A method for finding the optimum successive over-relaxation parameter

By J. K. Reid*

A proof is given here of the well-known relation between the eigenvalues of the Jacobi and S.O.R. iteration matrices in the case having Property A and consistent ordering. This proof also yields a relationship between the corresponding eigenvectors, and we use this relation to form a method of obtaining an approximation to the optimum relaxation parameter.

Analytical results

We consider here the solution by successive over-relaxation of the set of linear equations

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

and suppose that

$$A = D - L - U, \tag{2}$$

where D is diagonal, L is strictly lower-triangular and U is strictly upper-triangular. We will assume that the matrix A possesses Property A and is consistently ordered, that is that for any positive scalar p there exists a diagonal matrix G_p such that

$$p^{1/2}L + p^{-1/2}U = G_p(L + U)G_p^{-1}. \tag{3}$$

This is equivalent to the definitions given by Young (1954). He required the existence of an ordering vector $(q_1, q_2, \ldots q_n)$ with integer coefficients such that if the elements of A are a_{ij} and if $a_{ij} \neq 0$ and $i \neq j$ then either $q_i = q_j + 1$ and $i > j$ or $q_i = q_j - 1$ and $i < j$. If we set $G_p = \operatorname{diag}(p^{q_i/2})$ then equation (3) is satisfied and conversely, given a matrix G_p, it is a simple matter to construct an ordering vector with the necessary properties. Details of this construction are given in the appendix. Hereafter we assume G_p is of the form $\operatorname{diag}(p^{q_i/2})$.

The S.O.R. iteration matrix is

$$M_\omega = (D - \omega L)^{-1}((1 - \omega)D + \omega U). \tag{4}$$

If this has an eigenvalue λ_i then the corresponding eigenvector y_i satisfies the equation

$$((1 - \omega)D + \omega U)y_i = \lambda_i(D - \omega L)y_i \tag{5}$$

and

$$\omega(U + \lambda_i L)y_i = (\lambda_i + \omega - 1)Dy_i, \tag{6}$$

which may be written as

$$\omega\lambda_i^{1/2}(\lambda_i^{-1/2}U + \lambda_i^{1/2}L)y_i = (\lambda_i + \omega - 1)Dy_i \tag{7}$$

provided $\lambda_i \neq 0$. Using (3) and rearranging, we find

$$D^{-1}(L + U)G_{\lambda i}^{-1}y_i = \frac{\lambda_i + \omega - 1}{\omega\lambda_i^{1/2}} G_{\lambda i}^{-1}y_i \tag{8}$$

provided $\omega \neq 0$ (and the case $\omega = 0$ is of no interest to us). Now $D^{-1}(L + U)$ is the Jacobi iteration matrix and (8) shows that it has $G_{\lambda i}^{-1}y_i$ as an eigenvector. If the corresponding eigenvalue is μ_i we find the well-known relation

$$(\lambda_i + \omega - 1)^2 = \lambda_i\omega^2\mu_i^2. \tag{9}$$

It also holds for $\lambda_i = 0$ since in this case we find from (5) that

$$\det\{(1 - \omega)D + \omega U\} = 0 \tag{10}$$

that is

$$(1 - \omega)^n \prod_{i=1}^{n} d_i = 0 \tag{11}$$

if the elements of D are d_i. Now the iteration is not possible unless each d_i is non-zero. It follows that $\omega = 1$ and (9) is still valid.

Practical application for symmetric, positive-definite matrices

In the case where A is symmetric, it is a well-known deduction from equation (9) that the spectral radius of M_ω is minimized if ω is chosen as

$$\omega_{opt} = \frac{2}{1 + (1 - \mu^2)^{1/2}} \tag{12}$$

where $\mu = \max_i |\mu_i|$. This is shown by Varga (1962), for example.

This is very satisfactory as it stands if a good *a priori* estimate for μ is available, but otherwise we need an algorithm that finds it without increasing unduly the total amount of work. Carré (1961) and Kulsrud (1961) each describe useful techniques based on examination of the displacement vectors $\delta^{(i)}$ which satisfy the relation

$$\delta^{(i+1)} = M_\omega\delta^{(i)}. \tag{13}$$

Both rely on the use of a relaxation factor slightly less than ω_{opt} to ensure that the dominant λ_i corresponds to the dominant μ_i. Carré iterates a few times with parameter ω_k suggesting twelve times as suitable. He then makes an estimate, v_k, of the dominant latent root of

* *University of Sussex, Falmer, Brighton, Sussex.*

M_{ω_k} from the ratio of the norms of the last two displacement vectors or by Aitken extrapolation on the last three ratios of successive displacement vector norms. Hence using equations (9) and (12), he estimates μ and then ω_{opt}. If this estimate is ω_k' he continues the iteration using

$$\omega_{k+1} = \omega_k' - \tfrac{1}{4}(2 - \omega_k'). \tag{14}$$

For a wide range of problems he finds that this gives a good estimate for the value of ω for which the ratio of dominant to sub-dominant latent root of M_ω is largest. He continues in this way until successive estimates ω_{k+1} show good agreement and thereafter uses ω_k' as fixed relaxation parameter. Kulsrud's process is essentially the same except that he takes $\omega_{k+1} = \omega_k'$. As he shows in his paper, these estimates ω_k will steadily increase and it is difficult to guarantee that a gross over-estimate will not be obtained, particularly in view of the fact that all the eigenvalues of M_ω are complex for $\omega > \omega_{opt}$. However, he reports that for three test cases he found no more iterations were required with his technique than were needed with the use of ω_{opt} throughout.

An alternative procedure is to exploit the fact (noted just below (8)) that the eigenvectors z_i of the Jacobi matrix are related to the eigenvectors y_i of the S.O.R. matrix by the relation

$$z_i = G_{\lambda_i}^{-1} y_i. \tag{15}$$

Now the displacement vector gives us an estimate of the dominant eigenvector of M_{ω_k} and we can estimate G_λ^{-1} by using the ratio of the last two norms of displacement vectors as an estimate of λ. In this way we find an approximation for the dominant eigenvector of the Jacobi matrix, from which we may form a Rayleigh quotient. This will give a good estimate of μ on account of the well-known stationary property of the Rayleigh quotient. It will furthermore be an underestimate since

$$\mu = \max_{x \neq 0} \frac{x^T(L + L^T)x}{x^T D x}.$$

Hence if we use this approximation to μ to find an estimate ω_k of ω_{opt} via equation (12), then ω_k will be less than ω_{opt} and we will never have trouble with a complex dominant latent root of M_{ω_k}. It is possible, particularly near convergence where round-off errors play a significant role, that the new estimate for ω_{opt} will be smaller than the old one. In such a case the old estimate is certainly the better and should be used.

Numerical experiments

All three techniques for finding ω_{opt} have been tried on three test problems. Since the solutions were known we were able to calculate the norms of the error vectors and these together with the relaxation parameters are tabulated below. For comparison we also used $\omega = \omega_{opt}$ throughout. In each case the zero vector was taken as starting approximation and the iteration terminated by Carré's test, that

$$\frac{v_k \|\delta^{(i)}\|}{1 - v_k}$$

be less than the largest acceptable norm of the error vector. For vector norm we used the Euclidean norm, $\|x\| = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}$, throughout. We followed Carré's starting procedure, as described on pages 76 and 77 of his paper. Kulsrud's method requires an underestimate of ω_{opt} at the start and we used Carré's value, 1·375. For the new method we iterate with $\omega = 1$ just twice, the minimum number that permits us to find a new ω by the method already described. In all three procedures a value of ω_{k+1} was found after twelve iterations with ω_k unless the criteria shown in **Table 1** were satisfied in which case no further improved estimates were found, the iteration being completed with parameter ω_k' in Carré's case and ω_k in the other two cases. We make no claim that these criteria are the best that can be devised. The choice of numerical factors is particularly arbitrary; the figure of 1/20 was suggested by Carré and we have used the same factor in Kulsrud's technique to give a direct comparison. The test used in the new method is based on the assumption that $-\log v_k$ approximates the asymptotic convergence rate with $\omega = \omega_k$ and the fact that $-\log(\omega_k - 1)$ is certainly less than the optimum asymptotic convergence rate, $-\log(\omega_{opt} - 1)$, so that if $v_k^6 < (\omega_k - 1)^5$ we can expect the asymptotic convergence rate to be improved by not more than 20 per cent if iteration with $\omega = \omega_k$ is replaced by iteration with $\omega = \omega_{opt}$. In Kulsrud's method the dominant eigenvalue of M_ω may be complex, in which case it is likely that the ratios of the norms of successive displacement vectors will oscillate severely. To reduce this effect we took for v_k the geometric mean of the last eleven ratios of displacement norms.

Table 1

Criteria for stopping the process of improving ω

METHOD	CRITERION
Kulsrud	$\lvert\omega_k - \omega_{k-1}\rvert < \dfrac{2 - \omega_k}{20}$
Carré	$\lvert\omega_k' - \omega_{k-1}'\rvert < \dfrac{2 - \omega_{k-1}'}{20}$
New method	$v_k^6 < (\omega_{k-1} - 1)^5$

The first example considered was Laplace's equation in a rectangle with five by forty internal mesh-points, so that the matrix A has the block form

$$\begin{bmatrix} T & I & & & \\ I & T & I & & \\ & I & T & I & \\ & & I & T & I \\ & & & I & T \end{bmatrix}$$

Table 2

Results for Laplace's equation in a rectangle

ITERATION NUMBER	ω_{opt}		CARRÉ		KULSRUD		NEW METHOD	
	ω	$\log_{10}\|\|e\|\|$	ω	$\log_{10}\|\|e\|\|$	ω	$\log_{10}\|\|e\|\|$	ω	$\log_{10}\|\|e\|\|$
1	1·4667	0·96	1·0000	1·07	1·3750	0·99	1·0000	1·07
3	1·4667	0·66	1·3750	0·81	1·3750	0·74	1·1488	0·92
9	1·4667	−0·47	1·3750	−0·03	1·3750	−0·11	1·1488	0·43
15	1·4667	−1·63	1·4872	−0·98	1·4752	−1·10	1·4514	−0·14
21	1·4667	−2·80	1·4872	−2·22	1·4752	−2·29	1·4514	−1·20
27	1·4667	−3·98	1·4872	−3·49	1·5078	−3·56	1·4514	−2·30
33	1·4667	−5·19	1·4872	−4·78	1·5078	−4·96	1·4514	−3·41
39	1·4667	−6·46	1·4872	−6·12	1·5174	−6·42	1·4592	−4·54
45	1·4667	−8·01	1·4872	−7·64	1·5174	−7·86	1·4592	−5·72
51							1·4592	−6·97

Table 3

Results for Kulsrud's example

ITERATION NUMBER	ω_{opt}		CARRÉ		KULSRUD		NEW METHOD	
	ω	$\log_{10}\|\|e\|\|$	ω	$\log_{10}\|\|e\|\|$	ω	$\log_{10}\|\|e\|\|$	ω	$\log_{10}\|\|e\|\|$
1	1·7625	1·33	1·0000	1·35	1·3750	1·34	1·0000	1·35
3	1·7625	1·27	1·3750	1·31	1·3750	1·30	1·1218	1·33
15	1·7625	0·74	1·6706	1·11	1·5568	1·11	1·6149	1·20
27	1·7625	−0·06	1·7422	0·66	1·7918	0·74	1·7487	0·86
39	1·7625	−1·08	1·7856	−0·04	1·8253	−0·16	1·7499	0·19
51	1·7625	−2·46	1·7856	−1·09	1·8253	−1·23	1·7583	−0·63
63	1·7625	−3·50	1·7856	−2·65	1·8253	−2·13	1·7615	−1·65
75	1·7625	−5·03	1·7856	−3·52	1·8253	−3·03	1·7618	−2·79
87	1·7625	−6·38	1·7856	−5·05	1·8253	−3·99	1·7618	−3·97
99	1·7625	−7·75	1·7856	−6·15	1·8253	−5·13	1·7623	−5·24
111			1·7856	−7·37	1·8253	−6·23	1·7623	−6·52
123					1·8253	−7·03	1·7623	−7·83

where T is the tridiagonal matrix of order 40,

$$\begin{bmatrix} -4 & 1 & & & \\ 1 & -4 & 1 & & \\ & 1 & -4 & 1 & \\ & & \cdot & \cdot & \cdot \\ & & & 1 & -4 \end{bmatrix}$$

and I is the unit matrix of order 40. The results of this test are summarized in **Table 2**.

For the second problem we consider Kulsrud's example of the solution of Laplace's equation in cylindrical coordinates for the non-rectangular axially-symmetric region shown in **Fig. 1**. We have modified the grid slightly in order to obtain a symmetric matrix, and use the finite-difference approximation

$$rh^2\nabla^2\phi \simeq (r + \tfrac{1}{2}h)\phi(r + h, z) + (r - \tfrac{1}{2}h)\phi(r - h, z) + r\phi(r, z + h) + r\phi(r, z - h) - 4r\phi(r, z).$$

The results are summarized in **Table 3**.

As our third example we took the very ill-conditioned matrix considered by Engeli *et al.* (1959), page 100. This is of order 44 and of the block form

$$\begin{bmatrix} I & A & & & & & & \\ A^T & I & B & & & & & \\ & B^T & I & B & & & & \\ & & B^T & I & B & & & \\ & & & \cdot & \cdot & \cdot & & \\ & & & & B^T & I & B & \\ & & & & & B^T & I & C \\ & & & & & & C^T & I \end{bmatrix}$$

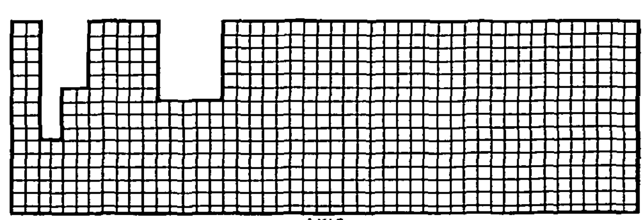

Fig. 1 The finite-difference grid in Kulsrud's example

Table 4

Results for Engeli's example

ITERATION NUMBER	ω_{opt}		CARRÉ		KULSRUD		NEW METHOD	
	ω	$\log_{10}\|\|e\|\|$	ω	$\log_{10}\|\|e\|\|$	ω	$\log_{10}\|\|e\|\|$	ω	$\log_{10}\|\|e\|\|$
1	1·9910	0·82	1·0000	0·81	1·3750	0·81	1·0000	0·81
3	1·9910	0·82	1·3750	0·81	1·3750	0·81	1·2461	0·81
15	1·9910	0·79	1·6852	0·80	1·4768	0·80	1·6513	0·80
27	1·9910	0·79	1·7664	0·79	1·7724	0·80	1·7889	0·80
39	1·9910	0·78	1·8379	0·79	1·9492	0·79	1·8524	0·79
51	1·9910	0·76	1·8985	0·78	1·9492	0·78	1·9074	0·78
63	1·9910	0·72	1·9185	0·78	1·9500	0·77	1·9118	0·78
75	1·9910	0·70	1·9334	0·77	1·9500	0·77	1·9118	0·77
99	1·9910	0·66	1·9334	0·76	1·9500	0·76	1·9151	0·77
123	1·9910	0·60	1·9334	0·76	1·9500	0·75	1·9518	0·76
147	1·9910	0·54	1·9334	0·75	1·9500	0·74	1·9784	0·75
171	1·9910	0·47	1·9334	0·74	1·9500	0·73	1·9847	0·73
195	1·9910	0·40	1·9334	0·74	1·9500	0·72	1·9876	0·71
219	1·9910	0·32	1·9334	0·73	1·9500	0·71	1·9892	0·67
243	1·9910	0·24	1·9334	0·72	1·9500	0·70	1·9900	0·63
267	1·9910	0·16	1·9334	0·71	1·9500	0·69	1·9903	0·59
315	1·9910	−0·01	1·9334	0·70	1·9500	0·67	1·9903	0·48
363	1·9910	−0·20	1·9334	0·68	1·9500	0·65	1·9903	0·35
411	1·9910	−0·41	1·9334	0·67	1·9500	0·63	1·9903	0·21
459	1·9910	−0·61	1·9334	0·66	1·9500	0·61	1·9903	0·07

where

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, A = \begin{bmatrix} -0\cdot15046\,61372 & 0\cdot05119\,54830 \\ -0\cdot87877\,37287 & -0\cdot09817\,61429 \end{bmatrix},$$

$$B = \begin{bmatrix} -0\cdot10000\,00000 & 0\cdot04721\,35955 \\ -0\cdot84721\,35955 & -0\cdot10000\,00000 \end{bmatrix},$$

and $$C = \begin{bmatrix} -0\cdot09817\,61429 & 0\cdot05119\,54830 \\ -0\cdot87877\,37287 & -0\cdot15046\,61372 \end{bmatrix}.$$

Here we found that v_k was often much larger than the dominant eigenvalue of M_{ω_k} and indeed was sometimes greater than unity. In this situation the parameter ω_{k+1} is likely to be greater than ω_{opt} or even complex. We avoided the latter situation by taking $\omega_{k+1} = \omega_k$ if $v_k \geqslant 1$, but made no attempt to avoid the former. The fact that this does occur for the results presented must be regarded as fortuitous. Presumably the trouble may be avoided if we continue the iteration until the ratios have, in some sense, settled down. Quite apart from the difficulty of devising an automatic criterion for this settling-down, we will have the disadvantage of a large number of iterations with ω less than its optimum, just what we are trying to avoid. For this problem the new method gave very satisfactory results, as shown in **Table 4**.

Conclusion

The advantage of the new method is simply that success can be guaranteed, and this advantage is shown clearly by the third example. For a particular accuracy the new method may require more steps than the earlier methods, as illustrated by examples one and two. We doubt, however, if the extra labour will ever be serious and feel that this is a reasonable price to pay for the additional security.

Appendix

Given a diagonal matrix G_p satisfying (3), with $p \neq 1$, we may construct an ordering vector $q = (q_1, q_2 \ldots q_n)$ as follows. Since any scaling of G_p will not alter the validity of (3), we first normalize G_p to have its first element unity, say $G_p = \text{diag}\,(1, g_2, g_3 \ldots g_n)$, and set $q_1 = 0$. Now suppose some off-diagonal element a_{1i_1}, is non-zero, then from (3) we find $g_{i_1} = p^{1/2}$ and may set $q_{i_1} = 1$. If some off-diagonal element $a_{i_1 i_2}$ is non-zero then $g_{i_2} = p$ if $i_2 > i_1$ and $g_{i_2} = 1$ if $i_2 < i_1$ and we may set $q_{i_2} = 2$ or 0. We continue in this way until we have found all q_i for which i belongs to some subset I_1 of the set N of integers $1, 2 \ldots n$, where I_1 is such that there is no non-zero off-diagonal element a_{ij} with only one of i and j belonging to I_1. If $I_1 \neq N$ then we may scale those g_j for which $j \epsilon (N - I_1)$, without altering the validity of (3), to make some chosen g_k unity and set the corresponding q_k to zero. We now find q_i for all $i \epsilon I_2$, another subset of N. Continuing, we eventually find q_i for $i \epsilon I_1 \cup I_2 \cup \ldots I_r = N$.

Acknowledgements

The author wishes to thank the directors of the Oxford Computing Laboratory and the S.R.C. Atlas Computer Laboratory for providing the necessary computing facilities.

References

CARRÉ, B. A. (1961). "The Determination of the Optimum Accelerating Factor for Successive Over-relaxation," *The Computer Journal*, Vol. 4, p. 73.

ENGELI, M., GINSBURG, TH., RUTISHAUSER, H., and STIEFEL, E. (1959). *Refined Iterative Methods for Computation of the Solution and the Eigenvalues of Self-adjoint Boundary Value Problems*, Birkhäuser, Basle.

KULSRUD, H. E. (1961). "A Practical Technique for the Determination of the Optimum Relaxation Factor of the Successive Over-relaxation Method," *Comm. ACM.*, Vol. 4, p. 184.

VARGA, R. S. (1962). *Matrix Iterative Analysis*, Prentice-Hall, London.

YOUNG, D. (1954). "Iterative Methods for Solving Partial Difference Equations of Elliptic Type," *Trans. Amer. Math. Soc.*, Vol. 76, p. 92.

Book Review

Numerical Solution of Partial Differential Equations, by G. D. Smith, 1965; 179 pages. (London: *Oxford University Press*, 25s.)

This book, intended mainly for students rather than for those already well versed in numerical methods, presents, through simple examples, the principal processes for obtaining numerical solutions to second-order quasi-linear partial differential equations, one chapter each being devoted to equations of Parabolic, Hyperbolic and Elliptic type. In addition there is an introductory chapter which includes the development of finite-difference approximations for derivatives, and one which covers the ideas of convergence, compatibility and stability of finite-difference schemes; and also iterative methods for solving sets of linear algebraic equations.

The author states in his preface that he has tried to make the main chapters independent of one another and admits that this has led to a certain amount of repetition. For example, the Jacobi, Gauss-Seidel and S.O.R. point iterative methods for solving sets of linear algebraic equations appear three times. In Chapter 2 they are applied in detail to a specific example, complete with numerical results; in Chapter 3 they are studied in more general form, and Chapter 5 presents them briefly in connection with Poisson's equation. A good understanding of these methods can be obtained from the sections in Chapters 2 and 3 and surely these would have been better presented together.

In Chapter 2 the main finite-difference methods for solving Parabolic equations are explained and illustrated clearly with detailed numerical calculations. Chapter 4, perhaps the weakest section of the book, presents both the method of characteristics and of finite differences for solving Hyperbolic equations but might have gained something by the inclusion of a section on first-order equations which appear only in the exercises at the end of the chapter. The fifth chapter gives the principal finite-difference methods for Elliptic equations, including a section on relaxation.

Each of the four main chapters includes a very valuable set of exercises with solutions outlined in most cases, and the volume concludes with a list of references for further reading.

Most students should find that this book gives them a good introduction to the subject but they may not be able to understand some of the more advanced concepts, several of which are not explained or illustrated as carefully as many of the simpler ideas. As examples we might cite parts of the section on characteristics of hyperbolic equations, the concept of consistent ordering for sets of algebraic equations, and the method of deferred correction which is dismissed in less than a page. There is, however, sufficient of value to recommend this as a student textbook, and it should also find its way on to the book-shelves of most teachers of the subject.

H. C. JOHNSTON