RECENT DEVELOPMENTS IN RELAXATION TECHNIQUES

E. STIEFEL

I. One of the usual procedures for the solution of a system of real linear equations

$$Ax = k \tag{1}$$

is iteration. Beginning with a trial point x_0, a sequence of points x_i is to be constructed which approaches the point of exact solution $A^{-1}k$. For the sake of simplicity we shall take x_0 as the origin. In order to check the goodness of any approximation, we set x_i into the equations and form the *residual*-vector

$$r_i = k - Ax_i. \tag{2}$$

The iteration procedure consists of following a set of rules, which determine how to compute x_{i+1} from the preceding approximations. We shall call this set of rules the iteration algorithm and we limit ourselves to so-called *linear iterations* of the type

$$x_{i+1} = \sum_{j=0}^{m-1} C_i^{(j)} x_{i-j} + v_i.$$

The $C_i^{(j)}$ are matrices, which may depend upon A; the v_i are vectors. m is known as the *order* of the iteration, for exactly m preceding approximations appear on the right hand side. In particular we shall investigate second order procedures. They may be put in the form

$$x_{i+1} = B_i x_i - C_i x_{i-1} + v_i. \tag{3}$$

As Forsythe [1]) has pointed out in his thorough report on the solution of linear equations, it should be required that the solution be a fixed point of this transformation; that is, the point $x_{i-1} = x_i = x_{i+1} = A^{-1}k$ must satisfy equation (3). This permits the elimination of the non-homogeneous terms v_i. After some simple juggling of the terms, (3) may be written in the form

$$\Delta x_{i+1} = A_i r_i + C_i \Delta x_i, \quad A_i = (1 - B_i + C_i)A^{-1}. \tag{4}$$

Here Δx_{i+1} is the correction $(x_{i+1} - x_i)$ and the matrices A_i and C_i are arbitrary as the B_i and C_i were.

As you see in the general second-order procedure we may compute the correction from the present residual together with the preceding correction. A first order algorithm does not contain the term with Δx_i; and therefore does not use the history of the iteration. In an algorithm of higher order than the second, additional terms arise:

$$D_i \Delta x_{i-1} + E_i \Delta x_{i-2} + \ldots \tag{5}$$

II. The classical methods of iteration are all of first order:

$$\Delta x_{i+1} = A_i r_i \tag{6}$$

They differ from each other only in the choice of the matrices A_i. The more important of these methods are the following.

1) In the "Cyclic single" step procedure (*Gauss, Seidel, Nekrasov, Liebmann*) the A_i are equal to each other and are chosen as the inverse of the lower triangle of the given matrix A.

2) Less familiar is the fact that the *elimination algorithm of Gauss* is also a first order iteration procedure. Here the A_i are the reciprocals of the triangular matrices which are generated during the reduction of A.

3) In the so-called gradient methods, the A_i are scalars a_i:

$$\Delta x_{i+1} = a_i r_i. \tag{7}$$

The correction arising at each step is thus proportional to the residual. (*Richardson, Temple, Hestenes, Stein.*)

The convergence of these first-order procedures has been thoroughly investigated (*v. Mises, Collatz, Ostrowski, Weissinger*).

III. Although the convergence of a given algorithm certainly suggests its use for practical computation, it is most desirable that the convergence be monotone. No one likes to walk around in circles or in spirals. This is the point at which the principle of *relaxation* enters. The fundamental idea is to reduce the value of an appropriate measure of error at each step. In the following we shall assume the given matrix A to be *symmetrical* and *positive* definite. A suitable error measure is then the length φ of the error vector $y_i = A^{-1}k - x_i$, measured not in the cartesian sense but in the sense of the A-metric:

$$\varphi_i = (Ay_i, y_i) = (A^{-1}r_i, r_i). \tag{8}$$

The comma means the ordinary cartesian scalar product. Although the following theory could be developed with a number of other error measures, the φ-measure has two advantages:

1) It is invariant under coordinate transformations.
2) If the given system (1) is the system of normal equations for a problem in the calculus of observations, then the error measure φ is nothing else than the sum of the squares of the errors (plus an additive constant).

The three examples of first order algorithms given above are indeed relaxation methods in this sense; that is, the value of φ actually diminishes with each step. With the gradient method however, the a_i must be positive and smaller than certain bounds α_i.

Using the terminology of the theory of economic behaviour, we may say that the basic idea of relaxation is a *tactic*, in that one attempts to reduce φ by as much as possible in each single step. What we really desire, though, is a *strategy*. Of all the algorithms which have a given number n of steps, we want

to choose the one yielding x_n closest to the true solution. The determination of such a strategy is a difficult mathematical problem and there is still much work to be done in this direction. Perhaps a closer study of the relationship to the theory of games would be useful. Certain partial results have already been found, however, and it appears to me that these represent the major advances in the theory of relaxation since the last international congress at Harvard. In particular the rules for *over-relaxation* have been investigated by *Ostrowski* [2]) and *Young* [3]).

Today I should like to outline a complete solution of the strategy problem in the special case of "*scalar*" iteration schemes. By scalar iteration is meant a linear algorithm in which the matrices involved in (4) and (5) are scalars. In the case of a second order procedure, the iteration algorithm becomes

$$\Delta x_{i+1} = a_i r_i + c_i \Delta x_i. \tag{9}$$

The method resulting from the choice of the a_i and c_i as independent of i, that is $a_i = a$, $c_i = c$, was first proposed by *Frankel* [4]) and further investigated by *Hochstrasser* [5]).

IV. Now it follows from (9) that any approximation x_i must be a linear combination of the iterated vectors

$$k, Ak, A^2k, \ldots, A^{i-1}k, \tag{10}$$

assuming $x_0 = 0$. In fact, this is true for scalar linear processes of any order and characterises these processes.

We may write

$$x_i = F_{i-1}(A)k, \tag{11}$$

where F_{i-1} is a polynomial with real coefficients of degree $(i - 1)$. The residual of x_i is

$$r_i = k - Ax_i = [1 - AF_{i-1}(A)]k. \tag{12}$$

Thus we shall have to do not only with the polynomials $F_{i-1}(A)$ but also with the *residual polynomials*

$$R_i(A) = 1 - AF_{i-1}(A). \tag{13}$$

Equation (12) becomes

$$r_i = R_i(A)k. \tag{14}$$

Let us consider the polynomials F_{i-1} and R_i for a real variable λ as argument instead of the matrix A. It follows from (13)

$$R_i(\lambda) = 1 - \lambda F_{i-1}(\lambda). \tag{15}$$

Thus for $\lambda = 0$

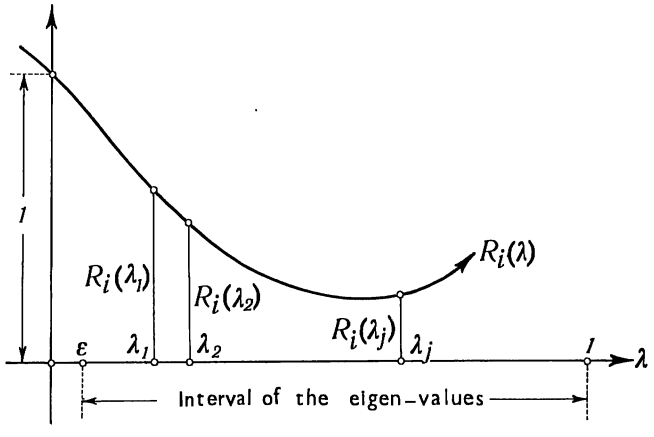
$$R_i(0) = 1. \tag{16}$$

It will turn out that this is the essential property of the residual polynomials. The residual polynomials have a simple and direct meaning in connection with the progressive approximations during the relaxation. This becomes clear if we write our equations in the coordinate system of the principle axes of A.

Since A was assumed symmetric and positive definite, its eigen-values λ_j are positive. Writing k_j for the components of k and r_{ij} for the j-*th* component of r_i we have in consequence of (14)

$$r_{ij} = R_i(\lambda_j)k_j. \tag{17}$$

Since $k = r_0$, the values $R_i(\lambda_j)$ of the residual polynomial $R_i(\lambda)$ tell by what percentage the components of the original residual have been reduced after the i-*th* iteration step (note the figure).



Our problem of finding the best strategy may now be stated somewhat inexactly as follows. A polynomial $R_n(\lambda)$ of given degree n must be found with the properties

a) $R_n(0)$ must be equal to one.

b) The values of the polynomial must be small over the interval of the λ-axis containing the eigen-values.

To state this precisely we use the error measure φ in the second form of equation (8).

The j-*th* component of $A^{-1}r_i$ is

$$\frac{R_i(\lambda_j)}{\lambda_j} k_j$$

and hence

$$\varphi_i = \sum_{(j)} \frac{R_i^2(\lambda_j)}{\lambda_j} k_j^2. \tag{18}$$

It will be convenient to use a more general measure of error ψ called the "*continuous*" error measure. In order to define it, we first need an *upper bound for the eigenvalues*, which without loss of generality may be set equal to 1. ψ is then given by

$$\psi_i = \int_0^1 \frac{R_i^2(\lambda)}{\lambda}\, \varrho(\lambda) d\lambda, \tag{19}$$

where $\varrho(\lambda)$ is an arbitrary density function defined over the interval $o \leqq \lambda \leqq 1$. The old *"discrete"* error measure φ is a special case of ψ corresponding to the density function

$$\varrho(\lambda) = \sum_{(j)} k_j^2 \cdot \delta(\lambda - \lambda_j), \tag{20}$$

where δ is the *Dirac*-function.

The following theorem can now be proved.

In the family of polynomials of *n-th* degree $R_n(\lambda)$ which satisfy $R_n(0) = 1$, the error measure ψ takes on its minimum value for the $(n + 1)$-*th* polynomial of the orthogonal set belonging to the density function $\varrho(\lambda)$.

V. This solves the strategy problem completely, but only if we are able to construct a scalar iteration algorithm yielding this minimal value of ψ after n steps.

Let $R_0(\lambda), R_1(\lambda), \ldots, R_n(\lambda)$ be the orthogonal set belonging to $\varrho(\lambda)$ such that each of the R_i satisfies (16). Note that the R_i can not be normalized in the usual sense. It is well known that three successive orthogonal polynomials are related by a recursion formula of the form

$$R_{i+1}(\lambda) = (d_i - a_i\lambda) R_i(\lambda) - c_i R_{i-1}(\lambda).$$

In consequence of (16) it follows that $d_i - c_i = 1$ and thus

$$R_{i+1}(\lambda) = (1 + c_i - a_i\lambda) R_i(\lambda) - c_i R_{i-1}(\lambda). \tag{21}$$

As given by (13) and (11) there are a sequence of polynomials $F_{i-1}(\lambda)$ and a sequence of approximations x_i associated with the $R_i(\lambda)$. Taking (14) and (21) into account, the corresponding residuals satisfy the recursion formula

$$r_{i+1} = (1 + c_i - a_i A) r_i - c_i r_{i-1} \tag{22}$$

and after proper substitutions it follows that

$$\Delta x_{i+1} = a_i r_i + c_i \Delta x_i. \tag{23}$$

By the definition (9) this is a scalar second-order process. In other words we must simply *carry out an iteration process of second order using coefficients taken from the recursion formula of the orthogonal set determined by the chosen density function.* Since this is the best possible scalar process, *it is not necessary to consider scalar procedures of higher order.*

As *Young* [6]) has remarked for a special case, the end result after n steps of our procedure may be reached by n steps of a *first order* scalar algorithm, that is, of a gradient method. Although the end values x_n coincide, the intermediate points do not. The second order procedure seems to be preferable in the sense of the adopted error-measure, since each of the intermediate points

x_i represents the better approximation after i steps. Furthermore the gradient algorithm requires the computation of the roots of $R_n(\lambda)$ but it needs less numerical work during the iteration.

VI. The selection of a density function $\varrho(\lambda)$ in the interval $0 \leqq \lambda \leqq 1$ must still be made. I wish to discuss three possibilities.

1) If no other information about the eigen-values is known except an upper bound, then we must take $\varrho(\lambda) > 0$ for $0 < \lambda < 1$ in order to cover the entire spectrum of A. Furthermore $\varrho(0)$ must be equal to 0 to insure the existence in the integral (19) and $\varrho(1)$ may vanish. These conditions suggest the form

$$\varrho(\lambda) = \lambda^\alpha (1-\lambda)^\beta f(\lambda), \quad \alpha > 0, \quad \beta > -1, \tag{24}$$

where we assume $f(\lambda)$ to have a continuous second derivative over $0 \leqq \lambda \leqq 1$ and to be bounded:

$$0 < c_1 \leqq f(\lambda) \leqq c_2, \quad 0 \leqq \lambda \leqq 1. \tag{25}$$

It can be shown that then the relaxation converges; that is

$$R_n(\lambda) \to 0 \text{ for } n \to \infty \text{ and } 0 < \lambda < 1. \tag{26}$$

The simplest possibility is $f(\lambda) \equiv 1$ yielding the *hypergeometric* or *Jacobian polynomials.*

$$R_n(\lambda) = F(-n, n+\alpha+\beta+1, \alpha+1; \lambda), \tag{27}$$

where F is the hypergeometric function of Gauss. α and β must still be chosen in consideration of the problem at hand. Some while ago *Lanczos* [7]) proposed the special case $\alpha = \frac{1}{2}, \beta = -\frac{1}{2}$. The behavior of the hypergeometric relaxation for large n is described asymptotically by the formulas;

a) for small λ we have

$$R_n(\lambda) \sim \Lambda_\alpha(2n\sqrt{\lambda}) \tag{28}$$

where Λ_α is *Lommel's* function.

$$\Lambda_\alpha(x) = \frac{2^\alpha \alpha!}{x^\alpha} J_\alpha(x) \tag{29}$$

tabulated in *Jahnke-Emde.*

b) In the interior of the interval, $R_n(\lambda)$ oscillates with an amplitude given by

$$\frac{\alpha!}{\sqrt{\pi}} (n\sqrt{\lambda})^{-(\alpha+\frac{1}{2})} (1-\lambda)^{-\frac{\beta}{2}-\frac{1}{4}} \tag{30}$$

β may be called the *parameter of over-relaxation* for the following reason. Suppose β to be large, than the density

$$\varrho(\lambda) = \lambda^\alpha (1-\lambda)^\beta$$

diminishes rapidly as λ approaches 1. This means that the components of the residuals corresponding to the small eigen-values are more heavily weighted than the higher ones and therefore will be more rapidly eliminated. This is highly desirable in the application of relaxation techniques to partial differential equations.

For example, assume that the given system (1) of equations has been roughly solved using the sum of μ terms of the *Neumann-series*

$$A^{-1} = \sum_{(\nu)} (1 - A)^{\nu}. \tag{31}$$

The residual given by this rough solution is

$$r_{\mu} = (1 - A)^{\mu} k \tag{32}$$

corresponding to the residual polynomial

$$R_{\mu}(\lambda) = (1 - \lambda)^{\mu}. \tag{33}$$

In order to improve this rough solution we continue with hypergeometric relaxation using $\beta = 2\mu - \frac{1}{2}$. From (30) it follows that during this procedure (33) will be multiplied asymptotically by

$$\frac{\alpha!}{\sqrt{\pi}} (n\sqrt{\lambda})^{-(\alpha+\frac{1}{2})} (1 - \lambda)^{-\mu}.$$

Hence the final amplitude of residuals is

$$\frac{\alpha!}{\sqrt{\pi}} (n\sqrt{\lambda})^{-(\alpha+\frac{1}{2})}.$$

The tendency of Neumann's series to neglect the lower eigen-values is thus removed.

Another method to improve the convergence of the Neumann's series has been proposed by *Rutishauser* [8]) using the transformation of a power series into a continued fraction by his so-called QD-algorithm.

We have had satisfactory results with hypergeometric relaxation in Zürich.

2) If a lower bound $\varepsilon > 0$ for the eigen-values is also known, then $\varrho(\lambda)$ should vanish in the interval $0 \leqq \lambda \leqq \varepsilon$, since there are no residuals there to be liquidated. This leads to the method of *Shortley* and *Flanders* [9]), who use Tschebyscheff-polynomials in the remaining interval $\varepsilon \leqq \lambda \leqq 1$ and arrive at the best strategy in the sense of Tschebyscheff-approximation.

3) Obviously we are particularly interested in choosing the density-function (20) such that the ψ-measure is identically to the old discrete φ-measure. Since the eigen-values must be considered unknown, the explicit construction of the iteration-algorithm given in section V must be modified. Using (17) we find

$$\int_0^1 R_p(\lambda) R_q(\lambda) \varrho(\lambda) d\lambda = \sum_{(j)} k_j^2 \int_0^1 R_p(\lambda) R_q(\lambda) \delta(\lambda - \lambda_j) d\lambda$$

$$= \sum_{(j)} R_p(\lambda_j) R_q(\lambda_j) k_j^2 = \sum_{(j)} r_{pj} \cdot r_{qj} = (r_p, r_q).$$

The left side is zero for $p \neq q$ because of the orthogonality of our polynomials. Thus this relaxation process has the special property that the residual vectors form an orthogonal set. The coefficients a_i, c_i in (23) must therefore be com-

puted in such a way that the residual r_{i+1} defined by (22) is orthogonal to r_i and r_{i-1}.

After finite many steps, the residual vector must vanish since there can only be a finite number of orthogonal vectors. The iteration thus reaches the exact solution after a finite number of steps, as is the case with the elimination method of Gauss. As *Hestenes* [10]) proved recently, every finite iteration is equivalent to the process of *conjugate gradients* developped by him and the author. In conclusion we may thus state:

Among all scalar iteration algorithms the method of conjugate gradients as given in the original papers [11]) *is the best strategy in the sense of the φ-measure of error.*

In particular for a problem of the calculus of observations this method gives the smallest sum of the squared errors which can be achieved in a given number of iteration steps by a scalar process [12]).

References

[1] G. Forsythe: Solving linear algebraic equations can be intersting. Bull. of the Amer. Math. Soc. 59/4 (1953), pp. 299—329.

[2] A. Ostrowski: On over and under relaxation in the theory of the cyclic single step iteration. MTAC VII, Nr. 43, (1953), pp. 153—158.

[3] D. Young: Iterative methods for solving partial difference equations of elliptic type. Trans. Amer. Math. Soc. 76/1 (1954), pp. 92—111.

[4] S. P. Frankel: Convergence rates of iterative treatments of partial differential equations. MTAC 4/30 (1950).

[5] U. Hochstrasser: Thesis, Federal Institut of Technology, Zürich, 1954.

[6] D. Young: On Richardson's method for solving linear systems with positive definite matrices. J. of Math. and Physics (1954), pp. 244—255.

[7] C. Lanczos: Chebyshev polynomials in the solution of large-scale linear systems. Proc. of the ass. for computing machinery. Toronto meeting 1952.

[8] H. Rutishauser: Der Quotienten-Differenzen-Algorithmus. Zeitschr. für angew. Math. und Physik V/3 (1954), pp. 233—251.

[9] G. Shortley: Use of Tschebyscheff-Polynomial Operators in the numerical solution of boundary-value problems. J. of applied physics 24, Nr. 4 (1953), pp. 392—396.

[10] M. Hestenes: The conjugate gradient method for solving linear systems. Report, Institute for numerical Analysis, Los Angeles, 1954.

[11] E. Stiefel: Über einige Methoden der Relaxationsrechnung. Zeitschr. für angew. Math. und Physik III (1952), pp. 1—33.

[12] M. Hestenes und E. Stiefel: Method of conjugate gradients for solving linear systems. J. of Research of the National Bureau of Standards, vol. 49 (1952), pp. 409—436.

[13] E. Stiefel: Ausgleichung ohne Aufstellung der Gauss'schen Normalgleichungen. Wissenschaftl. Zeitschr. der TH Dresden 2 (1952/53).