MATHEMATICS OF COMPUTATION, VOLUME 25, NUMBER 116, OCTOBER, 1971

On the Convergence Rates of Variational Methods. I. Asymptotically Diagonal Systems*

By L. M. Delves and K. O. Mead

Abstract. We consider the problem of estimating the convergence rate of a variational solution to an inhomogeneous equation. This problem is not soluble in general without imposing conditions on both the class of expansion functions and the class of problems considered; we introduce the concept of "asymptotically diagonal systems," which is particularly appropriate for classical variational expansions as applied to elliptic partial differential equations.

For such systems, we obtain a number of a priori estimates of the asymptotic convergence rate which are easy to compute, and which are likely to be realistic in practice. In the simplest cases these estimates reduce the problem of variational convergence to the simpler problem of Fourier series convergence, which is considered in a companion paper. We also produce estimates for the convergence rate of the individual expansion coefficients $a_i^{(n)}$, thus categorising the convergence completely.

I. Introduction. Variational methods for the solution of elliptic partial differential equations have a long and fruitful history in mathematical physics [1], [2]. More recently, the finite element method, as one particular variational procedure, has been widely used in engineering problems and its convergence properties have been studied.

There are, perhaps, two dominant features of any proposed algorithm for this type of problem:

(1) The rate of convergence of the numerical solution (to the exact solution).

(2) The ease of use of the method; that is, the cost of setting up the equations, and, in particular, the difficulty of treating awkward boundaries or boundary conditions.

The finite element method is notably favourable with respect to (2), and at least acceptable with respect to (1), when compared with alternative finite-difference formalisms.

It is well known that what may be called "classical" variational methods in which, typically, the expansion is made in terms of a set of orthogonal functions, such as, e.g., $\{\mathrm{Sin}\ nx\}$, can lead, in favourable cases, to extremely rapid convergence. However, no general analysis of the convergence rate problem for such classical expansions appears to have been made. In this paper, we present such an analysis. Although we make no specific reference to the expansion set used, the conditions placed on the systems analysed are motivated by those typifying the classical expansions. Our aim is to understand and to be able to predict the convergence rates to be expected, with

Received May 15, 1970, revised March 12, 1971.

AMS 1970 *subject classifications.* Primary 47H10, 15A57, 41A25; Secondary 65F99, 65J05, 65N30.

Key words and phrases. Variational methods, convergence rates, asymptotically diagonal systems.

* This work was supported in part by Science Research Council grants, to K. O. Mead, and to the University of Liverpool, for work on the Nuclear Three Body Problem.

an eye to condition (1) above; we do not take cognisance of condition (2). Such an analysis may be expected to be of value in two ways. When actually computing a variational solution to a system, a knowledge of the asymptotic convergence rate enables one to predict the number of expansion functions (the value of N) necessary to obtain an approximate solution of the required accuracy. The second, and more important, application arises in the initial choice of the expansion set. "A priori" convergence rate estimates facilitate selection of a realistic expansion set before commencing the computation and so help to eliminate the present trial-and-error evaluation of such sets.

The case for such an analysis was originally made by Schwartz [3] and illustrated by the examination of a particular system (having physical significance) and a limited class of expansion sets. In the present paper, we seek to provide the basis for a generalisation of this approach.

The procedure for finding numerical solutions to a differential (or other) equation by a variational method has two stages. First, we look for a functional defined over a Hilbert space containing the exact solution and possessing the property that it is stationary at such a solution. Throughout the present paper, we shall use as an example the inhomogeneous equation

$$\mathfrak{L}f = g \tag{1}$$

defined over a space R, and require that the operator $\mathfrak{L}$ is Hermitian with respect to the inner product used. For this equation, we can easily verify by differentiation that the functional

$$F[\omega] = (\omega, \mathfrak{L}\omega) - (\omega, g) - (g, \omega) \tag{2}$$

fulfills our requirements.

Secondly, we choose a complete, and perhaps orthogonal, set of functions $\{h_i\}$ in the Hilbert space, in terms of which the solution F may be expanded. We write

$$f = \sum_{i=1}^{\infty} b_i h_i \tag{3}$$

where the b_i are generalised Fourier coefficients of f. As an approximation to this solution, we take the finite sum

$$f_N = \sum_{i=1}^{N} a_i^{(N)} h_i. \tag{4}$$

Substitution of (4) into (2) leads to the well-known equations

$$L^{(N)}\mathbf{a}^{(N)} = \mathbf{g}^{(N)} \tag{4a}$$

where

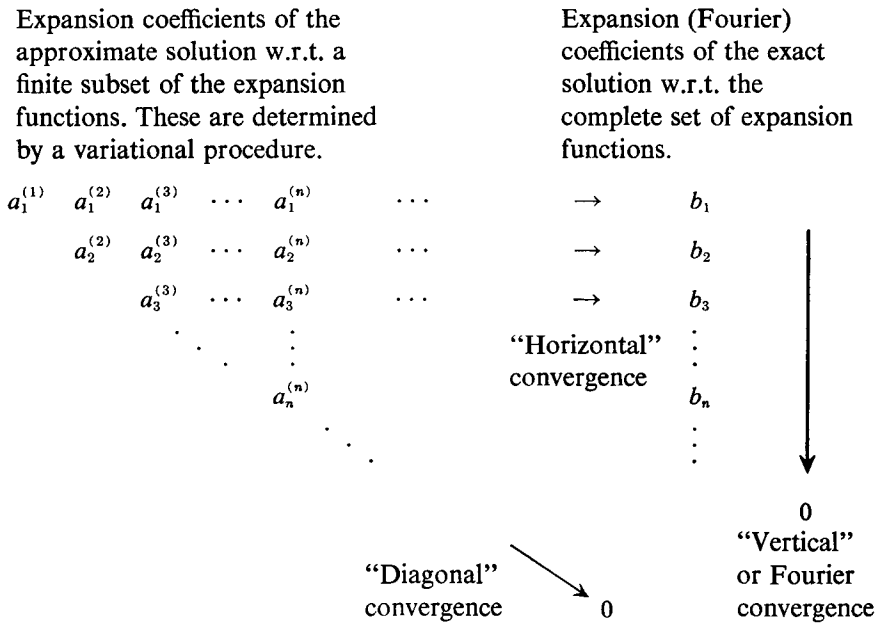
$$\begin{aligned} L_{i,j}^{(N)} &= (h_i, \mathfrak{L}h_j), \\ g_i^{(N)} &= (h_i, g), \end{aligned} \qquad i, j = 1, \cdots, N.$$

Where no ambiguity is likely to arise, we shall omit the superfix N.

The conditions under which this procedure converges to a solution of (1) have been considered by other authors, for example [1], [2]. We assume that convergence is assured and seek to estimate the *rate* of convergence for a given choice of an expansion set $\{h_i\}$.

Error Analysis. Throughout this section, reference is made to Diagram 1, in which the variational coefficients $a_i^{(N)}$ are displayed in a triangular array, together with the Fourier coefficients b_i of the exact solution.

DIAGRAM 1



The error in our Nth solution is given by

$$e_N = f - f_N = \sum_{i=1}^{N} (b_i - a_i^{(N)})h_i + \sum_{i=N+1}^{\infty} b_i h_i.$$

If we assume the h_i to be orthonormal with respect to some weight g, we have

$$||e_N||_g^2 = \sum_{i=1}^{N} (b_i - a_i^{(N)})^2 + \sum_{i=N+1}^{\infty} b_i^2 \tag{5}$$

$$\equiv S_1^{(N)} + S_2^{(N)}.$$

Since both $S_1^{(N)}$ and $S_2^{(N)}$ are defined as sums of squares and therefore positive quantities, convergence in this norm occurs if and only if both sums converge to zero with increasing N. Clearly, the convergence of $S_2^{(N)}$ depends only on the rate at which the Fourier coefficients of f tend to zero; this convergence is denoted by a vertical arrow in the diagram. Rapid convergence in this direction is seen to be a necessary condition for rapid convergence of the variational procedure and its consideration is therefore likely to be a major factor in the choice of a suitable expansion set $\{h_i\}$. Determining the convergence rate of $S_2^{(N)}$ is a problem lying purely within approximation theory and it may be resolved for a large class of systems on the basis of certain qualitative information concerning the solution f. We consider this question in detail for one-dimensional systems in a second paper. The problem of convergence of $S_1^{(N)}$ may be considered as the compound of two different types of convergence problems. First, we require that the variational coefficients $a_i^{(N)}$ tend to the corresponding

Fourier coefficients b_i with increasing N, i.e.

$$|b_i - a_i^{(N)}| \to 0 \quad \text{as } N \to \infty .$$

Secondly, we want

$$|b_i - a_i^{(i)}| \to 0 \quad \text{as } i \to \infty .$$

Since the series $\{b_i\} \to 0$, this will occur if $\{a_i^{(i)}\} \to 0$. We refer to Diagram 1, where these two convergence problems are, respectively, termed "horizontal" and "diagonal" convergence.

By considering bounding rates of convergence for these problems, it can easily be shown that the resulting convergence rate of $||e_N||_g$ is approximately that of the slowest of these separate convergence rates. For example, we take the case:

$$|b_i| \leqq k/i^r, \tag{6a}$$

$$|b_i - a_i^{(N)}| \leqq K/i^{(p-q)}N^q, \tag{6b}$$

i.e. the vertical convergence rate is $O(i^{-r})$ from (6a); putting $N = i$ in (6b), the diagonal convergence rate is $O(i^{-p})$, and with i constant the horizontal rate is $O(N^{-q})$. Then, by bounding the sums $S_1^{(N)}$ and $S_2^{(N)}$, it can easily be shown that

$$||e_N|| = O(N^{-s}) \quad \text{or} \quad O(N^{-s+1/2}) \quad \text{where } s = \min(p, q, r).$$

Similar results hold when one or more of the series converge exponentially. All three convergence problems must therefore be investigated to determine the net convergence rate of the solution.

Under certain conditions, however, it is possible to prove that variational convergence (horizontal and diagonal) is rapid, and hence that it is the Fourier convergence rate which characterises the overall convergence. As an extreme example, we consider choosing an expansion set which is orthogonalised with respect to the weight $\mathfrak{L}$, i.e., $(h_i, \mathfrak{L}h_j) = \delta_{ij}$.

(For a positive definite operator $\mathfrak{L}$, we can always orthogonalise an arbitrary expansion set in this way using a Gram-Schmidt process.) For such a choice, it is well known that the variational coefficients computed for the system (1) will be identically equal to the corresponding Fourier coefficients; this result is computationally obvious since the operator matrix L is a unit matrix. Thus, at the Nth stage of approximation, we obtain

$$a_i^{(N)} = b_i, \qquad i = 1, \cdots, N,$$

and an error

$$||e_N||_{\mathfrak{L}}^2 = \sum_{i=N+1}^{\infty} b_i^2.$$

The rate of decrease of the error term depends solely on the convergence rate of the Fourier series.

In principle, one may, therefore, always estimate the convergence rate, in the energy ($\mathfrak{L}$) norm, by estimating the coefficients b_i with respect to this set. In practice, this merely begs the question, first, because the appropriate orthogonalised expansion set is not known explicitly, and second, because one would often like the convergence rate in a fixed norm other than the energy norm. For any fixed choice of expansion set,

the operator matrix will not, in general, be diagonal; the convergence properties of the problem in a given norm then reside in the structure of the matrix $L^{(N)}$, and in the right-hand side vector $g^{(N)}$; both of these are reflected of course in the properties of the exact solution f, that is in the coefficients b_i. In many practical examples the matrix the matrix L, although not diagonal, does have a particularly simple structure.

Definition. A matrix L is said to be *asymptotically diagonal* (A.D.) of degree p if, for fixed j and all i,

$$\frac{|L_{i,j}|}{\{|L_{i,i}|\ |L_{j,j}|\}^{1/2}} \leqq C_j i^{-p}, \qquad C_j,\ p > 0.$$

It is *uniformly asymptotically diagonal* (U.A.D.) of degree p if, in addition, for some finite C,

$$C_j \leqq C, \qquad \forall j.$$

It should be noted that these properties are invariant under a diagonal transformation of L, since, if $L' = DLD^T$ and D is the diagonal matrix (d_i), we have

$$\frac{|L'_{ij}|}{(|L'_{ii}|\ |L'_{jj}|)^{1/2}} = \frac{|d_i L_{ij} d_j|}{(|d_i L_{ii} d_i|\ |d_j L_{jj} d_j|)^{1/2}} = \frac{|L_{ij}|}{(|L_{ii}|\ |L_{jj}|)^{1/2}}.$$

It follows that the properties are invariant under a renormalisation of the expansion set $\{h_i\}$.

As an example of an asymptotically diagonal system, we consider the solution of the elliptic equation over a closed region C:

$$Lf \equiv [\nabla^2 + V]f = g, \qquad f(C) = 0,$$

with expansion sets $\{h_n\}$ defined as the (orthonormal) solutions of

$$[\nabla^2 + W - \lambda_n]h_n = 0, \qquad h_n(C) = 0,$$

where W, V are point operators. In this case, we find immediately

$$L_{ij} = \lambda_i \delta_{ij} + (h_i(V - W)h_j).$$

Thus, for fixed $j \neq i$, the behavior of the matrix element L_{ij} is given by the Fourier coefficients with respect to the set $\{h_i\}$ of the fixed function $(V - W)h_i$. The considerations of paper II then lead to the result that the matrix L_{ij} is at least A.D. for a wide class of functions V, W. Similar results follow for a very wide class of problems of practical interest. More generally, we know that for any positive definite operator $\mathfrak{L}$ there exists at least one expansion set for which the matrix L is suitably asymptotically diagonal (namely, the set defined above which makes L the unit matrix).

The class of asymptotically diagonally systems is in fact very wide:

THEOREM 0. *Given that the operator $\mathfrak{L}$ maps $R \to R$, and the set $\{h_i\}$ is orthonormal in R, then the matrix L is A.D. of degree at least $\frac{1}{2}$, provided that $|L_{ii}|$ is bounded below.*

Proof. We have $L_{ij} = (h_i, \mathfrak{L}h_j) = (h_i, f_j)$ where, for fixed j, $f_j = \mathfrak{L}h_j$ is a fixed element of R. Hence, L_{ij} is the ith Fourier coefficient b_i of f_j and the series $\sum |b_i^2|$, therefore, converges. Whence the theorem follows, provided that, for some γ, $|L_{ii}| > \gamma$. Q.E.D.

The last restriction on L_{ii} is satisfied quite generally for partial differential equations.

In the following sections, we develop the theory of U.A.D. systems and obtain estimates of the convergence rate of $S_1^{(N)}$. These estimates yield sufficient conditions under which the convergence of the variational procedure may be characterised by the more easily determined Fourier series convergence. In particular, we show that, for A.D. systems of sufficiently high degree, the convergence of $||e_N||_2$ is directly related to the convergence of the coefficients g_i and also, that for such systems, the sum $S_1^{(N)}$ is negligible and $S_2^{(N)}$ dominates in the error expansion. We further derive relations between the convergence of g_i, b_i, and the structure of L_N, and characterise the convergence of the variational coefficients $a_i^{(N)}$ with respect to both i and N.

II. Theorems for U.A.D. Systems. In this section, we prove a number of theorems related to the solution of (4a). No reference will be made to the variational origin of these equations; the theorems are therefore valid for any numerical method (such as the method of moments) leading to such a set of equations with L symmetric and U.A.D.

First, we prove that if L belongs to a particular subset of U.A.D. matrices, and if there exists a lower triangular matrix T which diagonalises L, then T is itself U.A.D. and of the same degree. We note that, for positive definite matrices L, a suitable diagonalising T will always exist. (See [4], or consider the Cholesky decomposition of L^{-1}.)

In this paper, we shall assume that the operators under discussion are Hermitian, but not necessarily positive definite (see Theorem 5 for an exception to this). We do not, therefore, assume that the diagonal elements of the matrix L are positive; for a given expansion set, we normalise for convenience so that $L_{ii} = \pm 1$, and introduce the diagonal matrix J: $J_i \equiv J_{ii} = \text{sign}\,(L_{ii})$. This matrix relates to the triangular decomposition of L^{-1} in the form $L^{-1} = T^T J T$.

For nonorthogonal systems, the operator matrix L may not be asymptotically diagonal. However, the variational solution f_N is invariant under a nonsingular linear transformation of the set $\{h_i,\ i = 1, \cdots, N\}$; we may therefore transform to an orthogonal set before estimating the error.

THEOREM 1. *Let L be an $N \times N$ symmetric matrix having*

$$L_{ii} = J_{ii} = \pm 1, \text{ for } i = 1, \cdots, N,$$

and satisfying

$$|L_{ij}| \leqq Ci^{-p}, \quad \text{for all } i \neq j, \text{ with } p > 1 \text{ and } 0 < C \leqq C(p) \text{ [defined in the proof]}.$$

Also let T be a lower triangular $N \times N$ matrix such that

$$TLT^T = J.$$

Then if we write $T = I + X$, we have that X is also lower triangular and satisfies

$$|X_{ij}| \leqq Ki^{-p}, \qquad 0 < K < K(p, C).$$

Proof. Let us write $L = J + U + U^T$, where U is a lower triangular matrix (having zero diagonal). Then, the transformed matrix

$$\bar{L} = TLT^T = TJT^T + T(U + U^T)T^T = J.$$

Thus, $TJT^T = J - T(U + U^T)T^T$. Writing $T = I + X$,

$$(I + X)J(I + X^T) = J - (I + X)(U + U^T)(I + X^T),$$

i.e.

$$XJ + JX^T = -(U + U^T) - X(U + U^T) - (U + U^T)X^T - XLX^T. \tag{7}$$

We may write this as $Z = F(Z)$, where

$$\begin{aligned} Z_{ij} &= X_{ij}J_j \quad \text{for } i \geqq j, \\ &= X_{ji}J_i \quad \text{otherwise.} \end{aligned}$$

Since both Z and Eq. (7) are symmetric, we can look on F as a function in a $\frac{1}{2}N(N+1)$-dimensional vector space, with components $Z_{ij} = X_{ij}J$, $i \geqq j$. We seek a region in which the solution of (7) lies, i.e. we require a region R such that F: $R \to R$ and over which F has a Lipschitz constant which is less than unity. Consider the region R: $|Z_{ij}| \leqq Ki^{-q}$ where K, q are constants. From the conditions on L, we have $|U_{ij}| \leqq Ci^{-p}$. Also $|Z_{ij}| \leqq Ki^{-q}$ implies $|X_{ij}| \leqq Ki^{-q}$. Substituting into (7) for the case $i > j$:

$$\begin{aligned} |F_{ij}| &\leqq |U_{ij}| + \sum_{k=1}^{j-1} |X_{ij}|\,|U_{jk}| + \sum_{k=j+1}^{i} |X_{ik}|\,|U_{kj}| + \sum_{k=1}^{j} |U_{ik}|\,|X_{jk}| \\ &\quad + \sum_{k=1}^{i}\sum_{m=1}^{j} |X_{ik}|\,|L_{km}|\,|X_{jm}| \\ &\leqq Ci^{-p} + KCi^{-q}j^{-p+1} + \frac{KC}{(p-1)}\,i^{-q}j^{-p+1} + KCi^{-p}j^{-q+1} \\ &\quad + K^2i^{-q}j^{-q}\sum_{k=1}^{i}\sum_{m=1}^{j}|L_{km}|. \end{aligned}$$

Each term is maximised with $j = 1$, hence

$$|F_{ij}| \leqq \left[Ci^{-p+q} + KC\left(\frac{p}{p-1}\right) + KCi^{-p+q} + K^2 + \frac{K^2C}{(p-1)}\right]i^{-q}.$$

Similarly, for the case $i = j$:

$$|F_{ii}| \leqq \frac{1}{2}\left[2KCi^{-p+1} + K^2 + \frac{K^2C}{(p-1)}\,i^{-q+1}\right]i^{-q}.$$

A sufficient condition for F: $R \to R$ is that

$$|F_{ij}| \leqq Ki^{-q}, \quad \text{for all } i, j.$$

This holds if

$$\left[Ci^{-p+q} + KC\left(\frac{p}{p-1}\right) + KCi^{-p+q} + K^2 + \frac{K^2C}{(p-1)}\right] \leqq K$$

and

$$\frac{1}{2}\left[2KCi^{-p+1} + K^2 + \frac{K^2C}{(p-1)}\,i^{-q+1}\right] \leqq K \quad \text{for all } i.$$

These clearly imply

(a) $$1 \leqq q \leqq p \quad \text{and} \quad p \neq 1.$$

It follows that the inequalities will hold for all i if they hold for $i = 1$. In addition, the second inequality will hold if the first one does. Hence

(b) $$K^2\left[1 + \frac{C}{(p-1)}\right] + K\left[\left(\frac{2p-1}{p-1}\right)C - 1\right] + C \leqq 0.$$

(a) and (b) are sufficient conditions for F: $R \to R$. A Lipschitz constant for F will be given by

$$M = \sup_R \left\| \left| \frac{\partial F_\nu}{\partial Z_\mu} \right| \right\|_\infty = \sup_R \max_{i,j;i\geqq j} \sum_{l,n;l\geqq n} \left| \frac{\partial F_{ij}}{\partial Z_{ln}} \right|.$$

For the case $i > j$, (7) gives

$$F_{ij} = -U_{ij} - \sum_{k=1}^{j-1} X_{ik}U_{jk} - \sum_{k=j+1}^{i} X_{ik}U_{kj} - \sum_{k=1}^{j} U_{ik}X_{jk} - \sum_{k=1}^{i}\sum_{m=1}^{j} X_{ik}L_{km}X_{jm}.$$

Thus,

$$\begin{aligned}
\frac{\partial F_{ij}}{\partial X_{ln}} &= -\sum_{m=1}^{j} L_{jm}X_{jm} && \text{for } l = i,\ n = j,\\
&= -U_{jn} - \sum_{m=1}^{j} L_{nm}X_{jm} && \text{for } l = i,\ n < j,\\
&= -U_{nj} - \sum_{m=1}^{j} L_{nm}X_{jm} && \text{for } l = i,\ n > j,\\
&= -U_{in} - \sum_{k=1}^{i} X_{ik}L_{kn} && \text{for } l = j,\ n \leqq j,\\
&= 0 && \text{otherwise.}
\end{aligned}$$

Summing these terms,

$$\begin{aligned}
\sum_{l,n}\left|\frac{\partial F_{ij}}{\partial Z_{ln}}\right| &\leqq \left|\sum_{m=1}^{j} L_{jm}X_{jm}\right| + \sum_{n=1}^{j-1}\left|U_{jn} + \sum_{m=1}^{j} L_{nm}X_{jm}\right| + \sum_{n=j+1}^{i}\left|U_{nj} + \sum_{m=1}^{j} L_{nm}X_{jm}\right|\\
&\quad + \sum_{n=1}^{j}\left|U_{in} + \sum_{k=1}^{i} X_{ik}L_{kn}\right|\\
&\leqq Kj^{-q} + KCj^{-p-q+1} + \sum_{n=1}^{j-1}\left[Cj^{-p} + Kj^{-q} + KC\left(\frac{p}{p-1}\right)j^{-q}n^{-p+1}\right]\\
&\quad + \sum_{n=j+1}^{i}\left[Cn^{-p} + KCj^{-q+1}n^{-p}\right]\\
&\quad + \sum_{n=1}^{j}\left[Ci^{-p} + Ki^{-q} + KC\left(\frac{p}{p-1}\right)i^{-q}n^{-p+1}\right].
\end{aligned}$$

Clearly, the Lipschitz constant which results may be minimised by choosing q as large as possible. The maximum value of q consistent with inequality (a) is $q = p$, and this choice leaves (b) unaltered. Also

$$\sum_{n=1}^{j-1} n^{-p+1} \leqq \sum_{n=1}^{j} n^{-p+1} \leqq \sum_{n=1}^{j} 1 = j \leqq i.$$

Thus:

$$\sum_{l,n} \left|\frac{\partial F_{ij}}{\partial Z_{ln}}\right| \leqq Kj^{-p} + KCj^{-2p+1} + Cj^{-p+1} + Kj^{-p+1} + KC\left(\frac{p}{p-1}\right)j^{-p+1} + (C + KC)\frac{j^{-p+1}}{(p-1)} + Ci^{-p+1} + Ki^{-p+1} + KC\left(\frac{p}{p-1}\right)i^{-p+1}.$$

This is maximised for $i = 1, j = 1$, and

$$\max_{i,j;i>j} \sum_{l,n} \left|\frac{\partial F_{ij}}{\partial Z_{ln}}\right| \leqq 3K + \frac{(2p-1)C}{(p-1)} + 3\left(\frac{p}{p-1}\right)KC. \tag{8}$$

When $i = j$,

$$F_{ii} = -\sum_{k=1}^{i-1} X_{ik}U_{ik} - \frac{1}{2}\sum_{k=1}^{i}\sum_{m=1}^{i} X_{ik}L_{km}X_{im},$$

from which we similarly obtain the bound

$$\max_{i=j} \sum_{l,n} \left|\frac{\partial F_{ij}}{\partial Z_{ln}}\right| \leqq 2K + C + \left(\frac{2p-1}{p-1}\right)KC$$

which is clearly smaller than (8). Hence,

$$M = \sup_{R} \max_{i,j;i\geq j} \sum_{l,n;l\geq n} \left|\frac{\partial F_{ij}}{\partial Z_{ln}}\right| \leqq 3K + \left(\frac{2p-1}{p-1}\right)C + 3\left(\frac{p}{p-1}\right)KC.$$

So a sufficient condition for $M < 1$ is

$$3K + \left(\frac{2p-1}{p-1}\right)C + 3\left(\frac{p}{p-1}\right)KC < 1$$

which implies that

$$3K^2\left[1 + \left(\frac{p}{p-1}\right)C\right] + K\left[\left(\frac{2p-1}{p-1}\right)C - 1\right] < 0.$$

Comparing this with inequality (b), both will be satisfied if

$$3K^2\left[1 + \left(\frac{p}{p-1}\right)C\right] + K\left[\left(\frac{2p-1}{p-1}\right)C - 1\right] + C \leqq 0,$$

which in turn will hold if K lies between the zeroes of the L.H.S., provided that these zeroes are real, and that at least one of them is positive.

Reality of the zeroes requires that

$$\left[\left(\frac{2p-1}{p-1}\right)C - 1\right]^2 - 12C\left[1 + \left(\frac{p}{p-1}\right)C\right] \geqq 0.$$

This will be true if $C \leqq C(p)$, where

$$C(p) = \frac{(p-1)}{(8p^2 - 8p - 1)}\left[((8p-7)^2 + (8p^2 - 8p - 1))^{1/2} - (8p-7)\right]$$

and this is a condition of the theorem.

For at least one (and in fact both) of the zeroes to be positive we must have

$$1 - \left(\frac{2p-1}{p-1}\right)C > 0,$$

whence

$$C < \left(\frac{p-1}{2p-1}\right).$$

This inequality is also always satisfied as a result of the condition $C < C(p)$. We may take $K(p, C)$ to be the larger of the two zeroes. Q.E.D.

In the next few theorems, we use Theorem I to characterise the error norm $||e_N||_{\mathfrak{L}}$.

To simplify the statement of these theorems, we define the following class of systems:

Definition. Let $\mathfrak{L}f = g$ be an inhomogeneous equation, and let $L\mathbf{b} = \mathbf{g}$ be the corresponding infinite linear system, with $\{h_i\}$ as expansion set. We shall call this a "nice" system of degree p if for the given choice of $\{h_i\}$, every submatrix $L^{(N)}$ satisfies the conditions of Theorem 1.

We note that the requirement $L_{i,i} = \pm 1$ can be obtained by suitable normalisation of the functions h_i, and is not a real restriction since asymptotic diagonality is invariant under a diagonal transformation.

THEOREM 2. *For a "nice" system, the error in the Nth approximate solution, defined by*

$$e_N = f_N - f = \sum_{i=1}^{N} a_i^{(N)} h_i - f$$

satisfies the bounding inequality

$$||e_N||_{\mathfrak{L}} \leqq k \sum_{i=N+1}^{\infty} |a_i^{(i)}|, \quad \textit{where } 0 < k \leqq k(p, C).$$

Proof. Let us orthogonalise the first N expansion functions with respect to the operator $\mathfrak{L}$, using a Gram-Schmidt process. That is, we define

$$\bar{h}_i = \sum_{j=1}^{i} T_{ij} h_j, \qquad i = 1, \cdots, N,$$

where T is a lower triangular matrix such that

$$(\bar{h}_i, \mathfrak{L}\bar{h}_j) = \delta_{ij} J_j.$$

The transformed operator matrix is

$$\bar{L}^{(N)} = TL^{(N)}T^T = J^{(N)}.$$

The Nth approximate solution is invariant under this orthogonalisation and may be written

$$f_N = \sum_{i=1}^{N} \alpha_i^{(N)} \bar{h}_i = \boldsymbol{\alpha}^{(N)} \cdot \bar{h} \quad \text{(in an obvious notation)},$$

where $\boldsymbol{\alpha}^{(N)}$ is the (trivial) solution of

$$J^{(N)}\boldsymbol{\alpha}^{(N)} = \boldsymbol{\beta}^{(N)} \quad \text{and} \quad \beta_i^{(N)} = (\bar{h}_i, g).$$

Let us now add h_{N+1} to the expansion set. The operator matrix becomes

$$L^{(N+1)} = \begin{bmatrix} T & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} L^{(N)} & \gamma^{(N)} \\ \gamma^{(N)T} & J_{N+1} \end{bmatrix} \begin{bmatrix} T^T & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} J^{(N)} & T\gamma^{(N)} \\ \gamma^{(N)T}T^T & J_{N+1} \end{bmatrix},$$

where $\gamma_i^{(N)} = (h_i, \mathfrak{L}h_{N+1})$ and hence the system

$$\begin{bmatrix} J^{(N)} & T\gamma^{(N)} \\ \gamma^{(N)T}T^T & J_{N+1} \end{bmatrix} \begin{bmatrix} \alpha^{(N+1)} \\ a_{N+1}^{(N+1)} \end{bmatrix} = \begin{bmatrix} \beta^{(N)} \\ g_{N+1} \end{bmatrix}.$$

This implies

$$J^{(N)}\alpha^{(N+1)} + a_{N+1}^{(N+1)}T\gamma^{(N)} = \beta^{(N)} \tag{9}$$

and

$$\gamma^{(N)T}T^T\alpha^{(N+1)} + J_{N+1}a_{N+1}^{(N+1)} = g_{N+1}. \tag{10}$$

The $(N+1)$th approximate solution may be written:

$$f_{N+1} = \sum_{i=1}^{N} \alpha_i^{(N+1)}\bar{h}_i + a_{N+1}^{(N+1)}h_{N+1}.$$

Hence

$$\begin{aligned} e_{N+1} - e_N &= f_{N+1} - f_N \\ &= \alpha^{(N+1)}\cdot\bar{\mathfrak{h}} + a_{N+1}^{(N+1)}h_{N+1} - \alpha^{(N)}\cdot\bar{\mathfrak{h}} \\ &= J^{(N)}\beta^{(N)}\cdot\bar{\mathfrak{h}} - a_{N+1}^{(N+1)}J^{(N)}(T\gamma^{(N)})\cdot\bar{\mathfrak{h}} + a_{N+1}^{(N+1)}h_{N+1} - J^{(N)}\beta^{(N)}\cdot\bar{\mathfrak{h}}. \end{aligned}$$

Using (9) and $J^2 = I$, this becomes

$$= a_{N+1}^{(N+1)}[h_{N+1} - J^{(N)}T\gamma^{(N)}\cdot\bar{\mathfrak{h}}].$$

Therefore,

$$||e_N - e_{N+1}|| \leqq |a_{N+1}^{(N+1)}| \left\{||h_{N+1}|| + \sum_{i=1}^{N} |(T\gamma^{(N)})_i|\,||\bar{h}_i||\right\}.$$

Using the $\mathfrak{L}$ norm, $||h_{N+1}|| = ||\bar{h}_i|| = 1$, and hence

$$||e_N||_{\mathfrak{L}} - ||e_{N+1}||_{\mathfrak{L}} \leqq ||e_N - e_{N+1}||_{\mathfrak{L}} \leqq |a_{N+1}^{(N+1)}|\,\{1 + ||T\gamma^{(N)}||_1\}.$$

But

$$\begin{aligned} ||T\gamma^{(N)}||_1 &\leqq ||\gamma^{(N)}||_1\,\{||I||_1 + ||X||_1\} \quad \text{using Theorem 1 notation} \\ &\leqq CN^{-p+1}\left\{1 + \max_{j=1}^{N} \sum_{i=j}^{N} |X_{ij}|\right\}. \end{aligned}$$

From Theorem 1, $|X_{ij}| \leqq Ki^{-p}$, so

$$\sum_{i=j}^{N} |X_{ij}| \leqq \left(\frac{p}{p-1}\right)Kj^{-p+1}, \quad \text{maximised by } j = 1.$$

Thus

$$||T\gamma^{(N)}|| \leqq C + (p/(p-1))CK \quad \text{and} \quad ||e_N||_{\mathfrak{L}} - ||e_{N+1}||_{\mathfrak{L}} \leqq k\,|a_{N+1}^{(N+1)}|$$

with $k \leqq 1 + C + (p/(p-1))CK$. For $M > N$,

$$||e_N||_{\mathfrak{L}} - ||e_M||_{\mathfrak{L}} = \sum_{i=N+1}^{M} \{||e_{i-1}||_{\mathfrak{L}} - ||e_i||_{\mathfrak{L}}\} \leqq \sum_{i=N+1}^{M} k|a_i^{(i)}|.$$

As $M \to \infty$, $||e_M||_{\mathfrak{L}} \to 0$, and in the limit

$$||e_N||_{\mathfrak{L}} \leqq k \sum_{i=N+1}^{\infty} |a_i^{(i)}|. \qquad \text{Q.E.D.}$$

We have now shown that, for the class of problems under consideration, the convergence of the error norm is characterised by the diagonal convergence of the system. However, the diagonal elements $|a_i^{(i)}|$ are not readily computable, so we proceed by relating them to quantities which are. The next theorem shows a connection with the column of free terms g.

THEOREM 3. *For a "nice" system having*

$$g_i = (h_i, g) \leqq \mathcal{C} i^{-r} \quad \textit{with } r > 1,$$

it follows that

$$|a_{N+1}^{(N+1)}| \leqq \alpha N^{-r} + \beta N^{-p}$$

and hence that

$$||e_N||_{\mathfrak{L}} \leqq \alpha' N^{-r+1} + \beta' N^{-p+1},$$

where α, β, α', β' are positive constants.

Proof. From the proof of Theorem 2, Eq. (9) gives

$$\mathbf{a}^{(N+1)} = J^{(N)}\mathfrak{g}^{(N)} - a_{N+1}^{(N+1)} J^{(N)} T^{(N)} \boldsymbol{\gamma}^{(N)}.$$

Substituting into Eq. (10),

$$(T\boldsymbol{\gamma}^{(N)})^T \cdot J^{(N)}\mathfrak{g}^{(N)} - (T\boldsymbol{\gamma}^{(N)})^T \cdot a_{N+1}^{(N+1)} J^{(N)} T\boldsymbol{\gamma}^{(N)} + J_{N+1} a_{N+1}^{(N+1)} = g_{N+1}.$$

Hence

$$J_{N+1} a_{N+1}^{(N+1)} = \frac{g_{N+1} - (T\boldsymbol{\gamma}^{(N)})^T \cdot J^{(N)}\mathfrak{g}^{(N)}}{1 - J_{N+1}(T\boldsymbol{\gamma}^{(N)})^T \cdot J^{(N)} T\boldsymbol{\gamma}^{(N)}}$$

$$= \frac{g_{N+1} - (T\boldsymbol{\gamma}^{(N)})^T \cdot J^{(N)}(T\mathbf{g}^{(N)})}{1 - J_{N+1}(T\boldsymbol{\gamma}^{(N)})^T \cdot (T\boldsymbol{\gamma}^{(N)})}$$

since

$$\mathfrak{g}^{(N)} = (\bar{\mathbf{h}}, g) = (T\mathbf{h}, g) = T\mathbf{g}^{(N)}.$$

Therefore

$$|a_{N+1}^{(N+1)}| \leqq \frac{|g_{N+1}| + ||(T\boldsymbol{\gamma}^{(N)})^T \cdot J^{(N)}(T\mathbf{g}^{(N)})||_1}{1 - ||(T\boldsymbol{\gamma}^{(N)})^T \cdot J^{(N)}(T\boldsymbol{\gamma}^{(N)})||_1}$$

$$\leqq \frac{|g_{N+1}| + ||T\boldsymbol{\gamma}^{(N)}||_\infty \, ||T\mathbf{g}^{(N)}||_1}{1 - ||T\boldsymbol{\gamma}^{(N)}||_\infty \, ||T\boldsymbol{\gamma}^{(N)}||_1} \tag{11}$$

$$\leqq \frac{|g_{N+1}| + ||T||_1 \, ||T||_\infty \, ||\boldsymbol{\gamma}^{(N)}||_\infty \, ||g^{(N)}||_1}{1 - ||T||_1 \, ||T||_\infty \, ||\boldsymbol{\gamma}^{(N)}||_\infty \, ||\boldsymbol{\gamma}^{(N)}||_1},$$

provided the denominator of (11) is positive.

We now bound the norms as follows:

$$||T||_1 \leqq 1 + ||X||_1 \leqq 1 + \left(\frac{p}{p-1}\right)K,$$

$$||T||_\infty \leqq 1 + ||X||_\infty \leqq 1 + K,$$

$$||g^{(N)}||_1 \leqq \mathfrak{C} \sum_{i=1}^{N} i^{-r} \leqq \mathfrak{C}\left(\frac{r}{r-1}\right),$$

$$||\gamma^{(N)}||_\infty \leqq C(N+1)^{-p} \leqq CN^{-p},$$

$$||\gamma^{(N)}||_1 \leqq C(N+1)^{-p+1} \leqq CN^{-p+1}.$$

Hence, the denominator of (11)

$$\geqq 1 - \left[1 + \left(\frac{p}{p-1}\right)K\right][1 + K]C^2N^{-2p+1}$$

$$\geqq 1 - KCN^{-2p+1}$$

(using an inequality from the proof of Theorem 1)

$$\geqq 1 - KC > 0.$$

Therefore,

$$|a_{N+1}^{(N+1)}| \leqq \frac{\mathfrak{C}}{(1-KC)}\left[N^{-r} + K\left(\frac{r}{r-1}\right)N^{-p}\right] \leqq \alpha N^{-r} + \beta N^{-p}$$

where $\alpha \leqq \mathfrak{C}/(1-KC)$ and $\beta \leqq (K\mathfrak{C}/(1-KC))(r/(r-1))$.

Using Theorem 2,

$$||e_N||_{\mathfrak{L}} \leqq k \sum_{i=N}^{\infty} [\alpha i^{-r} + \beta i^{-p}]$$

$$\leqq \frac{k\alpha r}{(r-1)} N^{-r+1} + \frac{k\beta p}{(p-1)} N^{-p+1}$$

$$= \alpha' N^{-r+1} + \beta' N^{-p+1}. \qquad \text{Q.E.D.}$$

This result enables us to predict the asymptotic convergence rate of the variational procedure, provided we know the convergence properties of the terms $g_i = (h_i, g)$, i.e. the convergence rate of the generalised Fourier coefficients of the function g. It is frequently unnecessary, however, to compute these coefficients in order to determine their convergence rate. Given certain qualitative information about a function (such as its continuity, its differentiability, and its boundary behaviour) the convergence of generalised Fourier coefficients, with respect to a given expansion set, may frequently be predicted "a priori". This problem is considered further in a separate paper (II).

We conclude this section by considering the case in which, either through physical considerations or otherwise, we have qualitative information concerning the true solution of our equation and are thus in a position to predict the convergence rate of the coefficients b_i. The next theorem relates this "vertical" convergence to the convergence of the error norm.

Having determined the error convergence in terms of properties of a known function g, it may well seem unnecessary to obtain a similar result in terms of properties of the unknown solution f.

However, the step is of importance in considering the extension of these results to homogeneous systems (e.g. eigenproblems) for which, of course, no right-hand side function g exists. For an inhomogeneous system, the convergence rates of the b_i and g_i are related; we display the relation later.

THEOREM 4. *For a "nice" system having*

$$b_i = (h_i, f) \leqq \kappa i^{-q}, \quad \text{with } q > 1,$$

it follows that $|g_N| \leqq \gamma N^{-q} + \delta N^{-p}$ *and hence that* $||e_N||_{\mathfrak{L}} \leqq \gamma' N^{-q+1} + \delta' N^{-p+1}$, *where* γ, δ, γ', δ' *are positive constants.*

Proof.

$$\begin{aligned} g_N &= (h_N, g) = (h_N, \mathfrak{L}f) \\ &= \left(h_N, \mathfrak{L} \sum_{i=1}^{\infty} b_i h_i\right) = \sum_{i=1}^{\infty} b_i (h_N, \mathfrak{L}h_i) \\ &= \sum_{i=1}^{\infty} b_i L_{Ni} = J_N b_N + \sum_{i=1}^{N-1} b_i L_{Ni} + \sum_{i=N+1}^{\infty} b_i L_{Ni}. \end{aligned}$$

Thus,

$$\begin{aligned} |g_N| &\leqq \kappa N^{-q} + \left(\frac{q}{q-1}\right) C\kappa N^{-p} + \frac{C\kappa}{(p+q-1)} N^{-p-q+1} \\ &\leqq \kappa\left(1 + \frac{C}{q}\right) N^{-q} + \left(\frac{q}{q-1}\right) C\kappa N^{-p} \\ &= \gamma N^{-q} + \delta N^{-p}. \end{aligned}$$

From the proof of Theorem 3,

$$|a_{N+1}^{(N+1)}| \leqq \frac{|g_N|}{(1 - KC)} + \beta N^{-p}.$$

Hence

$$|a_{N+1}^{(N+1)}| \leqq \frac{\gamma}{(1 - KC)} N^{-q} + \left[\frac{\delta}{1 - KC} + \beta\right] N^{-p}$$

and so

$$\begin{aligned} ||e_N||_{\mathfrak{L}} &\leqq k \sum_{i=N+1}^{\infty} |a_i^{(i)}| \\ &\leqq \frac{k\gamma q}{(1 - KC)(q-1)} N^{-q+1} + \frac{kp}{p-1}\left[\frac{\delta}{1 - KC} + \beta\right] N^{-p+1} \\ &\leqq \gamma' N^{-q+1} + \delta' N^{-p+1}. \end{aligned}$$

Q.E.D.

An important conclusion from the above theorem is that the Fourier (vertical) convergence rate will dominate, provided $p > q$.

III. A Theorem Based on the Variation Principle (2). The theorems of Section II made no reference to the variational functional (2). If we assume that this functional exists, and in addition that the operator $\mathfrak{L}$ is positive, we may rederive Theorem 4 in a stronger form:

THEOREM 5. *$\mathfrak{L}$ is a positive Hermitian operator, and the solution of* (1) *is given by*

$$f\colon \min_{\omega\in\mathfrak{R}} F(\omega) = F(f).$$

In addition, for the suitably normalised expansion set $\{h_i\}$ the matrix L is U.A.D. (P, C) with $L_{ii} = 1$ and $p > \frac{1}{2}$ and $b_i = (h_i, f) \leqq \kappa i^{-q}$; $q > \frac{1}{2}$; $p + 2q > 2$.

We do not here require any restrictions on the constant C; further, the restrictions on p, q are weaker than in Theorem 4. *Under these assumptions,*

$$||\epsilon_N||_{\mathfrak{L}}^2 \leqq \gamma'' N^{-2q+1} + \delta'' N^{-(p+2q-2)}.$$

Proof. We have for any element $f_N = f + \epsilon_N$,

$$F(f_N) = F(f) + (\epsilon_N, \mathfrak{L}\epsilon_N) = F(f) + ||\epsilon_N||_{\mathfrak{L}}^2.$$

The minimum of $F(f_N)$, and hence of $||\epsilon_N||$, is therefore no greater than that given for any choice of the coefficients $a_i^{(N)}$. We choose

$$a_i^{(N)} = b_i, \qquad i = 1, 2, \cdots, N,$$

and hence find

$$\begin{aligned}
||\epsilon_N||_{\mathfrak{L}}^2 &\leqq \sum_{i,j=N+1}^{\infty} b_i^* b_j L_{ij} \\
&\leqq \sum_{i=N+1}^{\infty} |b_i|^2 + 2\sum_{i=N+1}^{\infty}\sum_{j=i+1}^{\infty} |b_i|\,|b_j|\,|L_{ij}| \\
&\leqq \kappa^2 \sum_{i=N+1}^{\infty} i^{-2q} + 2\kappa^2 C \sum_{i=N+1}^{\infty} i^{-q} \sum_{j=i+1}^{\infty} j^{-p-q} \\
&\leqq \frac{\kappa^2}{2q-1} N^{-2q+1} + \frac{2\kappa^2 C}{(p+q-1)(p+2q-2)} N^{-(p+2q-2)}. \qquad \text{Q.E.D.}
\end{aligned}$$

This result is slightly better than that given by Theorem 4.

IV. The Fourier and Variational Convergence Rates. The proof of Theorem 4 bounds the coefficients g_i in terms of the Fourier coefficients b_i; both Theorems 4 and 5 bound the norm $||\epsilon_N||_{\mathfrak{L}}$ in terms of the convergence rate of the b_i and the degree p of L. Both the b_i and the variational coefficients $a_i^{(N)}$ are, of course, uniquely determined by L and g. In this section we derive several theorems relating the convergence of b_i; and of $a_i^{(N)}$ to b_i; to the convergence of L and of g. We first give a bound on the coefficients b_i.

THEOREM 6. *For a "nice" system having*

$$g_i = (h_i, g) \leqq \mathcal{C} i^{-r} \quad \text{with } r > 1,$$

it follows that, for all i and some D,

$$|b_i| \leqq D i^{-s} \quad \text{where } s = \min(p, r).$$

Proof. The coefficients $\mathbf{b}$ satisfy the infinite set of equations

$$L\mathbf{b} = \mathbf{g}.$$

In terms of the triangular matrix T of Theorem 1, we set

$$\mathbf{b} = T^T\mathbf{c},$$

so that

$$\mathbf{c} = JT\mathbf{g},$$

whence

$$|c_i| \leqq \sum_{j=1}^{i} |T_{ij}|\,|g_j| = \sum_{j=1}^{i} |X_{ij}|\,|g_j| + |g_i|$$

$$\leqq K\mathcal{C}\sum_{j=1}^{i} i^{-p}j^{-r} + \mathcal{C}i^{-r}$$

$$\leqq \frac{K\mathcal{C}r}{r-1}\,i^{-p} + \mathcal{C}i^{-r}, \qquad r > 1.$$

We now, similarly, bound $b_i = c_i + \sum_{j=i}^{\infty} X_{ij}c_j$ to obtain the result

$$|b_i| \leqq \frac{K\mathcal{C}r}{r-1}\,i^{-p} + \mathcal{C}i^{-r} + \frac{2K^2\mathcal{C}rp}{(r-1)(2p-1)}\,i^{-2p+1} + \frac{K\mathcal{C}(p+r)}{(p+r-1)}\,i^{-(p+r-1)}.$$

Hence, for some D,

$$|b_i| \leqq Di^{-s} \quad \text{where } s = \min(p, r). \qquad \text{Q.E.D.}$$

Comment. We may feed this result back into Theorem 4; we then recover Theorem 3. This suggests that the bounds we obtain are the best possible so far as the predicted convergence rates are concerned. We see also that the Fourier (vertical) convergence rate dominates the convergence in the $\mathcal{L}$ norm provided that $p > r$ (see Theorem 4).

Finally, a similar procedure allows us to investigate the convergence of the individual variational coefficients $a_i^{(N)}$:

THEOREM 7. *Under the conditions of Theorem* 6 *it follows that*

$$|b_i - a_i^{(N)}| \leqq D_1N^{-(2p-1)}i^{-(r-1)} + D_2N^{-q'} \qquad \forall N;\ i = 1, 2, \cdots, N,$$

where $q' = \min\{p + r - 1, 2p - 1, 2p + r - 2\}$. *If* $r \geqq 2$, *this implies* $q' = \min\{p + r - 1, 2p - 1\}$.

Proof. We remark that the finite matrix T_N is a submatrix of T. We also have

$$\mathbf{a}^{(N)} = T_N^T J^{(N)} T_N \mathbf{g}_N,$$

$$\mathbf{b} = T^T JT\mathbf{g},$$

whence it follows, after some reduction, that

$$b_i - a_i^{(N)} = \sum_{j=N+1}^{\infty} J_j X_{ij} g_j + \left[\sum_{j=i}^{N}\sum_{k=N+1}^{\infty} + \sum_{j=N+1}^{\infty}\sum_{k=j}^{\infty} + \sum_{j=1}^{i-1}\sum_{k=N+1}^{\infty}\right] X_{ki}J_kX_{kj}g_j.$$

Inserting the bounds on X, $\mathbf{g}$ and bounding the series as before, we find

$$|b_i - a_i^{(N)}| \leqq \frac{K\mathcal{C}}{p+r-1} N^{-(p+r-1)} + \frac{K^2\mathcal{C}r}{(2p-1)(r-1)} N^{-(2p-1)} i^{-(r-1)}$$
$$+ \frac{2pK^2\mathcal{C}}{(2p-1)(2p+r-2)} N^{-(2p+r-2)} + \frac{K^2\mathcal{C}r}{(2p-1)(r-1)} N^{-(2p-1)},$$

whence the theorem follows by inspection.

Comments. (1) In view of the form of the second term in the theorem, no essential information is lost by the following simplification:

COROLLARY 1. *For some* D_3, $|b_i - a_i^{(N)}| \leqq D_3 N^{-q'}$.

(2) The uniform nature of the bound (with respect to i) implies rather surprisingly that *diagonal* convergence is as rapid as *horizontal* convergence.

(3) We have provided in Theorems 6, 7 *upper bounds* on b_i, $b_i - a_i^{(N)}$. If, in addition, we assume that the bounds on b_i are sufficiently tight we may obviously bound the *relative* error in $a_N^{(N)}$:

COROLLARY 2. *If for some subset* $\{m\}$ *of the integers* $\{i\}$, $|b_m| \geqq D_4 m^{-s}$ *where* s *is given in Theorem* 6, *then*

$$\left|\frac{b_m - a_m^{(m)}}{b_m}\right| \leqq D_5 m^{-p+1}.$$

Further, for all $N > m$,

$$\left|\frac{b_m - a_m^{(N)}}{b_m}\right| \leqq D_6 \left(\frac{N}{m}\right)^{-s} N^{-q''}$$

where $q'' = q' - s > 0$. *Hence, for all values of* $m \in \{m\}$, $a_m^{(N)}/b_m \to 1$, $N \to \infty$.

V. Convergence in the Natural Norm. Theorems 6 and 7 essentially characterise the convergence problem. From these, we can bound the error in norms other than the energy ($\mathcal{L}$) norm considered so far. As an example, we compute the natural norm

$$||\epsilon_N||^2 = (\epsilon_N, \epsilon_N)$$

under the assumption that the set $\{h_i\}$ is orthogonal in R. This norm is given by (5); we recall however that the normalisation implied there for the h_i is not that used in Theorems 1–7. We define an orthonormal set of functions $\bar{h}_i$:

$$\bar{h}_i = \gamma_i h_i; \qquad (\bar{h}_i, \bar{h}_j) = \delta_{ij}; \qquad (h_i, h_i) = \gamma_i^{-2}, \tag{12}$$

where the h_i have the normalisation of Theorems 1–7 and γ_i satisfies

$$(\bar{h}_i, \mathcal{L}\bar{h}_i) = \gamma_i^2. \tag{13}$$

In terms of the expansion coefficients $a_i^{(N)}$, b_i appropriate to the set $\{h_i\}$, (5) becomes

$$||\epsilon_N||^2 = \sum_{i=1}^{N} \gamma_i^{-2} |b_i - a_i^{(N)}|^2 + \sum_{i=N+1}^{\infty} \gamma_i^{-2} b_i^2. \tag{14}$$

Whence from Theorems 6 and 7, we obtain the following result.

THEOREM 8. *The conditions of Theorems* 6 *and* 7 *apply and in addition*

$$\gamma_i^{-2} \leqq \Gamma i^{\gamma}.$$

Then

$$||\epsilon_N||^2 \leqq C^2 \frac{\Gamma\gamma}{\gamma+1} N^{-2q'+\gamma+1} + \frac{D^2\Gamma}{2s-\gamma-1} N^{-2s+\gamma+1}.$$

Proof. As before, we bound the terms in (14), using the results of Theorems 6 and 7 and Corollary 1.

Comment. The two terms correspond to $S_1(N)$, $S_2(N)$ respectively; hence, since $q' > s$ we see that *for the systems considered, $S_2(N)$ dominates the convergence rate.*

VI. Conclusions. In this paper, we have been concerned with characterising the convergence of a variational calculation in terms of parameters which are easy to compute. We believe that, for systems of the structure considered, the bounds are realistic; they may be used in practice to give a priori estimates of the convergence rates for a given expansion set, and hence to influence the choice of this set.

Although our results are obtained only for U.A.D. systems, it is our belief that they can be extended to a much wider class of A.D. systems; we hope to report on suitable extensions at a later date.

Department of Computational and Statistical Science
University of Liverpool
Liverpool, England

1. L. V. KANTOROVIČ & V. I. KRYLOV, *Approximate Methods of Higher Analysis*, GITTL, Moscow, 1950; English transl., Noordhoff, Groningen, 1958. MR **13**, 77; MR **21** #5268.
2. S. G. MIKHLIN, *Variational Methods in Mathematical Physics*, GITTL, Moscow, 1957; English transl., Macmillan, New York, 1964. MR **22** #1981; MR **30** #2712.
3. C. SCHWARTZ, *Estimating Convergence Rates of Variational Calculations*, Methods in Computational Physics, vol. 2, Academic Press, New York, 1963.
4. J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965. MR **32** #1894.