See note inside cover

NPL Report NAC 69
September 1976

National Physical Laboratory

Division of Numerical Analysis and Computing

THE PRACTICAL BEHAVIOUR OF LINEAR ITERATIVE METHODS WITH PARTICULAR REFERENCE TO S.O.R.

by S.Hammarling & J H Wilkinson FRS

Department of Industry

Approved on behalf of Director, NPL by
Mr E L Albasiny, Superintendent, Division of Numerical Analysis and Computing

NPL Report NAC 69

September 1976

N A T I O N A L P H Y S I C A L L A B O R A T O R Y

THE PRACTICAL BEHAVIOUR OF LINEAR ITERATIVE METHODS WITH PARTICULAR REFERENCE TO S.O.R.

by

S Hammarling* and J H Wilkinson FRS

DIVISION OF NUMERICAL ANALYSIS AND COMPUTING

*on leave from Middlesex Polytechnic

1 INTRODUCTION

For many iterative methods of solving n x n systems of linear equations the error vector e_r of the rth approximate solution x_r satisfies a relation of the form

$$e_{r+1} = P_r e_r \tag{1.1}$$

where P_r is an n x n iteration matrix related to some splitting of the matrix of coefficients. Putting

$$Q_r = P_r \ldots P_1 P_o \tag{1.2}$$

equation (1.1) gives

$$e_{r+1} = Q_r e_o \, . \tag{1.3}$$

Convergence is then usually established by showing that either

$$\| P_r \| \leq \delta < 1 \tag{1.4}$$

for some convenient norm, or that

$$\rho(P_r) \leq \delta < 1 \tag{1.5}$$

where $\rho(P_r)$ is the spectral radius of P_r. In either case this ensures that

$$\lim_{r \to \infty} Q_r = 0 \text{ and } \lim_{r \to \infty} e_r = 0. \tag{1.6}$$

The asymptotic rate of convergence is governed by the speed at which $\delta^r \to 0$. It is not generally appreciated that this concentration on the asymptotic rate of convergence may be extremely misleading as far as the practical behaviour is concerned.

In some of the more important iterative methods $P_r = P$, independent of r; these are the so-called stationary iterative processes. We shall demonstrate our thesis by means of one of the best known stationary iterative processes, the successive over-relaxation (S.O.R.) method and a special case of this, the Gauss-Seidel method. We first describe the S.O.R. algorithm. Consider the system of equations

$$Ax = b \tag{1.7}$$

and write

$$A = L+D+U \tag{1.8}$$

where L, D and U are the matrices formed respectively by the sub-diagonal, diagonal and super-diagonal elements of A. The Gauss-Seidel method arises by writing equation (1.7) in the form

$$(L+D)x = b-Ux$$

and the Gauss-Seidel iteration is then defined by the relation

$$(L+D)x_{r+1} = b-Ux_r \ , \ r = 0,1,... \tag{1.9}$$

From a given x_r this relationship allows us to determine the elements of x_{r+1} in their natural order. The S.O.R. method is derived from the Gauss-Seidel method by taking the change in each element of x_r to be a multiple, ω, of the change that would be obtained by using the Gauss-Seidel method at that stage. Hence the S.O.R. iteration is

$$x_{r+1} = x_r + \omega[D^{-1}(b-Lx_{r+1}-Ux_r)-x_r], \ r = 0,1,... \tag{1.10}$$

which can be written as

$$(L+\tfrac{1}{\omega}D)x_{r+1} = b-[(1-\tfrac{1}{\omega})D+U]\ x_r \ , \tag{1.11}$$

so that S.O.R. corresponds to writing equation (1.7) in the form

$$(L+\tfrac{1}{\omega}D)x = b-[(1-\tfrac{1}{\omega})D+U]\ x \ .$$

S.O.R. reduces to Gauss-Seidel when $\omega = 1$. From equation (1.11) the error vector $e_r = x-x_r$ satisfies

$$(\omega L+D)e_{r+1} = [(1-\omega)D-\omega U]\ e_r$$

so that

$$e_{r+1} = (\omega L+D)^{-1}[(1-\omega)D-\omega U]\ e_r \ . \tag{1.12}$$

Thus the S.O.R. iteration matrix is the matrix $P(\omega)$ given by

$$P(\omega) = (\omega L+D)^{-1}\,[(1-\omega)D-\omega U]\ . \qquad (1.13)$$

It is well known that if A is a strictly diagonally dominant matrix then S.O.R. is convergent if $0 < \omega < 2$. The proof of convergence depends upon showing that

$$\rho\,[P(\omega)] < 1 \quad \text{for} \quad 0 < \omega < 2.$$

If $\rho[P(\omega)]$ is appreciably less than unity then the asymptotic rate of convergence must be very satisfactory; however we show in the next two sections that the actual behaviour of e_r may be extremely disappointing.

2 THE EARLY BEHAVIOUR OF S.O.R.

We first illustrate, by means of a simple example, that the 'early' behaviour of S.O.R. may be very poor. Consider the 2×2 matrix defined by

$$P = \begin{bmatrix} 0.5 & 0 \\ 10^{10} & 0.5 \end{bmatrix}\ .$$

The eigenvalues of P are given by $\lambda_1 = \lambda_2 = 0.5$ so that $\rho(P) = 0.5$. Hence if we have an iterative process for which $e_{r+1} = Pe_r$ its ultimate theoretical rate of convergence must be very satisfactory. However we have that

$$P^r = 0.5^r \begin{bmatrix} 1 & 0 \\ 2r \times 10^{10} & 1 \end{bmatrix}$$

so that if the initial error vector is $e_o^T = [1 \quad 0]$ then $e_1^T = [0.5 \quad 10^{10}]$, a vast increase. It is not until the 40th iteration that the components of e_r are both less than unity and we can really be said to be obtaining the benefit of the satisfactory theoretical rate of convergence.

There are norms for which $\|P\| < 1$, but they are highly artificial. For example the norm defined by

$$\|P\|_K = \|K^{-1}PK\|_\infty$$

where

$$K = \begin{bmatrix} 1 & 0 \\ 0 & 10^{11} \end{bmatrix}$$

gives

$$K^{-1}PK = \begin{bmatrix} 0.5 & 0 \\ 0.1 & 0.5 \end{bmatrix}$$

so that

$$\|P\|_K = 0.6.$$

However it is immediately evident that for any vector norm consistent with this matrix norm we have

$$\left\| \begin{matrix} 1 \\ 0 \end{matrix} \right\|_K = 10^{11} \left\| \begin{matrix} 0 \\ 1 \end{matrix} \right\|_K$$

which is clearly an undesirable feature.

The reader could be excused for regarding this example as very artificial. For a matrix of order 2 examples illustrating our point are indeed unsatisfactory, but for matrices of higher order, though not excessively high order, poor behaviour can arise with examples which are by no means bizarre.

Consider the case when A is the lower triangular matrix defined by

$$A = L + \alpha I \tag{2.1}$$

where L is such that

$$l_{ij} = \begin{cases} 1, & j = i-1 \\ 0, & j \neq i-1 \ . \end{cases} \tag{2.2}$$

A is illustrated when n = 4 by

$$A = \begin{bmatrix} \alpha & 0 & 0 & 0 \\ 1 & \alpha & 0 & 0 \\ 0 & 1 & \alpha & 0 \\ 0 & 0 & 1 & \alpha \end{bmatrix}.$$

When $\alpha > 1$, A is strictly diagonally dominant. The iteration matrix $P_n(\omega)$ of order n corresponding to S.O.R. is given by

$$P_n(\omega) = (\omega L+\alpha I)^{-1} [(1-\omega)\alpha I]. \qquad (2.3)$$

When $\omega = 1$ this is the null matrix so that the Gauss-Seidel method converges in one iteration starting from any initial approximation, this being true of course for any lower triangular matrix. Equation (2.3) can be written as

$$P_n(\omega) = (1-\omega)(I-\beta L)^{-1}, \quad \beta = -\frac{\omega}{\alpha}. \qquad (2.4)$$

The form of $P_n(\omega)$ is adequately illustrated by

$$P_4(\omega) = (1-\omega) \begin{bmatrix} 1 & 0 & 0 & 0 \\ \beta & 1 & 0 & 0 \\ \beta^2 & \beta & 1 & 0 \\ \beta^3 & \beta^2 & \beta & 1 \end{bmatrix}.$$

The eigenvalues of $P_n(\omega)$ are all $(1-\omega)$ so that

$$\rho [P_n(\omega)] = |1-\omega| \qquad (2.5)$$

and if $|1-\omega|$ is appreciably less than unity the asymptotic rate of convergence must be satisfactory. However successive powers of $P_n(\omega)$ are of the form illustrated by

$$P_4^2(\omega) = (1-\omega)^2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2\beta & 1 & 0 & 0 \\ 3\beta^2 & 2\beta & 1 & 0 \\ 4\beta^3 & 3\beta^2 & 2\beta & 1 \end{bmatrix}, \quad P_4^3(\omega) = (1-\omega)^3 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 3\beta & 1 & 0 & 0 \\ 6\beta^2 & 3\beta & 1 & 0 \\ 10\beta^3 & 6\beta^2 & 3\beta & 1 \end{bmatrix},$$

$$P_4^4(\omega) = (1-\omega)^4 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 4\beta & 1 & 0 & 0 \\ 10\beta^2 & 4\beta & 1 & 0 \\ 20\beta^3 & 10\beta^2 & 4\beta & 1 \end{bmatrix}, \quad \text{etc.}$$

In the general case the $(n,1)$ element, $p_{n1}^{(r)}$, of $P_n^r(\omega)$ is given by

$$p_{n1}^{(r)} = {}^{n+r-2}C_{r-1}(1-\omega)^r\beta^{n-1} \tag{2.6}$$

and even when $|(1-\omega)\beta|$ is appreciably less than unity the absolute value of this element continues to increase for some time. In fact it will increase until

$$\left(\frac{n+r-1}{r}\right)|1-\omega| < 1 ,$$

that is until

$$r > \frac{n-1}{\frac{1}{|1-\omega|}-1} .$$

Suppose, for example, that we have the values

$$\omega = 1.5, \; n = 100, \; \alpha = 1.5 . \tag{2.7}$$

Then despite the fact that $\rho[P_n(\omega)] = 0.5$, $|p_{n1}^{(r)}|$ will increase until $r = 100$ at which point

$$p_{100,1}^{(100)} = {}^{198}C_{99}(-0.5)^{100}(-1)^{99} \simeq -1.8 \times 10^{28}.$$

Thus an initial error vector of $e_o^T = [1 \; 0 \; \ldots 0]$ gives rise to an error vector e_{100} whose nth element is -1.8×10^{28}. It is not until iteration 330 that all the elements of e_r are less than unity in absolute value.

Once again we can find norms for which $\|P\| < 1$, but again they are highly artificial. For instance if we take

$$K = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & \gamma & \ldots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \ldots & \gamma^{n-1} \end{bmatrix} \tag{2.8}$$

then $K^{-1}P_n(\omega)K$ is of the form illustrated by

$$K^{-1}P_4(\omega)K = (1-\omega)\begin{bmatrix} 1 & 0 & 0 & 0 \\ \theta & 1 & 0 & 0 \\ \theta^2 & \theta & 1 & 0 \\ \theta^3 & \theta^2 & \theta & 1 \end{bmatrix}, \quad \theta = \frac{\beta}{\gamma} \tag{2.9}$$

and a suitable choice of γ will ensure that

$$\| P_n(\omega) \|_K = \| K^{-1}P_n(\omega)K \|_\infty < 1.$$

For example with $\omega = 1.5$, $\alpha = 1.5$ and $\gamma = 10$ we find that

$$\| P_n(\omega) \|_K = \frac{5}{9}\left[1-(0.1)^n\right] < 1.$$

Unfortunately any vector norm consistent with this will be such that

$$\| e_i \| = \gamma^{1-i} \| e_1 \|$$

where here e_i is used to denote the ith column of the identity matrix. Such a norm may therefore be far less sensitive to changes in the nth component of a vector than to changes in the first. Typical of such vector norms is $\| x \|_K$ defined by

$$\| x \|_K = \| K^{-1}x \|_\infty = \max | \gamma^{1-i} x_i | .$$

Clearly $\| x \|_K$ may be very small even when a more natural norm such as $\| x \|_\infty$ is quite large. (The matrix norm $\| A \|_K$ is, of course, the norm induced by the vector norm $\| x \|_K$).

It is worth noting that the matrix A of equation (2.1) is such that

$$\| A \|_\infty \| A^{-1} \|_\infty = \left(\frac{\alpha+1}{\alpha-1}\right)\left(1-\frac{1}{\alpha^n}\right)$$

and when $\alpha = 1.5$ this becomes

$$\| A \|_\infty \| A^{-1} \|_\infty = 5\left(1-\frac{1}{1.5^n}\right) < 5$$

so that A is very well-conditioned. Hence, for this example the poor early behaviour is certainly <u>not</u> due to A being ill-conditioned.

3 THE BEHAVIOUR OF S.O.R. IN THE PRESENCE OF ROUNDING ERRORS

In the previous section we showed that any error present in our initial approximation can grow extremely large before we begin to feel the effect of the asymptotic convergence. In practice since we are forced to use a finite arithmetic, such as t-digit binary arithmetic, we shall introduce rounding errors at each stage of the computation so that the "early" behaviour may in fact become the ultimate behaviour.

To illustrate the point S.O.R. was applied, using a KDF9 for which t = 39, to the two sets of equations

$$Ax = b \text{ and } Ay = c \tag{3.1}$$

where A is the matrix of equation (2.1) and b and c are given by

$$b_i = \begin{cases} 1.5, & i = 1 \\ 2.5, & i = 2,3,\ldots,n \end{cases} \quad ; \quad c_i = 2.5,\ i = 1,2,\ldots,n. \tag{3.2}$$

The exact solutions to these equations are given by

$$x_i = 1,\ i = 1,2,\ldots,n \ ; \ y_i = 1 - \left(\frac{-1}{1.5}\right)^i,\ i = 1,2,\ldots,n \tag{3.3}$$

so that in the first case the solution is exactly representable in t-digit binary arithmetic, but in the second case the solution is not exactly representable.

The values of ω, n and α were taken as in equation (2.7) and the initial approximations to the solutions were

$$x_i^{(o)} = \begin{cases} 1+10^{-8}, & i = 1 \\ 1, & i = 2,3,\ldots,n \end{cases} \quad ; \quad y_i^{(o)} = \bar{y}_i,\ i = 1,2,\ldots,n \tag{3.4}$$

where $\bar{y}_i$ is the best machine representation of y_i.

In the first example the computation proceeded very much as though rounding errors were absent. The maximum error occurred on the 100th iteration when

$$e_{100}^{(100)} \simeq 1.8002 \times 10^{20} = -1.8002 \times 10^{28} e_1^{(o)} .$$

After this $\| e_r \|_\infty$ gradually diminished until by iteration 329, $\| e_{329} \|_\infty < 10^{-8}$. The exact solution was obtained on iteration 347.

In the second example, since y_i is not exactly representable, even when $y_1^{(r)}$ is correct to working accuracy an error is still propogated down to $y_{100}^{(r)}$ so that we cannot escape the effect of rounding errors. Thus, despite an initial approximation which was correct to working accuracy, $\| e_r \|_\infty$ gradually increased and reached a maximum at iteration 214 at which point

$$\| e_{214} \|_\infty \simeq 2.881 \times 10^{17} .$$

From this point on $\| e_r \|_\infty$ remained fixed at this value with the element $e_{100}^{(r)}$ oscillating between $\pm 2.881 \times 10^{17}$. On this example we should need to use at least 97 digit binary arithmetic just to obtain one figure accuracy.

As a further illustration of poor practical performance we constructed an example for which we expected the method of Gauss–Seidel to perform badly. Consider the matrix A given by

$$A = L + \alpha I + U \qquad (3.5)$$

where L and U are strictly lower and upper triangular matrices respectively with

$$l_{ij} = \alpha, \; i > j \text{ and } u_{ij} = (-1)^{j-i+1}, \; j > i . \qquad (3.6)$$

A has the form illustrated when n = 6 by

$$A = \begin{bmatrix} \alpha & 1 & -1 & 1 & -1 & 1 \\ \alpha & \alpha & 1 & -1 & 1 & -1 \\ \alpha & \alpha & \alpha & 1 & -1 & 1 \\ \alpha & \alpha & \alpha & \alpha & 1 & -1 \\ \alpha & \alpha & \alpha & \alpha & \alpha & 1 \\ \alpha & \alpha & \alpha & \alpha & \alpha & \alpha \end{bmatrix} .$$

The corresponding Gauss–Seidel iteration matrix is given by

$$P_n = -\left(\frac{1}{\alpha}\right)\left(\frac{1}{\alpha} L+I\right)^{-1}U \tag{3.7}$$

and the form of P_n is adequately illustrated by

$$P_6 = \frac{1}{\alpha}\begin{bmatrix} 0 & -1 & 1 & -1 & 1 & -1 \\ 0 & 1 & -2 & 2 & -2 & 2 \\ 0 & 0 & 1 & -2 & 2 & -2 \\ 0 & 0 & 0 & 1 & -2 & 2 \\ 0 & 0 & 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} .$$

The spectral radius of P_n is given by

$$\rho(P_n) = 1/|\alpha| \tag{3.8}$$

so that if $|\alpha|$ is appreciably larger than unity the asymptotic convergence must be good.

The form of P_n suggests that some elements of P_n^r may get very large in absolute value before the effect of the asymptotic convergence is felt. If we take

$$\alpha = -3, \; n = 50 \tag{3.9}$$

then we find that the absolutely largest element occurs when $r = 36$ and is given by

$$| p_{2n}^{(r)} | \simeq 1.254 \times 10^{13} .$$

The method of Gauss–Seidel was applied, again using a KDF9, to the equations

$$Ax = b \tag{3.10}$$

where A, α and n are given by equations (3.5) and (3.9) and b is given by

$$b_i = \begin{cases} 1-3i, & i \text{ odd} \\ -3i, & i \text{ even.} \end{cases} \tag{3.11}$$

The exact solution to equations (3.10) is given by

$$x_i = 1, \quad i = 1,2,\ldots,n; \quad n \text{ even.} \tag{3.12}$$

An initial approximation of

$$x_i^{(o)} = 1, \quad i = 1,2,\ldots,49; \quad x_{50}^{(o)} = 1+10^{-8} \tag{3.13}$$

was taken. $\| e_r \|_\infty$ quickly increased and reached an initial maximum when $r = 36$ at which point

$$\| e_{36} \|_\infty \simeq 1.25 \times 10^{5} .$$

After this $\| e_r \|_\infty$ oscillated somewhat, but settled approximately at a level of the order of 10^3. As in the previous example, this poor practical behaviour is not due to A being particularly ill-conditioned because for this example

$$\| A \|_\infty \| A^{-1} \|_\infty = 3n\left(\frac{3}{4} - \frac{1}{2^{n-1}} \right), \quad \alpha = -3$$

$$< \frac{9n}{4} .$$

When $n = 50$, A is therefore quite well conditioned from the point of view of 39 binary digit computation.

4 ERROR ANALYSIS FOR THE PRACTICAL CONVERGENCE OF S.O.R.

In this section we wish to establish a bound on $\overline{\lim_{r \to \infty}} \| e_r \| / \| x \|$ where x is the exact solution of the equations $Ax = b$, $e_r = x_r - x$ and x_r is the rth computed iterate.

From equation (1.12) we can see that S.O.R. belongs to the class of iteration methods that satisfy

$$x_{r+1} = P_r x_r + (I-P_r)x \ , \ r = 0,1,... \tag{4.1}$$

Computationally, in place of equation (4.1), x_{r+1} will satisfy an equation of the form

$$x_{r+1} = P_r x_r + (I-P_r)x + \varepsilon_r \tag{4.2}$$

where ε_r is the difference between the computed x_{r+1} and the vector which would have been obtained with exact computation starting from the computed x_r. Let us suppose that the iteration matrix is such that for some norm

$$\| P_r \| \leq \delta < 1, \ r = 0,1,... \tag{4.3}$$

so that asymptotic convergence is assured and also let β be a value such that

$$\| \varepsilon_r \| \leq \beta \| x \|, \ r = 0,1,... \tag{4.4}$$

Then from equation (4.2) we find that

$$e_{r+1} = P_r e_r + \varepsilon_r \tag{4.5}$$

so that

$$\| e_{r+1} \| \leq \delta \| e_r \| + \beta \| x \| \tag{4.6}$$

from which it follows that

$$\| e_r \| \leq \delta^r \| e_o \| + \beta \left(\frac{1-\delta^r}{1-\delta} \right) \| x \| . \tag{4.7}$$

Hence

$$\overline{\lim_{r\to\infty}} \frac{\|e_r\|}{\|x\|} \leq \frac{\beta}{1-\delta} . \tag{4.8}$$

To look at the particular case of S.O.R. we now make some additional simplifying assumptions. We assume that t-digit binary arithmetic with a unit rounding error of 2^{-t} is used and that a bound of the form

$$(1-2^{-t})^r \leq 1+\varepsilon \leq (1+2^{-t})^r$$

can, without undue optimism, be replaced by

$$|\varepsilon| \leq r2^{-t} .$$

The computed values will not be of any practical interest unless from some point onwards $\|x_r\|$ is of the order of magnitude of $\|x\|$. We shall assume that there exists a positive integer k such that

$$\|x_r\| \leq 2\|x\| \quad (r \geq k) \tag{4.9}$$

though it will be appreciated that in some of the examples we discussed in section 3 the behaviour was so bad that even this was not true.

Now, from equation (1.10), we have computationally that

$$x_i^{(r+1)} = fl\{\omega(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(r+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(r)})/a_{ii} + (1-\omega)x_i^{(r)}\} , \quad i = 1,2,\ldots,n. \tag{4.10}$$

In most applications of S.O.R. the matrix of coefficients is sparse. Let us suppose that each row of A contains at most m non-zero elements. Equation (4.10) then leads to

$$x_i^{(r+1)} = \omega\{b_i(1+\gamma_i^{(r)}) - \sum_{j=1}^{i-1} a_{ij}x_j^{(r+1)}(1+\eta_{ij}^{(r)}) - \sum_{j=i+1}^{n} a_{ij}x_j^{(r)}(1+\eta_{ij}^{(r)})\}/a_{ii}$$

$$+(1-\omega)x_i^{(r)}(1+\theta_i^{(r)}), \tag{4.11}$$

where

$$|\gamma_i^{(r)}| \leq (m+1)2^{-t}, \quad |\eta_{ij}^{(r)}| \leq (m+2)2^{-t}, \quad |\theta_i^{(r)}| \leq 2.2^{-t} . \tag{4.12}$$

If we define the error vector f_r by the equation

$$x_{r+1} = \omega D^{-1}(b-Lx_{r+1}-Ux_r)+(1-\omega)x_r+f_r \tag{4.13}$$

then the ith element of f_r satisfies

$$|f_i^{(r)}| \leq \{|\omega|(|b_i|+\sum_{j=1}^{i-1}|a_{ij}x_j^{(r+1)}|+\sum_{j=i+1}^{n}|a_{ij}x_j^{(r)}|)(m+2)2^{-t}\} /|a_{ii}|+2|1-\omega||x_i^{(r)}|2^{-t}$$

and using assumption (4.9) this gives

$$\|f_r\| \leq 2\{(m+2)|\omega|[\|D^{-1}A\|+\|D^{-1}(L+U)\|]+2|1-\omega|\}2^{-t}\|x\| \tag{4.14}$$

provided we use an absolute norm, that is a norm for which $\|\,|A|\,\| = \|A\|$.

From equation (4.13) we obtain

$$x_{r+1} = P(\omega)x_r+(L+\tfrac{1}{\omega}D)^{-1}b+(I+\omega D^{-1}L)^{-1}f_r \tag{4.15}$$

and if we put

$$\varepsilon_r = (I+\omega D^{-1}L)^{-1}f_r \tag{4.16}$$

then this gives

$$\begin{aligned} e_{r+1} &= P(\omega)x_r-x+(L+\tfrac{1}{\omega}D)^{-1}Ax+\varepsilon_r \\ &= P(\omega)x_r-(L+\tfrac{1}{\omega}D)^{-1}(L+\tfrac{1}{\omega}D-L-D-U)x+\varepsilon_r \\ &= P(\omega)x_r-P(\omega)x+\varepsilon_r \\ &= P(\omega)e_r+\varepsilon_r \end{aligned} \tag{4.17}$$

which is of the same form as equation (4.5) so that equation (4.8) holds where β is such that

$$\beta \leq 2\|(I+\omega D^{-1}L)^{-1}\|\{(m+2)|\omega|[\|D^{-1}A\|+\|D^{-1}(L+U)\|]+2|1-\omega|\}2^{-t} . \tag{4.18}$$

In the case of Gauss–Seidel where $\omega = 1$ this becomes

$$\beta \leq 2\| (I+D^{-1}L)^{-1} \| \{ \| D^{-1}A \| + \| D^{-1}(L+U) \| \} (m+2)2^{-t} . \qquad (4.19)$$

We now give two examples of the use of these bounds. First we consider the matrix A of equation (2.1) with the choice of parameters given by equation (2.7). If we take the norm

$$\| A \|_K = \| K^{-1}AK \|_\infty \quad , \text{ K diagonal} \qquad (4.20)$$

then we have already seen that the matrix

$$K = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 10 & \dots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \dots & 10^{99} \end{bmatrix}$$

gives

$$\| P_n(\omega) \|_K < \frac{5}{9}$$

and so for this choice of norm we can certainly take

$$\delta = \frac{5}{9} .$$

For this particular matrix

$$(I+\omega D^{-1}L)^{-1} = \frac{1}{1-\omega} P_n(\omega)$$

so that

$$\| (I+\omega D^{-1}L)^{-1} \|_K < \frac{10}{9} .$$

We also have that

$$\| D^{-1}A \|_K = 1 + \frac{0.1}{|\alpha|} = \frac{16}{15} \text{ and } \| D^{-1}(L+U) \|_K = \frac{0.1}{|\alpha|} = \frac{1}{15}$$

so that substitution into (4.18) and (4.8) gives

$$\beta \leq \frac{52}{3} \times 2^{-t}$$

and

$$\overline{\lim_{r\to\infty}} \frac{\|e_r\|}{\|x\|} \leq 39 \times 2^{-t} .$$

At first sight this looks to be a very satisfactory result, but we have to remember that the result is true only for vector norms which are consistent with the matrix norm $\|A\|_K$. If we take the vector norm $\|x\|_K$ defined by

$$\|x\|_K = \|K^{-1}x\|_\infty = \max_i (10^{1-i}|x_i|)$$

our bound merely implies that

$$\overline{\lim_{r\to\infty}} \{\max_i (10^{1-i}|e_i^{(r)}|)\} \leq 39 \max_i (10^{1-i}|x_i|)2^{-t} .$$

If the elements of x are all of the same order of magnitude then

$$\overline{\lim_{r\to\infty}} \{\max_i (10^{1-i}|e_i^{(r)}|)\} \leq 39\ |x_1|\ 2^{-t}$$

and this bound can be satisfied even if

$$|e_n^{(r)}| = 39\ |x_1|10^{n-1}2^{-t} = 39\ |x_1|10^{n-1-t\mathrm{Log}_{10}2} .$$

As a second example we consider the use of the method of Gauss–Seidel on a strictly diagonally row dominant matrix A. That is A is such that

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| , \quad i = 1,2,\ldots,n. \tag{4.21}$$

The usual proof of convergence of Gauss–Seidel with such a matrix is based on showing that $\rho(P) < 1$ where P is the iteration matrix. However, a straightforward inductive proof shows that there exists a δ such that

$$\|P\|_\infty \leq \delta < 1 \tag{4.22}$$

and that

$$\|(I+D^{-1}L)^{-1}\|_\infty \leq n. \tag{4.23}$$

Since we also have that

$$\| D^{-1}A \|_\infty \leq 2 \text{ and } \| D^{-1}(L+U) \|_\infty \leq 1 \qquad (4.24)$$

substitution into equations (4.19) and (4.8) show that

$$\beta \leq 6n(m+2)2^{-t}$$

and

$$\overline{\lim_{r\to\infty}} \frac{\| e_r \|_\infty}{\| x \|_\infty} \leq \frac{6n(m+2)}{1-\delta} 2^{-t} . \qquad (4.25)$$

In terms of the elements of e_r and x this gives

$$\overline{\lim_{r\to\infty}} |e_i^{(r)}| \leq \frac{6n(m+2)}{1-\delta} (\max_i |x_i|)2^{-t}$$

and assuming that the elements of x are of roughly the same order of magnitude, we can guarantee a computed solution which is close to the exact solution so long as δ is not too close to unity. Notice also that equation (4.22) implies

$$\| P^r \|_\infty \leq \delta^r < 1 \qquad (4.26)$$

so that growth in the elements of P^r cannot occur. A bound on $\| P \|_\infty$ is therefore more satisfactory than a corresponding bound on $\rho(P)$, but of course we have always $\rho(P) \leq \| P \|$ in any norm and hence $\rho(P)$ gives the best estimate of the asymptotic rate of convergence.

The important distinction between these two examples is in the choice of norm. In the first case the size of the norm gives no sensible indication as to the size of elements, whereas in the second example the size of the norm is also a bound on the size of the elements.

From a practical point of view we must be wary of taking results on the asymptotic rate of convergence at their face value. The situation is much more satisfactory if we can also establish that $\| P_r \| < 1$ in some natural norm where we might define such a norm as being one for which

$$\max |a_{ij}| \leq \| A \| \leq n \max |a_{ij}| \qquad (4.27)$$

where A is a $p \times q$ matrix and $n = \max(p,q)$. The norms $\| A \|_1$, $\| A \|_2$, $\| A \|_\infty$ and $\| A \|_F$ all satisfy equation (4.27) whereas norms such as those of equation (4.20) will not in general do so.

It is perhaps worth mentioning that even when $\| P_r \|$ is marginally greater than unity in some natural norms then provided $\rho(P_r)$ is appreciably less than unity we can expect that the errors resulting from inexact arithmetic in any one iteration will not grow too large before the asymptotic rate of convergence begins to assert its authority.

5 CONCLUSION

We have shown that the practical convergence of a linear iterative method can be very different from the theoretical convergence. In particular we have given examples to show that the methods of Gauss-Seidel and S.O.R. can both be very poorly behaved. Although the condition

$$\rho(P_r) \leq \delta < 1$$

guarantees convergence in the mathematical sense we cannot take this at face value. To be <u>certain</u> of satisfactory convergence in practice one must show that there is a natural norm for which

$$\| P_r \| \leq \delta_1 < 1.$$

However, if $\rho(P_r) \leq \delta < 1$ and there is a natural norm for which $\| P_r \|$ is not significantly greater than unity practical convergence is reasonably probable.

REFERENCES

VARGA, R S, 1962, Matrix Iterative Analysis. Prentice-Hall, New Jersey.

WILKINSON, J H, 1963, Rounding Errors in Algebraic Processes, Notes on Applied Science No 32, Her Majesty's Stationery Office, London; Prentice-Hall, New Jersey.

YOUNG, David M, 1954, Iterative methods for solving partial difference equations of elliptic type. Trans.Amer. Math. Soc. 76, 92-111.

Distribution

Standard	150
Library	100

NPL Report NAC 69