ITERATIVE METHODS FOR SOLVING MATRIX EQUATIONS

R. S. VARGA, Case Institute of Technology

1. Introduction. Iterative methods are, in concept, well known to us all since it is likely that each of us has at one time or another used Newton's method to find square roots of numbers. It is probably not as well known that iterative methods, utilizing the great speeds of modern computing machines, are extensively used today in practical computations for solving matrix equations which arise from finite difference approximations to elliptic partial differential equations of reactor technology and petroleum technology.

The object of this paper is to illustrate the use of finite difference techniques, and to illustrate the nature of iterative methods for solving the associated matrix equations. We shall do this by means of a very simple one-dimensional problem. It is hoped that this simple example will serve as an elementary introduction to the theory of such iterative methods which is covered in much more detail in [1, 2, 3, 5, 6].

2. A simple two-point boundary value problem. Consider the solution of the two-point boundary value problem:

$$-y^{(2)}(x) + \sigma y(x) = f(x), \quad 0 < x < 1, \qquad y^{(2)} = d^2y/dx^2, \tag{1}$$

where

$$y(0) = \alpha, \qquad y(1) = \beta. \tag{2}$$

We assume that α, β, and σ are given constants with $\sigma \geqq 0$, and $f(x)$ is a given function such that $y^{(4)}(x)$ exists in $0 \leqq x \leqq 1$, and

$$|\, y^{(4)}(x) \,| \leqq M, \qquad 0 \leqq x \leqq 1. \tag{3}$$

By means of Taylor's Theorem, we now express $-y^{(2)}$ in terms of a three-point central difference approximation plus an error:

$$-y^{(2)}(x_i) = [2y(x_i) - (y(x_i + h) + y(x_i - h))]/h^2 + h^2 y^{(4)}(x_i + \theta_i h)/12, \tag{4}$$

where $|\theta_i| < 1$, $x_i \equiv ih$, $1 \leqq i \leqq N$, and $h \equiv 1/(N+1)$. With $y(x_i) \equiv y_i$, the differential equation (1) can be written for the particular values x_i, $1 \leqq i \leqq N$, in matrix form, as

$$A\mathbf{y} = \mathbf{k} + \tau(y), \tag{5}$$

where A is an $N\times N$ real matrix, and $\mathbf{y}$, $\mathbf{k}$, and $\tau(y)$ are vectors, all given explicitly by

$$A = \frac{1}{h^2}\begin{bmatrix} 2+\sigma h^2 & -1 & & \\ -1 & 2+\sigma h^2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2+\sigma h^2 \end{bmatrix}; \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}; \quad \mathbf{k} = \begin{bmatrix} f_1 + \alpha/h^2 \\ f_2 \\ \vdots \\ f_N + \beta/h^2 \end{bmatrix},$$

(6)

$$\tau = \frac{-h^2}{12}\begin{bmatrix} y^{(4)}(x_1+\theta_1 h) \\ \vdots \\ y^{(4)}(x_N+\theta_N h) \end{bmatrix}.$$

Neglecting the vector $\tau(y)$ in (5) gives us the matrix equation

$$A\mathbf{z} = \mathbf{k}, \tag{7}$$

whose solution $\mathbf{z}$ is defined to be our discrete approximation to the solution $y(x)$ of (1)–(2), i.e., z_j, the jth component of $\mathbf{z}$, is to approximate $y_j = y(jh)$.

It is obvious that A is real and symmetric. Noting that all the diagonal entries of A are equal, we can express the matrix A as

$$A = \frac{(2+\sigma h^2)}{h^2}[I - B], \tag{8}$$

where B is an $N\times N$ real symmetric matrix given explicitly by

$$B = \frac{1}{(2+\sigma h^2)}\begin{bmatrix} 0 & 1 & & \\ 1 & 0 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 0 \end{bmatrix}.$$

Because of the tridiagonal form of B, it is easy to verify that the vector $\mathbf{x}^{(j)}$, with k-th component $x_k^{(j)}$ defined by $x_k^{(j)} = \sin(jk\pi h)$, $1 \leqq k \leqq N$, is an eigenvector of B for *each* j, with the corresponding eigenvalue μ_j given by

$$\mu_j = 2\cos(j\pi h)/(2+\sigma h^2), \qquad 1 \leqq j \leqq N. \tag{10}$$

Then, it follows that the spectral radius $\rho(B) \equiv \max_{1\leqq j\leqq N} |\mu_j|$ is less than unity, and thus, from (8), A is evidently positive definite. Moreover, since $\rho(B) < 1$, we can express A^{-1} by means of the convergent matrix infinite series

$$A^{-1} = \gamma(I-B)^{-1} = \gamma\{I + B + B^2 + \cdots\}, \quad \gamma \equiv h^2/(2+\sigma h^2). \tag{11}$$

Note that all entries of B are nonnegative from (9), so that all powers of B and hence A^{-1} have only nonnegative real entries. This last fact will be useful in proving in the next section that the solution $\mathbf{z}$ of (7) is "close" to the solution $y(x)$ of (1)–(2), evaluated at the points $x_i = ih$, $1 \leqq i \leqq N$.

3. Convergence of the discrete approximation. To measure the discrepancy between the components z_i of the vector $\mathbf{z}$ of (7) and the numbers $y_i = y(ih)$, we subtract (7) from (5) giving $A(\mathbf{y}-\mathbf{z}) = \tau(y)$, or, since A is nonsingular,

$$\mathbf{y} - \mathbf{z} = A^{-1}\tau(y). \tag{12}$$

Now, from (3), each component of $\tau(y)$ is bounded above by $Mh^2/12$. Hence, since A^{-1} has only nonnegative components, $|y_i - z_i| \leqq (Mh^2/12)(A^{-1}\xi)_i$ for all $1 \leqq i \leqq N$, where ξ is the vector with all components unity. Thus, if $\mathbf{w}$ is any vector with $A\mathbf{w} \geqq \xi$, we obtain the bound $|y_i - z_i| \leqq Mh^2 w_i/12$. But it is easy to verify that the choice $w(x) \equiv x(1-x)/2$ with $w_i = w(ih)$ satisfies $A\mathbf{w} \geqq \xi$, so that

$$|y_i - z_i| \leqq Mh^2[ih(1-ih)]/24 \leqq Mh^2/96, \tag{13}$$

and last inequality following since $w(x) \leqq 1/8$ for $0 \leqq x \leqq 1$. This means that $\mathbf{z}$ is close to $\mathbf{y}$ if the mesh spacing h is sufficiently small.

4. The Jacobi iterative method. For h small, the result of the previous section shows that solving $A\mathbf{z} = \mathbf{k}$ will give us a reasonable approximation to the solution of the original differential equation (1)–(2). To solve $A\mathbf{z} = \mathbf{k}$ for $\mathbf{z}$, let us combine (7) with (8) to form

$$\mathbf{z} = B\mathbf{z} + \mathbf{g}, \tag{14}$$

where $\mathbf{g} \equiv (h^2\mathbf{k})/(2+\sigma h^2)$. Familiarity with the method of successive substitutions suggests that we consider the following iterative method, called the Jacobi iterative method,

$$\mathbf{z}^{(m+1)} = B\mathbf{z}^{(m)} + \mathbf{g}, \qquad m \geqq 0, \tag{15}$$

where $\mathbf{z}^{(0)}$ is some initial estimate of the unique solution of (14). Writing $\mathbf{z}^{(m)} = \mathbf{z} + \boldsymbol{\varepsilon}^{(m)}$ to define the error $\boldsymbol{\varepsilon}^{(m)}$ at each iteration, we see from (14) and (15) that $\boldsymbol{\varepsilon}^{(m+1)} = B\boldsymbol{\varepsilon}^{(m)}$, from which we deduce that

$$\boldsymbol{\varepsilon}^{(m)} = B^m\boldsymbol{\varepsilon}^{(0)}, \qquad m \geqq 0. \tag{16}$$

Since B is a real symmetric $N \times N$ matrix, the eigenvectors $\{\mathbf{x}^{(j)}\}_{j=1}^{N}$ of B can be normalized to form an orthonormal basis for the associated N dimensional vector space. Thus, there exist constants c_j such that $\boldsymbol{\varepsilon}^{(0)} = \sum_{j=1}^{N} c_j\mathbf{x}^{(j)}$, where $B\mathbf{x}^{(j)} = \mu_j\mathbf{x}^{(j)}$, and it follows from (16) that

$$\boldsymbol{\epsilon}^{(m)} = \sum_{j=1}^{N} (\mu_j)^m c_j\mathbf{x}^{(j)}, \qquad m \geqq 0. \tag{17}$$

Because all μ_j are less than unity in modulus, $\boldsymbol{\varepsilon}^{(m)}$ evidently tends to the zero vector for *any* initial $\boldsymbol{\varepsilon}^{(0)}$. Equivalently, the Jacobi iterative method converges for *any* initial $\mathbf{z}^{(0)}$. What is more, from (10) each component of the error $\boldsymbol{\varepsilon}^{(m)}$ with respect to the orthonormal basis $\{\mathbf{x}^{(j)}\}_{j=1}^{N}$ is reduced per iteration by a factor not exceeding in modulus

$$\rho(B) = 2\cos \pi h/(2+\sigma h^2) = 1 - (\pi^2+\sigma)h^2/2 + O(h^4), \qquad h \downarrow 0. \tag{18}$$

Having just shown that the iterative method of (15) is convergent, it is our sad duty to report that this iterative method can be *extremely* slowly convergent for small h. To illustrate this, let $h=10^{-2}$ and $\sigma=1$. Then, to reduce each component of $\varepsilon^{(0)}$ by 10 would require approximately m iterations, where $[\rho(B)]^m \approx 0.1$, and m turns out to be about 4200. Even for fast computing machines, this would be a slow process, and obviously there is a need for faster iterative methods.

5. The successive overrelaxation iterative method. To introduce just one such faster iterative method, we first express the matrix B of (9) as the sum $L+L^T$, where L is a strictly lower triangular matrix. Multiplying both sides of (14) by the real parameter ω, and then adding $\mathbf{z}$ to both sides of the result yields, after rearrangement,

$$(I - \omega L)\mathbf{z} = \{(1 - \omega)I + \omega L^T\}\mathbf{z} + \omega\mathbf{g}. \tag{19}$$

Since L is strictly lower triangular, then $(I-\omega L)$ is nonsingular for any choice of ω. Thus, we can multiply both sides of (19) on the left by $(I-\omega L)^{-1}$, which gives

$$\mathbf{z} = (I - \omega L)^{-1}[(1 - \omega)I + \omega L^T]\mathbf{z} + \omega(I - \omega L)^{-1}\mathbf{g}. \tag{19'}$$

Simply inserting iteration superscripts then serves to define the successive overrelaxation (SOR) iterative method:

$$\mathbf{z}^{(m+1)} = (I - \omega L)^{-1}[(1 - \omega)I + \omega L^T]\mathbf{z}^{(m)} + \omega(I - \omega L)^{-1}\mathbf{g}, \qquad m \geqq 0. \tag{20}$$

At first glance, this iterative method of (20) looks rather formidable, and would appear to be implicit. From the definition of B in (9), however, this iterative method can also be written equivalently as

$$z_i^{(m+1)} = z_i^{(m)} - \omega[(2 + \sigma h^2)z_i^{(m)} - z_{i-1}^{(m+1)} - z_{i+1}^{(m)} - h^2 f_i]/(2 + \sigma h^2), \tag{21}$$
$$1 \leqq i \leqq N.$$

Starting with $i=1$, $z_0=\alpha$ is fixed by the boundary condition (2), and we can solve for the new component $z_1^{(m+1)}$ in (21) in terms of the old components $z_1^{(m)}$ and $z_2^{(m)}$. Continuing from left to right, we see that this is an *explicit* iterative method.

In analogy with the Jacobi iterative method, we similarly seek the eigenvalues λ of the matrix

$$\mathcal{L}_\omega \equiv (I - \omega L)^{-1}\{(1 - \omega)I + \omega L^T\}. \tag{22}$$

Thus, we seek the roots of $\det\{\lambda I - \mathcal{L}_\omega\} = 0$. Since $\det(I-\omega L)=1$, it follows that

$$0 = \det(I - \omega L)\det(\lambda I - \mathcal{L}_\omega) = \det[(\lambda + \omega - 1)I - \omega(\lambda L + L^T)]. \tag{23}$$

Previously, we explicitly determined the eigenvectors $\mathbf{x}^{(j)}$ and eigenvalues μ_j of the matrix B of (9). In the same manner, one can verify that the vector $\mathbf{w}^{(j)}$, with kth component $w_k^{(j)}$ defined by $\tau^{k/2}\sin(jk\pi h)$, is an eigenvector of $\tau L+L^T$ for any τ, with corresponding eigenvalue $\tau^{1/2}\mu_j$. Consequently, with $\tau=\lambda$ there is a natural pairing from (23) of the eigenvalues λ of $\mathfrak{L}_\omega$ with the eigenvalues μ of B through

$$(\lambda+\omega-1)^2=\lambda\omega^2\mu^2, \tag{24}$$

which is the fundamental result of Young [8]. The real significance of this expression (24) is that it can be used to explicitly determine the real parameter ω which *minimizes* the spectral radius of $\mathfrak{L}_\omega$. Indeed, it is now well known [8] that

$$\min_\omega \rho(\mathfrak{L}_\omega)=\rho(\mathfrak{L}_{\omega_b})=\omega_b-1, \quad \text{where} \quad \omega_b=2/(1+\sqrt{1-\rho^2(B)}). \tag{25}$$

Using the result of (18), we see that

$$\rho(\mathfrak{L}_{\omega_b})=1-2(\sqrt{\pi^2+\sigma})h+O(h^2), \qquad h\downarrow 0. \tag{26}$$

Comparison of the exponents of h in the leading terms of (18) and (26) shows that an *order of magnitude* improvement in convergence rate has been made, simply by the adroit choice of the parameter ω. To illustrate this numerically, again let $h=10^{-2}$ and $\sigma=1$. Then, the least positive integer m such that $[\rho(\mathfrak{L}_{\omega_b})]^m \leqq 0.1$ now turns out to be approximately 35, as compared with 4200. For the slight increase in arithmetic requirements per iteration in passing from (15) to (21), a great over-all improvement in computational efficiency has been achieved.

6. Alternating direction implicit methods. After all is said and done, the solution of the discrete one-dimensional problem $A\mathbf{z}=\mathbf{k}$ in (7) would *not* be determined by means of iterative methods, except, of course, in expository papers on the subject! Rather, straight-forward Gaussian elimination applied to the matrix problem $A\mathbf{z}=\mathbf{k}$ is not only efficient, but it is very stable relative to the growth of rounding errors, thanks to the positive definite tridiagonal nature of A. The Gaussian elimination algorithm for solving this simple problem can be expressed as

$$\begin{cases} w_1=-1/b; & w_i=(-1/(b+w_{i-1})), \quad 2\leqq i\leqq N-1, \quad b\equiv 2+\sigma h^2, \\ g_1=h^2k_1/b; & g_i=(h^2k_i+g_{i-1})/(b+w_{i-1}), \quad 2\leqq i\leqq N, \\ z_N=g_N; & z_i=g_i-w_iz_{i+1}, \quad 1\leqq i\leqq N-1. \end{cases} \tag{27}$$

Our simple one-dimensional problem can be useful once more to us in quickly introducing a particular variant of the alternating direction implicit (ADI) methods, for the discussion above shows that discrete approximations to the one-dimensional problem $-u_{xx}+\sigma u=f$ can be directly solved. Consider then

the second order elliptic partial differential equation

$$-u_{xx}(x, y) - u_{yy}(x, y) + 2\sigma u(x, y) = f(x, y), \qquad 0 < x, y < 1, \tag{28}$$

in the unit square R with Dirichlet boundary conditions

$$u(x, y) = g(x, y), \qquad (x, y) \in \partial R, \tag{29}$$

where g is specified on ∂R, the boundary of R. Writing (28) as

$$[-u_{xx} + \sigma u] + [-u_{yy} + \sigma u] = f, \tag{30}$$

each term in brackets represents a differential operator in one of the space variables. Thus, with

$$\begin{cases} Hu(x_0, y_0) \equiv [(2 + \sigma h^2)u(x_0, y_0) - (u(x_0 + h, y_0) + u(x_0 - h, y_0))]/h^2, \\ Vu(x_0, y_0) \equiv [(2 + \sigma h^2)u(x_0, y_0) - (u(x_0, y_0 + h) + u(x_0, y_0 - h))]/h^2, \end{cases} \tag{31}$$

representing discrete approximations to these differential operators on a uniform mesh $h = 1/(N+1)$, the discrete matrix problem corresponding to (28)–(29) is

$$H\mathbf{u} + V\mathbf{u} = \mathbf{k}, \tag{32}$$

where H and V are real $N^2 \times N^2$ matrices. It is not difficult to see that, with suitable numbering of the mesh points of the square, H and V are direct sums of the particular matrix A of (6). An important application of the discussion above concerning Gaussian elimination is that matrix problems of the form $H\mathbf{x} = \mathbf{s}$ and $V\mathbf{x} = \mathbf{t}$ can be solved *directly* for $\mathbf{x}$, which is to be a basic part of the iterative method to be described. Writing (32) as the pair of equations

$$(H + rI)\mathbf{u} = (rI - V)\mathbf{u} + \mathbf{k}, \qquad (V + rI)\mathbf{u} = (rI - H)\mathbf{u} + \mathbf{k}, \tag{33}$$

we insert iteration superscripts to define the Peaseman-Rachford alternating direction iterative method [4]

$$\begin{aligned} (H + r_m I)\mathbf{u}^{(m+1/2)} &= (r_m I - V)\mathbf{u}^{(m)} + \mathbf{k}, \\ (V + r_m I)\mathbf{u}^{(m+1)} &= (r_m I - H)\mathbf{u}^{(m+1/2)} + \mathbf{k}, \qquad m \geqq 0, \end{aligned} \tag{34}$$

where $\{r_m\}$ is any sequence of positive acceleration parameters. Note that carrying out a single complete iteration of (34) requires the direct solution of matrix equations on horizontal mesh lines, then on vertical mesh lines, hence the name *alternating directions.*

For the particular problem (28)–(29), we shall now show that the iterative method (34) converges for *any* fixed positive acceleration parameter $r > 0$. In analogy with section 4, we first exhibit explicitly the eigenvectors of the $N^2 \times N^2$ matrices H and V. Defining the column vector $\boldsymbol{\alpha}^{(k,l)}$ with N^2 components $\alpha_{i,j}^{(k,l)}$ by

$$\alpha_{i,j}^{(k,l)} = \sin(k\pi ih)\sin(l\pi jh), \qquad 1 \leqq i, j \leqq N, \ 1 \leqq k, l \leqq N, \tag{35}$$

where $\alpha_{i,j}^{(k,l)}$ refers to the component of $\boldsymbol{\alpha}^{(k,l)}$ at the ith column and jth row of the uniform mesh on the unit square, we see from (31) that

$$(36)\qquad \begin{aligned} H\boldsymbol{\alpha}^{(k,l)} &= [4\sin^2(k\pi h/2) + \sigma h^2]\boldsymbol{\alpha}^{(k,l)}; \\ V\boldsymbol{\alpha}^{(k,l)} &= [4\sin^2(l\pi h/2) + \sigma h^2]\boldsymbol{\alpha}^{(k,l)} \qquad 1 \leqq k, l \leqq N. \end{aligned}$$

For any fixed $r>0$, the error vectors $\boldsymbol{\varepsilon}^{(m)}$ associated with (34) evidently satisfy, in analogy with (16),

$$(37)\quad \boldsymbol{\varepsilon}^{(m+1)} = T_r\boldsymbol{\varepsilon}^{(m)},\ m \geqq 0,\ \text{where}\ T_r \equiv (V + rI)^{-1}(rI - H)(rI + H)^{-1}(rI - V).$$

Thus, from (36), we have that $T_r\boldsymbol{\alpha}^{(k,l)} = \lambda_{k,l}\boldsymbol{\alpha}^{(k,l)}$ where

$$\lambda_{k,l} = \frac{[r - (4\sin^2(k\pi h/2) + \sigma h^2)][r - (4\sin^2(l\pi h/2) + \sigma h^2)]}{[r + (4\sin^2(k\pi h/2) + \sigma h^2)][r + (4\sin^2(l\pi h/2) + \sigma h^2)]}, \qquad 1 \leqq k, l \leqq N,$$

and clearly, $|\lambda_{k,l}| < 1$ for any $1 \leqq k, l \leqq N$ since $r>0$. Because the $\boldsymbol{\alpha}^{(k,l)}$ form, after normalization, an orthonormal basis for the associated N^2 dimensional vector space, it follows from $|\lambda_{k,l}| < 1$ that $\boldsymbol{\varepsilon}^{(m)}$ tends to the zero vector as $m \to \infty$ for *any* initial vector $\boldsymbol{\varepsilon}^{(0)}$, which proves that the iterative method of (34) converges for any fixed $r>0$. If we now choose the parameter r to be

$$r^* = \{(4\sin^2(\pi h/2) + \sigma h^2)(4\cos^2(\pi h/2) + \sigma h^2)\}^{1/2},$$

it turns out [6, p. 216] that the spectral radius $\rho(T_{r^*})$ is *exactly equal* to that of the successive overrelaxation iterative method applied to (28)–(29), so that from (26),

$$(38)\qquad \rho(T_{r^*}) = 1 - 2(\sqrt{\pi^2 + \sigma})\,h + O(h^2), \qquad h \downarrow 0.$$

The really interesting results for such ADI methods are obtained by using a sequence of optimally selected positive parameters [7; 6, p. 223], since the use of such parameters can lead to a further order of magnitude improvement in convergence rates. Specifically, with such optimally selected parameters, we merely state that the average error reduction in norm per iteration turns out to be asymptotically

$$(39)\qquad 1 - \frac{(\pi^2 + \sigma)}{\ln(1/h)}, \qquad h \downarrow 0,$$

which should be contrasted with (38). It must be mentioned, however, that the result of (39) depends heavily on the assumption that the matrices H and V have the common set of eigenvectors $\boldsymbol{\alpha}^{(k,l)}$, which is equivalent to the assumption that H and V commute, i.e., $HV = VH$. For cases in which $HV \neq VH$, the results concerning convergence rates of ADI methods are much less complete.

There is a rich and growing literature on the subject of iterative methods; our aim here was to introduce quickly some of these methods. More complete results, beyond the scope of this paper, can be found in [2, 3, 5, 6].

References

1. D. K. Faddeev and V. N. Faddeeva, Computational methods of linear algebra, translated by R. C. Williams, Freeman, San Francisco, 1963.

2. G. E. Forsythe and W. R. Wasow, Finite difference methods for partial differential equations, Wiley, New York, 1960.

3. A. S. Householder, The theory of matrices in numerical analysis, Blaisdell, New York, 1964.

4. D. W. Peaceman and H. H. Rachford, Jr., The numerical solution of parabolic and elliptic differential equations, J. Soc. Indust. Appl. Math., 3 (1955) 28–41.

5. John Todd, (editor) Survey of numerical analysis, McGraw-Hill, New York, 1962.

6. R. S. Varga, Matrix iterative analysis, Prentice-Hall, Englewood Cliffs, N. J., 1962.

7. E. L. Wachspress, Optimum alternating-direction-implicit iteration parameters for a model problem, J. Soc. Indust. Appl. Math., 10 (1962) 339–350.

8. D. M. Young, Iterative methods for solving partial difference equations of elliptic type, Trans. Amer. Math. Soc., 76 (1954) 92–111.