CS45

RELAXATION METHODS FOR SEMI-DEFINITE SYSTEMS

BY

W. KAHAN

TECHNICAL REPORT NO. CS45
AUGUST 9, 1966

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

RELAXATION METHODS FOR SEMI-DEFINITE SYSTEMS

By

W. KAHAN[*]

ABSTRACT

Certain non-stationary relaxation iterations, which are commonly applied to positive definite symmetric systems of linear equations, are also applicable to a semi-definite system provided that system is consistent. Some of the convergence theory of the former application is herein extended to the latter application. The effects of rounding errors and of inconsistency are discussed too, but with few helpful conclusions. Finally, the application of these relaxation iterations to an indefinite system is shown here to be ill-advised because these iterations will almost certainly diverge exponentially.

[*] Stanford University Computer Science Department and University of Toronto Departments of Mathematics and Computer Science.
Prepared under Contract Nonr-225(37) (NR-044-211) Office of Naval Research.
Reproduction in whole or in part is permitted for any purpose of the United States Government.

Relaxation Methods for Semi-Definite Systems

Relaxation methods like the Gauss-Seidel iteration are widely used to solve linear systems of the form

$$A\underline{y} = \underline{c}$$

when A is a Hermitian positive definite matrix, but their usefulness when A is semi-definite is less well appreciated. A recent paper by H. Keller (1965) has expanded earlier results of G. Forsythe (1960) and the author (1958, Ch. 2) concerning the convergence of stationary iterations when the system is consistent, though singular. The gist of Keller's paper is that those iterations of the form

$$\underline{y}_{n+1} = \underline{y}_n + T(\underline{c}-A\underline{y}_n)$$

which are usually used when A is definite also work when A is semi-definite. This note is concerned with a non-stationary iteration

$$\underline{y}_{n+1} = \underline{y}_n + T_n(\underline{c}-A\underline{y}_n)$$

and can be regarded as a supplement to Keller's work. In particular, the results here imply that some of his stationary iterations are numerically stable; but there are other applications too, like eigenvalue problems, for this note.

The hypotheses used in this work are intended to be as weak as will fit methods likely to be used in practice. Consequently, the results here do not completely generalize the work of A. Ostrowski (1954) or S. Schechter (1959). There is also some overlap with recent independent work of Ostrowski (1965).

The report is divided into four numbered sections. Section 1 is remeniscent of works by Kaczmarz (1937) and Agmon (1954) in that it characterizes the relaxation iteration to be considered here as a sequence of partial projections in a suitable space. The iteration is shown to converge at least as fast as some geometric series.

Section 2 defines that "suitable space" more precisely in terms of the given matrix A, and shows how convergence can be sustained in the face of certain rounding errors committed during the iteration.

Section 3 mentions two applications and discusses some open problems. The central problem is that the solution $\underline{y}$ of a singular but consistent system $A\underline{y} = \underline{c}$ is not a continuous function of A or $\underline{c}$. Therefore, it is no surprise that rounding errors in A and $\underline{c}$ may obscure the criteria by which one judges whether or not an iteration has come close enough to a desired solution that further iteration is worthless. Attempts have been made to iterate instead with a non-singular system that is practically equivalent to the given singular system $A\underline{y} = \underline{c}$; two such attempts are mentioned in this section.

Section 4 shows that relaxation diverges to infinity when the system $A\underline{y} = \underline{c}$ is inconsistent, and is almost certain to diverge exponentially if A is indefinite.

1.) Relaxation in a Simple Case

To begin, consider the solution by relaxation of the trivial equation

$$\underline{x} = \underline{b}$$

as follows:

Let $\{\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_M\}$ be a given set of non zero vectors which, though not necessarily linearly independent, do span $\underline{x}$'s space. In other words, M is no smaller than the dimension of $\underline{x}$ and both matrices

$$\{\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_M\} \text{ and } \{\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_M, \underline{x}\}$$

have the same rank for all vectors $\underline{x}$.

Next let $\underline{p}_1, \underline{p}_2, \ldots, \underline{p}_n, \ldots$ be a sequence constructed by choosing $\underline{p}_n = \underline{e}_j$ for some $j = J(n)$ so contrived that each set of L consecutive vectors $\{\underline{p}_n, \underline{p}_{n+1}, \ldots, \underline{p}_{n+L-1}\}$ spans $\underline{x}$'s space. L is some fixed integer exceeding the dimension of $\underline{x}$.

Now the iteration to solve $\underline{x} = \underline{b}$ for $\underline{x}$ can be defined. Beginning with an arbitrary $\underline{x}_1$, we define for $n = 1,2,3,\ldots$

$\eta_n = \underline{p}_n^*(\underline{b}-\underline{x}_n)/\underline{p}_n^*\underline{p}_n$,

δ_n is arbitrary except that $|\delta_n| \le d < 1$ for all n ,

$\xi_n = (1+\delta_n)\eta_n$,

$\Delta\underline{x}_n = \xi_n \underline{p}_n$, and

$\underline{x}_{n+1} = \underline{x}_n + \Delta\underline{x}_n$.

The numbers $(1+\delta_n)$ are frequently called ω_n in the literature. The values δ_n can be complex, but when they are real they have the following connotations:

$\delta_n = 0$ means exact relaxation ,

$\delta_n > 0$ means over-relaxation ,

$\delta_n < 0$ means under relaxation .

Each relaxation may be regarded as a near-projection;

$$\underline{b} - \underline{x}_{n+1} = \{I - (1+\delta_n)\, \underline{p}_n\, \underline{p}_n{}^*/\underline{p}_n{}^*\, \underline{p}_n\}\, (\underline{b}-\underline{x}_n) \quad .$$

Consequently, if the usual norm

$$\|\underline{x}\| \equiv \sqrt{\underline{x}^*\underline{x}}$$

is used, then $\|\underline{b}-\underline{x}_{n+1}\| \leq \|\underline{b}-\underline{x}_n\|$. To be more precise,

$$\|\underline{b}-\underline{x}_{n+1}\|^2 = \|\underline{b}-\underline{x}_n\|^2 - (1-|\delta_n|^2)\ |\underline{p}_n{}^*(\underline{b}-\underline{x}_n)|^2/\|\underline{p}_n\|^2$$

$$< \|\underline{b}-\underline{x}_n\|^2 \quad \text{unless} \quad \underline{p}_n{}^*(\underline{b}-\underline{x}_n) = 0$$

because $1 - |\delta_n|^2 \geq 1 - d^2 > 0$.

If we use the abbreviation

$$P(\underline{p},\delta) = I - (1+\delta)\ \underline{p}\ \underline{p}^*/\underline{p}^*\underline{p} \quad ,$$

then it becomes convenient to write

$$\underline{b} - \underline{x}_{n+L} = T_n(\underline{b}-\underline{x}_n) \quad \text{where}$$

$$T_n = \prod P(\underline{p}_m\ ,\ \delta_m) \quad \text{over} \quad n \leq m < n + L \quad .$$

The next step is to show that $\|T_n\| < 1$, where the matrix norm is defined by

$$\|T_n\| \equiv \max\ \|T_n\underline{v}\|/\|\underline{v}\| \quad \text{over} \quad \underline{v} \neq \underline{0} \quad .$$

Let us write $\lambda_n = \|T_n\|$. Obviously $\lambda_n \leq 1$. If $\lambda_n = 1$ then there must be some non-zero vector $\underline{v}$ such that $\|T_n\underline{v}\| = \|\underline{v}\|$. By examining the factors $P(\underline{p}_m , \delta_m)$ of T_n in turn, we conclude that

$$T_n \underline{v} = \underline{v} \text{ and}$$

$$\underline{p}_m{*}\underline{v} = 0 \text{ for } m = n , n{+}1 ,\ldots, n + L - 1 .$$

Because of the way the vectors $\underline{p}_m$ were chosen, there exists a complete linearly independent subset among them, and since $\underline{p}_m{*}\underline{v} = 0$ for all $\underline{p}_m$ in that subset, $\underline{v} = 0$. This contradiction shows that $\lambda_n < 1$.

Now, λ_n may be identified with a continuous function

$$\lambda_n = \lambda(\underline{p}_n , \underline{p}_{n+1} ,\ldots, \underline{p}_{n+L-1} ; \delta_n , \delta_{n+1} ,\ldots, \delta_{n+L-1})$$

of L vector and L scalar arguments. The vector arguments are constrained in such a way that there is only a finite number of permissible sets of L vector arguments. Indeed, of the L^M ways to choose the L vectors

$$\underline{p}_n , \underline{p}_{n+1} ,\ldots, \underline{p}_{n+L-1}$$

from the set $\{\underline{e}_1 , \underline{e}_2 ,\ldots, \underline{e}_M\}$, most will be rejected because the $\underline{p}$-vector must include a complete set spanning $\underline{x}$-space. The scalar arguments δ_m are constrained to the compact set $|\delta_m| \leq d < 1$. Therefore there must exist a number

$$\lambda = \max \lambda(\underline{p}_{(1)}, \underline{p}_{(2)}, \ldots, \underline{p}_{(L)} ; \delta_{(1)}, \delta_{(2)}, \ldots, \delta_{(L)})$$

over the set of allowed choices for $\underline{p}_{(m)}$ and $\delta_{(m)}$,

and this maximum is achieved, and $\lambda < 1$ by virtue of the same argument as was used to show $\lambda_n < 1$.

Therefore, the relaxation iteration converges at least as quickly as a geometric series with a common ratio $\lambda^{1/L}$;

$$\|\underline{b}-\underline{x}_{n+kL}\| \leq \lambda^k \|\underline{b}-\underline{x}_n\| \quad \text{for} \quad k = 1, 2, 3, \ldots .$$

2.) Relaxation in Practice

The foregoing theory is applicable to the solution by relaxation of the equation

$$A\underline{y} = \underline{c}$$

when A is Hermitian and positive definite or semi-definite, provided the equation is consistent when A is semi-definite. Only the semi-definite case is discussed here.

The relaxation process for solving $A\underline{y} = \underline{c}$ consists in choosing a set of spanning vectors $\underline{f}_j$ and a sequence of vectors $\underline{q}_n$ of which each consecutive L vectors include a spanning subset of the $\underline{f}$'s , and a sequence of values δ_n with $|\delta_n|^2 \leq d < 1$. Then

$$\eta_n = \underline{q}_n{}^*(\underline{c}-A\underline{y}_n)/\underline{q}_n{}^*A\underline{q}_n \quad ,$$

$$\xi_n = (1+\delta_n)\eta_n \quad ,$$

$$\Delta\underline{y}_n = \xi_n \underline{q}_n \quad \text{and}$$

$$\underline{y}_{n+1} = \underline{y}_n + \Delta\underline{y}_n \ , \quad \text{for} \quad n = 1 \ , 2 \ , 3 \ ,\ldots \quad .$$

There is a formal requirement that $\underline{q}_n{}^*A\underline{q}_n \neq 0$, which is tantamount to requiring $\underline{f}_j{}^*A\underline{f}_j > 0$ for all j . Alternatively, since $\underline{f}^*A\underline{f} = 0$ implies $A\underline{f} = \underline{0}$ and, because $\underline{c} = A\underline{u}$ for some $\underline{u}$, $\underline{f}^*\underline{c} = 0$, it suffices to define $\eta_n = 0$ instead of using the indeterminate $0/0$ when $\underline{q}_n{}^*A\underline{q}_n = 0$. In practice one is unlikely to have to worry about this contingency.

The relaxation process for $A\underline{y} = \underline{c}$ can be related to that for $\underline{x} = \underline{b}$ via a non-singular linear transformation U which satisfies

$$U^*AU = \text{diag}(1,1,1\ ,\ldots,\ 1,0,0\ ,\ldots,\ 0) \quad .$$

The number of 1's and zeros depends only upon A ; otherwise there is a substantial degree of freedom in the choice of U . One possibility is to choose U in such a way that U*U is diagonal too. That such a matrix U exists follows immediately from the fact that A is unitarily similar to a non-negative diagonal matrix.

Let us impose a partitioning upon

$$U^*AU = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$$

and, conformally, upon

$$\underline{y} = U\begin{pmatrix} \underline{x} \\ \hat{\underline{x}} \end{pmatrix} \ , \qquad U^*\underline{c} = \begin{pmatrix} \underline{b} \\ \hat{\underline{b}} \end{pmatrix} \ ,$$

$$\underline{f}_j = U\begin{pmatrix} \underline{e}_j \\ \hat{\underline{e}}_j \end{pmatrix}, \quad \text{and} \quad \underline{q}_j = U\begin{pmatrix} \underline{p}_j \\ \hat{\underline{p}}_j \end{pmatrix} \ .$$

Note that the vectors $\{\underline{e}_1 , \underline{e}_2 ,\ldots, \underline{e}_M\}$ span the space of $\underline{x}$ because the vectors $\{\underline{f}_1 ,\ldots, \underline{f}_M\}$ span the space of $\underline{y}$. Now, $A\underline{y} = \underline{c}$ is precisely equivalent to

$$\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \underline{x} \\ \hat{\underline{x}} \end{pmatrix} = \begin{pmatrix} \underline{b} \\ \hat{\underline{b}} \end{pmatrix} , \quad \text{or}$$

$$\underline{x} = \underline{b} \quad \text{and} \quad \underline{0} = \hat{\underline{b}} \ .$$

The last condition is necessarily satisfied because the equation $A\underline{y} = \underline{c}$ is consistent by hypothesis. Furthermore,

$$\underline{q}_n{}^*A\underline{q}_n{}^* = \underline{p}_n{}^* \underline{p}_n > 0 \quad \text{and}$$

$$\eta_n = \underline{q}_n{}^*(\underline{c}-A\underline{y}_n)/\underline{q}_n{}^*A\underline{q}_n$$

$$= \underline{p}_n{}^*(\underline{b}-\underline{x}_n)/\underline{p}_n{}^* \underline{p}_n \ .$$

Therefore the theory developed for $\underline{x} = \underline{b}$ is almost applicable to $A\underline{y} = c$, the defect being that while $\underline{x}_n \to \underline{b}$ we do not yet know what happens to $\hat{\underline{x}}_n$.

Now,

$$\hat{\underline{x}}_n = \hat{\underline{x}}_1 + \sum_1^{n-1} \xi_m \hat{\underline{p}}_m ,$$

so we can deduce that $\hat{\underline{x}}_n$ converges if we can prove that $\sum_1^\infty |\xi_n|$ converges. Because

$$|\xi_n| \le (1+d)\ \|\underline{p}_n\|\ \|\underline{b}-\underline{x}_n\| \quad , \text{ and}$$

$$\|\underline{b}-\underline{x}_{n+kL}\| \le \lambda^k\ \|\underline{b}-\underline{x}_n\| \quad \text{for all } n \text{ and } k = 1,2,3,\ldots \ ,$$

and λ is some number which depends upon $\{\underline{f}_j\}$, d , L and U (actually upon A instead of U) , and $\lambda < 1$, the series

$$\hat{\underline{x}}_1 + \sum_1^\infty \xi_n \hat{\underline{p}}_n$$

converges at least as quickly as a geometric series with common ratio $\lambda^{1/L} < 1$. Therefore the sequence $\underline{y}_n$ converges too to one of the finite solutions of $A\underline{y} = \underline{c}$.

The foregoing theory is easily generalized to include block relaxation as well as point relaxation, but the theory is already too general to permit anything useful to be said about the rate of convergence of the iteration, nor upon how to choose $\{\underline{f}_j\}$ and $\{\delta_n\}$ to minimize λ . However, convergence may be retarded if δ_n is chosen in a way which keeps $|\delta_n|$ too large. More precisely, if $\beta < |\delta_n| \le d$ for all n , then $\lambda^{1/L} > \beta$. This is so because

$$\|\underline{b}-\underline{x}_{n+1}\|^2 = \|\underline{b}-\underline{x}_n\|^2 - (1-|\delta_n|^2)|\underline{p}_n{}^*(\underline{b}-\underline{x}_n)|^2/\|\underline{p}_n\|^2$$

$$\ge |\delta_n|^2\ \|\underline{b}-\underline{x}_n\|^2 \quad \text{by the Schwartz inequality} \ .$$

And if the values δ_n and vectors $\underline{q}_n$ are chosen cyclically ($\underline{q}_{n+L} = \underline{q}_n$ and $\delta_{n+L} = \delta_n$ for all n) then

$$\lambda \ge \prod_{m=1}^{L} |\delta_m|^{1/R} \quad \text{where} \quad R = \text{rank } (A) \ .$$

The last inequality can be proved by generalizing a theorem due to the author (1958); the theorem and proof can also be found in Varga's book (1962). The proof of the generalization is a tedious computation too long to include in this report.

Finally, one aspect of numerical stability is considered here. Let a semi-norm for the error $(\underline{y}_n-\underline{y})$ be defined by

$$\|\underline{y}_n-\underline{y}\|_A \equiv [(\underline{y}_n-\underline{y})^*A(\underline{y}_n-\underline{y})]^{1/2}$$

$$= \|\underline{x}_n-\underline{b}\| \ .$$

Since $\|\underline{y}_{n+1}-\underline{y}\|_A \leq \lambda^{1/L} \|\underline{y}_n-\underline{y}\|_A$ and $\lambda^{1/L} < 1$,

$\underline{y}_{n+1}$ can be replaced by a perturbed vector $\underline{y}'_{n+1}$ which still satisfies

$$\|\underline{y}'_{n+1}-\underline{y}\|_A < K\|\underline{y}_n-\underline{y}\|_A \quad \text{for some} \quad K < 1$$

provided the perturbation $(\underline{y}'_{n+1}-\underline{y}_{n+1})$ is small enough to satisfy

$$\|\underline{y}'_{n+1}-\underline{y}_{n+1}\|_A < (K-\lambda^{1/L})\ \|\underline{y}_n-\underline{y}\|_A \ .$$

Therefore, the relaxation iteration will appear to converge as long as roundoff is kept sufficiently small. But the foregoing argument is too superficial to be of much use in practice because one does not normally know $\|\underline{y}_n-\underline{y}\|_A$ nor λ , and therefore cannot tell when rounding errors are small enough to be unimportant. When an iteration converges slowly (because λ is very nearly 1), it can be difficult to supply criteria whereby a computer program will be stopped <u>after</u> the iteration has achieved as accurate an approximate solution $\underline{y}_n$ as is desired or possible, but before a large quantity of time has been wasted on iterations whose effect has been nullified by roundoff. For further discussion of these difficulties, consult papers by Golub (1962) and Descloux (1963).

Fortunately, the relaxation iteration need not suffer intolerably from these difficulties, because the iteration can be carried on usefully until $\Delta\underline{y}_n$ remains smaller in magnitude than two or three units in the last place of $\underline{y}_n$, after which there is no point in continuing. This is so for the following reasons, which are adapted from Ch. 4 of the author's thesis (1958).

Let each vector $\underline{f}_j$ be one of the coordinate vectors

$$f_j = (0,0 ,\dots, 0,1,0 ,\dots, 0)^T$$

with a 1 in the j^{th} position. Since $\underline{q}_n$ is chosen from the set of $\underline{f}$'s , it too is a coordinate vector. Therefore

$$\eta_n = \underline{q}_n^*(\underline{c}-A\underline{y}_n)/\underline{q}_n^*A\underline{q}_n$$

can be computed to almost full single precision; to do so one must compute the relevant component of the residual $\underline{c} - A\underline{y}_n$ with the aid of double-precise accumulation of products of single precision numbers before the residual is rounded to single precision. Next, the number $\xi_n = (1+\delta_n)\eta_n$ can be computed tentatively to provide a value for the formula

$$\underline{y}_{n+1} = \underline{y}_n + \xi_n \underline{q}_n \ .$$

However, the vector $\underline{y}_{n+1}$ will be rounded before it is stored. Therefore, the value of ξ_n actually used will be defined in fact by the equation

$$\xi_n \underline{q}_n \equiv \underline{y}_{n+1} - \underline{y}_n$$

in which only one component can be non-zero, and that component is the difference between the new value and the previous value stored in the array $\underline{y}_n$. In other words, when we let the symbol $\underline{y}_n$ stand for a vector which is precisely represented by an array of numbers stored in the computer, then all of the foregoing theory remains applicable provided we understand that ξ_n is finally defined after $\underline{y}_{n+1}$ is rounded and stored. Therefore ξ_n may differ from the tentative value $(1+\delta_n)\eta_n$ which had been intended for it. Even so, convergence is assured if the final value of ξ_n satisfies

$$|\xi_n/\eta_n - 1| \leq d < 1 \quad \text{for all } n \ .$$

This last condition can be satisfied easily unless η_n is not much larger than a unit in the last place of the affected component of $\underline{y}_n$. Therefore, rounding errors may slow the iteration down, but they need not prevent the iteration from progressing to a point where the scaled residual

$$(\mathrm{diag}(A))^{-1} (\underline{c} - A\underline{y}_n)$$

is scarcely larger than a unit in the last place of $\underline{y}_n$. This is as small a residual as might reasonably be hoped for, but whether it is worth waiting for is a harder question. In my opinion, a good relaxation program can confidently be expected to reduce the scaled residual to about ten units in the last place in $\underline{y}_n$, beyond which point further progress is likely to be too slow to be economical.

3.) Open Problems Connected with Applications

There are two important applications of the foregoing theory. One is to the solution of

$$(A-\lambda B)\,\underline{u} = 0$$

for an eigenvector $\underline{u}$ corresponding to the smallest eigenvalue λ of A with respect to B when both A and B are Hermitian and B is positive definite. Since this application usually entails the simultaneous calculation of λ as well as $\underline{u}$, the details are deferred to a later report (Kahan, 1966).

The second application is to the solution of the Neumann problem in potential theory. Here the semi-definite matrix A represents a discrete approximation to a partial differential operator, and $\underline{c}$ in the equation

$$A\underline{y} = \underline{c}$$

depends upon boundary values assigned to a normal derivative. The boundary values must satisfy a compatibility condition to permit a solution $\underline{y}$ to exist. Unfortunately, roundoff in $\underline{c}$ may prevent the compatibility condition from being precisely satisfied. What happens to the iteration in this case?

This question was considered in the author's earlier work (1958) only for the stationary case of constant $\delta_n = \delta$ and a cyclic choice of $\underline{q}_n = \underline{q}_{n+L}$ for all n. There it was shown that the sequence of residuals

$$\underline{c} - A\underline{y}_n$$

converged like a geometric series even though the sequence $\underline{y}_n$ diverged like an arithmetic progression. The implication was that if $\underline{c}$ deviated only slightly from consistency, then the sequence $\underline{y}_n$ would diverge fairly slowly and, for n large enough, would adequately approximate the solution of a nearby consistent system. Besides, if the general solution of

$$A\underline{n} = \underline{0}$$

were known then some iterates $\underline{y}_m$ could be replaced by $\underline{y}_m - k\underline{n}$ with k chosen to diminish $\|\underline{y}_m - k\underline{n}\|$ conveniently. In particular, if A came from the Neumann problem then $\underline{n}$ would represent a function everywhere constant.

But the situation is not so clear for the non stationary relaxation process. The best I can do is prove that for all large enough values of n the residuals $\underline{c} - A\underline{y}_n$ will be bounded by some expression of the form

$$\|\underline{c} - A\underline{y}_n\| \leq K \, \|\underline{r}\|$$

where $\|\underline{r}\|$ is the minimum possible value of $\|\underline{c}-A\underline{y}\|$ for all $\underline{y}$ and K depends upon the same data as determines λ, i.e. upon d, L, A and the set $\underline{f}_j$. (Unfortunately, $\|\underline{r}\|$, λ and K are discontinuous functions of A .) This is enough to establish numerical stability in the face of errors in $\underline{c}$ and in $\Delta\underline{y}_n$, but not enough to tell a computer program when to stop iterating. The problem is acute when "convergence" is slow, because the effect of inconsistency in $\underline{c}$ is scarcely distinguishable from the effects of a value λ very near 1 or a rounding error in A .

One way to sidestep the problem of a slightly inconsistent right-hand side $\underline{c}$ is to use a restricted relaxation iteration; a selected component of $\underline{y}_n$ is forced to be constant for all n and then relaxation is restricted to the other components. For example, in the Neumann problem the value of the desired solution at one point in the region of interest could be fixed arbitrarily, and the values of the solution elsewhere could be obtained by solving the relevant difference equations by relaxation. Such a procedure is described unenthusiastically by Forsythe and Wasow (1960). The scheme is open to two criticisms:

First, the effect of an inconsistent right-hand side $\underline{c}$ is concentrated in the one equation of the system whose residual is never relaxed. For the Neumann problem this can mean a cusp-like intrusion in the solution at the artificially fixed point.

Second, the rate of convergence of restricted relaxation can compare unfavourably with that of unrestricted relaxation. This possibility is clear in those cases, as when A has Young's "property A" , when the rate of convergence of successive overrelaxation can be computed directly in terms of the smallest non-zero eigenvalue of A ; these cases must have been the ones that Forsythe and Wasow had in mind when they advised against the restricted relaxation scheme. But in general there is no way to estimate the rate of convergence of successive overrelaxation in terms only of the non-zero eigenvalues of A, and there are rare cases in which a restricted iteration is faster than the corresponding unrestricted iteration. The following example is one for which, if δ is fixed at its best constant value for each iteration separately, the restricted successive overrelaxation converges almost twice as quickly as the unrestricted overrelaxation.

Let $A = \{a_{ij}\}$ be the symmetric semi-definite $N \times N$ circulant matrix defined by

$$a_{ij} = 0 \quad \text{except that}$$

$$a_{ii} = 1 \quad \text{and}$$

$$a_{ij} = -1/2 \quad \text{whenever} \quad i \equiv j \pm 1 \mod N \ .$$

The equation $A\underline{y} = \underline{b}$ is consistent whenever

$$\sum_1^N b_m = 0 \quad ;$$

and successive overrelaxation with constant δ and arbitrary $\underline{y}^1$ produces a sequence of iterates

$$\underline{y}^1 , \underline{y}^2 , \underline{y}^3 ,\ldots, \underline{y}^n ,\ldots$$

via the recurrence

$$y_m^{n+1} = y_m^n + (1+\delta)(b_m + \frac{1}{2} y_{m-1}^{n+1} - y_m^n + \frac{1}{2} y_{m+1}^n)$$

in which it is to be understood that $y_{m+N} \equiv y_m$ for all m. The sequence of error-vectors $\underline{u}^n$ satisfy the same recurrence except that b_m is replaced by zero. The iteration is stationary; its eigenvalues ϵ are the N complex numbers for which there exist complex error-vectors $\underline{u}^n$ satisfying

$$\underline{u}^{n+1} = \epsilon\, \underline{u}^n \quad \text{and} \quad \underline{u}^n \neq \underline{0} \ .$$

It can be shown that these eigenvalues satisfy

$$\epsilon = z^N \quad \text{and} \quad 2z^N - (1+\delta)(z^{N-1}+z) + 2\delta = 0 \ ;$$

this is done in Ex. 5 of the author's thesis (1958), and by further tedious work along the lines discussed there and in Ex. 3 it is possible to approximate each of the N eigenvalues ϵ as functions of δ with sufficient accuracy to support the author's claims. But the same conclusions can be drawn more elegantly from an argument patterned upon Garabedian's (1956):

When N is large the equation $A\underline{y} = \underline{b}$ can be approximated by the differential equation

$$\frac{d^2y}{dx^2} = b$$

with periodic boundary conditions

$$y(x+\pi) \equiv y(x) \quad \text{for all} \quad x \ .$$

Then the error vectors u_m^n can be approximated by the function $u(mh\ ,\ nk)$, where $u(x\ ,\ t)$ satisfies the hyperbolic partial differential equation

$$\tau\, u_t = \rho\, u_{xx} - u_{xt}$$

with constants

$$\rho = h/k \quad \text{and} \quad \tau = ((1-\delta)/(1+\delta))/h \quad .$$

Here $h = \pi/N$ can be made arbitrarily small by making N large enough, and the error of approximation tends to zero as $h \to 0$; note that $k \to 0$ and $\delta \to 1$ as $N \to \infty$.

We deviate slightly from Garabedian by neglecting to transform the partial differential equation into canonical form before separating the variables. Instead, a trial solution

$$u(x\ ,\ t) \equiv X(x)\ T(t)$$

yields

$$\tau(T'/T) = \rho(X''/X) - (X'/X)(T'/T)$$

for which the solutions are $T = \exp(\lambda t)$ and $X = \exp(\mu x)$ provided

$$\tau\, \lambda = \rho\, \mu^2 - \mu\, \lambda\ .$$

Now we write the general solution $u(x\ ,\ t)$ symbolically in the form

$$u(x\ ,\ t) = \sum_\mu \alpha(\mu)\ \exp(\mu x + \lambda(\mu) t)$$

summed over all permissible complex numbers μ , with

$$\lambda(\mu) \equiv \rho\mu^2/(\tau+\mu)$$

and $\alpha(\mu)$ chosen to satisfy the periodic boundary conditions

$$u(x+\pi\ ,\ t) \equiv u(x\ ,\ t) \quad \text{for all} \quad x \quad \text{and} \quad t \ .$$

To any value λ corresponds at most two values of μ for which $\lambda = \lambda(\mu)$, and if we call these values μ' and μ'' then

$$f_\lambda(x) \equiv \alpha(\mu') \exp(\mu' x) + \alpha(\mu'') \exp(\mu'' x)$$

must be periodic too. This leads promptly to the conclusion that the only permissible values of μ are

$$\mu_n = 2ni$$

where n is an integer (positive, negative or zero) and $i^2 = -1$. The corresponding values of λ are

$$\lambda_n = -4n^2 \rho(\tau-2ni)/(\tau^2+4n^2) \ .$$

Our object now is to choose τ , and hence δ , in such a way that

$$\max_{n\neq 0} |\exp(\lambda_n)|$$

is minimized, thus ensuring that even if $u(x\ ,\ 0)$ (corresponding to u_m^1) is chosen in the worst possible way, the convergence of u to its limiting form

$$u(x\ ,\ \infty) = \text{constant for all} \quad x$$

will be as fast as possible. Since

$$\max_{n \neq 0} |\exp(\lambda_n)| = |\exp(\lambda_1)|$$

it is soon concluded that the best value for τ is 2 ; then

$$|\exp(\lambda_1)| = \exp(-\rho) \quad .$$

The restricted successive overrelaxation differs from the foregoing only in that the boundary conditions for $u(x , t)$ are

$$u(0 , t) \equiv u(\pi , t) \equiv 0 \quad \text{for all} \quad t \quad ,$$

whence the values λ_n are

$$\lambda_n = -2\rho(\tau+ni \sqrt{n^2-\tau^2}) \quad \text{for} \quad \pm n = 1,2,3 ,\ldots \quad ;$$

now the best value for τ is $\tau = 1$, and

$$|\exp(\lambda_1)| = \exp(-2\rho) \quad .$$

The conclusion is that the restricted iteration is twice as fast as the unrestricted iteration.

The example should not be taken too seriously; it is a counterexample to a plausible conjecture, and atypical of most cases encountered in practice.

Finally, a trick due to Riley (1955) deserves some attention. The idea here is to approximate the semi-definite system $A\underline{y} = \underline{c}$ by a definite system

$$(A+\Delta A)\ \underline{z} = \underline{c}$$

in which ΔA is a suitably chosen positive definite matrix. To be most useful, ΔA should be just large enough to swamp the uncertainties in A and $\underline{c}$, but not too large lest $\underline{z}$ be useless. An attractive choice for ΔA is an $N \times N$ diagonal matrix each of whose diagonal elements is of the order of N or N^2 units in the last significant decimal place of the corresponding element of A . This choice may be useful when ΔA is negligible compared with the smallest positive eigenvalue of the semi-definite matrix A . Otherwise one may be forced to embed the relaxation iteration within an outer iteration process defined by

$$(A+\Delta A)\, \Delta \underline{z}_n = \underline{c} - A\underline{z}_n$$

and designed to replace $\underline{z}_n$ by a vector $\underline{z}_{n+1} = (\underline{z}_n + \Delta\underline{z}_n)$ which better approximates $\underline{y}$. The relaxation process would be used to compute each $\Delta\underline{z}_n$. Unfortunately, these wheels within wheels can be troublesome, especially when they all turn very slowly, because it is so hard to tell a computer which wheel should be turned and when to stop. I am not convinced that Riley's trick is worth while when relaxation methods must be used to calculate $\Delta\underline{z}_n$, although it has proved valuable when used with direct methods like Gaussian elimination where it is much easier to deal with a matrix $(A+\Delta A)$ that stays positive definite despite perturbation by roundoff than to deal with a matrix A which may be made indefinite by perturbation. (cf. sections 2 and 6 of the paper by Martin, Peters and Wilkinson (1965).)

4.) Some Negative Results

It is widely known that the relaxation iteration described in section 2 of this report may fail to converge when the system $A\underline{y} = \underline{c}$ is inconsistent or when A is indefinite. (cf. Keller (1965, theorem 2) or Ostrowski (1954, theorem II).) The results proved here are somewhat stronger. We find that if $A\underline{y} = \underline{c}$ has no solution $\underline{y}$ then the sequence of iterates $\underline{y}_n$ must diverge to infinity. We find that if A is indefinite (has both positive and negative eigenvalues) then the sequence $\underline{y}_n$ is almost certain to diverge to infinity like an exponential function of n. The proofs involve heavy computations, so only brief outlines are sketched in here.

The assumptions about $\underline{f}_j$, $\underline{q}_n$ and δ_n in section 2 are repeated here with one extra restriction; we assume $\underline{f}_j{}^*A\underline{f}_j > 0$ for all j. Moreover, to simplify the computations we shall assume that each $\underline{f}_j$ has been scaled so that $\underline{f}_j{}^*A\underline{f}_j = 1$.

The first step in the computation is the construction of matrices

$$D_n \equiv \mathrm{diag}(\delta_n, \delta_{n+1}, \ldots, \delta_{n+L-1}) \text{ and}$$

$$Q_n \equiv (\underline{q}_n, \underline{q}_{n+1}, \ldots, \underline{q}_{n+L-1}) .$$

The matrix Q_n may have more columns than rows. Since its columns contain a spanning subset, the equation $Q_n\underline{u} = \underline{y}$ can always be solved for $\underline{u}$, albeit not uniquely, whatever $\underline{y}$ may be. Also, $Q_n{}^*\underline{y} = \underline{0}$ implies $\underline{y} = \underline{0}$ for the same reason.

Since $\underline{q}_n{}^*A\underline{q}_n = 1$ for all n, it is possible to write

$$Q_n{*}AQ_n = I - R_n - R_n{*}$$

where R_n is an upper triangular matrix with zero diagonal. The relaxation iteration can now be described conveniently in a closed form; the reader is asked to verify that

$$\underline{y}_{n+L} = \underline{y}_n + Q_n[(I+D_n)^{-1} + R_n{*}]^{-1} Q_n{*}(\underline{c}-A\underline{y}_n) \quad .$$

(This relation may seem less mysterious after one observes that it is possible to solve the equations

$$Q_n\, \underline{u}_{n+m} = \underline{y}_{n+m} \qquad \text{for} \qquad m = 0,1,2\ ,\ldots,\ L$$

for vectors $\underline{u}_{n+m}$ corresponding to the L intermediate steps from $\underline{u}_n$ to $\underline{u}_{n+L}$ of one iteration of the extrapolated Gauss-Seidel Method for solving

$$(I-R_n-R_n{*})\ \underline{u} = Q_n{*}\ \underline{c} \quad .$$

Compare Kahan (1958), or Varga (1962) p. 59 where $\delta_n = \omega-1$ is held constant for all n .)

The second step in the computation is to define the quadratic functional

$$W(\underline{v}) = \underline{v}{*}A\underline{v} - \underline{v}{*}\underline{c} - \underline{c}{*}\underline{v}$$

The significance of W is clear when A is positive definite because then $\underline{c} = A\underline{y}$ and

$$W(\underline{v}) = (\underline{v}-\underline{y})^* A(\underline{v}-\underline{y}) - \underline{y}^* A\underline{y}$$

$$= \|\underline{v}-\underline{y}\|_A^2 - \|\underline{y}\|_A^2 \ .$$

But if A is indefinite, then $\|\cdots\|_A$ is not a norm; and if $A\underline{y} = \underline{c}$ has no solution then $(\underline{v}-\underline{y})$ cannot be computed. Even so, W is always computable. The reader is asked to verify the following connection between W and the sequence of iterates $\underline{y}_n$:

$$W(\underline{y}_{n+L}) - W(\underline{y}_n) = -(\underline{c}-A\underline{y}_n)^* B_n(\underline{c}-A\underline{y}_n) \quad \text{where}$$

$$B_n = Q_n[I + (I+D_n^*)R_n]^{-1} (I-D_n^* D_n)[I + (I+D_n)R_n^*]^{-1} Q_n^* \ .$$

Since $|\delta_n| \leq d < 1$ for all n , B_n is soon shown to be positive definite (<u>not</u> semi-definite); and another compactness argument, like that used in section 1 to establish the existence of $\lambda < 1$, establishes the existence of some constant $\beta > 0$ such that

$$\underline{v}^* B_n \underline{v} \geq \beta \underline{v}^* \underline{v} \quad \text{for all } \underline{v} \text{ and all } n \ .$$

Therefore

$$W(\underline{y}_{n+L}) \leq W(\underline{y}_n) - \beta\|\underline{c}-A\underline{y}_n\|^2 \ .$$

The discussion branches here to deal separately with each of two cases.

<u>Case 1</u>: Suppose $A\underline{y} = \underline{c}$ has no solution. Then some positive constant γ exists such that

$$\|A\underline{y}-\underline{c}\|^2 \geq \gamma > 0 \quad \text{for all} \quad \underline{y} \ .$$

Consequently,

$$W(\underline{y}_{n+kL}) \leq W(\underline{y}_n) - k\,\beta\,\gamma \quad \text{for} \quad k = 1,2,3, \ldots$$

$$\rightarrow -\infty \quad \text{as} \quad k \rightarrow \infty \ .$$

This implies that the sequence $\underline{y}_n$ diverges to infinity at least as quickly as $\sqrt{n}$ as $n \rightarrow \infty$.

(If A is positive semi-definite, the preceding statement can be elaborated slightly; $\underline{y}_n$ diverges to infinity no faster than an arithmetic progression. This follows from a lengthy computation in which $(\underline{c}-A\underline{y}_n)$, and hence $(\underline{y}_{n+L}-\underline{y}_n)$, is shown to be bounded. The crux of the computation consists in pre-multiplying the equation

$$\underline{c} - A\underline{y}_{n+L} = (I-AQ_n[(I+D_n)^{-1} + R_n^*]^{-1}\,Q_n^*)(\underline{c}-A\underline{y}_n)$$

by the matrix U of section 2 to obtain

$$\underline{b} - \underline{x}_{n+L} = T_n(\underline{b}-\underline{x}_n) + S_n\,\hat{\underline{b}} \ .$$

The matrix T_n was defined in section 1 where it was shown to be bounded by $\|T_n\| \leq \lambda < 1$. The matrix S_n can be shown to be bounded too, say by

$$\|S_n\| \leq \sigma \quad \text{for all} \quad n \ .$$

Therefore

$$\|\underline{b}-\underline{x}_{n+L}\| \leq \lambda\|\underline{b}-\underline{x}_n\| + \sigma\|\hat{\underline{b}}\| \quad ,$$

and hence

$$\|\underline{b}-\underline{x}_{n+kL}\| \leq \lambda^k\|\underline{b}-\underline{x}_n\| + \sigma\|\hat{\underline{b}}\|/(1-\lambda) \quad ,$$

whence comes the desired result.)

Case 2: Suppose now that A is indefinite. We already know that $\underline{y}_n$ diverges when $A\underline{y} = \underline{c}$ has no solution, so suppose too that a solution $\underline{y}$ exists. Now there can be no guarantee of divergence because setting $\underline{y}_1 = \underline{y}$ yields a convergent sequence $\underline{y}_n = \underline{y}$. However, if any member of the sequence $\underline{y}_n$ should fall into the open cone

$$(\underline{y}_n-\underline{y})^* A(\underline{y}_n-\underline{y}) < 0 \quad ,$$

then the sequence $\underline{y}_n$ will subsequently remain in the cone as it diverges exponentially to infinity. This is so because

$$\begin{aligned}
&(\underline{y}_{n+L}-\underline{y})^* A(\underline{y}_{n+L}-\underline{y}) - (\underline{y}_n-\underline{y})^* A(\underline{y}_n-\underline{y}) \\
&\quad = W(\underline{y}_{n+L}) - W(\underline{y}_n) \\
&\quad = - (\underline{y}_n-\underline{y})^* A B_n A(\underline{y}_n-\underline{y}) \\
&\quad \leq - \beta(\underline{y}_n-\underline{y})^* A^2(\underline{y}_n-\underline{y}) \\
&\quad \leq \alpha\beta(\underline{y}_n-\underline{y})^* A(\underline{y}_n-\underline{y}) < 0 \quad ,
\end{aligned}$$

where α is a positive constant defined by

$$\alpha \equiv \min(-\ \underline{v}^*A^2\underline{v}/\underline{v}^*\ A\ \underline{v}\ ) \quad \text{over} \quad \underline{v}^*A\ \underline{v} < 0$$

$$= -\ (\text{the negative eigenvalue of } A \text{ closest to zero})\ .$$

Therefore it is possible to place a lower bound upon

$$\begin{aligned} \|A\|\ \|\underline{y}_{n+kL}-\underline{y}\|^2 &\geq -\ (\underline{y}_{n+kL}-\underline{y})^*\ A(\underline{y}_{n+kL}-\underline{y}) \\ &\geq -\ (1+\alpha\beta)^k(\underline{y}_n-\underline{y})^*\ A(\underline{y}_n-\underline{y}) \end{aligned}$$

which shows that $\underline{y}_{n+kL}$ diverges to infinity at least as quickly as $(1+\alpha\beta)^{k/2}$ as $k \rightarrow \infty$. The foregoing argument can be refined slightly to show that exponential divergence takes place whenever some member of the sequence $\underline{y}_n$ falls into the slightly larger open cone

$$(\underline{y}_n-\underline{y})^*\ (A-\beta A^2)(\underline{y}_n-\underline{y}) < 0$$

Conversely, whenever the sequence $\underline{y}_n$ diverges exponentially, the sequence ultimately falls into the cone $(\underline{y}_n-\underline{y})^*\ A(\underline{y}_n-\underline{y}) < 0$ and never gets out. This is so because the exponential divergence of $\underline{y}_n$ implies exponential divergence of $(\underline{y}_{n+L}-\underline{y}_n)$, which implies exponential divergence of $\underline{c} - A\underline{y}_n$, which implies exponential divergence of $W(\underline{y}_n)$ to $-\infty$.

Exponential divergence is possible, but how likely is it? In a sense to be made more precise later, divergence is almost certain provided the sequences $\underline{q}_n$ and δ_n are chosen in advance, as indicated in section 2, before $\underline{y}_n$ is known. The proof is based upon the linear relation between $\underline{y}_1$ and $\underline{y}_n$;

$$\underline{y}_n - \underline{y} = E_n(\underline{y}_1 - \underline{y}) \quad \text{where}$$

$$E_1 = I \quad \text{and, for} \quad n = 1,2,3, \ldots ,$$

$$E_{n+1} = [I - (1+\delta_n)\, \underline{q}_n\, \underline{q}_n{}^*A/\underline{q}_n{}^*A\, \underline{q}_n]\, E_n \quad .$$

Note that E_n is independent of $\underline{y}_1$, though it does depend upon A and the sequences $\underline{q}_n$ and δ_n. Also, $\|E_n\| \to \infty$ exponentially as $n \to \infty$ because, if $\underline{y}_1$ is any vector in the cone mentioned above,

$$\|E_n\| \geq \|E_n(\underline{y}_1 - \underline{y})\| / \|\underline{y}_1 - \underline{y}\|$$

$$= \|\underline{y}_n - \underline{y}\| / \|\underline{y}_1 - \underline{y}\| \to \infty \quad \text{exponentially} \quad .$$

On the other hand, suppose $\underline{y}_1$ could be so chosen that $\|\underline{y}_n - \underline{y}\|$ did not diverge exponentially. In other words, suppose

$$\|\underline{y}_n - \underline{y}\| \leq e_n \|\underline{y}_1 - \underline{y}\| \quad \text{for each } n , \quad \text{where}$$

$$e_n \exp(-nt) \to 0 \quad \text{as} \quad n \to \infty \quad \text{for all} \quad t > 0 \quad .$$

Could such a $\underline{y}_1$ exist; and if so, where?

The region in which $\underline{y}_1$ must lie to satisfy

$$\|\underline{y}_n - \underline{y}\| = \|E_n(\underline{y}_1 - \underline{y})\| \leq e_n \|\underline{y}_1 - \underline{y}\|$$

is a closed cone C_n :

$$(\underline{y}_1 - \underline{y})^* \, (E_n{}^* E_n)(\underline{y}_1 - \underline{y}) \leq e_n^2 \, (\underline{y}_1 - \underline{y})^* \, (\underline{y}_1 - \underline{y}) \quad .$$

The shape of C_n depends upon the eigenvalues of

$$e_n^{-2} E_n^* E_n - I \quad ;$$

at least one eigenvalue must be negative if C_n is not to collapse to a point $\underline{y}_1 = \underline{y}$. But no negative eigenvalue can be less than -1 , while the largest eigenvalue is

$$\|E_n\|^2/e_n^2 \to \infty \quad \text{as} \quad n \to \infty \quad .$$

Therefore, as $n \to \infty$, C_n tends to become flat like a hyperplane, and so the region common to all cones C_n must be either the point $\underline{y}$ or a hyperplane $\mathcal{H}_1$ through $\underline{y}$. $\mathcal{H}_1$ depends only upon A and the sequences $\underline{q}_n$ and δ_n . The dimensionality of $\mathcal{H}_1$ is definitely less than the dimension of the whole $\underline{y}$-space .

Now it is clear that $\underline{y}_n$ will diverge exponentially unless $\underline{y}_1$ lies in $\mathcal{H}_1$. Since $\mathcal{H}_1$ is of measure zero relative to the space in which $\underline{y}_1$ might otherwise be chosen, the probability seems small that $\underline{y}_1$ will lie in $\mathcal{H}_1$. And even if one were to succeed in placing $\underline{y}_1$ in $\mathcal{H}_1$, one faces a considerable risk that roundoff will throw some $\underline{y}_n$ out of its corresponding hyperplane $\mathcal{H}_n = E_n \mathcal{H}_1$. From a practical point of view there is ample justification for concluding that, when A is indefinite, the relaxation iteration is almost certain ultimately to diverge exponentially unless the sequences $\underline{q}_n$ and δ_n are carefully correlated with $\underline{y}_n$.

REFERENCES

S. Agmon (1954) "The Relaxation Method for Linear Inequalities", Can. Journ. Math. VI p. 382-392.

J. Descloux (1963) "Note on the Round-off Errors in Iterative Processes" Math. of Comp. XVII p. 18-27.

G. E. Forsythe and W. Wasow (1960) "Finite-Difference Methods for Partial Differential Equations" New York, p. 240, p. 376-7.

P. R. Garabedian (1956) "Estimation of the Relaxation Factor for Small Mesh Size" Math. Tables and Aids to Comput'n 62 p. 219-235.

G. H. Golub (1962) "Bounds for the Round-off Errors in the Richardson Second Order Method" BIT 2 p. 212-223.

S. Kaczmarz (1937) "Angenäherte Auflösung von Systemen linearer Gleichungen", Bull. Internat. Acad. Polon. Sci. Cl. A. p. 355-357.

W. Kahan (1958) "Gauss-Seidel Methods for Solving Large Systems of Linear Equations" A Microfilm copy of this Ph.D. thesis is available from the Library of the University of Toronto.

W. Kahan (1966) "Relaxation Methods for an Eigenproblem" CS44, Stanford University. Computer Science Department.

H. B. Keller (1965) "On the Solution of Singular and Semidefinite Linear Systems by Iteration" J. SIAM Ser. B: Numer. Anal. 2 p. 281-290.

R. S. Martin, G. Peters, J. H. Wilkinson (1965) "Symmetric Decomposition of a Positive Definite Matrix" Numerische Math. 7 p. 362-383.

A. M. Ostrowski (1954) "On the Linear Iteration Procedure for Symmetric Matrices" Rend. Mat. e App. (Roma) XIV p. 140-163.

A. M. Ostrowski (1965) "Contributions to the Method of Steepest Descent" (to appear).

J. D. Riley (1955) "Solving Systems of Linear Equations with a Positive Definite Symmetric, but possibly Ill-conditioned Matrix" MTAC IX p. 96-101.

S. Schechter (1959) "Relaxation Methods for Linear Equations" Comm. Pure and Appl. Math. XII p. 313-335.

R. S. Varga (1962) "Matrix Iterative Analysis", New Jersey, p. 75.