On the solution of systems of linear equations by iteration

Helmut Wittmeyer

Zamm.-Z Aengew. Math. Me., v 16, (1936), pp. 301-310

Introduction

1. Objective. In the paper "Practical Methods for Equation Solving" by R. v. Mises and H. Pollaczek-Geiringer in Zeitschrift für Angewandte Mathematik und Mechanik, Vol. 9 (1929), pp. 58 to 77, the authors investigate the following iteration process (0):

$$\begin{aligned} x_1^{(2)} &= x_1^{(1)} + c_1 L_1(x^{(1)}) \\ \vdots &= \vdots \\ x_n^{(2)} &= x_n^{(1)} + c_n L_n(x^{(1)}) \end{aligned}$$

to solve the linear, inhomogeneous system of equations

$$\begin{aligned} L_1(x) &= a_{1,1}x_1 + a_{1,2}x_2 + \cdots a_{1,n}x_n - r_1 = 0 \\ \vdots &= \vdots \\ L_n(x) &= a_{n,1}x_1 + a_{n,2}x_2 + \cdots a_{n,n}x_n - r_n = 0 \end{aligned}$$

They derive some convergence conditions and give an estimate of the error when stopping the iteration process at the ith step (Theorems 1 to 5). - In the following iteration methods for systems of linear equations are considered from another general point of view, giving some new results.

2. Notation.

$A, B, C, \ldots$ = real[1] square, n-row matrices. The elements of the matrices are denoted by the corresponding small Latin letters, e.g.

$$A = (a_{i,k})$$

E = identity matrix,
A' = transposed matrix,

[1]The treatment of complex matrices does not bring any further difficulties.

λ^A = any characteristic number of the matrix A. This number is solution of the "characteristic equation": $\det(A - \lambda E) = 0$,
$\lambda^{A'A}_{\max}$ = largest of the (all non-negative) characteristic numbers of the matrix $A'A$,
$|\lambda^A|_{\max}$ = maximum of the absolute values of the characteristic numbers of the matrix A.

3. Important formulas[2]

a) From $\mathfrak{y} = A\mathfrak{x}$ it follows[3]:

$$|\mathfrak{x}| \cdot (\lambda^{A'A}_{\max})^{1/2} \leq |\mathfrak{y}| \leq |(\lambda^{A'A}_{\min})^{1/2} \cdot |\mathfrak{x}| \qquad (1)$$

b) For (real) normal (especially symmetric) matrices it holds[4]:

$$(\lambda^{A'A}_{\max})^{1/2} = |\lambda^A|_{\max}, \quad (\lambda^{A'A}_{\min})^{1/2} = |\lambda^A|_{\min} \qquad (2)$$

c) The following estimates for characteristic numbers hold[5],[6]

$$|\lambda^A|_{\min} \geq (\lambda^{A'A}_{\min})^{1/2} \geq \left(\min_{i=1,\ldots,n} |a_{i,i}|\right) - \left(\sum_{i,k=1,i\neq k}^{n} a^2_{i,k}\right)^{1/2} \qquad (3)$$

$$|\lambda^A|_{\min} \geq \left(\lambda^{A'A}_{\min}\right)^{1/2} \geq \frac{1}{2}\{\min_{i=1,\ldots,n}(2|a_{i,i}| - \sum_{k=1,k\neq i}^{n} |a_{i,k} + a_{k,i}|) - \max_{i=1,\ldots,n} \sum_{k=1}^{n} |a_{i,k} - a_{k,i}|\} \qquad (4)$$

$$|\lambda^A|_{\max} \leq \left(\lambda^{A'A}_{\max}\right)^{1/2} \leq \frac{1}{2}\left\{\max_{i=1,\ldots,n} \sum_{k=1}^{n} |a_{i,k} + a_{k,i}| + \max_{i=1,\ldots,n} \sum_{k=1}^{n} |a_{i,k} - a_{k,i}|\right\} \qquad (5)$$

$$|\lambda^A|_{\max} \leq \max_{i=1,\ldots,n} \sum_{k=1}^{n} |a_{i,k}| \qquad (6)$$

[2]Proofs can be found in: H. Wittmeyer: Influence of the change of a matrix on the solution of the associated system of equations, as well as on the characteristic numbers and the eigenvectors, Dissertation Darmstadt, 1934 (in the following cited as "dissertation"). An abbreviated version of the dissertation can be found under the same title p. 287 ff. of this issue.

[3]see dissertation, page 6.

[4]see dissertation, page 4.

[5]see dissertation, page 35 to 36. There you will find further estimations, also those of A. Hirsch, I. Schur and U. Wegner. The estimate for $|\lambda^A|_{\min}$ contained in (3) was first proved by U. Wegner. The inequalities contained in the inequalities (3), (4), (5) and (8) between the characteristic numbers of the matrices A and $A'A$ derive from E.T. Browne: Bulletin of the American Math. Soc. 34 (1928), pages 363 to 368.

[6]Further estimates follow in a forthcoming paper.

and[7]

$$|\lambda^A|_{\max} \leq \max_{k=1,\ldots,n} \sum_{i=1}^{n} |a_{i,k}| \qquad (7)$$

Moreover

$$|\lambda^A|_{\max} \leq \left(\lambda_{\max}^{A'A}\right)^{1/2} \leq (\max_{i=1,\ldots,n} |a_{i,i}|) + \left(\sum_{i,k=1,i\neq k}^{n} a_{i,k}^2\right)^{1/2} \qquad (8)$$

§1. Sufficient convergence conditions for an iteration process of general form.

1. The above system of equations is in matrix notation:

$$A\mathfrak{x} = \mathfrak{r} \qquad (9)$$

with

$$A = (a_{i,k}); \quad \det(A) \neq 0; \quad \mathfrak{x} = (x_1, \ldots, x_n); \quad \mathfrak{r} = (r_1, \ldots, r_n),$$

and the iteration process is:

$$\mathfrak{x}^{(i+1)} = (E + DA)\mathfrak{x}^{(i)} - D\mathfrak{r}. \qquad (10)$$

In the above cited work of R. v. Mises and H. Pollaczek-Geiringer, in Eq. (10) a diagonal matrix with diagonal elements $c_1, \ldots, c_n$ is chosen for D.

2. We want to show that the iteration process (10) leads to the goal even if D is an arbitrary matrix - i.e. not necessarily a diagonal matrix - as long as the elements of D satisfy certain "convergence conditions".

a) We set[8]

$$\mathfrak{x}^{(0)} = -D\mathfrak{r} \text{ with } \det D \neq 0.$$

Then the iteration process (10) yields (11):

$$\begin{aligned}
\mathfrak{x}^{(1)} &= -(E + DA)D\mathfrak{r} - D\mathfrak{r}, \\
\mathfrak{x}^{(2)} &= -(E + DA)^2 D\mathfrak{r} - (E + DA)D\mathfrak{r} - D\mathfrak{r}, \\
\vdots &= \vdots \\
\mathfrak{x}^{(i)} &= -(E + DA)^i D\mathfrak{r} - (E + DA)^{i-1} D\mathfrak{r} - \cdots - (E + DA)D\mathfrak{r} - D\mathfrak{r}.
\end{aligned}$$

If the series arising for $i \to \infty$ converges, then

$$\mathfrak{x} = \lim_{i\to\infty} \mathfrak{x}^{(i)} = -[E + (E + DA) + (E + DA)^2 + (E + DA)^3 + \cdots]D\mathfrak{r} \qquad (12)$$

[7] This estimate has not yet been proven in the dissertation. However, it can be easily deduced from the preceding one if one notes that the matrices A and A' have the same characteristic numbers.

[8] The theorems I and II derived under the assumption of these special initial values are equally valid for arbitrary initial values $\mathfrak{x}^{(0)}$, The special choice of $\mathfrak{x}^{(0)} = -D\mathfrak{r}$ only simplifies our proofs.

For this $\mathfrak{x}$ it holds:

$$\begin{aligned} \mathfrak{x} &= -[E-(E+DA)]^{-1}D\mathfrak{r}, \\ &= [DA]^{-1}D\mathfrak{r}, \\ DA\mathfrak{x} &= D\mathfrak{r}, \\ A\mathfrak{x} &= \mathfrak{r}. \end{aligned}$$

I.e.: If the series (12) converges for any matrix D with non-vanishing determinant, then the limit vector $\mathfrak{x}$ of the iteration process (10) is the solution of the system of equations (9).

b) When does the series (12) converge?

The series

$$\left(E-\frac{1}{\lambda}B\right)^{-1} = E+\frac{1}{\lambda}B+\frac{1}{\lambda^2}B^2+\frac{1}{\lambda^3}B^3+\cdots$$

converges, as is well known, if $\left|\frac{1}{\lambda}\right| < \frac{1}{|\lambda^B|_{\max}}$, i.e. the series

$$E+B+B^2+B^3+\cdots$$

converges if

$$\frac{1}{1} < \frac{1}{|\lambda^B|_{\max}} \text{ or } |\lambda^B|_{\max} < 1.$$

We apply this criterion to the relation (12):

The series (12) converges if $|\lambda^{E+DA}|_{\max} < 1$, i.e. if all characteristic numbers of $(E+DA)$ lie entirely in the unit circle of the complex plane.

c) How is this criterion expressed with the characteristic numbers of DA?

The characteristic numbers λ of $(E+DA)$ satisfy the equation

$$\det((E+DA)-\lambda E) = 0 \text{ or } \det(DA-(\lambda-1)E) = 0.$$

The characteristic numbers μ of DA satisfy the equation

$$\det(DA-\mu E) = 0.$$

Thus, between these numbers μ and λ there is the relation

$$\mu = \lambda - 1 \text{ or } \lambda = \mu + 1.$$

In particular, it follows from $|\mu+1| < 1$ that $|\lambda| < 1$. I.e.: If the characteristic numbers μ of DA lie in the complex plane in the circle with the center -1 and the radius 1, then the characteristic numbers λ of $(E+DA)$ satisfy the inequality $|\lambda| < 1$.

3. We summarize the results of paragraph 2 in the following theorem[9]:

[9] Theorem 5 on page 68 of the paper by R. v. Mises and H. Pollaczek-Geiringer cited in the introduction is a special case of Theorem I for $D = \text{const } E$.

Theorem I. The vectors of the iteration

$$\mathfrak{x}^{(0)} = -D\mathfrak{r}, \quad \mathfrak{x}^{(i+1)} = (E + DA)\mathfrak{x}^{(i)} - D\mathfrak{r}$$

with a non-vanishing determinant of D converge to the solution $\mathfrak{x}$ of the system of equations $A\mathfrak{x} = \mathfrak{r}$ with non-vanishing determinant,

1. if $|\lambda^{E+DA}|_{\max} < 1$, i.e. if the largest absolute value of the characteristic numbers of the matrix $E + DA$ is smaller than one,

or

2. if the characteristic numbers of DA in the complex plane lie entirely in the circle with center -1 and radius 1.

Remark: If D is equal to the negative inverse of A, then $DA = -E$, and the characteristic numbers of DA are all equal to -1. Thus, if D is sufficiently close to the negative inverse of A, then for continuity reasons the characteristic numbers of DA in the complex plane will still lie entirely within the circle of radius 1 around the point -1. Therefore we can take the following illustrative rule from the second part of Theorem I:

Addition to Theorem I: The iteration process of Theorem I converges to the sought solution if the matrix D is "sufficiently close" to the negative inverse of A.

In the following, this addition will be our guide in finding suitable matrices D for the iteration process of Theorem I, and with the first part of Theorem I we will investigate whether the iteration process really converges with these matrices.

§2. Sufficient convergence conditions for the usual iteration method for systems of linear equations.

If the matrix A has a "predominant principal diagonal", it can be assumed that the inverse of the diagonal matrix

$$\begin{pmatrix} a_{1,1} & & \\ & \ddots & \\ & & a_{n,n} \end{pmatrix}$$

is sufficiently close to the inverse of the matrix A in the sense of the addition to Theorem I, i.e., that the iteration process of Theorem I converges with

$$D = -\begin{pmatrix} \frac{1}{a_{1,1}} & & \\ & \ddots & \\ & & \frac{1}{a_{n,n}} \end{pmatrix} = C.$$

This iteration process is equal to the process (0) of R. v. Mises and H. Pollaczek-Geiringer with $c_i = -1/a_{i,i}$ which, since it is the most common, we will call the "usual iteration process". - The investigation shows:

Theorem II,1. If in the matrix A the diagonal elements predominate in such a way that

$$a) \qquad \max_{i=1,\ldots,n} \left(\frac{1}{|a_{i,i}|} \sum_{k=1,k\neq i}^{n} |a_{i,k}| \right) < 1$$

or

$$b) \qquad \max_{k=1,\ldots,n} \left(\sum_{i=1,i\neq k}^{n} \left| \frac{a_{i,k}}{a_{i,i}} \right| \right) < 1$$

or

$$c) \qquad \sum_{k=1,k\neq i}^{n} \left(\frac{a_{i,k}}{a_{i,i}} \right)^2 < 1$$

then the iteration process of Theorem I converges with

$$D = C = - \begin{pmatrix} \frac{1}{a_{1,1}} & & \\ & \ddots & \\ & & \frac{1}{a_{n,n}} \end{pmatrix}.$$

Proof: One has

$$E + CA = \begin{pmatrix} 0 & -\frac{a_{1,2}}{a_{1,1}} & -\frac{a_{1,3}}{a_{1,1}} & \cdots & -\frac{a_{1,n}}{a_{1,1}} \\ -\frac{a_{2,1}}{a_{2,2}} & 0 & -\frac{a_{2,3}}{a_{2,2}} & \cdots & -\frac{a_{2,n}}{a_{2,2}} \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ -\frac{a_{n,1}}{a_{n,n}} & \cdots & \cdots & -\frac{a_{n,n-1}}{a_{n,n}} & 0 \end{pmatrix}.$$

For $|\lambda^{E+CA}|_{\max}$, therefore, according to (8), the following bound results

$$|\lambda^{E+CA}|_{\max} \leq \left(\sum_{i,k=1,k\neq i}^{n} \left(\frac{a_{i,k}}{a_{i,i}} \right)^2 \right)^{1/2}.$$

If now according to c) in Theorem II,1 the right-hand side is smaller than 1, then also the left-hand side is smaller than 1 and the iteration process in question converges on the basis of Theorem I,1. In the same way the convergence of the iteration process results from the inequality a) or b), if one uses the estimation (6) or (7) instead of (8).

The inequalities b) and c) of Theorem II, 1 are essentially the two convergence conditions found (by other means) in the work of R. v. Mises and H. Pollaczek-Geiringer cited at the beginning. There, b) is first expressed in a somewhat more general form, the proof of which, however, would have posed no new difficulties according to our method.

§3. Error estimates for the usual iteration method.

Theorem II,2. For the iteration process of Theorem I with

$$D = -\begin{pmatrix} \frac{1}{a_{1,1}} & & \\ & \ddots & \\ & & \frac{1}{a_{n,n}} \end{pmatrix}.$$

the following error estimate holds for the ith approximate solution x with respect to the true solution $\mathfrak{x}$:

$$\max_{k=1,\ldots,n} |x_k - x_k^{(i)}| \leq |\mathfrak{x} - \mathfrak{x}^{(i)}| \leq \frac{|\mathfrak{x}^{(i+1)} - \mathfrak{x}^{(i)}|}{1 - \mu_r}, \quad r = 1, 2.$$

Here

$$\mu_1 = \frac{1}{2}\left\{ \max_{i=1,\ldots,n} \left(\sum_{k=1,k\neq i}^{n} \left| \frac{a_{i,k}}{a_{i,i}} + \frac{a_{k,i}}{a_{k,k}} \right| \right) + \max_{i=1,\ldots,n} \left(\sum_{k=1,k\neq i}^{n} \left| \frac{a_{i,k}}{a_{i,i}} - \frac{a_{k,i}}{a_{k,k}} \right| \right) \right\} \tag{13}$$

and

$$\mu_2 = \left(\sum_{i,k=1,i\neq k}^{n} \left(\frac{a_{i,k}}{a_{i,i}} \right)^2 \right)^{1/2} \tag{14}$$

Proof: From (11) and (12) it follows[10]

$$\begin{aligned} \mathfrak{x} - \mathfrak{x}^{(i)} &= -(E + DA)^{i+1} D\mathfrak{r} - (E + DA)^{i+2} D\mathfrak{r} - (E + DA)^{i+3} D\mathfrak{r} - \cdots \\ &= -[E + (E + DA) + (E + DA)^2 + \cdots](E + DA)^{i+1} D\mathfrak{r}, \\ &= -[E - (E + DA)]^{-1}(E + DA)^{i+1} D\mathfrak{r}, \\ \mathfrak{x} - \mathfrak{x}^{(i)} &= [DA]^{-1}(E + DA)^{i+1} D\mathfrak{r}. \qquad (15) \end{aligned}$$

Now it follows from (11) that

$$\mathfrak{x}^{(i+1)} - \mathfrak{x}^{(i)} = -(E + DA)^{i+1} D\mathfrak{r}. \tag{16}$$

If we put this into (15), we get

$$\mathfrak{x} - \mathfrak{x}^{(i)} = -[DA]^{-1}(\mathfrak{x}^{(i+1)} - \mathfrak{x}^{(i)}). \tag{17}$$

With the help of the inequality (1) it yields

$$|\mathfrak{x} - \mathfrak{x}^{(i)}| \leq \frac{|\mathfrak{x}^{(i+1)} - \mathfrak{x}^{(i)}|}{(\lambda_{\min}^{(DA)'(DA)})^{1/2}}.$$

[10]The formula (15), which is derived here under the condition that the series (12) converges, is also valid without this condition. We verify this by substituting for $\mathfrak{x}^{(i)}$ the value from (11), which we still sum up according to the summation formula for a finite geometric series.

We can further estimate the root in the denominator by observing $D = C$ using inequality (3) or (4), whereupon the assertion is easily obtained.

Theorem II,3. Using the terms of Theorem II,2 one has the estimate:

$$\max_{k=1,\ldots,n} |x_k - x_k^{(i)}| \leq |\mathfrak{x} - \mathfrak{x}^{(i)}| \leq \frac{\mu_r}{1-\mu_r} |\mathfrak{x}^{(i)} - \mathfrak{x}^{(i-1)}|, \quad r = 1, 2. \tag{18}$$

Proof: From (16) we get

$$\mathfrak{x}^{(i)} - \mathfrak{x}^{(i-1)} = -(E + DA)^i D\mathfrak{r}.$$

If we put this into (15), we get

$$\mathfrak{x} - \mathfrak{x}^{(i)} = -[DA]^{-1}(E + DA)(\mathfrak{x}^{(i)} - \mathfrak{x}^{(i-1)}). \tag{19}$$

By applying the inequality (1) twice, it follows from this:

$$|\mathfrak{x} - \mathfrak{x}^{(i)}| \leq \frac{(\lambda_{\max}^{(E+DA)'(E+DA)})^{1/2}}{(\lambda_{\min}^{(DA)'(DA)})^{1/2}} |\mathfrak{x}^{(i)} - \mathfrak{x}^{(i-1)}|.$$

In this we note that $D = C$ and estimate the numerator of the fraction using (5) and (8), respectively, and the denominator using (4) and (3), respectively. Thus we obtain the inequality of Theorem II,3 with $r = 1$ and $r = 2$, respectively.

§4. Comparison of the estimates of §3 with an earlier one and calculation of an example.

1. The estimates of Theorem II, 2 differ in principle from the estimation in the paper cited above in that in our case the quantity $|\mathfrak{x} - \mathfrak{x}^{(i)}| = (\sum_{k=1}^n (x_k - x_k^{(i)})^2)^{1/2}$ is estimated and there the expression $\sum_{k=1}^n |x_k - x_k^{(i)}|$ is estimated.

Let us compare the estimates of our Theorem II,2 with the estimate of the cited paper by using all three to compute an upper bound on the absolute values of the components of the error vector $(\mathfrak{x} - \mathfrak{x}^{(i)})$. The estimation of the cited work results in

$$\max_{k=1,\ldots,n} |\mathfrak{x} - \mathfrak{x}^{(i)}| \leq \sum_{k=1}^n |x_k - x_k^{(i)}| \leq \frac{1}{1-\mu} \left(\sum_{k=1}^n |x_k^{(i+1)} - x_k^{(i)}| \right) \tag{20}$$

with

$$\mu = \max_{k=1,\ldots,n} \left(\sum_{i=1, i\neq k}^n \left| \frac{a_{i,k}}{a_{i,i}} \right| \right). \tag{20a}$$

As an example we choose the system of equations on p. 65 of the cited paper (some notation is changed):

$$\begin{aligned} 3x + 0.15y - 0.09z &= 6 \\ 0.08x + 4y - 0.16z &= 12 \\ 0.05x - 0.3y + 5z &= 20 \end{aligned}$$

	$x_1^{(i)}$	$x_2^{(i)}$	$x_3^{(i)}$
0. approximation=$\mathfrak{x}^{(0)}$	2	3	4
1. approximation=$\mathfrak{x}^{(1)}$	1.97	3.12	4.16
2. approximation=$\mathfrak{x}^{(2)}$	1.9688	3.127	4.1675
3. approximation=$\mathfrak{x}^{(3)}$	1.96868	3.12732	4.16793
4. approximation=$\mathfrak{x}^{(4)}$	1.96867	3.12734	4.16795

with these approximations for the usual iteration method.

2. First we calculate μ and write down the "divided" matrix $\left(\frac{a_{i,k}}{a_{i,i}}\right)$

$$\begin{pmatrix} 1 & 0.05 & -0.03 \\ 0.02 & 1 & -0.04 \\ 0.01 & -0.06 & 1 \end{pmatrix}.$$

a) Calculation of μ according to formula (20a):

The numbers $\sum_{i=1,i\neq k}^{n} \left|\frac{a_{i,k}}{a_{i,i}}\right|$ for $i = 1, 2, 3$ are 0.08, 0.06, 0.07. μ is their maximum, that is, $\mu = 0.08$.

b) Calculation of μ_1, according to formula (13)[11]:

The numbers $\sum_{k=1,k\neq i}^{n} |\frac{a_{i,k}}{a_{i,i}} + \frac{a_{k,i}}{a_{k,k}}|$ for $i = 1, 2, 3$ are 0.06, 0.17, 0.12. The maximum is 0.17.

The numbers $\sum_{k=1,k\neq i}^{n} |\frac{a_{i,k}}{a_{i,i}} - \frac{a_{k,i}}{a_{k,k}}|$ for $i = 1, 2, 3$ are 0.07, 0.05, 0.06. The maximum is 0.07.

Hence $\mu_1 = \frac{0.17+0.07}{2} = 0.12$.

c) Calculation of μ_2 according to formula (14):

$$\mu_2 = (5^2 + 3^2 + 4^2 + 2^2 + 1^2 + 6^2)^{1/2} \cdot 10^{-2} = 0.096.$$

3. We estimate the error of the 3rd approximation:

a) Using the estimation of the cited work (see formula (20)):

$$\max |x_k - x_k^{(3)}| \leq \frac{(1 + 2 + 2) \cdot 10^{-5}}{1 - 0.08} = 5.5\ 10^{-5}.$$

b) Using the first estimation of theorem II, 2:

$$\max |x_k - x_k^{(3)}| \leq \frac{(1^2 + 2^2 + 2^2)^{1/2} \cdot 10^{-5}}{1 - 0.12} = 3.5\ 10^{-5}.$$

c) Using the second estimation of theorem II, 2:

$$\max |x_k - x_k^{(3)}| \leq \frac{(1^2 + 2^2 + 2^2)^{1/2} \cdot 10^{-5}}{1 - 0.0096} = 3.4\ 10^{-5}.$$

[11] For larger matrices, it is convenient to calculate μ_1, by writing down the two symmetric matrices $\left(\frac{a_{i,k}}{a_{i,i}} + \frac{a_{k,i}}{a_{k,k}}\right)$ and $\left(\frac{a_{i,k}}{a_{i,i}} - \frac{a_{k,i}}{a_{k,k}}\right)$. Cf. dissertation, p. 41.

4. A disadvantage of all estimations mentioned so far is that one can estimate with their help after the calculation of the 4th approximation only the error of the 3rd (not the 4th). The theorem II,3 on the other hand, allows under the same conditions the estimation of the error of the 4th (last) approximation (we set $r = 2$ in the formula (18)):

$$\max |x_k - x_k^{(4)}| \leq \frac{0.096}{1 - 0.096}(1^2 + 2^2 + 2^2)^{1/2} \cdot 10^{-5} = 3.2\ 10^{-5}.$$

§5. Convergence conditions and error estimates for Hertwig's iteration method.

1. It is obvious to extend the train of thought at the beginning of the 2nd paragraph in the following way: In the 2nd paragraph the negative inverse C of the diagonal matrix

$$\begin{pmatrix} a_{1,1} & & \\ & \ddots & \\ & & a_{n,n} \end{pmatrix}$$

was chosen as matrix D in the iteration process of Theorem I. A matrix which is in general even "closer" to the matrix $-A^{-1}$ would be obtained e.g. for $n = 6$ as the negative inverse of

$$\begin{pmatrix} a_{1,1} & a_{1,2} & 0 & 0 & 0 & 0 \\ a_{2,1} & a_{2,2} & 0 & 0 & 0 & 0 \\ 0 & 0 & a_{3,3} & a_{3,4} & 0 & 0 \\ 0 & 0 & a_{4,3} & a_{4,4} & 0 & 0 \\ 0 & 0 & 0 & 0 & a_{5,5} & a_{5,6} \\ 0 & 0 & 0 & 0 & a_{6,5} & a_{6,6} \end{pmatrix}$$

which is also easy to calculate. With this as matrix D the iteration process of the theorem I would probably converge in general even faster than with $D = C$.

2. A method of this kind has already been given, namely by Hertwig[12]. To show that Hertwig's method can be regarded as a special case of our method in §1, we translate it into matrix notation:

Given the system of equations

$$A\mathfrak{x} = \mathfrak{r}, \quad \det A \neq 0 \qquad (21)$$

From the matrix A a simpler matrix B is formed by omitting especially small coefficients in A in such a way that the above system of equations decomposes into several systems of equations which can be solved unambiguously

[12]Festschrift Heinrich Müller-Breslau 1912, pp. 37 to 59.

by themselves. Now we want to represent the solution $\mathfrak{x}$ of (21) in the form of an infinite series:

$$\mathfrak{x} = \mathfrak{x} + \mathfrak{x}_1 + \mathfrak{x}_2 + \cdots$$

The $\mathfrak{x}_i$ are determined as follows[13]:

$$\begin{aligned} \mathfrak{x}_0 &= B^{-1}\mathfrak{r}, \\ \mathfrak{x}_1 &= B^{-1}(\mathfrak{r} - A\mathfrak{x}_0), \\ \vdots &= \vdots \\ \mathfrak{x}_n &= B^{-1}[\mathfrak{r} - A(\mathfrak{x}_0 + \mathfrak{x}_1 + \cdots + \mathfrak{x}_{n-1})]. \end{aligned}$$

We now make some transformations:

$$\mathfrak{x}_n = \mathfrak{x}_{n-1} - B^{-1}A\mathfrak{x}_{n-1} = (E - B^{-1}A)\mathfrak{x}_{n-1} = (E - B^{-1}A)^n B^{-1}\mathfrak{r}.$$

We still set

$$\mathfrak{x}^{(i)} = \sum_{k=0}^{i} \mathfrak{x}_k \qquad (22)$$

then

$$\mathfrak{x}^{(i)} = [E + (E - B^{-1}A) + (E - B^{-1}A)^2 + \cdots + (E - B^{-1}A)^i]B^{-1}\mathfrak{r},$$

and we see from (11) that the Hertwig iteration process follows from our general process of Theorem l, if we put

$$D = -B^{-1} \qquad (23)$$

there. Thus, by Theorem I, Hertwig's iteration method converges when

$$|\lambda^{E-B^{-1}A}|_{\max} < 1.$$

The fulfillment of this convergence condition can be verified numerically with the inequalities (6), (7) or (8).

3. Let us consider in (17) the relation (22) and set[14]

$$B = A - C \qquad (24)$$

so we obtain

$$\mathfrak{x} - \mathfrak{x}^{(i)} = \sum_{k=i+1}^{\infty} \mathfrak{x}_k = [(A-C)^{-1}A]^{-1}\mathfrak{x}_{i+1} = (E - A^{-1}C)\mathfrak{x}_{i+1} = \mathfrak{x}_{i+1} - A^{-1}C\mathfrak{x}_{i+1} \qquad (25)$$

[13]The propositions derived under this assumption are exactly the same for other initial conditions

[14]So from now on, the letter C no longer has the meaning it had in §2 and §3.

By applying the inequality (1) twice, we obtain

$$|\mathfrak{x} - \mathfrak{x}^{(i)}| \leq |\mathfrak{x}_{i+1}| \left(1 + \frac{(\lambda_{\max}^{C'C})^{1/2}}{(\lambda_{\min}^{A'A})^{1/2}}\right) \qquad (26)$$

4. Similarly, from relation (19), if we still consider that according to (23) and (24)

$$E + DA = E - B^{-1}(B + C) = -B^{-1}C,$$

we obtain

$$|\mathfrak{x} - \mathfrak{x}^{(i)}| \leq \frac{(\lambda_{\max}^{C'C})^{1/2}}{(\lambda_{\min}^{B'B})^{1/2}} \left(1 + \frac{(\lambda_{\max}^{C'C})^{1/2}}{(\lambda_{\min}^{A'A})^{1/2}}\right) |\mathfrak{x}_i|.$$

5. Quite similarly to paragraph 4, we can also obtain the following inequality

$$|\mathfrak{x} - \mathfrak{x}^{(i)}| \leq \left(\frac{(\lambda_{\max}^{C'C})^{1/2}}{(\lambda_{\min}^{B'B})^{1/2}}\right)^i \left(1 + \frac{(\lambda_{\max}^{C'C})^{1/2}}{(\lambda_{\min}^{A'A})^{1/2}}\right) |\mathfrak{x}_i|.$$

From this it follows that Hertwig's method converges when

$$\lambda_{\max}^{C'C} \leq \lambda_{\min}^{B'B}.$$

6. We summarize the results of the first five paragraphs in the following two results:

Theorem III,1 . The Hertwig iteration process

$$\begin{aligned} \mathfrak{x}_0 &= B^{-1}\mathfrak{r}, \\ \mathfrak{x}_n &= B^{-1}[\mathfrak{r} - A(\mathfrak{x}_0 + \mathfrak{x}_1 + \cdots + \mathfrak{x}_{n-1})] \\ \vdots &= \vdots \\ \mathfrak{x} &= \mathfrak{x}_0 + \mathfrak{x}_1 + \mathfrak{x}_2 + \cdots \end{aligned}$$

with $B = A - C$ and $\det B \neq 0$ converges against the solution of the system of equations $A\mathfrak{x} = \mathfrak{r}$ with $\det A \neq 0$ if one of the following inequalities exists:

$$|\lambda^{E-B^{-1}A}|_{\max} < 1 \quad \text{or} \quad \lambda_{\max}^{C'C} \leq \lambda_{\min}^{B'B}.$$

Theorem III,2. For the Hertwig iteration process with the notation as in Theorem III, 1 and

$$\mathfrak{x}^{(i)} = \sum_{k=0}^{i} \mathfrak{x}_k$$

ones has the following bounds:

$$|\mathfrak{x} - \mathfrak{x}^{(i)}| \leq \frac{(\lambda_{\max}^{C'C})^{1/2}}{(\lambda_{\min}^{B'B})^{1/2}} \left(1 + \frac{(\lambda_{\max}^{C'C})^{1/2}}{(\lambda_{\min}^{A'A})^{1/2}}\right) |\mathfrak{x}_i|. \qquad (27)$$

$$|\mathfrak{x} - \mathfrak{x}^{(i)}| \leq \left(1 + \frac{(\lambda_{\max}^{C'C})^{1/2}}{(\lambda_{\min}^{A'A})^{1/2}}\right) |\mathfrak{x}_{i+1}| \qquad (28)$$

Addition to theorems III, 1 and III,2:

If A and B are real symmetric matrices, then in the inequalities of the last two propositions one can replace $(\lambda_{\min}^{A'A})^{1/2}$, $(\lambda_{\max}^{C'C})^{1/2}$ in the above formulas the characteristic numbers by their estimates listed in the introduction.

We highlight one more important special case:

Theorem III,3. The Hertwig iteration process stated in Theorem III,1 converges to the sought solution $\mathfrak{x}$ for real symmetric matrices $A = (a_{i,k})$ and $B = (b_{i,k})$ with nonvanishing determinants when $\gamma < \beta$. Then, for the ith approximation $\mathfrak{x}^{(i)} = \sum_{k=0}^{i} \mathfrak{x}_k$ the estimation

$$\max |x_k - x_k^{(i)}| \leq |\mathfrak{x} - \mathfrak{x}^{(i)}| \leq \frac{\gamma}{\beta}\left(1 + \frac{\gamma}{\alpha}\right) |\mathfrak{x}_i|,$$

with

$$C = A - B,\ \alpha = \min_{i=1,\ldots,n}(|a_{i,i}| - \sum_{k=1,k\neq i}^{n} |a_{i,k}|),$$

$$\beta = \min_{i=1,\ldots,n}(|b_{i,i}| - \sum_{k=1,k\neq i}^{n} |b_{i,k}|),\ \gamma = \max_{i=1,\ldots,n} \sum_{k=1}^{n} |c_{i,k}|.$$

§6 Other possible iteration methods.

1. In some practical cases a matrix is offered in another way than the one treated so far, which is "sufficiently close" to the negative inverse of the coefficient matrix of the given system of equations in the sense of the addition to Theorem I. This matrix can then be used as matrix D in the iteration process of Theorem I. In structural analysis, for frequently occurring loading tasks (e.g. solution of systems of linear equations in the calculation of a continuous beam on several supports), the inverse matrix is occasionally available for a given matrix in a ready-calculated form, so that with its help, given any right-hand side of the system of equations to be solved, one can quickly calculate the unknowns. If now the accuracy of this "Zahlenrechtecks" (as it is called there) is not sufficient, or if one wants to calculate a system, which is close to the present calculated one, then one can use the iteration process of theorem I for this calculation. For D the negative of the existing "Zahlenrechtecks" is to be taken[15].

2. We find a completely different suitable matrix D by the following consideration: The matrix $A'A$ with nonvanishing $\det A$ has only positive

[15]For the practical execution of the iteration process in the case treated here, it is best to proceed according to Hertwig's method (see Theorem III, I), where one then has to use the "Zahlenrechtecks" for B^{-1}.

characteristic numbers. The characteristic numbers of the matrix $c\,A'A$ with $c < 0$ therefore lie in the negative part of the complex plane. If now in addition $|c| < 2/\lambda_{\max}^{A'A}$ is chosen, then all characteristic numbers of $c\,A'A$ lie in the complex number plane completely in the circle with the radius one and the center -1. For the suitable choice of c we can use the well-known bounds of $\lambda_{\max}^{A'A}$ (see introduction). Then, according to Theorem I, the iteration process must converge with $D = c\,A'$.

Similar considerations can be made for the case that the matrix A is real symmetric and positive definite. We then set $D = c\,E$ with $c < 0$ and $|c| < 2/\lambda_{\max}^{A'A}$. Thus, given the inequalities (5), (6), and (8), we can express the theorems:

Theorem IV,1. The iteration process of Theorem I converges for $D = c\,A'$ with $c < 0$ and

$$|c| \leq \frac{8}{\left(\max_{i=1,\ldots,n} \sum_{k=1}^{n} |a_{i,k} + a_{k,i}| + \max_{i=1,\ldots,n} \sum_{k=1}^{n} |a_{i,k} - a_{k,i}|\right)^2}$$

or

$$|c| \leq \frac{2}{\left[\left(\max_{i=1,\ldots,n} |a_{i,i}|\right) + \left(\sum_{i,k=1,\, i\neq k}^{n} a_{i,k}^2\right)^{1/2}\right]^2}.$$

Theorem IV, 2. The iteration process of Theorem I converges for all real symmetric positive definite[16] matrices A for $D = c\,E$ with

$$|c| \leq 2/\max_{i=1,\ldots,n} \sum_{k=1}^{n} |a_{i,k}|$$

and $c < 0$ or $c > 0$ for negative definite matrices.

[16] i.e. matrices that correspond to a definite quadratic form.