \tag{30}$$

operation called decomposition or factorization (we shall use both terms).

For the same matrix, one can search the relation

$$\mathbf{A} = \mathbf{LSU} - \mathbf{R}, \tag{31}$$

operation called incomplete factorization, where the following symbols have been used: $\mathbf{L}$ – lower left matrix (sub-diagonal triangular matrix) with unit diagonal; $\mathbf{U}$ – upper right matrix (over-diagonal triangular matrix) with unit diagonal; $\mathbf{S}$ – diagonal matrix; $\mathbf{R}$ – matrix containing the elements deliberately not included in the matrix product factors. In the case of a symmetric matrix, the relation $\mathbf{U} = \mathbf{L}^T$ is fulfilled.

At first sight, it might seem, that the incomplete factorization would not have advantages as compared to the complete factorization, if the last were possible. In fact, the incomplete factorization is more advantageous, because it requires a number of arithmetical operations, and a storage zone, both much smaller.

The matrix assumed to serve as a pre-conditioning matrix is

$$\mathbf{A} = \mathbf{LSU}, \tag{32}$$

and the matrix $\mathbf{R}$ must not be stored. The computation of the elements of the matrices of relation (32) are performed by the known procedures [2], [3], [4], [6], [8].

3.1. Procedure for performing an incomplete Cholesky decomposition (factorization)

In this decomposition one keeps only those elements which correspond to those places of the matrix $\mathbf{A}$, the elements of which are different from zero; then, to store matrices $\mathbf{L}, \mathbf{S}, \mathbf{U}$, a storage zone, equal to that for storing the matrix $\mathbf{A}$ is necessary. In some cases, when the convergence is slow, a certain increase, called shifting, of the elements of the principal diagonal of the matrix $\mathbf{A}$ [4, p. 482] is recommendable. In the cases that we have examined, this increase has not been necessary.

4. The number of iterations necessary for obtaining the solution

To establish the number of iterations necessary for obtaining the solution, we shall search the number of the iteration for which the rest (residual) is null.

The derivation will be in accordance with the proof of the formulae and differs from those given in literature which concern the formulae without pre-conditioning, or are based on other considerations [5, p. 181].

We shall use relation (29) in the form

$$(\mathbf{r}^{(m)})^T \mathbf{M}^{-1} \mathbf{r}^{(m+1)} = 0. \tag{33}$$

The last equation may be written

$$(\mathbf{r}^{(m)})^T \mathbf{M}^{-1/2} \mathbf{M}^{-1/2} \mathbf{r}^{(m+1)} = 0. \tag{34}$$

We shall denote

$$\hat{\mathbf{r}}^{(m)} = \mathbf{M}^{-1/2} \mathbf{r}^{(m)}, \tag{35}$$

and from relations (33) and (35) we get

$$(\hat{\mathbf{r}}^{(m)})^T \hat{\mathbf{r}}^{(m+1)} = 0. \tag{36}$$

If one utilizes relations (4), (5), (3), successively, one obtains for any iteration, expressions of the form

$$\mathbf{r}^{(m)} = \sum_{k=0}^{m} c_{mk} (\mathbf{A}\mathbf{M}^{-1})^k \mathbf{r}^{(0)}. \tag{37}$$

From relations (37) and (35) we obtain

$$\hat{\mathbf{r}}^{(m)} = \sum_{k=0}^{m} c_{mk} \mathbf{M}^{-1/2} (\mathbf{A}\mathbf{M}^{-1})^k \mathbf{M}^{1/2} \mathbf{M}^{-1/2} \mathbf{r}^{(0)}. \tag{38}$$

Considering the general relation [1, p. 377]

$$\mathbf{S}^{-1}(\mathbf{A})^k\mathbf{S} = (\mathbf{S}^{-1}\mathbf{A}\mathbf{S})^k, \tag{39}$$

where $\mathbf{S}$ is a non-singular matrix, we get

$$\hat{\mathbf{r}}^{(m)} = \sum_{k=0}^{m} c_{mk}(\mathbf{M}^{-1/2}\mathbf{A}\mathbf{M}^{-1/2})^k\hat{\mathbf{r}}^{(0)}. \tag{40}$$

We shall suppose that $\hat{\mathbf{r}}^{(0)}$ is not an eigenvector of the matrix $(\mathbf{M}^{-1/2}\mathbf{A}\mathbf{M}^{-1/2})$, otherwise, one would have obtained the solution after the first iteration, $m = 1$. It may be noted that the matrices $(\mathbf{M}^{-1}\mathbf{A})$ and $(\mathbf{M}^{-1/2}\mathbf{A}\mathbf{M}^{-1/2})$, have the same eigenvalues, but different eigenvectors.

It can be mentioned that the vectors

$$\hat{\mathbf{A}}^k\hat{\mathbf{r}}^{(0)}, \quad (k = 0, 1, 2, \ldots, s-1); \quad \hat{\mathbf{A}} = \mathbf{M}^{-1/2}\mathbf{A}\mathbf{M}^{-1/2}, \tag{41}$$

form a vector basis of the linear space S of $s \leq n$ dimensions, and if the eigenvalues of the matrix $\hat{\mathbf{A}}$ are distinct, $s = n$. It follows that all the vectors $\hat{\mathbf{r}}^{(m)}$ for $m \geq s-1$ are in the same linear space.

For $m = s-1$, according to relation (36), we obtain

$$(\hat{\mathbf{r}}^{(s-1)})^T\hat{\mathbf{r}}^{(s)} = 0. \tag{42}$$

At the same time, the vector $\hat{\mathbf{r}}^{(s)}$ must satisfy a relation of the form (29), written in the form of (42), for $j = s$, and $i = 0, 1, 2, \ldots, s-1$. Taking into account that the vector $\hat{\mathbf{r}}^{(s)}$ must be in the same linear space with the vector $\hat{\mathbf{r}}^{(s-1)}$, it follows that it must be null. Hence

$$\hat{\mathbf{r}}^{(s)} = 0, \tag{43}$$

and

$$\mathbf{r}^{(s)} = 0. \tag{44}$$

Therefore the solution is obtained after $s \leq n$ iterations at the most.

5. The computation errors

The conjugate directions methods permits, if the rounding errors are not taken into consideration, i.e. in the idealized case, to obtain the solution after a number of iterations equal, at the most, to the number of equations. That is why, in the idealized case the examined methods may be considered as direct methods.

From a practical point of view, the solution may be obtained after a much smaller number of iterations, but sometimes the number of iterations may be larger than the number of equations. That is why, in a non-idealized case, the examined methods may be considered as iterative methods. Thus, in the idealized case, if in a system of n equations, one introduces $\mathbf{x}^{(0)}$ and

the solution will be obtained after exactly n iterations, then the solution will be $\mathbf{x}^{(n)}$. As it is shown in paper [8], different measures of the error can be used for evaluating the accuracy of the results and for stopping the iterative process. In this study, the errors have been computed by the formulae we have proposed and utilized previously [6].

Firstly, the values of the functional must be verified, in order to ensure that this decreases after every iteration, regardless of the values of the considered errors, meaning that the process is convergent, else the process must be interrupted. Then, the following errors have been considered.

The global relative error with respect to the right-hand side

$$e_{rb} = \sum_{i=1}^{n} |r_i^{(m)}| / \sum_{i=1}^{n} |b_i|, \quad \mathbf{r}^{(m)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(m)}. \tag{45}$$

The maximum relative error corresponding to the equation of the row i of the system, with respect to the right-hand side

$$e_{ri} = \max\Big(n|r_i^{(m)}| / \sum_{i=1}^{n} |b_i|\Big). \tag{46}$$

The maximum absolute error corresponding to the equation of the same row of the system

$$e_{ri} = \max(|r_i^{(m)}|). \tag{47}$$

6. Numerical experiments

6.1. The experimented cases

We have experimented on numerical computer, the methods examined in the present paper, for the case of the system of equations obtained from a type problem. The type problem has been represented by the computing example examined in the study [6].

The example of paper [6] refers to an enough complicated configuration of a rotating electrical machine with ferromagnetic parts for which some electromagnetic performances are to be computed. The problem results in computing the vector potential of the magnetic field for a two-dimensional domain. For this purpose it is necessary to solve a partial differential equation of the second order of elliptic type. For obtaining the solution, a finite element method, involving a discretization net (mesh) was used. The application of this method requires solving of large systems of linear equations.

6.2. The properties of the matrix of the coefficients

To facilitate the understanding, we recall that the discretization net for the mentioned example has 701 nodes, and if one leaves out the equations

corresponding to the nodes at which first kind boundary conditions (namely the vector potential) are given, there remain 636 equations.

1^0. The matrix is symmetric and positive definite and weak conditioned. The first two properties derive from the type of the mathematical problem, whereas the third depends on the conditioning degree which depends on the properties of the parts of the domain.

We have established the conditioning degree (the condition number) of the matrix **A** and we have obtained [6] the value K =0.2043377E+08.

2^0. The matrix has 636 rows and 636 columns and each row contains at the most 9 elements different from zero, thus it is a sparse matrix. The fact that each row contains 9 elements is due to the finite element method.

3^0. These 9 elements are distributed on each row of the matrix, so that the matrix does not represent a band matrix with a small width. This distribution yields from the numbering of the net nodes utilized in the finite element method. For obtaining a band matrix it is necessary to renumber the unknowns, i.e. the same, to renumber the nodes of the net nodes.

4^0. The width of the band matrix after the renumbering of the nodes was $2 \times 98 + 1$.

5^0. The order of magnitude of the matrix elements $\max|a_{ij}|$ and $\min|a_{ij}|$ considering only the elements $a_{ij} \neq 0$ are 10^7 and 10^0 respectively. In the case of the experimented method, the input data, i.e. the coefficients matrix, have been introduced in simple precision (4 bytes per word), whereas the computations have been carried out in double precision (8 bytes per word).

6.3. Considerations regarding the results

1^0. The beginning value (starting value) of the iterative procedure was always $\mathbf{x}^{(0)} = 0$.

2^0. The conjugate gradient method, in each of the two variants given by formulae (3) and (26) respectively, ensures the same precision of the results for the same number of equations.

3^0. The conjugate gradient method with pre-conditioning by incomplete Cholesky factorization, leads to very small errors. We have examined two situations: **a**. The matrix of the coefficients is sparse; **b**. The matrix of the coefficients of the same system of equation is in the form of a band matrix of width $2 \times 98 + 1$ obtained by renumbering the nodes (hence the unknowns) in the previously described manner [6]. In both situations the necessary storage zone was the same, a supplementary vector excepted (for the case of the band matrix).

One can observe, that the errors are sensibly smaller for the case of the sparse matrix, what is convenient, the renumbering being not necessary.

6.4. Numerical results

The numerical results that we have obtained are given in Table 1.

Table 1: The results obtained by solving the system of equations for the test problem

Nb.	Method	Number of iterations	Duration in time units	Relative global error	Relative maximal error	Absolute maximal error
1	Cholesky	-	1782	0.5425692E-02	0.6295747E-01	0.3263733E+01
2	Conjugate gradient	66	808	0.5829649E-04	0.1090070E-02	0.5650950E-01
	I	104	1141	0.7410503E-06	0.2041079E-04	0.1058100E-02
3	Conjugate gradient	66	808	0.6780558E-01	0.9768320E+00	0.5063920E+02
	II	104	1141	0.3460083E-04	0.4289385E-03	0.2223627E-01

Cholesky (SCHB of [6]: method for unsymmetric band matrix, simple precision.

Conjugate Gradient I (SOLVE3 or SOLVE31 of [6]): **a.** Pre-conditioning by incomplete Cholesky factorization, from paragraph 3.1. Duration of pre-conditioning 211 time units (included in the table in duration). **b.** The matrix of the system is sparse.

Conjugate Gradient II (SOLVE3 or SOLVE31 of [6]): **a.** Pre-conditioning by incomplete Cholesky factorization, from paragraph 3.1. Duration of pre-conditioning 211 time units (included in the table in duration). **b.** The matrix is of the band type.

References

[1] B. Démidovitch, I. Maron, *Éléments de calcul numérique.* (Fundamentals of Numerical Computation). Éditions Mir, Moscou, 1973.

[2] G. H. Golub, G. A. Meurant, *Résolution numérique des grands systèmes linéaires.* (Numerical Solution of Large Systems of Linear Equations). Edit. Eyrolles, Paris, 1983.

[3] P. Lascaux, R. Théodor, *Analyse numérique matricielle appliquée à l'art de l'ingénieur.* Tome 2. (Numerical Matrix Analysis Applied to the Art of Engineering). Ed. Masson, Paris, 1987.

[4] T. A. Manteuffel, *An incomplete factorization technique for positive definite linear systems,* Math. Comp. 34 (150) (1980), 473–497.

[5] G. I. Marchouk, *Méthodes de calcul numérique.* (Methods of Numerical Analysis). Editions Mir, Moscou, 1980.

[6] A. Nicolaide, *The Application of the Pre-Conditioned Conjugate Gradient Methods for Computer Aided Solving of Large Systems of Linear Equations in Electrical Engineering.* In: Proceedings of the International Conference on Optimization of Electrical and Electronic Equipments, Brasov, May 15–17 (1996), Vol. I, pp. 3–22.

A. Nicolaide, *Programs for Solving Large Systems of Equations with Sparse Coefficients Matrices in Fortran Language.* Program Package. Research and Modelling Electrotechnical Laboratory. Transilvania University of Brasov, 1991–1994.

[7] H. Werner, *Praktische Mathematik I. Methoden der linearen Algebra.* (Practical Mathematics I. Methods of Linear Algebra). Springer-Verlag, Berlin, Heidelberg, New York, 1970.

[8] Z. I. Woźnicki, *On numerical analysis of the conjugate gradient method*, Japan J. Indust. Appl. Math. 10 (3) (1993), 487–519.

[9] R. Zurmühl, S. Falk, *Matrizen und ihre Anwendungen. Teil 1: Grundlagen. Teil 2: Numerische Methoden.* (Matrices and their Applications. Part 1: Fundamentals. Part 2: Numerical Methods). Springer-Verlag, Berlin, Heidelberg, New York, Tokio, 1986.

DEPARTMENT OF ELECTRICAL ENGINEERING
TRANSILVANIA UNIVERSITY OF BRASOV
2200 BRASOV, ROMANIA

Received July 2nd, 1996.