WHAT DO YOU MEAN BY A SOLUTION TO AN OPERATOR EQUATION?

Tapan K. Sarkar and Ercument Arvas
Department of Electrical Engineering
Syracuse University
Syracuse, New York 13244-1240

ABSTRACT: The main thrust of this presentation is to illustrate that for many electromagnetic field problems the quality of the solution is very subjective and not objective at all. Therefore to classify solutions on a subjective criteria which is not scientific will create more problems than it would solve. The underlying feature to all this is what do "we mean" by a solution.

1. **INTRODUCTION**: There is a renewed interest in recent days on Computer Code Validation. In our opinion, this is a healthy sign, however, we must distinguish between the terms "verify" and "validate". Verification implies the following: Does the code do what it is supposed to do? and validation implies: Does it solve the physical problems? The Penguin English Dictionary specifies validate to mean "soundly based on having legal force and authority" (p.780) whereas verify means "show to be true, check, confirm, authenticate" (p. 784). When performing verification it makes sense to compare the output of various computer codes and compare numbers. However, the only validation check one can make is to compare with experimental results. It is our objective to demonstrate in this short paper, that several codes might give similar results but they may be no where near the "truth". The whole class of "ill-posed" problems fall in this category [1] and interestingly enough almost all physical real life problems are "ill-posed". Perhaps we should consider this as a blessing as we will always be gainfully employed developing numerical codes!

2. ILLUSTRATION OF OUR THESIS

As a first example consider the following integral equation

$$Ku = f \tag{1}$$

where K is the integral operator

$$\int_0^1 k(x,y)\, u(y)\, dy = f(x) \tag{2}$$

For illustration purposes consider the kernel k(x,y) as

$$k(x,y) = \begin{matrix} 1 - y & 0 \le x \le y \\ 1 - x & y \le x \le 1 \end{matrix} \tag{3}$$

Now observe the following eigenvalue problem

$$Ku = \lambda u \tag{4}$$

It is seen that the eigenvalues of the operator K described in (1) are

$$\lambda_n = \frac{4}{(2n-1)^2\pi^2} \tag{5}$$

and the eigenvectors are

$$e_n(x) = \sqrt{2}\cos(2n-1)\ \frac{\pi x}{2} \tag{6}$$

Suppose, now we solve (1) by utilizing an eigenfunction expansion. In that case, we expand the unknown u in terms of known expansion functions e_n with unknown coefficients a , i.e.

$$u = \sum_j a_j e_j \tag{7}$$

Equation (7) is now substituted in (1) and an inner product in the classical "method of moment" context is taken with respect to the weighting functions e_i. Therefore

$$\begin{aligned} < Ku ; e_i > &= < \sum_j a_j Ke_j ; e_i > = < \sum_j a_j \lambda_j e_j ; e_i > \\ &= < f ; e_i > \ \triangleq \int_0^1 f(x)\, e_i(x)\, dx \end{aligned} \tag{8}$$

It is seen that the unknown coefficients are given by

$$a_i = \frac{< f;e_i >}{\lambda_i} \tag{9}$$

and the unknown solution u is given by

$$u = \sum_j \frac{< f;e_j >}{\lambda_j} e_j = \sum_j \frac{(2j-1)^2\pi^2}{4} < f;e_j > e_j \tag{10}$$

Therefore, the solution is obtained in form of a series. Now observe that for many values of f (for example: a constant) that the series in (10) is a divergent series, due to the presence of j in

the numerator. So the solution of (1) is given by a divergent series. The question now is what do you mean by a solution? If we have another divergent series u' as a solution how can we compare the two solutions u and u' and say one is better than the other? This is the crux of the problem.

Mathematically, what has happened is that $\lambda = 0$ is not an eigenvalue of the operator K, (so that we always have a unique solution) but it is the limit point of the eigenvalues of λ_n. (Therefore the eigenfunction expansion is divergent). This is a typical "characteristic" of an ill-posed problem. Next we will see how this carries over to electromagnetic problems.

3. **PROBLEM OF INTEREST.** Let us consider the first problem that everybody tries to solve numerically on their first encounter with electromagnetics. This is the electrostatic problem. As an example consider the charge distribution on a disk. The charge distribution $\sigma(r)$ satisfies the simple integral equation:

$$\int\int_s \frac{\sigma(\vec{r'})r'\,dr'\,d\theta'}{4\pi\epsilon_0\,|\,\vec{r}(\theta,r) - \vec{r'}(\theta',r')\,|} =$$

$$\int\int_s \frac{\sigma(\vec{r'})r'\,dr'\,d\theta'}{4\pi\epsilon_0[r^2 + (r')^2 - 2rr'\cos(\theta-\theta')]^{\frac{1}{2}}} = (\text{a constant})\ V \qquad (11)$$

In our presentation we shall look at this equation only and analyze what we mean by the solution to this problem. There are several factors that need to be considered even in this simple equation. To prove things one has to be as general as possible. However, a counter example would suffice to disprove claims.

(A) QUALITY OF THE SOLUTION

Let us consider the charge distribution on a circular disk of radius "a" charged to a constant potential. It is well known that the charge distribution can be solved for exactly and the solution is given by

$$\sigma_{exact} = \frac{c}{\sqrt{r^2 - a^2}} \qquad c = \text{a constant} \qquad (12)$$

Now two different computer programs, let us say, generate two different solutions which are of the form

$$\sigma_I = \frac{c}{(r^2 - a^2)^{0.6}} \qquad (13)$$

$$\sigma_{II} = \frac{c}{(r^2 - a^2)^{0.7}} \qquad (14)$$

Now the question that is raised is which of the two solutions is better. In order to say definitely which is the better solution, one has to define an error criterion which is quantitative: This implies

that one must define an error criteria before one can say which is the better solution. The most common error criteria that we use in everyday life are introduced through the concept of the two norms:

$$L_\infty = \max_{\text{for all } r} \left| \sigma_I(r) - \sigma_{exact}(r) \right| \tag{15}$$

$$L_2 = \int_s [\sigma_I(r) - \sigma_{exact}(r)]^2 \, dr \tag{16}$$

The first one is the min-max error criterion and the second is the mean squared error. Unfortunately neither of these error criteria can be used in evaluating the quality of the solution as (12) - (14) are not square integrable. Therefore, even though from a "subjective" view point (13) is a better solution than (14) there is absolutely no scientific grounds on judging the better solution. To go a step further, in this example we know what the exact solution is and therefore we can devise subjective criteria as the nature of the approximate solution. Now the question is: what subjective criteria to use when the exact solution is unknown? It may therefore be very dangerous to validate codes based on heuristic subjective judgement, like the shape of the charge distribution. Sometimes the value of the capacitance is used as a basis of comparison. This is also heuristic as nowhere any error related to the value of the capacitance is minimized to yield (11). Verification then based upon the value of the capacitance may not be sound as then the datum becomes problem dependent.

(B) THEORETICAL NATURE OF THE SOLUTION

The next problem that we look into is the nature of the solution of this Fredholm integral equation of the first kind. This problem has been analyzed exhaustively and is summarized by Stakgold [Green's functions and boundary value problems pp. 516-517]. Stakgold states, "Equation (11) is a Fredholm equation of the first kind with a symmetric Hilbert-Schmidt kernel. We know that 0 is in the continuous spectrum of the corresponding integral operator so that (11) cannot have an L_2 solution for each $f \in L_2(s)$. Nevertheless, even in these cases one can interpret the 'solution' in a distributional sense, and although the formal eigenfunction expansion for σ may diverge, the corresponding expressions for potential will be well-behaved".

The above paragraph clearly states that it will be totally meaningless to compare the charge distribution of various solution techniques as the solution is "divergent". Secondly, the integral operator has a continuous spectrum of eigenvalues and not discrete. Since 0 is in the continuous spectrum, the inverse operator is unbounded. The implication of this is (for the solution of AX=Y)

$$\begin{aligned} X_{exact} - X_{approximate} &= A^{-1}(AX_e - AX_a) \\ &= A^{-1}(Y - AX_a) \end{aligned} \tag{17}$$

therefore

$$\| X_e - X_a \| \leq \| A^{-1} \| \cdot \| Y - AX_a \| \tag{18}$$

Thus minimization of the residual, implies a better solution only when $\|A^{-1}\|$ is bounded. This is known as the stability of the solution. Example will be presented to illustrate what happens when $\|A^{-1}\|$ is large.

Distributional convergence implies that a sequence $\{U_n\}$ converges distributionally to the distribution $\{U\}$ as $n \to n_o$ and we write $U_n \to U$ as $n \to n_o$ if

$$n \to n_o, \quad < U_{no}; \phi > \to < U; \phi >$$

or

$$\lim_{n \to no} < U_{no}; \phi > = < U; \phi > \qquad (19)$$

where ϕ is called a test function [p 153 Stakgold].

A test function ϕ must be

(i) infinitely differentiable, and

(ii) ϕ with all its derivatives, vanish at $|x| = \infty$ faster than any negative powers of x. This means

$$\lim_{|x| \to \infty} \left| x^k \frac{dP\phi}{dxP} \right| = 0 \quad \text{for all P.}$$

Therefore the test functions are very restrictive. Observe now that covergence of distributions has been reduced from covergence of functions U_n to covergence of numbers $< U_n; \phi >$. Hence one can ascertain that for each ϕ the left hand side of (19) has a limit as $n \to n_o$ but one is not sure that the limiting values are indeed values of a distribution u. In simple terms it implies covergence is "subjective" as it depends on the test functions ϕ.

From the above discussions, it becomes quite clear that the delta function is excluded from being a possible test function as it is not differentiable. This is because $< f; \delta > = f(o)$. Hence if the function f is discontinuous then the inner product is undefined. So the delta function weighting makes sense from a purely mathematical point of view when the operator generates a continuous function in the range (for open problems, this cannot happen as the field is infinity at the edges). The delta function can be used as weighting functions for closed bodies. However, for open bodies delta functions are not possible test functions. Even though, from a practical point of view one can always use delta functions as test functions, there is no mathematical basis of comparison as to which technique provides superior results. The comparison then becomes an art!

Hence the solution of (11) has to be interpreted in a distributional sense. Almost everything is convergent in a distributional sense!! This is a mathematical artifact and it is difficult to present it physically. In summary, Stakgold points out, that the representation of the charge distribution can be represented only by a divergent series! However, the error in the potential can be a basis of comparison.

(C) NUMERICAL NATURE OF THE SOLUTION

From our experience of numerical computation, we all know that the solution of the charge distribution of (11) is quite stable. The reason for that is again outlined by Stakgold on p. 517 [2]. Stakgold points out "Another advantage of (11) is that the corresponding integral operator is positive i.e. zero is not a solution". When we solve a problem numerically, we are computing in a finite dimensional space. In that space all the eigenvalues become discrete and the continuous spectrum transforms to a discrete spectrum. It turns out that zero is not an eigenvalue of the operator, and all the eigenvalues are positive. However, for more accurate solution one had to do a fine discretization and the size of the matrix will increase.

What Stakgold points out is that as the size of the problem increases, the smallest eigenvalue of the moment matrix approaches zero. Hence, if we solve a small size problem, computation is quite stable and as the dimension of the matrix increases, the numerical stability of the computation comes into question. The crux of the problem is: For more accurate solution, one needs a larger matrix, but one also faces the problem of numerical stability!

So the question raised is how can one perform meaningful code validation under these circumstances?

We illustrate this point through a simple numerical example.

Consider solution of the matrix equation $Ax=y$ where A is the 4X4 symmetric matrix of [3]:

$$A = \begin{bmatrix} 36.86243 & 51.23934 & 53.50338 & 50.49425 \\ 51.23934 & 71.22350 & 74.37005 & 70.18714 \\ 53.50338 & 74.37005 & 77.66275 & 73.29752 \\ 50.49425 & 70.18714 & 73.29752 & 69.17882 \end{bmatrix}$$

and

$$y^T = [192.09940 \quad 267.02003 \quad 278.83370 \quad 263.15773]$$

where the superscript T denotes transpose. It is readily verified that the exact solution is given by

$$x^T = [1 \ 1 \ 1 \ 1].$$

However, it can be shown that A is an ill-conditioned matrix in the sense that the ratio of its maximum to minimum eigenvalues is on the order of 10^{18} .

Let us study the effect on the solution x when only one component of y is changed in the fifth decimal place. Specifically, we obtain the results presented in Table III. Clearly, extremely small perturbations in y result in large variations in x .

y^T[192.09939	267.02003	278.88370	263.15773]
x^T[-6,401,472,429	3,866,312,299	1,607,634,613	-953,521,374]
y^T[192.09940	267.02002	278.83370	263.15573]
x^T[3,866,312,299	-2,335,145,694	-970,966,842	575,900,539]
y^T[192.09940	267.02003	278.83369	263.15573]
x^T[1,607,634,615	970,966,842	-403,733,529	239,462,717]
y^T[192.09940	267.02003	278.83370	263.15772]
x^T[-953,521,374	574,900,539	239,462,717	-143,030,294]

4. SUMMARY

It has been demonstrated that even the simplest integral equation in electromagnetics need not provide a numerically stable solution. The instability is not in the computation process but with the operator itself. Hence, when one is dealing with an "ill-posed" operator, the comparison of various entities for such problems has to be performed subjectively and not objectively, such as error in the solution or in the residuals. This makes code validation a rather difficult problem as there is no "fixed northern star" to get the datum from for an arbitrary problem. It is important to point out that for "ill-posed" problems the quality of the solution is always "subjective" and not "objective" at all. Therefore, it is wrong and mathematically incorrect to say that one has a mathematically valid solution to an ill-posed problem (see [1] for details).

5. EPILOGUE

Given the history on how "methods of moments" became a part of the electromagnetic community (the first paper was rejected) one has to be extremely careful in making "subjective" decisions about computer codes!

REFERENCES

1. T.K. Sarkar, D.D. Weiner and V.K. Jain, "Some Mathematical Considerations in Dealing with an Inverse Problem" IEEE Transactions on Antennas and Propagation, March 1981, pp. 373-379

2. Ivar Stakgold, "Green's Functions and Boundary Value Problems", 1979.

3. A.E. Hoerl and R.W. Kennard, "Ridge Regression-Past, Present and Future", at the International Symposium on Ill-Posed Problems: Theory and Practice, University of Delaware, Newark, October 1979