An Initial Value Method for Solving Fredholm Integral Equation of the First Kind*

by v. vemuri

School of Advanced Technology State University of New York at Binghamton, Binghamton, New York

and FANG-PAI CHEN

School of Aeronautics and Astronautics Purdue University, Lafayette, Indiana

ABSTRACT: The difficulties in solving Fredholm integral equations of the first kind are well known. A classical method has been to convert the equation into a set of m linear algebraic equations in n unknowns ($m \le n$). For computational convenience, it is customary to force m = n and solve the resulting ill-conditioned system using one technique or other. In the general case, a feasible solution, if it exists, can be found by determining the generalized inverse of the coefficient matrix. One method of finding the generalized inverse is to reformulate the problem and observe the steady state response of a system of ordinary differential equations with prescribed initial conditions. Results obtained from this reformulation are found to be comparable in quality to those obtained earlier by other methods. Analog and digital computer implementations are discussed.

I. Introduction

The Fredholm integral equation of the first kind

$$\int_{a}^{b} K(x,y) f(x) dx = g(y), \quad a \leqslant y \leqslant b,$$
(1)

occurs frequently in many branches of physical and biological sciences. The hapten-binding equation of immunology (1), for example, is an equation of this kind. The problem of locating tumors in a body using radiographic techniques (2) can be formulated as an integral equation of the first kind. The problem of deducing the structure of a planetary atmosphere from satellite observations (3) can also be reduced to a mathematical problem similar to Eq. (1). Indeed Eq. (1), and variations thereof, plays a central role in many experimental sciences whenever physical data are gathered by indirect sensing devices as is the case with many remote sensing experiments (4).

The convolution type integral equation, which is a special case of Eq. (1) with x = t, $K(x, y) \triangleq h(t - y)$, plays a central role in the identification of

* The work reported in this paper is supported in part by Grant GK31786 from the National Science Foundation.

linear systems. Since the input–output relationships for a time-invariant linear system can be formulated as an integral equation of the convolution type, identification of the system's impulse response, h(t), from a record of input and output functions f(t) and g(t), requires the solution of an integral equation such as

$$\int_{a}^{b} h(t-\tau) f(\tau) d\tau = g(t).$$

This type of problem is called a *system identification* problem. An equivalent problem is to find the system input, f(t), given the impulse response h(t) and the output g(t). This is referred to as *input identification*. Problems belonging to the latter class are of general interest in remote sensing experiments, and attention in this paper will therefore be focused on such types of problems.

Treatment of Eq. (1) is difficult and complicated. For some types of kernels, a solution is possible only for a restricted class of functions g(y). For instance, it can be seen, by successive differentiation of Eq. (1), that if K(x, y) satisfies the linear differential equation of the form

$$\left\{\frac{\partial^n}{\partial y^n} + p_1(y)\frac{\partial^{n-1}}{\partial y^{n-1}} + \ldots + p_n(y)\right\}K(x,y) = 0,$$

then solution to Eq. (1) exists only if g(y) satisfies the same differential equation. Thus for kernels of this kind there is no well-behaved solution for arbitrarily well-behaved g(y). The problem is further complicated by the fact that the solution is extremely sensitive to the presence of noise, such as measurement errors or rounding errors. Phillips (5) demonstrated that it is possible to add a finite quantity to any solution f(x) adding only an infinitesimal amount to the observed function g(y). Hence, if K(x,y) and g(y) contain any errors at all, the solution f(x) will be unreliable. Typically, the solution one obtains is unstable and oscillatory. It can also be shown (5) that the existence of errors converts the problem from one with a unique solution to one with an infinite number of solutions. To choose one out of many solutions, it is necessary to impose constraints on the problem. One suggested constraint is to pick a solution that exhibits a minimum second difference over the family of solutions. Following this lead, Twomey (6) formulated the problem as a constrained optimization problem using Lagrange multipliers. One common feature of both the above methods is that they require a matrix inversion which is generally a time-consuming operation. To overcome the storage and time restrictions, Hunt (7) used fast Fourier transforms to solve the convolution type integral equation. As pointed out by Hunt, the transform method lacks the generality of the method suggested by Phillips and Twomey.

The importance of this problem prompted others to try various other approaches. For instance, Bellman *et al.* (8) treated this problem using dynamic programming in conjunction with a successive approximation

technique. While agreeing that no one technique can satisfactorily resolve the fundamental problem of obtaining sensible solutions from ill-conditioned systems, the authors strived to point out that a reasonable compromise is to combine various techniques such as dynamic programming, successive approximations, extrapolation and smoothing.

II. Motivation

The method described in this paper is new and radically different from all the earlier methods. Essentially, this method is an application of a recent result obtained by Menhertz *et al.* (9). This method converts the problem of solving the integral equation into an equivalent one of solving a system of ordinary differential equations with prescribed initial conditions. Thus, the problem becomes amenable for solution on digital, analog or hybrid computers.

The idea of reformulating an integral equation as an initial value problem is not entirely new. Fredholm integral equations of the second kind have been reformulated as systems of ordinary differential equations using invariant imbedding (10). These equations have also been solved on an analog/hybrid computer (11). To date, however, Fredholm integral equations of the *first kind* have not been solved after reformulating them as initial value problems. Also, the possibility of using an analog computer to solve Fredholm equations of the first kind appears to be new.

A first step in the new procedure is to replace, as is usually done in conventional methods, the continuous variable y with a finite set of mesh points $y_1, y_2, ..., y_m$,

$$a \leqslant y_1 < y_2 < \ldots < y_m \leqslant b$$

and to write

$$g(y_1) = \int_a^b K(x, y_1) f(x) dx$$

$$\vdots$$

$$g(y_m) = \int_a^b K(x, y_m) f(x) dx.$$
(2)

In experimental sciences $g(y_i) \triangleq g_i$ (i = 1, 2, ..., m) generally correspond to m experimentally observed data points. For simplicity, these data are assumed to be noise-free even though the contrary is generally true.

The second step of the discretization process is to replace the continuous variable x by a finite set of mesh points $x_1, x_2, ..., x_n$ such that

$$a \le x_1 < x_2 < \dots < x_n \le b$$
.

To carry out this step, each of the integrals in Eq. (2) is replaced by a

suitably chosen numerical quadrature formula to yield

$$g(y_1) \approx \sum_{j=1}^{n} w_j K(x_j, y_1) f(x_j),$$

$$\vdots \qquad \vdots$$

$$g(y_m) \approx \sum_{j=1}^{n} w_j K(x_j, y_m) f(x_j),$$
(3)

where $w_1, w_2, ..., w_n$ are the weights associated with the quadrature formula. Equation (3), in vector-matrix notation, is

$$g \approx Af$$
, (4)

where $g = [g(y_1), g(y_2), g(y_3), \dots, g(y_m)]^T$, $f = [f(x_1), f(x_2), \dots, f(x_n)]^T$ in which the superscript T denotes the transpose and A is an $(m \times n)$ matrix given by

$$A = \left[\begin{array}{cccc} w_1 \, K(x_1, y_1) & w_2 \, K(x_2, y_1) & \dots & w_n \, K(x_n, y_1) \\ w_1 \, K(x_1, y_2) & w_2 \, K(x_2, y_2) & \dots & w_n \, K(x_n, y_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ w_1 \, K(x_1, y_m) & w_2 \, K(x_2, y_m) & \dots & w_n \, K(x_n, y_m) \end{array} \right]$$

As Eq. (4) represents m equations in n unknowns, a usual procedure is to force m=n and recover f by inverting the resulting square matrix A. However, it can be shown (12) that A is ill-conditioned and indeed approaches, in the limit as $n\to\infty$ and $m\to\infty$, the Hilbert matrix. Indeed, choosing n larger and larger to reduce the discretization error is not a panacea to the ill-conditioning problem. On the one hand, large n demands that we make m correspondingly larger (i.e. in effect make more measurements). On the other hand, large n tends to make A more and more ill-conditioned. Making n too small is also not desirable as it renders the approximation very crude. Indeed, almost all the techniques reviewed earlier are aimed at the problem of inverting the ill-conditioned matrix A or at by-passing the inversion problem completely.

The method suggested in this paper offers several advantages: (1) It no longer requires A to be a square matrix; (2) The problem of inverting an ill-conditioned matrix is completely by-passed by reformulating the problem as one of solving a set of ordinary differential equations as initial value problems; (3) For the first time, Fredholm integral equations of the first kind can be solved on analog/hybrid computers; (4) The problem of estimating the kernel K(x,y) from input-output data is now relatively straightforward as several estimation techniques can be readily applied to the new formulation.

III. Description of the Initial Value Method

The starting point of the method is Eq. (4) which is reproduced here for convenience

$$Af = g$$
.

While solving Eq. (4), the system may exhibit one of three characteristics: (1) the system may be infeasible (i.e. no solution may exist); (2) the system may have an infinite number of solutions; or (3) the system may have a unique solution. Regardless of the nature of the system, a reasonable approach is to minimize the error (Af-g) in a least square sense. Among the class of vectors f that minimize the Eucledean norm of the error, there exists a unique vector A^+g which is of least norm where A^+ is the generalized inverse of A (13, 14). Thus the problem of solving the integral Eq. (1) is equivalent to solving Eq. (4), which in turn is tantamount to finding the generalized inverse of A. This observation is the point of departure between the conventional methods and the new method.

There exist many elegant iterative methods of determining the generalized inverse of an arbitrary rectangular matrix (15). However, for an initial value formulation, the task of finding A^+ is made possible by the following theorems (9, 16).

Theorem I. For any arbitrary real $(m \times n)$ matrix A the inverse is given by

$$A^{+} = \lim_{\lambda \to 0} \left[(\lambda I + A^{\mathrm{T}} A)^{-1} A^{\mathrm{T}} \right]. \tag{5}$$

Proof of Theorem I. This can be found in the paper of Manhertz et al. (9). Theorem II. For any real, finite $(m \times n)$ matrix A and column vector g, the system

has a unique asymptotic solution given by

$$\lim_{t \to \infty} f(t) = A + g. \tag{7}$$

Theorem II can be readily proved by taking Laplace transforms on both sides of Eq. (6) and applying the Final Value Theorem:

$$F(s) = (1/s)(sI + A^{T}A)^{-1}A^{T}g,$$
(8)

$$\lim_{t \to \infty} f(t) = \lim_{s \to 0} sF(s) = \lim_{s \to 0} (sI + A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}g.$$
 (9)

Equations (9) and (5) give the result shown in Eq. (7).

Thus, if Eq. (6) is programmed, say, on an analog computer, and if the computer is allowed to reach steady state (i.e. the integrator outputs become sufficiently constant), the value of f(t) obtained constitutes a solution of the given integral equation.

Note that $A^{\mathrm{T}}A$ is at best positive semidefinite (9). Therefore an unforced version of Eq. (6) is not necessarily stable. The forced system, however, is stable due to the form of the forcing function $A^{\mathrm{T}}g$. Since $A^{\mathrm{T}}g$ cannot be programmed exactly anyway, Eq. (6) can be slightly modified to ensure stability, i.e.

$$\dot{f} = -(A^{T} A + \delta I) f + A^{T} g,
f(0) = 0,$$
(10)

where δ is a small positive scalar and I is the identity matrix. If δ is small compared to the nonzero eigenvalues of $A^{\rm T}A$, then the error introduced is small. In practice, a small δ can be chosen first and as the solution reaches its asymptotic value, the value of δ may be made smaller and smaller until no further improvement in the solution is possible.

IV. Illustrative Examples

The procedure suggested in the preceding section was applied to solve three different problems. The results are compared with solutions of the same problems obtained by others. To facilitate comparison, only the cases in which the matrix A is chosen to be square (i.e. m=n) are presented here even though these three problems were solved using various combinations of m and n with $m \neq n$. Secondly, while implementing the new method, Simpson's quadrature formula was used whenever the number of grid points in the integration formula (i.e. n) is odd and Newton-Coates closed integration formula was used whenever n is even. The resulting systems of ordinary differential equations were solved (1) on a digital computer in conjunction with a Runge-Kutta fourth-order method and (2) on an analog computer. Even though all the examples were also solved on an EAI 680 analog computer only the results obtained from one problem are reported in this paper.

Example 1 (the Bellman problem)

$$\int_0^1 (x-y)^2 f(x) \, \mathrm{d}x = \frac{1}{2} y^2 - \frac{2}{3} y + \frac{1}{4}, \quad 0 \leqslant y \leqslant 1.$$
 (11)

This problem was first discussed by Bellman *et al.* (8). Solution of this equation is readily seen to be f(x) = x by direct integration of the left side. Bellman *et al.* discretized this problem into an 11×11 linear system

$$Ax = b$$

using Simpson's quadrature formula where

$$b_{i} = \frac{1}{2}y_{i}^{2} - \frac{2}{3}y_{i} + \frac{1}{4}, \quad i = 1, 2, ..., 11,$$

$$\alpha_{ij} = \frac{1}{10}w_{j}(x_{j} - y_{i})^{2}, \quad i, j = 1, 2, ..., 11,$$
(12)

with the weighting coefficients w_i given by

$$[w_j] = \frac{1}{3}(1, 4, 2, 4, ..., 2, 4, 1), \quad j = 1, 2, ..., 11.$$

The solution obtained by direct inversion (see Fig. 1) shows wide oscillations. This problem is solved once again using the new method for three different values of m = n = 6, 7, 11. The results are shown in Fig. 2.

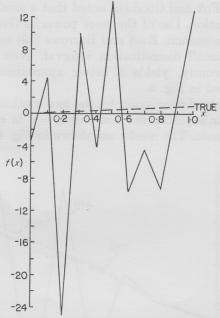


Fig. 1. Solution of Eq. (11), for n = 11, by direct inversion of the coefficient matrix A.

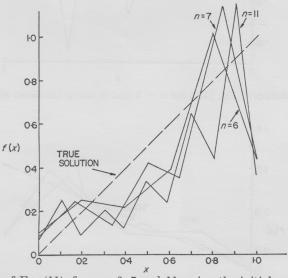


Fig. 2. Solution of Eq. (11), for n = 6, 7 and 11, using the initial value formulation.

Example 2 (the Fox-Goodwin problem)

$$\int_0^1 (x^2 + y^2)^{\frac{1}{2}} f(x) \, \mathrm{d}x = \frac{1}{3} [(1 + y^2)^{\frac{3}{2}} - y^3], \quad 0 \le y \le 1.$$
 (13)

This equation was first discussed by Fox and Goodwin (17). The solution of this equation is f(x) = x. Using Simpson's quadrature, respectively, with

5 and 9 grid points Fox and Goodwin noted that a smaller interval led to a more oscillatory solution. Use of the more powerful Gaussian formulae also led to a similar phenomenon. Rust and Burrows (18) repeated this problem and noted that a small discretization interval, even though it tends to oscillate more vigorously, yields a better approximation. Some typical results are reproduced in Fig. 3.

This problem is solved once again by the new method with n=5 in conjuction with Simpson's quadrature and with n=6 in conjunction with the Newton-Coates formula. The results are shown in Fig. 4.

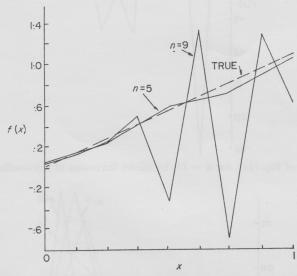


Fig. 3. Solution of Eq. (13), for n = 5 and 9, using Gaussian elimination.

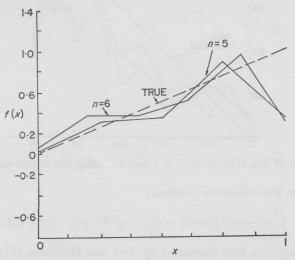


Fig. 4. Solution of Eq. (13), with n=5 and 6, using the initial value formulation.

Example 3 (the Phillips problem)

$$\int_{-6}^{+6} K(x, y) f(x) dx = g(y), \quad |y| \le 6,$$
(14)

where

$$K(x,y) = \begin{cases} 1 + \cos\frac{\pi(x-y)}{3}, & |x-y| \le 3, & |y| \le 6, \\ 0, & |x-y| \ge 3, & |y| \le 6, \end{cases}$$
 (15)

and

$$g(y) = \begin{cases} (6 - |y|) \left[1 + \frac{1}{2} \cos \frac{\pi y}{3} \right] + \frac{9}{2\pi} \sin \frac{\pi |y|}{3}, & |y| \leqslant 6, \\ 0, & |y| \geqslant 6. \end{cases}$$
 (16)

This problem was first suggested by Phillips (5). The solution of the above equation is

$$f(x) = \begin{cases} 1 + \cos\frac{1}{3}\pi x, & |x| \le 3, \\ 0, & |x| \ge 3. \end{cases}$$
 (17)

Phillips solved this problem by discretizing the problem using Simpson's rule and solving the linear system via Gaussian elimination. His results are shown in Fig. 5.

Results obtained by the new method are shown in Fig. 6.

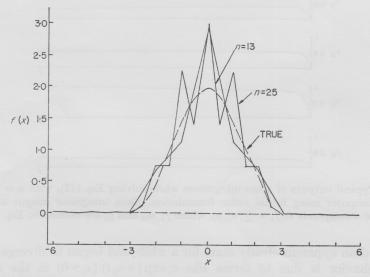


Fig. 5. Phillips' solution of Eq. (14), for n = 13 and 25, using Gaussian elimination.

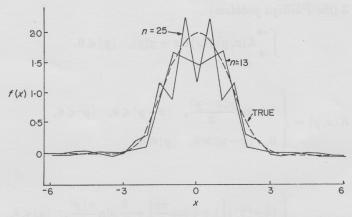


Fig. 6. Solution of Eq. (14), with n = 13 and 25, using the initial value formulation.

Example 4 (the Fox-Goodwin problem on an analog computer)

To demonstrate the feasibility of solving these equations on an analog computer, Example 2 is repeated once again on an analog computer. The solutions are shown in Fig. 7. As time increased, the integrator outputs

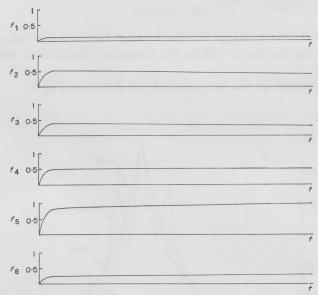


Fig. 7. Typical outputs of some integrators while solving Eq. (13), with n=6, on an analog computer using initial value formulation. Each integrator output at steady state corresponds to $f_i = \sum_i a_{ij}^+ g_j$, where f_i , a_{ij} and g_j are defined in Eq. (4).

exhibited an apparent steady state for a while and began to diverge slowly. This behavior is due to terms like $c_i \exp(+\alpha_i t)$ ($\alpha_i > 0$) in the solution which are due to small errors such as truncation errors and errors due to

approximation of the kernel. Attempts to minimize errors due to approximating the kernel resulted in a wild oscillation of the solution, perhaps because of the poor conditioning of the resulting system. Filtering techniques to impart stability, by removing the undesirable components, are discussed in the sequel.

V. Improvements

The preceding section essentially demonstrates the feasibility of the new method of solving Fredholm integral equations of the first kind as an initial value problem on digital or analog computers. As can be seen, by comparison, the results are comparable in quality to those obtained by other methods. In the course of several experiments with various grid sizes, it was noted, as expected, that as the grid size used in integration formulas decreases the system becomes more and more ill-conditioned and the solution more and more oscillatory. Secondly, the solution of the ordinary differential equations, viz. Eq. (6), never achieves a true steady state. The reason for this is the presence of small unwanted eigenvalues introduced by the approximation process. Both of these problems can be remedied to some extent by a suitably chosen filtering process. Phillips and Towmey (5, 6) attempted this filtering by writing the solution f of Eq. (4), as

$$f = (A^{T} A + \gamma H)^{-1} A^{T} g, \tag{18}$$

where H is a constraint matrix of some convenient form. Specifically, Phillips and Towmey chose H to be the second difference matrix, i.e.

$$H = \begin{bmatrix} 1 & -2 & 1 \\ -2 & 5 & -4 & 1 \\ 0 & 1 & 4 & -6 & -4 & 1 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \end{bmatrix}. \tag{19}$$

Following this lead, the initial value formulation shown in Eq. (6) is rewritten as

$$\dot{f}(t) = -(A^{T} A + \gamma H) f(t) + A^{T} g,$$

$$f(0) = 0.$$
(20)

Note that this equation is the same as Eq. (10), obtained earlier from a stability argument, if H=I and $\gamma=\delta$. The three model problems are solved once again using Eq. (20), rather than Eq. (6). Each problem is repeated with various values of γ , and some selected results are shown in Figs. 8–11. It can be seen that Phillips' problem, with the best behaved kernel of all three, required only a small value of γ (i.e. required the least amount of smoothing) in order to match the exact solution.

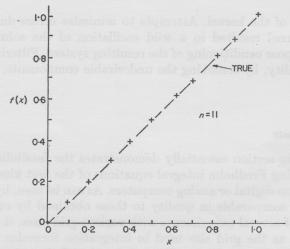


Fig. 8. Solution of Eq. (11), for n=11, using a smoothing technique, i.e. by implementing Eq. (20).

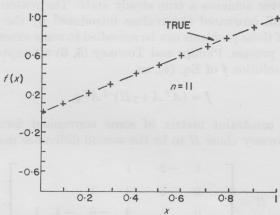


Fig. 9. Solution of Eq. (13), for n = 11, using a smoothing technique, i.e. by implementing Eq. (20).

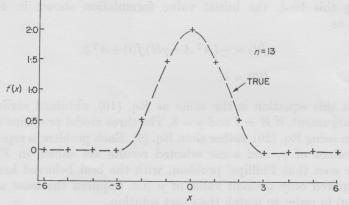


Fig. 10. Solution of Eq. (14), for n = 13, using a smoothing technique, i.e. by implementing Eq. (20).

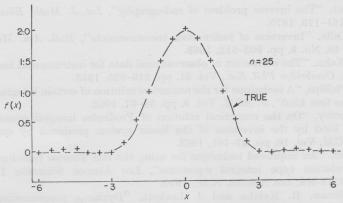


Fig. 11. Solution of Eq. (14), for n=25, using a smoothing technique, i.e. by implementing Eq. (20).

VI. Discussion and Conclusions

This paper has demonstrated the utility of a new method of solving Fredholm integral equations of the first kind. The first step of the method involves, as usual, the replacement of the integral by a summation and rewriting the integral equation as Af = g. However, it is no longer required that the resulting coefficient matrix A be square. In practical problems this means freedom in choosing the step size of the integration algorithm independent of the number of data points available on g. The next stage involves solving for f as f = A + g where A + is the generalized inverse of A. Instead of finding A^+ using iterative methods, the method determined A^+ by looking at the steady state solution of a set of ordinary differential equations with specified initial conditions. This reformulation allows one to use the powerful techniques associated with ordinary differential equations. Just as A (or $A^{\mathrm{T}}A$) is often ill-conditioned, the ordinary differential equations obtained after the reformulation also tend to be unstable. At this point, the welldeveloped stability theory of ordinary differential equations can be readily used to tackle the stability problem. Furthermore, the initial value formulation allows one to solve the given equation in near real time by obtaining the solution using an analog or hybrid computer.

To improve the quality of solutions obtained, a stabilizing technique has been used. The crux of the idea is to render $A^{\rm T}A + \gamma H$ positive definite by choosing a sufficiently large γ using a trial and error procedure. It was noted that the bigger the value of γ , the slower the convergence. If γ is too small, the solution is oscillatory. The present investigation, among other things, is aimed at determining an optimum value of γ for any given problem.

References

(1) F. Aladjem and M. T. Palmiter, "The antigen—antibody reaction—V: A quantitative theory of antigen—antibody reactions which allows for heterogeneity of antibodies", J. Theor. Biol., Vol. 8, pp. 8–21, 1965.

- (2) B. Hunt, "The inverse problem of radiography", Int. J. Math. Biosci., Vol. 8, pp. 161–179, 1970.
- (3) J. C. Gille, "Inversion of radiometric measurements", Bull. Am. Meterol. Soc., Vol. 49, No. 9, pp. 903–912, 1968.
- (4) F. D. Kahn, "The correction of observational data for instrumental band width", Proc. Cambridge Phil. Soc., Vol. 51, pp. 519–525, 1955.
- (5) D. L. Phillips, "A technique for the numerical solution of certain integral equations of the first kind", J. ACM, Vol. 9, pp. 84-97, 1962.
- (6) S. Twomey, "On the numerical solution of Fredholm integral equation of the first kind by the inversion of the linear system produced by quadrature", J. ACM, Vol. 10, pp. 97-101, 1963.
- (7) B. Hunt, "An improved technique for using the fast Fourier transform to solve convolution type integral equations", Los Alamos Scientific Lab. Rep., LA-4515-MS, Los Alamos, N.M., 1970.
- (8) R. Bellman, R. Kalaba and J. Lockett, "Dynamic programming and ill-conditioned linear systems", J. Math. Analysis Appl., Vol. 10, pp. 206–215, 1965.
- (9) R. K. Manherz, B. W. Jordan and S. L. Hakimi, "Analog methods for computation of the generalized inverse", *IEEE Trans. Auto. Control*, Vol. AC-13, No. 5, pp. 582–585, 1968.
- (10) H. H. Kagiwada and R. E. Kalaba, "An initial value theory for Fredholm's integral equations with semi-degenerate kernals", Rand Corp. Memo. RM5602-PR, Santa Monica, Calif., 1968.
- (11) V. Vemuri, "Hydrid computation in hydrosystems", in Adv. Hydrosci. (Ed. Ven te Chow), Vol. 9, pp. 1–98, Academic Press, New York, 1973.
- (12) J. Y. Wang, "Theory and applications of inversion techniques: a review", Purdue Univ. Rep., AA & ES 70-6, Sept. 1970.
- (13) R. Penrose, "On best approximate solutions of linear algebraic equations", Proc. Cambridge Phil. Soc., Vol. 52, pp. 17–19, 1956.
- (14) R. Penrose, "A generalized inverse of matrices", Proc. Cambridge Phil Soc., Vol. 51, pp. 406-413, 1955.
- (15) R. P. Tewarson, "An iterative method for computing generalized inverses", Int. J. Computer Math., Vol. 3, No. 1, pp. 65-74, Dec. 1971.
- (16) C. A. den Broeder, Jr., and A. Charnes, "Contributions to the theory of generalized inverses of matrices", Purdue Univ., Lafayette, Indiana, 1957.
- (17) L. Fox and E. T. Goodwin, "The numerical solution of nonsingular linear integral equations", *Phil. Trans. Roy. Soc.*, Lond., Vol. A245, pp. 501–534, 1953.
- (18) B. W. Rust and W. R. Burrows, "Mathematical Programming and the Numerical Solution of Linear Equations, Chap. 1, American Elsevier, New York, 1972.