JOURNAL OF COMPUTATIONAL PHYSICS 60, 119–134 (1985)

A Modified SOR Method for the Poisson Equation in Unsteady Free-Surface Flow Calculations

EUGEN F. F. BOTTA

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

AND

MARCEL H. M. ELLENBROEK

Space Division, Fokker B.V., 1117 ZJ Schiphol, The Netherlands

Received June 13, 1984; revised November 13, 1984

Convergence difficulties that sometimes occur if the successive overrelaxation (SOR) method is applied to the Poisson equation on a region with irregular free boundaries are analyzed. It is shown that these difficulties are related to the treatment of the free boundaries and caused by the appearance of complex eigenvalues in the system of discrete equations, when standard centered differences are used. After a modification of this system of equations such that the complex eigenvalues become small, a modified SOR method is presented where two relaxation factors are used alternately. The method leads to fast convergence without requiring specific information about the complex eigenvalues. © 1985 Academic Press, Inc.

1. INTRODUCTION

When the unsteady Navier–Stokes equations formulated in primitive variables are solved by a time marching technique, the pressure is usually determined from a Poisson equation. Since we have to solve this Poisson equation in each time step, which is the most time consuming part of the computational effort, and since a typical run may consist of up to 100,000 time steps, it is worthwhile to look for fast numerical methods for solving a Poisson equation. The choice of a suitable method depends on the details of the flow problem on hand. Our interest is focussed on the simulation of liquid motion in the presence of a free liquid surface. An established method for solving flow problems with free surfaces is the MAC (marker-and-cell) method, developed in the mid-sixties by Welch *et al.* [1]. This method forms the basis for a very popular family of methods, of which each member is dedicated to a specific physical situation. For flows involving arbitrarily shaped free surfaces, the SOLA–VOF technique, introduced by Hirt and Nichols [2], is probably the most

0021-9991/85 $3.00

powerful. The method has been used successfully for a large number of applications, e.g., bubble dynamics [3], sloshing of liquid in baffled ship tanks [4], and liquid sloshing in spacecraft under micro-gravity [5].

In the course of the latter application the desire arose to replace the Poisson solver of the SOLA–VOF code. This solver belongs to the class of point-iterative methods, accelerated by means of a single relaxation parameter. Per iteration, point-iterative methods are very cheap in terms of computational effort. Hence, when not too many iterations are required they are sufficiently efficient. This situation occurs when the fluid motion is almost stationary, since then a very good initial guess for the solution of the Poisson equation is available from the previous time step. Under more general circumstances many iterations may be required even with an optimal choice of the relaxation parameter. We will see in the sequel that the boundary condition on the free liquid surface is responsible for the slow convergence. To be more precise, this boundary condition introduces complex eigenvalues in the discrete Poisson equation, for which standard relaxation strategies like SOR are not particularly suited.

In the last decade a number of other solution techniques have been developed. Within the class of direct (= noniterative) methods, those based on Fourier analysis and cyclic reduction (see the review paper by Schwartztrauber [6]) have gained great popularity. For internal flow problems without moving free surfaces, they have been combined successfully with the MAC method [7]. Irregularities in the computational domain can be handled by the capacitance matrix technique. However, when the shape of the domain is changing each time step, as is the case with a moving free surface, the computational overhead required by the latter technique makes these direct methods less attractive.

Also within the class of iterative techniques two types of methods have matured recently: methods based on matrix preconditioning, such as the ICCG method [8], and the multi-level methods [9]. In the former methods, prior to starting the iterations, an incomplete decomposition of the discrete matrix is constructed which has very good convergence properties. In the multi-level methods the iterative solution is pursued by using various grids, of which the finest one corresponds to the desired solution. A high convergence rate can be obtained; however, the use of various grids requires that on each grid a version of the discrete solution matrix has to be determined before the iterations can start. Both types of methods perform very well on fine grids; the start-up overhead and the higher computational effort per iteration, as compared to point-iterative schemes like SOR, are clearly compensated by a much higher convergence rate of the iterative process. On the coarser grids, which are used in the fluid flow calculations we are interested in, this comparison is less pronounced.

In recent years a new aspect is playing a role in the trade-off between the various solution methods: the amount of parallelism in the algorithm. The current supercomputers like CDC Cyber 205 and Cray-1 differ in their internal architecture from the traditional computers. In scalar machines the central processor performs its operations one at a time (sequentially), whereas in the supercomputers operations

can be carried out in parallel. To profit from this property it is necessary that the computational process can be divided into a number of independent parts that can go on concurrently. For instance, when the arithmetic to be performed for one grid point is independent of what is going on in the other grid points, all grid points can be treated at the same time. The extent to which this is possible depends upon the structure of the numerical algorithm used. As an example, point-iterative methods can profit relatively more from this, making them more competitive in comparison to other methods.

In view of these arguments it was decided to investigate point-iterative methods particularly suited for systems of equations with complex eigenvalues. Special emphasis is given to the Poisson equation present in the SOLA–VOF code for the simulation of free-surface liquid dynamics. The results of this investigation are presented in this paper. Section 2 contains the description of the Poisson problem we are interested in. In Section 3 it is explained why the complex eigenvalues slow down the convergence of the SOR method. Thereafter the discrete equations are reformulated in order to decrease the size of the complex eigenvalues (Section 4). Based on the theory given in Section 5, it is suggested in Section 6 that alternate use of two relaxation parameters be made. One parameter is selected to decrease the iteration residuals in the subspace spanned by the eigenvectors corresponding to the complex eigenvalues. The other parameter does the same for the residuals in the subspace corresponding to the real eigenvalues. When all eigenvalues are real the optimum SOR parameter is selected. When complex eigenvalues are present it will be shown that the use of two relaxation parameters, in the way described above, leads to much faster convergence than can be obtained with optimum SOR.

2. Formulation of the Problem

We consider a two-dimensional Poisson equation for the pressure distribution of the fluid inside a partly filled container. For analysing the difficulties mentioned in the Introduction we consider configurations as given in Fig. 1, where the container

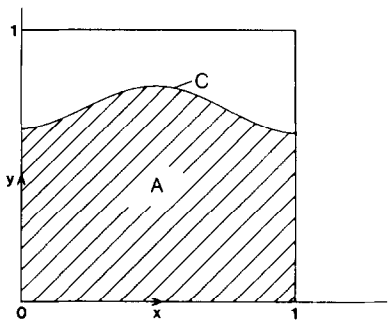

FIG. 1. The region A occupied by the fluid and its free surface C.

corresponds with the square $[0, 1] \times [0, 1]$ and the height of the free surface C can be described as a single-valued function of x.

In the MAC-type calculation methods the Poisson problem can be stated as

$$\Delta p = f \quad \text{on } A, \tag{1}$$

$$p = p_0 \quad \text{on } C, \tag{2}$$

$$\frac{\partial p}{\partial n} = g \quad \text{on the container wall,} \tag{3}$$

where f, p_0 and g are given functions and $\partial/\partial n$ means differentiation in the direction of the outward normal. The problem is discretized via a rectangular grid covering the whole of the container. For ease of presentation in the analysis the grid spacing is chosen to be the same in both directions, thus the grid points are (ih, jh); $i, j = 0, 1, \ldots, m$ $(h = 1/m)$. The cells of this grid are labeled to indicate their position with respect to the liquid. Three types of cells can be distinguished: empty, surface and full cells. Empty cells contain no liquid at all, surface cells are cells containing liquid but have at least one empty neighbour cell, and the other cells containing fluid are termed full. The pressure p is defined in cell centers $(x_i, y_j) = ((i - \frac{1}{2})h, (j - \frac{1}{2})h)$.

In the center of a full cell Eq. (1) is applied, which in discrete form (using second-order central differences) reads

$$p_{i-1,j} + p_{i,j-1} - 4p_{i,j} + p_{i+1,j} + p_{i,j+1} = h^2 f(x_i, y_j). \tag{4}$$

For points adjacent to the wall the Neumann condition (3) is used to eliminate pressure values corresponding to points outside the container. For example, in the column $i = 1$ the Poisson equation becomes

$$p_{1,j-1} - 3p_{1,j} + p_{2,j} + p_{1,j+1} = h^2 f(x_1, y_j) - hg(0, y_j).$$

The Dirichlet condition (2) is used to derive a discrete equation to be applied in the center of surface cells. In this equation also the pressure in an adjacent full cell shows up. When the surface cell has more than one neighbouring full cell, that one is selected for which the line connecting the respective cell centers is most orthogonal to the free surface. For details see [2]. Thus we end up with a typical

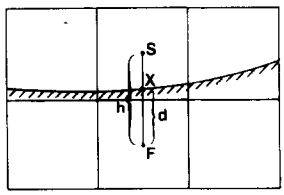

FIG. 2. Implementation of the pressure condition at the free surface.

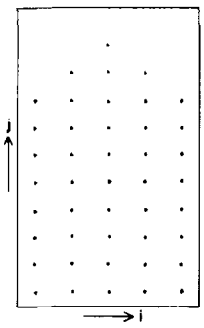

FIG. 3. The computational domain for $m = 5$ and $n_1 = 8$, $n_2 = 9$, $n_3 = 10$, $n_4 = 9$, $n_5 = 8$.

configuration as displayed in Fig. 2. Condition (2) will be applied at the point of intersection X of the free surface with the line connecting the center S of the surface cell and the center F of the full cell. The pressure in the point X is set equal to the value obtained from a linear interpolation (or extrapolation) between the pressure at the points S and F,

$$p(S) - (1 - \eta)\, p(F) = \eta p_0(X), \tag{5}$$

where $\eta = h/d$ is the ratio of the distance between the free surface and F (see Fig. 2). The resulting system of equations will be denoted in vector notation as

$$Ap = b. \tag{6}$$

The empty cells are omitted from the computation. Thus the computational domain possesses a shape governed by the position of the free surface. For configurations like the one in Fig. 1, the domain can be characterized by numbers n_i $(i = 0, 1, \ldots, m)$, where n_i is equal to the number of cells in the ith column which take actively part in the calculation (full cells and surface cells). An example is given in Fig. 3. Our investigations in the sequel of the paper will be performed for various choices of n_i.

3. THE DIFFICULTIES WITH SOR

If SOR is employed to solve the system (6), it turns out that the optimum relaxation factor ω_{opt} strongly depends upon the values of η and for larger values of η even the Gauss–Seidel process diverges. Thus we may conclude that the eigenvalues of the corresponding Jacobi matrix are greatly influenced by the values of η. In the special case of a free surface with constant height and a non-staggered grid, it is possible to apply separation. After some analytic manipulation we find that the

TABLE I

The Six Real Eigenvalues of Largest Modulus and the Complex Eigenvalues of the Jacobi Matrix for Various Configurations

	$\eta_1, \eta_2, \ldots, \eta_5$		
$\eta_1, \eta_2, \ldots, \eta_5$	8, 8, 8, 8, 8	8, 9, 10, 9, 8	6, 7, 8, 9, 10
5, 10, 5, 10, 5	± 0.983	± 0.985	± 0.978
	± 0.856	± 0.874	± 0.867
	± 0.855	± 0.857	± 0.820
	$\pm 1.410i$	± 1.453	$\pm 1.403i$
	$\pm 0.042 \pm 1.256i$	$\pm 0.023 \pm 1.227i$	$\pm 0.086 \pm 1.274i$
	$\pm 1.165i$	$\pm 1.113i$	$\pm 1.123i$
	$\pm 1.009i$	$\pm 1.077i$	$\pm 1.026i$
1, 5, 10, 50, 100	± 0.983	± 0.985	± 0.981
	± 0.860	± 0.882	± 0.867
	± 0.853	± 0.857	± 0.834
	$\pm 5.723i$	$\pm 5.719i$	$\pm 5.729i$
	$\pm 3.487i$	$\pm 3.485i$	$\pm 3.487i$
	$\pm 1.450i$	$\pm 1.489i$	$\pm 1.450i$
	$\pm 0.960i$	$\pm 0.968i$	$\pm 0.968i$
1, 10, 50, 100, 500	± 0.983	± 0.986	± 0.980
	± 0.858	± 0.881	± 0.866
	± 0.853	± 0.856	± 0.832
	$\pm 12.889i$	$\pm 12.887i$	$\pm 12.892i$
	$\pm 4.958i$	$\pm 4.957i$	$\pm 4.956i$
	$\pm 3.489i$	$\pm 3.501i$	$\pm 3.489i$
	$\pm 1.460i$	$\pm 1.466i$	$\pm 1.466i$

eigenvalues become complex for $\eta > 1$ and that for large values of η the imaginary part of the eigenvalues is given approximately by

$$\tfrac{1}{2}\sqrt{\eta - 1} \tag{7}$$

while the real parts remain relatively small.

The method of separation fails on the staggered grid, but for various configurations we have determined the eigenvalues numerically. Some typical examples are given in Table I for $m = 5$ and interpolation in y-direction. All results indicate that even for nonhorizontal surfaces there is a similar relation between the values of η and the imaginary parts. We also notice that the real parts of the eigenvalues are in absolute value smaller than 1 and that the dominant ones correspond with real eigenvalues of the Jacobi matrix. A well known result from SOR theory is the relation

$$(\lambda + \omega - 1)^2 = \lambda \omega^2 \mu^2 \tag{8}$$

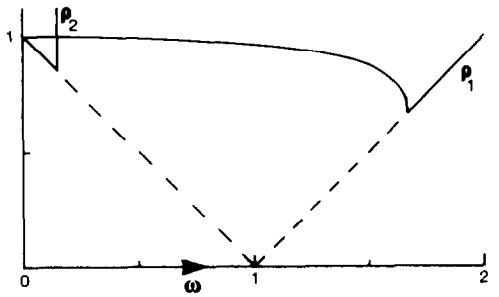

FIG. 4. The graphs of $\rho_1(\omega)$ and $\rho_2(\omega)$ defined by (9).

between the eigenvalues λ of the SOR matrix and the eigenvalues μ of the Jacobi matrix. From this relation it follows that if all eigenvalues μ satisfy $|\mathrm{Re}\,\mu| < 1$, the SOR method converges for ω sufficiently small. However, even for the optimal choice ω_{opt}, the convergence can be extremely slow. This is easily demonstrated in the last case of Table I by restricting the spectrum of the Jacobi matrix to the dominant real and complex eigenvalues $\pm\mu_i$, $i=1,2$, with $\mu_1 = 0.980$ and $\mu_2 = 12.892i$. Let $\lambda_i^+(\omega)$ and $\lambda_i^-(\omega)$ be the eigenvalues of the SOR matrix corresponding with $\pm\mu_i$ and define

$$\rho_i(\omega) = \max\{|\lambda_i^+(\omega)|, |\lambda_i^-(\omega)|\}, \qquad i=1,2. \tag{9}$$

Both graphs are shown in Fig. 4. In the point of intersection we have $\omega = \omega_{\mathrm{opt}} = 0.144$ and $\rho_i(\omega_{\mathrm{opt}}) = 0.997$, clearly illustrating the slow convergence. More details and further references can be found in Botta and Veldman [10].

4. The Modified System of Equations

In the previous section we have seen that the slow convergence of the SOR method is caused by the appearance of complex eigenvalues in the Jacobi matrix due to values $\eta > 1$. Let us consider a single equation of type (5),

$$p(S) - (1-\eta)\,p(F) = b_1, \tag{10}$$

and suppose that point F, for which we have an equation of type (4),

$$p(F_w) + p(F_s) - 4p(F) + p(F_e) + p(S) = b_2, \tag{11}$$

is the only point connected with S. Since the numbering of points has no influence on the spectrum of the Jacobi matrix, we may start with the points F and S. Then, aside from the second element, the first row and column of this matrix are zero; see Fig. 5. It will be clear that only the product $\frac{1}{4}(1-\eta)$ of these elements is of interest

$$\begin{bmatrix} 0 & 1-\eta & 0\cdots 0 \\ \frac{1}{4} & & \\ 0 & & \\ \vdots & & \\ 0 & & \end{bmatrix}$$

FIG. 5. The first row and column of the Jacobi matrix.

for the eigenvalues of the matrix. However, this product can become arbitrarily large and the only way to avoid this difficulty is to replace the element $\frac{1}{4}$ by 0. This can be done by subtracting Eq. (10) from Eq. (11), thus replacing (11) by

$$p(F_w) + p(F_s) - (3+\eta)\, p(F) + p(F_e) = b_2 - b_1. \tag{12}$$

For the Jacobi matrix this means the following modification of the second row: the first element $\frac{1}{4}$ becomes 0 and the other elements $\frac{1}{4}$ must be replaced by $1/(3+\eta)$. Hence, the sum of the elements in the second row usually becomes smaller and the element $1-\eta$ in the first row is no longer of influence. For a horizontal surface each surface cell has only one neighbouring full cell and therefore the various modifications lead in fact to the elimination of all surface cells in the equations for the full cells. For this special configuration it can be shown analytically that all eigenvalues become real and in absolute value smaller than 1. Unfortunately, complex eigenvalues appear again for more general surfaces, as is shown in Table II where the corresponding results from Table I are given for the modified system. We observe that the absolute value of the complex eigenvalues is significantly lower as compared with the old situation. The dominant eigenvalues are all real and their absolute value is less than 1. This last fact can be proven for all surfaces where the center S of a surface cell is connected with at most two centers of full cells, which is the normal situation; see Fig. 6. It can be done by replacing the Jacobi matrix by a similar matrix (thus with the same eigenvalues) with a row norm smaller than 1. If the point S is connected with 3 or 4 centers of full cells we neither did find a simple proof, nor did we find a counterexample.

If SOR is applied to the modified system the convergence is much faster as compared with the old situation. This is demonstrated by a discussion of the same kind as given at the end of Section 3 for the corresponding case in Table II. With $\mu_1 = 0.980$ and $\mu_2 = 0.097 + 0.215i$ we now find a spectral radius of 0.857 instead of 0.997. Although this is a considerable improvement, the convergence remains rather slow and in the next section we will give the theoretical framework for a faster method.

We finally remark that, in general, the five-point structure is not retained after a complete elimination of the surface cells. This prohibits efficient programming and the application of SOR theory and therefore complete elimination is not recommendable.

TABLE II

Eigenvalues as in Table I for the Modified System

	$n_1, n_2, ..., n_5$		
$\eta_1, \eta_2, ..., \eta_5$	8, 8, 8, 8, 8	8, 9, 10, 9, 8	6, 7, 8, 9, 10
5, 10, 5, 10, 5	± 0.983	± 0.985	± 0.978
	± 0.857	± 0.875	± 0.867
	± 0.856	± 0.857	± 0.822
		$\pm 0.406 \pm 0.259i$	$\pm 0.297i$
		$\pm 0.235i$	$\pm 0.229 \pm 0.282i$
		$\pm 0.264 \pm 0.118i$	$\pm 0.457 \pm 0.264i$
		$\pm 0.109 \pm 0.046i$	$\pm 0.172 \pm 0.060i$
		$\pm 0.132 \pm 0.020i$	$\pm 0.448 \pm 0.042i$
1, 5, 10, 50, 100	± 0.983	± 0.986	± 0.981
	± 0.861	± 0.882	± 0.867
	± 0.854	± 0.857	± 0.834
		$\pm 0.255i$	$\pm 0.111 \pm 0.224i$
		$\pm 0.363 \pm 0.239i$	$\pm 0.321 \pm 0.156i$
		$\pm 0.155 \pm 0.044i$	$\pm 0.084i$
		$\pm 0.289 \pm 0.026i$	$\pm 0.041i$
		$\pm 0.084 \pm 0.022i$	
		$\pm 0.515 \pm 0.007i$	
1, 10, 50, 100, 500	± 0.983	± 0.986	± 0.980
	± 0.859	± 0.881	± 0.866
	± 0.853	± 0.856	± 0.832
		$\pm 0.228i$	$\pm 0.097 \pm 0.215i$
		$\pm 0.282 \pm 0.173i$	$\pm 0.272 \pm 0.137i$
		$\pm 0.148 \pm 0.039i$	$\pm 0.083i$
		$\pm 0.085 \pm 0.018i$	
		$\pm 0.291 \pm 0.017i$	

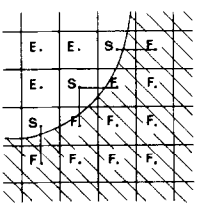

FIG. 6. The cell labeling and the connection between surface cells and full cells.

5. Invariant Subspaces and Convergence

If we take a chessboard ordering of the grid points with l "white" points and m "black" points, we can write the Jacobi matrix B in partitioned form as

$$B = \begin{bmatrix} 0 & R \\ L & 0 \end{bmatrix},$$

with R an $l \times m$ matrix, L an $m \times l$ matrix and 0 a zero matrix. Throughout this section we will use this partitioning for vectors and matrices.

We start considering the case that the eigenvectors of B form a basis for $\mathbb{C}^N$. If μ_i is an eigenvalue of B and

$$\begin{pmatrix} v_i \\ w_i \end{pmatrix}$$

the corresponding eigenvector then it is straightforward to verify that also

$$\begin{pmatrix} v_i \\ -w_i \end{pmatrix}$$

is an eigenvector of B and $-\mu_i$ the associated eigenvalue. Thus, the eigenvalues of B occur in pairs. Let all nonzero eigenvalues of B be given by the k pairs $\pm\mu_i$, $i = 1, 2, \ldots, k$, with $\operatorname{Re} \mu_i \geqslant 0$ and define the subspaces

$$K_i = \left\{ \begin{pmatrix} v_i \\ w_i \end{pmatrix}, \begin{pmatrix} v_i \\ -w_i \end{pmatrix} \right\}, \qquad i = 1, 2, \ldots, k, \tag{13}$$

spanned by the eigenvectors associated with $\pm\mu_i$. Finally, we define the subspace K_0 as the nullspace of B, i.e., the set of all vectors x for which $Bx = 0$. Clearly,

$$\mathbb{C}^N = K_0 \oplus K_1 \oplus \cdots \oplus K_k, \tag{14}$$

and all subspaces K_i are invariant under B.

For the same ordering of grid points, we now consider the SOR matrix

$$H(\omega) = \begin{bmatrix} I_1 & 0 \\ -\omega L & I_2 \end{bmatrix}^{-1} \begin{bmatrix} (1-\omega) I_1 & \omega R \\ 0 & (1-\omega) I_2 \end{bmatrix}, \tag{15}$$

where I_1 and I_2 are identity matrices. In Section 3 we have already given the relation (8) between the eigenvalues of B and $H(\omega)$. Let λ_i^+ and λ_i^- be the roots of Eq. (8) for $\mu^2 = \mu_i^2 \neq 0$. Evidently, $|\lambda_i^+ \lambda_i^-| = (\omega - 1)^2$ and therefore we can always take $|\lambda_i^+| \geqslant |\omega - 1| \geqslant |\lambda_i^-|$. Now it can directly be verified that the eigenvectors of $H(\omega)$ associated with the eigenvalues λ_i^+ and λ_i^- are given by

$$\begin{pmatrix} v_i \\ (\lambda_i^+)^{1/2} w_i \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} v_i \\ \operatorname{sign}(\omega - 1)(\lambda_i^-)^{1/2} w_i \end{pmatrix}, \tag{16}$$

respectively. For $\lambda_i^+ \neq \lambda_i^-$ the eigenvectors (16) form a basis for the subspace K_i, whereas for $\lambda_i^+ = \lambda_i^-$ these eigenvectors become dependent, but it can be shown that

$$\begin{pmatrix} 0 \\ \frac{1}{2}(\lambda_i^+)^{-1/2} w_i \end{pmatrix} \tag{17}$$

is a principal vector of grade 2 and hence the subspace K_i is invariant under $H(\omega)$ as well as under B.

We will now consider the convergence of the SOR method in connection with the subspaces K_i. The error after n iterations is denoted by $\varepsilon^{(n)}$ and since $\varepsilon^{(n+1)} = H(\omega)\,\varepsilon^{(n)}$, we have

$$\varepsilon^{(n)} = (H(\omega))^n\,\varepsilon^{(0)}.$$

From (14) we know that the error $\varepsilon^{(0)}$ in the starting values can be written as the sum of its components in the subspaces K_i, i.e.,

$$\varepsilon^{(0)} = \sum_{i=0}^{k} \varepsilon_i^{(0)}, \qquad \varepsilon_i^{(0)} \in K_i.$$

If $H_i(\omega)$ is the restriction of $H(\omega)$ to the subspace K_i, we obtain

$$\varepsilon^{(n)} = \sum_{i=0}^{k} (H_i(\omega))^n\,\varepsilon_i^{(0)},$$

and for studying the convergence of the iteration process we only have to examine the convergence within a single subspace K_i. Therefore, we restrict ourselves to one component

$$\varepsilon_i^{(n)} = (H_i(\omega))^n\,\varepsilon_i^{(0)}$$

and, for brevity, to the case $\omega \geqslant 1$ and $\lambda_i^+ \neq \lambda_i^-$. Now the eigenvectors (16) of $H(\omega)$ form a basis for K_i on which we can represent $H_i(\omega)$ by the diagonal matrix

$$\Lambda_i(\omega) = \begin{bmatrix} \lambda_i^+(\omega) & 0 \\ 0 & \lambda_i^-(\omega) \end{bmatrix}, \tag{18}$$

but this basis depends on ω through $\lambda_i^{\pm}(\omega)$. As follows from the construction of (13), we can define another basis by

$$\begin{pmatrix} v_i \\ w_i \end{pmatrix}, \qquad \begin{pmatrix} v_i \\ -w_i \end{pmatrix} \tag{19}$$

which is independent of ω. Using (16) and (18) we can write $H_i(\omega)$ on the basis (19) as

$$\tilde{\Lambda}_i(\omega) = Q^{-1}(\omega)\,\Lambda_i(\omega)\,Q(\omega),$$

with

$$Q(\omega)=\begin{bmatrix} 1 & 1 \\ (\lambda_i^+(\omega))^{1/2} & (\lambda_i^-(\omega))^{1/2} \end{bmatrix}^{-1}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Let $\varepsilon_i^{(0)}$ on the basis (19) be given by

$$\varepsilon_i^{(0)}=\begin{pmatrix} c_1 \\ c_2 \end{pmatrix},$$

then

$$\varepsilon_i^{(n)}=(\tilde{\Lambda}_i(\omega))^n\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}=Q^{-1}(\omega)\,\Lambda_i^n(\omega)\,Q(\omega)\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}.$$

Since we have taken $|\lambda_i^+(\omega)| \geqslant |\lambda_i^-(\omega)|$, we can find for a given norm $\|\cdot\|$ a constant C such that for all n

$$\|\varepsilon_i^{(n)}\| \leqslant C|\lambda_i^+(\omega)|^n \tag{20}$$

and, more generally, after n_1 iterations with a relaxation factor ω_1 and, in addition, n_2 iterations with a relaxation factor ω_2, we find similarly

$$\|\varepsilon_i^{(n_1+n_2)}\| \leqslant C'(\rho_i(\omega_1))^{n_1}\,(\rho_i(\omega_2))^{n_2} \tag{21}$$

with $\rho_i(\omega)$ defined as in (9).

In the foregoing we have made the restrictions $\omega \geqslant 1$ and $\lambda_i^+ \neq \lambda_i^-$. The first restriction was made for reasons of shortness and is not essential. If $\lambda_1^+ = \lambda_1^-$, however, the vectors (16) are dependent and to get a basis for K_i we must add (17). This leads to a slight modification of (20) in the sense that the constant C now depends on n and must be replaced by n times a new constant, similar to what happens within SOR for the optimum relaxation factor; see [11].

In the beginning of this section we made the assumption that the eigenvectors of B form a basis for $\mathbb{C}^N$, a quite common situation. But even if this is not true, it remains possible to define in a similar way subspaces K_i which are invariant under $H(\omega)$ and where $H(\omega)$ has only the eigenvalues λ_i^+ and λ_i^-. Therefore the convergence within K_i is again determined by the spectral radius $\rho_i(\omega)$ of $H_i(\omega)$. In the next section we will show how to exploit this property.

6. Numerical Method and Results

We consider first, in the notation of the foregoing theory, the example given at the end of Section 4, where the spectrum of the Jacobi matrix is restricted to the eigenvalues $\pm\mu_i$, $i=1,2$, with $\mu_1=0.980$ and $\mu_2=0.097+0.215i$. As we have seen, the SOR convergence within the subspace K_i belonging to $\pm\mu_i$, is determined by

the spectral radius $\rho_i(\omega)$ of $H_i(\omega)$. The function $\rho_1(\omega)$ is well known from standard SOR theory, but the function $\rho_2(\omega)$ is more complicated and for explicit formulae we refer to [10]. If ω_i denotes the value of ω for which $\rho_i(\omega)$ attains its minimum, it turns out that

$$\omega_1 = 1.668, \qquad \rho_1(\omega_1) = 0.668, \qquad \rho_2(\omega_1) = 1.034,$$
$$\omega_2 = 0.988, \qquad \rho_1(\omega_2) = 0.961, \qquad \rho_2(\omega_2) = 0.045.$$

Thus, the SOR process with $\omega = \omega_1$ leads to convergence within the subspace K_1 and to slow divergence within K_2, whereas for $\omega = \omega_2$ SOR converges slowly within K_1 but extremely fast within K_2. In order to get fast convergence in as well K_1 as K_2, we combine both iterations: alternately we do n_1 iterations with ω_1 and n_2 iterations with ω_2. If we require the errors in K_1 and K_2 to be of about the same size, we must take

$$(\rho_1(\omega_1))^{n_1}(\rho_1(\omega_2))^{n_2} \approx (\rho_2(\omega_1))^{n_1}(\rho_2(\omega_2))^{n_2}.$$

In our example this leads to $n_2 \approx 0.14 n_1$ and an average spectral radius of 0.699 which is only slightly above $\rho_1(\omega_1)$ and much better than the value 0.857 given in Section 4 which was obtained for a fixed optimum relaxation factor.

We will now show that for the modified system of equations given in Section 4, this approach can even be applied to the full Jacobi spectrum. Let us first add the real eigenvalues $\pm\mu_r$ with $|\mu_r| < \mu_1$. From SOR theory we know that $\rho_r(\omega) \leqslant \rho_1(\omega)$ and therefore these eigenvalues do not effect the convergence. Hence, as in the classical case of overrelaxation, the choice of ω_1 must be based on the largest real eigenvalue μ_1 or, more practically, on a slight overestimation μ^* hereof. Therefore we replace ω_1 by

$$\omega^* = \frac{2}{1 + [1 - (\mu^*)^2]^{1/2}}.$$

In Table II we can see that the dominant real eigenvalues are hardly influenced by the values of η and hence we let μ^* only depend upon the number of grid points.

With respect to all complex eigenvalues $\pm\mu_c$, the situation is less predictable since there is no corresponding ordening of the curves $\rho_c(\omega)$. Moreover, the complex eigenvalues are greatly influenced by the position of the free surface. In Section 4 we have seen, however, that $|\mu_c|^2 \ll 1$ and as a consequence the curves $\rho_c(\omega)$ reach their minima for values of ω close to 1. Consequently we take $\omega = 1$ (i.e., Gauss–Seidel with $\rho_c(1) = |\mu_c|^2$) as an effective choice for reducing the errors within the subspaces associated with complex eigenvalues.

The ratio of the number of iterations with $\omega = \omega^*$ and $\omega = 1$ depends upon the (usually unknown) maximum values of $\rho_c(1)$ and $\rho_c(\omega^*)$. Fortunately, the decision whether to iterate with $\omega = \omega^*$ or $\omega = 1$ can be made during the iteration. Roughly speaking, we can iterate with $\omega = \omega^*$ until the convergence becomes "significantly slower" than might be expected for a spectral radius $\omega^* - 1$. Since this slowing-

down is caused by the (eventually growing) errors in the subspaces associated with the complex eigenvalues, we next iterate with $\omega = 1$ as long as the convergence is "significantly faster" than might be expected for a spectral radius μ_1^2. Hereafter we continue again with $\omega = \omega^*$. Of course we must be more specific about what is called significantly slower and faster. Although the implementation hereof is not very critical, it is recommendable not to exaggerate the number of changes in ω.

In our computer program we calculate, as is most common, the maximum norm $\|\delta^{(n)}\|_\infty$ of the difference of two successive iterates $p^{(n)}$ and $p^{(n+1)}$. The iteration is always started with $\omega = 1$. During the iteration with $\omega = 1$ we use $\lambda_n = \|\delta^{(n+1)}\|_\infty / \|\delta^{(n)}\|_\infty$ as a measure of the rapidity of convergence. Theoretically λ_n tends to μ_1^2 and therefore we switch to $\omega = \omega^*$ as soon as λ_n becomes larger than $0.9(\mu^*)^2$, which stands for a value somewhat smaller than μ_1^2. To eliminate the influence of irregularities we perform the first 10 iterations after a change in ω without tests. The choice $\omega = \omega^*$ is retained during the next p iterations as long as $\|\delta^{(n)}\|_\infty$ remains smaller than $10(\omega^* - 1)^p \|\delta^{(n-p)}\|_\infty$.

For a comparison of the present method and SOR we have tested both methods by solving the modified system of equations given in Section 4 with all right-hand sides equal to zero; thus the exact solution $p^{(\infty)}$ is zero as well. In Table III we have given, for several configurations and starting values 1, the number of iterations required to get $\|p^{(n)}\|_\infty < 10^{-8}$. With respect to the SOR method we remark that an accurate estimation of the optimum relaxation factor requires full knowledge of the complex eigenvalues in the Jacobi spectrum and this is unrealistic. Therefore we have only listed the SOR results for some fixed values of ω. The results in Table III clearly demonstrate the advantage of the present method over the SOR method and

TABLE III

Comparison of the Number of Iterations for SOR and the Present Method

		SOR				Present method	
$n_1, n_2, \ldots, n_5$	$\eta_1, \eta_2, \ldots, \eta_5$	$\omega = 1.4$	$\omega = 1.5$	$\omega = 1.6$	$\omega = 1.65$	$\omega = 1$ or ω^*	ω^*
8, 9, 10, 9, 8	5, 10, 5, 10, 5	258	195	div	div	93	1.75
	1, 5, 10, 50, 100	270	204	div	div	80	
	1, 10, 50, 100, 500	268	203	669	div	87	
	1, 10, 100, 10, 1	304	231	164	131	75	
6, 7, 8, 9, 10	5, 10, 5, 10, 5	168	293	div	div	82	1.70
	1, 5, 10, 50, 100	195	146	472	div	63	
	1, 10, 50, 100, 500	192	143	372	div	63	
	1, 10, 100, 10, 1	240	div	div	div	82	
17, 18, 19, 18, 17	5, 10, 5, 10, 5	1435	1111	div	div	208	1.88
	1, 5, 10, 50, 100	1462	1132	div	div	168	
	1, 10, 50, 100, 500	1458	1129	839	div	184	
	1, 10, 100, 10, 1	1541	1194	888	747	150	

are in full agreement with the theory. For a number of configurations the eigenvalues of interest can be found in Table II. The complex ones determine a value $\omega_{max} \leqslant 2$ above which SOR diverges and as we see in Table III, ω_{max} varies with the values of η. Consequentially, to avoid divergence of SOR in the calculations with a varying free surface we are forced to take a rather low choice of ω, which causes an additional slowing-down. The last four cases of Table III show the effect of an increase in the number of grid points in the y-direction. The results may be well understood if we remark that the complex eigenvalues of the Jacobi matrix remain nearly the same as in the corresponding first four cases of Table III, where we have a similar form of the free surface. As is usual for a grid refinement the real eigenvalues are changed by which the largest real eigenvalue comes closer to 1. In the present method the value of ω^* is adjusted but in SOR the complex eigenvalues limit the relaxation factor and therefore the advantage of the present method is even enlarged for an increasing number of grid points.

7. Conclusion

We have presented an iterative method that efficiently eliminates the difficulties caused by the appearance of small complex eigenvalues in the Jacobi matrix. In the SOR method such eigenvalues have great influence on the value of the optimum relaxation factor and an overestimation easily leads to divergence. In the present method the optimum choice of the relaxation factor ω^* is much less critical and does not depend upon the complex eigenvalues. A number of test calculations have demonstrated the advantage of the present method over the SOR method.

Acknowledgment

The authors are particularly grateful to Dr. A. E. P. Veldman at the National Aerospace Laboratory in Amsterdam for stimulating and valuable discussions.

References

1. J. E. Welch, F. H. Harlow, J. P. Shannon, and B. J. Daly, "The MAC Method: A Computing Technique for Solving Viscous, Incompressible, Transient Fluid-Flow Problems Involving Free Surfaces," Los Alamos Scientific Laboratory Report LA-3425, 1966.
2. C. W. Hirt and B. D. Nichols, *J. Comput. Phys.* **39** (1981), 201–225.
3. C. W. Hirt and B. D. Nichols, "A Computational Method for Free Surface Hydrodynamics," ASME paper 80-C2/PVP-144, 1980.
4. T. J. Bridges, *in* "Proceedings, 3rd International Conference on Numerical Ship Hydrodynamics, Paris, 1981."
5. A. E. P. Veldman and M. E. S. Vogels, Axisymmetric liquid sloshing under low-gravity conditions, IAF paper 82–141, 33rd IAF conference, Paris, 1982.
6. P. N. Schwarztrauber, *SIAM Rev.* **19** (1977), 490–501.

7. U. SCHUMANN, *in* "Proceedings, 5th International Conference on Numerical Methods in Fluid Dynamics," pp. 398–403, Lecture Notes in Physics, Vol. 59, Springer-Verlag, New York, 1976.
8. J. A. MEIJERINK AND H. A. VAN DER VORST, *Math. Comp.* **31** (1977), 148–162.
9. W. HACKBUSCH AND U. TROTTENBERG (Eds.), *in* "Proceedings, Köln-Porz 1981," Lecture Notes in Mathematics, Vol. 960, Springer-Verlag, Berlin, 1982.
10. E. F. F. BOTTA AND A. E. P. VELDMAN, *J. Comput. Phys.* **48** (1982), 127–149.
11. D. M. YOUNG, "Iterative Solution of Large Linear Systems," Academic Press, New York, 1971.