THE SOLUTION OF LINEAR SIMULTANEOUS EQUATIONS BY MATRIX ITERATION

By J. GUEST†

[*Manuscript received April* 22, 1955]

Summary

In a recent paper Stiefel presented a method designed for a high speed computer, for solving simultaneous linear algebraic equations of the type

$$\sum_{k=1}^{n} a_{ik}u_k + l_i = 0, \quad i=1, 2, 3, \ldots, n.$$

The method proposed here arose from that paper. Moreover, since Stiefel fully examined the symmetric case, $a_{ik}=a_{ki}$, it seemed natural to develop the present theory for non-symmetric matrices also. Actually, Stiefel and Hestenes also touched on the non-symmetric problem but did not pursue the subject very far. A comparison between their method and that proposed here is given. As in Stiefel's theory the iteration ends at step n, which actually represents the exact solution provided no rounding-off errors have been committed. However, a different type of orthogonality and conjugate relation is used here as both D (i.e. the matrix $[a_{ik}]$) and its transpose D^* are operated with simultaneously. Formulae have been found for the characteristic polynomial of D and for its inverse.

I. INTRODUCTION

The problem is to solve a set of linear non-singular simultaneous algebraic equations

$$\sum_{k=1}^{n} a_{ik}u_k + l_i = 0, \quad (i=1, 2, 3, \ldots, n). \qquad \ldots\ldots\ldots\ (1)$$

For values of n up to 10 this is probably best done by well-known methods such as Crout's. For n greater than 10, and especially when automatic equipment is available, iteration methods with accelerated convergence are superior. These methods have the advantage that inevitable rounding-off errors are kept in check and at the same time iteration methods are more suitable for digital computers.

The method outlined here is based essentially on the method of " minimal iterations " as described by C. Lanczos (1950). The problem dealt with here has been recently discussed by Hestenes and Stiefel (1952). When the matrix $[a_{ik}]$ is non-symmetric it appeared advantageous to depart from his suggested procedure and an alternative method is investigated.

† Aeronautical Research Laboratories, Department of Supply, Melbourne.

II. The Definitions of the Fundamental Vectors and Parameters

The following six vectors and two parameters are introduced to start with :

$$\left.\begin{aligned} p_k &= -r_k + \varepsilon_{k-1} p_{k-1}, \qquad k \geqslant 1,\ p_0 = -r_0, \\ v_{k+1} &= v_k + \lambda_k\ p_k, \\ r_{k+1} &= r_k + \lambda_k\ D p_k, \end{aligned}\right\} \quad \ldots\ldots \quad (2)$$

$$\left.\begin{aligned} p^*_k &= -r^*_k + \varepsilon_{k-1} p^*_{k-1}, \qquad k \geqslant 1,\ p^*_0 = -r^*_0, \\ v^*_{k+1} &= v^*_k + \lambda_k\ p^*_k, \\ r^*_{k+1} &= r^*_k + \lambda_k\ D^* p^*_k, \end{aligned}\right\} \quad \ldots\ldots\ldots \quad (3)$$

$$\varepsilon_{k-1} = \frac{(r_k,\ D^* p^*_{k-1})}{(p_{k-1},\ D^* p^*_{k-1})} = \frac{(r^*_k,\ D p_{k-1})}{(p^*_{k-1},\ D p_{k-1})}, \quad \ldots\ldots\ldots\ldots \quad (4)$$

$$\lambda_k = -\frac{(r^*_k,\ p_k)}{(p_k,\ D^* p^*_k)} = -\frac{(r_k,\ p^*_k)}{(p^*_k,\ D p_k)}, \quad \ldots\ldots\ldots\ldots \quad (5)$$

where p_k is called the kth direction vector, v_k is called the kth solution vector, and r_k is called the kth residue vector. The above three vectors operate on D alone whilst p^*, v^*, and r^* operate only on D^* and carry the same names. Finally, ε_k and λ_k are, as will be shown shortly, suitable orthogonality parameters.

With the above definitions it is now possible to develop an algorism to solve a system of n equations in n unknowns.

III. The Solution of Linear Equations

It is required to solve (**1**) or, in matrix notation,

$$Du + l = 0, \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (6)$$

where D is the square matrix with elements a_{ik},

$$u \equiv (u_1,\ u_2,\ u_3,\ .\ .\ .,\ u_n),$$
$$l \equiv (l_1,\ l_2,\ l_3,\ .\ .\ .,\ l_n\ ).$$

In order to solve (**6**) it appears desirable to treat simultaneously

$$D^* u^* + l^* = 0, \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (7)$$

where D^* is the transposed matrix of D,

u^* is a different solution vector (i.e. the one associated with D^*) and usually of no interest,

l^* is a conveniently chosen vector.

The first step in the analysis is to make a guess for u. This first approximation to the solution vector u is denoted by v_0 ; whilst successive approximations will be denoted by v_k. It then follows that

$$Dv + l = r, \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (8)$$

where r is called the residue vector. Likewise for (**7**) it follows that

$$D^* v^* + l^* = r^*. \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (9)$$

Using the definitions of v_{k+1} and v^*_{k+1} of (**2**) and (**3**) it follows, using (**8**), that

$$r_k = Dv_k + l,$$

and also

$$r_{k+1} = Dv_{k+1} + l, \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (\mathbf{10})$$

whence on subtraction

$$\begin{aligned} r_{k+1} - r_k &= D(v_{k+1} - v_k) \\ &= \lambda_k D p_k \text{ by equation } (\mathbf{2}), \qquad \ldots\ldots \quad (\mathbf{11}) \end{aligned}$$

which shows that, once v_k and p_k are defined, relation (**11**) is a direct consequence, that is, of the six defining vectors only four are independent.

Likewise, therefore,

$$r^*_{k+1} = r^*_k + \lambda_k D^* p^*_k. \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (\mathbf{12})$$

The λ_k are to be chosen in such a manner that successive residuals r_{k+1} will be orthogonal to p^*_j for $j=0, 1, 2, \ldots, k$. It will be shown that this can actually be achieved by orthogonalizing r_{k+1} merely against p^*_k.

To fix λ_k it follows therefore that

$$\left. \begin{aligned} (r_{k+1}, p^*_k) &= 0, \\ (r^*_{k+1}, p_k) &= 0. \end{aligned} \right\} \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (\mathbf{13})$$

Using (**11**) and (**12**) this gives the first orthogonality parameter

$$\lambda_k = -\frac{(r^*_k, p_k)}{(p_k, D^* p^*_k)} = -\frac{(r_k, p^*_k)}{(p^*_k, Dp_k)}, \qquad \ldots\ldots\ldots\ldots \quad (\mathbf{14})$$

provided $(p_k, D^*p^*_k)$ is non-vanishing. If the denominator vanishes then either r_k is orthogonal to p^*_k or it is required to start with a new vector v^*_0. It should be remembered, however, that it is only for very exceptional v^*_0 that the above inner product would actually vanish. Of course it is still undesirable for this product to be very small. To be on the safe side in the choice of v^*_0 one should try to choose a vector which is a linear combination of all the eigenvectors of D^*; therefore it is usually best to choose for v^*_0 a vector like

$$v^*_0 = \{1, 1, 1, \ldots, 1\}.$$

In order to fix ε_{k-1}, use the defining equations of p_k and p^*_k and postulate that

$$(r_k, p^*_{k-2}) = 0 = (r^*_k, p_{k-2}).$$

Post-multiplying (**11**) by p^*_k, it follows that

$$0 = (r_{k+1}, p^*_{k-1}) = (r_k, p^*_{k-1}) + \lambda_k\{-(Dr_k, p^*_{k-1}) + \varepsilon_{k-1}(Dp_{k-1}, p^*_k)\}.$$

Using (**13**) it follows that

$$\varepsilon_{k-1} = \frac{(r_k, D^* p^*_{k-1})}{(p_{k-1}, D^* p^*_{k-1})} = \frac{(r^*_k, Dp_{k-1})}{(p^*_{k-1}, Dp_{k-1})}, \qquad \ldots\ldots\ldots \quad (\mathbf{15})$$

and therefore also

$$(p_k,\ D^*p^*_{k-1}) \quad =0, \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (16)$$

provided

$$(p_{k-1},\ D^*p^*_{k-1}) \neq 0.$$

The remarks made above on the vanishing of this product are still applicable here.

It is now possible to prove two fundamental theorems.

Theorem I

The system of residue vectors

$$\{r_0,\ r_1,\ r_2, \ldots, r_{n-1}\}$$

is mutually orthogonal to

$$\{r^*_0,\ r^*_1,\ r^*_2, \ldots, r^*_{n-1}\},$$

that is,

$$(r_i,\ r^*_j)=0=(r^*_i,\ r_j),$$

where $i,\ j=0,\ 1,\ 2, \ldots, n-1$ and $i \neq j$.

Theorem II

The system of direction vectors

$$\{p_0,\ p_1,\ p_2, \ldots, p_{n-1}\}$$

is mutually conjugate to

$$\{p^*_0,\ p^*_1,\ p^*_2, \ldots, p^*_{n-1}\},$$

that is,

$$(p_i,\ D^*p^*_j)=0=(p^*_j,\ Dp_i),$$

where $i,\ j=0,\ 1,\ 2, \ldots, n-1$ and $i=j$.

These theorems will be proved by induction. Let it be assumed that Theorem II be true for $n=k$, that is,

$$(p_k,\ D^*p^*_{k-1})=0, \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (17)$$

$$(p_k,\ D^*p^*_{k-2})=0, \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (18)$$

$$(p_k,\ D^*p^*_{k-3})=0, \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (19)$$

$$\vdots$$

$$(p_k,\ D^*p^*_0) \quad =0.$$

It is required to prove it to be true for $n=k+1$. First, it is useful to establish the following

Lemma

Prove that

$$\text{(i)}\ (r^*_k,\ r_{k+1})=0, \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (20)$$

$$\text{(ii)}\ (p^*_k,\ r_{k+1})=0, \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (21)$$

$$\text{(iii)}\ (r^*_k,\ r_{k+1})=0. \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (22)$$

(ii) follows from the definition of ε_{k-1}.

Also, taking the defining equation of p_k^* and forming the scalar product with r_{k+1} gives

$$(r_{k+1}, r_k^*) = -(r_{k+1}, p_k^*) + \varepsilon_{k-1}(r_{k+1}, p_{k-1}^*)$$

$$= 0, \qquad \qquad (23)$$

using (**13**) and (**11**).

Finally, form the scalar product with r_{k-1}^* in (**11**):

$$(r_{k+1}, r_{k-1}^*) = (r_k, r_{k-1}^*) + \lambda_k(Dp_k, r_{k-1}^*) \qquad \qquad (23\text{A})$$

$$= \lambda_k(Dp_k, r_{k-1}^*),$$

using (**23**).

But, by definition,

$$r_{k-1}^* = -p_{k-1}^* + \varepsilon_{k-2}p_{k-2}^*,$$

which, upon substitution in (**23A**), gives

$$(r_{k+1}, r_{k-1}^*) = -\lambda_k(Dp_k, p_{k-1}^*) + \lambda_k\varepsilon_{k-2}(Dp_k, p_{k-2}^*)$$

$$= 0, \qquad \qquad (24)$$

using (**17**) and (**18**), which proves (**22**).

Now, form the scalar product with $D^*p_{k-1}^*$ in

$$p_{k+1} = -r_{k+1} + \varepsilon_k p_k,$$

that is,

$$(p_{k+1}, D^*p_{k-1}^*) = -(r_{k+1}, D^*p_{k-1}^*) + \varepsilon_k(p_k, D^*p_{k-1}^*)$$

$$= -(r_{k+1}, D^*p_{k-1}^*).$$

But

$$\lambda_k D^*p_{k-1}^* = r_k^* - r_{k-1}^*,$$

whence

$$\lambda_{k-1}(p_{k+1}, D^*p_{k-1}^*) = -(r_{k+1}, r_k^*) + (r_{k+1}, r_{k-1}^*)$$

$$= 0, \qquad \qquad (25)$$

using (**23**) and (**24**).

Now

$$(p_{k+1}, D^*p_k^*) = 0 \text{ by } (\mathbf{16}),$$

and

$$(p_{k+1}, D^*p_k^*) = 0 \text{ by } (\mathbf{25}).$$

Likewise, it can be proved that p_{k+1} and p_{k-2}^* are mutually conjugate, and it can be shown at the same time that r_{k+1} and r_{k-2}^* are mutually orthogonal and so on until it is shown that p_{k+1} and p_0^* are mutually conjugate and r_{k+1} and r_0^* are mutually orthogonal. So, if the theorems be true for $n=k$, they will also be true for $n=k+1$. But the theorems are true for $k=1$, for

$$(r_1, r_0^*) = 0 \text{ by the choice of } p_0 \text{ and } (\mathbf{13})$$

and also

$$(p_1, D^*p_0^*) = 0 \text{ by } (\mathbf{16}).$$

Hence the theorems are true for all n.

Theorem III

Asterisks can always be interchanged from one side of an inner product to the other.

Using the definition of p_k and p_k^* it follows immediately by using (**13**) that

$$(p_k, r_k^*)=(p_k^*, r_k).$$

It follows by induction that the stars are interchangeable in the product (p_k, r_j^*) for $j>k$ and it will be proved presently that this product vanishes for $j<k$ in either case. The orthogonality relations of Theorem I ensure that $(r_k, r_j^*)=(r_k^*, r_j)$. That $(p_k, p_j^*)=(p_k^*, p_j)$ can be shown by induction by using the definitions of p_k and p_j^* and the fact that $p_0=-r_0$ and $p_0^*=-r_0^*$.

As regards the interchangeability of the stars in expressions like $(r_k, D^*p_j^*)$, the definition of p_k and induction again easily lead to the results :

$$\left.\begin{aligned}(r_k, D^*p_j^*)&=(r_k^*, Dp_j)=(Dr_k, p_j^*)=(D^*r_k^*, p_j),\\ (p_k, D^*p_j^*)&=(p_k^*, Dp_j),\\ (r_k, D^*r_j^*)&=(r_k^*, Dr_j).\end{aligned}\right\} \quad \ldots. \quad (\mathbf{26})$$

Therefore, it is always permissible to interchange asterisks from one side of an inner product to the other.

An interesting result of lesser importance is the following :

Theorem IV

The residue vector r_{k+1}^* is mutually conjugate to the system $\{r_i\}$ with $i=0, 1, 2, \ldots, k-1$.

This is easily proved with the help of Theorem II.

By definition

$$r_i=-p_i+\varepsilon_{i-1}p_{i-1}.$$

Forming the scalar product with r_{k+1}^* in the above and operating with D gives

$$\begin{aligned}(r_{k+1}^*, Dr_i)&=-(r_{k+1}^*, Dp_i)+\varepsilon_{i-1}(r_{k+1}^*, Dp_{i-1}),\\ &=-\{-p_{k+1}^*+\varepsilon_k(p_k^*, Dp_i)\}+\varepsilon_{i-1}\{-p_{k+1}^*+\varepsilon_k(p_k^*, Dp_{i-1})\}\\ &=0,\end{aligned}$$

since p_k^* is mutually conjugate to the system $\{p_i\}$ $i=0, 1, 2, \ldots, k-1$ by Theorem II. Hence the result.

Theorem V

For a system of n unknowns this iteration method will give the exact solution in n steps.

Every r_k is a linear combination of $r_0, Dr_0, D^2r_0, \ldots, D^{k-1}r_0$ and similarly r_k^* is a linear combination of $r_0^*, D^*r_0^*, (D^*)^2r_0^*, \ldots, (D^*)^{k-1}r_0^*$.

If r_0 and r_0^* have components along all eigenvectors and principal vectors then these chains of vectors will be linearly independent up to $k=n$. Since r_n is orthogonal to all elements of the chain $r_0^*, D^*r_0^*, (D^*)^2r_0^*, \ldots, (D^*)^{n-1}r_0^*$, it must therefore be zero. Consequently the problem must be solved in n steps.

IV. The Inverse of a Square Matrix D

We shall show that the general element of D, i.e. the inverse of D, is

$$a_{ij}=\sum_{k=0}^{n-1}\frac{p_{ki}p^*_{kj}}{(p_k,\ D^*p^*_k)}, \qquad (27)$$

where p_k is the direction vector defined in Section II and p_{ki} its ith component.

This is done by formally solving

$$Du_i=e_i$$

(where e_i is the unit vector with a 1 in the ith place and zeros elsewhere) for $i=1,\ 2,\ 3,\ .\ .\ .,\ n$.

The solution

$$Du=-f$$

can always be expressed as a linear combination of $p_0,\ p_1,\ p_2,\ .\ .\ .,\ p_{n-1}$ in the form

$$u=\alpha_0p_0+\alpha_1p_1+\alpha_2p_2+.\ .\ .+\alpha_np_{n-1}. \qquad (28)$$

Assume this to be the case, for it will always happen unless the iteration procedure terminates before n steps, i.e. very exceptionally. Now find the α_i by using the biconjugate relation of the p_k and p^*_k (i.e. Theorem II).

For, by post-multiplying (28) by $D^*p^*_j$ it follows that

$$\alpha_j(p_j,\ D^*p^*_j)=(u,\ D^*p^*_j)=(Du,\ p^*_j)=-(f,\ p^*_j), \qquad (29)$$

whence

$$\alpha_j=-\frac{(f,\ p^*_j)}{(p_j,\ D^*p^*_j)}. \qquad (30)$$

Let now

$$-f=e_i,$$

with $i=1,\ 2,\ 3,\ .\ .\ .,\ n$ in turn, and let the corresponding u be u_i, i.e. $Du_i=e_i$. Then the matrix whose columns are u_i is really the inverse of D. If the jth component of p_k is p_{kj} then the jth component of u_{ij} is given by

$$\sum_{\rho=0}^{n-1}\alpha_\rho p_{\rho j}=\sum_{\rho=0}^{n-1}\frac{(e_i,\ p^*_\rho)}{(p_\rho,\ D^*p^*_\rho)}p_{\rho j}.$$

But $(e_i,\ p^*_\rho)\equiv p^*_{\rho i}$, which, on substitution, gives the right-hand side of expression (27).

V. The Characteristic Equation of D

Let $q_k(x)$ be a polynomial of degree k, and let $q_{k+1}(x)$ be related to $q_k(x)$ in the same way as r_{k+1} is related to r_k. Thus

$$r_{k+1}=(1+\gamma_k-\lambda_kD)r_k-\gamma_kr_{k-1} \qquad (31)$$

is replaced by

$$q_{k+1}(x)=(1+\gamma_k-\lambda_kx)q_k(x)-\gamma_kq_{k-1}(x), \qquad (32)$$

where

$$\gamma_k=\frac{\lambda_k\varepsilon_{k-1}}{\lambda_{k-1}}, \qquad (33)$$

and the λ_k and ε_k are as defined earlier. Equation (**31**) follows from the recurrence relations

$$r_{k+1}=r_k+\lambda_k D p_k, \quad \ldots \ldots \quad (\mathbf{11\ bis})$$
$$p_k=-r_k+\varepsilon_{k-1}p_{k-1}, \quad \ldots \ldots \quad (\mathbf{2\ bis})$$

and the transformed version of the first by replacing $k+1$ by k, that is,

$$Dp_{k-1}=\frac{1}{\lambda_{k-1}}\{r_k-r_{k-1}\}, \quad \ldots \ldots \quad (\mathbf{11\ ter})$$

on substituting p_k of (**2 bis**) into (**11 bis**) and by subsequently substituting Dp_{k-1} as given in (**11 ter**).

It is seen that $r_n=0$, but also $r_n=\varphi_n(D)r_0$, where φ_n is a polynomial of degree n in D. If r_0 has components in the directions of all the eigenvectors of D then by the argument of Silberstein (1952) it follows that

$$\varphi_n(D)=0, \quad \ldots \ldots \quad (\mathbf{34})$$

and hence, by the Cayley-Hamilton theorem,

$$\varphi_n(x)=0 \quad \ldots \ldots \quad (\mathbf{35})$$

is the characteristic equation of D. By definition (**32**)

$$q_n(x)=\varphi_n(x).$$

Hence the characteristic equation of D is given by

$$q_n(x)=0,$$

with

$$q_0(x)=1,$$
$$q_1(x)=1-\lambda_0 x,$$

and the later $q_k(x)$ are developed by the recurrence relation as given by equation (**32**).

VI. An Illustrative Example

It is desired to solve the following system :†

$$\begin{bmatrix} 22 & -14 & 2 \\ -7 & 15 & -5 \\ 2 & -10 & 6 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}-\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}=0. \quad \ldots \quad (\mathbf{36})$$

The corresponding transpose equation is

$$\begin{bmatrix} 22 & -7 & 2 \\ -14 & 15 & -10 \\ 2 & -5 & 6 \end{bmatrix}\begin{bmatrix} v_1^* \\ v_2^* \\ v_3^* \end{bmatrix}-\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}=0. \quad \ldots \quad (\mathbf{37})$$

For convenience, start with $v_0=0$ and $v_0^*=0$. This gives now rise to the following system of vectors which are recorded in Table 1.

† This relates to the deflection of a clamped square plate. The finite difference equivalent of the governing differential equation $\nabla^4 w-p/D=0$ had to be satisfied at nine equally spaced inner points. This made the original matrix of order 9×9. By symmetry considerations this was condensed to the above 3×3 non-symmetric matrix. For convenience the constant $625p/Da^4$ was put equal to one.

TABLE 1

ALGORISM FOR SOLVING THREE LINEAR SIMULTANEOUS EQUATIONS IN THREE UNKNOWNS

k	r_k^*	p_k	$D^*p_k^*$	r_k	p_k^*	Dp_k	v_k	v_k^*
	−1	+1	+17	−1	+1	+10	0	0
0	−1	+1	− 9	−1	+1	+ 3	0	0
	−1	+1	+ 3	−1	+1	− 2	0	0
	+3·636363636	+0·669421488	−63·074380158	+1·727272727	−1·239669421	−13·487603302	+0·2727272727	+0·2727272727
1	−3·454545455	+2·578512397	+79·338842974	−0·1818181818	+5·851239670	+14·280991739	+0·2727272727	+0·2727272727
	−0·1818181818	+3·942148760	−16·264462810	−1·545454545	+2·578512397	− 0·793388434	+0·2727272727	+0·2727272727
	−0·9801762125	−0·4136505650	+10·017659958	+0·7400881057	+0·3756622489	− 7·563896053	+0·3217235683	+0·1819933920
2	+2·3524229095	+0·3939529208	−24·042383900	+0·8634361242	+0·5008830003	− 8·824545403	+0·4614537446	+0·7009911896
	−1·372246697	+3·525878634	+14·024723948	−1·6035242294	+2·629635742	+16·388441466	+0·5612610133	+0·4614537446
	0	0	0	0	0	0	+0·28125	+0·21875
3	0	0	0	0	0	0	+0·5	+0·75
	0	0	0	0	0	0	+0·90625	+0·71875

$\varepsilon_0 = +2\cdot396694215$ $\lambda_0 = +0\cdot2727272727$

$\varepsilon_1 = +0\cdot4876412642$ $\lambda_1 = +0\cdot07319199709$

$\varepsilon_2 = 0$ $\lambda_2 = +0\cdot09784482755$

It is seen that by operating simultaneously with D and D^* not only the desired solution for the $\mathbf{v}_k$ is obtained but also the one for $\mathbf{v}_k^*$ as well. This is as it should be because of the mutual orthogonality relation with the residue vectors ; $\mathbf{r}_3^*$ had to be zero here, hence $\mathbf{v}_3^*$ gave the exact solution to (**37**).

Furthermore, now that all ε_i and λ_i have been computed it is an easy matter to obtain the inverse of D.

Let the inverse matrix be given by

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

Then find successively (using the main result of Section IV) :

$$a_{11}=\sum_{\mu=0}^{2}\frac{p_{\mu 1}p^*_{\mu 1}}{(p_\mu,\ D^*p^*_\mu)}=\frac{1}{11}-\frac{0\cdot 8298612384}{98\cdot 23591284}-\frac{0\cdot 1553929015}{35\cdot 83409646}$$
$$=+0\cdot 078125,$$
$$a_{12}=\sum_{\mu=0}^{2}\frac{p_{\mu 1}D^*p^*_2}{(p_\mu,\ D^*p^*_\mu)}=\frac{1}{11}+\frac{3\cdot 916945567}{98\cdot 23591284}-\frac{0\cdot 2071905361}{35\cdot 83409646}$$
$$=+0\cdot 125.$$

Similarly, it is found that

$$\begin{aligned} a_{13}&=+0\cdot 078125, \\ a_{21}&=+0\cdot 0625\quad, \\ a_{22}&=+0\cdot 25\quad, \\ a_{23}&=+0\cdot 1875\quad, \\ a_{31}&=+0\cdot 078125, \\ a_{32}&=+0\cdot 375\quad, \\ a_{33}&=+0\cdot 453125. \end{aligned}$$

Finally, let the characteristic polynomial for the above matrix be computed (using the main result of Section V).

It is found that

$$\begin{aligned} q_0(x)&=1, \\ q_1(x)&=1-0\cdot 2727272727x, \\ q_2(x)&=1-0\cdot 5213381059x+0\cdot 01996145375x^2, \\ q_3(x)&=1-0\cdot 78125x+0\cdot 083984375x^2-0\cdot 001953125x^3, \end{aligned}$$

that is,

$$x^3-43x^2+400x-512=0,$$

whilst

$$\begin{aligned} \gamma_1&=+0\cdot 6432023988, \\ \gamma_2&=+0\cdot 6518906069. \end{aligned}$$

VII. A Comparison with Stiefel's Method

Hestenes and Stiefel (1952) have also briefly discussed the non-symmetrical case. They arrived at the following iteration formulae :

$$\left.\begin{aligned} r_0&=k-Ax_0, \quad p_0=A^*r_0, \\ a_i&=\frac{|A^*r_i|^2}{|Ap_i|^2}, \\ v_{i+1}&=v_i+a_ip_i, \\ r_{i+1}&=r_i-a_iAp_i, \\ b_i&=\frac{|A^*r_{i+1}|^2}{|A^*r_i|^2}, \\ p_{i+1}&=A^*r_{i+1}+b_ip_i. \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots \quad (38)$$

It was next attempted to solve the following system of six equations in six unknowns† by

(i) Stiefel's method,
(ii) the present method.

$$\begin{bmatrix} +22 & -16 & +2 & +2 & 0 & 0 \\ -8 & +23 & -7 & -8 & +3 & 0 \\ +1 & -7 & +13 & +2 & -6 & +1 \\ +2 & -16 & +4 & +20 & -14 & +2 \\ 0 & +3 & -6 & -7 & +14 & -5 \\ 0 & 0 & +2 & +2 & -10 & +6 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{bmatrix} = \begin{bmatrix} +1 \\ +1 \\ +1 \\ +1 \\ +1 \\ +1 \end{bmatrix} \quad \ldots (39)$$

To save space only the successive $\mathbf{v}_k$ vectors will be shown in Table 2. $\mathbf{v}_0$ was taken as 0 in both cases.

The correct solution as given by Crout's method is (with a possible error of one in the last figure) :

$$\begin{bmatrix} +0\cdot385284810 \\ +0\cdot837816454 \\ +1\cdot10007911 \\ +1\cdot86431962 \\ +2\cdot47587025 \\ +3\cdot30498417 \end{bmatrix}.$$

The reason for the slower convergence in Stiefel's case is due to the fact that Stiefel's procedure is essentially a procedure with the matrix D^*D. The eigenvalues of this matrix are necessarily more widely spaced than for D alone. Consequently, the rate of convergence will be adversely affected if the constant vector l has large components along the largest and smallest eigenvectors.

† The problem with which this equation is associated is the same as that described in the footnote of Section V, the subdivision now having 25 inner points. By symmetry only six prove to be independent.

TABLE 2

SUCCESSIVE $\mathbf{v}_k$ VECTORS

$\mathbf{v}_k$	Stiefel's Method	Present Method
$\mathbf{v}_1$	+0·009353718895 −0·007152843861 +0·004401750068 +0·006052406344 −0·007152843861 +0·002200875034	+0·4285714286 +0·4285714286 +0·4285714286 +0·4285714286 +0·4285714286 +0·4285714286
$\mathbf{v}_2$	+0·06372821316 +0·03593381053 +0·009822136321 +0·003034374902 −0·03071552448 +0·01537565739	+0·4572531715 +0·6580253718 +0·6293436290 +0·8014340864 +0·7727523435 +0·7440706006
$\mathbf{v}_3$	+0·06620838207 +0·05309477145 +0·01698714733 +0·02239227478 −0·03424002185 +0·01012951991	+0·3722517070 +0·8530106900 +0·8942678642 +1·558693553 +1·473973065 +1·405684445
$\mathbf{v}_4$	+0·1161590155 +0·1209167665 +0·05414183470 +0·1122802989 +0·00582644703 −0·04490672156	+0·4946151738 +0·8280575085 +0·9873382239 +1·709336742 +2·062209289 +2·115900105
$\mathbf{v}_5$	+0·1248919599 +0·1409036315 +0·0958567300 +0·1253287238 +0·0201016119 −0·05565182276	+0·3845229637 +0·8383734999 +1·122971474 +1·841998113 +2·466182473 +3·294146561
$\mathbf{v}_6$	+0·3844214078 +0·8380317059 +1·098388256 +1·861448352 +2·473336988 +3·300624954	+0·3852848081 +0·8378164566 +1·100079112 +1·864319617 +2·475870252 +3·304984174

Numerous checking facilities are available for either method. It is, however, pointless to carry out more than the most essential checks and these are :

(i) column checks for all the included vectors,

(ii) λ checks, e.g. $(r_k^*, r_0) = 0$,

(iii) ε checks, e.g. $(p_k, D^* p_0^*) = 0$.

VIII. The Correction of Rounding-off Errors

Rounding-off errors may become quite serious, in particular for large n. These types of errors can, however, be minimized by using an artifice due to Stiefel.

Let it be assumed that step i has just been completed in the computation and it is subsequently found that

$$(Dp_{i-1}, p_i^*) \neq 0, \qquad \text{(40)}$$

but is fairly small of course. (If this is not so the error is due to the computer.)

It is now desirable to redefine λ_i and ε_i in such a manner that

$$(r_i, r_{i+1}^*) = 0, \qquad \text{(41)}$$

that is, assuring that r_{i+1}^* will be orthogonal to the old r_i vector, and

$$(Dp_i, p_{i+1}^*) = 0, \qquad \text{(42)}$$

that is, assuring that p_{i+1}^* will be orthogonal to the old Dp_i vector. It is necessary to prove the following

Lemma

(i) $$(r_i, r_{i+1}^*) = (r_i, r_i^*) - \lambda_i(p_i, D^*p_i^*) + \varepsilon_{i-1}\lambda_i(p_{i-1}, D^*p_i^*), \qquad \text{(43)}$$

(ii) $$\lambda_i(Dp_i, p_{i+1}^*) = -(r_{i+1}, r_{i+1}^*) + (r_i, r_{i+1}^*) + \varepsilon_i\lambda_i(Dp_i, p_i^*). \qquad \text{(44)}$$

Using the definition of r_i and post-multiplying with r_{i+1}^* gives

$$(r_i, r_{i+1}^*) = -(p_i, r_{i+1}^*) + \varepsilon_{i-1}(p_{i-1}, r_{i+1}^*). \qquad \text{(45)}$$

Now substitute for $r_{i+1}^* = r_i^* + \lambda_i D^* p_i^*$,

$$\begin{aligned}(r_i, r_{i+1}^*) &= -(p_i, r_i^*) - \lambda_i(p_i, D^*p_i^*) \\ &\quad + \varepsilon_{i-1}(p_{i-1}, r_i^*) + \varepsilon_{i-1}\lambda_i(p_{i-1}, D^*p_i^*) \\ &= (r_i, r_i^*) - \lambda_i(p_i, D^*p_i^*) + \varepsilon_{i-1}\lambda_i(p_{i-1}, D^*p_i^*), \qquad \text{(46)}\end{aligned}$$

using the definition of r_i^*. This proves (**43**).

Also by pre-multiplying the definition of p_{i+1}^* by Dp_i the following relation results:

$$(Dp_i, p_{i+1}^*) = -(Dp_i, r_{i+1}^*) + \varepsilon_i(Dp_i, p_i^*). \qquad \text{(47)}$$

Now substitute for $\lambda_i Dp_i = r_{i+1} - r_i$ in (**47**). This gives

$$\lambda_i(Dp_i, p_{i+1}^*) = -(r_{i+1}, r_{i+1}^*) + (r_i, r_{i+1}^*) + \varepsilon_i\lambda_i(Dp_i, p_i^*),$$

which proves (**44**).

Next introduce a λ_i' which is slightly different from λ_i and choose it such that

$$(r_i, r_{i+1}^*) = 0. \qquad \text{(41 bis)}$$

Using result (i) of the lemma and also the fact that $(r_i, r_i^*)=(r_i^*, p_i)$ in (**10**) yields at once

$$\lambda_i' = \frac{+(r_i, r_i^*)}{(p_i, D^*p_i^*)-\varepsilon_{i-1}(p_{i-1}, D^*p_i^*)} \quad \dots\dots\dots\dots \quad (\mathbf{48})$$

$$= \frac{\lambda_i}{d_i}, \quad \dots\dots\dots\dots \quad (\mathbf{49})$$

where

$$d_i = 1-\varepsilon_{i-1}\frac{(p_i^*, Dp_{i-1})}{(p_i, D^*p_i^*)} \quad \dots\dots\dots\dots \quad (\mathbf{50})$$

There is now a refined λ_i' at our disposal which assures that $(r_i, r_{i+1}^*)=0$ or at any rate is much smaller than it had been before the correction term was applied.

Further, let an ε_i' be now introduced—slightly different from the old ε_i—which shall be chosen such that

$$(Dp_i, p_{i+1}^*)=0. \quad \dots\dots\dots\dots \quad (\mathbf{42\ bis})$$

Now equation (**44**) holds for *both* ε_i and ε_i' and also λ_i and λ_i', i.e.†

$$\varepsilon_i\lambda_i(Dp_i, p_i^*) = +(r_{i+1}, r_{i+1}^*), \quad \dots\dots\dots\dots \quad (\mathbf{51})$$

$$\varepsilon_i'\lambda_i'(Dp_i, p_i^*) = \quad (r_{i+1}, r_{i+1}^*). \quad \dots\dots\dots\dots \quad (\mathbf{52})$$

Hence using (**51**), (**52**), and (**49**) we obtain

$$\varepsilon_i' = \varepsilon_i d_i. \quad \dots\dots\dots\dots \quad (\mathbf{53})$$

Thus both λ_i and ε_i have been refined for rounding-off errors.

IX. Conclusions

The above method of solving systems of n equations in n unknowns seems to be well suited for an electronic high speed computer, since once a programme for an affine transformation has been devised the rest is quite straightforward. However, the method is not very fast. In fact, compared with one of the pivotal condensation methods the present approach requires nearly three times as many more multiplications. Against that should be weighed the undoubted advantage of having control of round-off errors. The method is therefore not suitable for desk machines for that reason. A good computer may complete a 10×10 matrix in about 8 working hours when using the usual Crout's method approach but would spend about five times that time on the above method. It is important to keep some checking facilities going when proceeding from one step to the next. It is considered desirable to carry all column checks, one bi-orthogonality test, and one test for the biconjugate relation. It will be found that the effect of rounding-off errors becomes rather appreciable as n increases, but this can

† The second term on the right-hand side must vanish by (**41**) for λ_i' or by the bi-orthogonality relations of the r_k and r_j^* for λ_i.

be overcome to a large extent by going beyond n steps. Lanczos (1950) suggests a test for orthogonality by adding to b_i a correction term ε_{ij} as defined by

$$\varepsilon_{ij}=-\frac{(b_i,\ b_j)}{(b_j,\ b_j)}b_j,$$

if b_i is appreciably lacking in orthogonality with another vector b_j of whose orthogonality we are certain. This, however, has the obvious weakness that while $(b_i,\ b_j)=0$ now, the new b_i will disturb the previous orthogonality so that in fact nothing better has been gained in the end.

The present method, outlined above, is, in general, superior to Stiefel's as pointed out in Section VI, but some disadvantages of the method must also be mentioned.

(i) As compared with Stiefel's method, its computing time is slightly longer.

(ii) The method may fail altogether if

$$(p_k,\ D^*p_k^*)=0,$$

which is, however, rather unlikely.

X. Acknowledgment

The author would like to take this opportunity to thank the Chief Scientist, Department of Supply, for permitting that this paper be published in an outside Australian journal.

XI. References

Hestenes, M. R., and Stiefel, E. (1952).—Method of conjugate gradients for solving linear systems. U.S. Nat. Bur. Stand. Rep. 1659.

Lanczos, C. (1950).—An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bur. Stand., Wash.* **45**: 255–82.

Silberstein, J. P. O. (1952).—On the method of minimal iterations. Dep. Supply Aust. Aeron. Res. Lab. Rep. SM 200.