Short Papers

On the Performance of the Least Squares Method for Waveguide Junctions and Discontinuities

R. JANSEN

Abstract—**The high computational expenditure of the least squares boundary residual method restricts its application to certain problems. It is therefore necessary to utilize the inherent simplicity of special cases together with convergence-optimization criteria in order to reduce computational time and storage requirements. The problem of the coaxial-to-circular waveguide junction is presented as an illustrative example of how this is performed. A selection criterion is also suggested to determine the optimum weighting factor.**

INTRODUCTION

Attention has lately been directed to the solution of boundary-value problems by the method of least squares [1], [2], an essentially variational method which was first presented by Picone in 1928 [3]. This method is useful because of its universal applicability, its ability to produce upper and lower bounds, and its property of ensured (nonrelative) convergence [1]. It allows bounds to be generated simply by means of varying a weighting factor [1], and in this compares favorably with the direct application of the calculus of variations, a technique which requires separate formulations for each bound [6], [7]. On the other hand, convergence is slower than that obtained from the Ritz method [3], [4], if one does not take care of singularities explicitly, and usually large complex matrices must be dealt with. Therefore, as long as an *a priori* determination of the optimum-convergence weighting factor does not exist, it is of essential importance with least squares methods for the program and the formulation to utilize the simplifications possible in special cases.

Such special situations prevail, for example, in the following cases: the planar infinitely thin discontinuity in an otherwise homogeneous waveguide, structures which exhibit special symmetries, and waveguide junctions with pure boundary reduction or enlargement. The coaxial-to-circular waveguide junction belongs to the latter class and will now serve to illustrate the time and storage saving performance of the method.

PERFORMANCE OF THE METHOD

Consider the waveguide junction in Fig. 1 together with the following conventional truncated modal expansions for the transverse field at $z = 0$:

$$E_{tP}^{\mathrm{I}} = t_0^{\mathrm{I}}(1 + r_0) + \sum_{i=1}^{P} t_i^{\mathrm{I}} A_i^- \qquad \text{in region I}$$

$$H_{tP}^{\mathrm{I}} = (1/Z_{F0}^{\mathrm{I}})(e_z \times t_0^{\mathrm{I}})(1 - r_0) - \sum_{i=1}^{P} (1/Z_{Fi}^{\mathrm{I}})(e_z \times t_i^{\mathrm{I}}) A_i^-$$

$$E_{tN}^{\mathrm{II}} = \sum_{k=1}^{N} t_k^{\mathrm{II}}(1 + r_k) B_k^+ \qquad \text{in region II}$$

$$H_{tN}^{\mathrm{II}} = \sum_{k=1}^{N} (1/Z_{Fk}^{\mathrm{II}})(e_z \times t_k^{\mathrm{II}})(1 - r_k) B_k^+. \qquad (1)$$

Manuscript received April 22, 1974; revised January 14, 1975.

The author is with the Institute für Hochfrequenztechnik, 51 Aachen, Templergraben 55, Germany.

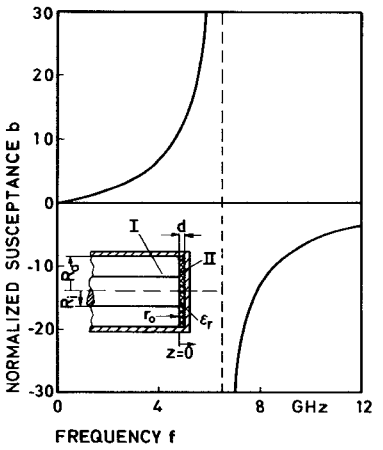

Fig. 1. Normalized susceptance b versus frequency. Dimensions are: $d = 1$ mm, $R_i = 3$ mm, $R_a = 7$ mm, relative dielectric constant $\epsilon_r = 9$.

Here $t_i^{\mathrm{I}}, t_k^{\mathrm{II}}$ are the real orthonormalized transverse-vector mode functions (t_0^{I} for the TEM mode which is the excitation with amplitude $A_0^+ = 1$; $A_0^- = r_0$); A_i^- and B_k^+ denote the corresponding complex amplitudes; $r_k = B_k^-/B_k^+ = -\exp(-2\gamma_k \cdot d)$ are the mode reflection coefficients; Z_{Fi}^{I} and Z_{Fk}^{II} represent wave impedances, and e_z is the z-directed unit vector. Because of the rotational symmetry of the structure and the excitation field only TM_{0n} modes have to be taken into account.

For the calculation of r_0 or the normalized input admittance $y = jb = (1 - r_0)/(1 + r_0)$, the matrix M_{P+N} which corresponds to the positive Hermitian form

$$F_{P+N} = \int_0^{2\pi}\int_0^{R_i} |E_{tN}^{\mathrm{II}}|^2 r\, dr\, d\varphi + \int_0^{2\pi}\int_{R_i}^{R_a} (|E_{tP}^{\mathrm{I}} - E_{tN}^{\mathrm{II}}|^2 + Z^2 \cdot |H_{tP}^{\mathrm{I}} - H_{tN}^{\mathrm{II}}|^2) r\, dr\, d\varphi \qquad (2)$$

has to be constructed [1], [2]. The quantity Z in (2) is a convergence-optimizing electric-to-magnetic weighting factor.

In addition to reducing the set of mode functions by virtue of structural symmetry, it is possible to save computer time by making use of the fact that the fields in (1) have been split into real orthonormal vector functions and complex scalars. This is performed by first generating a real symmetric auxiliary matrix G with elements

$$G_{kl} = \int_0^{2\pi}\int_{R_i}^{R_a} t_k^m t_l^n r\, dr\, d\varphi; \qquad m = \mathrm{I,II} \text{ and } n = \mathrm{I,II} \qquad (3)$$

containing the complete information about the geometry of the problem, whereas the frequency-dependent information is included in the complex values Z_{Fi}^{I}, Z_{Fk}^{II}, and r_k which are stored in linear arrays.

The splitting procedure indicated above applies for least squares methods in general. For the special case under consideration, there is an additional advantage since G contains the $(P + 1) \times (P + 1)$ identity matrix as a submatrix, indicating that $(P + 1)$ of the $(P + 1 + N)$ unknown amplitudes can be eliminated. This is a consequence of the pure boundary enlargement that prevails in the problem. This property of the matrix G is transferred to the positive definite Hermitian matrix C_{P+N} (M_{P+N} with its first row and column omitted) insofar as C_{P+N} possesses a $(P + 1) \times (P + 1)$ diagonal submatrix D. With reference to [1], the set of amplitudes which

minimizes F_{P+N} could be calculated from the system of linear equations

$$C_{P+N}\cdot\begin{bmatrix} r_0 \\ A_i^- \\ B_k^+ \end{bmatrix} = \begin{bmatrix} D & A \\ A^{*T} & B \end{bmatrix}\cdot\begin{bmatrix} r_0 \\ A_i^- \\ B_k^+ \end{bmatrix} = \begin{bmatrix} v_0 \\ 0 \\ v_k \end{bmatrix} = v_{P+N} \qquad (4)$$

with the excitation vector v_{P+N} extracted from the first column of M_{P+N}. However, elimination of the A_i^- leads to the equivalent system

$$C_N\cdot[B_k^+] = [B - A^{*T}\cdot D^{-1}\cdot A]\cdot[B_k^+] = [v_k] - [A^{*T}\cdot D^{-1}]\cdot\begin{bmatrix} v_0 \\ 0 \end{bmatrix}$$

$$r_0 = (1/D_{11})\,(v_0 - \sum_{k=1}^{N} A_{1k}B_k^+) \qquad (5)$$

which is governed by the likewise positive definite Hermitian matrix C_N of reduced order $N \times N$. Hence the problem may be viewed as that of minimizing a modified functional F_N with corresponding matrices M_N and C_N, respectively. As a result, in its solution there is considerable saving of storage and running time.

SELECTION CRITERION FOR THE WEIGHTING FACTOR Z

Let N denote the order of the matrix C_N, and E_{tN},H_{tN}, the corresponding approximate field solutions. E_t,H_t is the exact solution of the boundary value problem considered. Since for any finite Z the method ensures convergence of the fields in the mean [1], [3], [4], there exists an integer number N_0 such that

$$0 \leq |\,\|E_t\| - \|E_{tN}\|\,| \leq \|E_t - E_{tN}\| \leq \epsilon_E(Z)$$

and

$$0 \leq |\,\|H_t\| - \|H_{tN}\|\,| \leq \|H_t - H_{tN}\| \leq \epsilon_H(Z) \qquad (6)$$

is satisfied for any $N \geq N_0$ and a given range of Z, with the norms defined in accordance with (2). In addition, as the boundary operator representing the problem is defined on the complete set of functions of finite norm t_i^{I},t_k^{II}, the norms of E_t,H_t must also be bounded [4] so that the edge condition is not violated as indicated by Davies in [1].

Convergence of E_{tN} as defined in (6) implies that scalar multiplication of $E_{tN}{}^{\mathrm{I}}$ by $t_0{}^{\mathrm{I}}$ and term-by-term integration of the series in (1) are allowed even if the upper summation limit goes to infinity [4]. Together with the Cauchy–Bunyakovsky inequality [4] this yields

$$|\,r_0 - r_{0N}\,| = \left|\int_{\text{region I}} (E_t{}^{\mathrm{I}} - E_{tN}{}^{\mathrm{I}})t_0{}^{\mathrm{I}}\,dF\right| \leq \|E_t{}^{\mathrm{I}} - E_{tN}{}^{\mathrm{I}}\| \leq \epsilon_E(Z). \qquad (7)$$

With $jb = (1 - r_0)/(1 + r_0)$ it can be concluded that the convergence of the approximate susceptance b_N to its exact value b for any finite range of Z exists [uniform convergence of $b_N(Z)$], and the convergence rate of b_N is in some way governed by that of the field $E_{tN}{}^{\mathrm{I}}$.

The slope of the curves b_N versus matrix size N is defined as

$$\frac{db_N(Z)}{dN} = b_N(Z) - b_{N+1}(Z). \qquad (8)$$

Hence it follows from Cauchy's criterion for the existence of a limit [5], that with

$$\left|\frac{db_N(Z)}{dN}\right| = |\,b_N - b_{N+1}\,| \leq 2\cdot|\,b_N - b\,| \leq 2\cdot\epsilon_b(Z) \qquad (9)$$

a minimization of the modulus of the slope with respect to the weighting factor is necessary in order for the error in b_N to be small. This minimum exists because b_N with Z is stationary, in agreement with the remarks of Oraizi and Perini [2].

Consider then a range of weighting factors $Z_l \leq Z \leq Z_u$ including the optimum $Z = Z_{\mathrm{opt}}$ and an integer $N = N_0$ which leads to an inequality of the form (9) for all of these values of Z. N_1 denotes a matrix size which is large enough to ensure that $|\,b_{N1} - b\,| \leq \Delta\epsilon_b$ with $\Delta\epsilon_b \ll \min[\epsilon_b(Z_l),\epsilon_b(Z_u)]$ is valid for a given required accuracy of $\Delta\epsilon_b$ which is not a function of Z. Under the assumption of monotonical behavior of $b_N(Z)$ as a function of N, except in the close vicinity of Z_{opt}, the slope of b_N at the beginning of the interval $N_0 \leq N \leq N_1$ is greater than its mean value given by

$$\left|\frac{b_{N0} - b_{N1}}{N_0 - N_1}\right| \approx \frac{1}{N_1 - N_0}\,|\,b_{N0} - b\,|. \qquad (10)$$

This assumption can be justified by the ability of the method to produce variational bounds [1], and has also been verified by experience. Hence it follows that

$$\left|\frac{db_N(Z)}{dN}\right|_{N_0} \leq 2\cdot|\,b_N(Z) - b\,|_{N_0} \lesssim 2\cdot(N_1 - N_0)\cdot\left|\frac{db_N(Z)}{dN}\right|_{N_0} \qquad (11)$$

where the symbol $\lesssim$ denotes approximate inequality. The relation expressed in (11) justifies the selection of the weighting factor Z for a not too small value of N so that the modulus of the slope of $b_N(Z)$ is minimized.

From the curves of the calculated susceptance $b_N(Z)$ and the corresponding reflection coefficients r_{0N} illustrated in Fig. 2 it can be seen that the power-conservation principle of Oraizi and Perini [2] is not an optimum criterion for the selection of Z. However, there is agreement for the values of weighting factors in the range of $0.02\cdots 50Z_{\mathrm{opt}}$.

Another more intuitive approach for determining an optimum value of Z is based on the fact that if the single error contributions in (2) have the correct weighting physically, then the convergence must be rapid. Otherwise, the terms containing the electric field alone would also have to be weighted relative to each other. This suggests that the weighting factor should be chosen as

$$Z_N{}^{\mathrm{I}} = \frac{\|E_{tN}{}^{\mathrm{I}}\|}{\|H_{tN}{}^{\mathrm{I}}\|}. \qquad (12)$$

$Z_N{}^{\mathrm{I}}$ converges automatically to the physical Z^{I} as the mode number N is increased. $Z_N{}^{\mathrm{I}}$ can be found by the following iterative procedure:

$$Z_{N\nu+1}{}^{\mathrm{I}} = \frac{\|E_{tN}{}^{\mathrm{I}}(Z_{N\nu}{}^{\mathrm{I}})\|}{\|H_{tN}{}^{\mathrm{I}}(Z_{N\nu}{}^{\mathrm{I}})\|}, \qquad Z_{N0}{}^{\mathrm{I}} = 120\pi. \qquad (13)$$

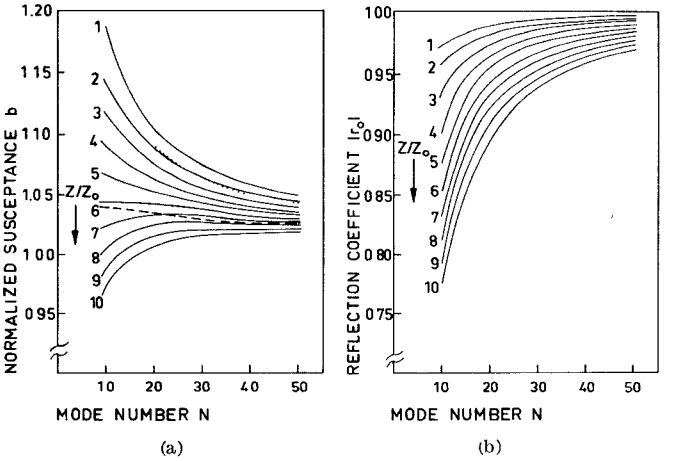

Fig. 2. (a) Normalized susceptance b versus matrix size N for different Z and near optimum N/P. The dashed and the dotted curves refer to optimization with Z_{opt} and $Z_N{}^{I}$, respectively. $N/P = 10/6$; $f = 1$ GHz. (b) Corresponding $|\,r_0\,|$ as a measure of power conservation. $f = 1$ GHz.

In actual use it is observed that (13) converges very rapidly, so that three iterations are sufficient with $N = 30$ for a wide range of dimensions and frequencies. Furthermore, in any of the cases considered the iterated weighting factor Z_N^{I} turned out to be of the order of that Z_{opt} which was found by minimizing $|\, db_N(Z)/dN \,|$. For example, in Fig. 2 the values of Z_N^{I} and Z_{opt} differ by a factor of about 3.5.

NUMERICAL RESULTS

The input admittance $y = jb$ of the resonating structure shown in Fig. 1 was computed and is plotted as a function of frequency. Series resonance occurs when the capacitive gap compensates the input impedance of the propagating circular E_{01} mode. The validity of the program has been tested by comparison with Marcuvitz' results for the capacitive gap at low frequencies, which is same as the problem considered here with $\epsilon_r = 1$ [8]. Generation of the orthonormalized functions t_i^{I}, t_k^{II}, and the elements of the matrix G takes about 30 s of CPU time on a CDC-6400 computer for the sum of the modes $N + P = 80$. The CPU time required for the generation and inversion of a 50×50 complex C_N-matrix is about 5 s.

Fig. 2 shows the dependence of the least squares solution b_N on weighting factor Z and mode number N. At the same time an error estimate is provided, since a change of sign exists for the slope of $b_N(Z)$. For comparison, the deviation of $|\, r_{0N}(Z) \,|$ from its ideal value $|\, r_0 \,| = 1$ is also plotted as a measure of power conservation in Fig. 2.

The dashed curve in Fig. 2 refers to susceptances which were computed by minimizing $|\, db_N(Z)/dN \,|$ with respect to Z at given values of N. The utilization of the iterated weighting factor Z_N^{I} requires less computer time and results in the dotted curve.

In addition to this, the influence of the upper summation limits N and P on the convergence rate has been studied because the ratio N/P plays an important role for methods which exhibit relative convergence phenomena [9], [10]. In Fig. 3 the dependence of b_N on this ratio is shown as a function of the mode number N. The trend of the minimum value of F_N decreasing with N for a fixed $Z = Z_{\mathrm{opt}}$ as seen from Fig. 3 serves as a measure of the convergence rate. It has been found generally that it is sufficient to choose N/P in accordance with point-matching methods, so that in the present case, the near-optimum value of $N/P = R_a/(R_a - R_i)$ lies between 10/5 and 10/6. This coincides with the behavior of F_N shown in Fig. 3 having its maximum average slope F_{10}/F_{50} around this ratio of N/P.

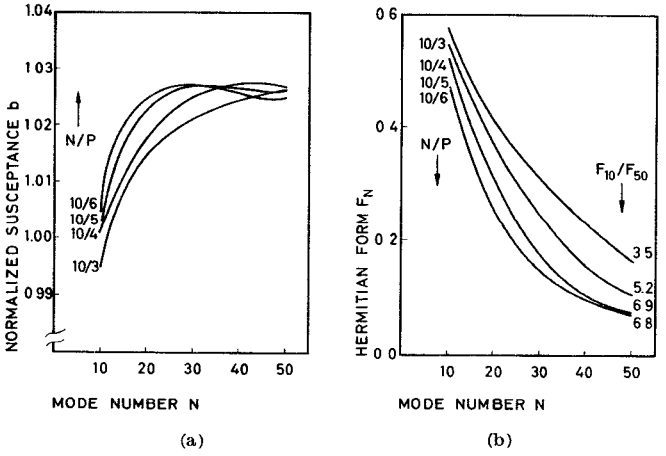

Fig. 3. (a) Influence of the N/P ratio on the convergence of b for near optimum Z. $Z/Z_0 = 8$; $f = 1$ GHz. (b) Corresponding Hermitian forms F_N. The average slope F_{10}/F_{50} serves as a measure of convergence. $f = 1$ GHz.

ACKNOWLEDGMENT

The author wishes to thank Prof. H. Döring for his interest and useful advice in the course of this work. The numerical computations were run at the computer center of the Technical University, Aachen, Germany.

REFERENCES

[1] J.B. Davies, "A least-squares boundary residual method for the numerical solution of scattering problems," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-21, pp. 99–104, Feb. 1973.
[2] H. Oraizi and J. Perini, "A numerical method for the solution of the junction of cylindrical waveguides," *IEEE Trans. Microwave Theory Tech.* (Short Papers), vol. MTT-21, pp. 640–642, Oct. 1973.
[3] B.A. Finlayson and L.E. Scriven, "The method of weighted residuals—A Review," *App. Mech. Rev.*, vol. 19, pp. 735–748, Sept. 1966.
[4] S.G. Mikhlin, *Variational Methods in Mathematical Physics.* New York: Macmillan, 1964, pp. 16, 47, 300–308, 491–504.
[5] W.I. Smirnow, *Lehrgang der hoheren Mathematik*, vol. 1. Berlin: VEB Deutscher, 1971, p. 72.
[6] R.E. Collin, *Field Theory of Guided Waves.* New York: McGraw-Hill, 1960, ch. 8.
[7] E.W. Risley, Jr., "Discontinuity capacitance of a coaxial line terminated in a circular waveguide: Part II—Lower bound solution," *IEEE Trans. Microwave Theory Tech.* (Short Papers), vol. MTT-21, pp. 564–566, Aug. 1973.
[8] N. Marcuvitz, *Waveguide Handbook.* New York: McGraw-Hill, 1951, sec. 3.5 and p. 178.
[9] S.W. Lee, W.R. Jones, and J.J. Campbell, "Convergence of numerical solutions of iris-type discontinuity problems," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-19, pp. 528–536, June 1971.
[10] A. Wexler, "Solution of waveguide discontinuities by modal analysis," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-15, pp. 508–517, Sept. 1967.

The Solution of Electromagnetic Eigenvalue Problems by Least Squares Boundary Residuals

HUGH J. A. LARIVIERE AND
J. BRIAN DAVIES, MEMBER, IEEE

***Abstract*—The least squares boundary residual technique as used for the numerical solution of scattering problems is extended to the solution of electromagnetic eigenvalue problems. The theory is described and numerical results are given for the solution of an L-shaped membrane and microstrip in a hollow conducting guide. The microstrip example was chosen as a test case to compare with Fourier matching. This least square error mimimization technique is of the same family as point matching and Fourier matching; however, it is shown to have three potentially important advantages: 1) it is rigorously convergent, 2) the choice of optimum weighting factors greatly accelerates convergence between a decreasing upper bound and an increasing lower bound, and 3) it is free from problems of relative convergence.**

I. INTRODUCTION

Recently, there has been a surge of interest in the least squares boundary residual technique for the numerical solution of scattering problems [1],[16],[19]. In this short paper, the same approach is extended to the solution of eigenvalue problems, and examples

Manuscript received May 1, 1974; revised October 21, 1974. The work of H. J. A. LaRiviere was supported by the British Department of Trade and Industry and the National Research Council of Canada.

H. J. A. LaRiviere was with the Department of Electronic and Electrical Engineering, University College, London, England. He is now with Bell Northern Research Ltd., Ottawa, Ont., Canada.

J. B. Davies is with the Department of Electronic and Electrical Engineering, University College, London, England.