JOURNAL OF COMPUTATIONAL PHYSICS 48, 127–149 (1981)

On Local Relaxation Methods and Their Application to Convection-Diffusion Equations

EUGEN F. F. BOTTA

Mathematical Institute, University of Groningen, 9700 AV Groningen, The Netherlands

AND

ARTHUR E. P. VELDMAN

National Aerospace Laboratory NLR, 1006 BM Amsterdam, The Netherlands

Received November 5, 1981

This paper discusses local relaxation (LR) methods which can be regarded as generalizations of the successive overrelaxation (SOR) method. The difference is that within an LR method the relaxation factor is allowed to vary from equation to equation. A number of existing methods are found to be in fact special LR methods. Moreover, based on SOR theory, a new LR method is developed. The performance of LR methods is illustrated by applying them to central difference approximations of convection–diffusion equations. It is found that equations with small diffusion coefficients can be handled without difficulty. For equations with strongly varying coefficients, and for nonlinear equations, a properly selected LR method can be significantly more efficient than the optimum SOR method. As a special example, a 16×16 driven cavity problem for a Reynolds number of 10^6 can be solved in just a few seconds on a modern computer.

1. INTRODUCTION

As is well known, iterative schemes for solving second-order central difference approximations of convection–diffusion equations exhibit convergence difficulties when the diffusion coefficient, i.e., the inverse of the Reynolds number (Re) or the Péclet number, becomes small. The extent of the difficulties is related to the cell Reynolds number Re h (h being the mesh size). A typical example of these difficulties can be found in a study by Burggraf [1] of the driven cavity problem. In spite of using underrelaxation, Burggraf was not able to obtain a converged solution for Reynolds numbers larger than about 1000. From studies by Tuann and Olson [2] and Khosla and Rubin [3] it is recognized that Burggraf's problems are partly due to the use of the convective formulation for the vorticity transport equation (which enhances the occurrence of nonlinear instabilities); however, using the divergence

0021-9991/82/100127-23$02.00/0

581/48/1-9

formulation, the solution of the flow equations still requires considerable effort when the Reynolds number is large [4].

A way to avoid these difficulties is the use of upwind differencing of the convective terms. Early references to the application of this idea in meteorological problems can be found in Forsythe and Wasow [5]. In fluid flow problems Greenspan [6] and Gosman, *et al.* [7] have been early advocators. Upwind differencing leads to a diagonally dominant matrix, hence the discrete equations can be solved by standard techniques such as Gauss–Seidel and SOR. A disadvantage of upwind differencing, in general, is a loss of accuracy of the discrete solution as compared with the solution of the central difference approximation. Therefore, when available, the latter solution generally is preferred [2, 4, 8–13]. An exception has to be made for situations in which boundary layers or shock layers are present whose thickness is comparable to or smaller than the mesh size; in these cases the solution of the central difference approximation often exhibits oscillations, whereas the upwind difference solution is smoother [3, 14]. Even in this case, however, Gresho and Lee [15] strongly warn against a false sense of security, as regards to accuracy, which can emanate from the smooth upwind results.

In order to increase the accuracy without losing the convergence of the Gauss–Seidel method, discretization schemes have been developed which approach the central scheme when $\mathrm{Re}\, h \to 0$ and which tend to the upwind scheme when $\mathrm{Re}\, h \to \infty$. An example of such a scheme, rediscovered many times, has been used by Allen and Southwell [16] (see also Steele and Barrett [17] and references herein for applications). Other schemes of this type have been designed by Spalding [18] and Raithby and Torrance [9]. Accuracy comparisons performed by Runchal [8] and Raithby and Torrance [9] reveal that for small $\mathrm{Re}\, h$ these schemes are comparable to the central scheme, whereas for large $\mathrm{Re}\, h$ they are as inaccurate as the upwind scheme.

Dennis and Chang [19] have chosen another approach, in which the accuracy of the central difference scheme is retained, but where the convergence of the iterative process is no longer guaranteed theoretically. To the upwind difference term they add a correction term such that, after convergence of the iterative process, the discrete solution satisfies the central difference equations. The correction term is calculated with values from previous iterations. Dennis and Chang keep this term fixed for about 30 iteration sweeps. In more recent applications the correction term has been determined using values from the preceding sweep [12, 20–22], or using the latest available values [13, 23–25]. A slightly different correction term has been used in [10] and [26]; here it is chosen such that the converged solution satisfies the equations discretized with a three-point backward scheme for the convective terms (which is also of second-order accuracy).

The way in which the correction term is treated can have a large influence on the convergence of the iterative process. This is illustrated for the Gauss–Seidel method in the following one-dimensional example: Consider the equation

$$(d^2u/dx^2) - \mathrm{Re}\, f(x)(du/dx) = 0, \quad 0 \leqslant x \leqslant 1, \tag{1}$$

with $u(0)$ and $u(1)$ prescribed, on a grid $x_i = ih$, $i = 0, 1, \ldots, N$, where $h = 1/N$. After upwind differencing with a correction term as used in [12, 20–22], the discrete equation at the point x_i for the $(n+1)$th sweep can be written as

$$(1 - r_i + |r_i|)\, u_{i+1}^n - 2(1 + |r_i|)\, u_i^{n+1} + (1 + r_i + |r_i|)\, u_{i-1}^{n+1} = |r_i|\,(u_{i+1}^n - 2u_i^n + u_{i-1}^n), \tag{2}$$

where $r_i = \frac{1}{2}\,\mathrm{Re}\, hf(x_i)$. When the correction term is calculated with the most recent values, as in [13, 23–25], the index n of the term u_{i-1}^n in the right-hand side of (2) changes into $n+1$, whereas the other terms remain unchanged. Hence this scheme becomes

$$(1 - r_i + |r_i|)\, u_{i+1}^n - 2(1 + |r_i|)\, u_i^{n+1} + (1 + r_i + |r_i|)\, u_{i-1}^{n+1} = |r_i|\,(u_{i+1}^n - 2u_i^n + u_{i-1}^{n+1}). \tag{3}$$

The influence of the small difference between schemes (2) and (3) can be seen in Table I. Here the analytically determined spectral radius of the Gauss–Seidel matrix has been tabulated for a case where $f(x) \equiv 1$, $N = 20$, and for various values of Re.

It is remarked that (for the case of constant f) changing the sign of Re is equivalent to reversing the sweep direction of the Gauss–Seidel process. Hence we see that the convergence of scheme (2) can depend on the sweep direction. On the other hand, for scheme (3), which only differs from scheme (2) in the treatment of the correction term, the convergence is independent of the sweep direction. Due to this difference in behaviour, scheme (2) will not be discussed in this paper.

A closer inspection of scheme (3), which can be rewritten as

$$(1 - r_i)\, u_{i+1}^n - 2(1 + |r_i|)\, u_i^{n+1} + (1 + r_i)\, u_{i-1}^{n+1} + 2\,|r_i|\, u_i^n = 0, \tag{3'}$$

reveals that it can be regarded as an SOR method for the central difference approximation to (1), in which the relaxation factor $\omega_i = (1 + |r_i|)^{-1}$ can be different in each grid point. The favourable experience with this scheme reported by Veldman [23] and Dijkstra [24] led us to investigate generalizations of the SOR method. In

TABLE I

Spectral Radii Related to Schemes (2) and (3) for $f \equiv 1$ and $h = \frac{1}{20}$

	Re					
	10^3	10^2	10	-10	-10^2	-10^3
Scheme (2)	0.96	0.71	0.93	0.93	1.10	1.36
Scheme (3)	0.96	0.71	0.94	0.94	0.71	0.96

each grid point the relaxation factor will be determined from the coefficients of the corresponding discrete equation and therefore this factor no longer needs to be constant throughout the mesh. The methods thus obtained will be called *local relaxation* (LR) *methods.*

The underlying idea is not new; it was already in use by Russell [27] in the early sixties, but apart from a few applications, e.g., Apelt [28], it has not attained widespread recognition. Only recently does there seem to be a revival of the idea of spatially varying relaxation factors, as may be inferred from papers by Benjamin and Denny [4], Takemitsu [29], and Strikwerda [30].

For equations with constant coefficients the LR strategy leads to relaxation factors which are the same in each grid point; hence in this case an LR method is equivalent to an SOR method. By this correspondence, SOR theory can be used to study the behaviour of LR methods. Furthermore, this relation suggests a special choice for the local relaxation factor ω_i: it seems reasonable to select ω_i as the optimum relaxation factor ω_{opt} of the SOR method applied to a system in which all equations have the same coefficients as the ith equation. It will appear, in Section 2, that ω_{opt} is a complicated function of the coefficients which can be rather expensive to evaluate. Therefore, in Section 3, we introduce approximations of ω_{opt}, chosen such that the rate of convergence is not much affected, but which can be computed more economically.

Section 4 is devoted to the study of the performance of LR methods. It turns out that for equations with spatially varying coefficients and for nonlinear equations, a properly selected LR method is more effective than the optimum SOR method; the difference can be several orders of magnitude. In particular, LR methods are very effective when solving central difference approximations of convection–diffusion equations, even at very large cell Reynolds numbers, as is demonstrated by a driven cavity problem in Section 5.

2. Analysis

In this section we will first give some theory on the determination of the optimum relaxation factor ω_{opt} of the SOR method for solving a system of real linear equations $Ax = b$, in which the matrix A can be written as $A = D(I - L - U)$, where L is a strictly lower triangular matrix, U is a strictly upper triangular matrix, and D is a nonsingular diagonal matrix. The Jacobi matrix B and the SOR matrix L_ω, corresponding with relaxation factor ω, can then be written as

$$B = L + U, \qquad L_\omega = (I - \omega L)^{-1}\,[(1 - \omega)I + \omega U].$$

When the matrix A is consistently ordered, the following fundamental relation exists between the eigenvalues μ of B and the eigenvalues λ of L_ω ($\lambda \neq 0$, $\omega \neq 0$) [31]

$$(\lambda + \omega - 1)^2 = \omega^2 \mu^2 \lambda. \tag{4}$$

When μ is an eigenvalue of B, so are $-\mu$ and $\pm\bar{\mu}$. Thus we will consider the frequently occurring case where the eigenvalues of B are known and lying in a rectangle of which the vertices are the eigenvalues $\pm\mu_R \pm i\mu_I$ with $\mu_R \geqslant 0$ and $\mu_I \geqslant 0$. Once ω is given, (4) can be used to determine the eigenvalues λ of L_ω. Then also the spectral radius $\rho(\omega)$ of L_ω follows as the maximum of the moduli of the eigenvalues of L_ω. When $\rho(\omega) < 1$ the SOR method is called convergent. The value of ω for which $\rho(\omega)$ attains its minimum is called the optimum relaxation factor ω_{opt}. For given μ and ω, (4) yields two eigenvalues of L_ω, the product of whose moduli equals $(\omega - 1)^2$. Hence $\rho(\omega) \geqslant |\omega - 1|$, and since we are only interested in $\rho(\omega) < 1$, we restrict ourselves to $0 < \omega < 2$.

By rewriting (4) (the central symmetry allows us to consider only one sign of the square root) as

$$\mu = \omega^{-1}[\lambda^{1/2} + (\omega - 1)\lambda^{-1/2}], \qquad \lambda \neq 0, \tag{5}$$

we obtain a conformal mapping from the complex $\lambda^{1/2}$-plane to the complex μ-plane. The circle $|\lambda^{1/2}| = r$ is mapped onto the ellipse $E_{\omega,r}$

$$\frac{[\mathrm{Re}\,\mu]^2}{[(r + (\omega - 1)/r)/\omega]^2} + \frac{[\mathrm{Im}\,\mu]^2}{[(r - (\omega - 1)/r)/\omega]^2} = 1. \tag{6}$$

Furthermore, when $r^2 \geqslant |\omega - 1|$ the exterior of the circle is mapped onto the exterior of the ellipse. Hence when the exterior of ellipse (6) does not contain an eigenvalue of B, then $\rho(\omega) \leqslant r^2$. The equality sign holds if and only if at least one of the eigenvalues of B lies on the ellipse [32]. It follows that L_ω is convergent if and only if all eigenvalues μ lie in the interior of the ellipse $E_{\omega,1}$

$$[\mathrm{Re}\,\mu]^2 + [\mathrm{Im}\,\mu]^2/[(2 - \omega)/\omega]^2 = 1. \tag{7}$$

A necessary condition for convergence is therefore $\mu_R < 1$, whereas in our case ω has to be chosen such that $\mu = \mu_R + i\mu_I$ lies inside ellipse (7). Hence the SOR method is convergent if and only if $0 < \omega < \omega_{max}$, where

$$\omega_{max} = 2/[1 + \mu_I(1 - \mu_R^2)^{-1/2}]. \tag{8}$$

Further, when r is chosen such that $\mu = \mu_R + i\mu_I$ lies on the ellipse $E_{\omega,r}$, the spectral radius of L_ω can be found from $\rho(\omega) = r^2$.

By minimizing $\rho(\omega)$ the optimum relaxation factor can be obtained. This requires some tedious algebra [33], unless μ_R or μ_I equals zero. We will only present the results here. Abbreviating

$$a = \mu_R^2 + \mu_I^2, \qquad b = \mu_R^2 - \mu_I^2, \qquad c = a^2 - b^2, \qquad d = a^2 - b, \qquad e = (c + d^2)^{1/2},$$

we can write

$$\rho(\omega) = \tfrac{1}{4}(\alpha + [\alpha^2 - 16(1 - \omega)^2]^{1/2}), \tag{9}$$

where

$$\alpha = \omega^2 a + [\omega^4 a^2 + 8\omega^2(1-\omega)b + 16(1-\omega)^2]^{1/2}.$$

The optimum relaxation factor is given by

$$\begin{aligned} \omega_{\text{opt}} &= -\tfrac{1}{2}[\beta - (\beta^2 + 4\beta)^{1/2}], && \text{if} \quad d > 0, \\ &= 1, && \text{if} \quad d = 0, \\ &= -\tfrac{1}{2}[\beta + (\beta^2 + 4\beta)^{1/2}], && \text{if} \quad d < 0, \end{aligned} \tag{10}$$

where

$$\beta = [(3d+e)(e-d)^{1/3}\, c^{1/3} - (3d-e)(e+d)^{1/3}\, c^{1/3} + c - 4bd]/(a^2 d).$$

The case $d > 0$ corresponds to $\omega_{\text{opt}} < 1$, whereas $d < 0$ corresponds to $\omega_{\text{opt}} > 1$. The case $d = 0$, where $\omega_{\text{opt}} = 1$, occurs when the determining eigenvalue μ lies on a Bernoulli lemniscate [34].

The formulas of Eq. (10) are rather uneconomical to use due to the appearance of the fractional powers. Therefore we will look for simple approximations. This requires knowledge of the behaviour of $\rho(\omega)$ in the vicinity of ω_{opt}. An impression can be obtained from Fig. 1, where $\rho(\omega)$ has been plotted for some values of μ_{R} and μ_{I}. Using this figure the following observations can be made:

Case 1. $\mu_{\text{R}} = 0$. The optimum relaxation factor becomes

$$\omega_{\text{opt}} = 2/[1 + (1 + \mu_{\text{I}}^2)^{1/2}] \leqslant 1, \tag{11}$$

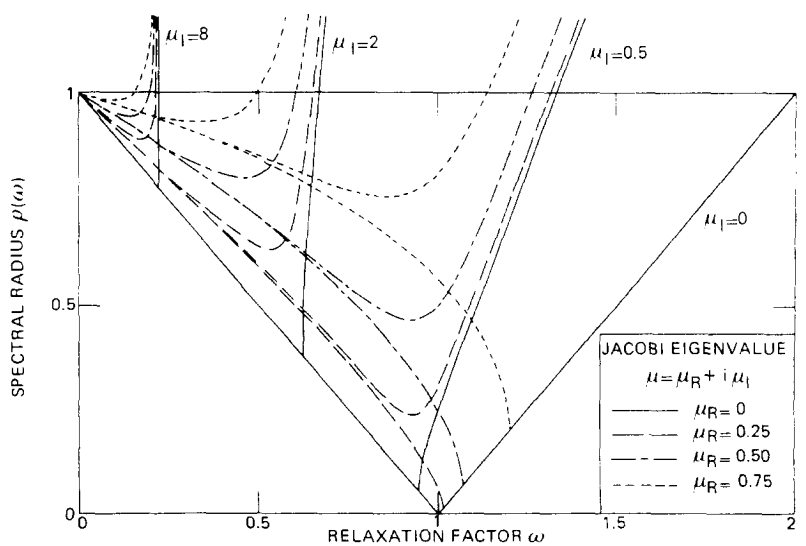

FIG. 1. Relation between spectral radius, relaxation factor, and Jacobi spectrum.

hence underrelaxation has to be applied. When $\mu_I \gg 1$ it can be seen in Fig. 1 that ω_{max} is only slightly larger than ω_{opt} (compare also (8) and (11)), so a very small overestimate of ω_{opt} can lead to divergence. Hence it is essential that any approximation of ω_{opt} be an underestimate. Note that (9) reduces to $\rho(\omega) = 1 - \omega$ when $\omega \leqslant \omega_{opt}$.

Case 2. $\mu_I = 0$. This is the classical case of overrelaxation [31] with

$$\omega_{opt} = 2/[1 + (1 - \mu_R^2)^{1/2}] \geqslant 1. \tag{12}$$

Now it is better to overestimate ω_{opt} than to underestimate, but the choice is much less critical than in Case 1. Here $\rho(\omega) = \omega - 1$ when $\omega \geqslant \omega_{opt}$.

Case 3. $\mu_R \neq 0$, $\mu_I \neq 0$. When μ_R is not too small the choice of an approximation for ω_{opt} is not very critical. For μ_R small and $\mu_I \gg 1$, however we must underestimate as in Case 1. Now, for fixed μ_I, ω_{opt} decreases with increasing μ_R.

From (10) it is difficult to extract analytical information; therefore we shall first present a simple but very good approximation $\tilde{\omega}_{opt}$ of ω_{opt}. It can be derived that $\omega_{opt} \sim 2\mu_I^{-1}(1 - \mu_R^{2/3})^{1/2}$ when $\mu_I \gg 1$. By combining this with (11) and (12) we are led to consider

$$\tilde{\omega}_{opt} = 2/\{1 + [1 - \mu_R^2 + \mu_I^2(1 - \mu_R^{2/3})^{-1}]^{1/2}\}. \tag{13}$$

TABLE II

Exact and Approximate Values of the Optimum Relaxation Factor and Corresponding Spectral Radius

μ_R	μ_I	ω_{opt}	$\tilde{\omega}_{opt}$	ω_{max}	$\rho(\omega_{opt})$	$\rho(\tilde{\omega}_{opt})$
0	0	1	1	2	0	0
0.25	0	1.016133	1.016133	2	0.016133	0.016133
0.50	0	1.071797	1.071797	2	0.071797	0.071797
0.75	0	1.203777	1.203777	2	0.203777	0.203777
0	0.5	0.944272	0.944272	1.333333	0.055728	0.055728
0.25	0.5	0.928228	0.924748	1.318915	0.237603	0.237759
0.50	0.5	0.923371	0.911583	1.267949	0.463002	0.463703
0.75	0.5	0.854061	0.844778	1.138998	0.757638	0.757788
0	2	0.618034	0.618034	0.666667	0.381966	0.381966
0.25	2	0.533561	0.533156	0.652403	0.633531	0.633533
0.50	2	0.455602	0.454551	0.604339	0.800693	0.800697
0.75	2	0.343309	0.342878	0.497053	0.929915	0.929915
0	8	0.220696	0.220696	0.222222	0.779304	0.779304
0.25	8	0.176279	0.176268	0.215928	0.889668	0.889668
0.50	8	0.141073	0.141047	0.195358	0.945322	0.945322
0.75	8	0.099208	0.099199	0.152732	0.981942	0.981942

We note that $\tilde{\omega}_{\text{opt}}$ equals ω_{opt} when $\mu_R = 0$ or $\mu_I = 0$. Furthermore, for other values of μ_R and μ_I, $\tilde{\omega}_{\text{opt}}$ and $\rho(\tilde{\omega}_{\text{opt}})$ are very good approximations of ω_{opt} and $\rho(\omega_{\text{opt}})$. This can be inferred from Table II where, for some values of μ_R and μ_I, we have given the values of ω_{opt}, $\tilde{\omega}_{\text{opt}}$, $\omega_{\max}$, $\rho(\omega_{\text{opt}})$, and $\rho(\tilde{\omega}_{\text{opt}})$.

3. Local Relaxation

With the basic formulas from the preceeding section the optimum relaxation factor for the SOR method can be calculated for any matrix satisfying $\mu_R < 1$ (a necessary and sufficient condition for convergence of the optimum SOR method). When the eigenvalues of the Jacobi matrix are irregularly distributed, however, it is unknown on which eigenvalue the optimum relaxation factor should be based. Recently, Rigal [33] has proposed an algorithm to determine the optimum for an arbitrary Jacobi spectrum. This requires full knowledge of the spectrum which, in general, is hard to obtain.

The above drawback of the SOR strategy can be circumvented by switching to the LR strategy. Unlike SOR, in which the (uniform) relaxation parameter depends on the total discrete system, an LR method bases the (nonuniform) relaxation factor for a given equation on this equation only. The present method, for instance, bases the relaxation factor ω_i for the ith equation on a Jacobi matrix in which all coefficients are the same as in the ith equation. The eigenvalues of such a constant-coefficient matrix can be calculated analytically, which allows us to express ω_i explicitly in the coefficients of the ith equation.

By its construction, for equations with constant coefficients the LR method just proposed is equivalent to the optimum SOR method. In the case of a linear equation its relaxation factors, determined by (10) (alternatively (13) may be used), have to be evaluated once in each grid point, but for nonlinear equations this has to be repeated each iteration sweep. Consequently we shall look for less complicated approximations of ω_{opt}. How these can be obtained is demonstrated next in the important example of a second-order convection–diffusion equation. The extension to more general equations is discussed in Section 6.

Consider on a domain $\Omega = \{(x, y) \mid 0 \leqslant x \leqslant l_1, 0 \leqslant y \leqslant l_2\}$ the differential equation

$$\Delta u - f(x, y)\frac{\partial u}{\partial x} - g(x, y)\frac{\partial u}{\partial y} = 0, \tag{14}$$

where u is prescribed on the boundary $\partial\Omega$. The domain Ω is covered with a grid $(x_i, y_j) = (ih, jk)$; $i = 0, 1, \ldots, N(h = l_1/N)$; $j = 0, 1, \ldots, M(k = l_2/M)$. The equation is discretized using second-order central differences, which yields for the grid point (x_i, y_j) the following discrete equation

$$C_W u_{i-1,j} + C_S u_{i,j-1} - u_{i,j} + C_E u_{i+1,j} + C_N u_{i,j+1} = 0, \tag{15a}$$

where

$$C_W = \tfrac{1}{2}\alpha(1+a), \qquad C_E = \tfrac{1}{2}\alpha(1-a), \qquad C_S = \tfrac{1}{2}\beta(1+b), \qquad C_N = \tfrac{1}{2}\beta(1-b) \quad (15b)$$

with

$$\alpha = k^2/(h^2+k^2), \quad \beta = h^2/(h^2+k^2), \quad a = \tfrac{1}{2}hf(x_i, y_j), \quad b = \tfrac{1}{2}kg(x_i, y_j). \quad (15c)$$

The eigenvalues of the Jacobi matrix formed from (15a) using the LR strategy can be found readily

$$\mu = 2(C_E C_W)^{1/2} \cos(p\pi/N) + 2(C_N C_S)^{1/2} \cos(q\pi/M), \quad (16a)$$
$$(p = 1, 2, \ldots, N-1;\ q = 1, 2, \ldots, M-1).$$

Hence, using (15b),

$$\mu_R + i\mu_I = \alpha(1-a^2)^{1/2} \cos(\pi/N) + \beta(1-b^2)^{1/2} \cos(\pi/M). \quad (16b)$$

Note that the condition $\mu_R < 1$, necessary and sufficient for convergence of the optimum SOR method, is satisfied for equations of type (15) with constant coefficients.

Now we shall derive approximations ω_{ij}^* for the optimum relaxation factor ω_{opt} as given by Eq. (10). The same three cases as in Section 2 are considered.

Case 1. $\mu_R = 0$. This case applies when $C_E C_W \leqslant 0$ and $C_N C_S \leqslant 0$, or equivalently, $a^2 \geqslant 1$ and $b^2 \geqslant 1$. The optimum relaxation factor is given by Eq. (11). Using (16) we estimate

$$\begin{aligned} 1 + \mu_I^2 &< 1 + \alpha^2(a^2-1) + \beta^2(b^2-1) + 2\alpha\beta(a^2-1)^{1/2}(b^2-1)^{1/2} \\ &= 2\alpha\beta + \alpha^2 a^2 + \beta^2 b^2 + 2\alpha\beta(a^2-1)^{1/2}(b^2-1)^{1/2} \\ &\leqslant 2\alpha\beta + \alpha^2 a^2 + \beta^2 b^2 + 2\alpha\beta(|ab|-1) \\ &= (\alpha|a| + \beta|b|)^2, \end{aligned}$$

hence ω_{opt} is underestimated by

$$2/(1 + \alpha|a| + \beta|b|), \quad (17)$$

which, moreover, is a very good approximation of ω_{opt} when $\alpha|a|$ or $\beta|b|$ is large.

Case 2. $\mu_I = 0$. This case applies when $C_E C_W \geqslant 0$ and $C_N C_S \geqslant 0$, or equivalently, $a^2 \leqslant 1$ and $b^2 \leqslant 1$. The optimum relaxation factor is now given by Eq. (12). Similarly to Case 1, it can be argued that (17) (slightly) overestimates ω_{opt}. For $a = b = 0$, however, (17) takes the value 2, and the iterative process is no longer convergent. Therefore we shall bound this approximation from above by the optimum relaxation

factor ω_0 corresponding to $a = b = 0$ (ω_0 can be found by substitution of (16b) into (12)). Thus we choose

$$\omega_{ij}^* = \min\{\omega_0, 2/(1 + \alpha|a| + \beta|b|)\}. \tag{18}$$

It is remarked that in Case 1 the value of ω_{ij}^* given by (18) coincides with the value from (17); hence (18) can be used in both cases, i.e., when $C_E C_W C_N C_S \geqslant 0$.

Case 3. $\mu_R \neq 0$, $\mu_I \neq 0$. The remaining case can be split into a case with $C_W C_E > 0$ and $C_N C_S < 0$, and a case with $C_W C_E < 0$ and $C_N C_S > 0$. Only the first case will be treated in detail. Therefore let

$$\mu_R = \alpha(1 - a^2)^{1/2} \cos(\pi/N), \qquad \mu_I = \beta(b^2 - 1)^{1/2} \cos(\pi/M). \tag{19}$$

A good approximation of ω_{opt} is given by $\tilde{\omega}_{opt}$ in (13). This equation can be used to show that, for fixed $\mu_I > \frac{1}{4}\sqrt{3}$, $\tilde{\omega}_{opt}$ is a decreasing function of μ_R. The choice of ω_{ij}^* is most critical when μ_I is large. In fact, it is better to underestimate ω_{opt} by 50% or more, than to overestimate it by only a few percent. The approximation for ω_{opt} will therefore be based on the maximum value of μ_R, i.e., $\mu_R = \alpha$, in which case ω_{opt} can be approximated very well by

$$\omega_{ij}^* = 2/(1 + \gamma_1 \beta|b|), \qquad \text{with} \quad \gamma_1 = (1 - \alpha^{2/3})^{-1/2}.$$

This approximation can also be used for smaller values of $|b|$ since the situation is not critical then.

When applied to a system of equations with constant coefficients, for which the SOR theory is valid, the above choices for ω_{ij}^* lead to a convergent LR method. In Cases 1 and 2 the convergence follows straightforwardly. In Case 3 convergence can be inferred from the following estimate:

$$\begin{aligned}\omega_{ij}^* &= 2/(1 + (1 - \alpha^{2/3})^{-1/2} \beta|b|) < 2/(1 + (1 - \alpha^2)^{-1/2} \beta|b|) \\ &< 2/(1 + [1 - \alpha^2(1 - a^2)]^{-1/2} \beta(b^2 - 1)^{1/2}) < \omega_{max}.\end{aligned}$$

In the last step (8) and (19) have been used.

Summarizing, to solve equations of type (15) we propose the following choice of the local relaxation factor (for which, when applied to equations with constant coefficients, convergence has been proved): when

$$C_E C_W C_N C_S \geqslant 0: \qquad \omega^* = \min\left\{\omega_0, \frac{2}{1 + |C_E - C_W| + |C_N - C_S|}\right\}, \tag{20a}$$

when

$$C_E C_W C_N C_S < 0: \qquad \omega^* = \frac{2}{1 + \gamma_1 |C_N - C_S|}, \qquad \text{if} \quad C_W C_E > 0, \tag{20b}$$

$$= \frac{2}{1 + \gamma_2 |C_E - C_W|}, \qquad \text{if} \quad C_W C_E < 0, \tag{20c}$$

where $\gamma_1 = [1 - (C_E + C_W)^{2/3}]^{-1/2}$, $\gamma_2 = [1 - (C_N + C_S)^{2/3}]^{-1/2}$. It is noted that for equations of type (14), γ_1 and γ_2 depend only on the mesh sizes h and k, and not on the coefficients of the first-order derivatives; in the special case of equal mesh sizes we have $\gamma_1 = \gamma_2 = 1.644$.

Remark. A one-dimensional local relaxation choice can be obtained from (20) by putting $C_N = C_S = 0$ (hence only (20a) applies).

4. Performance of Local Relaxation Methods

In this section we shall discuss the performance of the LR method defined in (20), and of some other methods—reported in the literature—which belong to the class of LR methods. Also a comparison with the optimum SOR method is made. The performance of the relaxation methods is tested by solving discrete equations of type (15), repeated here

$$C_W u_{i-1,j} + C_S u_{i,j-1} - u_{i,j} + C_E u_{i+1,j} + C_N u_{i,j+1} = 0, \tag{15a}$$

in which the coefficients are characterized by

$$C_E + C_W + C_N + C_S = 1, \tag{21a}$$

$$C_E + C_W \geqslant 0, \qquad C_N + C_S \geqslant 0. \tag{21b}$$

Note that all estimates in Section 3 remain valid for this type of equation.

The following LR methods are considered:

(1) A method, apparently first described by Veldman [23] and Dijkstra [24], but later rediscovered in [13] and [25]. The relaxation factor is chosen as

$$\omega_{VD} = 1/[1 + |C_E - C_W| + |C_N - C_S|]. \tag{22}$$

(2) A related method used by Takemitsu [29]:

$$\omega_T = 2/[2 + |C_E - C_W| + |C_N - C_S|]. \tag{23}$$

(3) The method suggested about two decades ago by Russell [27], who, however, restricted himself to situations with $C_E + C_W = C_N + C_S = \frac{1}{2}$, i.e., $h = k$ in (15):

$$\omega_R = 2/[1 + (2|C_E - C_W|^2 + 2|C_N - C_S|^2 + K)^{1/2}]. \tag{24}$$

$K = \frac{1}{2}\pi^2(N^{-2} + M^{-2})$ plays the same role as ω_0 in (20a): it guarantees optimum convergence when first-order terms are absent.

(4) A method, similar to the previous one but covering the case $h \neq k$, presented by Strikwerda [30]:

$$\omega_S = 2 \Big/ \left[1 + \left\{ \frac{(C_E - C_W)^2}{C_E + C_W} + \frac{(C_N - C_S)^2}{C_N + C_S} \right\}^{1/2} \right]. \tag{25}$$

(5) The method defined in Eq. (20) of this paper.

4.1 *Equations with Constant Coefficients*

A theoretical discussion of the performance of the above LR methods can be given when the coefficients in (15a) are independent of the grid point (this occurs, e.g., when f and g in (14) are constant). Once again the three cases are considered.

Case 1. $\mu_R = 0$ ($C_E C_W \leqslant 0$, $C_N C_S \leqslant 0$). It is not difficult to show that in this case $\omega_{VD} \leqslant \omega_T \leqslant \omega^*$, $\omega_R \leqslant \omega^*$, and $\omega_S \leqslant \omega^*$. Since from its construction $\omega^* < \omega_{opt} < \omega_{max}$ it follows that all methods are convergent, and that the present method has the smallest spectral radius, i.e., the fastest convergence.

The method of Strikwerda can sometimes be very inefficient. Such a situation occurs when $C_E + C_W \ll 1$ (or $C_N + C_S \ll 1$), which is tantamount to $h \gg k$ ($k \gg h$) in (15c). Let us consider an example in which $C_E + C_W = \varepsilon \ll \frac{1}{2}$, chosen such that the Jacobi eigenvalues lie within the unit circle (hence Gauss–Seidel converges); for instance

$$C_E = \tfrac{1}{2}(\varepsilon - \tfrac{1}{2}), \qquad C_W = \tfrac{1}{2}(\varepsilon + \tfrac{1}{2}), \qquad C_N = 0, \qquad C_S = 1 - \varepsilon. \tag{26}$$

Setting the cosines in (16) equal to unity, we have $\mu_R = 0$, $\mu_I = \frac{1}{2} + O(\varepsilon^2)$. Table II gives the optimum relaxation factor for this case as $\omega_{opt} = 0.944 + O(\varepsilon^2)$, corresponding to the optimum spectral radius $\rho(\omega_{opt}) = 0.056 + O(\varepsilon^2)$. The Strikwerda method for this case, however, converges arbitrarily slowly, when $\varepsilon \to 0$ since $\omega_S \sim 4\varepsilon^{1/2}$, corresponding to $\rho(\omega_S) \sim 1 - 4\varepsilon^{1/2}$. For comparison, the other methods which are applicable give $\omega_{VD} \sim \frac{2}{5}$, $\omega_T \sim \frac{4}{7}$ and $\omega^* \sim \frac{4}{5}$, leading to $\rho(\omega_{VD}) \sim \frac{3}{5}$, $\rho(\omega_T) \sim \frac{3}{7}$ and $\rho(\omega^*) \sim \frac{1}{5}$.

Case 2. $\mu_I = 0$ ($C_E C_W \geqslant 0$, $C_N C_S \geqslant 0$). As we have seen in Section 2, any choice with $0 < \omega < \omega_{max} = 2$ leads to convergence. All methods satisfy this relation, and hence are convergent, except the method of Strikwerda in case $C_E = C_W$, $C_N = C_S$ (which occurs when in (14) the first-order terms are absent). When $C_E \approx C_W$ and $C_N \approx C_S$, the optimum relaxation factor is close to 2. Therefore, since ω_{VD} and ω_T cannot exceed 1, the methods of Veldman–Dijkstra and Takemitsu are less efficient when the coefficients of the first-order terms are small.

Case 3. $\mu_R \neq 0$, $\mu_I \neq 0$ ($C_E C_W C_N C_S < 0$). For this case the present method has been proved convergent; also the method of Strikwerda can readily be shown to be convergent. Further, when $C_E + C_W = C_N + C_S = \frac{1}{2}$ convergence can be proved for the methods of Veldman–Dijkstra and Russell. The latter methods can be divergent when $C_E + C_W \ll 1$ (or $C_N + C_S \ll 1$). This is apparent from the following example

which is closely related to (26); the difference is that μ_R is chosen close to 1 instead of equal to zero:

$$C_E = \tfrac{1}{2}(\varepsilon - \tfrac{1}{2}), \qquad C_W = \tfrac{1}{2}(\varepsilon + \tfrac{1}{2}), \qquad C_N = C_S = \tfrac{1}{2}(1 - \varepsilon).$$

Replacing the cosines in (16) by unity we have $\mu_R = 1 - \varepsilon$, $\mu_I = \frac{1}{2} + O(\varepsilon^2)$, therefore from (8) it follows that $\omega_{max} \sim 4(2\varepsilon)^{1/2}$. When ε approaches zero, however, $\omega_{VD} \sim 2/3$ and $\omega_R \sim 1.17$. Here it should be remarked that Russell has not intended to apply his method to this type of problems.

The method of Takemitsu also diverges on this example (since $\omega_T > \omega_{VD}$), but additionally his method can be divergent even when $C_E + C_W = C_N + C_S = \frac{1}{2}$. Take, for instance, a problem with $C_N = C_S$ and $|C_E - C_N|$ sufficiently large. The relaxation factor chosen by Takemitsu behaves like $\omega_T \sim 2\,|C_E - C_W|^{-1}$, whereas from (8) we can derive $\omega_{max} \sim \sqrt{3}\,|C_E - C_W|^{-1}$.

4.2 *Equations with Variable Coefficients*

For equations with variable coefficients a theoretical comparison is not yet possible since insufficient theory is available. Therefore we shall compare the various methods by applying them to a number of carefully selected examples which are believed to be representative of the type of equations that can be encountered. We begin with a few one-dimensional cases.

One-dimensional versions of the methods of Veldman–Dijkstra (22), Takemitsu (23), and the present method (20) can be obtained simply by substituting $C_N = C_S = 0$ into their expressions for the relaxation parameter. For the methods of Russell (24) and Strikwerda (25) this is not so straightforward. Following their philosophy, however, we have derived one-dimensional analogues of their formulas, which read

$$\omega_R = 2/[1 + (|C_E - C_W|^2 + K)^{1/2}], \tag{24'}$$

where $K = \pi^2 N^{-2}$, and

$$\omega_S = 2/[1 + |C_E - C_W|], \tag{25'}$$

respectively.

The one-dimensional situation will be treated by solving the following equation for some choices of $f(x)$:

$$d^2u/dx^2 - f(x)\,du/dx = 0, \qquad 0 \leqslant x \leqslant 1, \quad u(0) = 0, \quad u(1) = 0.$$

The equation is discretized, using central differences, on a grid with $h = \frac{1}{20}$ (unless stated otherwise). Starting with $u^0(x) = x(1 - x)$, the discrete equations are iterated according to

$$u_i^{n+1} = (1 - \omega_i)\,u_i^n + \omega_i(C_W u_{i-1}^{n+1} + C_E u_{i+1}^n)$$

until $\max_i |u_i| < 10^{-6}$. In the tables to be presented below, the number of iterations required is indicated.

In the first example $f(x) = \mathrm{Re}\, x^2$ ($\mathrm{Re} = 1$, 10, 10^2, 10^3, and 10^4) has been chosen—a function which has a zero at one of the end points of the interval. Therefore the ratio between the maximum and minimum value of $|f(x)|$ is infinite. Table III shows that now, for large Re, the optimum SOR method is clearly outperformed by any of the LR methods. When the zero is removed, e.g., choosing $f(x) = \frac{1}{2}\,\mathrm{Re}(1 + x^2)$, the situation changes significantly towards the situation with constant coefficients for which optimum SOR is known to be equivalent to the optimum LR method. It is remarked that the optimum SOR results tabulated are the minima we obtained by scanning the ω axis with small steps $\Delta\omega$.

Both examples show for small Re, when Case 2 applies, the inefficiency of the methods of Veldman–Dijkstra and of Takemitsu caused by prohibiting overrelaxation (see Section 4.1). Also visible is a great resemblance between the present results and those of Russell and Strikwerda (especially for large Reynolds numbers). This is easily explained by comparing (20a), (24′), and (25′). For small Reynolds numbers the method of Strikwerda is less efficient because ω_S is chosen too close to 2 (Section 4.1).

TABLE III

One-Dimensional Comparison of Point Iterative Methods

$u_{xx} - fu_x = 0$	Method	Re=1	Re=10	Re=10^2	Re=10^3	Re=10^4
$f(x) = \mathrm{Re}\, x^2$	Optimum SOR	48	53	258	716	1030
	Veldman–Dijkstra	536	740	277	116	561
	Takemitsu	532	695	232	79	455
	Russell	57	93	38	58	331
	Strikwerda	825	80	14	58	331
	Present method	56	77	26	58	331
$f(x) = \frac{1}{2}\mathrm{Re}(1 + x^2)$	Optimum SOR	46	35	15	128	1222
	Veldman–Dijkstra	527	382	39	206	1950
	Takemitsu	519	335	21	104	953
	Russell	54	43	11	97	921
	Strikwerda	369	38	11	97	921
	Present method	52	37	11	97	921
$f(x) = \mathrm{Re}\, u^2$	Optimum SOR	46	46	43	405	5050
	Veldman–Dijkstra	504	504	506	493	div
	Takemitsu	504	504	504	483	455
	Russell	52	52	50	48	41
	Strikwerda	[a]	[a]	[a]	[a]	[a]
	Present method	51	51	48	41	44

[a] Strikwerda's method requires more than a million iterations.

TABLE IV

Comparison of Point Iterative Methods for Decreasing Mesh Size

$u_{xx} - fu_x = 0$	Method	$h = \frac{1}{10}$	$h = \frac{1}{40}$	$h = \frac{1}{160}$
$f(x) = 10^4 x^2$	Optimum SOR	1525	3409	15595
	Veldman–Dijkstra	846	395	744
	Takemitsu	540	352	609
	Russell	433	227	109
	Strikwerda	433	227	109
	Present method	433	227	109

The effect of decreasing the mesh size is shown in Table IV for a case with a large Reynolds number. It is observed that the number of iterations required for SOR increases; a phenomenon not unfamiliar. But the table also shows that the LR methods perform better in this problem. This can be explained from the decrease of μ_I (when h decreases), which leads to a smaller spectral radius, i.e., faster convergence. For reference, notice in Table II the behaviour of $\rho(\omega_{opt})$ for large μ_I and $\mu_R = 0$. The increase in the number of iterations required by the methods of Veldman–Dijkstra and of Takemitsu, when h changes from $\frac{1}{40}$ to $\frac{1}{160}$, can be attributed to the fact that in the latter methods overrelaxation is prohibited. It is remarked that eventually for all methods the number of iterations required will increase with decreasing mesh size.

A further advantage of an LR strategy over the SOR method is apparent when nonlinear equations are solved. The amount of relaxation applied in an LR method can be changed each iteration sweep, and thus adapt itself to the present magnitude of the matrix coefficients. In contrast, in the usual SOR method the relaxation factor has to be tailored to the "worst" situation which is encountered during the iteration process. The difference in efficiency is visible in an example with $f(x) = \mathrm{Re}\, u^2$ (Table III). We note that $f(x)$ approaches zero towards the end of the iteration process, allowing the LR methods to use overrelaxation, whereas at the start of the iterations underrelaxation is required (when Re is large).

In the latter example the method of Strikwerda performs very poorly. This is also due to the fact that $f(x)$ approaches zero, since in such a situation Strikwerda chooses his relaxation factor too close to 2 (see Section 4.1). By comparison with the method of Russell and the present method, the effect of K in (24′) and ω_0 in (20a) is clearly demonstrated.

In the two-dimensional examples the following equation is solved on the domain $\Omega = [0, 1] \times [0, 1]$:

$$\Delta u - f(x, y)\frac{\partial u}{\partial x} - g(x, y)\frac{\partial u}{\partial y} = 0, \qquad u = 0 \quad \text{on} \quad \partial\Omega.$$

Central differences are used in the discretization on a grid with mesh sizes

$h = k = \frac{1}{20}$. The initial guess is chosen as $u^0(x, y) = xy(1 - x)(1 - y)$, after which the iterations

$$u_{i,j}^{n+1} = (1 - \omega_{i,j})\, u_{i,j}^{n} + \omega_{i,j}(C_W u_{i-1,j}^{n+1} + C_E u_{i+1,j}^{n} + C_S u_{i,j-1}^{n+1} + C_N u_{i,j+1}^{n})$$

are performed until $\max_{i,j} |u_{i,j}| < 10^{-6}$.

We begin with an example in which $|C_E - C_W| = |C_N - C_S|$ and where Cases 1 and 2 apply: $f(x, y) = g(x, y) = \mathrm{Re}\, x^2$ ($\mathrm{Re} = 10^n$, $n = 0, 1, \ldots, 4$). From the number of iterations required, given in Table V, it is seen that the behaviour of all methods is very much like the one-dimensional situation.

As a second example we choose $f(x, y) = \frac{1}{2}\,\mathrm{Re}(1 + x^2)$ and $g(x, y) = 100$. An interesting situation occurs when Re is large: Case 1 applies with $|C_E - C_W| \gg |C_N - C_S| > 1$. An analytical indication of the performance of the various methods

TABLE V

Two-Dimensional Comparison of Point Iterative Methods

$\Delta u - f u_x - g u_y = 0$	Method	Re=1	Re=10	Re=10^2	Re=10^3	Re=10^4
$f(x, y) = \mathrm{Re}\, x^2$	Optimum SOR	46	43	310	1056	2053
$g(x, y) = \mathrm{Re}\, x^2$	Veldman–Dijkstra	465	516	264	117	530
	Takemitsu	462	486	221	78	478
$h = k = 1/20$	Russell	51	59	30	60	300
	Strikwerda	761	90	34	60	300
	Present method	50	47	26	60	300
$f(x, y) = \frac{1}{2}\mathrm{Re}(1 + x^2)$	Optimum SOR	27	26	17	94	878
$g(x, y) = 100$	Veldman–Dijkstra	46	47	53	164	1402
	Takemitsu	28	27	25	79	633
$h = k = 1/20$	Russell	24	22	14	91	947
	Strikwerda	24	22	14	91	947
	Present method	25	24	13	67	606
$f(x, y) = \frac{1}{2}\mathrm{Re}(1 + x^2)$	Optimum SOR	8	7	10	52	602
$g(x, y) = 100$	Veldman–Dijkstra	68	69	74	157	981
	Takemitsu	36	36	38	84	494
$h = 1/10$, $k = 1/40$	Strikwerda	9	7	15	174	1870
	Present method	9	8	11	56	464
$f(x, y) = \mathrm{Re}\, x^2$	Optimum SOR	46	41	202	658	1328
$g(x, y) = 0$	Veldman–Dijkstra	463	542	311	113	535
	Takemitsu	461	524	280	180	div
$h = k = 1/20$	Russell	51	66	45	64	355
	Strikwerda	1036	108	38	64	355
	Present method	50	58	36	75	366

can be found from the asymptotic behaviour of the relaxation factors given in Eqs. (20)–(25), viz.,

$$\omega_{VD} \sim |C_E - C_W|^{-1},$$

$$\omega_R \quad \text{and} \quad \omega_S \sim \sqrt{2}\,|C_E - C_W|^{-1}, \qquad \omega_T \quad \text{and} \quad \omega^* \sim 2\,|C_E - C_W|^{-1}.$$

In the case of constant coefficients, where for these relaxation factors $\rho = 1 - \omega$, we expect that the present method and the one of Takemitsu are faster by a factor $\sqrt{2}$ than Strikwerda's method, and about twice as fast as the Veldman–Dijkstra method when Re is large. Table V confirms this behaviour for this example with variable coefficients.

In the latter example equal mesh sizes are used, i.e., $C_E + C_W = \frac{1}{2}$. When $C_E + C_W \ll \frac{1}{2}$, however, Strikwerda's method loses efficiency as discussed in Section 4.1. For instance, when $h = 4k$, i.e., $C_E + C_W = \frac{1}{17}$, the asymptotic behaviour of Strikwerda's relaxation factor $\omega_S \sim 2(C_E + C_W)^{1/2}\,|C_E - C_W|^{-1}$ predicts this method to be the slowest of the LR methods considered (when $|C_E - C_W| \gg 1$). The figures in Table V, where the latter example has been treated with $h = \frac{1}{10}$ and $k = \frac{1}{40}$, are in agreement with this prediction. Russell's method has not been included in this example with unequal mesh sizes because it was not designed to cover this type of problem.

When Case 3 applies with large μ_1 the situation again is changed, as illustrated by an example with $f(x, y) = \mathrm{Re}\, x^2$ and $g(x, y) = 0$ (Table V). Takemitsu's method is seen to diverge for $\mathrm{Re} = 10^4$; the reason has already been discussed in Section 4.1. Applying unequal mesh sizes is not as interesting as in the previous example. The only feature which is worth mentioning is that, as predicted in Section 4.1, the method of Veldman–Dijkstra can become divergent for large values of Re.

More difficult are situations in which one of the coefficients switches sign whereas

$g(x, y) = 0$ (Table VI). For large Re not only does Takemitsu's method diverge, but so do Russell's method, Strikwerda's method, and the present one. The latter methods can be made convergent, however, by restricting the relaxation factor to be less than unity. For the present method this is realized by replacing ω_0 by 1 in (20a). The numbers marked with an asterisk have been obtained this way. As a possible explanation of this behaviour it is observed that when $x = \frac{1}{2}$ (20a) recommends overrelaxation with $\omega = \omega_0$ close to 2, whereas, for large values of Re, in the grid points adjacent to $x = \frac{1}{2}$ underrelaxation is prescribed. It is believed that this large difference between neighbouring ω values is responsible for the divergence, since reducing the difference, by restricting ω to values less than or equal to 1, leads to convergence.

Also interesting are situations in which both coefficients switch sign in the interior—especially those in which f and g have common zeros, i.e., internal turning points. De Groen [35] has given a classification of two-dimensional turning points, together with a discussion of their intrinsic properties. Two cases will be treated here. The first case is chosen such that only interior boundary layers can exist. An example

TABLE VI

Point Iterative Methods Applied to Turning-Point Problems ($h = k = \frac{1}{20}$)

$\Delta u - fu_x - gu_y = 0$	Method	Re=1	Re=10	Re=10^2	Re=10^3	Re=10^4
$f(x,y) = \mathrm{Re}(2x-1)^3$	Optimum SOR	46	53	223	4406	39356
$g(x,y) = 0$	Veldman–Dijkstra	458	556	1015	941	881
	Takemitsu	458	550	964	876	div
	Russell	51	69	169	164	408*
	Strikwerda	4165	370	99	94	408*
	Present method	50	67	141	112	608*
$f(x,y) = \mathrm{Re}(1-2x)$	Optimum SOR	43	37	43	106	1019
$g(x,y) = \mathrm{Re}(1-2y)$	Veldman–Dijkstra	414	230	52	133	1241
	Takemitsu	411	215	40	74	674
	Russell	42	37	24	69*	679*
	Strikwerda	634	63	26	69*	679*
	Present method	43	41	26	70*	666*
$f(x,y) = \mathrm{Re}(2x-1)$	Optimum SOR	44	77	—	div	div
$g(x,y) = \mathrm{Re}(2y-1)$	Veldman–Dijkstra	503	1674	—	div	div
	Takemitsu	500	1572	—	div	div
	Russell	59	249	—	div	div
	Strikwerda	638	76	—	div	div
	Present method	58	215	—	div	div

Note. For numbers marked with an asterisk see text.

is provided by $f(x,y) = \mathrm{Re}(1-2x)$ and $g(x,y) = \mathrm{Re}(1-2y)$. From Table VI it is seen that this problem can be solved; for large values of Re overrelaxation again has to be prohibited. The second case is chosen such that a boundary layer is formed all around the perimeter of the domain: $f(x,y) = \mathrm{Re}(2x-1)$, $g(x,y) = \mathrm{Re}(2y-1)$. De Groen [35] has proved that the continuous problem, in the limit $\mathrm{Re} \to \infty$, possesses an eigenvalue zero, and hence cannot be solved uniquely. The discrete approximations show similar behaviour (Table VI). For $\mathrm{Re} = 100$ the discrete matrix appears to be singular (zero eigenvalue), and for larger values of Re the iterations slowly diverge for all methods tried.

4.3 *Summary of Comparative Test Results*

The above comparative tests show the following properties of the LR methods, when compared with the optimum SOR method:

An explicit choice for the relaxation parameter is available.

For equations with constant coefficients several LR methods are as efficient as the optimum SOR method.

For equations with varying coefficients, and for nonlinear equations, most LR methods considered are more efficient than optimum SOR; the difference can be several orders of magnitude.

Like any point iterative method, an LR method is very simple to programme.

Comparing the LR methods considered with each other we can conclude:

The methods of Veldman–Dijkstra and of Takemitsu are inefficient when the coefficients of the first-order derivatives are small. Moreover, when one of the coefficients is small and the other is large the methods can diverge in some cases.

Russell's method is a very good one when applied to grids with $h = k$ (for which the method was designed originally): it should have gotten much more attention. Strikwerda's related method which covers the case $h \neq k$ can be extremely inefficient. A (small) disadvantage of both methods is that a square root has to be calculated.

The present method is found to converge whenever one of the other methods converges; moreover, it is found to be competitive with the other methods.

After completion of the present investigation, a paper by Ehrlich [37] has appeared which is based on the same philosophy as used in the present paper, i.e., the starting point is the formula for the optimum SOR factor given in (10). To define the local relaxation factor Ehrlich [37] evaluates (10) using (16b) for Dirichlet boundary conditions, or similar formulas valid for Neumann or periodic boundary conditions. In the present paper a simpler approximation of the resulting expression is proposed; for nonlinear problems, where the relaxation factors have to be recalculated each iteration sweep, this can lead to an appreciable decrease in computational effort. Due to the close resemblance, the convergence of Ehrlich's method and of the present method will be about the same.

5. A Driven Cavity Example

We thought it unavoidable to test the performance of the present LR method by means of the driven cavity problem. A review of driven cavity calculations up to 1978 has been given by Tuann and Olson [2]. The maximum Reynolds number for which they reported central difference solutions is 5000. More recently larger Reynolds numbers have been treated: up to $\mathrm{Re} = 50{,}000$ by Kurtz, *et al.* [36]. They used the method of lines on a 16×16 grid. To enable a fair comparison, we solved the same system of discrete equations as they did. Additionally, we increased the Reynolds number to $\mathrm{Re} = 10^6$.

In short, the driven cavity problem asks to solve, on the square $\{(x, y) \mid 0 \leqslant x, y \leqslant 1\}$, the incompressible Navier–Stokes equations, which read in divergence form

$$\frac{\partial}{\partial x}\left(\frac{\partial \psi}{\partial y}\Omega\right) - \frac{\partial}{\partial y}\left(\frac{\partial \psi}{\partial x}\Omega\right) = \frac{1}{\mathrm{Re}}\Delta\Omega, \qquad -\Omega = \Delta\psi,$$

with boundary conditions

$$\begin{aligned} x = 0 \quad \text{and} \quad & x = 1: \ \psi = 0, \ \partial\psi/\partial x = 0; \\ & y = 0: \ \psi = 0, \ \partial\psi/\partial y = 0; \\ & y = 1: \ \psi = 0, \ \partial\psi/\partial y = -1. \end{aligned}$$

These equations have been discretized on a grid $(x_i, y_j) = (ih, jh)$, $i, j = 0, 1, \ldots, N$ $(h = 1/N)$ using central differences. The discrete equations have been solved in the following way, scanning the grid along horizontal lines from left to right and starting at $y = 1$:

$$\begin{aligned}
&\Omega_{i,N}^{n+1} = (1-\omega_b)\,\Omega_{i,N}^{n} + 2\omega_b h^{-2}(h - \psi_{i,N-1}^{n}), && i = 1,\ldots,N-1;\\
&\Omega_{0,j}^{n+1} = (1-\omega_b)\,\Omega_{0,j}^{n} - 2\omega_b h^{-2}\psi_{1,j}^{n}, && j = 1,\ldots,N-1;\\
&\Omega_{i,j}^{n+1} = (1-\omega^*)\,\Omega_{i,j}^{n} + \omega^*(C_N\Omega_{i,j+1}^{n+1} + C_W\Omega_{i-1,j}^{n+1} + C_E\Omega_{i+1,j}^{n} + C_S\Omega_{i,j-1}^{n}),\\
&\psi_{i,j}^{n+1} = (1-\omega_\psi)\,\psi_{i,j}^{n} + \tfrac{1}{4}\omega_\psi(h^2\Omega_{i,j}^{n+1} + \psi_{i,j+1}^{n+1} + \psi_{i-1,j}^{n+1} + \psi_{i+1,j}^{n} + \psi_{i,j-1}^{n}),\\
&\qquad\qquad && i,j = 1,\ldots,N;\\
&\Omega_{N,j}^{n+1} = (1-\omega_b)\,\Omega_{N,j}^{n} - 2\omega_b h^{-2}\psi_{N-1,j}^{n+1}, && j = 1,\ldots,N-1;\\
&\Omega_{i,0}^{n+1} = (1-\omega_b)\,\Omega_{i,0}^{n} - 2\omega_b h^{-2}\psi_{i,1}^{n+1}, && i = 1,\ldots,N-1.
\end{aligned}$$

The relaxation factor ω^* has been chosen according to the present LR method as indicated in (20), where for simplicity $\gamma_1 = \gamma_2 = 1.644$ has been used. The coefficients in the vorticity equation are given by

$$C_N = \tfrac{1}{4} + \tfrac{1}{16}\,\mathrm{Re}(\psi_{i+1,j+1}^{n+1} - \psi_{i-1,j+1}^{n+1}),$$

and similar expressions for the others, with the understanding that for $\psi_{i+1,j-1}$ and $\psi_{i-1,j-1}$ only the values from the nth sweep are available. The streamfunction equation and the boundary conditions have been combined with a relaxation factor too; these were chosen constant throughout the field. The iteration process with starting values zero was terminated when

$$\max_{i,j} |\psi_{i,j}^{n+1} - \psi_{i,j}^{n}| < 5 \times 10^{-6}.$$

The two relaxation parameters ω_ψ and ω_b have been varied; the most efficient ones encountered are listed.

For $\mathrm{Re} = 5 \times 10^4$ and $h = \frac{1}{16}$, the minimum number of iterations we have obtained is 494 ($\omega_\psi = 0.7$, $\omega_b = 0.02$). The calculation time required on a CDC Cyber 170-760 amounts 2 CPU seconds. For comparison, the method of Kurtz, *et al.* [36] requires 8261 CPU seconds on a CDC 6600; their stopping criterion is roughly comparable to ours. Taking into account the difference in computer speed (a factor 3–4), the present LR method is faster by a factor of about 10^3.

For $\mathrm{Re} = 10^6$ and $h = \frac{1}{16}$, the minimum number of iterations obtained is 1836 ($\omega_\psi = 0.1$, $\omega_b = 0.005$) which requires 7 CPU seconds on the Cyber 170-760.

Of course, the discrete solutions for both Reynolds numbers on such a coarse grid have little to do with the continuous solutions; therefore we do not present any results (for $\mathrm{Re} = 5 \times 10^4$ see [36]). These examples merely serve to show that using the present LR method it is not difficult to obtain the discrete solution of centrally discretized Navier–Stokes equations.

6. Discussion

A local relaxation method can be considered as a generalization of the successive overrelaxation method. For equations with constant coefficients they are equivalent. in which case they can be used only if the eigenvalues μ of the Jacobi matrix satisfy $-1 < \mathrm{Re}\,\mu < 1$. For equations with nonconstant coefficients the LR methods prescribe nonconstant relaxation factors: no theory is available for these situations at the moment. The LR methods differ among themselves in the way the relaxation factors are chosen.

In this paper, the LR methods have been applied to (one-dimensional and) two-dimensional convection–diffusion equations, discretized with central differences on a five-point molecule leading to discrete equations of type (15a). The restrictions of Eqs. (21), which often are fulfilled, are sufficient to guarantee that the eigenvalues of the local Jacobi matrix satisfy $-1 < \mu_R < 1$. Under these restrictions the present LR method (20) has been designed. Nevertheless the parameter choice given in (20) can also be useful in neighbouring situations where restrictions (21) are slightly violated. An example of this is given by the driven cavity problem in Section 5 where (21b) need not be satisfied.

For larger deviations from (21), and for generalizations to three or more dimensions, the analysis leading to (20) has to be revisited. The starting point remains the relation between the optimum SOR factor given in (10) and the eigenvalues of the Jacobi matrix. The requirement $-1 < \mu_R < 1$, necessary for convergence in the constant coefficient case, ensures that relaxation factor (10) and its approximation (13) take real values; whether it is satisfied has to be checked in each situation. Further, the relation between the coefficients of the discrete equation and the eigenvalues of the local Jacobi matrix can still be given by an expression of type (16a). This relation combined with (10) (or (13)) gives a choice of the local relaxation factor ω which is (near) optimal in the constant coefficient case. If the

expression thus obtained is regarded as too expensive to evaluate, a simple approximation may be sought leading to an analogue of (20). We think we can leave these steps to the interested reader.

7. Conclusion

A large variety of equations has been used to test the power of the LR strategy. Our experience thus far is that, apart from some notorious turning-point problems, it is always possible to choose the local relaxation factors such that the LR method converges. The present choice given in (20) can be used when restrictions (21) are (approximately) satisfied; for equations with constant coefficients it is as efficient as the optimum SOR method. Further, it is our experience that for equations with strongly varying coefficients, and for nonlinear equations, a properly chosen LR method will be more efficient than the optimum SOR method; the difference can be several orders of magnitude.

In conclusion, it has been shown that an LR strategy (of which a very fine example was already available in the early days of the upwind era) can easily solve central difference approximations of convection–diffusion equations in cases of a small diffusion coefficient, thereby eliminating the need for the usually made trade-off between the accuracy of the central difference solution in favour of the convergence of the upwind-type methods.

References

1. O. R. Burggraf, *J. Fluid Mech.* **24** (1966), 113–151.
2. S. Y. Tuann and M. D. Olson, *J. Comput. Phys.* **29** (1978), 1–19.
3. P. K. Khosla and S. G. Rubin, *J. Engrg. Math.* **13** (1979), 127–141.
4. A. S. Benjamin and V. E. Denny, *J. Comput. Phys.* **33** (1979), 340–358.
5. G. E. Forsythe and W. R. Wasow, "Finite Difference Methods for Partial Differential Equations," Wiley, New York, 1960.
6. D. Greenspan, "Lectures on the Numerical Solutions of Linear, Singular, and Nonlinear Differential Equations," Prentice–Hall, Englewood Cliffs, N.J., 1968.
7. A. D. Gosman, W. M. Pun, A. K. Runchal, D. B. Spalding, and M. Wolfshtein, "Heat and Mass Transfer in Recirculating Flows," Academic Press, New York, 1969.
8. A. K. Runchal, *Internat. J. Numer. Methods Engrg.* **4** (1972), 541–550.
9. G. D. Raithby and K. E. Torrance, *Comput. and Fluids* **2** (1974), 191–206.
10. M. Atias, M. Wolshtein, and M. Israeli, *in* "Proceedings, AIAA 3rd Computational Fluid Dynamics Conference," Hartford, 1975.
11. G. de Vahl Davis and G. D. Mallinson, *Comput. and Fluids* **4** (1976), 29–43.
12. A. Moult, D. Burley, and H. Rawson, *Internat. J. Numer. Methods Engrg.* **14** (1979), 11–35.
13. C. W. Richards and C. M. Crane, *Appl. Math. Modelling* **3** (1979), 205–211.
14. S. I. Cheng and G. Shubin, *J. Comput. Phys.* **28** (1978), 315–326.
15. P. M. Gresho and R. L. Lee, *Comput. and Fluids* **9** (1981), 223–253.
16. D. N. de G. Allen and R. V. Southwell, *Q. J. Mech. Appl. Math.* **8** (1955), 129–145.
17. N. C. Steele and K. E. Barrett, *Internat. J. Numer. Methods Engrg.* **12** (1978), 405–414.

18. D. B. SPALDING, *Internat. J. Numer. Methods Engrg.* **4** (1972), 551–559.
19. S. C. R. DENNIS AND G. Z. CHANG, *Phys. Fluids* **12** (II)(1969), 88–93.
20. D. A. H. JACOBS, *in* "Numerical Methods in Fluid Dynamics" (C. A. Brebbia and J. J. Connor, Eds.), Pentech, London, 1974.
21. P. K. KHOSLA AND S. G. RUBIN, *Comput. and Fluids* **2** (1974), 207–209.
22. C. W. RICHARDS AND C. M. CRANE, *Appl. Math. Modelling* **2** (1978), 59–61.
23. A. E. P. VELDMAN, *Comput. and Fluids* **1** (1973), 251–271.
24. D. DIJKSTRA, "The Solution of the Navier–Stokes Equations near the Trailing Edge of a Flat Plate," Ph. D. Thesis, University of Groningen, 1974.
25. H. FASEL, "Untersuchungen zum Problem des Grenzschichtumschlages durch numerische Integration der Navier–Stokes Gleichungen" (in German), Ph. D. Thesis, University of Stuttgart, 1974.
26. R. F. WARMING AND R. M. BEAM, *in* "Proceedings, AIAA 3rd Computational Fluid Dynamics Conference," Hartford, 1975.
27. D. B. RUSSELL, "On Obtaining Solutions to the Navier–Stokes Equations with Automatic Digital Computers," Aeronautical Research Council Report R & M 3331, Oxford, 1963.
28. C. J. APELT, *J. Fluid Mech.* **37** (1969), 209–229.
29. N. TAKEMITSU, *J. Comput. Phys.* **36** (1980), 236–248.
30. J. STRIKWERDA, *SIAM J. Sci. Stat. Comput.* **1** (1980), 119–130.
31. D. M. YOUNG, *Trans. Amer. Math. Soc.* **76** (1954), 92–111.
32. D. M. YOUNG, "Iterative Solution of Large Linear Systems," Chap. 6.4, Academic Press, New York, 1971.
33. A. RIGAL, *J. Comput. Phys.* **32** (1979), 10–23.
34. G. KJELLBERG, *Ericsson Technics Stockholm* **2** (1958), 245–258.
35. P. P. N. DE GROEN, "Singularity Perturbed Differential Operators of Second Order," MC Tract 68, Mathematical Centre, Amsterdam, 1976.
36. L. A. KURTZ, R. E. SMITH, C. L. PARKS, AND L. R. BONEY, *Comput. and Fluids* **6** (1978), 49–70.
37. L. W. EHRLICH, *J. Comput. Phys.* **44** (1981), 31–45.