QUARTERLY OF APPLIED MATHEMATICS
JULY, 1973

MEAN CONVERGENCE OF APPROXIMATION TO A FUNCTION BY GENERAL FINITE SUMS*

By
HWA-SHAN HO
University of Southern California

Abstract. The approximation of a function by a general finite sum (linear combination of non-orthogonal functions) is considered here. It is shown that the mean error of such an approximation, defined in the sense of any weighted inner product in the Hilbert space, is positive semi-definitely decreasing as the number of terms in the expansion increases. Conditions under which the mean error is stationary are thoroughly discussed. Some interesting properties of such approximations are revealed by related theorems. The theorems are proven for complex variables, and are valid of course for real variables.

Introduction. Among the methods of approximation of a function, series or finite-sum approximations are those most commonly used. While orthonormal series approximations prove to be convenient and easier most of the time, approximations in terms of general finite sums (sometimes denoted loosely as non-orthogonal "series", the terminology used by Kantorovich and Krylov [1]) are sometimes more useful. The Weierstrass approximation, the various variational methods, as well as most perturbation methods fall into the latter category.

The mean convergence in the L_2 sense of orthonormal series is easily proven in a fairly straightforward fashion. But the proof of such mean convergence of general finite-sum approximations is far from trivial. In fact, in the various methods where such approximations are utilized (e.g. [1]), mean convergence in the infinite sum is simply stated as the results of the "completeness" of the functions $[\varphi_n]$ used in the approximation. This presents two problems: first, a criterion for the "completeness" of the functions $\{\varphi_n\}$ is not known a priori. Secondly, completeness regulates only the behavior of $\{\varphi_n\}$ as the number of terms n goes to infinity. Since n has to be finite in such an expansion, due to restrictions to be seen later (as is generally the case in actual applications), completeness in such a sense loses its significance in such mean convergence.

In this paper approximations by general non-orthogonal expansions are considered. Approximations are obtained to minimize the "mean error" of the expansion. Here the mean error is defined in the sense of any weighted inner product as defined in an infinite functional space (called Hilbert space if complete). (See, e.g., Dettman [2], and Courant and Hilbert [3].)

* Received January 27, 1971; revised version received October 25, 1972. Some preliminary results of this paper were obtained when the author was writing his Ph.D. dissertation at Brown University under the supervision of Prof. Paul S. Symonds.

It is shown that when the number of terms in the expansion is increased, the mean error of expansion is never increased. In fact, the mean error is shown to decrease always, except for a few cases when it becomes stationary. The "stationary" conditions are those to be avoided if completeness of the series is to be achieved. In the proof, no special requirement on the "modal" functions is required except that they be square-integrable.

This paper is the outcome of a study on general mode approximation theorems in the study of impulsive loading problems in continuum mechanics, as a generalization of the simple mode approximation by Martin and Symonds [6], and subsequently extended by Ho [7]. In such problems, the function expanded is the velocity distribution, and the quadratic form is directly proportional to the kinetic energy of the approximate velocity field. The weighting function $m(x, y, z)$ is the mass density distribution of the structural system.

Formulation of the approximation and related theorems.

(A) *Formulations of the approximation.* Let us approximate any square-integrable function $f(x, y, z)$ by a function $f_n(x, y, z)$ in the form of a finite sum of n arbitrary functions $\{\varphi_n\}$, i.e.:

$$\begin{aligned} f_n(x, y, z) &= a_p^n \varphi_p(x, y, z) \qquad p = 1, \cdots n \\ &= f(x, y, z) - F_n(x, y, z). \end{aligned} \tag{1}$$

Here a_p^n is the coefficient of the "modal function" φ_n and $F_n(x, y, z)$ is the residue, both dependent upon the number of terms n used in the expansion. For convenience, the summation convention will be used throughout the paper; i.e., whenever a subscript is repeated twice, that subscript is to be summed over its entire range (for p, q and r, the range is from 1 to n; for i and j, it is from 1 to $n + 1$), unless otherwise mentioned. Here and throughout the paper, an equality means equality "almost everywhere" in the region of the physical system. The independent variables (x, y, z) will be omitted at times.

The "modal functions" φ_p need not be orthogonal, but will be normalized in the modified inner product in the Hilbert space, i.e.

$$c_{pq} = (\varphi_p , \varphi_q) = \int m(x, y, z)\varphi_p^* \varphi_q \, dV \tag{2}$$

with

$$c_{pq} = 1 \quad \text{if} \quad p = q. \tag{3}$$

Here the asterisk denotes the complex conjugate, the integration is over V, the region of the physical system, and $m(x, y, z)$ is a real and positive weighting function of the physical system. To begin with, we assume that the n modes φ_p are linearly independent. This condition will be relaxed later to examine the behavior ensuing when φ_{n+1} is linearly dependent on the n lower modes φ_1 to φ_n . The "interaction matrix" (c_{pq}) possesses the following properties:

$$c_{pq} = c_{qp}^* \quad \text{(Hermitian)}, \tag{4}$$

$$0 \leq |c_{pq}| \leq 1 \quad -1 \leq \operatorname{Re} c_{pq} \leq 1. \tag{5}$$

Relation (5) comes from the triangular inequality

$$0 \leq \int m\,|\varphi_p \pm \varphi_q|^2\,dV = c_{pp} + c_{qq} \pm (c_{pq} + c_{qp})$$

$$= 2 \pm (c_{pq} + c_{pq}^*) = 2 \pm 2(\mathrm{Re}\, c_{pq}), \qquad p,\ q \text{ not summed} \tag{6}$$

and from the Schwarz inequality:

$$|c_{pq}| \leq \int |m\varphi_p^*\varphi_q|\,dV \leq \left(\int m\,|\varphi_p|^2\,dV\right)^{1/2}\left(\int m\,|\varphi_p|^2\,dV\right)^{1/2} = 1. \tag{7}$$

The first equality in (5) is true iff φ_p and φ_q are orthogonal. The other equalities in (5) hold iff $\varphi_p = \pm\varphi_q$.

It is noted here, once and for all, that the orthonormal expansions (such as Fourier series) are special cases of the present problem when the interaction matrix (c_{pq}) is a unit matrix, i.e.

$$c_{pq} = 0, \qquad \text{if } p \neq q; \qquad c_{pq} = 1 \text{ if } p = q.$$

Hence all theorems established here will be valid for orthonormal expansions. (In fact, Theorem IV is equivalent to Bessel's inequality.) For a discussion of Hilbert space and orthonormal expansion, the reader is referred to any book on mathematical methods or vector spaces ([2] and [3]).

To obtain the expansion coefficients a_p^n , we minimize the "mean error" of the expansion

$$I_n = \int m\,|f - f_n|^2\,dV = \int m(f - a_p^n\varphi_p)^*(f - a_q^n\varphi_q)\,dV \tag{8}$$

with respect to the coefficients a_p^n ; i.e.

$$\partial I_n/\partial a_p^n = 0, \qquad p = 1, \cdots, n. \tag{9}$$

Or we can decompose a_p^n into real and imaginary parts, a_{pr}^n and a_{pi}^n respectively, and minimize I as follows:

$$\partial I/\partial a_{pr}^n = \partial I/\partial a_{pi}^n = 0 \quad p = 1, \cdots, n. \tag{9a}$$

We then obtain the following system of linear algebraic equations for the solution of $\{a_p^n\}$:

$$c_{pq}a_q^n = b_p \quad p, q = 1, \cdots n, \tag{10}$$

with

$$b_p = \int m\varphi_p^* f\,dV, \tag{11}$$

$$|b_p| \leq \left(\int m\,|f|^2\,dV\right)^{1/2}. \tag{12}$$

The equality in (12) holds iff $\varphi_p \propto f$. It is seen that the process of finding a_p^n is the same as in orthonormal expansion and as that used by Kantorovich and Krylov [1] in non-orthogonal expansion without explanation. Since we assumed that the n modes $\{\varphi_p\}$ are linearly independent, the interaction matrix (c_{pq}) is non-singular. Hence the system

(10) possesses a unique solution for $\{a_p^n\}$:

$$a_p^n = d_{pq} b_q \tag{13}$$

where (d_{pq}) (also a hermitian matrix) is the inverse of (c_{pq}). Using the Kronecker delta, we have:

$$c_{pq} d_{qr} = d_{pq} c_{qr} = \delta_{pr} . \tag{14}$$

Our objective now is to prove that as we increase the number of terms in the approximation, the mean error decreases:

$$I_{n+1} \leq I_n . \tag{15}$$

To prove (15), first let us examine some of the properties of such approximations.

(B) *Basic properties of such approximations.* The following theorems can be established for the approximation described in (A).

THEOREM I: The residue F_n in the approximation is orthogonal to all the modes $\{\varphi_p\}$, i.e.

$$\int m\varphi_p^* F_n \, dV = 0, \qquad p = 1, \cdots n.$$

Proof:

$$\int m\varphi_p^* F_n \, dV = \int m\varphi_p^* (f - a_q^n \varphi_q) \, dV$$

$$= b_p - c_{pq} a_q^n = 0 \qquad p, q = 1, \cdots n. \qquad \text{Q.E.D.}$$

THEOREM II: The mean square error I_n can be written as the difference between the square integrals of the actual and the approximate functions.

Proof:

$$\int m f_n^* f \, dV = \int m a_p^{n*} \varphi_p^* f \, dV = a_p^{n*} b_p = a_p^{n*} c_{pq} a_q^n$$

$$= \int m (a_p^{n*} \varphi_p^*)(a_q^n \varphi_q) \, dV = \int m \, |f_n|^2 \, dV \left(= \int m f^* f_n \, dV \right).$$

Hence

$$I_n = \int m \, |f - f_n|^2 \, dV = \int m \, |f|^2 \, dV - \int m (f_n^* f + f^* f_n) \, dV + \int m \, |f_n|^2 \, dV$$

$$= \int m \, |f|^2 \, dV - \int m \, |f_n|^2 \, dV. \qquad \text{Q.E.D.} \tag{16}$$

This also implies

$$\int m F_n^* f \, dV = \int m (f - f_n)^* f \, dV = \int m \, |f|^2 \, dV - \int m \, |f_n|^2 \, dV$$

$$= I_n \left(= \int m f^* F_n \, dV \right). \tag{17}$$

THEOREM III: The associated quadratic form

$$Q_n = c_{pq} a_p^{n*} a_q^n = a_p^{n*} b_p = d_{pq} b_p^* b_q \tag{18}$$

is real, positive-semi-definite and bounded; i.e.

$$0 \leq Q_n = Q_n^* \leq \int m\,|f|^2\,dV. \tag{19}$$

Proof: From the proof of Theorem II,

$$Q_n = \int m\,|f_n|^2\,dV \geq 0 \quad \text{(real, positive-semi-definite)}. \tag{20}$$

Equality occurs if f is orthogonal to all n modes. Also

$$0 \leq I_n = \int m\,|f|^2\,dV - Q_n\,. \tag{21}$$

Hence, combining (20) and (21), we have

$$0 \leq Q_n \leq \int m\,|f|^2\,dV,$$

$$I_n = 0 \Rightarrow f \equiv f_n\,.$$

The quadratic form Q_n (now a "Hermitian form") is real and positive-semidefinite (relation (20)) due to the properties of the matrix (c_{pq}). For a thorough investigation of this subject, the reader is referred to texts on matrix theories such as those by Gantmacher [4], and Hoffman and Kunze [5].

(C) *Stepwise convergence in mean.* Let us introduce an additional mode φ_{n+1} such that

$$c_{n+1,n+1} = \int m\varphi_{n+1}^*\varphi_{n+1}\,dV = 1 \tag{22}$$

and

$$f = f_{n+1} + F_{n+1} = a_i^{n+1}\varphi_i + F_{n+1}\,, \quad i = 1, \cdots n+1. \tag{23}$$

Also, φ_{n+1} may be expanded in terms of $\{\varphi_p\}$:

$$\varphi_{n+1} = h_p^n \varphi_p + R_n \tag{24}$$

where R_n, F_{n+1} are the residues and h_p^n, a_i^{n+1} are the expansion coefficients, determined in a similar manner to a_p^n (Eq. (10)), i.e.:

$$c_{ij} a_j^{n+1} = b_i\,; \qquad b_{n+1} = \int m\varphi_{n+1}^* f\,dV \tag{25}$$

$$c_{pq} h_q^n = c_{p,n+1}\,; \quad h_p^n = d_{pq} c_{q,n+1}\,; \quad i, j = 1, \cdots, n+1. \tag{26}$$

The interaction coefficient $c_{p,n+1}$ satisfy properties (4)–(7). The quadratic form associated with (26) will be denoted by g_n, called the "augmental quadratic form":

$$g_n = c_{pq} h_p^{n*} h_q^n = d_{pq} c_{p,n+1}^* c_{q,n+1}$$

$$= d_{pq} c_{n+1,p} c_{q,n+1}\,. \tag{27}$$

LEMMA I: The augmental quadratic form satisfies the constraints:

$$0 \leq g_n \leq 1, \tag{28}$$

$$g_n = 1, \quad \text{iff} \quad R_n = 0, \tag{29}$$

i.e., iff φ_{n+1} is linearly dependent on the n lower modes.

Proof: g_n is a special form of Q_n when f is φ_{n+1}. Since φ_{n+1} is normalized (22), substitution for f of φ_{n+1} in Theorem III leads to (28) and (29).

THEOREM IV: The stepped-up quadratic form is no less than Q_n, i.e.

$$Q_{n+1} = c_{ij} a_i^{(n+1)*} a_j^{n+1} \geq Q_n = c_{pq} a_p^{n*} a_q^n, \qquad i, j = 1, \cdots n+1; \quad p, q = 1, \cdots n. \tag{30}$$

Notice that, in general,

$$a_i^{n+1} \neq a_p^n \quad \text{for} \quad i = p. \tag{31}$$

Proof: The first n equations in (25) can be written as

$$a_p^{n+1} = d_{pq}(b_q - c_{q,n+1} a_{n+1}^{n+1}) \quad p, q = 1, \cdots n. \tag{32}$$

The last equation in (25) is

$$c_{n+1,p} a_p^{n+1} + a_{n+1}^{n+1} = b_{n+1}. \tag{33}$$

Using (32), we can write (33) as

$$\begin{aligned} a_{n+1}^{n+1} &= (b_{n+1} - c_{n+1,p}\, d_{pq} b_q)/(1 - c_{n+1,p}\, d_{pq} c_{q,n+1}) \\ &= (b_{n+1} - c_{n+1,p}\, d_{pq} b_q)/(1 - g_n). \end{aligned} \tag{34}$$

Hence

$$\begin{aligned} Q_{n+1} &= a_i^{n+1} b_i^* = a_p^{n+1} b_p^* + a_{n+1}^{n+1} b_{n+1}^* \\ &= b_p^*\, d_{pq}(b_q - c_{q,n+1} a_{n+1}^{n+1}) + b_{n+1}^* a_{n+1}^{n+1} \\ &= d_{pq} b_p^* b_q + a_{n+1}^{n+1}(b_{n+1}^* - b_p^* d_{pq} c_{q,n+1}). \end{aligned} \tag{35}$$

Noting the Hermitian properties of d_{pq} and $c_{q,n+1}$, and using (18), we then have

$$\Delta Q_n = Q_{n+1} - Q_n = a_{n+1}^{n+1}(b_{n+1} - c_{n+1,p} d_{pq} b_q)^*. \tag{36}$$

Combining (34) and (36), we have

$$\Delta Q_n = |b_{n+1} - c_{n+1,p} d_{pq} b_q|^2/(1 - g_n). \tag{37}$$

Eq. (37) and Lemma I imply that

$$\Delta Q_n \geq 0 \quad \text{if} \quad \varphi_{n+1} \neq h_p^n \varphi_p. \tag{38}$$

The special case when φ_{n+1} is linearly dependent on the n lower modes ($g_n = 1$) will be examined in the next theorem. (Q.E.D.)

THEOREM V (stationary conditions): ΔQ_n is zero iff φ_{n+1} is orthogonal to the residue F_n of f, i.e.

$$\int m \varphi_{n+1}^* F_n \, dV = \int m F_n^* \varphi_{n+1} \, dV = 0. \tag{39}$$

Proof. In order that $\Delta Q_n = 0$, it is necessary (by (37)) and sufficient (by (36)) that

$$b_{n+1} - c_{n+1,p} d_{pq} b_q = 0.$$

The result now follows, since

$$b_{n+1} - c_{n+1,p}\, d_{pq} b_q = b_{n+1} - c_{n+1,p} a_p^n$$
$$= \int m\varphi_{n+1}^* (f - a_p^n \varphi_p)\, dV = \int m\varphi_{n+1}^* F_n\, dV. \qquad \text{(Q.E.D.)}$$

Since F_n is orthogonal to $\varphi_1, \cdots, \varphi_n$, (24) shows that (39) is equivalent to

$$\int m R_n^* F_n\, dV = \int m F_n^* R_n\, dV = 0. \tag{40}$$

There are two special cases: (i) $F_n = 0$, i.e. $f = a_p^n \varphi_p$. In this case, the approximation is exact and Q_n reaches its upper bound $\int m f^2\, dV$. Hence, no improvement can be made. (ii) $R_n = 0$, i.e. $\varphi_{n+1} = h_p^n \varphi_p$. In this case, the φ_{n+1} mode is linearly dependent on the n lower modes. No advantage is gained by the introduction of this mode, since $(f - F_n)$ is already taken care of by the n lower modes. In fact, we will have $F_{n+1} = F_n$. We now note that, by (21) and Theorem IV, the mean error I_n satisfies $I_n - I_{n+1} = Q_{n+1} - Q_n \geq 0$; consequently we have:

THEOREM VI (stepwise convergence in mean): The mean error I_n is positive-definitely decreasing when the number of terms in the approximation is increased, except for the stationary condition stated in Theorem V.

If closure at infinity is desired, i.e. $\lim_{n\to\infty} I_n = 0$, the sequence of functions $\{\varphi_n\}$ must form a "complete" set similar to the orthogonal expansions ([1], [2]). However, since the expansion coefficients are obtained from the solution of a system of linear algebraic equations, it is generally impossible, except for special cases, to handle infinite values of n under this type of expansion. This is precisely why the stepwise convergence of such an approximation is so important.

Geometric interpretations. The quantities and the theorems proven in the previous section can be interpreted clearly from their counterparts in geometry. The square integral of a function is the "square of the length" of the "vector" of the function in an infinite-dimensional vector space, and is real and positive definite. Thus Theorem I implies the residue vector is "perpendicular" to the base vectors $\varphi_1, \cdots \varphi_n$ and f_n is the projection of the vector f in the subspace defined by $\{\varphi_1, \cdots \varphi_n\}$. This is a direct result from minimizing the length of the residue vector. Theorem II is the Pythagorean theorem. Theorem III implies that the length of the projected vector cannot be larger than the original vector. Theorem IV implies that the length of the projection of a vector in an $(n + 1)$-dimensional space cannot be less than that in an n-dimensional space. Theorem V implies that the lengths of the projections in n- and $(n + 1)$-dimensional spaces are equal if the base vector φ_{n+1} is perpendicular to the residue F_n. All these are simple geometric relations. They can be used to assist in the understanding of their algebraic counterparts.

Conclusion. The proof given here for the "stepwise" mean convergence of any non-orthogonal expansion justifies the applicability of such approximations. The "sta-

tionary" conditions given in Eqs. (39), (40), (42) and (43) are to be avoided if completeness is to be achieved. Other related theorems demonstrate clearly the behaviors of such expansions.

The inclusion of the weighting function here, as compared to other standard treatments (e.g. [1] and [2]), makes this expansion very suitable for problems of statistical nature where the weighting function is identified with the distribution function. The validity of the theorems in complex as well as real variables makes them applicable in wave mechanics, both quantum and classical. It is hoped that the theorems presented here will encourage studies of various physical phenomena by way of approximations by some types of non-orthogonal expansions.

In a future paper [8] a mathematical system described by a symmetric differential and/or integral operator is examined. It is also shown there that the solution can be approximated by a finite sum, and that such an approximation also exhibits the stepwise mean convergence proven here.

References

[1] L. V. Kantorovich and V. I. Krylov, *Priblizhennye metody vysshego analiza*, English translation by C. D. Benster, *Approximate methods of higher analysis*, Interscience, Noordhoff, Groningen, Netherlands, 1958

[2] J. W. Bettman, *Mathematical methods in physics and engineering*, McGraw-Hill, N. Y., 1962

[3] R. Courant and D. Hilbert, *Methods of mathematical physics*, Vol. I, Interscience, N. Y., 1953

[4] F. R. Gantmacher, *Teoriya matrits*, English translation by K. A. Hirsch, *The theory of matrices*, Vol. I, Chelsea, N. Y., 1959

[5] K. Hoffman and R. Kunze, *Linear algebra*, Prentice-Hall, Englewood, N. J., 1964

[6] J. B. Martin and P. S. Symonds, *Mode approximations for impulsively loaded rigid-plastic structures*, J. Engineering Mechanics Division, Amer. Soc. Civil Engineers 92, 43–66 (1966)

[7] Hwa-Shan Ho, *Convergent approximations of problems of impulsively loaded structures*, J. Appl. Mech. 38, 852–860 (1971)

[8] Hwa-Shan Ho, *Stepwise convergence in linear systems employing non-orthogonal finite sums* (manuscript)