The Approximate Solution of Matrix Problems*

A. S. Householder

Oak Ridge National Laboratory, Oak Ridge, Tennessee

1. *Statement of the Problem; Notational Conventions*

These notes have to do with methods of obtaining and methods of appraising approximate solutions of matrix problems. If the problem is to solve a system of linear equations, or to invert a matrix, one might suppose that an appraisal of the error can be made directly by substitution. But consider the system

$$Ax = h, \tag{1.1}$$

and let $\lambda = \lambda(A)$ be a proper value of A, and u a proper vector belonging to λ:

$$Au = \lambda u.$$

Then

$$A(x + u) = h + \lambda u.$$

Hence if λ is small, any component of error along u can be completely obscured in the rounding process, so that an "approximate" solution $x^* = x + u$, even if crude, may satisfy the system exactly to within machine errors.

The situation is the same in the inversion of a matrix; in fact, the following theorem is of interest:

THEOREM 1.1. *For any $\lambda > 0$ and any $\mu > 0$, there exist matrices, A and C such that every element of $AC - I$ is numerically less than λ, whereas there are elements of $CA - I$ which equal μ in magnitude.*

In other words, C could be a good right-hand inverse of A but a poor left-hand inverse. Let

$$Au = \lambda u, \qquad A^T v = \mu v.$$

For any λ, μ, u and v such a matrix A exists. Then

$$A(A^{-1} + uv^T) = I + \lambda uv^T,$$
$$(A^{-1} + uv^T)A = I + \mu uv^T,$$

and one has only to take

$$C = A^{-1} + uv^T.$$

The observation of certain notational conventions will save repeated explanations. Either Greek or Roman lower case letters will be used for indices and dimensions. Otherwise, lower case Greek letters will denote scalars; lower

* Received March, 1957; revised October, 1957.

case Roman letters will denote vectors, and these will be column vectors unless the contrary is indicated; capitals, either Greek or Roman, will denote matrices. In general, matrices will be square and of order n, unless it is indicated otherwise, and vectors will be of dimension n. The elements of a matrix A will generally be α_{ij}; the column vectors $a_{.j}$ and the row vectors $a_{i.}$. Its proper values will be $\lambda_i(A)$, and the singular values are the non-negative quantities σ_i defined by

$$\sigma_i^2(A) = \sigma_i^2(A^*) = \lambda_i(AA^*) = \lambda_i(A^*A).$$

The spectral radius is

$$\rho(A) = \max_i |\lambda_i(A)| .$$

The identity matrix will be denoted I, however, and its ith column vector is e_i and its ith row vector e_i^T. Also

$$e = \sum e_i$$

is the vector with 1 in each position.

The matrix

$$J = \begin{bmatrix} 0 & 0 & 0 & \cdot & \cdot & \cdot \\ 1 & 0 & 0 & \cdot & \cdot & \cdot \\ 0 & 1 & 0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

has 1 just below a diagonal element, and is zero elsewhere. The matrix

$$K = J + J^T$$

has 1 just below a diagonal element and 1 just above, and is elsewhere zero When necessary a subscript will indicate the order: J_n and K_n are of order n. Evidently J^2 has 1 two places below a diagonal element; J^3 three places; $\cdots$; and $J^n = 0$.

Absolute value signs with matrices and vectors signify the replacement of each element by its absolute value. Inequality signs between matrices or between vectors signify that corresponding elements everywhere satisfy the inequality. Thus

$$A \leqq B$$

signifies that

$$\alpha_{ij} \leqq \beta_{ij}$$

for every i and j. Moreover,

$$A \leqq |B| .$$

signifies that

$$\alpha_{ij} \leqq |\beta_{ij}|$$

for every i and j. Note that the two statements

$$A \leqq B, \qquad A \neq B$$

do *not* imply $A < B$. They imply only that the relation

$$\alpha_{ij} \leqq \beta_{ij}$$

holds for every i and j, and that for some i, j it is true that

$$\alpha_{ij} < \beta_{ij}.$$

But such implications as:

$$\text{if } A \leqq B, \qquad B \leqq C, \qquad \text{then } A \leqq C;$$

$$\text{if } A \leqq B, \qquad C \geqq 0, \qquad \text{then } AC \leqq BC;$$

are fairly obvious, as is the inequality

$$|A + B| \leqq |A| + |B|.$$

Finally, in discussing systems of equations it is sufficient, when convenient, to assume all quantities real, since a complex system can be replaced by a real system of twice the order. Thus consider the system

$$(A + iB)(x + iy) = h + ik.$$

On multiplying out and equating real and imaginary parts one has

$$Ax - By = h,$$

$$Bx + Ay = k.$$

Hence the complex system is equivalent to the real system

$$\begin{bmatrix} A & -B \\ B & A \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} h \\ k \end{bmatrix}.$$

Unfortunately, though, in the discussion of proper values and vectors there is, in general, no escape from the complex plane.

It will be convenient on occasion to speak of "the point x." By this is meant the point at the terminus of x when x is drawn from the origin. It will also be convenient to speak of "the space A." By this will be meant the space of all possible linear combinations of the columns of A.

2. *Norms*

Two vectors can differ in any or all of their n elements; two matrices in any or all of their n^2 elements. In either case it is desirable to have a single number to measure the departure of the one from the other or to measure the magni-

tude of a vector as a whole or a matrix as a whole. This implies the need for a suitable real valued, non-negative function of all the elements. The functions that are most useful for present purposes possess certain other special properties and are called norms. Using pairs of vertical bars to denote a norm, the additional properties can be stated as follows:

$$(V1): \qquad \| x \| > 0 \quad \text{unless} \quad x = 0;$$

$$(V2): \qquad \text{if } \alpha \geqq 0, \quad \| \alpha x \| = \alpha \| x \| ;$$

$$(V3): \qquad \| x + y \| \leqq \| x \| + \| y \| .$$

With $\alpha = 0$, $V2$ implies that $\| 0 \| = 0$. Any real-valued function of the elements of a vector will be called a norm if it possesses these three properties.

It is sufficient, for the moment, to consider only real vectors, since if

$$x = u + iv,$$

a complex vector, one can think of the vector $\begin{pmatrix} u \\ v \end{pmatrix}$ in $2n$-space.

THEOREM 2.1. *Given any norm, the points satisfying*

$$K: \qquad \| x \| \leqq 1,$$

form a closed, bounded convex body which contains the origin in its interior.

To say that K is a convex body means that any segment joining two points in K contains only points of K. If x and y are two points of K, then all points of the segment between them are represented by

$$\alpha x + (1 - \alpha)y, \qquad 0 \leqq \alpha \leqq 1.$$

But

$$\| \alpha x + (1 - \alpha)y \| \leqq \| \alpha x \| + \|(1 - \alpha)y \|$$
$$= \alpha \| x \| + (1 - \alpha)\| y \| \leqq \alpha + (1 - \alpha) = 1.$$

If K were not finite, any finite vector x in the direction of a point at infinity would satisfy $\| x \| = 0$, contrary to $V1$.

THEOREM 2.2. *Let K be any closed, bounded convex body containing the origin in its interior. Any ray through the origin intersects the boundary of K in exactly one point. If the ray is in the direction x, let the intersection be X', and let X represent the terminus of x. Define*

$$\nu(x) = \frac{OX}{OX'}.$$

Then the function $\nu(x)$ is a norm. Hereafter it will be understood that the origin is interior to any convex body to be considered, and that the convex body is closed and bounded.

It is clear that the function possesses properties $V1$ and $V2$. To show that it also possesses $V3$, if x and y are any two points, then

$$x' = \frac{x}{\nu(x)}, \quad y' = \frac{y}{\nu(y)}$$

terminate on the boundary of K, since $\nu(x') = \nu(y') = 1$. The point

$$\frac{x'\nu(x) + y'\nu(y)}{\nu(x) + \nu(y)} = \frac{(x + y)}{\nu(x) + \nu(y)}$$

is on the segment joining x' and y' and hence, by the convexity property, lies within K or on its boundary. Hence

$$\nu\left[\frac{(x + y)}{\nu(x) + \nu(y)}\right] \leqq 1,$$

and hence, by property $V2$ which the function is already known to possess

$$\nu(x + y) \leqq \nu(x) + \nu(y).$$

But this is $V3$.

Thus there is a one-to-one correspondence between norms and convex bodies K. On occasion it will be convenient to let $\| x \|_K$ signify the norm associated with K. Most useful norms have the property

$$(V2') \qquad \| \alpha x \| = | \alpha | \cdot \| x \|,$$

somewhat stronger than $V2$. For this to hold in the real case, K must be symmetric with respect to the origin:

$$\| x \| = \| -x \|.$$

In the complex case this implies that

$$\| e^{i\omega} x \| = \| x \|,$$

which implies that K is bounded by certain cylindrical surfaces.

It is sometimes convenient to consider any K as a member of a nested family of convex bodies κK, for $\kappa \geqq 0$, where κK consists of all points x satisfying

$$\| x \| \leqq \kappa.$$

More generally, for any scaler κ (real or complex), κK can be defined as the set of all points $y = \kappa x$ where $x \varepsilon K$. Evidently if the origin is strictly interior to K, it is strictly interior to κK when $\kappa \neq 0$. Throughout this discussion ordinary Euclidean geometry is presupposed, with an orthonormal set of basis vectors.

THEOREM 2.3. *If the sequence $\{u_i\}$ of vectors u_i vanishes in the limit in one norm it vanishes in every norm. In this event the vectors u_i will be said to approach 0 as a limit, and if $u_i = x_i - x$, the vectors x_i will be said to approach x as a limit.*

Let K and K' be two convex bodies and $\| u_i \|$ and $\| u_i \|'$ the associated norms. Suppose the sequence of norms $\| u_i \|$ is known to vanish in the limit. Let $\kappa > 0$ satisfy $\kappa \leqq \| x \|$ for all points x with $\| x \|' = \kappa'$. Such a κ exists since the origin is strictly interior to $\kappa' K'$. Hence

$$\kappa K \subset \kappa' K'.$$

Since the sequence of $\| u_i \|$ vanishes in the limit, for every $\kappa > 0$ there exists a p such that when $i \geqq p$, $\| u_i \| \leqq \kappa$. Hence $\| u_i \|' \leqq \kappa'$, which proves the theorem.

It is usual to say that a sequence of vectors vanishes in the limit if each sequence of corresponding elements vanishes in the limit. But the magnitude of an element of maximal magnitude is a norm, and its K is the hypercube whose faces are

$$\xi_i = \pm 1.$$

Hence the theorem implies that a sequence of vectors converges in the ordinary sense if and only if, for any choice of norm, the sequence of norms vanishes. An obvious corollary is:

COROLLARY 2.3. *A norm $\| x \|$ is a continuous function of the n variables ξ_i.*

For real vectors u and x, consider the relation

$$u^T x = \nu$$

with fixed u and variable x.

If ν is held to a fixed value the vectors x define a plane having u as normal. For $\nu = 0$, the plane passes through the origin, and intersects any K. As ν increases, it will pass through a certain maximal value above which the plane no longer contains points of K. The plane corresponding to this maximal value of ν is called a support plane of K. This plane divides the space into two half-spaces, one containing K and one not. For all points x of the first half-space,

$$u^T x \leqq \nu.$$

The value of ν defining the support plane is a function $\nu(u)$, called the support function, and the plane is called the support plane in the direction u. For complex vectors one considers the real part of u^*x,

$$\mathrm{Re}(u^*x) \leqq \nu,$$

since evidently if $u = u_1 + iu_2$, $x = x_1 + ix_2$, then

$$\mathrm{Re}(u^*x) = u_1^T x_1 + u_2^T x_2 .$$

THEOREM 2.4. *Given a convex body K and the associated norm, the support function $\nu(u)$ is a norm $\| u \|'$ and will be said to be dual to the norm of K. This can be otherwise defined as*

$$\| u \|' = \max_{x \neq 0} \frac{\mathrm{Re}(u^*x)}{\| x \|} .$$

In the proof, the notation $\nu(u)$ will be retained in order not to prejudice the issue, and it will be sufficient to consider only real vectors. It has already been remarked that for all x in the half-space containing K it is true that

$$u^T x \leqq \nu(u),$$

and this includes all vectors x' with $\| x' \| = 1$. Hence take

$$x' = \frac{x}{\| x \|},$$

and the final assertion follows. Since $\nu(u)$ obviously possesses properties $V1$ and $V2$, it remains to prove $V3$. But

$$\nu(u + v) = \max_{\|x\|=1} (u + v)^T x \leqq \max_{\|x\|=1} u^T x + \max_{\|x\|=1} v^T x = \nu(u) + \nu(v).$$

Moreover, the maximum is always attained since $\| x \| = 1$ defines a closed set. Hence $\nu(u)$ is indeed a norm.

THEOREM 2.5. $(\| x \|')' = \| x \|$. That is to say, duality is a reciprocal relation. It is to be shown that

$$\| x \| = \max_{\|u\|'=1} u^T x,$$

and it is sufficient to consider a vector x for which $\| x \| = 1$. Let u be any vector with $\| u \|' = 1$. Then, as y varies,

$$u^T y = 1$$

is the equation of the support plane in that direction. Since $\| x \| = 1$, x is a point of the boundary of K, and either it lies on that support plane, in which case $u^T x = 1$, or else it lies in the half-space for which $u^T x < 1$.

THEOREM 2.6. *For any pair of dual norms, and any two vectors x and y,*

$$\mathrm{Re}(x^* y) \leqq \| x \|' \cdot \| y \|.$$

This is essentially a restatement of the last part of theorem 2.4.

In defining matrix norms, since a matrix can be regarded as a vector of n^2 elements, it is natural to impose the same conditions, and possibly others. The same conditions will indeed be imposed, and hence matrix norms will have the same continuity properties, in particular, as vector norms. It is convenient to impose also a fourth condition, yielding altogether the following set:

$(M1)$: $\| A \| > 0$ unless $A = 0$;

$(M2)$: if $\alpha \geqq 0$, $\| \alpha A \| = \alpha \| A \|$;

$(M3)$: $\| A + B \| \leqq \| A \| + \| B \|$;

$(M4)$: $\| AB \| \leqq \| A \| \cdot \| B \|$.

The matrices are assumed to be square, which is no real restriction since one can always adjoin null rows or columns. A particular matrix norm will be said to be consistent with a given vector norm in case, for every A and x,

(C): $\| Ax \| \leqq \| A \| \cdot \| x \|$.

Furthermore, the matrix norm is said to be subordinate to a vector norm in case it is consistent and possesses the further property,

(S): For every A there exists an $x \neq 0$ such that $\| A \| = \| A \| \cdot \| x \|$.

Such an x would, of course, vary from matrix to matrix.

THEOREM 2.7. *To every vector norm there corresponds a unique subordinate matrix norm defined by*

$$\| A \| = \max_{\|x\|=1} \| Ax \| = \max_{x \neq 0} \frac{\| Ax \|}{\| x \|}.$$

The following lemma will perhaps aid the intuition:

LEMMA 2.7. *For any convex set K, and any matrix A, let AK denote the set of points $z = Ax$ where $x \varepsilon K$. Then AK is convex.*

In fact, if Ax and Ay are in AK, then, for $0 \leqq \alpha \leqq 1$, the point

$$\alpha Ax + (1 - \alpha)Ay = A[\alpha x + (1 - \alpha)y]$$

is also in AK. This proves the lemma. Now let $\kappa = \kappa(A)$ be the smallest number for which $AK \subset \kappa K$. Then $\kappa(A)$ is the function of A defined in the theorem, and if $\kappa(A)$ is a norm it certainly satisfies condition S. Moreover, $\kappa(A)$ clearly satisfies $M1$ and $M2$. For $M3$,

$$\| (A + B)x \| = \| Ax + Bx \| \leqq \| Ax \| + \| Bx \|$$

by $V3$, and the maximum of a sum cannot exceed the sum of the maxima. For $M4$, if $Bx \neq 0$,

$$\frac{\| ABx \|}{\| x \|} = \frac{\| ABx \|}{\| Bx \|} \frac{\| Bx \|}{\| x \|},$$

and both sides can be maximized. The maximum on the left precludes $Bx = 0$ unless $B = 0$, and for this case $AB = 0$. Hence $\kappa(A)$ is a norm and is clearly unique.

THEOREM 2.8. *Given a vector norm $\| x \|$, the subordinate matrix norm $\| A \|$, and the dual vector norm $\| x \|'$, the matrix norm $\| A \|'$ that is subordinate to the vector norm $\| x \|'$ is*

$$\| A \|' = \| A^* \|.$$

It is immediately obvious that $\nu(A) = \| A^* \|$ is a norm, and it is sufficient to consider the real case. It will be shown that

$$\| A \| = \| A^T \|'.$$

In fact, for all $y \neq 0$ and $x \neq 0$,

$$y^T x \leqq \| A^T y \|' \| x \| \leqq \| A^T \|' \cdot \| y \|' \cdot \| x \|, \qquad \frac{| y^T Ax |}{\| y \|'} \leqq \| A^T \|' \cdot \| x \|.$$

Hence, when the left member is maximized with respect to $y \neq 0$,

$$\| Ax \| \leqq \| A^T \|' \cdot \| x \|,$$

or

$$\frac{\| Ax \|}{\| x \|} \leqq \| A^T \|'.$$

Again, when the left member if maximized with respect to $x \neq 0$,

$$\| A \| \leqq \| A^T \|'.$$

But in like manner one can show that

$$\| A^T \|' \leqq \| A \|.$$

Hence equality must follow.

THEOREM 2.9. *Given a matrix norm* $\| A \|$, *a vector norm with which it is consistent can be defined by*

$$\| x \| = \| (x, 0, 0, \cdots) \|,$$

where on the right one takes the matrix everywhere null except in the first column, which is equal to x.

Consistency is a consequence of $M4$, and the other M-properties imply the corresponding V-properties.

THEOREM 2.10. *Given any nonsingular matrix* G, *if* $\| x \|$ *is a vector norm, then*

$$\| x \|_G = \| G^{-1}x \|$$

is a vector norm; given any matrix norm $\| A \|$, *then*

$$\| A \|_G = \| G^{-1}AG \|$$

is a matrix norm. These will be called the G-transforms of the original norms and relations of consistency and subordination are preserved under such transformations.

This is verified directly.

THEOREM 2.11. *Dual to the G-transform of a norm is the dual norm transformed by* $(G^{-1})^*$. *Hence duality is preserved only under unitary transformations.*

This follows from the fact that

$$x^*y = (x^*G)(G^{-1}y) = (G^*x)^*G^{-1}y.$$

THEOREM 2.12. *If the convex body* K *is bounded by planes, and if* $\| x \|$ *is the associated vector norm and* $\| A \|$ *the subordinate matrix norm, then at least one of the vectors* x *for which* $\| x \| = 1$ *and* $\| Ax \| = \| A \|$ *represents a corner of* K.

For if κ is the smallest number for which $\kappa K \supset AK$, then the boundary of κK must contain at least one corner point $y = Ax$ of AK. But then x is a corner point of K.

THEOREM 2.13. *If* A *is nonsingular, then* AK *is of dimension* n. *Hence there is a largest* $\kappa \neq 0$ *for which* $\kappa K \subset AK$, *and* $\kappa^{-1} = \| A^{-1} \|$.

The existence of the $\kappa \neq 0$ is obvious. Let $y = Ax$ be any point common to the boundaries of κK and AK. Then $\| x \| = 1$, and if $\| x' \| = 1$, $\| Ax' \| \geqq \kappa$, since the point $y' = Ax'$ can only lie outside κK or on its boundary. Hence

$$\kappa = \min_{\|x'\|=1} \| Ax' \| = \min_{x' \neq 0} \frac{\| Ax' \|}{\| x' \|} = \min_{y' \neq 0} \frac{\| y' \|}{\| A^{-1}y' \|}.$$

But since A is nonsingular, $x' \neq 0$ if and only if $y' \neq 0$, whence

$$\kappa^{-1} = \max_{y' \neq 0} \frac{\| A^{-1}y' \|}{y'} = \| A^{-1} \|.$$

The fundamental problem is, given a matrix A, to construct a norm for which $\| A \|$ is as small as possible. It will be shown that $\| A \|$ can be made arbitrarily close to the spectral radius $\rho(A)$, and, indeed, this can be done with symmetric norms. But practical methods for constructing such norms are not available except for $A \geqq 0$.

3. *Examples of Norms*

The most commonly used norm is the Euclidean norm $\| x \|_E$, defined by

$$\| x \|_E^2 = x^*x,$$

and associated with the unit sphere as K. This norm is self-dual, and the inequality of theorem 2.6 is the Schwarz inequality. The Euclidean matrix norm $\| A \|_E$, defined by

$$\| A \|_E^2 = \sum \| a_{i.} \|^2 = \sum \| a_{.j} \|^2$$

is consistent but not subordinate. That it is consistent follows from the relation

$$\| Ax \|_E^2 = \sum (a_{i.}x)^2 \leqq \| x \|_E^2 \sum \| a_{i.} \|_E^2 = \| x \|_E^2 \| A \|_E^2,$$

the inequality coming from the Schwarz inequality. That it cannot be subordinate follows from the fact that

$$\| I \|_E = n^{\frac{1}{2}},$$

together with the following lemma:

LEMMA 3.1. *If there exists a vector norm to which the matrix norm* $\| A \|$ *is subordinate, then* $\| I \| = 1$.

To see this, apply $M4$. Then if

$$\| Ix \| = \| I \| \cdot \| x \|,$$

the lemma follows, since $Ix = x$.

The matrix norm subordinate to the Euclidean vector norm is the spectral norm $\| A \|_S$, defined as the largest singular value. This can be seen most easily by applying theorem 2.10: Observe first that if V is any unitary matrix, then

$$\| Vx \|_E = \| x \|_E,$$

since

$$(Vx)^*Vx = x^*V^*Vx = x^*x.$$

Hence to maximize x^*A^*Ax subject to $\| x \|_E = 1$ is equivalent to maximizing $x^*V^*A^*AVx$ subject to the same conditions, where V is any unitary matrix. In particular V can be the matrix of proper vectors of A^*A, so that

$$V^*A^*AV = \Lambda(A^*A) = \operatorname{diag} [\lambda_1(A^*A), \lambda_2(A^*A), \cdots].$$

Hence the problem is reduced to that of maximizing $x^*\Lambda x$, subject to $x^*x = 1$.

But if

$$\lambda_1(A^*A) \geqq \lambda_2(A^*A) \geqq \cdots,$$

then $x^*\Lambda x$ is a weighted mean of the $\lambda_i(A^*A)$ which takes on its maximum of $\lambda_1(A^*A) = \sigma_1^2(A)$ for $x = e_1$.

If G is nonsingular, the G-transform of $\| x \|_E$ is given by

$$\| G^{-1}x \|_E^2 = (G^{-1}x)^*G^{-1}x = x^*Hx,$$

where H is a positive definite matrix,

$$H = (GG^*)^{-1}.$$

This norm is not self-dual unless G is unitary, and in that case $H = I$ and the norm is unchanged. The subordinate matrix norm is $\| G^{-1}AG \|_S$, i.e. the largest singular value of $G^{-1}AG$, or the square root of the largest root of

$$\det (G^*A^*HAG - \lambda I) = 0,$$

or, equivalently, of

$$\det (A^*HA - \lambda H) = 0.$$

The dual norms are similarly expressed in terms of H^{-1}.

For the next example, consider first the real case, and let K be the cube with faces

$$\xi_i = \pm 1.$$

The associated norm will be called the e-norm, and is defined

$$\| x \|_e = \max_i | \xi_i |.$$

More abstractly it can be defined by the two conditions

(i) $$| x | \leqq e\| x \|_e ;$$

(ii) $$\text{if } \varepsilon e \geqq | x |, \qquad \text{then } \varepsilon \geqq \| x \|_e .$$

It is easy to verify that the dual norm is

$$\| x \|_e' = e^T| x |,$$

and the faces of K' are defined by the equations

$$\sum \pm \xi_i = 1,$$

for all possible choices of the signs. Subordinate to the vector e-norm is the matrix e-norm,

$$\| A \|_e = \| | A | e \|_e ,$$

and to the e'-norm

$$\| A \|_e' = \| | A^T | e \|_e .$$

In words, one sums the absolute values in each row for the e-norm, or each column for the e'-norm, and the largest of these is the value of the norm.

Let $G = \operatorname{diag}(\gamma_1, \gamma_2, \cdots, \gamma_n) \geqq 0$ be any non-negative, nonsingular diagonal matrix, and let

$$g = Ge.$$

Then $g > 0$. The G-transform will be called a g-norm. It satisfies

$$\| x \|_g = \max_i \frac{|\xi_i|}{\gamma_i}.$$

Or, it can be defined by

(i) $$| x | \leqq g \| x \|_g ;$$

(ii) $$\text{if } \gamma g \geqq | x |, \qquad \text{then } \gamma \geqq \| x \|_g .$$

The associated K has the faces

$$\xi_i = \pm \gamma_i .$$

Dual to the g-norm is the g'-norm:

$$\| x \|_g' = g^T | x |.$$

The faces of the associated K' have the equations

$$\sum \pm \gamma_i \xi_i = 1.$$

Subordinate to the vector g-norm is

$$\| A \|_g = \| \, | A | \, g \|_g .$$

For complex matrices the e-norm has two natural generalizations. One comes by applying the ordinary e-norm in the $2n$-space of real vectors $\begin{pmatrix} x \\ y \end{pmatrix}$ defined by the space of complex vectors $x + iy$. The subordinate matrix norm for complex matrices $A + iB$ is the e-norm of $| A | + | B |$. The other is perhaps more natural and will be assumed here, and this applies the formulas of the real case. Thus

$$\| x \|_e = \max_i | \xi_i |,$$

where ξ_i is complex and the absolute value signs signify the modulus. In the $2n$-dimensional space the associated K is no longer bounded by plane faces, but by circular cylinders

$$| \xi_i | = 1.$$

Certain relations of inequality among norms can be established directly:

THEOREM 3.1 *For any vector x,*

$$\| x \|_e \leqq \| x \|_E \leqq \| x \|_e' \leqq n^{\frac{1}{2}} \| x \|_E \leqq n \| x \|_e .$$

The third inequality follows from the ordinary Schwarz inequality:

$$| x^T | e \leqq \| x \|_E \cdot \| e \|_E = n^{\frac{1}{2}} \| x \|_E .$$

THEOREM 3.2. *If V is unitary and $g > 0$, then*

$$\| V \|_g = \| g \|_E \max_i \gamma_i^{-1}.$$

In particular

$$\| V \|_e \leqq n^{\frac{1}{2}}.$$

If V is unitary, and v^* any row of V, then

$$v^* v = | v^T | \cdot | v | = 1.$$

But by the ordinary Schwarz inequality

$$| v^T | g \leqq \| v \|_E \cdot \| g \|_E = \| g \|_E .$$

Hence

$$| v | g \leqq \nu g$$

if

$$\nu \leqq \| g \|_E \max_i \gamma_i^{-1}.$$

THEOREM 3.3. *Given any vector norm $\| x \|$, with associated convex body K_1, for fixed positive κ let $K_2 = K_1$, and define the vector norm $\| x \|_2$ associated with K_2. Then*

$$\| x \|_2 = \kappa^{-1} \| x \|_1,$$

but

$$\| A \|_2 = \| A \|_1 ,$$

where $\| A \|_1$ and $\| A \|_2$ are the subordinate matrix norms.

The proof is immediate from geometrical considerations.

THEOREM 3.4. *For any matrix A,*

$$n^{-\frac{1}{2}} \| A \|_S \leqq \| A \|_e \leqq n^{\frac{1}{2}} \| A \|_S , \qquad n^{-\frac{1}{2}} \| A \|_S \leqq \| A \|_e' \leqq n^{\frac{1}{2}} \| A \|_S .$$

First, let $x^* x = 1$, and $\| Ax \|_E = \| A \|_S$. Such a vector exists since the spectral norm is subordinate to the Euclidean vector norm. Moreover $| x | \leqq e$. Hence, by theorem 3.1,

$$\| Ax \|_E \leqq n^{\frac{1}{2}} \| Ax \|_e \leqq n^{\frac{1}{2}} \| A \|_e \| x \|_e \leqq n^{\frac{1}{2}} \| A \|_e .$$

Next, let $\| x \|_e = 1$ and $\| Ax \|_e = \| A \|_e$. Then $\| x \|_E \leqq n^{\frac{1}{2}}$. But

$$\| Ax \|_e \leqq \| Ax \|_E \leqq \| A \|_S \| x \|_E .$$

This completes the proof for $\| A \|_e$, and for $\| A \|_e'$ apply the same argument to A^*.

For each of the four relations of the theorem, there exists a matrix A for which the equality holds. Hence these relations are the sharpest possible.

THEOREM 3.5. *If A is Hermitian, then*

$$\| A \|_S \leqq \| A \|_e = \| A \|_e' \leqq n^{\frac{1}{2}} \| A \|_S .$$

For if A is Hermitian, then $\| A \|_S = \rho(A)$, and it will be shown below that $\rho(A) \leqq \| A \|$ for any norm and any matrix A.

THEOREM 3.6. *For any matrix A and any norm,*

$$\| A \|_s^2 \leqq \| A \| \cdot \| A \|'.$$

For

$$\| A \|_s^2 = \rho(AA^*) \leqq \| AA^* \| \leqq \| A \| \cdot \| A^* \| = \| A \| \cdot \| A \|'.$$

4. *Norms and the Spectral Radius*

If the proper values of a matrix B are $\lambda_i(B)$ and these are ordered in magnitude,

$$\lambda_1(B) \mid \geqq \mid \lambda_2(B) \mid \geqq \cdots \geqq \mid \lambda_n(B) \mid,$$

then

$$\rho(B) = \mid \lambda_1(B) \mid$$

is the spectral radius of B, since all proper values of B lie in or on the circle of radius ρ. Often iterative methods require the formation, implicitly or explicitly, of sequences of vectors of the form

$$s_\nu = Bs_{\nu-1} = B^\nu s_0 , \tag{4.1}$$

with some matrix B and an arbitrary s_0, where

$$s_\nu = x^* - x_\nu$$

represents the deviation of a current approximation from the true solution.

THEOREM 4.1. *The sequence of vectors s_ν defined by* (4.1) *vanishes in the limit independently of s_0 if and only if the sequence of powers B^ν vanishes in the limit. For this it is sufficient that there be some norm such that*

$$\| B \| < 1.$$

There exists a vector norm with which the matrix norm is consistent by theorem 2.9 and if the s_ν vanish in some norm they vanish in all. The sufficiency therefore follows from

$$\| s_\nu \| \leqq \| B^\nu \| \cdot \| s_0 \| \leqq \| B \|^\nu \cdot \| s_0 \|.$$

The value of $\| B \|$ provides also a measure of the rate of convergence when convergence occurs. Actually convergence occurs if and only if $\rho(B) < 1$, but in general ρ is not easily computed. It is desirable, therefore, to know how closely ρ may be approximated by a norm. First, however, the following partial theorem will be proved:

THEOREM 4.2. *For the sequence of matrix powers B^ν to vanish in the limit, it is necessary that $\rho(B) < 1$.*

For let

$$V^{-1}BV = L$$

where L is the Jordan normal form of B. Then

$$B^\nu = VL^\nu V^{-1}.$$

The elements of V and V^{-1} are independent of ν, whereas λ_1^ν occurs in the diagonal of L^ν, and elsewhere there occur lower powers of λ_1 multiplied by binomial coefficients which become infinite as ν becomes infinite. Hence elements of L^ν, and therefore of B^ν, become infinite if $|\lambda_1| \geqq 1$.

THEOREM 4.3. *For any norm and any matrix B,*

$$\| B \| \geqq \rho(B).$$

For suppose $\beta = \| B \| < \rho(B)$. Choose $\epsilon > 0$ so that $\beta + \epsilon < \rho$, and let $B' = (\beta + \epsilon)^{-1}B$. Then $\| B' \| < 1$ and the sequence of powers B'^ν vanishes in the limit. But

$$\rho(B') = \frac{\rho(B)}{\beta + \epsilon} > 1$$

and this is impossible.

COROLLARY 4.3. *If B is symmetric, then $\| B \|_s \leqq \| B \|_e = \| B \|_e'$.*

In fact, if B is symmetric, $\| B \|_s = \rho(B)$.

LEMMA 4.4. *Let T be a triangular matrix and τ its diagonal element of greatest magnitude. Then for any $\epsilon > 0$ there exists a norm such that*

$$\| T \| \leqq \tau + \epsilon.$$

For definiteness let T be a lower triangle, and consider

$$D = \operatorname{diag}(1, \delta, \delta^2, \cdots, \delta^{n-1}).$$

In $D^{-1}TD$ the elements in the diagonal are the same as those in T; all elements in the first subdiagonal are divided by δ, all those in the second by $\delta^2, \cdots$. By choosing δ sufficiently large, the sum of the magnitudes of the off-diagonal elements in any row can be made less than ϵ. Hence apply the e-norm or the spectral norm to $D^{-1}TD$.

THEOREM 4.4 *For any matrix B and any $\epsilon > 0$, there exists a norm for which*

$$\| B \| \leqq \epsilon + \rho(B).$$

Moreover, if for each $\lambda_i(B)$ such that $|\lambda_i(B)| = \rho$, the number of independent proper vectors belonging to it is equal to its multiplicity, then there exists a norm for which $\| B \| = \rho(B)$.

The last condition holds, in particular, for all diagonalizable matrices. The theorem follows from lemma 4.4, theorem 2.10, and the fact that the Jordan normal form is a triangular matrix.

THEOREM 4.5. *For the sequence of matrix powers B^r to vanish in the limit it is necessary and sufficient that $\rho(B) < 1$.*

This supplements theorem 4.2 which stated only the necessity. If $\rho < 1$, let $0 < \epsilon < 1 - \rho$. Then the norm of theorem 4.4 satisfies $\| B \| \leqq \rho + \epsilon < 1$, which establishes convergence.

THEOREM 4.6. *For any g-norm and any x and B,*

$$\| x \|_g = \| \, | x | \, \|_g , \qquad \| B \|_g = \| \, | B | \, \|_g .$$

This follows from the fact that only absolute values occur in the definitions. In many practical situations the matrix B of interest is non-negative, $B \geqq 0$. Hence some consideration of the g-norms of non-negative matrices is in order.

A matrix B is said to be reducible in case for some permutation matrix P,

$$P^T B P = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix},$$

where B_{11} and B_{22} are square submatrices. If no such permutation matrix exists, the matrix is irreducible. The examination of a reducible matrix reduces to the separate examination of the submatrices B_{11} and B_{22}.

THEOREM 4.7. *Let $B \geqq 0$ be irreducible, and let $g > 0$ be any positive vector. If g is not a proper vector of B, then a $g' > 0$ can be found such that*

$$\| B \|_{g'} < \| B \|_g .$$

Let $\| B \|_g = \gamma$. Then

$$Bg \leqq \gamma g,$$

and since g is not a proper vector,

$$Bg \neq \gamma g.$$

If the inequalities are written in scalar form, some will express equalities, since otherwise γ would exceed $\| B \|_g$, and some will express strict inequalities. Then there exists a permutation matrix P such that

$$P^T B P = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, \qquad P^T g = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix},$$

$$B_{11} g_1 + B_{12} g_2 < \gamma g_1 , \qquad B_{21} g_1 + B_{22} g_2 = \gamma g_2 .$$

Since B is irreducible, $B_{21} \neq 0$. Then for sufficiently small $\epsilon > 0$,

$$(1 - \epsilon) B_{11} g_1 + B_{12} g_2 < (1 - \epsilon) \gamma g_1 ,$$

$$(1 - \epsilon) B_{21} g_1 + B_{22} g_2 \leqq \gamma g_2 ,$$

where the strict inequalities remain inequalities, while at least one of the equalities becomes an inequality. That is to say, if

$$g_1' = (1 - \epsilon) g_1 ,$$

then

$$B_{11}g_1' + B_{12}g_2 < \gamma g_1',$$
$$B_{21}g_1' + B_{22}g_2 \leqq \gamma g_2 .$$

If any equalities remain, perform a new permutation and repeat. Eventually one arrives at a g' for which all relations are of strict inequality.

Hereafter a transformation by a permutation matrix will be called simply a permutational transformation.

THEOREM 4.8. *Let $B \geqq 0$ be a non-negative, irreducible matrix. The spectral radius $\rho(B) = \beta$ is a proper value, and belonging to it is a unique proper vector b, and $b > 0$. Moreover,*

$$\| B \|_b = \beta.$$

The process described in theorem 4.7 can be repeated to provide a sequence of vectors g', g'', $\cdots$, and associated norms γ', γ'', $\cdots$ with $\gamma' > \gamma'' > \cdots$. The sequence of γ's is properly monotonically decreasing and is bounded below by β, hence can only terminate on reaching β as a limit, with a non-negative vector b such that

$$Bb = \beta b.$$

Hence β is a proper value. If there are null elements in b, perform a permutational transformation if necessary so that

$$b = \begin{bmatrix} b_1 \\ 0 \end{bmatrix}, \qquad b_1 > 0,$$

and denote the permuted matrix by B. Let

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}.$$

Then since $Bb = \beta b$, one has

$$B_{11}b_1 = \beta b_1$$
$$B_{21}b_1 = 0.$$

But $b_1 > 0$, $B_{21} \geqq 0$, and the last relation can hold only if $B_{21} = 0$, which is contrary to the hypothesis of irreducibility. Hence $b > 0$ and $\| B \|_b$ is defined. If $b' > 0$ is any other proper vector belonging to β, then for suitable α,

$$b - \alpha b' \geqq 0$$

and has at least one null element. But this is also a proper vector belonging to β. Hence $b - \alpha b' = 0$.

THEOREM 4.9. *If $B \geqq B' \geqq 0$, then $\rho(B) \geqq \rho(B')$. In particular, B' can be any submatrix of B augmented by null rows and columns to make the order that of B.*

If B is reducible, then by a permutational transformation it can be given the

form

$$B = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}$$

and the same permutational transformation applied to B' gives it the form

$$B' = \begin{bmatrix} B'_{11} & B'_{12} \\ 0 & B'_{22} \end{bmatrix}.$$

Since

$$\det(\lambda I - B) = \det(\lambda I - B_{11}) \det(\lambda I - B_{22}),$$

it follows that

$$\rho(B) = \max[\rho(B_{11}), \rho(B_{22})].$$

Hence it is sufficient to prove the theorem for the case of B irreducible. But if $\beta = \rho(B)$ and $Bb = \beta b$, then $B'b \leqq \beta b$. Hence

$$\| B' \|_b \leqq \beta,$$

and therefore

$$\rho(B') \leqq \beta.$$

THEOREM 4.10. *Let* $| A | \leqq B$. *Then* $\rho(A) \leqq \rho(B)$. *If* B *is irreducible, then* $\rho(A) = \rho(B)$ *implies* $| A | = B$.

Since, for any $g > 0$, $\| A \|_g = \| | A | \|_g$, the first part follows from theorem 4.9. Now let

$$Aa = \alpha a,$$

where $\alpha = \lambda(A)$ is any proper value of A. Then

$$| \alpha | \cdot | a | \leqq | A | \cdot | a | \leqq B | a |.$$

If B is irreducible, there exists a positive row-vector $b' > 0$ such that

$$b'B = \beta b', \qquad \beta = \rho(B),$$

and b' can be normalized so that $b' | a | = 1$. On multiplying the first and last members of the above inequalities by b', one has that

$$| \alpha | \leqq \beta.$$

Suppose $| \alpha | = \beta$. Then equalities must hold throughout:

$$| \alpha | \cdot | a | = | A | \cdot | a | = B | a |,$$

$| a |$ is a proper vector of B belonging to B, whence $| a | > 0$. Therefore

$$(B - | A |) | a | = 0,$$

and since $B \geqq | A |$, this implies $B = | A |$.

The g-norms are easiest to compute, but unfortunately any g-norm of A may be far in excess of $\rho(A)$. Thus if A is unitary, $\rho(A) = 1$, and orthogonal matrices can be constructed for which $\rho(|A|) = n^{\frac{1}{2}}$. But fortunately one is often interested in non-negative matrices, in which case the above theorems show that g-norms provide all needed information.

THEOREM 4.11. *If, with any norm* $\| B \| < 1$, *then* $I - B$ *is nonsingular and*

$$(I - B)^{-1} = I + B + B^2 + \cdots,$$

the series on the right converging. Conversely, if the series on the right converges, then $\rho(B) < 1$ *and for some norm it is true that* $\| B \| < 1$.

If, for any norm $\| B \| < 1$, then $\rho(B) < 1$ and $1 - \lambda(B) \neq 0$ for every proper value $\lambda(B)$. Hence $I - B$ is nonsingular. Since

$$(I - B)^{-1} - (I + B + \cdots + B^\nu) = (I - B)^{-1}B^{\nu+1}$$

identically, therefore

$$\| (I - B)^{-1} - (I + B + \cdots + B^\nu) \| \leqq \| (I - B)^{-1} \| \cdot \| B \|^{\nu+1}$$

and the right member vanishes in the limit.

Conversely, let

$$S_\nu = I + B + \cdots + B^\nu,$$

and let the S_ν approach the limit S. Then

$$SB^{\nu+1} = S - S_\nu,$$

and since the right member has the limit 0, so has the left. If both members vanish for any ν, they vanish for any $\nu' \geqq \nu$, whence $B^{\nu+1} = 0$, B is nilpotent, and every proper value $\lambda(B) = 0$. Hence $\rho(B) = 0$. Otherwise a slight modification of the argument given for theorem 4.2 shows that $\rho(B) < 1$.

THEOREM 4.12. *Let* $B \geqq B_1' \geqq 0$, *equalities excluded. Let* B *be irreducible and* $B_1 = B - B_1'$. *If* $\rho(B) \leqq 1$, *then* $C_1 = (I - B_1)^{-1}B_1'$ *exists and either* $\rho(C_1) = \rho(B) = 1$ *or* $\rho(C_1) < \rho(B) < 1$. *If* $\rho(B) > 1$ *but* C_1 *exists and* $C_1 \geqq 0$, *then* $\rho(C_1) \geqq \rho(B)$. *In particular, if* B_1 *is nilpotent, then* C_1 *always exists and* $C_1 \geqq 0$.

In practice B_1 is usually null except in elements below the diagonal, and hence is nilpotent. In general, let $\gamma = \rho(C_1)$ and

$$\gamma c = C_1 c = (I - B_1)^{-1}B_1'c \geqq 0.$$

Then

$$\gamma(I - B_1)c = B_1'c,$$

$$\gamma c = (\gamma B_1 + B_1')c.$$

Hence $\gamma = \rho(C_1) = \rho(\gamma B_1 + B_1')$, and $\gamma B_1 + B_1'$ is irreducible if B is irreducible. But either

$$\gamma = 1, \qquad \gamma B_1 + B_1' = B, \qquad \rho(B) = \gamma,$$

or

$$\gamma > 1, \qquad \gamma B_1 + B_1' \geqq B, \qquad \rho(\beta) < \gamma,$$

or else

$$\gamma < 1, \qquad \gamma B_1 + B_1' \leqq B, \qquad \rho(B) > \gamma.$$

Now let $\beta = \rho(B) \leqq 1$, and let

$$\beta b = Bb > 0.$$

Then $(I - \beta_1^{-1}B_1)^{-1} = I + \beta_1^{-1}B_1 + \beta_1^{-2}B_1^2 + \cdots \geqq I + B_1 + B_1^2 + \cdots = (I - B_1)^{-1} = C_1$. But

$$\beta(I - \beta^{-1}B_1)b = B_1'b,$$

$$\beta b = (I - \beta^{-1}B_1)^{-1}B_1'b \geqq C_1 b.$$

Hence $\rho(C_1) \leqq \rho(B)$, and the equality can hold only if $\beta = 1$. Existence is assured in this case since

$$\rho(\beta^{-1}B_1) = \beta^{-1}\rho(B_1) < \beta^{-1}\rho(B).$$

Note that as in the proof of theorem 4.8 one can show that $c > 0$.

THEOREM 4.13. *Let $B \geqq B_1' \geqq B_2' \geqq 0$, equalities excluded, with B irreducible and $\rho(B) < 1$. Let $B_i = B - B_i'$, $C_i = (I - B_i)^{-1}B_i'$. Then $\rho(B) > \rho(C_1) > \rho(C_2)$.*

By theorem 4.12, $\gamma_i = \rho(C_i) < 1$. Let

$$\gamma_1 c_1 = C_1 c_1 > 0,$$

where the strict inequality follows from the remark made above. Then

$$\gamma_1 c_1 = (\gamma_1 B_1 + B_1')c_1 .$$

Let

$$B' = B_1' - B_2' = B_2 - B_1 \geqq 0.$$

Then

$$\gamma_1 B_1 + B_1' = \gamma_1 B_2 + B_2' + (1 - \gamma_1)B' \geqq \gamma_1 B_2 + B_2'.$$

Hence

$$\gamma_1 c_1 \geqq (\gamma_1 B_2 + B_2')c_1 .$$

From the proof of theorem 4.12 it can be seen that

$$\rho(C_2) = \rho(\gamma_2 B_2 + B_2'),$$

and from the preceding result it appears that

$$\rho(\gamma_1 B_2 + B_2') < \gamma_1 .$$

Let

$$\varphi(\gamma) = \rho(B_2 + \gamma^{-1}B_2').$$

Then

$$\varphi(\gamma_1) < 1$$

and φ decreases monotonically as γ increases. But if

$$\gamma_2 = \rho(C_2)$$

then

$$\rho(\gamma_2 B_2 + B_2') = \gamma_2 ,$$

$$\varphi(\gamma_2) = 1.$$

Hence $\gamma_2 < \gamma_1$. Hence the theorem.

THEOREM 4.14. *Let* $B = B_1 + B_2$, $|B| = |B_1| + |B_2|$. *Then if* $\rho(|B|) < 1$, $C = (I - B_1)^{-1}B_2$ *exists and* $\rho(C) < \rho_1(|B|)$.

This is a corollary to theorem 4.12.

A system of equations can generally be written in the form

$$(I - B)x = h,$$

with B having a null diagonal. The most common iterative methods make use of either of the two sequences

$$x_{\nu+1} = h + Bx_\nu ,$$

or

$$x_{\nu+1} = k + Cx_\nu ,$$

where

$$B = B_1 + B_2 , \qquad k = (I - B_1)^{-1}h, \qquad C = (I - B_1)^{-1}B_2 .$$

Generally (in the so-called Gauss-Seidel method) B_1 is a lower triangle, B_2 an upper triangle. The theorems 4.12 and 4.13 say that when $B \geqq 0$, either both iterations converge or both diverge, for arbitrary x_0, and that if both converge, the second converges more rapidly. Moreover, the transfer of non-null elements from B_2 to B_1 accelerates convergence whenever convergence occurs. When the condition $B \geqq 0$ fails, however, the results are much less precise, and, in fact, examples can be constructed for which either iteration converges while the other diverges. Intuitively, though, one expects the second iteration to converge more rapidly, and in most practical cases this is so. Some other fairly general cases will be examined now.

THEOREM 4.15. *Let* B *have the form*

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

where each 0 *represents a square submatrix. Let*

$$B_1 = \begin{bmatrix} 0 & 0 \\ Q & 0 \end{bmatrix}, \qquad B_2 = \begin{bmatrix} 0 & R \\ 0 & 0 \end{bmatrix}.$$

Then $C = (I - B_1)^{-1}B_2$ exists and $\rho(C) = \rho^2(B)$. Moreover, if $\lambda(B)$ is any proper value of B, then $\mu = \lambda^2(B)$ is a proper value of C; and if $\mu(C) = \lambda^2$ is a non-null proper value of C, then $\pm\lambda$ are proper values of B.

One verifies that $I + B_1 = (I - B_1)^{-1}$ which proves the existence of C. Proof of the other assertions is based upon a well known property of determinants:

LEMMA 4.15. *If A_{11} is nonsingular, then*

$$\det\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \det(A_{11}) \cdot \det(A_{22} - A_{21} A_{11}^{-1} A_{12}).$$

In fact,

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} I & -A_{11}^{-1} A_{12} \\ 0 & I \end{bmatrix} = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} - A_{21} A_{11}^{-1} A_{12} \end{bmatrix}.$$

Hence take determinants of both sides.

Now it follows directly from the lemma that det $(\lambda I - B)$ is equal to a power of λ multiplied by det $(\lambda^2 I - QR)$. Next

$$C = \begin{bmatrix} 0 & R \\ 0 & QR \end{bmatrix}$$

and det $(\mu I - C)$ is equal to a power of μ multiplied by det $(\mu I - QR)$. The theorem now follows immediately.

Note that if $R = Q^T$, B is symmetric, and $I - B$ is positive definite if and only if $\rho(B) < 1$, hence if and only if $\rho(C) < 1$.

A matrix A is said to have property (A) if there exists a permutation matrix P such that

$$P^T A P = D - B,$$

where D is diagonal and B has the form of theorem 4.15. This transformation permutes the diagonal elements among themselves. For purposes of equation solving or inverting, if D is nonsingular, there is no restriction in assuming $D = I$.

THEOREM 4.16. *If $\rho(A) < 1$, then there exists a positive definite matrix H such that $H - A^T HA$ is positive definite. Conversely, if H and $H - A^T HA$ are both positive definite, then $\rho(A) < 1$.*

If H is positive definite, then there exists a matrix G such that

$$H = G^T G.$$

If $H - A^T HA$ is positive definite, then for every $x \neq 0$,

$$x^T(H - A^T HA)x > 0,$$

$$(Gx)^T(Gx) > (GAx)^T(GAx).$$

Let

$$y = Gx.$$

Then

$$y^T y > (GAG^{-1}y)^T(GAG^{-1}y),$$

$$\| y \|_E > \| GAG^{-1}y \|_E ,$$

for every $y \neq 0$. Hence $\| GAG^{-1} \|_S < 1$ and $\rho(A) < 1$.

Conversely if $\rho(A) < 1$, there exists a nonsingular G such that $\| GAG^{-1} \|_S < 1$ (theorem 4.3). The above proof goes through in reverse.

LEMMA 4.17. *If A and S are symmetric, $A = S - B - B^T$, and*

$$C = (S - B)^{-1}B^T$$

exists, then

$$A - C^TAC = (I - C^T)S(I - C).$$

This is verified directly.

THEOREM 4.17. *If A, S, B and C are as in the lemma, and S is positive definite, then $\rho(C) < 1$ if and only if A is positive definite.*

If A is positive definite, then $I - C$ is nonsingular. For suppose

$$x = Cx = (S - B)^{-1}B^Tx.$$

Then

$$\begin{aligned} (S - B)x &= B^Tx, \\ (S - B - B^T)x &= 0, \\ Ax &= 0, \end{aligned}$$

and hence $x = 0$. Hence $(I - C^T)S(I - C)$ is positive definite, and by theorem 4.16, $\rho(C) < 1$.

Next, suppose $\rho(C) < 1$. Then

$$P = (I - C^T)S(I - C)$$

is positive definite, and

$$\begin{aligned} A &= P + C^TAC \\ &= P + C^TPC + (C^2)^TAC^2 \\ &= P + C^TPC + (C^2)^TPC^2 + \cdots \end{aligned}$$

Now if

$$C = V\Lambda V^{-1},$$

where Λ is the Jordan normal form, it can be shown that there exists W such that

$$C^T = W\Lambda W^{-1}.$$

Hence if $U = WV^{-1}$, then

$$C^T = UCU^{-1}.$$

Hence

$$A = P + UCU^{-1}PC + UC^2U^{-1}PC^2 + UC^3U^{-1}PC^3 + \cdots .$$

Therefore A is expressible as a converging series in positive definite matrices. Hence A is itself positive definite.

In most applications, S is the diagonal of A, and $S - B$ the lower triangle of A. However, S can be made up of blocks of submatrices along the diagonal, and B taken to be the lower triangle of $S - A$.

5. *Errors and Iterations*

Given a system of equations

$$Ax = h, \tag{5.1}$$

let x_ν be supposed to approximate the true solution x, and let

$$s_\nu = x - x_\nu, \qquad r_\nu = h - Ax_\nu = A(x - x_\nu) = As_\nu.$$

Then

$$s_\nu = A^{-1}r_\nu,$$

and, for any consistent norms,

$$\| s_\nu \| \leqq \| A^{-1} \| \cdot \| r_\nu \|. \tag{5.3}$$

A rigorous appraisal of the error s_ν requires knowing something of A^{-1} as well as of r_ν. If A is known only as a numerical matrix, one can, in general, evaluate $\| A^{-1} \|$ only after evaluating A^{-1}. But in many cases the matrices A which arise in the solution of linear differential equations have a rather simple structure, and certain norms can be evaluated, at least approximately, without knowing A^{-1} explicitly.

For matrices known only numerically, the computed inverse is not necessarily sufficiently close to the true inverse to yield a value of $\| A^{-1} \|$ directly. However, a rigorous upper bound is available.

THEOREM 5.1. *Let $H = I - AC$, and, in any norm, let $\| H \| < 1$. Hence C is an approximation to A^{-1}. Then*

$$\| A^{-1} \| \leqq \frac{\| C \|}{1 - \| H \|}, \qquad \| A^{-1} - C \| \leqq \frac{\| CH \|}{1 - \| H \|}.$$

COROLLARY 5.1. *If $K = I - CA$ and $\| K \| < 1$, then*

$$\| A^{-1} \| \leqq \frac{\| C \|}{1 - \| K \|}, \qquad \| A^{-1} - C \| \leqq \frac{\| KC \|}{1 - \| K \|}.$$

Evidently

$$\begin{aligned} A^{-1} &= C(I - H)^{-1} \\ &= C + CH + CH^2 + \cdots \end{aligned}$$

and since $\| H \| < 1$ the series converges. Hence

$$\| A^{-1} \| \leqq \| C \| + \| C \| \cdot \| H \| + \| C \| \cdot \| H \|^2 + \cdots = \| C \| [1 - \| H \|]^{-1}.$$

The other relations are obtained in a similar way.

In actual application one should be warned that because of rounding the computed H will, in general differ from the true one, and there is actually a case on record in which the norm of the computed H gave too small an estimate of $\| A^{-1} - C \|$. For the e-norms, at least, it is possible to place a bound upon the difference between the norms of the true H and the computed H, by considering the programming of the matrix product AC. Hence a rigorous, though possibly pessimistic, bound for $\| A^{-1} - C \|$ is still available.

The main objective in this section, however, will be to consider certain special matrices which arise in the finite difference schemes for solving differential equations. Only linear partial differential equations will be considered. Assume, first, that the equation is two dimensional, and that the solution is required over a rectangular region. It is no restriction to take two of the sides along the two axes. For hyperbolic and parabolic equations, the values of the required solution, or its derivatives, or both, are normally prescribed along three of the sides (initial and boundary values); for elliptic equations the function is prescribed along all four sides.

To set up the approximating difference equations, one subdivides the region by equally spaced vertical lines, and equally spaced horizontal lines, say n vertical lines with spacing $\triangle x$, and m horizontal lines with spacing $\triangle t$ or $\triangle y$. These lines intersect in nm points forming a lattice, and one seeks to evaluate the required function at these points. To do this, one approximates the derivatives in the equation by finite differences according to some suitable method of interpolation, and so obtains a set of, altogether, nm difference equations to be solved for the nm functional values. These equations are linear and algebraic.

In abbreviated symbolic form, let v be the function defined by the differential equations, and let the equation be written

$$P(D)v = f,$$

where $P(D)$ represents a differential operator operating upon v, and f is some given function, possibly zero. Let u represent the function defined by the difference equations, hence defined only at the lattice points, and let

$$Q(\triangle)u = f \tag{5.4}$$

represent the system of difference equations. Finally let $w = v - u$ represent the truncation error. This, like u, is defined only at the lattice points. Then

$$Q(\triangle)w = [Q(\triangle) - P(D)]v.$$

Now the difference operator $Q(\triangle)$ can be expanded by known methods in terms of the derivatives, and hence $Q(\triangle) - P(D)$ can be expressed as a differential operator, say

$$Q(\triangle) - P(D) = R(D).$$

Hence

$$Q(\triangle)w = R(D)v. \tag{5.5}$$

While v is the required function, and hence not known, it may be that its existence and continuity properties are known to such an extent that by invoking the theorem of the mean, $R(D)v$ is known to be bounded and even has known bounds. These bounds are, of course, functions of n and m, or, equivalently, of $\triangle x$ and $\triangle y$. If so, then w satisfies a system of difference equations for which the right-hand members can at least be bounded. It is important to note that if we consider the right member of (5.5) to be known, then (5.4) and (5.5) are of the same form.

Now the lattice points are arranged in a rectangular array with m rows of n per row. Let these be numbered from 1 to n along the lower line; from $n + 1$ to $2n$ along the next, $\cdots$, and from $(m - 1)n + 1$ to mn along the top line. Without ambiguity we can now let u represent the vector whose elements are the values of the function u arranged in the order just described, and let w represent the vector whose elements are the values of the function w. Then the vectors u and w satisfy equations of the form

$$Au = h$$

and

$$Aw = k,$$

where only the right members differ. Three questions are of importance. The first is, does

$$w = A^{-1}k$$

approach zero in a suitable norm as n and m become infinite, and if so, how rapidly? This is the question of convergence and truncation errors. In general, what kind of bounds can be established for $\| w \|$ as a function of n and m? Second, given any approximation u^* to u, however obtained, what bounds are available for $\| u^* - u \|$ as a function of n and m? Evidently

$$A(u - u^*) = h - Au^* \equiv d,$$

$$u - u^* = A^{-1}d,$$

$$\| u - u^* \| \leqq \| A^{-1} \| \cdot \| d \| .$$

Also

$$\| w \| \leqq \| A^{-1} \| \cdot \| k \| .$$

Hence in both instances one needs to evaluate, at least approximately, the norm $\| A^{-1} \|$. In general $\| A^{-1} \|$ becomes infinite as n and m become infinite. It is important to know whether or not the norm grows so rapidly with n and m (e.g., exponentially) that carrying a reasonable number of extra digits in the computation would suffice to hold $\| u - u^* \|$ within bounds. This is the problem of numerical stability.

Finally, since these questions are independent of any method of solving the difference equations, this question remains. Since the systems are generally

large, one usually resorts to an iterative method of solving, and it is important to know that the selected method converges, and at what rate. In some schemes (the "marching" schemes), the problem does not arise, since the matrix A is already triangular in form and the value of u can be given explicitly at each point in terms of its values at previously computed points. In others, however, the implicit (or "jury") schemes, the matrix is not triangular, and this is always the case for elliptic equations. But for parabolic and hyperbolic equations the matrix is reducible, and one has to solve repeatedly a system of equations of order n (instead of nm) all having an identical submatrix.

In the methods to be discussed, the matrix A is expressible as an $m \times m$ array of submatrices, each of order n. Moreover these submatrices are themselves expressible as rational functions of the matrix K defined in section 1. Hence the submatrices are symmetric and commutative, their proper vectors are the same as those of K, and their proper values are rational functions of those of K. It is possible to give explicit expressions for the proper vectors and for the proper values, but only the latter will be used here. These expressions follow immediately from the following theorem:

THEOREM 5.2. *Define the polynomials* $\psi_\nu(\lambda, \rho)$ *by*

$$\psi_0 = 1$$

$$\psi_1 = \lambda,$$

$$\psi_\nu = \lambda\psi_{\nu-1} - \rho^2\psi_{\nu-2}, \qquad \nu \geqq 2.$$

If μ_1 *and* μ_2 *satisfy the quadratic*

$$\mu^2 - \lambda\mu + \rho^2 = 0,$$

then

$$\psi_\nu = \frac{\mu_1^{\nu+1} - \mu_2^{\nu+1}}{\mu_1 - \mu_2}, \qquad \lambda^2 \neq 4\rho^2,$$

$$\psi_\nu = (\nu + 1)\rho^\nu, \qquad \lambda^2 = 4\rho^2.$$

Moreover, if

$$\lambda = 2\rho \cos 2\theta,$$

then

$$\psi_\nu = \frac{\rho^\nu \sin 2(\nu + 1)\theta}{\sin 2\theta}.$$

and if

$$\lambda = 2\rho \cosh 2\omega,$$

then

$$\psi_\nu = \frac{\rho^\nu \sinh 2(\nu + 1)\omega}{\sinh 2\omega}.$$

The proof is made by induction. Now one verifies also by induction that

$$\psi_\nu(\lambda, 1) = \det(\lambda I - K_\nu).$$

Hence

THEOREM 5.3. *The proper values of* $K = K_n$ *are*

$$\lambda_\nu(K) = 2\cos 2\theta_\nu, \qquad 2\theta_\nu = 2\nu\theta_1 = \frac{\nu\pi}{n+1}.$$

Many of the matrices A to be considered are of the form described in the following theorem.

THEOREM 5.4. *Let* A *have the form*

$$A = \begin{bmatrix} P & 0 & 0 & \cdots \\ -Q & P & 0 & \cdots \\ 0 & -Q & P & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{bmatrix},$$

where $P = P(K)$ *and* $Q = Q(K)$ *are polynomials in* K. *Then if* $\kappa_\nu = \lambda_\nu(K)$ *is any proper value of* K, *then*

$$P(\kappa_\nu) = \lambda_\nu(P), \qquad Q(\kappa_\nu) = \lambda_\nu(Q)$$

are proper values of P *and* Q, *respectively. Let*

$$|\lambda_\nu(P)| \geqq \rho^{-1} > 0$$

and

$$|Q(\kappa_\nu)| \leqq |P(\kappa_\nu)|$$

for every ν. *Then*

$$\|A^{-1}\|_e = \|A^{-1}\|_{e'} \leqq \rho n^{\frac{1}{2}} m.$$

In fact, if $M = P^{-1}Q$, then

$$A^{-1} = \begin{bmatrix} P^{-1} & 0 & 0 & \cdots \\ MP^{-1} & P^{-1} & 0 & \cdots \\ M^2P^{-1} & MP^{-1} & P^{-1} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}.$$

Hence

$$\|A^{-1}\|_e = \|A^{-1}\|_{e'} \leqq \|P^{-1}\|_e + \|MP^{-1}\|_e + \cdots + \|M^{m-1}P^{-1}\|_e.$$

But $\nu' = 0, 1, \cdots, m-1$,

$$|\lambda_\nu(M^{\nu'}P^{-1})| = \left|\frac{\lambda_\nu^{\nu'}(Q)}{\lambda_\nu^{\nu'+1}(P)}\right| \leqq |\lambda_\nu^{-1}(P)| \leqq \rho.$$

Hence

$$\|M^{\nu'}P^{-1}\|_e \leqq \rho n^{\frac{1}{2}}$$

which proves the theorem.

THEOREM 5.5. *With hypotheses the same as in theorem* 5.4, *for large m the inequality*

$$\| A^{-1} \|_s^{-2} > [\, | P(\kappa) | - | Q(\kappa) | \,]^2 + \frac{| P(\kappa)Q(\kappa) | \pi^2}{4m^2}$$

holds, where $\kappa = \kappa_\nu$ *is that proper value of* K *that minimizes the right member.*

To prove this it is necessary to consider the matrix

$$\lambda I - A^T A = \begin{bmatrix} \lambda I - P^2 - Q^2 & PQ & 0 & \cdots & 0 \\ PQ & \lambda I - P^2 - Q^2 & PQ & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \lambda I - P^2 \end{bmatrix}$$

In this matrix every submatrix along the diagonal is of the form

$$\lambda I - P^2 - Q^2$$

except for the last one, which is $\lambda I - P^2$; every block above and every block below the diagonal is PQ; and every other block is null. Evidently all submatrices are commutative. Hence consider the following lemma:

LEMMA 5.5. *With obvious nonsingularity assumptions, a triple-diagonal matrix has the following factorization:*

$$\begin{bmatrix} P_1 & Q_1 & 0 & \cdots \\ R_1 & P_2 & Q_2 & \cdots \\ 0 & R_2 & P_3 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} = \begin{bmatrix} I & 0 & 0 & \cdots \\ \Gamma_1 & I & 0 & \cdots \\ 0 & \Gamma_2 & I & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \begin{bmatrix} B_1 & Q_1 & 0 & \cdots \\ 0 & B_2 & Q_2 & \cdots \\ 0 & 0 & B_3 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

where

$$\begin{aligned} B_1 &= P_1, & \Gamma_1 B_1 &= R_1, \\ \Gamma_1 Q_1 + B_2 &= P_2, & \Gamma_2 B_2 &= R_2, \\ \Gamma_2 Q_2 + B_3 &= P_3, & \Gamma_3 B_3 &= R_3, \\ &\cdots, & &\cdots \end{aligned}$$

Moreover, if the submatrices P_ν, Q_ν, R_ν *are all commutative, then one can define*

$$\Psi_0 = I,$$

$$\Psi_\nu = B_\nu \Psi_{\nu-1}, \qquad \nu = 1, 2, \cdots, m,$$

and the Ψ_ν *satisfy the recursion*

$$\Psi_0 = I$$

$$\Psi_1 = P_1,$$

$$\Psi_\nu = P_\nu \Psi_{\nu-1} - R_{\nu-1} Q_{\nu-1} \Psi_{\nu-2}, \qquad \nu = 2, 3, \cdots, m.$$

This is proved by induction. Evidently the determinant of the matrix on the left is equal to the determinant of Ψ_ν.

Now let the submatrices P_ν, Q_ν and R_ν be identified with those which occur in $\lambda I - A^TA$, and consider the polynomials $\psi_\nu(\lambda, P)$ of theorem 5.2. Then

$$\Psi_\nu = \psi_\nu(\lambda I - P^2 - Q^2, PQ), \qquad \nu < m,$$

$$\Psi_m = \psi_m(\lambda I - P^2 - Q^2, PQ) + Q^2\psi_{m-1}(\lambda I - P^2 - Q^2, PQ).$$

Since the matrices P and Q are commutative and diagonalized by the same orthogonal matrix V, the same is true of the matrices Ψ_ν, and $V^T\Psi_\nu V$ is diagonal. Moreover, each diagonal element is a polynomial in λ of degree ν. Hence every proper value of A^TA is a zero of a polynomial of the form

$$\psi = \psi_m(\lambda - \sigma^2 - \tau^2, \sigma\tau) + \tau^2\psi_{m-1}(\lambda - \sigma^2 - \tau^2, \sigma\tau),$$

where

$$\sigma = P(\kappa_\nu), \qquad \tau = Q(\kappa_\nu),$$

for some ν. Let

$$\lambda - \sigma^2 - \tau^2 = 2\sigma\tau\cos\theta, \tag{5.6}$$

$$\lambda = (\sigma - \tau)^2 + 4\sigma\tau\cos^2(\theta/2) \tag{5.7}$$

$$= (\sigma + \tau)^2 - 4\sigma\tau\sin^2(\theta/2).$$

Then by theorem 5.2,

$$\psi\sin\theta = \sigma^m\tau^m[\sin(m+1)\theta + (\tau/\sigma)\sin m\theta]. \tag{5.8}$$

To prove the theorem it will be shown that

$$(|\sigma| - |\tau|)^2 + 4|\sigma\tau|\frac{\sin^2\pi}{4(m+1)}$$

is a lower bound for the zeros of ψ. The result will then follow from the fact that $\|A\|_s^2 = \lambda$ is the smallest proper value. From the two expressions for λ in (5.7) it is clear that when σ and τ have like signs λ is a decreasing function of θ, and when they have opposite signs it is an increasing function. Suppose the signs are opposite. Since $|\tau/\sigma| < 1$, the quantity within brackets in (5.8) is positive when

$$0 < 2\theta \leqq \frac{\pi}{m+1}.$$

For large m,

$$\sin\frac{\pi}{4(m+1)} \doteq \frac{\pi}{4m},$$

which proves the theorem for this case. If the signs are alike, set $\theta = \pi - \theta'$ and the result again follows. The smallest proper value of A^TA is the smallest zero λ of all the ψ's formed as κ_ν ranges over all proper values of K.

In case $P = I$, one has a "marching" method and there is no need for an iterative solution of the equations. But if $P \neq I$, the equations can be solved

by repeatedly solving systems with the matrix P, and if n is large one may wish to resort to an iterative method of solving each time. Often P, or a scalar multiple of P, has the form

$$P = I - \mu K, \qquad \mu > 0. \tag{5.9}$$

For large n,

$$\rho(K) \doteq 2\left[1 - \frac{\pi^2}{2n^2}\right]$$

whence the simple iteration converges if and only if $\mu \leqq \frac{1}{2}$. If

$$C(\mu) = \mu(I - \mu J)^{-1}J^T,$$

then

$$\rho(C) = \rho^2(\mu K).$$

In fact, the principal minors of $[\lambda(I - \mu J) - \mu J^T]$ have a representation like that of $\psi_\nu(\lambda, \mu)$ but with $\lambda = 4\mu^2 \cos^2 \theta$.

THEOREM 5.6. *If P has the form* (5.9), *then the simple iteration converges, independently of the order of P, if and only if $\mu \leqq \frac{1}{2}$. In this event the Seidel type iteration converges twice as fast.*

Turn, now, to some special methods, and consider first the *parabolic equation*:

$$\partial^2 u/\partial x^2 = \partial u/\partial t.$$

The simplest method is to take

$$\gamma\delta_x^2 u = \Delta_y u, \qquad \gamma = \frac{\Delta y}{(\Delta x)^2}. \tag{5.10}$$

This leads to

$$P = I, \qquad Q = (1 - 2\gamma)I + \gamma K.$$

Hence

$$\lambda_\nu(Q) = 1 - 4\gamma \sin^2 \theta_\nu,$$

and

$$|\lambda_\nu(Q)/\lambda_\nu(P)| \leqq 1$$

for every ν if and only if $\gamma \leqq \frac{1}{2}$. In this event theorem 5.4 applies.

To apply theorem 5.5, note that

$$\lambda_1(P) - \lambda_1(Q) = 4\gamma \sin^2 \theta_1 \doteq \frac{\pi \Delta y}{[(n+1)\Delta x]^2}$$

and this certainly minimizes the first term on the right of the inequality. But $(n + 1)\Delta x$ remains fixed as n increases, while $\Delta y \sim m^{-1}$. Hence there exists a constant κ' independent of n and m such that

$$\| A^{-1} \|_S < \kappa' m.$$

If the region is bounded but not rectangular, it can be enclosed in a rectangular region, and by omitting certain rows and columns from the matrix associated with the rectangular region, one obtains that associated with the actual region. It is sufficient to omit only the off-diagonal elements. If A_t represents the matrix associated with the true region, then

$$I - A \geqq I - A_t \geqq 0.$$

Hence

$$\rho(I - A_t) < \rho(I - A) < 1,$$

and therefore

$$\| A_t^{-1} \|_e < \| A^{-1} \|_e, \qquad \| A_t^{-1} \|_S < \| A^{-1} \|_S.$$

Thus the appraisals for the rectangular region provide upper bounds for the norms associated with more general regions.

The method

$$\gamma \delta_x^2 u = \nabla_y u, \tag{5.11}$$

where $\nabla_y u$ represents the backward difference and γ is the same as before, is an implicit scheme leading to

$$P = (1 + 4\gamma)I - 2\gamma K, \qquad Q = I.$$

Here one has

$$\lambda_\nu(P) = 1 + 8\gamma \sin^2 \theta_\nu,$$

and the conditions of theorems 5.4 and 5.5 are satisfied independently of the value of γ. Since

$$\lambda_1(P) - \lambda_1(Q) = 8\gamma \sin^2 \theta_1,$$

therefore, for some κ',

$$\| A^{-1} \|_S < \kappa' m.$$

To apply theorem 5.6 for convergence, one has

$$\mu = \frac{2\gamma}{1 + 4\gamma} < \frac{1}{2}.$$

so that both types of iteration converge.

The Crank-Nicolson method makes use of

$$\gamma \delta_x^2 (1 + E_y) u = 2 \Delta_y u. \tag{5.12}$$

Hence

$$P = (1 + 2\gamma)I - \gamma K,$$

$$Q = (1 - 2\gamma)I + \gamma K,$$

$$\lambda_\nu(P) = 1 + 4\gamma \sin^2 \theta_\nu,$$

$$\lambda_\nu(Q) = 1 - 4\gamma \sin^2 \theta_\nu .$$

Hence the conditions of the two theorems are fulfilled for all γ. For this case

$$\lambda_1(P) - \lambda_1(Q) = 8\gamma \sin^2 \theta_1 ,$$

and again, for some κ',

$$\| A^{-1} \|_S < \kappa' m.$$

For theorem 5.6 one has

$$\mu = \frac{\gamma}{1 + 2\gamma} < \frac{1}{2}.$$

More generally, consider

$$\gamma \delta_x^2(\beta + \alpha E_y)u = \Delta_y u, \qquad 0 \leqq \alpha = 1 - \beta \leqq 1. \tag{5.13}$$

Then

$$\begin{aligned} P &= (1 + 2\gamma\alpha)I - \gamma\alpha K, \\ Q &= (1 - 2\gamma\beta)I - \gamma\beta K, \\ \lambda_\nu(P) &= 1 + 4\gamma\alpha \sin^2 \theta_\nu , \\ \lambda_\nu(Q) &= 1 - 4\gamma\beta \sin^2 \theta_\nu . \end{aligned}$$

The requirements of the theorems are satisfied if

$$2\gamma\alpha \geqq \gamma - \tfrac{1}{2}, \qquad 2\gamma\beta \leqq \gamma + \tfrac{1}{2}.$$

Then

$$\lambda_1(P) - \lambda_1(Q) = 4\gamma \sin^2 \theta_1 ,$$

and, for some κ',

$$\| A^{-1} \|_S < \kappa' m.$$

For the iteration one has

$$\mu = \frac{\gamma\alpha}{1 + 2\gamma\alpha} < \frac{1}{2}.$$

The matrices required by the hyperbolic equation are slightly more complicated. Some additional preparation is required.

THEOREM 5.7. *A matrix M expressible in the form*

$$M = \begin{bmatrix} I & 0 & 0 & 0 & \cdots \\ -B & I & 0 & 0 & \cdots \\ C^2 & -B & I & 0 & \cdots \\ 0 & C^2 & -B & I & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

in which B and C are commutative, has as its inverse

$$M^{-1} = \begin{bmatrix} I & 0 & 0 & 0 & \cdots \\ \Gamma_1 & I & 0 & 0 & \cdots \\ \Gamma_2 & \Gamma_1 & I & 0 & \cdots \\ \Gamma_3 & \Gamma_2 & \Gamma_1 & I & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

where

$$\Gamma_{\nu'} = \psi_{\nu'}(B, C), \qquad \nu' = 0, 1, \cdots, m - 1.$$

The polynomials ψ_ν are those defined in theorem 5.2, and the proof is made by direct verification.

COROLLARY 5.7. *If B and C are diagonalized by the same transformation, so, likewise, is each $\Gamma_{\nu'}$, and*

$$\lambda_\nu(\Gamma_{\nu'}) = \psi_{\nu'}(\beta_\nu, \gamma_\nu) = \frac{\gamma_\nu^{\nu'} \sin(\nu' + 1)\varphi_\nu}{\sin \varphi_\nu}$$

where

$$\beta_\nu = \lambda_\nu(B), \qquad \gamma_\nu = \lambda_\nu(C),$$
$$\beta_\nu = 2\gamma_\nu \cos \varphi_\nu .$$

The theorem and corollary permit estimates of the e-norms for the usual methods of solving the simple hyperbolic equation:

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial y^2}.$$

Evaluation of the spectral norm is more difficult and will be omitted.

The simplest method is represented by

$$\tau^2 \delta_x^2 u = \delta_y^2 u, \qquad \tau = \frac{\Delta y}{\Delta x}. \tag{5.14}$$

This leads to a matrix $A = M$ of the form given in theorem 5.7 with

$$B = 2(1 - \tau^2)I + \tau^2 K, \qquad C = I. \tag{5.15}$$

Hence

$$\beta_\nu = 2 - 4\tau^2 \sin^2 \theta_\nu , \qquad \gamma_\nu = 1. \tag{5.16}$$

From the corollary it follows that

$$|\lambda_\nu(\Gamma_{\nu'})| < \csc \varphi_\nu .$$

If $\tau \leqq 1$, then

$$\cos \varphi_\nu = 1 - 2\tau^2 \sin^2 \theta_\nu ,$$

and csc φ_ν assumes its greatest value with $\nu = 1$. Hence, neglecting terms of higher order,

$$1 - \frac{\varphi_1^2}{2} = 1 - \frac{2\tau^2\pi^2}{4(n+1)^2}, \qquad \varphi_1 = \frac{\tau\pi}{n+1} \doteq \frac{\tau\pi}{n}.$$

Hence

$$|\lambda_\nu(\Gamma_{\nu'})| < \frac{n}{\tau\pi}.$$

$$\|\Gamma_{\nu'}\|_e = \|\Gamma_{\nu'}\|_e' < \frac{n^{3/2}}{\tau\pi},$$

and therefore

$$\|A^{-1}\|_e = \|A^{-1}\|_e' < \frac{mn^{3/2}}{\tau\pi}. \tag{5.17}$$

If $\tau > 1$, the trigonometric functions must, in some cases, be replaced by hyperbolic functions and the proper values can become arbitrarily large.

Analogous to the final method considered for parabolic equations is the scheme

$$\tau^2(\alpha E_y + \beta E_y^{-1})\delta_x^2 u = \delta_y^2 u, \qquad 0 \leqq \alpha = 1 - \beta \leqq 1. \tag{5.18}$$

Let

$$\begin{aligned} P &= (1 + 2\alpha\tau^2)I - \alpha\tau^2 K, \\ Q &= (1 + 2\beta\tau^2)I - \beta\tau^2 K. \end{aligned} \tag{5.19}$$

Then the matrix which arises in this case is of the form

$$A = M \operatorname{diag}(P, P, \cdots, P) \tag{5.20}$$

where M is the matrix of theorem 5.7 with

$$B = 2P^{-1}, \qquad C^2 = QP^{-1}. \tag{5.21}$$

All submatrices are symmetric and commutative. The proper values of P and Q are

$$\lambda_\nu(P) = 1 + 4\alpha\tau^2 \sin^2\theta_\nu,$$

$$\lambda_\nu(Q) = 1 + 4\beta\tau^2 \sin^2\theta_\nu.$$

Hence all matrices are positive definite, and if $\beta \leqq \alpha$,

$$\gamma_\nu^2 = \lambda_\nu(C^2) \leqq 1.$$

Reference to corollary 5.7 therefore shows that this condition ensures that

$$|\lambda_\nu(\Gamma_{\nu'})| < \csc\varphi_\nu.$$

A little manipulation shows that

$$\cos^2\varphi_\nu = \lambda_\nu(P^{-1})\lambda_\nu(Q^{-1}),$$

and the largest value occurs with $\nu = 1$. Neglecting terms of higher order,

$$\varphi_1 = \frac{\tau\pi}{n},$$

whence

$$|\lambda_1(\Gamma_{\nu'})| < \frac{n}{\tau\pi}.$$

Moreover

$$|\lambda_\nu(P^{-1})| < 1,$$

whence, again,

$$\|A^{-1}\|_e = \|A^{-1}\|_e' < \frac{mn^{3/2}}{\tau\pi}. \tag{5.22}$$

Slightly more complicated is the scheme

$$\tau^2[\alpha E_y + (1 - 2\alpha) + \alpha E_y^{-1}]u = \delta_y^2 u, \qquad 2\alpha < 1. \tag{5.23}$$

For this method take

$$\begin{aligned} P &= (1 + 2\alpha\tau^2)I - \alpha\tau^2 K, \\ Q &= 2I - (1 - 2\alpha)\tau^2(2I - K). \end{aligned} \tag{5.24}$$

Then A has the form (5.20) with

$$B = QP^{-1}, \qquad C = I. \tag{5.25}$$

Then

$$\begin{aligned} \lambda_\nu(P) &= 1 + 4\alpha\tau^2 \sin^2\theta_\nu, \\ \lambda_\nu(Q) &= 2 - 4(1 - 2\alpha)\tau^2 \sin^2\theta_\nu, \end{aligned}$$

and

$$\beta_\nu = \lambda_\nu(B) = \frac{\lambda_\nu(Q)}{\lambda_\nu(P)} < 2.$$

Hence φ_ν is real and

$$|\lambda_\nu(\Gamma_{\nu'})| < \csc\theta_\nu.$$

The greatest value occurs at $\nu = 1$, and

$$\cos\varphi_1 = 1 - \frac{2(1 - 2\alpha)\tau^2 \sin^2\theta_1}{[1 + 4\alpha\tau^2 \sin^2\theta_1]}.$$

Neglecting terms of higher order,

$$\varphi_1^2 = 4(1 - 2\alpha)\tau^2\pi^2 n^{-2}, \qquad \|A^{-1}\|_e = \|A^{-1}\|_e' < \frac{mn^{3/2}}{2\tau\pi(1 - 2\alpha)^{\frac{1}{2}}}. \tag{5.26}$$

For the elliptic equation, $\partial^2 u/\partial x^2 + \partial^2 u/\partial y^2 = 0$, the simplest scheme is

$$\tau^2 \delta_x^2 u + \delta_y^2 u = 0, \qquad \tau = \frac{\Delta y}{\Delta x}. \tag{5.27}$$

The matrix has the form

$$A = \begin{bmatrix} B & -I & 0 & \cdots \\ -I & B & -I & \cdots \\ 0 & -I & B & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}, \qquad B = 2(1 + \tau^2)I - \tau^2 K. \tag{5.29}$$

The proper values of A are readily found by applying lemma 5.5 to the matrix $\lambda I - A$. In the recursion there defined,

$$P_\nu = \lambda I - B, \qquad R_\nu = Q_\nu = I.$$

Hence

$$\Psi_\nu = \psi_\nu(\lambda I - B, I).$$

Let

$$\beta_\nu = \lambda_\nu(B) = 2 + 4\tau^2 \sin^2 \theta_\nu .$$

Then the zeros of $\psi_m(\lambda - \beta_\nu, 1)$ are the proper values of A. If β is any β_ν, and

$$\lambda - \beta = 2 \cos 2\theta,$$

then ψ_m vanishes for

$$2\theta = 2\theta_{\nu'} = 2\nu'\theta_1 = \frac{\nu'\pi}{m + 1},$$

and hence for

$$\lambda = 4(\cos^2 \theta_{\nu'} + \tau^2 \sin^2 \theta_\nu).$$

This is least when $\nu' = m$, $\nu = 1$. Hence, neglecting terms of higher order, if $\lambda(A)$ is any proper value of A, then

$$\lambda(A) \geqq \pi^2(m^{-2} + \tau^2 n^{-2}),$$

hence, for some constant κ, if τ remains fixed,

$$\lambda(A) \geqq (\kappa m)^{-2}.$$

Hence

$$\| A^{-1} \|_S \leqq \kappa^2 m^2,$$

and

$$\| A^{-1} \|_\sigma = \| A^{-1} \|_\sigma' \leqq \kappa^2 m^{5/2} n^{\frac{1}{2}}.$$

The examination of problems in three or more independent variables can be handled in the same way although, as one may expect, the details are more

complicated. If we refer to the matrices treated above as two-story matrices, then in three dimensions one must treat three-story matrices. That is, if, in the z-direction there are p parallel planes in the lattice, then there arises a matrix A of order $n\,m\,p$, and this matrix breaks up into p^2 submatrices of a pattern similar to those appearing above, but each of these submatrices itself falls into m^2 submatrices, each of order n. The patterns are repeated, and the analysis is similar but requires more steps to carry through.

REFERENCES

Theorem 1.1 was stated by Mendelsohn (1955).

Norms are well known and much used in functional analysis, and certain special norms are common in numerical analysis, but the more general axiomatic approach adopted here was suggested by Faddeeva (1950), who defined vector and matrix norms independently and linked them by the notions of consistency and subordination. A norm is a convex function of the arguments, and as such is known to be continuous. For the general properties of convex bodies and functions see Bonnesen and Fenchel (1934) Dual norms are defined by von Neumann (1937) who, however, defines norms in a way somewhat different from ours. For a more general treatment of norms see Ostrowski (1955). Kolmogoroff (1934) associated norms and convex bodies in more general spaces.

A somewhat different definition of spectral norm is given by von Neumann and Goldstine (1947), who, however, do not use this term. See also Householder (1953). The use of g-norms was suggested by a paper of de la Garza (1953). Theorem 3.4 is given by Ostrowski.

For theorem 4.2 and its converse, see Oldenburger (1940). The case of nonnegative matrices has been investigated extensively by Frobenius (1908, 1909, 1912). See also a more recent paper by Wielandt (1950). Fiedler and Pták (1956) give theorems somewhat similar to theorem 4.12 and 4.13 but more limited. These and the related theorems are largely due to Stein and Rosenberg (1948). Weissinger (1952) gives theorem 4.14. For theorem 4.15, see Young (1954). It has been known for some time that when A is positive definite, then $\rho(C) < 1$, but the converse, as in theorem 4.18, is more recent and due to Reich (1949), whose proof is rather involved. The proof given here is based upon theorem 4.17, due to Stein (1952), and Lemma 4.18, due to Weissinger (1953). Actually Weissinger's lemma is slightly more general than lemma 4.18.

The proposal to examine the entire matrix in studying generated error in the solution of partial difference equations was made by Todd (1956). Theorems on proper values in section 5 are contained in or suggested by Rutherford (1945, 1951). Methods indicated by equations (5 13), (5.18) and (5.23) are studied by Mitchell (1956).

1. T. Bonnesen and W. Fenchel, *Theorie der konvexen Korper*. Berlin, Julius Springer, 1934.
2. V. N. Faddeeva, *Vyčislitel'nye metody lineĭnoĭ algebry*. Moscow, Gosudarstvennoe Izdatel'stvo, 1950.
3. Miroslav Fiedler and Vlastimil Ptak, Über die Konvergenz des verallgemeinerten Seidelschen Verfahrens fur Lösung von Systemen linearer Gleichungen. *Math. Nachr. 15* (1956), 31–38.
4. G. Frobenius, Über Matrizen aus positiven Elementen. *Berlin Sitz.* (1908) 471–76, (1909) 514–18.
5. ———, Über Matrizen aus nicht negativen Elementen. *Berlin Sitz.* (1912), 456–77
6. A. de la Garza, Error bounds on approximate solutions to systems of linear algebraic equations. *MTAC 7* (1953), 81–84.
7. Alston S. Householder, *Principles of numerical analysis*. New York, McGraw-Hill 1953.
8. ———, On the convergence of matrix iterations. Oak Ridge National Laboratory, ORNL-1883, 1955.

9. ———, On the convergence of matrix iterations. *J. Assoc. Comp. Mach.* 3 (1956), 314–24.
10. ———, Generated error in the solution of certain partial difference equations. Oak Ridge National Laboratory, ORNL-2230, 1956.
11. A. Kolmogoroff, Zur Normierbarkeit eines allgemeinen topologischen linearen Raumes. *Studia Math. 5* (1934), 29–33.
12. N. S. Mendelsohn, Some elementary properties of ill-conditioned matrices and linear equations. CARDE Technical Memorandum No. 120/55, 1955.
13. A. R. Mitchell, Round-off errors in implicit finite difference methods. *Qu. J. Mech. Appl. Math. 9* (1956), 111–21.
14. Richard von Mises and Hilda Pollaczek-Geiringer, Praktische Verfahren der Matrixauflösung, *Z. angew. Mat. Mech. 9* (1929), 58–77, 152–64.
15. John von Neumann, Some matrix inequalities and metrization of matrix-space. *Bull. Inst. Math. Mech. Univ. Tomsk 1* (1937), 286–99.
16. John von Neumann and Herman Goldstine, Numerical inverting of matrices of high order. *Bull. Am. Math. Soc. 53* (1947), 1021–99.
17. Rufus Oldenburger, Infinite powers of matrices and characteristic roots. *Duke Math. J. 6* (1940), 357–61.
18. Alexander Ostrowski, Über Normen von Matrizen, *Math. Z. 63* (1955), 2–18.
19. Edgar Reich, On the convergence of the classical iterative method of solving linear simultaneous equations. *Ann. Math. Stat. 20* (1949), 448–51.
20. D. E. Rutherford, Some continuant determinants arising in physics and chemistry. *Proc. Roy. Soc. Edinburgh A 62* (1945), 229–36, *63* (1951), 232–41.
21. P. Stein, Some general theorems on iterants. *J. Res. Natl. Bur. Stand. 48* (1952), 82–83.
22. ——— and R. L. Rosenberg, On the solution of linear simultaneous equations by iteration. *J. London Math. Soc. 23* (1948), 111–18.
23. John Todd, A direct approach to the problem of stability in the numerical solution of partial differential equations. *Comm. Pure and Appl. Math. 9* (1956), 597–612.
24. Johannes Weissinger, Über das Iterationsverfahren *Z. angew. Math. Mech. 31* (1951), 245–46.
25. ———, Zur Theorie und Anwendung des Iterationsverfahrens. *Math. Nachr. 8* (1952), 193–212.
26. ———, Verallgemeinerung des Seidelschen Iterationsverfahrens. *Z. angew. Math. Mech. 33* (1953), 155–63.
27. Helmut Wielandt, Unzerlegbare, nichtnegative Matrizen. *Math. Z. 52* (1950), 642–48.
28. David Young, Iterative methods for solving partial difference equations of elliptic type. *Trans. Amer. Math. Soc. 76* (1954), 92–111.