NEW METHOD TO SOLVE A LINEAR FUNCTIONAL EQUAION AND IMPROVED SCHMIDT'S ORTHONORMALIZATION, WITH THEIR APPLICATION TO ANALYSE SCATTERING BY A HOLLOW PIPE OF FINITE LENGTH

Yoshio HAYASHI

2-14-12, Shimoigusa, Suginami-ku, Tokyo, 167-0022, Japan

E-mail name hayashiy@cg.mbn.or.jp

1. INTRODUCTIION

The method of moments is the most acknowledged and widely accepted technique to solve a linear functional equation. However, solutions obtained by the method are not necessarily the best ones. Contrary to this, the new method developed in this paper enables us to obtain, in a similar programming and computing time as a moment method, always the best approximate solutions which make error minimum, which hence converge to the true solution most rapidly. Furthermore, our method is the most stable against "ill-posed computational conditions" as is explained in section 4.

Schmidt's orthonormalization formula which transforms a set of functions to an orthonormal system plays an important role in our method. However, a formula usually given in a text book takes too long computing time and not practical. We improve the formula so that it carries out a transformation in the least time. The orthonormalization formula named Schmidt should be replaced, not only in this paper but also in a usual text book, by this formula.

In order to show the usefulness of our method, a scattering of a scalar wave by a hollow pipe of a finite length is analysed, and the merits mentioned above are proved numerically.

2. NEW METHOD TO SOLVE A LINEAR FUNCTIONAL EQUATION

$L_2(S)$-space is a set of all functions square integrable on a domain S, where inner product $< f, g >= \int_S f \cdot \overline{g} dS$ and a norm $\|f\|=< f, f >^{1/2}$ are defined for f and g in $L_2(S)$.

We consider a linear functional equation in $L_2(S)$

$$L(\tau) = g \tag{1}$$

where g is a given function, while τ is unknown which is looked for.

To begin with, an algorithm of our method is given as follows.

[Step 1] Take a complete system $\{\tau_n\}$ in $L_2(S)$, and set $u_n=L(\tau_n)$.

[Step 2] Calculate $a_{mn}=< u_n, u_m >$, and, by row operations (4) below, obtain $a_{nn}^{(l)}$, $a_{nl}^{(l)}$ etc., and then orthonormalization coefficients $\{\alpha_{ln}\}$ by (5), which transform the system $\{u_n\}$ to an orthonormal system $\{v_l\}$.

[Step 3] Calculate $\{c_n\}$ by

$$c_n = \Sigma_{m=1}^{n} b_m \Sigma_{l=n}^{N} (-1)^{m+n} \alpha_{ln} \overline{\alpha_{lm}} + \Sigma_{m=n+1}^{N} b_m \Sigma_{l=m}^{N} (-1)^{m+n} \alpha_{ln} \overline{\alpha_{lm}} \tag{2}$$

where $b_m=< g, u_m >$, and $\overline{\alpha}$ is the complex conjugate of α.

[Step 4] Set $\tau(N; c, u)=\Sigma_1^N c_n \tau_n$, then it gives a N-terms approximate solution of (1) which makes the error $E(N)=\|g - \Sigma_1^N c_n u_n\|$ minimum.

Next, we shall prove the validity of the algorithm.

By one of the basic theorems of the authors theory, any element g in $L_2(S)$ is approximated by a linear combination of $\{u_n\}$ as closely as wanted. That is, for arbitrarily given positive constant ϵ, we can find an integer N and a set of constants $\{d_n\}$ so that

$$E(N;d,u) \equiv \|g - \Sigma_1^N d_n u_n\| = \|g - L(\Sigma_1^N d_n \tau_n)\| < \epsilon \qquad (3)$$

holds.

On the other hand, our equation (1) is rewritten as (1)′ $\|g - L(\tau)\|$=0. By a comparison (3) with (1)′, we see that $\Sigma_1^N d_n \tau_n$ is a N-terms approximate solution of (1) with a norm error $E(N;d,u)$. This is our interpretation of an approximate solution. Furthermore, since ϵ in (3) is arbitrary, we may take it as small as we want. Then, $E(N;d,u)$ tends to zero and (3) tends to (1)′. In this sense, $\Sigma_1^N d_n \tau_n$ tends to the true solution τ.

In a practical case, we prescribe an integer N first and then consider a N-terms approximate solution. For a fixed N, $E(N;d,u)$=$\|g - \Sigma_1^N d_n u_n\|$ may vary with $\{d_n\}$, and hence there may exist the best choice of $\{d_n\}$ which makes $E(N;d,u)$ minimum. Steps 3 and 4 of the algorithm show how the best choice $\{c_n\}$ is obtained. This is proved as follows.

Suppose that, by a transformation v_l= $\Sigma_1^l (-1)^{(l+n)} \alpha_{ln} u_n$ which transforms a set $\{u_n\}$ to an orthonormal system $\{v_l\}$, $\Sigma_1^N d_n u_n$ is transformed to, say, $\Sigma_1^N a_l v_l$, while $\Sigma_1^N c_n u_n$ is transformed to $\Sigma_1^N g_l v_l$, where g_l= $< g, v_l >$. The latter transformation is made because the formula (2) for c_n has been obtained by solving $\Sigma_1^N g_l v_l$= $\Sigma_1^N g_l \Sigma_1^l (-1)^{(l+n)} \alpha_{ln} u_n$= $\Sigma_1^N c_n u_n$.

On the other hand, it is easy to see that $E(N,a_n) \equiv \|g - \Sigma_{n=1}^N a_n v_n\|^2 = < g - \Sigma_{n=1}^N a_n v_n, g - \Sigma_{n=1}^N a_n v_n >$= $\|g\|^2$-$\Sigma_{n=1}^N |g_n|^2$+ $\Sigma_{n=1}^N |a_n - g_n|^2$, which takes the minimum value $E(N,g_n) \equiv \|g - \Sigma_{n=1}^N g_n v_n\|^2$= $\|g\|^2$-$\Sigma_{n=1}^N |g_n|^2$ when a_n=g_n. Therefore, we have $\|g - \Sigma_1^N c_n u_n\|$=$\|g - \Sigma_1^N g_l v_l\|$ $\leqq$ $\|g - \Sigma_1^N a_l v_l\|$= $\|g - \Sigma_1^N d_n u_n\|$. That is, (2) gives the solution $\{c_n\}$ which makes the error $\|g - \Sigma_1^N c_n u_n\|$ minimum.

Note that this holds whichever d_n is substituted. That is, any $\{d_n\}$ including those obtained by the moment methods do not necessarily make $E(N : d, u)$ minimum. While, $\{c_n\}$ given by (2) makes $E(N : c, u)$ minimum and hence gives the best approximate solution $\Sigma_1^N c_n \tau_n$.

3. IMPROVED SCHMIDT'S ORTHONORMALIZATION

Schmidt's formula is usually given as v_l=$\phi_l / \|\phi_l\|$, where ϕ_1=u_1 and ϕ_l=u_l-$\Sigma_1^{l-1} < u_l, v_n > v_n$. However, this is very inconvenient for our use, so we rewrite it as follows. v_l=$\Sigma_{n=1}^l (-1)^{l+n} \alpha_{ln} u_n$, α_{ln}=$D_{ln}/\sqrt{D_l D_{l-1}}$, where D_l is a determinant of order l such as D_l=$|a_{mn}|$, $(m,n = 1,\ldots,l)$, while D_{ln} is the minor of an element a_{ln} obtained by omitting the $l-th$ row and the $n-th$ column of D_l. This formula still requires to calculate the values of many determinants. So we employ the following technique to diminish calculation labor to only one determinant.

Set $a_{mn}^{(1)}$=a_{mn}=$< u_n, u_m >$, and let $D_N^{(1)}$=$|a_{mn}^{(1)}|$, $(m,n = 1,\ldots,N)$, be a determinant of order N.

If we apply row operations

$$a_{kn}^{(k+1)} = a_{kn}^{(k)}, \qquad a_{mn}^{(k+1)} = a_{mn}^{(k)} - a_{mk}^{(k)} a_{kn}^{(k)} / a_{kk}^{(k)}, \ (m \neq k) \qquad (4)$$

to $a_{mn}^{(k)}$ in succession for $k = 1,\ldots,l$, we have $D_N^{(l)}$=$|a_{mn}^{(l)}|$, $(m,n = 1,\ldots,N)$, where $a_{mn}^{(l)}$=0 for $1 \leq n \leq l-1$ and $m \neq n$. Especially, D_l=$D_l^{(1)}$=$|a_{mn}^{(1)}|$, $(m,n = 1,\ldots,l)$ is reduced to $D_l^{(l)}$=$|a_{mn}^{(l)}|$ without changing their value. Hence, we have D_l=$a_{11}^{(l)}\ a_{22}^{(l)} \ldots a_{ll}^{(l)}$, which is not zero because $u_1,\ldots,u_l$ are linearly independent and their Grammian D_l is not zero.

Similarly,it is easy to see that D_{ln}=$a_{11}^{(l)} \ldots a_{n-1n-1}^{(l)}\ (-1)^{(l-m+1)} a_{nl}^{(l)} a_{n+1n+1}^{(l)} \ldots a_{l-1l-1}^{(l)}$, where $a_{nl}^{(l)}$ is the nl-element of the determinant $D_{ln}^{(l)}$. Consequently, we have $D_{ln} = \frac{(-1)^{(l-n+1)} D_l a_{nl}^{(l)}}{a_{nn}^{(l)} a_{ll}^{(l)}}$.

On substituting these values in the formula for α_{ln} shown above, we have the very simple formula

$$\alpha_{ln} = \frac{(-1)^{(l-n+1)} a_{nl}^{(l)}}{\sqrt{a_{ll}^{(l)} \cdot a_{nn}^{(l)}}} \tag{5}$$

Note that the number of multiplications and divisions required by the above mentioned Gauss-Jordan type operation to derive (5) is about $N^3/2$.

4. SCATTERING OF A SCALAR WAVE BY A HOLLOW PIPE OF A FINITE LENGTH

Let S be a cylindrical hollow pipe of radius a and length $2h$ [1] which is, by a cylindrical coordinates (r, ϕ, z), $S=\{(a, \phi, \zeta) | 0 \leq \phi \leq 2\pi, -1 \leq \zeta = z/h \leq +1\}$.

A three-dimensional scalar wave u scattered by S is obtained if we solve an integral equation $L(\tau) \equiv \iint_S \Psi(P,Q)\tau(Q)dS_Q = g(P)$, where $\Psi(P,Q)=e^{-iR}/4\pi R$, $R=\overline{PQ}$. This is a linear functional equation which is to be solved by the method mentioned above.

To begin with, we take a complete set in $L_2(S)$ as $\tau_n(P)=T_l(\zeta)\cdot e^{im\phi}$, $(0 \leq l \leq L-1$, $0 \leq m \leq M-1)$, where $T_l(\zeta)=\cos(l cos^{-1}\zeta)$ is the Chebyshev' function of the l-th order. The number n is related to (l, m) by $n=mL+l+1$, and $1 \leq n \leq N = L * M$

Then, we calculate integrals $u_n(P)=L(\tau_n)$ where the integration with respect to ζ is carried out by a Z-point Gauss-Legendre formula, while the integration with respect to ϕ is carried out by a trapezoidal rule with P-subdivisions.

Then, $a_{mn}=< u_n, u_m >$, and $\{\alpha_{mn}\}$ are calculated, and $\{c_n\}$ are obtained by (2) of Step 3, where $b_m=< g, u_m >$ are known if g is given.

Throughout the calculation, we set parameters to be $L=6$, $M=12$, $N=L*M=72$, $Z=9$ and $P=12$, and assume a plane wave incidence $g=-cos\theta\cdot e^{i\lambda(\phi,\zeta)}$, where $\lambda(\phi,\zeta)=asin\phi cos\theta+h\zeta sin\theta$ whose plane of incidence and the angle of incidence are $yz-pl$ and $\pi+\theta$, respectively. Parameters a, h and θ are given at each case as shown below.

Our purpose is to compare the results obtained by our method with those obtained by other method, say, by the Galerkin method, where the only criterion of comparison is a norm error. Hence, we give a list of the norm errors below.

$HE=\|g - \Sigma_1^N c_n u_n\|/\|g\|$ is the norm error of a solution obtained by our method, while $GE=\|g - \Sigma_1^N d_n u_n\|/\|g\|$ is that of Galerkin method. (In Galerkin method, expansion functions and weighting functions are taken to be our complete system $\{\tau_n\}$.)

[I] Case 1; a=4, h=6

$[I-1]$ when $\theta = 0°$

HE=0.166406D − 02=0.825 × GE

GE=0.201773D − 02=1.213 × HE

$[I-2]$ when $\theta = 30°$

HE=0.261265D − 02=0.811 × GE

GE=0.322180D − 02=1.233 × HE

$[I-3]$ when $\theta = 60°$

HE=0.273249D − 02=0.798 × GE

GE=0.342526D − 02=1.254 × HE

[II] Case 2; a=2, h=12

$[II-1]$ when $\theta = 0°$

HE=0.818021D − 02=0.681 × GE

GE=0.120052D − 01=1.468 × HE

$[II-2]$ when $\theta = 30°$

HE=0.929437D − 01=0.605 × GE

GE=0.153762D − 00=1.654 × HE

We explain what we see from this numerical results. To begin with, it is obvious that our results HE are always smaller than GE. That is, our results are the best.

[1] A quantity with dimension of length, say, r, is normalized by denoting kr simply as r, where k is a wave number.

We mean by an "ill-posed condition" a computational environment or setting which makes an error increase. Here, we have two kind of ill-posedness. The first one of which comes from an increase of an angle of incidence θ , while the other comes from a mismatch of parameters L, M, Z, P, a and h.

First, we consider an ill-posedness caused by an increase of θ.

In a N-terms approximation, a function g is approximated by N functions $\{\tau_n\}_{n=1,\cdots,N}$. When θ increases, a boundary data g may become more complex, and more functions τ_n may be required to keep a same degree of accuracy of approximation. In other words, if N is fixed, errors HE as well as GE may increase with an increase of θ. This fact is obviously true from the above list. However, what is important is to note that the ratio HE/GE decreases with the increase of θ. This means that, though both of HE and GE increase with the increase of θ, GE increases more rapidly than HE, or, HE is more stable than GE against the ill-posed condition caused by the increase of θ. This holds in Case I as well as in Case II.

Next, by the comparison of Case I with Case II, we shall examine an ill-posedness caused by a mismatch of parameters.

In a N-terms approximation by $\{\tau_n\}_{n=1,\cdots,N}$, L functions $\{T_l(\zeta)\}_{l=0,\cdots,L-1}$ are employed in a range Ξ=$\{-h \leqq z = h\zeta \leqq +h\}$ whose length is $2h$, and M functions $\{e^{im\phi}\}_{m=0,\cdots,M-1}$ are employed in a range Φ=$\{0 \leqq a\phi \leqq 2\pi a\}$ whose length is $2\pi a$. While, Z points of integration are taken in the range Ξ, and P points of integration are taken in the range Φ.

This means that, in Case I, we take six functions and nine points of integration in the range Ξ whose length is twelve, and twelve functions and twelve points of integration in the range Φ whose length is about twenty five.

Contrary to this, in Case II, six functions and nine points of integration are employed in the wide range Ξ whose length is twenty four, while twelve functions and twelve points of integration are employed in the narrow range Φ whose length is twelve.

That is, in Case I, there exists a kind of harmony between the length of a range and the numbers of functions and points of integration employed there. While, in Case II, there exists a mismatch between the length of a range and the numbers of functions and points of integration employed there.

As a consequence of the mismatch of parameters, as is obviously seen from the above list, errors in Case II are far larger than the corresponding errors in Case I.

However, what is important is to note that the ratios HE/GE in Case II are considerably small in comparison with the corresponding ones in Case I. This implies that, though errors HE and GE increase by ill-posedness caused by mismatch of parameters, GE increases more rapidly than HE, or, HE is more stable than GE against the ill-posedness caused by the mismatch of parameters.

5. CONCLUSION

In this paper, the traditional Schmidt's formula was improved and a practically useful formula (5) was derived. It is apparent that an expansion of a function is the best when an orthonormal system of functions is employed as an expanding functions. Therefore, our method to solve a linear functional equation, which was based on our orthonormalization formula, yields always the best solution with minimum error, and is the most stable against ill-posed computational conditions.

A more detailed version of this paper is expected to be published in some journal soon.