ERROR ESTIMATES FOR SOME VARIATIONAL METHODS APPLICABLE TO SCATTERING AND RADIATION PROBLEMS*

BY

M. SZAŁEK

Institute for Fundamental Problems of Technology, Polish Academy of Sciences, Warsaw, Poland

Abstract. We consider variational expressions helpful in calculating the approximate value of a scalar product, in Hilbert space, of an arbitrary vector g with a solution u^0 of an arbitrary inhomogeneous linear equation. Error bounds for this approximate value are given. For the case where an approximate solution of an inhomogeneous equation is sought in an arbitrary subspace of a space containing u^0, conditions are specified for a best estimate of the error by the use of two trial vectors. A method is presented for an additional improvement of the error estimate by using four trial vectors.

1. Introduction. In the problems of scattering or radiation we are usually interested in finding out the values of only a few functionals, depending on the solutions of the equations pertinent to the problem. For example, in the problem of scattering on a waveguide junction the quantities of interest are, in most cases, only the amplitudes of the propagating modes, but not the amplitudes of the evanescent modes. When calculating the approximate values of such functionals it may be useful to apply variational methods.

To state the problem, consider an equation

$$k = Du^0, \tag{1}$$

where $u^0 \in \mathrm{H}_1$, $k \in \mathrm{H}_2$, $D \in \mathrm{M}$, H_1 and H_2 are Hilbert spaces, and M is the set of linear operators $(\mathrm{H}_1 \to \mathrm{H}_2)$. For $g \in \mathrm{H}_1$, we want to find the value of the scalar product

$$\langle u^0, g\rangle. \tag{2}$$

That product is assumed to satisfy the conditions for a scalar product in H spaces.

An approximate value of (2) can be found by using some functionals, such that their stationary values are equal to $\langle u^0, g\rangle$. Let D^+ denote the adjoint of D, so that for any $w \in \mathrm{H}_2$, $u \in \mathrm{H}_1$ the equality $\langle Du, w\rangle = \langle u, D^+w\rangle$ holds. In general, when $D \neq D^+$, as happens for a waveguide junction in the presence of anisotropic media, or where $k \neq \text{const} \cdot g$, as is the case if the waveguides at the junction are of different cross-section, the above functionals depend on two trial vectors, u and w. We shall consider two simplest functionals of that kind:

$$\mathrm{R}_1[u, w] = \langle u, g\rangle + \langle k, w\rangle - \langle u, D^+w\rangle, \tag{3}$$

$$\mathrm{R}_1'[u, w] = \frac{\langle u, g\rangle\langle k, w\rangle}{\langle u, D^+w\rangle}. \tag{4}$$

Functional (3) appears in [1], [2], [3] and [4]. Functional (4) was used in [5] and is also equivalent to an expression discussed in [6] in connection with eigenvalue problems. For

* Received January 18, 1967; revised version received January 21, 1969.

$D = D^+$ and $g = k$ it reduces to the homogeneous Schwinger's expression

$$\frac{\langle u, k\rangle\langle k, u\rangle}{\langle u, Du\rangle}, \tag{5}$$

introduced, e.g., in [7] and [8].

In this paper the error bounds for the expressions (3) and (4) are determined (Theorems 2.1 and 2.2). The error estimate for (3) turns out to be better than for (4). These estimates are meaningless when the parameter m_0, as defined in (6), equals zero, and they may be of little advantage if m_0 differs only slightly from zero. It is shown in Sec. 4, however, that in such cases one can often obtain a sufficiently large m_0 by transforming Eq. (1).

For the case where an approximate solution is sought in an arbitrary subspace $H_3 \subset H_1$, conditions are specified to guarantee a best error estimate (Theorem 2.3). Theorem 3.1 allows for an additional improvement by introducing four trial vectors.

2. Error bounds. For any $g, u \in H_1$, $k, w \in H_2$, let

$$\begin{aligned} |u| &= [\langle u, u\rangle]^{1/2}, \qquad |w| = [\langle w, w\rangle]^{1/2}, \\ R_2[u] &= \langle k - Du, k - Du\rangle = |k - Du|^2, \\ R_3[w] &= \langle g - D^+w, g - D^+w\rangle = |g - D^+w|^2, \\ m_1 &= \inf_{|u|=1} |Du|, \qquad m_2 = \inf_{|w|=1} |D^+w|, \qquad m_0 = \max(m_1, m_2), \end{aligned} \tag{6}$$

where $\langle z, x\rangle$, for $z, x \in H_1$ or $z, x \in H_2$, is the scalar product in H_1 or H_2 respectively. We notice that if for every $z \in H_2$ there exists one and only one $x \in H_1$ such that $z = Dx$, then $m_1 = m_2$, because for $m_1 > 0$ or $m_2 > 0$ they are the inverses of the norms of the adjoint operators D^{-1} and $[D^{-1}]^+ = [D^+]^{-1}$.

The following theorems can be stated:

THEOREM 2.1. *Let $D \in M$, $g, u \in H_1$, $k, w \in H_2$, $0 < m \leq m_0$. If there exists $u^0 \in H_1$ satisfying Eq. (1), and $w^0 \in H_2$ satisfying the equation*

$$g = D^+w^0, \tag{7}$$

then

$$|R_1[u, w] - \langle u^0, g\rangle| \leq m^{-1} |R_2[u]R_3[w]|^{1/2} = A[u, w] \tag{8}$$

where $A[u, w] = B\,|u - u^0|\,|w - w^0|$, and $B \geq 0$ depends only on the directions of the vectors $u - u^0$ and $w - w^0$.

Proof. Substituting $u = u^0 + x$, $w = w^0 + z$ into R_1, R_2, R_3 and taking into account (1) and (7), we obtain

$$R_1[u, w] = \langle u^0, g\rangle - \langle x, D^+z\rangle, \tag{9}$$

$$R_2[u] = |Dx|^2, \tag{10}$$

$$R_3[w] = |D^+z|^2. \tag{11}$$

From (6)

$$|x| \leq m_1^{-1}\,|Dx|, \qquad |z| \leq m_2^{-1}\,|D^+z|.$$

Consequently, by Schwarz' inequality, and from Eqs. (10) and (11), we have

$$|\langle x, D^+z\rangle| = |\langle Dx, z\rangle| \leq m^{-1}\,|Dx|\,|D^+z| = m^{-1}\{R_2[u]R_3[w]\}^{1/2}.$$

This relation, along with (9), proves (8).

REMARK. Since we know $Dx = Du - k$ and $D^+z = D^+w - g$, we can sometimes improve the estimate of $|x|$ or $|z|$ as compared to that obtained in the proof of Theorem 2.1. We can replace there m_1 and m_2 by m_1' and m_2', respectively, where $m_1' > m_1$ or $m_2' > m_2$. It can then easily be seen that m in (8) can be replaced by $m_0' = \max(m_1, m_1', m_2, m_2')$, which may improve the estimate.

In connection with Theorem 2.1 it might be pointed out for comparison that a typical value of a linear estimate of the error in the scalar product (2), as obtained by estimating the error in the approximate solution of the linear equation (1), is $m_1^{-1}\{R_2[u]\}^{1/2}\,|g|$. One can readily see the clear advantage of estimate (8), which is of second order with respect to the errors in u and w.

THEOREM 2.2. *If* $m_1m_2 > 0$, $\langle u, D^+w\rangle \neq 0$ *then Theorem* 2.1 *remains true when inequality* (8) *is replaced by*

$$|R_1'[u, w] - \langle u^0, g\rangle| \leq |\langle u, D^+w\rangle|^{-1}\,[(m_1m_2)^{-1}\,|k|\,|g| + m^{-1}\,|\langle u^0, g\rangle|]\,|R_2[u]R_3[w]|^{1/2}, \tag{12}$$

where $R_1'[u, w]$ *is given by* (4).

This can be proved in analogy to Theorem 2.1, taking into account that

$$\frac{\langle u, g\rangle\langle k, w\rangle}{\langle u, D^+w\rangle} = \langle u^0, g\rangle + \frac{\langle k, z\rangle\langle x, g\rangle - \langle u^0, g\rangle\langle x, D^+z\rangle}{\langle u, D^+w\rangle}.$$

Using the relation $\langle k, z\rangle\langle x, g\rangle = \langle u^0, D^+z\rangle\langle Dx, w^0\rangle$, we can also write

$$|R_1'[u, w] - \langle u^0, g\rangle| \leq |\langle u, D^+w\rangle|^{-1}\,[|u^0|\,|w^0| + m^{-1}\,|\langle u^0, g\rangle|]\,|R_2[u]R_3[w]|^{1/2}. \tag{13}$$

Having in mind that $\langle u^0, g\rangle \cong \langle u, D^+w\rangle$ for u, w approximating closely enough u^0 and w^0, we see that estimates (12) and (13) are worse than (8).

Consider now an operator P_1 projecting every $u \in H_1$ onto a subspace $H_3 \subset H_1$: $P_1u \in H_3$; $P_1P_1 = P_1$; $P_1^+ = P_1$. In the case where D is a $N \times N$ matrix, P_1 can be defined, e.g., as

$$P = (P_{ir}) = (\delta_{ir}) \qquad r \leq n_0$$
$$= (0) \qquad n_0 < r \leq N,$$

where δ_{ir} is the Kronecker delta. Hence an equation $P_1k = P_1DP_1u$, for example, is an approximating set of n_0 algebraic equations obtained from the set of N equations by rejecting $N - n_0$ equations and setting $N - n_0$ unknowns equal to zero.

Similarly, let P_2 be such that for $w \in H_2$, $H_4 \subset H_2$ the relations $P_2w \in H_4$; $P_2P_2 = P_2$; $P_2^+ = P_2$ hold. We can then formulate

THEOREM 2.3. *Let* $D \in M$, $g \in H_1$, $k \in H_2$, $u' \in H_3$, $w' \in H_4$. *If for a certain* $u'^0 \in H_3$

$$P_1D^+k = P_1D^+DP_1u'^0, \tag{14}$$

then $R_2[u']$ *reaches its least value for* $u' = u'^0$.

Similarly, if for a certain $w'^0 \in H_4$

$$P_2Dg = P_2DD^+P_2w'^0, \tag{15}$$

then $R_3[w']$ *reaches its least value for* $w' = w'^0$.

Proof. Substituting $u' = u'^0 + x' (x' \in H_3)$ into $R_2[u']$ we get $R_2[u'] = |k - Du'^0|^2 + |Dx'|^2$, which proves the first part of Theorem 2.3. The second part can be proved by substituting D^+, g, w, P_2, z for D, k, u, P_1, x, respectively.

Eqs. (14) and (15) are, of course, identical with the equations of the method of least squares, as applied to Eqs. (1) and (7), (see e.g. [9], [10]).

From Theorem 2.3 it follows that estimate (8) is best for $u \in H_3$, $w \in H_4$ if $u = u'^0$, $w = w'^0$.

3. Additional estimates. There exists a way of improving estimate (8) if in place of the two trial vectors u and w we consider four vectors. This results from the following theorem in which the spaces H_1 and H_2 are for simplicity assumed real; this is not an essential restriction since with complex H_1 and H_2 the problem can always be reduced to that with real H_1 and H_2.

THEOREM 3.1. *Let* H_1, H_2 *be real,* $D \in M$; $g, u^0 \in H_1$; $k, w^0 \in H_2$; $u'^0, x' \in H_3$; $z', w'^0 \in H_4$. *If* u'^0 *and* w'^0 *satisfy* (14) *and* (15), *respectively, then*

$$\max_{x', z'} (a_1 - a_2) \leq \langle u^0, g \rangle \leq \min_{x', z'} (a_1 + a_2), \tag{16}$$

where

$$a_1 = R_1[u'^0, w'^0] + \langle x', g - D^+w'^0 \rangle + \langle k - Du'^0, z' \rangle - \langle x', D^+z' \rangle,$$

$$a_2 = m^{-1}\{R_2[u'^0] + \langle x', D^+Dx' \rangle\}^{1/2}\{R_3[w'^0] + \langle z', DD^+z' \rangle\}^{1/2}.$$

Proof. Let $u = u'^0 + x'$, $w = w'^0 + z'$. Taking into account relations (14) and (15) we get $a_1 = R_1[u, w]$, $a_2 = m^{-1} |R_2[u]R_3[w]|^{1/2}$. Hence, inequality (8) can be written in the form

$$a_1 - a_2 \leq \langle u^0, g \rangle \leq a_1 + a_2 . \tag{16'}$$

This is valid for every $x' \in H_3$, $z' \in H_4$, which proves (16).

For $x' = z' = 0$ inequality (16') goes into (8) with $u = u'^0$, $w = w'^0$. Inequality (16) gives, therefore, an estimate not worse than (8) with $u \in H_3$, $w \in H_4$.

The exact values of $\max (a_1 - a_2)$ and $\min (a_1 + a_2)$ apparently are not worth searching for in view of mathematical difficulties. The method of steepest descent can be applied here. This corresponds to setting $z' = d_1P_2y_1$; $x' = d_2P_1y_2$, where d_1, d_2 are real numbers, $y_1 = k - Du'^0$, $y_2 = g - D^+w'^0$. One can also put $z' = d_1'y_1'$, $x' = d_2'y_2'$, where $y_1' \in H_4$ and $y_2' \in H_3$ are the eigenvectors belonging to the lowest eigenvalues of the operators $P_2DD^+P_2$ and $P_1D^+DP_1$, respectively. Such a choice will guarantee the slowest increase of a_2 with the increase of $|d_1'|$ and $|d_2'|$. Combining both the substitutions we can put $z' = d_1P_2y_1 + d_1'P_2y_1'$; $x' = d_2P_1y_2 + d_2'P_1y_2'$.

Applying the method of steepest descent we obtain from (16) the inequality

$$\max_{d_1, d_2} (a_1 - a_2) \leq \langle u^0, g \rangle \leq \min_{d_1, d_2} (a_1 + a_2), \tag{17}$$

where

$$a_1 = R_1[u'^0, w'^0] + d_1 |P_2 y_1|^2 + d_2 |P_1 y_2|^2 - d_1 d_2 \langle y_2 , P_1 D^+ P_2 y_1 \rangle,$$

$$a_2 = m^{-1} \{ |y_1|^2 + d_2^2 |DP_1 y_2|^2 \}^{1/2} \{ |y_2|^2 + d_1^2 |D^+ P_2 y_1|^2 \}^{1/2},$$

because $R_2[u'^0] = |y_1|^2$, $R_3[w'^0] = |y_2|^2$. Knowing $D_1 = P_1 D^+ D P_1$ and $D_2 = P_2 D D^+ P_2$, needed in (14) and (15), we can express here $|DP_1 y_2|^2 = \langle y_2 , P_1 D_1 P_1 y_2 \rangle$; $|D^+ P_2 y_1|^2 = \langle y_1 , P_2 D_2 P_2 y_1 \rangle$.

The functions a_1 and a_2 depend on two variables, d_1 and d_2. For such functions it is relatively easy to calculate their approximate extreme values. Apart from that, to determine the coefficients defining a_1 and a_2 it is sufficient to calculate some products of "small" (n_0-dimensional) vectors with matrices, since all time-consuming calculations have been already performed in connection with Eqs. (14), (15) and (8). It therefore seems reasonable to use estimate (17), even if there is only a relatively small chance that this might essentially improve the error estimate.

So far, this author knows of only one case where estimate (17) has been applied; this was in numerical calculations concerning scattering on a waveguide junction in the presence of an anisotropic medium. In that case no noticeable improvement in the error estimate was obtained despite the fact that the necessary conditions, as given below, for estimate (17) to be useful were satisfied.

In order to find the above-mentioned conditions consider a quantity a_3 defined as the ratio of the shift in the upper bound of the error to the distance between the bounds

$$a_3 = [(a_1 + a_2)\,|_{d_1 = d_2 = 0} - \min_{d_1, d_2} (a_1 + a_2)][2a_2\,|_{d_1 = d_2 = 0}]^{-1} \tag{18}$$

$$= \tfrac{1}{2} \max_{d_1, d_2} [-d_1 |P_2 y_1|^2 - d_2 |P_1 y_2|^2 + d_1 d_2 \langle y_2 , P_1 D^+ P_2 y_1 \rangle + m^{-1} |y_1|\,|y_2| - a_2] \frac{m}{|y_1|\,|y_2|}.$$

This quantity can serve as a measure of the achieved improvement in the error estimate. The inequality $0 \leq a_3 \leq 1$ must be satisfied. The larger a_3, the greater the improvement in the estimate. We always have $|y_1|^2 \geq |P_2 y_1|^2$, $|y_2|^2 \geq |P_1 y_2|^2$, $|DP_1 y_2| \geq m_1 |P_1 y_2|$, $|D^+ P_2 y_1| \geq m_2 |P_2 y_1|$.

Taking for example $|y_1| = |y_2|$, $b\,|y_1| = |P_2 y_1| = |P_1 y_2|$, $m_0 = m$, $|DP_1 y_2| = |D^+ P_2 y_1| = nm\,|y_1|$, $\langle y_2 , P_1 D^+ P_2 y_1 \rangle = 0$, $d_1 = d_2$, $n \geq b$, $0 \leq b \leq 1$, we obtain

$$a_3 = \frac{m}{2} \max_{d_1} [-2d_1 b^2 - d_1^2 n^2 m].$$

The maximum is reached for $d_1 = -b^2/n^2 m$; then $a_3 = b^4/2n^2$. Hence the outlined method based on Theorem 3.1 can be sensibly applied only when n can be small, e.g. $b \leq n \leq 2$, and, simultaneously, $b \cong 1$. This can be the case only when $m_1''/m_0 < 2,5$ and $m_2''/m_0 < 2,5$, where

$$m_1'' = \inf_{|P_1 u| = 1} |DP_1 u| \qquad m_2'' = \inf_{|P_2 w| = 1} |D^+ P_2 w|$$

with $u \in H_1$, $w \in H_2$. These conditions remain also valid in a more general case, in particular, in every case where $\langle y_2 , P_1 D^+ P_2 y_1 \rangle = 0$.

4. Remarks concerning m_0 and $A[u, w]$. In connection with the estimates given it is useful to examine how transformations of Eq. (1), with $m_0 = 0$ or m_0 approaching zero, may influence the values of m_0 and $A[u, w]$ in (8).

Suppose that Eq. (1) is equivalent to an infinite set of algebraic equations. Usually

this set can be approximated, with an accuracy sufficient for the physical problem under consideration, by a finite set of N algebraic equations with $m_0 \neq 0$. This procedure is analogous to that where Eq. (1) is transformed by multiplying it by such a linear operator L that for the operator LD the value of m_0 is positive. To obtain a suitably large m_0 we can also apply these two procedures successively. Let $A_1[u] = m_0^{-1}\{R_2[u]\}^{1/2}$. Putting for simplicity $m = m_0$, we have $A[u, w] = A_1[u]\{R_3[w]\}^{1/2}$. In Theorem 4.1 we shall specify the operators L for which $A_1[u]$ reaches a minimum.

Suppose that D can be written in a matrix representation as a finite N × N matrix (D_{ir}), with $m_0 \neq 0$. Let $H_1 = H_2$, $\det(D) \neq 0$, $\det(L) \neq 0$. We have in place of Eq. (1) an equation $Lk = LDu^0$. Thus $R_2[L, u] = |L(k - Du)|^2$, $m_0(L) = \min_{|u|=1} |LDu|$, $A_1[L, u] = \{m_0(L)\}^{-1}\{R_2[L, u]\}^{1/2}$. Let the components of an arbitrary vector $x \in H_1$ be $x_i (i = 1, 2, \cdots, N)$, and $\langle x, t\rangle = \sum_{i=1}^{N} x_i^* t_i$ for any $x, t \in H_1$ where x_i^* denotes the complex conjugate of x_i. We can always adopt such a definition of a scalar product, defining accordingly the vector g. Let $u^0, u, x \in H_1$; $u = u^0 + x$.

THEOREM 4.1. *If* $\det(D) \neq 0$, $\det(L) \neq 0$, $H_1 = H_2$, *then for every* $u \in H_1$ *the quantity* $A_1[L, u]$ *assumes a smallest value for* $LD = cS$ *where* c *is any number different from zero and* S *is any unitary matrix.*

Proof.

$$A_1[L, u] = [m_0(L)]^{-1}[\langle x, [LD]^+ LDx\rangle]^{1/2}.$$

Since $L_1 = [LD]^+ LD$ is a Hermitian positive definite matrix we can pass by means of a unitary transformation S_1 to a coordinate system where L_1 is diagonal and can be written $L_1 = (a_i^2\, \delta_{ir})$, where $a_i > 0$. In this new coordinate system the definition of the scalar product remains unchanged and $m_0(L) = \min_i a_i$. Hence

$$A_1[L, u] = \left\{\sum_{i=1}^{N} |x_i|^2 a_i^2\right\}^{1/2} \frac{1}{\min\limits_i a_i}.$$

Thus $A_1[L, u]$ assumes a smallest value for $a_i = \text{const}$ $(i = 1, 2, \cdots, N)$. This is so if $LD = cS$, e.g. if $L = D^{-1}$.

By making N tend to infinity, Theorem 4.1 can be generalized to the case of infinite matrices, even with $\det(D) = 0$, provided that to every $z \in H_2$ there corresponds only one $x \in H_1$ such that $z = Dx$.

Regarding $R_3[w]$ the following facts can be noted. In place of Eq. (7) we have $g = D^+L^+w^0(L)$. Hence $R_3[L, w(L)] = |g - D^+L^+w(L)|^2$. We can expect that the closer LD approaches cS, the smaller $R_3[L, w(L)]$, because it is then easier to find a good approximation to the solution of the equation $LDg = LD[LD]^+w^0 \cong |c|^2 w^0$.

We can, therefore, in general say that the closer LD approaches cS, that is the less the norm of LD differs from $m_0(L)$, the smaller $A[u, w]$. We can also notice that, in accordance with the inequalities given at the end of Sec. 3, most favorable conditions for applying formulas (16) and (17) exist when $LD \cong cS$, as in that case the norm of LD differs little from $m_0(L)$.

If D does not differ much from a diagonal matrix we can, in particular, set $L = (c_i\, \delta_{ir})$ with suitably chosen numbers c_i. Assuming L in that form may also turn out useful in other cases.

In connection with our estimates the question arises of determining the number m_0.

This question will not be discussed here. It may, however, be noticed that even the knowledge of a very rough approximation of m_0 will be useful. In addition, if we want to find the values of m_0 for many operators D or LD not much different one from another, it is sufficient to find them for some of these operators and apply to the rest the estimates connecting the numbers m_0 of two operators and the norm of their difference (cf. e.g. [11]). If the operator LD is given in a matrix form and that matrix is not much different from diagonal, we can also apply other estimates such as the Hadamard estimate [11].

References

[1] P. Roussopoulos, *Méthodes variationnelles en théorie des collisions*, C. R. Acad. Sci. Paris **236**, 1858–1860 (1953).

[2] M. Becker, *The principles and application of variational methods*, Research monograph No. 27, MIT Press, Cambridge, Massachusetts.

[3] G. Pomraning, *A variational principle for linear systems*, J. Soc. Indust. Appl. Math. **13**, 511 (1965).

[4] W. W. Nikol'skij, *Variacionnye metody dlja vnutrennych zadač élektrodynamiki* (*Variational methods for electromagnetic boundary-value problems*), "Nauka", Moskva 1967, p. 25.

[5] L. A. Vajnštejn, *Volny toka v tonkom cil'indričeskom provodnike, III. Variacionnyj metod* (*Waves of current in a thin cylindrical conductor, III. Variational method*), J. Tech. Phys. **31**, 29–44 (1961).

[6] L. Cairo and T. Kahan, *Techniques variationnelles en radioélectricité*, Dunod, Paris 1962, p. 54.

[7] F. E. Borgnis and C. H. Papas, *Randwertprobleme der Mikrowellenphysik*, Springer-Verlag, Berlin 1955.

[8] R. E. Collin, *Field theory of guided waves*, McGraw-Hill, New York, 1960, p. 314–363.

[9] S. G. Michlin, *Variacionnye metody v matematičeskoj fizike* (*Variational methods in mathematical physics*), GITTL, Moskva 1957, pp. 410–417.

[10] M. Altman, *Metody przybliżone analizy funkcjonalnej* (*Approximate methods of functional analysis*), Zakład Narodowy im. Ossolińskich, Wydawnictwo PAN, Wrocław-Warszawa-Kraków 1963, pp. 35–38.

[11] E. Bodewig, *Matrix calculus*, North-Holland, Amsterdam, 1956, pp. 57–69.