LECTURE 3

Difference Approximations

1. Five-point difference equation. As was stated in Lecture 2, § 1, numerical methods have the great advantage of being applicable in principle to linear partial DE's with *variable* coefficients on *general domains* (but see § 6). This is because one can approximate such DE's by difference equations (ΔE's); the present lecture will be devoted to a discussion of such approximating ΔE's. As was explained in the preface, the resulting "difference methods" (whose discussion will occupy Lectures 3–5) were used almost exclusively to solve elliptic problems on computers until recently.

We begin with the special case of the Laplace DE. It is classic that, knowing the values of a function $u \in C^4(R)$ at the mesh-points $(x_i, y_j) = (ih, jh)$ of a uniform square mesh, the Laplacian of u is approximated with $O(h^2)$ accuracy by the second central difference quotient

$$\text{(1)} \qquad \nabla_h^2 u(x_i, y_j) \doteqdot \frac{1}{h^2}[u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j}].$$

Clearly, $\nabla_h^2 u = (\delta_x{}^2 + \delta_y{}^2)u/h^2$, where $\delta_x{}^2$ and $\delta_y{}^2$ signify second central difference operators.

To compute the *truncation error* $\nabla_h^2 u - \nabla^2 u$, we assume $u \in C^6(R)$ and expand in Taylor series, getting

$$\text{(2)} \qquad h^2\nabla^2 u = (\delta_x{}^2 + \delta_y{}^2)u - (\delta_x{}^4 + \delta_y{}^4)u/12 + O(h^6).$$

Hence, dividing by h^2 and noting that $\delta_x{}^4 u + \delta_y{}^4 u = O(h^4)$, we see that the truncation error in (1) is $O(h^2)$ for any $u \in C^4(R)$.

A function which satisfies $\nabla_h^2 u = 0$ on a uniform mesh is called a "discrete harmonic function" (see § 2). This is evidently equivalent to the condition that its value at each interior mesh-point is the *arithmetic mean* of its values at the four adjacent mesh-points.

More generally, consider the *source problem* in a bounded plane region R. As in Lecture 1, § 3, this problem consists in solving the self-adjoint elliptic DE

$$\text{(3)} \qquad -\nabla \cdot [p(x, y)\nabla u] + q(x, y)u = s(x, y), \qquad p > 0, q \geqq 0,$$

for suitable boundary conditions. This DE may be approximated at the interior mesh-points of any rectangular mesh by the following five-point central *difference equation* (ΔE):

$$\text{(4)} \qquad D_{i,j}u_{i,j} = L_{i,j}u_{i-1,j} + R_{i,j}u_{i+1,j} + T_{i,j}u_{i,j+1} + B_{i,j}u_{i,j-1} + q_{i,j}u_{i,j} + S_{i,j},$$

where

$$(4') \qquad \begin{aligned} D_{i,j} &= R_{i,j} + L_{i,j} + T_{i,j} + B_{i,j}, \quad \text{with} \\ R_{i,j} &= -[p(x_{i+1}, y_j) + p(x_i, y_j)]/2(x_{i+1} - x_i); \end{aligned}$$

and $L_{i,j}$, $T_{i,j}$ and $B_{i,j}$ are given by similar formulas [FW, p. 201], [V, p. 186]. As we shall see in § 5, the error in the approximation (4) to the DE (3) is $O(h^2)$ if a uniform mesh is used, but only $O(|\Delta x_i|)$ if a nonuniform mesh is used. That is, an order of accuracy is lost in passing from the case of *constant* mesh-length to the general case of *variable* mesh-length.

These approximations "reduce" the analytical source problem defined by the DE (3) and the Dirichlet boundary condition $u(x, y) = f(x, y)$ on ∂R, the boundary of R, to an approximately equivalent *algebraic* problem of solving a system of n simultaneous linear equations in n unknowns, where n is the number of interior mesh-points. In vector notation, this algebraic problem consists in solving a vector equation of the form

$$(5) \qquad A\mathbf{u} = \mathbf{b}.$$

Here $\mathbf{u}$ is the vector of unknown values of the $u_{i,j}$ at interior mesh-points, while A and $\mathbf{b}$ (the vector of boundary values and source terms) are known.

To solve large systems (5) of simultaneous linear equations efficiently and accurately is not easy; techniques for doing this when A is a $10^4 \times 10^4$ matrix, say, will be the main theme of my next two lectures. The success of such techniques depends basically on a number of special properties of A—and especially on the fact that A is a Stieltjes matrix whose off-diagonal entries form a 2-cyclic matrix, in the sense of the following definitions.

DEFINITION. A *Stieltjes matrix* is a symmetric matrix whose diagonal elements are positive, whose off-diagonal elements are negative or zero, and which is positive definite [V, p. 85]. An $n \times n$ matrix B is *2-cyclic* (or has "Property A") when its indices can be partitioned into two nonvoid subsets S and T such that $b_{kl} \neq 0$ implies $k \in S$ and $l \in T$ or vice versa.

We shall now verify that the square matrix $A = \|a_{kl}\|$ of coefficients[1] of the system (5) is *symmetric*, since

$$R_{i,j} = L_{i+1,j} = c_{i+1/2,j}, \quad T_{i,j} = B_{i,j+1} = c_{i,j+1/2}.$$

Next, decompose A as follows:

$$(6) \qquad A = D - E - F, \quad F = E^T,$$

where D is its diagonal component, $-E$ its subdiagonal component, and $-F$ its superdiagonal component. All three matrices D, E, F are nonnegative. Moreover, in particular, (i) A has positive diagonal entries and negative or zero off-diagonal entries ($-A$ is "essentially nonnegative"); also (ii) A is *diagonally dominant*, in the sense that each (positive) "diagonal" coefficient $D_{i,j}$ in (4) is equal to the sum of the magnitudes of all the other coefficients in its row, and greater than this sum for points where $q > 0$, and adjacent to points satisfying boundary conditions $u = g(\mathbf{x})$ or $\partial u/\partial n + b(\mathbf{x})u = g(\mathbf{x})$, $b(\mathbf{x}) > 0$; (iii) A is *positive definite*, i.e., $\mathbf{x}A\mathbf{x} > 0$ unless $\mathbf{x} = 0$; and (iv) it is a *Stieltjes matrix* with strictly positive inverse; finally, (v) A is *sparse* in that it has at most 4 nonzero off-diagonal entries in each row, and (vi) it is 2-*cyclic* since one can take for S those $k = (i,j)$ with $i + j$ odd, and for T the $k = (i,j)$ with $i + j$ even. For a sufficiently fine mesh in a connected domain with smooth boundary, (vii) the matrix A is also *irreducible*.

[1] Here each index (k or l) stands for a mesh-point (i, j).

2. Network analogies. Solutions of linear systems having a Stieltjes coefficient-matrix are of interest not only as approximate solutions of problems of continuum physics; they also represent exact solutions to interesting *network problems* arising in various branches of physics.

Specifically, let $C = \|c_{kl}\|$ be any Stieltjes matrix. We can construct a D.C. *network* whose kth node is connected with its lth node by a wire of conductance $-c_{kl}$ if $c_{kl} < 0$, and is not connected with node l when $c_{kl} = 0$ (i.e., otherwise). We let each jth node have an input lead with controlled current S_j and a resistive connection to "ground" with conductance $c_{jj} - \sum_k c_{jk} \geqq 0$. Then Kirchhoff's laws are equivalent to the vector equation

$$(7) \qquad S_j = \sum_k c_{jk}(u_j - u_k) = \sum I_{jk} = -\sum I_{kj}.$$

The sparseness of C is reflected in a sparseness of links.

By inspection, we find that the 5-point difference approximation (4) to $-\nabla \cdot (p\nabla u) = f(x, y)$ leads to a rectangular D.C. network, whose nodes are the mesh-points and whose conducting elements are the mesh-segments. In this *network analogy*, $u_{i,j}$ is the voltage at the terminal (i,j), $R_{i,j} = L_{i+1,j} = c_{i+1/2,j}$ in (6) is the conductance of the wire connecting node (i,j) to node $(i + 1, j)$, $s_{i,j}$ is the current flowing into node (i,j), and so on. As a result of this analogy, one can build rectangular networks for solving the difference equations (4) by analogy. (Similarly, one can use an electrolytic tank or telegraphic "teledeltos" paper as analogue computer to solve the DE $\nabla^2 u = 0$.)

A mechanical analogy is provided by locating taut strings under constant tension T on the mesh-lines of a rectangular network, loaded at the mesh-points where these lines are joined, and looking for static equilibrium (minimum strain energy): the stationary state of minimum strain energy, with $\delta J = 0$ for

$$(8) \qquad J = \tfrac{1}{2}(\mathbf{u}, A\mathbf{u}) - \mathbf{b} \cdot \mathbf{u}.$$

This analogy suggested to Hardy Cross and to R. V. Southwell the idea of solving the resulting equations (i.e., of minimizing J) by iterative "relaxation" methods to be described in Lectures 4 and 5, in which J is repeatedly reduced by changing one $u_{k,l}$ at a time.

Though using variable mesh-length and nonrectangular meshes ("irregular stars" [9]) improve the accuracy of the network analogy in regions where the exact solution is rapidly varying, they also greatly complicate the writing of computer programs for solving the resulting systems of linear algebraic equations.

Discrete harmonic functions. The difference approximation (1) on a uniform mesh also defines a fascinating class of "discrete harmonic functions" $u(i,j)$, defined as solutions of the ΔE

$$u(i,j) = \tfrac{1}{4}[u(i+1,j) + u(i,j+1) + u(i-1,j) + u(i,j-1)]. \tag{9}$$

Discrete harmonic functions have been extensively studied by Duffin and others.

Note that (9) is analogous to Gauss' theorem of the arithmetic mean. Again, solutions of (9) minimize the "Dirichlet sum" $\sum (\Delta u)^2$ of the squares of difference ("jumps" in u between adjacent mesh-points) for given boundary values.

3. Solution by elimination. When the number N of mesh-points is moderate (when $N < 1000$, say), it is usually feasible to solve the system of ΔE's $A\mathbf{v} = \mathbf{k}$ by Gaussian elimination in single-precision arithmetic. However, this involves approximately $N^3/3$ multiplications [FW, Chap. 25], as well as storing up to $N^2 \sim 10^6$ numerical coefficients. The situation is very different from that in the one-dimensional case, in which the 3-point $O(h^2)$ approximation leads to only a *tridiagonal* matrix.

If the number of mesh-points on any horizontal line is bounded by M, then the matrix A is a *band matrix* with bandwidth at most $2M + 1$. Gaussian elimination then requires only about M^2N multiplications.[2]

Alternatively, one can regard the ΔE (1) (for example) as a *two-endpoint problem* for a two-level system of M second order ΔE's:

$$u_{j+1}(i) = 4u_j(i) - u_{j-1}(i) - u_j(i+1) - u_j(i-1), \tag{10}$$

which can then be integrated using "multiple shooting" techniques. These have been studied by H. B. Keller[3] and others. However, the DE (10) is unstable, and this approach may well lead to a need for double precision [FW, loc. cit.].

Optimal elimination. Reduction to minimum bandwidth is only one of several techniques which have been developed for exploiting the sparseness of matrices arising from ΔE's and network problems. Reduction to minimum bandwidth does not always minimize the work of achieving exact solutions (in "rational arithmetic"): it is by no means always optimal. Indeed, the whole subject of optimizing elimination for sparse matrices is currently a very active research area; I can only give you a few major references.[4]

[2] For more details, see G. E. Forsythe and C. B. Moler, *Computer Solution of Linear Algebraic Systems*, Prentice-Hall, Englewood Cliffs, New Jersey, 1967, pp. 115–119.

[3] *Two-Endpoint Problems*, Blaisdell, Waltham, Massachusetts, 1968.

[4] See D. V. Steward, SIAM J. Numer. Anal., 2 (1965), pp. 345–365; R. A. Willoughby, editor, *Sparse Matrix Proceedings*, RA-1, IBM Res Publ., March, 1969; [4], [8], and Part D of my article to appear in Proc. SIAM–AMS Symp. IV (1971).

A very special elimination method, which is brilliantly successful for solving the Poisson equation $-\nabla^2 u = f$ in rectangular regions is the Tukey–Cooley fast Fourier transform on a uniform $2^m \times 2^n$ mesh.[5]

However, for most very large problems ($N > 10{,}000$, say) in general regions, and especially for those which involve multiple interfaces such as occur in nuclear reactors, stable and self-correcting *iterative* methods seem to be preferable. My next two lectures will be largely devoted to iterative and semi-iterative methods for solving large systems of simultaneous linear equations. These have the further advantage of being more readily adaptable to nonlinear problems.

4. Nonlinear problems. By the simple device of replacing derivatives by (approximately equal) difference quotients, *nonlinear* DE's can also be approximated by (nonlinear) systems of algebraic equations.

Methods for solving the resulting systems of nonlinear equations are typically iterative, beginning with *Newton's method*, which is the method most commonly proposed in textbooks. (The usual expositions of this method take for granted the triviality of solving *linear* systems, incidentally.)

For this reason, I shall postpone the study of (iterative) methods for solving systems of nonlinear algebraic equations to Lecture 4 (and to Lecture 8, § 3); their success for large systems usually depends on quite special considerations.

Nonlinear networks. For example, they may depend on variational properties, such as hold for a wide class of nonlinear networks[6] analogous to the linear networks discussed in § 2. From this principle, one can derive existence and uniqueness theorems for flows.

5. Local truncation errors. For the rest of this lecture, I shall ignore the practical difficulties of solving accurately large systems of algebraic equations, and describe what is known about the *accuracy* of difference approximations, assuming that the difference equations can be solved.

As I said in § 1, the 5-point central difference quotient approximation for $\nabla^2 u$ on a uniform mesh introduces an error of $O(h^2)$ at each mesh-point. Unfortunately, its generalization (4) to a nonuniform mesh (or even with a uniform mesh unless (3) has constant coefficients) introduces an error of $O(h)$. Moreover, this order of accuracy is "best possible": with only five mesh-points, one cannot match more than the five coefficients corresponding to u, u_x, u_y, u_{xx} and u_{yy} in the Taylor series expansion of u. It is sheer luck when other derivatives have no influence. Indeed, one cannot express u_{xy} even approximately in terms of the 5 values of u in (3). For this reason, difference approximations to elliptic problems in which

[5] R. W. Hockney, J. Assoc. Comput. Mach., 12 (1965), pp. 95–113; F. W. Dorr, SIAM Rev., 12 (1970), pp. 248–263; B. L. Buzbee et al., SIAM J. Numer. Anal., 7 (1970), pp. 623–656.

[6] G. Birkhoff and J. B. Diaz, Quart. Appl. Math., 13 (1956), pp. 432–443; see also G. Birkhoff and R. B. Kellogg, Proc. Symp. Generalized Networks, MRI Symposium Series 16, Brooklyn Polytechnic Press, New York, 1966, and the references of Lecture 8, footnote 8.

u_{xy} enter normally use a 9-point formula since, for example,

$$h^2u_{xy} = u_{i+1,j+1} + u_{i-1,j-1} - u_{i+1,j-1} - u_{i-1,j+1} + O(h^4). \tag{11}$$

Instead of using truncated Taylor series to derive difference approximations to derivatives, one can use integral formulas. Careful discussions of this approach may be found in [FW], [KK], [V] and [W].

In either case, the most useful fact to be deduced from such a priori error estimates is the principle that the error (for a uniform mesh) is typically asymptotic to $Mh^n + O(h^{n+1})$ for some positive integer n.

Order of convergence. For square meshes with mesh-length h, the truncation error is typically of the form $Mh^n + O(h^{n+1})$ for some positive integer n, the "order of convergence." More generally, this is true of rectangular meshes with mesh-length $h\theta_k$ in the x_k-direction, and in many other cases. In such cases, the changes $\Delta u_{i,j}$ in computed values when the mesh-length is halved from h to $h/2$ are approximately proportional to h^n. Though M is unknown, one can use Richardson's method of "deferred approach to the limit" [FW, p. 307] to improve the accuracy of results obtained by mesh-halving (until roundoff takes over). See also [11].

6. Higher order accuracy. One can always approximate difference quotients of very smooth functions with higher order accuracy by using stencils with enough mesh-points; this follows from Taylor's formula. In the case of partial DE's with constant coefficients and a uniform mesh, the process yields some very elegant (and sometimes useful) formulas. I shall mention a few such formulas, giving references[7] and assuming high order differentiability.

Thus, formula (2) leads to a difference approximation

$$\nabla^2 u = \frac{1}{6h^2}[4\delta^2 u + \tilde{\delta}^2 u] + O(h^4),$$

where

$$\tilde{\delta}^2 u = [u_{i+1,j+1} + u_{i+1,j-1} + u_{i-1,j+1} + u_{i-1,j-1} - 4u_{ij}]$$

having $O(h^4)$ accuracy on a 9-point *square* of mesh-points (see [KK, p. 179] and J. Bramble and B. Hubbard [2]). This is not to be confused with the difference approximation

$$\nabla^2 u = \bar{\delta}_{xx}u + \bar{\delta}_{yy}u + O(h^4),$$

where

$$\bar{\delta}_{xx}u = [16(u_{i+1} + u_{i-1}) - (u_{i+2} + u_{i-2}) - 30u_0]/24h^2$$

on a 9-point *cross* of mesh-points [KK, p. 184], obtained by minimizing the Dirichlet integral on the piecewise bilinear function interpolated between values of u at these mesh-points.[8] This 9-point difference approximation with $O(h^4)$ accuracy applies also to DE's of the form $Au_{xx} + Cu_{yy} = 0$ and to

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = 0$$

if $A = C$ or $B = 0$.[9]

One can obtain a difference approximation to $\nabla^2 u$ having $O(h^{10})$ accuracy by using a 13-point stencil [KK, p. 184], while 17-point stencils for $\nabla^2 u$ and 25-point stencils for $\nabla^4 u$ have also been worked out.[10]

Finally, accurate difference approximations to ∇^2 on triangular and hexagonal nets have been worked out by various authors.[11] Unfortunately, although such higher order methods are intriguing, the use of the associated larger stencils almost invariably leads to serious complications near the boundary.

7. Global error bounds. The errors referred to in § 1 and § 6 were discrepancies between difference quotients ("divided differences") and derivatives. The question arises: how are such errors related to those in the *values* of the functions? If we write the difference approximation in the form $A\mathbf{v} = \mathbf{k}$, then we have $A\mathbf{u} = \mathbf{k} + \mathbf{r}$, where $\mathbf{r}$ is the vector whose components r_i are these discrepancies. The r_i are also called *residuals* for the system $A\mathbf{u} - \mathbf{k}$.

For the Laplace equation, and in some other cases, one can achieve higher order (local) accuracy a posteriori by estimating the r_i from numerical data (e.g., by estimating $\nabla^4\mathbf{u}$ from the computer printout). If $\bar{\mathbf{r}}$ is the estimated dominant error term, then by subtracting the solution of $A\mathbf{v} = \bar{\mathbf{r}}$ as a "differential correction" from the solution of the difference approximation, one should reduce the error. This is Fox's "method of differential corrections" [5].[12]

Discrete Green's function. Alternatively, one can combine remainder formulas with a priori knowledge of the derivatives of the exact solution, obtained by analytic considerations (cf. Lecture 2), to bound the residuals r_i. Since the actual error vector $\mathbf{e} = \mathbf{v} - \mathbf{u}$ satisfies $\mathbf{e} = G\mathbf{r}$ where $G = A^{-1}$, this leads to an *a priori error bound* in terms of the norm of G. Here G may be called the *Green's matrix* because it acts like a *discrete Green's function* [FW, pp. 315–318] for the source problem being solved. It is a positive matrix for (4). Finally, again using analytical considerations discussed in Lecture 2, one can often bound the norm of G.

[7] A useful compendium is contained in Collatz, Table VI [C, pp. 505–509]; see also W. G. Bickley et al., Proc. Roy. Soc. London, A262 (1961), pp. 219–236.

[8] R. Courant, Bull. Amer. Math. Soc., 49 (1943), pp. 1–27; B. Epstein, Math. Comp., 16 (1962), pp. 110–112.

[9] J. Bramble and B. Hubbard, Contributions to Differential Equations, 2 (1963), pp. 319–340; Young and Dauwalder, Rep. TNN-46, Univ. of Texas Comp. Lab.

[10] B. Meister and W. Prager, Z. Angew Math. Phys., 16 (1965), pp. 403–410; see also G. Fairweather et al., Numer. Math., 10 (1967), pp. 56–66; A. Hadjimos, Ibid., 13 (1969), pp. 396–403; and F. D. Burgoyne, Math. Comp., 22 (1968), pp. 589–594.

[11] [KK, pp. 187–188]; R. B. Kellogg, Math. Comp., 18 (1964), pp. 203–210; [C]; [9]; D. N. de G. Allen, *Relaxation Methods in Engineering and Science*, McGraw-Hill, New York, 1954. I. Babuška, M. Prager and M. Vitasek, *Numerical Processes in DE's*, SNTL-Interscience, 1966, § 5.4.2.

[12] See also [Az, p. 203], and E. A. Volkov, Vychisl. Mat., 1 (1957), pp. 34–61 and 62–80.

Using such considerations, global convergence as $h \downarrow 0$ was first proved for the Laplace ΔE on a square mesh by R. G. D. Richardson in 1917 and by Phillips and Wiener in 1922; the aim of these authors was to establish *existence theorems* for solutions of the Dirichlet problem for $\nabla^2 u = 0$ from algebraic existence theorems for $\nabla_h^2 u = 0$. In 1927, Courant, Friedrichs and Lewy showed that all difference quotients of given order converged to the appropriate derivatives, as $h \downarrow 0$.

The maximum principle of Lecture 2, § 3, was applied to the Poisson equation by Gerschgorin [6] in 1930 to prove $O(h)$ global accuracy. Using linear interpolation on the boundary, Collatz[13] sharpened this result in 1933, under appropriate differentiability assumptions, to prove $O(h^2)$ accuracy. Further work was also done by Walsh and Young and by Wasow in 1954–5, and by P. Laasonen, who discussed carefully the loss of accuracy introduced by corners, where local singularities occur.[14] This literature is reviewed in [FW, p. 302], and in [C, pp. 326–327]. When mesh-points on the boundary are extremely close together, errors can be greatly magnified. A way to resolve this difficulty has been described by Babuška, Prager and Vitasek (op. cit., p. 274).

The whole subject was carefully reconsidered by Bramble and Hubbard, who used the Green's function approach systematically. They published their results in a series of papers written in 1964–5, especially in [1]–[2] and the references given there.[15] A significant question is whether or not A must be "monotone," i.e., whether the inverse G of A needs to be nonnegative. On this point, see [3] and recent work by Harvey Price.[16] The preceding authors have shown that, by using higher order differences, one can obtain higher order accuracy (for $\nabla^2 u = f$ and $\nabla^4 u = f$ on a square mesh).

The accuracy of the 5-point difference approximation with variable coefficients has been studied by Bramble, Hubbard and Thomée,[17] under weakened assumptions of smoothness on the boundary. For $u \in C^4(R) \cap C^2(\bar{R})$, for example, one obtains $O(h^2)$ accuracy. Finally, the $O(h^2)$ convergence of *all* difference quotients to the appropriate derivatives has been proved for the Laplace DE on a square mesh by V. Thomée and Achi Brandt.[18] Making increased smoothness assumptions, Thomée also showed that difference quotients converge at the same rate as the solution in the interior (giving discrete Harnack-type inequalities).

Many other more general results have been proved. Thus V. Thomée has proved convergence to order $O(h^{1/2})$ for simple difference approximations to the Dirichlet problem for any linear, constant-coefficient equation of elliptic type, and McAllister has obtained global error bounds for difference approximations to certain mildly nonlinear elliptic problems.[19] Finally, Bramble [BV, pp. 201–209] has shown that by appropriately smoothing f, one can get improved convergence of difference approximations to $L[u] = f$, for uniformly elliptic L.

[13] L. Collatz, Z. Angew Math. Mech., 13 (1933), pp. 56–57.

[14] See [7]; J. Assoc. Comput. Mach., 5 (1958), pp. 32–38; also E. Batschelet, Z. Angew Math. Phys., 3 (1952), pp. 165–193; N. M. Wigley, SIAM J. Numer. Anal., 3 (1966), pp. 372–383.

[15] Including Contributions to Differential Equations, 2 (1963), pp. 229–252; 3 (1963), pp. 319–340; SIAM J. Numer. Anal., 2 (1965), pp. 1–14; J. Assoc. Comput. Mach., 12 (1965), pp. 114–123; Numer. Math., 4 (1962), pp. 313–332; Ibid., 9 (1966), pp. 236–249.

[16] H. Price, Math. Comp., 22 (1968), pp. 489–516.

[17] BIT, 8 (1968), pp. 154–173. See also N. S. Bahalov, Vestnik Moskov Univ., 5 (1959), pp. 171–195, and J. R. Kuttler, SIAM J. Numer. Anal., 7 (1970), pp. 206–232.

[18] Math. Comp., 20 (1966), pp. 473–499. See also P. G. Ciarlet, Aequat. Math., 4 (1970), pp. 206–232.

REFERENCES FOR LECTURE 3

[1] J. H. Bramble and B. E. Hubbard, *Approximation of derivatives by difference methods in elliptic boundary value problems*, Contributions to Differential Equations, 3 (1964), pp. 399–410.

[2] ———, *New monotone type approximations for elliptic problems*, Math. Comp., 18 (1964), pp. 349–367.

[3] ———, *On a finite difference analogue of an elliptic boundary problem which is neither diagonally dominant nor of non-negative type*, J. Math. and Phys., 43 (1964), pp. 117–132.

[4] G. E. Forsythe and C. B. Moler, *Computer Solutions of Linear Algebraic Systems*, Prentice-Hall, Englewood Cliffs, New Jersey, 1967.

[5] L. Fox, *Some improvements in the use of relaxation methods for the solution of ordinary and partial differential equations*, Proc. Roy. Soc. London Ser. A, 190 (1947), pp. 31–59. (See also Philos. Trans. Roy. Soc. London. Ser. A, 242 (1950), pp. 345–378; Quart. J. Mech. Appl. Math., 1 (1948), pp. 253–280.)

[6] G. Gerschgorin, *Fehlerabschätzung für das Differenzenverfahren . . .*, Z. Angew. Math. Mech., 10 (1930), pp. 373–382.

[7] P. Laasonen, *On the degree of convergence of discrete approximations for the solutions of the Dirichlet problem*, Ann. Acad. Sci. Fenn. Ser. A, 246 (1957), 19 pp.

[8] A. Ralston and H. S. Wilf, editors, *Numerical Methods for Digital Computers*, vol. II, John Wiley, New York, 1967.

[9] R. V. Southwell, *Relaxation Methods in Theoretical Physics*, Clarendon Press, Oxford, 1946.

[10] J. Nitsche and J. C. C. Nitsche, *Error estimates for the numerical solution of elliptic differential equations*, Arch. Rational Mech. Anal., 5 (1960), pp. 293–306; pp. 307–314.

[11] V. Pereyra, *Accelerating the convergence of discretization algorithms*, SIAM J. Numer. Anal., 4 (1967), pp. 508–533. (See also Numer. Math., 8 (1966), pp. 376–391, and 11 (1968), pp. 111–125.)

[19] V. Thomée, Contributions to Differential Equations, 3 (1964), pp. 301–324; G. T. McAllister, J. Math. Anal. Appl., 27 (1969), pp. 338–366.

LECTURE 4

Relaxation Methods

1. Point-Jacobi method. This lecture and the next will be devoted to *iterative* and *semi-iterative* methods for solving systems of linear equations (vector equations) of the form

$$Au = \mathbf{b}. \tag{1}$$

For very large systems involving 10^4 unknowns, these are usually more efficient than the elimination methods described in Lecture 3, § 3 (see § 7).

A great variety of such methods have been proposed; those involving "relaxation methods" are especially applicable when A is a *Stieltjes matrix* of the form

$$A = D - E - F, \quad F = E^T. \tag{2}$$

As was shown in Lecture 3, § 2, such matrices arise naturally from D.C. *network problems*, including those which correspond to the 5-point difference approximation to a source problem (with or without leakage). As we saw in that section, they also arise from the usual difference approximation to *second order* self-adjoint elliptic DE's of the form $-\nabla \cdot (p\nabla u) + qu = f$. When applied to such problems, many iterative methods are suggested by concepts of *relaxation* or *overrelaxation*, which may be motivated as follows.

The solution of (1) is that (column) vector $\mathbf{u}$ which minimizes the (positive definite) quadratic functional

$$J(\mathbf{u}) = \tfrac{1}{2}\mathbf{u}^T A\mathbf{u} - \mathbf{b} \cdot \mathbf{u}. \tag{3}$$

In the loaded membrane physical interpretation, this functional is just the total *potential energy* of the system.

Example 1. For the 5-point discretization of $-\nabla^2 u = f(\mathbf{x})$, the functional to be minimized is the sum

$$\tfrac{1}{2}\sum [(u_{i,j} - u_{i,j-1})^2 + (u_{i,j} - u_{i-1,j})^2] + \sum f_{i,j}u_{i,j}.$$

As was explained in § 2 of the previous lecture, one may simplify interpretation of relaxation methods by thinking of $J(\mathbf{u})$ as the "strain" energy of a configuration $\mathbf{u}$, whose coordinates u_j are "relaxed" cyclically so as to reduce J at each step. Finding this minimum by successively "relaxing" components u_j at $\mathbf{u}$, so as to reduce $J(\mathbf{u})$, is a simple way of looking at relaxation methods. This was also the idea of Poincaré's "méthode de balayage" for solving the Dirichlet problem, which likewise reduces the Dirichlet integral at each sweep.[1]

If one "scales" the quantities b_i, replacing them by $d_i^{-1}b_i = k_i$, the equation (1) is premultiplied by the diagonal matrix D^{-1}, which transforms it to the equivalent vector equation

$$u = D^{-1}(E + F)\mathbf{u} + D^{-1}\mathbf{b}.$$

This suggests the iterative process

$$\mathbf{u}^{(n+1)} = D^{-1}(E + F)\mathbf{u}^{(n)} + D^{-1}\mathbf{b}, \tag{4}$$

which *is* the *point-Jacobi* method, also called the "method of simultaneous displacements."

If A is a Stieltjes matrix, the point-Jacobi method always converges [V, Theorems 3.3 and 3.6]. In particular, it converges for the matrix problems associated with any connected (irreducible) network, except when the current is specified at all boundary nodes, and there is no leakage.

2. Rate of convergence. Not only the fact of convergence but the *rate* of convergence is of crucial importance for an iterative method. For (4), this depends on the spectrum of $D^{-1}(E + F)$, which may also be written as

$$B = D^{-1}(E + F) = D^{-1/2}[D^{-1/2}(E + F)D^{-1/2}]D^{-1/2}.$$

In this notation, (4) simplifies to

$$\mathbf{u}^{(n+1)} = B\mathbf{u}^{(n)} + \mathbf{k}, \quad B = D^{-1}(E + F), \quad \mathbf{k} = D^{-1}\mathbf{b}. \tag{5}$$

Since B is similar to a symmetric matrix $D^{-1/2}(E + F)D^{-1/2}$ (is "symmetrizable"), all eigenvalues of B are *real*.

In general, the matrix A underlying the point-Jacobi method for any well-designed difference approximation to a *self-adjoint* elliptic boundary value problem should be *symmetric*. Hence it and B should be similar to a *real diagonal* matrix.

Spectral radius. We now consider in some detail the questions of the *convergence* and the asymptotic *rate of convergence* of the point-Jacobi iterative method (4). The relevant concept is the *spectral radius* of B, $\rho(B)$. This is defined as the maximum of the magnitudes (absolute values) of the eigenvalues λ_i of $B: \rho(B) = \max |\lambda_i(B)|$.

By considering the Jordan canonical form $J = PBP^{-1}$ of B (P nonsingular), which is real and diagonal in the present case, it is easy to prove that (5) gives for any $\mathbf{u}^{(0)}$ a sequence of $\mathbf{u}^{(n)}$ which converge as $n \to \infty$ to the (unique) solution $\mathbf{u}$. In fact, the *error* $\mathbf{e}^{(n)} = \mathbf{u}^{(n)} - \mathbf{u}$ satisfies $\mathbf{e}^{(n)} = B^n\mathbf{e}^{(0)}$. When $\rho(B) < 1$, the norm of the error thus tends to zero, asymptotically like $[\rho(B)]^n$. Hence, the *asymptotic rate of* convergence as $n \to \infty$ is asymptotically proportional to $-\log \rho(B)$ if $\rho(B) < 1$;[2] if $\rho(B) \geqq 1$, the method fails to converge.

[1] H. Poincaré, Amer. J. Math., 12 (1890), pp. 216–237; [K, p. 283].

[2] In the sense that, asymptotically, the error decreases by a factor e every $1/(-\log \rho(B))$ iterations (cf. [FW, p. 218]).

Remark. In some cases, one can interpret (4) as the Cauchy polygon method for integrating $\mathbf{u}_t = -A\mathbf{u} + \mathbf{k}$ with a small time step (see [V, § 8.4]). Thus, this is true for the usual 5-point approximation to $-\nabla^2 u = f$ on a uniform mesh; in this case, the point-Jacobi method (4) gives the Schmidt process for integrating the heat equation with source, $u_t = \nabla^2 u + f$.

More generally, all eigenvalues of A are *positive* for suitable mathematical models of most source problems, "passive" D.C. electrical networks, and other conservative or dissipative physical systems in the linear (small amplitude) range, including those of elasticity. Hence, $\mathbf{u}^{(r+1)} = \mathbf{u}^{(r)} + \Delta t(A\mathbf{u}^{(r)} - \mathbf{k})$ is convergent for sufficiently small Δt. If one takes the eigenvectors of A for coordinate axes, one can interpret (4) as integrating the system $dv_i/dt = -\lambda_i v_i + g_i$, where the λ_i are the (positive) eigenvalues of A.

The optimum Δt depends on the ratio $\lambda_{\max}/\lambda_{\min}$. In general, this is not easy to estimate, but see § 4.

3. Gauss–Seidel method. The point-Jacobi method yields every component of $\mathbf{u}^{(n+1)}$, the $(n + 1)$st approximation to the solution vector of (1), as a (linear) function of components of $\mathbf{u}^{(n)}$, which are nearby in the case of difference schemes.

Thus, for the 5-point approximation to the Laplace DE, the point-Jacobi scheme at interior mesh-points is

$$u_{i,j}^{(n+1)} = \tfrac{1}{4}[u_{i+1,j}^{(n)} + u_{i-1,j}^{(n)} + u_{i,j+1}^{(n)} + u_{i,j-1}^{(n)}]. \tag{6}$$

Alternatively, sweeping through the components cyclically, one can *use improved values as soon as available*. Thus, for the natural ordering of mesh-points, one can use

$$u_{i,j}^{(n+1)} = \tfrac{1}{4}[u_{i+1,j}^{(n)} + u_{i-1,j}^{(n+1)} + u_{i,j+1}^{(n)} + u_{i,j-1}^{(n+1)}]. \tag{7}$$

The resulting method is called the *Gauss–Seidel method* (also the method of "successive displacements"). In general, the (point) Gauss–Seidel method is defined for (4) by

$$\mathbf{u}^{(n+1)} = (D - E)^{-1}F\mathbf{u}^{(n)} + (D - E)^{-1}\mathbf{b}. \tag{7'}$$

It requires only half as much storage as point-Jacobi.

Stein–Rosenberg theorem. A very general theorem, due to Stein and Rosenberg, asserts that the preceding Gauss–Seidel method converges at least as fast as the point-Jacobi method. The proof depends only on the fact that the iteration matrix B is *nonnegative* with *zero diagonal entries*; thus B need not be symmetrizable for it to apply.

In the 2-*cyclic* case of § 5 (e.g., for (7)), Gauss–Seidel converges exactly *twice* as fast as Jacobi for any given B. This is a fundamental result of David Young [V, p. 107].

The stopping problem. With iterative methods, a basic question is when to stop iterating. Criteria may be given in terms of either $\|\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)}\|$ (in any norm) or, better, of the residual $\|A\mathbf{u} - \mathbf{b}\|$ and the rate of convergence of the process. We shall not enter into this question, beyond noting that, for ill-conditioned matrices, roundoff can pose surprising problems with Gauss–Seidel iteration.

Thus Wilkinson (J. Assoc. Comput. Mach., 8 (1961)) takes

$$A = \begin{bmatrix} .96326 & -.81321 \\ -.81321 & .68654 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} .88824 \\ -.74968 \end{bmatrix},$$

with the initial trial $\mathbf{x}^{(1)} = \begin{bmatrix} 0 \\ -.7 \end{bmatrix}$. Then, to five decimal digits,

$$\mathbf{x}^{(2)} = \mathbf{x}^{(3)} = \cdots = \begin{bmatrix} .33116 \\ -.70000 \end{bmatrix},$$

yet $A^{-1}\mathbf{b} = \begin{bmatrix} .39473\ldots \\ -.62470\ldots \end{bmatrix}$. Professor Moler kindly called this example to my attention.

4. Rates of convergence. Although the spectral radius $\rho(B)$ plays a central role in the theory of the rate of convergence of iterative methods, it is very hard to compute accurately. In practice, $\rho(B)$ must usually be estimated from numerical experiments (see § 7). However, there are a few exceptional "model problems" in which not only $\rho(B)$ but the entire spectrum is known.

Example 2. The eigenfunctions of the Laplace and Poisson equations in the rectangle $[0, a] \times [0, b]$ are $\sin(j\pi x/a)\sin(k\pi y/b)$. If this rectangle is subdivided by a uniform mesh into $M \times N$ subrectangles, the values of any of the above eigenfunctions at mesh-points define an eigenvector for the 5-point difference approximation to $-\nabla^2$. The case of a square (and the approximation $-\nabla_h^2$) is typical; the eigenvalues (for $j = 1, \cdots, M - 1$ and $k = 1, \cdots, N - 1$) are

$$4[\sin^2(j\pi/2M) + \sin^2(k\pi/2N)];$$

they range from $\lambda_{\min} = 4[\sin^2(\pi/2M) + \sin^2(\pi/2N)]$ to $2 - \lambda_{\min}$. Those of B range from $\lambda_{\min} - 1$ to $1 - \lambda_{\min}$; hence the spectral radius of the corresponding point-Jacobi iteration matrix is

$$\rho(B) = \tfrac{1}{2}[\cos(\pi/M) + \cos(\pi/N)] = 1 - \sin^2(\pi/M) - \sin^2(\pi/N)$$

and the eigenvalues of A correspondingly are $\pi^2[(j/a)^2 + (k/b)^2]$, $j = 1, \cdots, M$, $k = 1, \cdots, N$. Hence the eigenvalues of A range from $\sin^2(\pi/M) + \sin^2(\pi/N) = \lambda_{\min}$ to $\sin^2[(M - 1)\pi/M] + \sin^2[(N - 1)\pi/N] = 2 - \lambda_{\min}$.

Similar formulas can be written whenever $A = A' \otimes A''$ is a tensor product of tridiagonal matrices: $\lambda_{kl}(A) = \lambda_k(A')\lambda_l(A'')$, and the eigenvalues of tridiagonal matrices can be estimated as in Example 2 of Lecture 2. Moreover, the eigenvalues of A depend monotonically on its coefficients and the domain, so that comparison theorems can be invoked.[3] Finally, in diffusion problems with absorption, when the diffusion length is only a few mesh-lengths (e.g., when $h^2\sigma/p > 0.1$, say, for the DE $\nabla(p\nabla u) - \sigma u = s(x, y)$), and more generally when D strongly dominates $E + F$ in (2), the spectral radius can be estimated from this fact alone.

[3] See P. R. Garabedian, Math. Tables Aid. Comput., 10 (1956), pp. 183–185.

Example 3. For a uniform mesh on a square and Dirichlet (clamped plate) boundary conditions, the eigenfunctions of ∇^4 and the fourth order central difference approximation $\nabla_h^4 = (\nabla_h^2)^2$ to it can again be found by inspection. Using the results of Example 2, we can verify that the eigenvalues of ∇_h^2 satisfy

$$\mu_{\min} = \lambda_{\min}^2 \leqq \mu_k \leqq 2 - \mu_{\min},$$

where $\lambda_{\min} = 2\sin^2 h$, whence $\mu_{\min} = 4\sin^2 h$. Although the matrix A associated with the operator ∇_h^4 is not a Stieltjes matrix, one can apply to it the block overrelaxation methods to be discussed in § 7.

5. Point SOR. More generally, we can define *point* SOR for any "relaxation factor" ω as follows (SOR is an acronym for "successive overrelaxation"):

(8) $$(D - \omega E)\mathbf{u}^{(n+1)} = \{(1 - \omega)D + \omega F\}\mathbf{u}^{(n)} + \omega\mathbf{b}.$$

Setting $L = D^{-1}E$ and $U = D^{-1}F$, this becomes

(8′) $$\mathbf{u}^{(n+1)} = (1 - \omega L)^{-1}[(1 - \omega)I + \omega U]\mathbf{u}^{(n)} + \omega(1 - \omega L)^{-1}D^{-1}\mathbf{b}.$$

When this method is applied over a complete cycle of mesh-points, the errors are transformed linearly in conformity to the formula

(9) $$\mathbf{e}^{(n+1)} = (1 - \omega L)^{-1}\{(1 - \omega)I + \omega U\}\mathbf{e}^{(n)},$$

which we rewrite as

(9′) $$\mathbf{e}^{(n+1)} = L_\omega[\mathbf{e}^{(n)}], \quad L_\omega = (1 - \omega L)^{-1}\{(1 - \omega)I + \omega U\}.$$

For Stieltjes matrices, the Ostrowski–Reich theorem [V, p. 77] asserts that $\rho(L_\omega) < 1$ (in other words, point SOR converges) if and only if A is positive definite and $0 < \omega < 2$.

Kahan's thesis. In his unpublished thesis, W. Kahan (1958) extended to general Stieltjes matrices, in less sharp form, many of the results on point SOR which had been obtained by Young for the 2-cyclic case. Specifically, he showed that Young's best optimal overrelaxation factor ω_b was still good. We summarize his results as follows (for details, see [V, Theorems 4.9 and 4.12]).[4]

Let $A\mathbf{x} = \mathbf{b}$, where A is a Stieltjes matrix. Then we can rescale the known b_i so as to get an equivalent system

(10) $$D^{-1}A\mathbf{x} = \mathbf{c}, \quad \mathbf{c} = D^{-1}\mathbf{b}, \quad D = \operatorname{diag} A.$$

Though $D^{-1}A$ is of course similar to the Stieltjes matrix

$$D^{1/2}AD^{1/2} = D^{-1/2}(DA)D^{1/2},$$

it is not itself generally a Stieltjes matrix. Both $D^{-1}A$ and $D^{1/2}AD^{1/2}$ have 1's on the main diagonal. Now rewrite (1) in the form

(11) $$\mathbf{x} = B\mathbf{x} + \mathbf{c}, \quad B = I - D^{-1}A,$$

[4] We also thank David Young for the exposition abstracted here (personal communication).

most suitable for iteration. Let $\mu = \rho(B)$ be the spectral radius of B:

$$\mu = \rho(D^{1/2}AD^{1/2} - 1) < 1$$

by § 4. Apply *successive point-overrelaxation* (point SOR) to (11), with the particular (over)relaxation factor

$$\omega_b = 2/(1 + \sqrt{1 - \mu^2}) = 1 + [\mu/(1 + \sqrt{1 - \mu^2})]^2. \tag{12}$$

Kahan has proved that for this ω_b,

$$\omega_b - 1 \leqq \rho(L_{\omega_b}) \leqq \sqrt{\omega_b - 1};$$

hence this ω_b is a *good* relaxation factor, since $\rho(L_\omega) \geqq \omega_b - 1$ for *any* relaxation factor. For $\rho(B) = 1 - \varepsilon$, where ε is small, the asymptotic convergence rate $\gamma = -\log \rho(L_{\omega_b})$ therefore satisfies

$$\sqrt{2\varepsilon} = -\tfrac{1}{2}\log(\omega_b - 1) \leqq \gamma \leqq \log(\omega_b - 1) = 2\sqrt{2\varepsilon}. \tag{13}$$

Rates of convergence. By combining the preceding considerations with those of § 4, one can show that the rate of convergence of SOR is $O(h)$ for second order and $O(h^2)$ for fourth order elliptic problems (see again [V], [9]).

Two-cyclic case. The original and simplest class of applications of point SOR was to the case of 5-point difference approximations to *self-adjoint* elliptic problems on a rectangular mesh. In this case, the matrix B for the point-Jacobi method is (weakly) 2-*cyclic*, in the sense that for an appropriate ordering of the entries (indices), it has the form sketched below:

$$B = \begin{bmatrix} O & C \\ \tilde{C} & O \end{bmatrix}.$$

This is most easily visualized by interpreting the nodes as forming a *checkerboard* of red and black squares, such that no link (nonzero entry of B) joins two squares of the same color. The matrix displayed above is obtained by listing first all red squares and then all black squares. Although this ordering gives Gauss-Seidel and SOR their optimal (minimum) spectral radius, it is complicated as regards transfer of data from tape to core; this is handled better by a straightforward row-by-row (or column-by-column) sweep of mesh-points.

However, the 2-cyclic form of B displayed can be made to yield a significant economy: it suffices to store values at red mesh-points during *even* cycles (half-iterations) and values at black mesh-points during *odd* cycles. In symbols, write $\mathbf{u} = \mathbf{v} + \mathbf{w}$, where $\mathbf{v}$ and $\mathbf{w}$ are the vectors whose components are the values of $\mathbf{u}$ at red and black mesh-points, respectively (see [V, p. 150]). Then $\mathbf{v}^{(n+2)} = \tilde{C}C\mathbf{v}^{(n)}$, and data transfer becomes efficient if one sweeps through all red mesh-points row-by-row, and then all black mesh-points row-by-row.

Optimum overrelaxation parameters. In the 2-cyclic case which he originally considered, Young gave an exact formula for the optimum overrelaxation factor ω_b:

$$\omega_b = 2/[1 + \sqrt{1 - \rho^2(B)}] \qquad [\text{V, p. } 110] \tag{14}$$

and the asymptotic rate of convergence [V, p. 106], in terms of the optimum SOR parameter ω_b. The eigenvalues μ of point SOR are related to those λ of the point-Jacobi method by

$$(\lambda + \omega - 1)^2 = \lambda\omega_b^2\mu^2 \qquad [\text{V}, (4.18)]. \tag{15}$$

They lie on a circle in the complex plane which is mapped 2–1 and conformally onto the slit of real eigenvalues of B.

To illustrate the effectiveness of the SOR method, consider the case when A is a 2-cyclic Stieltjes matrix and $\rho^2 = .9999$. For this problem the Gauss-Seidel method would require an average of 25,000 iterations (neglecting roundoff) to get an extra decimal place of accuracy, whereas the SOR method, using optimum ω, would require only 115 iterations. However, to achieve this rapid convergence the overrelaxation parameter ω, or equivalently ρ^2, must be estimated accurately. For the above example, if an estimate of .999 were used for ρ^2 in computing ω, the SOR method would require 704, instead of 115, iterations to reduce the error by a factor of ten. When ρ^2 is close to unity, small changes in the estimate for ρ^2 can drastically affect the rate of convergence, especially if ω is underestimated.

In practice, two different numerical schemes have been widely used to obtain estimates for ρ^2 (or equivalently ω_b). One approach is to attack the eigenvalue problem directly and calculate ρ^2 prior to starting the main SOR iterations (see, for example, [4]). The second approach is to start with the SOR iterations with some $\omega < \omega_b$ and then obtain new estimates for ω based on numerical results.[5] The second approach is of the "semi-iterative" type to be discussed in Lecture 5.

p-cyclic matrices. Varga has generalized many properties of 2-cyclic matrices (matrices having "Property A" in Young's terminology) to p-cyclic matrices, such as arise in the "outer iterations" of the multigroup diffusion equations [V, Chap. 4]. In particular [V, Theorem 4.5], the spectral radius of SOR is again $1 - O(h)$ with the optimum overrelaxation parameter ω_b. However, it is the case $p = 2$ which arises most frequently in applications.

6. Richardson's method. There are several variants of SOR which have approximately the same rate of convergence. One of these is the second order Richardson's method,[6] a two-step method of "simultaneous displacements" which expresses $\mathbf{u}^{(n+1)}$ in terms of $\mathbf{u}^{(n)}$ and $\mathbf{u}^{(n-1)}$. It has the disadvantage of requiring twice as much storage as SOR.

After setting $\mathbf{v}^{(0)} = \mathbf{Bu}^{(0)} + \mathbf{k}$, one can replace (1) by the following larger system [V, pp. 142–143]:

$$\mathbf{u}^{(n+1)} = \omega[B\mathbf{v}^{(n)} + \mathbf{k} - \mathbf{u}^{(n)}] + \mathbf{u}^{(n)}, \tag{16a}$$

$$\mathbf{v}^{(n+1)} = \omega[B\mathbf{u}^{(n+1)} + \mathbf{k} - \mathbf{v}^{(n)}] + \mathbf{v}^{(n)}, \tag{16b}$$

[5] See [V, Chap. 9]; Hageman and Kellogg [4]; and J. K. Reid, Comput. J., 9 (1966), pp. 200–204.

[6] L. F. Richardson, Philos. Trans. Roy. Soc. London Ser. A, 210 (1910), pp. 307–357; see [V, p. 159]. Richardson did not use the 2-cyclic concept.

which is 2-*cyclic* even if B is not. For the choice $\omega = \omega_b = 2/[1 + (1 - \rho^2(B))^{1/2}]$, one achieves a rate of convergence which is about half the optimal rate, as in Kahan's analysis.

SSOR overrelaxation. Another variant of SOR is Sheldon's "symmetric" SOR (or SSOR), in which sweeps are alternately made in the forward and backward directions of the ordering.[7] This is about as efficient as SOR. However, it has the advantage over SOR of having real eigenvalues and can be combined with *semi-iterative* methods (see Lecture 5) so as to achieve $O(h^{1/2})$ order of convergence with the 5-point approximation to the Dirichlet problem,[8] and D. M. Young has obtained a formula for the optimum overrelaxation factor (unpublished result).

Consistent ordering. In a similar vein, it has been shown that having a "consistent ordering" does not dramatically improve the rate of convergence [8],[9] and that all consistent orderings have exactly the same asymptotic rate of convergence.

7. Line and block overrelaxation.[10] The convergence of iterative methods can often be accelerated by using elimination to obtain a whole string of improved values at once. This is especially easy to achieve in the case of line or, more generally, k-line groupings of values on a rectangular network (2-cyclic case). The factor of acceleration for k-line overrelaxation is $\sqrt{k}$ [6], but not an order of magnitude (as $h \downarrow 0$). More important, such groupings make difference approximations *block tridiagonal* for sufficiently large k, permitting the use of block SOR.

In general, one can prove for (irreducible) Stieltjes matrices the much weaker result that block Gauss–Seidel converges more rapidly than point Gauss–Seidel; the proof involves "regular splittings" of matrices [V, p. 78]. More important, matrices arising from higher order problems such as the biharmonic equation of Example 3, § 4, have block tridiagonal form relative to suitable k-line groupings of the mesh-lines. By applying block SOR to the resulting system, one can reduce the rate of convergence for the biharmonic equation from $O(h^4)$ to $O(h^2)$, for example.[11]

A much more intimate combination of partial elimination with iteration has been recently used by H. L. Stone and others[12] on problems arising from 5-point difference approximations to source problems. The basic idea is to construct a matrix B such that $A - B$ is readily factored, as $A - B = LU$, into lower (resp.

upper) triangular matrices L and U, and then to iterate

$$LU\mathbf{u}^{(r+1)} = B\mathbf{u}^{(r)} + \mathbf{k}.$$

Such "strongly implicit" iterative approaches deserve further study partly because of their potential adaptability to the variational formulations with piecewise polynomial approximations to be discussed in Lectures 7 and 8.

REFERENCES FOR LECTURE 4

[1] G. Birkhoff and R. S. Varga, *Reactor criticality and non-negative matrices*, J. Soc. Indust. Appl. Math., 6 (1958), pp. 354–377.

[2] G. H. Golub and R. S. Varga, *Chebyshev semi-iterative methods, successive overrelaxation iterative methods and second order Richardson iterative methods, I, II*, Numer. Math., 3 (1961), pp. 147–156; pp. 157–168.

[3] L. A. Hageman and R. S. Varga, *Block iterative methods for cyclically reduced matrix equations*, Ibid., 6 (1964), pp. 106–119.

[4] L. A. Hageman and R. B. Kellogg, *Estimating optimum overrelaxation parameters*, Math. Comp., 22 (1968), pp. 60–68.

[5] W. Kahan, *Gauss–Seidel methods of solving large systems of linear equations*, Doctoral thesis, University of Toronto, 1958.

[6] S. V. Parter, *Multi-line iterative methods for elliptic difference equations and fundamental frequencies*, Numer. Math., 3 (1961), pp. 305–319.

[7] H. L. Stone, *Iterative solution of implicit approximations of multidimensional partial differential equations*, SIAM J. Numer. Anal., 5 (1968), pp. 530–558. (See also T. Dupont, R. P. Kendall and H. H. Rachford, Ibid., pp. 559–573; and [BV, pp. 168–174].)

[8] R. S. Varga, *Ordering of the successive overrelaxation scheme*, Pacific J. Math., 9 (1959), pp. 925–939.

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

[10] ———, *The numerical solution of elliptic and parabolic partial differential equations*, Modern Mathematics for the Engineer, Second Series, McGraw-Hill, New York, 1961, pp. 373–419.

[7] J. W. Sheldon, Math. Tables Aid. Comput., 9 (1955), pp. 101–112; J. Assoc. Comput. Mach., 6 (1959), pp. 494–505.

[8] G. J. Habetler and E. L. Wachspress, Math. Comp., 15 (1961), pp. 356–362.

[9] See also C. G. Broyden, Numer. Math., 12 (1968), pp. 47–56.

[10] See [V, § 6.4], [3], and the references given there. Early relevant papers include J. Schröder, Z. Angew Math. Mech., 34 (1954), pp. 241–253; R. J. Arms, L. D. Gates and B. Zondek, J. Soc. Indust. Appl. Math., 4 (1956), pp. 220–229; J. Heller, Ibid., 8 (1960), pp. 150–173.

[11] See S. V. Parter, Numer. Math., 1 (1959), pp. 240–252, and [6]; also J. Assoc. Comput. Mach., 8 (1961), pp. 359–365 and [V, p. 208].

[12] See [7]; also J. E. Gunn, SIAM J. Numer. Anal., 2 (1964), pp. 24–25; T. Dupont, Ibid., 4 (1968), pp. 753–782.

LECTURE 5

Semi-iterative Methods

1. Chebyshev semi-iteration. In Lecture 4, I discussed purely iterative methods for solving $D\mathbf{u} = (E + F)\mathbf{u} + \mathbf{b}$, which can be reduced to $\mathbf{u} = B\mathbf{u} + \mathbf{k}$ with symmetrizable B by "scaling." These methods consist in applying repeatedly the linear inhomogeneous (affine) operator $L:\mathbf{u} \to B\mathbf{u} + \mathbf{k}$, i.e., in "iterating"

$$\mathbf{u}^{(r)} = L[\mathbf{u}^{(r-1)}], \tag{1}$$

where (for example) we might have $L[\mathbf{u}] = B\mathbf{u} + \mathbf{k}$. In practice, optimal methods are seldom purely iterative, because numerical information obtained from previous iterations can usually be used as "feedback" to improve on L.

This leads to the study of *semi-iterative* methods of the more general form

$$\mathbf{u}^{(r)} = L_r[\mathbf{u}^{(0)}, \mathbf{u}^{(1)}, \cdots, \mathbf{u}^{(r-1)}]. \tag{1'}$$

Specifically, we shall consider in this section the rate of convergence of methods of the form

$$\mathbf{v}^{(r)} = \sum_{j=0}^{r} c_j^r \mathbf{u}^{(j)}, \tag{2}$$

where $\mathbf{u}^{(j)} = B\mathbf{u}^{(j-1)} + \mathbf{k}$. To measure this, we define the *error* of an approximate solution $\mathbf{v}$ of $\mathbf{u} = B\mathbf{u} + \mathbf{k}$ as $\mathbf{e} = \mathbf{v} - \mathbf{u}$, where $\mathbf{u}$ is the exact solution of $\mathbf{u} = B\mathbf{u} + \mathbf{k}$. We shall consider only "solution preserving" methods such that $\mathbf{v}^{(0)} = \mathbf{u}$ implies $\mathbf{v}^{(r)} = \mathbf{u}$; hence $\mathbf{e}^{(r)} = \mathbf{v}^{(r)} - \mathbf{u} = 0$, for all $r > 0$. For this,

$$\sum_{j=0}^{r} c_j^r = 1$$

is necessary and sufficient. We shall then have

$$\mathbf{e}^{(r)} = \sum_{j=0}^{r} c_j^r B^r[\mathbf{e}^{(0)}] = p_r(B)[\mathbf{e}^{(0)}], \quad p_r(x) = \sum_{j=0}^{r} c_j^r x^j. \tag{3}$$

For a general (random) initial $\mathbf{u}^{(0)} = \mathbf{v}^{(0)}$, the spectral radius of $p_r(B)$ provides the best measure of the rate of convergence. This leads one to ask: What choice of the c_j^r (i.e., of $p_r(B)$) will *minimize* the spectral radius of $p_r(B)$, among all polynomials p_r with $p_r(1) = 1$ (i.e., $\sum_{j=0}^{r} c_j^r = 1$)?

Since the eigenvalues of B are real and on $[-\rho(B), \rho(B)]$, the *Chebyshev polynomial*

$$C_m(x/\beta)/(C_m(\beta^{-1}), \quad \beta = \rho(B) \tag{4}$$

has the desired property. This was shown by L. F. Richardson, Lanczos [5], and Stiefel [8]. (Here $C_m(t) = \cos(m \cos^{-1} t)$ on $(-1, 1)$.) Moreover, from classic recursion formulas for the Chebyshev polynomials, it follows that[1]

$$\mathbf{u}^{(r)} = \omega_r\{B\mathbf{u}^{(r-1)} + \mathbf{k} - \mathbf{u}^{(r-2)}\} + \mathbf{u}^{(r-1)}, \tag{5}$$

where the rth relaxation factor is

$$\omega_r = 1 + C_{r-2}(1/\rho)/C_r(1/\rho), \quad \rho = \rho(B).$$

Furthermore, as was first observed by Golub and Varga (Lecture 4, [2]),

$$\lim_{r \to \infty} \omega_r = \omega_b = 2/[1 + (1 - \rho^2(B))^{1/2}].$$

More generally, "for very large numbers of iterations, there is very little difference" between SOR and Chebyshev [V, p. 143], and Chebyshev "requires an additional vector of storage," which makes semi-iterative Chebyshev by itself asymptotically no better than SOR, as $r \to \infty$.

This additional storage can be eliminated by the semi-iterative *cyclic* Chebyshev methods which will now be described.[2] Recall that the semi-iterative method of (5) works whenever B is convergent ($\rho(B) < 1$) and Hermitian. When B is also weakly 2-*cyclic*, i.e., when

$$B = \begin{bmatrix} 0 & F \\ F^T & 0 \end{bmatrix}, \ F \text{ Hermitian}, \tag{5a}$$

we can partition the $\mathbf{u}^{(r)}$ of (5) as $\begin{bmatrix} \mathbf{u}_1^{(r)} \\ \mathbf{u}_2^{(r)} \end{bmatrix}$, corresponding to the splitting of B in (5a). Furthermore, taking appropriate components of the $\mathbf{u}^{(r)}$, the iteration of (5) reduces to

$$\begin{aligned} u_1^{(2m+1)} &= \omega_{2m+1}\{Fu_2^{(2m)} + k_1 - u_1^{(2m-1)}\} + u_1^{(2m-1)}, \qquad m \geqq 1, \\ u_2^{(2m+2)} &= \omega_{2m+2}\{F^T u_1^{(2m+1)} + k_2 - u_2^{(2m)}\} + u_2^{(2m)}, \qquad m \geqq 0, \end{aligned} \tag{5b}$$

where ω_r is as before, with $\omega_1 = 1$, and where $u_1^{(1)} = Fu_2^{(0)} + k_1$. This semi-iterative cyclic Chebyshev method of Golub and Varga (Lecture 4, [2]) then requires no extra vector storage, and retains the superior norm characteristics of the cyclic Chebyshev method.

For very large systems of linear equations, combinations of multiline and block techniques with this semi-iterative cyclic Chebyshev method are probably the most effective methods in widespread use today. Here by multiline techniques, we mean direct inversion on sets of k adjacent lines.[3] Such multiline techniques permit one to adapt the cyclic Chebyshev method to the biharmonic and other higher order difference equations,[4] as well as to variational methods using piecewise polynomial approximations with patch bases (see Lecture 8).

[1] See A. Blair, N. Metropolis, et al., Math. Tables Aid. Comput., 13 (1959), pp. 145–184.

[2] The following exposition was kindly supplied by Professor Varga.

[3] E. Cuthill and R. S. Varga, J. Assoc. Comput. Mach., 6 (1959), pp. 236–244.

2. Matrices H **and** V. A much more novel family of semi-iterative methods for solving plane elliptic problems is provided by *alternating direction implicit* (ADI) schemes, to which most of this lecture will be devoted.

As in Lecture 3, let the self-adjoint elliptic partial DE

$$G(x, y)u - \frac{\partial}{\partial x}\left[A(x, y)\frac{\partial u}{\partial x}\right] - \frac{\partial}{\partial y}\left[C(x, y)\frac{\partial u}{\partial y}\right] = S(x, y) \tag{6}$$

be approximated by the 5-point difference equation

$$(H + V + \Sigma)\mathbf{u} = \mathbf{b}, \tag{7}$$

where, for a uniform rectangular mesh with mesh lengths h and k, we have

$$Hu(x, y) = -a(x, y)u(x + h, y) + 2b(x, y)u(x, y) - c(x, y)u(x - h, y), \tag{8}$$

$$Vu(x, y) = -\alpha(x, y)U(x, y + h) + 2\beta(x, y)u(x, y) - \gamma(x, y)u(x, y - k). \tag{9}$$

The most common choices for $a, b, c, \alpha, \beta, \gamma$ are

$$\begin{aligned} a &= kA(x + h/2, y)/h, \quad c = kA(x - h/2, y)/h, \quad 2b = a + c, \\ \alpha &= hC(x, y + k/2)/k, \quad \gamma = hC(x, y - k/2)/k, \quad 2\beta = \alpha + \gamma. \end{aligned} \tag{10}$$

These choices[5] make H and V *symmetric* matrices acting on the vector space of functions $u = u(x_i, y_i)$ defined on interior mesh-points.

We shall assume that A and C are *positive* functions in (6) which makes the DE elliptic, while G is nonnegative. The matrix Σ is then a nonnegative diagonal matrix with diagonal entry $hkG(x_i, y_i)$ at (x_i, y_i). The vector $\mathbf{b}$ is computed by adding to the source terms $hkS(x_i, y_i)$ the terms in (8)–(9) associated with points on the *boundary* of the domain.

Our concern here is with the rapid solution of the vector equation (7) for large networks. For this purpose, it is essential to keep in mind some general properties of the matrices Σ, H and V.

As already stated, Σ is a nonnegative diagonal matrix. Moreover, H and V have positive diagonal entries and nonpositive off-diagonal entries. Because of the Dirichlet boundary conditions for (6), the diagonal dominance of H and V implies that they are positive definite [V, p. 23]; as in Lecture 4, such real symmetric and positive definite matrices with nonpositive off-diagonal entries are called *Stieltjes matrices.*

If the network $\mathscr{R}(h, h) = \mathscr{R}_h$ of interior mesh-points is connected, then $H + V$ and $H + V + \Sigma$ are also *irreducible*: it is known[6] that if a Stieltjes matrix is irreducible, then its matrix inverse has all positive entries.

[4] J. Heller, J. Soc. Indust. Appl. Math., 8 (1960), pp. 150–173; and S. V. Parter, Numer. Math., 3 (1961), pp. 305–319; see also Hageman and Varga (Lecture 4, [3]).

[5] See [1, § 2] for choices of a, b, c; also [W, pp. 70–74], Spanier [7, Part **d**] derives the appropriate difference approximations in a cylindrical or (r, z)-geometry.

[6] See [V, § 3.5]; irreducibility is defined in [V, § 1.4].

The matrices H and V are also *diagonally dominant*, by which we mean that the absolute value of the diagonal entry in any row is greater than or equal to the sum of the off-diagonal entries. For any $\theta \geqq 0$, the same is true a fortiori of $H + \theta\Sigma$, $V + \theta\Sigma$, and for $\theta_1 H + \theta_2 V + \theta\Sigma$ if $\theta_1 > 0$, $\theta_2 > 0$. The above matrices are all *diagonally dominant Stieltjes matrices*.

By ordering the mesh-points by rows, one can make H tridiagonal; by ordering them by columns, one can make V tridiagonal. That is, both H and V are similar to tridiagonal matrices, but one cannot make them both tridiagonal simultaneously.

3. Basic ADI operators. From now on, we shall consider only the iterative solution of the vector equation (7). Since it will no longer be necessary to distinguish the approximate solutions u from the exact solution $u(x, y)$, we shall cease to use boldface type, and will write u_n instead of $\mathbf{u}^{(n)}$.

Equation (7) is clearly equivalent, for any matrices D and E, to each of the two vector equations

$$(H + \Sigma + D)u = k - (V - D)u, \tag{11}$$

$$(V + \Sigma + E)u = k - (H - E)u. \tag{12}$$

This was first observed by Peaceman and Rachford in [6] for the case $\Sigma = 0$, $D = E = \rho I$ a scalar matrix. In this case, (11) and (12) reduce to

$$(H + \rho I)u = k - (V - \rho I)u, \quad (V + \rho I)u = k - (H - \rho I)v.$$

The generalization to $\Sigma \neq 0$ and arbitrary $D = E$ was made by Wachspress and Habetler [8].

For the case $\Sigma = 0$, $D = E = \rho I$ which they considered, Peaceman and Rachford proposed solving (7) by choosing an appropriate sequence of positive numbers ρ_n, and calculating the sequence of vectors $u_n, u_{n+1/2}$ defined from the sequence of matrices $D_n = E_n = \rho_n I$, by the formulas

$$(H + \Sigma + D_n)u_{n+1/2} = k - (V - D_n)u_n, \tag{13}$$

$$(V + \Sigma + E_n)u_{n+1} = k - (H - E_n)u_{n+1/2}. \tag{14}$$

Provided that $(H + \Sigma + D)$ and $(V + \Sigma + E)$ are nonsingular, and that the matrices to be inverted are similar under conjugation by permutation matrices (and scaling) to tridiagonal Stieltjes matrices, each of the equations (13) and (14) can be rapidly solved by Gauss elimination. The aim is to choose the initial trial vector u_0 and the matrices $D_1, E_1, D_2, E_2, \cdots$ so as to make the sequence $\{u_n\}$ converge rapidly.

Peaceman and Rachford considered the iteration of (13) and (14) when D_n and E_n are given by $D_n = \rho_n I$ and $E_n = \tilde{\rho}_n I$. This defines the Peaceman–Rachford method:

$$u_{n+1/2} = (H + \Sigma + \rho_n I)^{-1}[k - (V - \rho_n I)u_n], \tag{15}$$

$$u_{n+1} = (V + \Sigma + \tilde{\rho}_n I)^{-1}[k - (H - \tilde{\rho}_n I)u_{n+1/2}]. \tag{16}$$

The rate of convergence will depend strongly on the choice of the iteration parameters $\rho_n, \tilde{\rho}_n$.

An interesting variant of the Peaceman–Rachford method was suggested by Douglas and Rachford [3, p. 422, (2.3)], again for the case $\Sigma = 0$. It can be defined for general $\Sigma \geqq 0$ by

$$u_{n+1/2} = (H_1 + \rho_n I)^{-1}[k - (V_1 - \rho_n I)u_n], \tag{17}$$

$$u_{n+1} = (V_1 + \rho_n I)^{-1}[V_1 u_n + \rho_n u_{n+1/2}], \tag{18}$$

where H_1 and V_1 are defined as $H + \frac{1}{2}\Sigma$ and $V + \frac{1}{2}\Sigma$, respectively. This amounts to setting $D_n = E_n = \rho_n I - \frac{1}{2}\Sigma$ in (13) and (14) and making some elementary manipulations. Hence (17) and (18) are also equivalent to (7) if $u_n = u_{n+1/2} = u_{n+1}$.

For higher-dimensional ADI methods, see J. Douglas, Numer. Math., 4 (1962), pp. 41–63, and J. Douglas, B. Kellogg and R. S. Varga, Math. Comp., 17 (1963), pp. 279–282.

4. Model problems. The power of ADI methods is greatest for model problems in which the preceding difference equations involve *permutable operators*, so that[7]

$$HV = VH, \quad H\Sigma = \Sigma H, \quad \text{and} \quad V\Sigma = \Sigma V. \tag{19}$$

This is the case if (1) reduces to the (modified) Helmholtz equation in a rectangle: $\sigma u - \nabla^2 u = S(x, y)$. More generally, H and V are permutable when the variables x and y are "separable" (in the sense discussed in Lecture 2, § 1) for the given elliptic problem.

If (19) holds, one can achieve an order-of-magnitude gain in the rate of convergence with the ADI methods described in § 3 by letting the ρ_n and $\tilde{\rho}_n$ be distributed in the intervals containing the (real) eigenvalues of $(H + \Sigma + \rho_n I)^{-1}$ $(V - \rho_n I)$ and $(V + \Sigma + \tilde{\rho}_n I)^{-1}$ $(H - \tilde{\rho}_n I)$ with equal proportionate spacing (see [7, g]). As the mesh-length h decreases, the number of (semi-) iterations required to reduce the error by a prescribed factor is (asymptotically and neglecting roundoff) only $O(\log h^{-1})$, as compared with $O(h^{-1})$ for SOR using the optimum relaxation parameter ω_b, or $O(h^{-2})$ as with Gauss–Seidel (or point-Jacobi).

A very interesting precise determination of the *optimum* parameters for such model problems has in fact been made by Jordan, in terms of elliptic functions; we shall omit the details.[8]

Unfortunately, it seems to be impossible to make rigorous extensions of the preceding theoretical results to most problems with variable coefficients or in nonrectangular regions [1], [2]. Whereas the theory of SOR applies to the 5-point approximation to general source problems, the experimentally observed success of ADI is in general hard to explain and even harder to predict.

[7] For discussions of (19), see [2, Part II], and R. E. Lynch, J. R. Rice and D. H. Thomas, Bull. Amer. Math. Soc., 70 (1964), pp. 378–384.

[8] See [W, p. 185], or E. Wachspress, J. Soc. Indust. Appl. Math., 10 (1962), pp. 339–350; 11 (1963), pp. 994–1016. Also C. de Boor and John Rice, Ibid., 11 (1963), pp. 159–169 and 12 (1964), pp. 892–896; and R. B. Kellogg and J. Spanier, Math. Comp., 19 (1965), pp. 448–452.

5. Iterative ADI. Even purely iterative (or "stationary") ADI methods using a single parameter ρ have the same order of convergence as optimized SOR [2, Theorem 20.1]. More generally [9, Theorem 1], simple iteration of the Peaceman–Rachford method (15)–(16) is always *convergent* if one chooses $D = E = \rho I - \Sigma/2$, where ρ is a positive number. This makes $D + \Sigma/2$ positive definite and symmetric and $H + V + \Sigma$ positive definite. A few sample proofs will be sketched below (see [1] and [2] for more details).

As in Lecture 4, we define the error vector as the difference $e_n = u_n - u_\infty$ between the *approximate solution* u_n after the nth iteration and the *exact solution* u_∞ of (7). For simplicity, we set $D = E = \rho I$. A straightforward calculation shows that, for the Peaceman–Rachford method, the effect of a single iteration of (15)–(16) is to multiply the error vector e_n by the error reduction matrix T, defined by

$$T_\rho = (V + \Sigma + \rho I)^{-1}(H - \rho I)(H + \Sigma + \rho I)^{-1}(V - \rho I). \tag{20}$$

Likewise, the error reduction matrix for the Douglas–Rachford method (18)–(19) with all $\rho_n = \rho$ is given by

$$\begin{aligned} W_\rho &= (V_1 + \rho I)^{-1}(H_1 + \rho I)^{-1}(H_1 V_1 + \rho^2 I) \\ &= [H_1 V_1 + \rho(V_1 + H_1) + \rho^2 I]^{-1}(H_1 V_1 + \rho^2 I). \end{aligned} \tag{21}$$

If one assumes that $D_n = -\Sigma/2 + \rho I = E_n$ also for the generalized Peaceman–Rachford method (13)–(14), then from (15), we have

$$T_\rho = (V_1 + \rho I)^{-1}(H_1 - \rho I)(H_1 + \rho I)^{-1}(V_1 - \rho I), \tag{22}$$

and the matrices W_ρ and T_ρ are related by

$$2W_\rho = I + T_\rho. \tag{23}$$

We next prove a lemma which expresses the algebraic content of a theorem of Wachspress and Habetler [9, Theorem 1].

LEMMA 1. *Let P and S be positive definite real matrices, with S symmetric. Then $Q = (P - S)(P + S)^{-1}$ is norm-reducing*[9] *for real row vectors x relative to the norm $\|x\| = (xS^{-1}x^T)^{1/2}$.*

Proof. For any norm $\|x\|$, the statement that Q is norm-reducing is equivalent to the statement that $\|(S - P)y\|^2 < \|(S + P)y\|^2$ for every nonzero vector $y = (P + S)^{-1}x$. In turn, this is equivalent for the special Euclidean norm $\|x\| = (xS^{-1}x^T)^{1/2}$ to the statement that

$$y(P + S)S^{-1}(P^T + S^T)y^T > y(P - S)S^{-1}(P - S)^T y^T \tag{24}$$

for all nonzero y. Expanding the bilinear terms, cancelling, and dividing by two, this is equivalent to the condition that $y(P + P^T)y^T > 0$ for all nonzero y. But this is the hypothesis that P is positive definite.[10]

COROLLARY. *In Lemma* 1, $\rho(Q) < 1$.

[9] The phase "norm-reducing" here refers to Euclidean norm only in special cases.

[10] Note that P is *not* assumed to be symmetric, but only to be such that $x^T(P + P^T)x > 0$, for all real $x \neq 0$.

This follows from Lemma 1 and the following general result on matrices:

$$\rho(M) \leqq \max_{\|x\|=1} (\|Mx\|/\|x\|) \quad \text{for any norm } \|\cdot\|.$$

Actually, $\rho(M)$ is the infimum of $\max(\|Mx\|/\|x\|)$ taken over all Euclidean (inner product) norms.

THEOREM 1. *Any iterative ADI process* (13)–(14) *with all $D_n = D$ and all $E_n = E$ is convergent, provided $\Sigma + D + E$ is symmetric and positive definite, and $2H + \Sigma + D - E$ and $2V + \Sigma + E - D$ are positive definite.*[11]

Proof. It suffices to show that $\rho(T) < 1$. But since similar matrices have the same eigenvalues and hence the same spectral radius, the error reduction matrix

$$T = (V + \Sigma + E)^{-1}(H - E)(H + \Sigma + D)^{-1}(V - D) \tag{25}$$

of (13)–(14) has the same spectral radius as

$$\begin{aligned} \tilde{T} &= (V + \Sigma + E)T(V + \Sigma + E)^{-1} \\ &= [(H - E)(H + \Sigma + D)^{-1}][(V - D)(V + \Sigma + D)^{-1}]. \end{aligned} \tag{26}$$

By Lemma 2, both factors in square brackets reduce the norm $[x^T(\Sigma + D + E)^{-1}x]^{1/2} = \|x\|$, provided $\Sigma + D + E = 2S$, $R_H = [H + \Sigma/2 + (D - E)/2]$ and $R_V = [V + \Sigma/2 + (E - D)/2]$ are positive definite, and $\Sigma + D + E$ is also symmetric.

It is easy to apply the preceding result to difference equations (8)–(9) arising from the Dirichlet problem for the self-adjoint elliptic differential equation (6). In this case, as stated in § 2, H and V are diagonally dominant (positive definite) *Stieltjes matrices.* The same properties hold a fortiori for $\theta_1 H + \theta_2 V + \theta_3 \Sigma$ if all $\theta_i \geqq 0$ and $\theta_1 + \theta_2 > 0$.

Hence the hypotheses of Theorem 1 are fulfilled for $D = \rho I - \theta\Sigma$, $E = \tilde{\rho} I - \tilde{\theta}\Sigma$ for any ρ, $\tilde{\rho} > 0$ and any θ, $\tilde{\theta}$ with $0 \leqq \theta, \tilde{\theta} \leqq 2$. Substituting into (13)–(14), we obtain the following result.

COROLLARY 1. *If $\rho, \tilde{\rho}, > 0$ and $0 \leqq \theta, \tilde{\theta} \leqq 2$, then the stationary ADI method defined with $\theta' = 2 - \theta$ by*

$$(H + \theta\Sigma/2 + \rho I)u_{n+1/2} = k - (V + \theta'\Sigma/2 - \rho I)u_n, \tag{27}$$

$$(V + \theta\Sigma/2 + \rho I)u_{n+1} = k - (H + \theta'\Sigma/2 - \rho I)u_{n+1/2} \tag{28}$$

is convergent.

In fact, it is norm-reducing for the norm defined by

$$\|x\|^2 = x^T(\Sigma + D + E)^{-1}x = x^T[(\rho + \tilde{\rho})I + (\theta + \tilde{\theta})\Sigma/2]^{-1}x.$$

6. Final remarks. One of the most interesting treatments of a reasonably general case is that of Guilinger [4]. Utilizing the *smoothness* of solutions of elliptic DE's (see Lecture 2), Guilinger proved that the Peaceman–Rachford semi-iterative method could be made to reduce the error by a given factor in a number of steps which was *independent of the mesh-length.*[12]

[11] This result, for $D - E = 0$, was first given in [9]. For the analogous result on W, see [1].

[12] See also R. E. Lynch and J. R. Rice, Math. Comp., 22 (1968), pp. 311–335.

Also, Widlund [10] has obtained some theoretical results for ADI with variable ρ (i.e., as a semi-iterative method) in the noncommutative case. Moreover, Spanier [7] and Kellogg[13] have applied ADI methods to ΔE's on nonrectangular meshes.

Finally, the reader's attention is called to the existence of a carefully documented[14] HOT-1 code, written at the Bettis Atomic Power Laboratory, whose relation to the theoretical principles described in this chapter has been the subject of a careful and lucid exposition by Spanier [7].

REFERENCES FOR LECTURE 5

[1] G. Birkhoff and R. S. Varga, *Implicit alternating direction methods*, Trans. Amer. Math. Soc., 92 (1959), pp. 13–24.

[2] G. Birkhoff, R. S. Varga and David Young, *Alternating direction implicit methods*, Advances in Computers, 3 (1962), pp. 189–273.

[3] J. Douglas, Jr. and H. Rachford, *On the numerical solution of heat conduction problems in two and three space variables*, Trans. Amer. Math. Soc., 82 (1956), pp. 421–439.

[4] W. H. Guilinger, *Peaceman-Rachford method with small mesh-increments*, J. Math. Anal. Appl., 11 (1964), pp. 261–277.

[5] C. Lanczos, *Solution of systems of linear equations by minimized iterations*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 33–53.

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

[7] J. Spanier, *Alternating direction methods applied to heat conduction problems*, Mathematical Methods for Digital Computers, A. Ralston and H. S. Wilf, eds., vol. II, John Wiley, New York, 1967, pp. 215–245.

[8] E. Stiefel, *On solving Fredholm integral equations*, J. Soc. Indust. Appl. Math., 4 (1956), pp. 63–85.

[9] E. L. Wachspress and G. J. Habetler, *An alternating-direction-implicit iteration technique*, Ibid., 8 (1960), pp. 403–424.

[10] O. Widlund, *On the rate of convergence of an alternating direction implicit method*, Math. Comp., 20 (1966), pp. 500–515.

[11] David Young, *On the solution of linear systems by iteration*, Proc. Symposia Applied Math., vol. VI, American Mathematical Society, Providence, 1956, pp. 283–298.

[13] R. B. Kellogg, Math. Comp., 18 (1964), pp. 203–210.

[14] R. B. Smith and J. Spanier, Bettis Atomic Power Laboratory Report WAPD-TM-465, 1964.

LECTURE 6

Integral Equation Methods

1. Introduction. The last three lectures were devoted to difference methods. These quickly reduce elliptic problems to an approximately equivalent algebraic form by elementary considerations from analysis; the main job is to solve the resulting system of algebraic equations.

The next three lectures will make much deeper use of classical analysis, including especially more sophisticated *approximation* methods (in Lecture 7) and *variational* methods (in Lecture 8). The present lecture will introduce the subject by describing briefly a number of numerical techniques which are especially closely related to the results from classical analysis reviewed in Lecture 2.

These techniques are especially applicable to homogeneous linear elliptic DE's with constant coefficients, such as $\nabla^2 u = 0$ or $\nabla^4 u = 0$. One of the most powerful classical techniques consists in expanding in series. We already saw in Lecture 2 how effective this technique was for solving the Dirichlet problem in the unit disc (by Fourier series). In § 2, we shall discuss its extension on to other domains.

A related but more sophisticated approach consists in expressing functions in terms of definite integrals of their boundary values or other quantities (e.g., their normal derivatives on the boundary). This approach will be discussed in § 3 and § 4.

Both techniques rely essentially on the principle that any (discrete or continuous) *superposition* (linear combination or integral) of solutions of a given homogeneous linear DE is again a solution. Hence, if one has a *basis* of elementary solutions, one can take the coefficients w_j or weight-function $w(\mathbf{s})$ as unknowns in an expression for the general solution

$$u(\mathbf{x}) = \Sigma w_j \varphi_y(\mathbf{x}) \quad \text{or} \quad u(\mathbf{x}) = \int \varphi(\mathbf{s}, \mathbf{x})\, dw(\mathbf{s}),$$

respectively, and then try to obtain enough equations on the w_j or on $w(\mathbf{s})$ to determine which of them represents the solution.

This lecture and Lecture 8 will contain several illustrations of ways to obtain numerical results by implementing the above principle. In general, one must use (approximate) numerical quadrature to obtain such results, although in exceptional cases formal integration may be possible.

2. Superposition of elementary solutions. By the maximum principle, any harmonic function $u(\mathbf{x})$ which is uniformly approximated on the boundary ∂R of a