THE PENNSYLVANIA
STATE UNIVERSITY

IONOSPHERIC RESEARCH

Scientific Report No. 344

DECONVOLUTION OF PHYSICAL DATA

by

J. P. Rarick

November 30, 1969

IONOSPHERE RESEARCH LABORATORY

University Park, Pennsylvania NASA Grant NGL 39-009-032

Ionospheric Research

NASA Grant NGL 39-009-032

Scientific Report

on

"Deconvolution of Physical Data"

by

J. P. Rarick

November 30, 1969

Scientific Report No. 344

Ionosphere Research Laboratory

Submitted by: BRF Kendall

B. R. F. Kendall, Assoc. Prof. of Physics

Approved by: A.H. Waynick

A. H. Waynick, Director, IRL

The Pennsylvania State University

TABLE OF CONTENTS

ABSTRACT

Various methods useful for deconvolution of physical data in order to remove systematic distortions are discussed. Digital, numerical and analog techniques are described along with experimental results which indicate the merits of different methods. An analog method is mathematically analyzed in detail demonstrating that it performs a modified symmetric Gauss-Seidel iteration. Convergence criteria and the effects of noise are also discussed briefly.

CHAPTER I

INTRODUCTION

1.1 Convolution and Deconvolution

In the process of measuring and recording any physical observable in experimental physics, the quantity being measured is filtered by the measuring process. The optimum instrument is the one which records the quantity being measured with a minimum of distortion; i.e., the instrument which has the highest frequency response or resolving power. However in many instruments the information being sought is necessarily obtained in a distorted form. When the quantity being sought is measured as a function of another parameter, e.g., as in the case of spectral information or distribution functions, quite often this filtering process can be expressed as the convolution of the function being measured with another function, which represents the characteristic distortion produced by a certain instrument. The criteria necessary for this definition to be applicable are discussed in section II of this thesis. The process of convolution for functions of one variable can be expressed in integral form as is shown in section III. Emslie and King[1] and Frei and Gunthard[2] discuss the representation of instrumental distortions by convolutions. Roseller[3] and Schrack[4] are typical of works which calculate the convolutions of some common spectral functions.

The more difficult process of "deconvolution" is the one with which the experimenter is often faced when he must analyze recorded data. It is not uncommon for the convolution process of measurement to have completely masked details or fine structure in the measured function. An example of this can be found in spectral measurements in which fine structures were discovered when instruments with greater resolving power were first used. In order to recover these masked details or fine structure one must reverse the convolution process or "de-smear" the data. This can be done by solving the convolution integral equation when the characteristic distortion function of the measuring apparatus is known. However, the actual implementation of deconvolution is a non-trivial operation, and perhaps for this reason it has not yet been widely applied for the interpretation of experimental data.

1.2 Numerical Deconvolution

There are many methods for solving the convolution integral equation, but they are generally numerical techniques, because rarely is the apparatus function known in closed form and almost never is the recorded output of an instrument defined analytically. It is shown in section IV that the problem can be formulated numerically in terms of linear simultaneous equations, or a matrix equation. Direct methods can be used[5,6,7] to solve the simultaneous equations but these methods generally are not successful for higher-order systems. This is due to the fact that the system of equations is ill-conditioned[8,9,10] resulting from the fact that the coefficients are points on a continuous apparatus function.[11]

Because of this ill-conditioning, iterative techniques have been widely used to approximate the desired solution.

One of the earliest iterative techniques to be applied to deconvolution is a method of successive approximations due to van Cittert,[12] discussed by Burger and van Cittert,[13,14] and called the simplest method of steepest descent by Tal.[11] Much literature is available on the use of van Cittert's iteration for deconvolution.[6,15-21] Another iterative technique which can be used for deconvolution is the Jacobi iteration (see for example Ralston[22]). The Jacobi iteration can be 'over-relaxed" resulting in Von Mises' iteration,[23] which can always be made to converge for a positive definite coefficient matrix. Another iterative technique, which is generally favored over the Jacobi iteration because of its more rapid convergence, is called the Gauss-Seidel iteration and seems to be due to Seidel.[24] A symmetric or double sweep version of this iteration was reported by Aitken.[25] Both the Gauss-Seidel and its modification by Aitken can be shown to converge for a positive definite coefficient matrix.[26,23] The Gauss-Seidel method can also be 'over-relaxed' to hasten convergence. This technique is called successive over-relaxation or SOR, and is credited to Young.[27]

Also there are the "gradient" iterative techniques derived from the minimization of the quadratic functional. The standard method of steepest descent first proposed by Temple[28] and discussed by Stiefel[29] and Tal[11] can be used for deconvolution and will always converge for a positive definite coefficient matrix.[11] An accelerated steepest descent method[30] was also used by Tal.[11] The method of conjugate gradients is a special modification of the steepest descent method, and was first used by Hestenes and Stiefel.[31] It is actually a direct method since it theoretically converges in a finite number of steps. However, in practice due to round-off errors the iteration is continued until the desired accuracy is obtained. These iterative techniques are discussed in section VI.

By Fourier transforming the convolution integral equation a simple algebraic equation can be obtained which is easily transformed to give the desired solution. In actual calculations a Fourier series is used and only a finite number of terms are considered. The Fourier transform method has become quite popular and much literature is available relating to the use of the Fourier transform method for deconvolution.[6,21,32,41] Another method for deconvolution which is easily derived from the Fourier transform approach is called the 'derivative method.'[42-46] The Fourier transform and derivative methods are discussed in section V of this work.

Another deconvolution technique, which should be noted because of its increasing popularity, is a technique in which the system of linear equations defining the convolution process is overspecified and then some method of least squares fitting is used to solve for the unknowns.[39,40,47-51] Both direct and iterative techniques have been used to calculate the least squares fit.

Deconvolution using the properties of the eigenvalues and eigenfunctions of the convolution integral operator is discussed in section VII. This method of deconvolution has been used by several workers[16,52-54] to analyze the deconvolution process. In actual calculations one deals with the eigenvalues and eigenvectors of the coefficient matrix.

Perhaps at this point some general references on numerical deconvolution should be noted. An excellent review of iterative techniques is found in Martin and Tee,[23] and many books on the topic of numerical analysis (see for example, Ralston,[22] Hildebrand,[55] and John[56]) discuss them in some detail. Several more references on numerical techniques are available;[57-62] and, Mikusinski[63] and Berg[64] talk about convolution transforms in terms of operational calculus.

1.3 Errors in Deconvolution

Since the deconvolution process is prone to noise or errors due to the ill-conditioning or instability of the equations, many of the works on numerical deconvolution draw attention to the effects of errors and/or discuss the problem theoretically.[5,16,19,32,33,36,50,51,53,54,65-67] Worth particular mention are the works of Rautian[32] and Rushforth and Harris[54] on this topic. In this vein as well, several authors discuss convolution and deconvolution in terms of information theory.[1,68-70] The effects of noise and errors are discussed in section VIII of this work.

1.4 Deconvolution by Analog Methods

Up to this point the methods listed have been of the numerical type which are usually implemented with the aid of a high-speed digital computer. There also has been a fair amount of work using electronic analog devices to effect deconvolution. The work relating to analog methods can be broken down into two general areas. Analog devices which simply convolute comprise the first area, which will be called indirect methods. In order to deconvolute with an analog device which is capable only of convolution, the accepted technique is to have a trained operator who proposes a solution, convolutes it with the apparatus function, and then checks to see if it matches the problem to be deconvoluted. If it doesn't, another solution is proposed and the process is repeated. This continues until the operator is convinced that the solution fits. This amounts to having a human in a feed back loop, and, unfortunately there is no way in which a solution so obtained can be mathematically demonstrated to be unique. In the area of indirect analog methods, French, et al[71] and Noble, et al[72] presented instruments which essentially sum curves in order to simulate the convolution process; and Profos[73] reported an electro-mechanical analog device for convolution. Diamantides[74] and Kindlemann[75] developed high speed electronic correlation computers which are somewhat digital in nature; Zverev and Orlaf,[76] and Breton and Hirschberg[77] used electro-optical devices for performing convolutions.

The second area is composed of analog devices which use a direct method for deconvolution; i.e., devices whose operation can be shown to be mathematically equivalent to one of the direct or iterative numerical techniques. In the area of direct methods, the simplest approach is to use standard analog computer techniques to solve simultaneous equations (see for example Goldberg.[78]) Dolby,[79,80] and Dolby and Cosslett[81] used this approach to the problem; but, for more than a few unknowns (they had only three) the circuitry becomes formidable and somewhat unstable, again due to the ill-conditioning of the system of equations. Allen, Gladney, and Glarum[82] and Glarum[83] used analog methods for resolution enhancement which have for a mathematical basis the derivative method described earlier. Zörner[46] proposed an instrument using a special magnetic tape apparatus whose basis was also the derivative method. Krishnamurty,[84] and Korsunskii and Genkin[85] reported the use of analog devices which were based upon iterative numerical techniques. The most sophisticated direct analog methods are found in the automatic analog devices proposed and developed by Kendall.[86-90] In a patent, Kendall[88] suggested the modification of many of the indirect analog methods (e.g. the optical methods of convolution [76,76]) in order to have them deconvolute automatically using a direct method. A later electrostatic analog computer was proposed by Kendall, developed by Zabielski,[91] and reported by Kendall and Zabielski.[92] Section IX of this work will be devoted to a description of this instrument, its operation and performance, and various experimental results. It will be shown that this instrument performs a modified symmetric Gauss-Seidel iteration, and the effects of various parameters on convergence will be demonstrated.

CHAPTER II

STATEMENT OF THE PROBLEM

The general problem is that of recovering information which has been transformed in the process of its measurement or detection in a characteristic manner which can be described by a linear convolution operation. For the measuring or detection process to be shown to be a linear operation, the law of superposition must apply (e.g. an input which is a sum of several distinct signals must result in an output which is the sum of the outputs which each one of the distinct inputs would produce separately), and the characteristic distortion function or apparatus function must remain the same to within a multiplicative constant over the entire region of the independent variables over which the problem is defined.

The specific goal of this work is to review and evaluate available techniques for performing deconvolution. Results obtained using the RM-5 analog device which has been developed in this laboratory will be reported, along with a mathematical description of its operation.

CHAPTER III

THE CONVOLUTION INTEGRAL

3.1 Some Properties of the Convolution Integral

Assuming that the broadening or distortion of spectral data can be described as the convolution of the true data input to an instrument and the characteristic distortion function of the instrument, the problem becomes one of solving the integral equation which represents the convolution process. Equation (1) describes the convolution of two functions A(x) and T(x).

$$F(x) = \int_{-\infty}^{\infty} A(x-x') \; T(x') \; dx' \qquad (1)$$

F(x) represents the recorded output data from an instrument; T(x') describes the true input data to the instrument, and A(x-x') the characteristic function of the instrument. Equation (1) can be thought of as representing a transformation which takes a function T(x') in the x' space and maps it onto another space spanned by x, resulting in F(x). It is instructive to examine some special cases of equation (1). If one lets

$$A(x-x') = \delta(x-x') \qquad (2)$$

where δ is the Dirac delta function; then one finds

$$F(x) = T(x) \qquad (3)$$

This means that equation (2) describes the "ideal" instrument; i.e., one which simply reproduces the input data exactly. Further if one lets

$$T(x') = \delta(x') \tag{4}$$

the result is:

$$F(x) = A(x) \tag{5}$$

and in this case the output of the instrument, F(x), is just the characteristic function of the instrument, A(x), or its response to a delta function input. Another interesting property of the convolution integral can be found by integrating equation (1) with respect to x

$$\int_{-\infty}^{\infty} F(x)dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} A(x-x')\ T(x')dx'dx \tag{6}$$

and rewriting equation (6) as:

$$\int_{-\infty}^{\infty} F(x)dx = \int_{-\infty}^{\infty} A(z)dz \int_{-\infty}^{\infty} T(x')dx' \tag{7}$$

Now if one requires A(x-x') to be normalized; i.e.,

$$\int_{-\infty}^{\infty} A(x)dx = 1 \tag{8}$$

then equation (7) reduces to

$$\int_{-\infty}^{\infty} F(x)dx = \int_{-\infty}^{\infty} T(x')dx' \tag{9}$$

what this implies is that if $A(x-x')$ is normalized, then area is preserved by the transformation of equation 1. All of the above properties are well known[63].

3.2 Real Convolution Problems with Finite Limits

In all of the practical cases with which this report will deal, $F(x)$ will be defined over some finite interval (α,β), and be positive definite and continuous over that interval; which implies that $F(x)$ goes to zero at $x=\alpha$ and $x=\beta$, and is identically zero for $X\leq\alpha$ and $x\geq\beta$. This, along with the fact that A and T will be assumed to be functions defined in the same manner as F, further implies that equation 1 becomes:

$$F(x)= \int_{a}^{b} A(x-x') \; T(x') \; dx' \tag{10}$$

which is immediately recognized as a special case of the Fredholm Equation of the First Kind, which is usually written:

$$F(x)= \int_{a}^{b} A(x,x') \; T(x') \; dx' \tag{11}$$

Figure 1 illustrates, schematically, a typical convolution as described by equation (10).

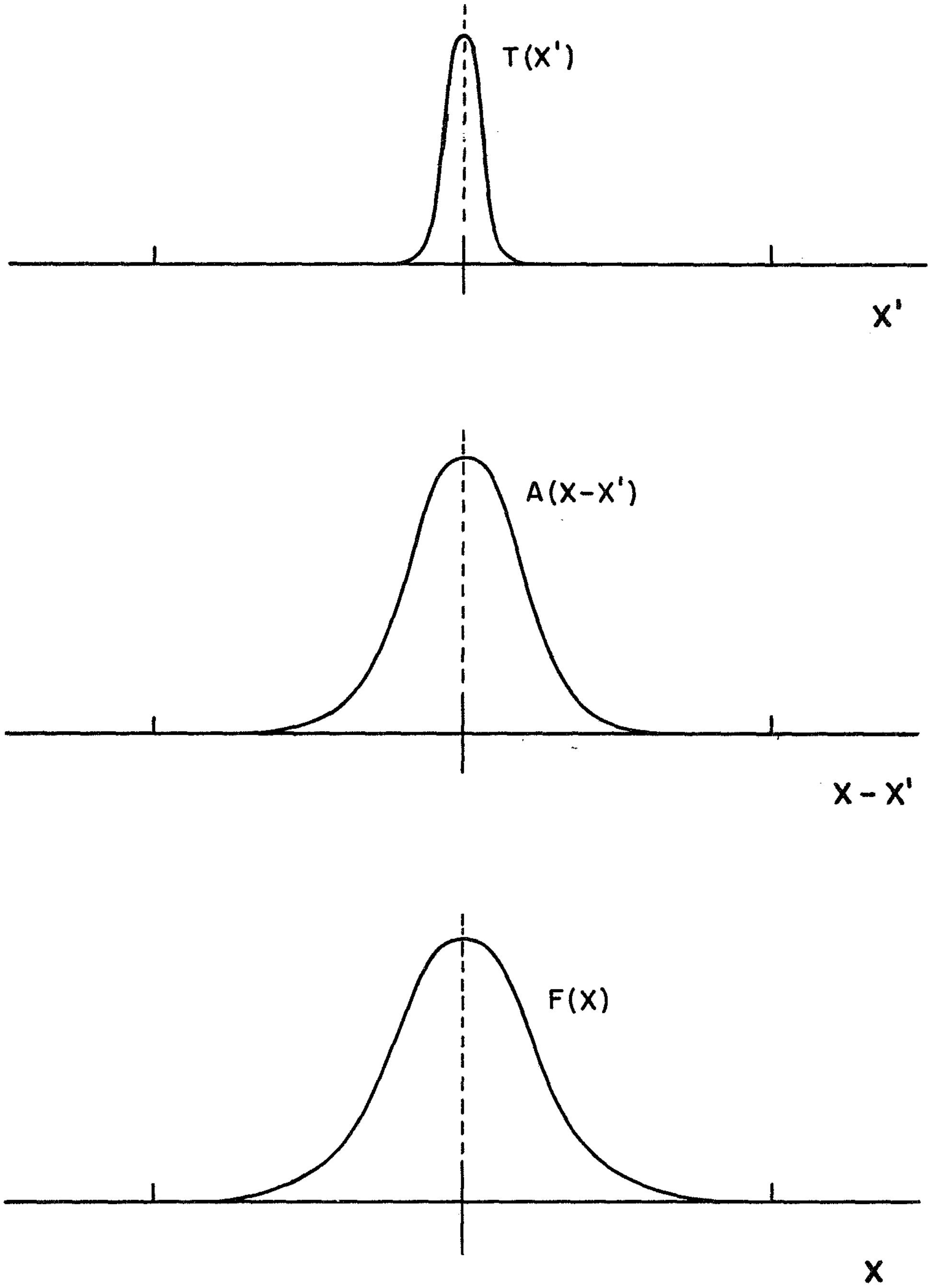


Figure 1

A schematic example of the convolution process.

CHAPTER IV

NUMERICAL FORMULATION OF THE PROBLEM

In all but very special practical cases, the apparatus function $A(x-x')$ is not known in closed form, but rather is obtained by recording the response of the instrument to a δ function input (i.e., an input which is as nearly like a δ function as is practical). Then the problem becomes one of solving equation (10) numerically when $F(x)$ and $A\ (x-x')$ are known. To solve the problem numerically, one must take a finite number of points to describe $F(x)$ and $A(x-x')$ and also solve for a finite number of points which will describe $T(x')$. If one takes N equally spaced points in the interval (a,b) over which $F(x)$ is defined, then letting

$$x_n=a+[(b-a)/(N-1)]\ (n-1) \quad n=1\ \dots\ N \tag{12}$$

one can define a set $\{F_n\}$ of N points which represent $F(x)$, where

$$F_n \equiv F(x_n) \tag{13}$$

Then if one similarly defines

$$x'_n=a+[(b-a)/(N-1)]\,(n-1);\ n=1,\cdots,N \quad [1] \tag{14}$$

and

$$T_n=T(x'_n)\ , \tag{15}$$

[1] If $A(x)$ is symmetric about $x=0$ and $F(x)\in(a,b)$, then $T(x)\in(a,b)$

one can also define

$$A_{nm}=A(x_n-x_m) \tag{16}$$

Substituting the results of equations (13), (15), and (16) into equation (10) and replacing the integral by a sum, one obtains:

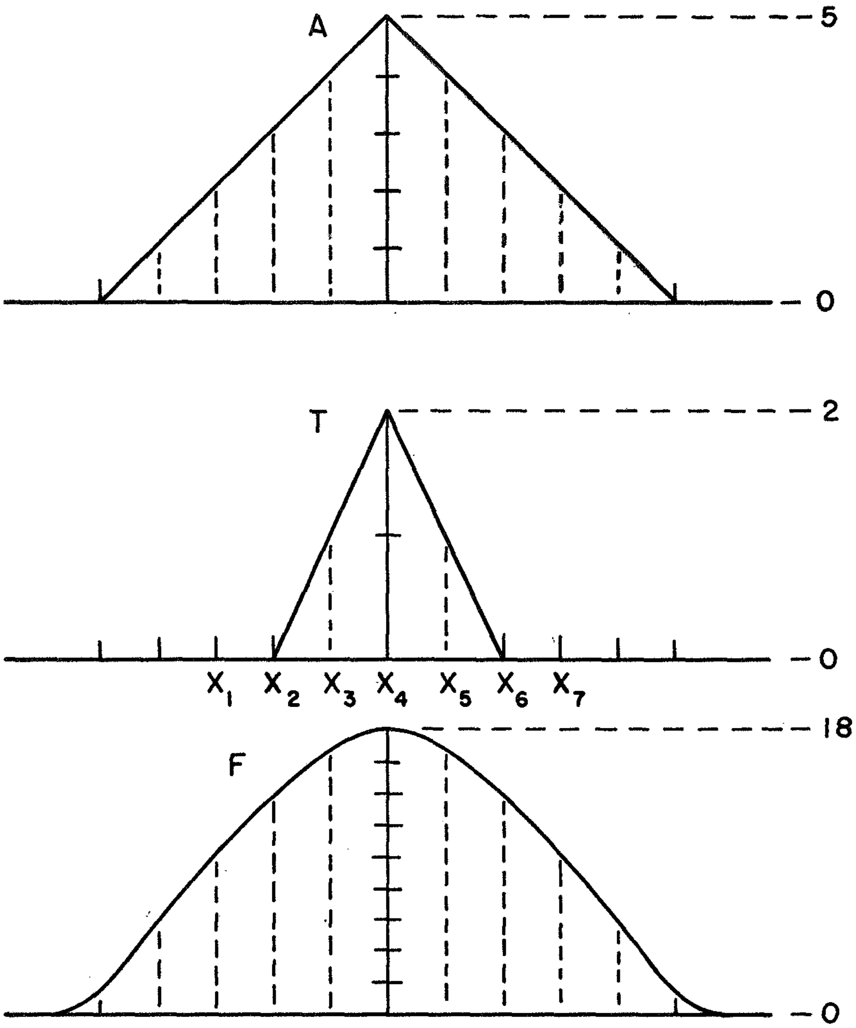
$$F_n=K\sum_{m=1}^{N} A_{nm}T_m \qquad [2] \tag{17}$$

But equation (17) simply defines the matrix equation which is commonly used to describe a system of N linear simultaneous equations in N unknowns. The T_m are the N unknown elements of a N x 1 matrix (vector), the A_{nm} are the elements of the N x N coefficient matrix, and the F_n are the elements of a N x 1 matrix (vector) which is known. Equation (17) can be written in standard matrix form:

$$Ax=b \tag{18}$$

where the elements of A are $A_{nm,}$ those of x are T_m, and those of b are F_n. Figure 2 illustrates this procedure for a typical problem. (Note that x in equation (18) now represents an unknown vector, not an independent variable as it did previously.)

[2] K is a scale factor equal to the increment between points of T(x). For simplicity it will be assumed equal to one.



$$\begin{pmatrix} 5 & 4 & 3 & 2 & 1 & 0 & 0 \\ 4 & 5 & 4 & 3 & 2 & 1 & 0 \\ 3 & 4 & 5 & 4 & 3 & 2 & 1 \\ 2 & 3 & 4 & 5 & 4 & 3 & 2 \\ 1 & 2 & 3 & 4 & 5 & 4 & 3 \\ 0 & 1 & 2 & 3 & 4 & 5 & 4 \\ 0 & 0 & 1 & 2 & 3 & 4 & 5 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \\ X_7 \end{pmatrix} = \begin{pmatrix} 8 \\ 12 \\ 16 \\ 18 \\ 16 \\ 12 \\ 8 \end{pmatrix}$$

FIGURE 2

An example of convolution done numerically using discrete points.

Theoretically now the problem is just that of inverting an N x N matrix, for as one multiplies equation (18) by the inverse of A, A^{-1} the result is:

$$x = A^{-1}b \tag{19}$$

However, as soon as the order of A gets moderately large (N=30 to 40), A becomes more nearly singular; i.e.,

$$\lim_{N \to \infty} (\det A) = 0 \tag{20}$$

due to the continuous nature of A(x-x')[11,53,65,66], and direct methods break down due to finite round-off error in any numerical calculation. This problem will be discussed in detail in a later section. The fact that det A is small means that the system of equations is "ill-conditioned"[8-10]. For solving ill-conditioned systems, iterative techniques are generally the most successful.

CHAPTER V

DECONVOLUTION BASED ON FOURIER TECHNIQUES

5.1 The Fourier Transform or Series Approach

Many workers[6,21,32,41] have reported on the application of the Fourier transform to solve the convolution integral. Using an approach similar to that of Rautian[32], if one Fourier-transforms equation (1),

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} F(x)e^{i\omega x}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} A\ (x-x')T(x')dx'\right]e^{i\omega x}dx$$

Upon manipulation this becomes:

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} F(x)e^{i\omega x}dx = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} A(x-x')e^{i\omega(x-x')}dx \quad T(x')e^{i\omega x'}dx'$$

which can be further reduced to

$$f(\omega) = \int_{-\infty}^{\infty} a(\omega)\ T(x')e^{i\omega x'}dx'$$

and finally

$$f(\omega) = \sqrt{2\pi}\ a(\omega)\ t(\omega) \tag{21}$$

where $f(\omega)$, $a(\omega)$, and $t(\omega)$ represent, respectively, the Fourier transforms of $F(x)$, $A(x)$, and $T(x)$. Solving equation (21) for $t(\omega)$ and doing an inverse Fourier transform, the result is

$$T(x) = \frac{1}{2\pi} \int \frac{f(\omega)}{a(\omega)} e^{-i\omega x} d\omega \qquad (22)$$

Hence, in theory, one needs only to Fourier transform $F(x)$ and $A(x)$, and then use equation (22) to obtain $T(x)$. However, in practice, this technique is found to be highly susceptible to noise[32,36] and only moderate resolution enhancement (factors of 2 to 4) is usually obtained by this method.

A similar treatment can be used for analyzing equation (1) in terms of a discrete Fourier expansion. If one starts with

$$F(x) = \sum_{n=-\infty}^{\infty} f_n e^{inx}$$

$$A(x-x') = \sum_{n=-\infty}^{\infty} a_n e^{in(x-x')}$$

$$T(x') = \sum_{n=-\infty}^{\infty} t_n e^{inx'} \qquad (23)$$

where $F(x)$, $A(x)$, and $T(x)$ are defined only on the interval $(-\pi,\pi)$ on x, then one can find the relation:

$$t_n = \frac{f_n}{2\pi a_n} \qquad (24)$$

in analogy with equation (21). This general technique can be used to analyze equation (1) in terms of any linear integral transform or expand the functions in terms of any complete orthonormal set of functions.

5.2 The Derivative Method

Another method for deconvolution which is frequently reported in the literature, the 'Derivative Method'[42,46] is derived from the Fourier transform method. First one expands the reciprocal of the Fourier transform of A(x) in a Taylor series of the form

$$\frac{1}{a(\omega)} = \sum_{n=0}^{\infty} c_n (i\omega)^n \tag{25}$$

and then this result is substituted into equation 22:

$$T(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \sum_{n=0}^{\infty} c_n (i\omega)^n \, f(\omega) \, e^{-i\omega x} d\omega \tag{26}$$

Next, using the definition of f(ω), equation (26) can be rewritten

$$T(x) = \frac{1}{(2\pi)^{3/2}} \int_{-\infty}^{\infty} \sum_{n=0}^{\infty} c_n (i\omega)^n \left[\int_{-\infty}^{\infty} F(x')e^{i\omega x'} dx' \right] e^{-i\omega x} d\omega$$

or rearranging

$$T(x) = \frac{1}{\sqrt{2\pi}} \sum_{n=0}^{\infty} c_n \int_{-\infty}^{\infty} \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} (i\omega)^n e^{i\omega(x'-x)} d\omega \right] F(x') dx' \tag{27}$$

But the term in brackets in equation (27) can be expressed as the n'th derivative of the Dirac delta function[46]:

$$\delta^{(n)}(x'-x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} (i\omega)^n e^{i\omega(x-x')} d\omega \tag{28}$$

With this information, equation (27) becomes

$$T(x) = \frac{1}{\sqrt{2\pi}} \sum_{n=0}^{\infty} c_n F^{(n)}(x) \tag{29}$$

which expresses T(x) in terms of the derivatives of F(x), hence the name 'Derivative Method'. The constants $\{c_n\}$ can be expressed in terms of the moments of the apparatus function. Equation (25) can be written:

$$\sum_{n=0}^{\infty} c_n (i\omega)^n = \frac{1}{\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} A(x) e^{i\omega x} dx} \tag{30}$$

Then using the power series expansion for an exponential, equation (31) is obtained

$$\sum_{n=0}^{\infty} c_n (i\omega)^n = \frac{\sqrt{2\pi}}{\sum_{n=0}^{\infty} \left[\frac{1}{n!}\int_{-\infty}^{\infty} x^n A(x) dx\right] (i\omega)^n} \tag{31}$$

from which the c_n can be evaluated in terms of the moments of A(x) (see, for example, Zörner[46]). It can be easily seen that this method also is prone to noise, as the process of differentiation enhances noise; and hence in practice the series for T(x) in equation (29) must be truncated before the noise level becomes intolerable.

CHAPTER VI

NUMERICAL ITERATIVE TECHNIQUES

6.1 The Jacobi and Gauss-Seidel Iterations

A numerical iterative method which is commonly used to solve ill-conditioned equations of the form shown in equation (18) uses the following formulae:

$$x_1^{(i+1)} = \frac{1}{a_{11}} (b_1 - a_{12}x_2^{(i)} - a_{13}x_3^{(i)} - \ldots - a_{1N}x_N^{(i)}$$

$$x_2^{(i+1)} = \frac{1}{a_{22}} (b_2 - a_{21}x_1^{(i)} - a_{23}x_3^{(i)} - \ldots - a_{2N}x_N^{(i)}$$

$$\vdots$$

$$x_N^{(i+1)} = \frac{1}{a_{NN}} (b_N - a_{N1}x_1^{(i)} - a_{N2}x_2^{(i)} - \ldots - a_{NN-1}x_{N-1}^{(i)} \qquad (32)$$

In equations (32), the superscript denotes the iteration number and the subscript denotes the component of the vector x. The iterative method defined by equations (32) is sometimes called the Jacobi iteration[22].

Another iterative technique which is commonly called the Gauss-Seidel technique[24] uses formulae similar to equations (32) with the modification that the most recently calculated value of each x_i is used when calculating the new value of any component, x_j.

$$x_1^{(i+1)} = \frac{1}{a_{11}} (b_1 - a_{12}x_2^{(i)} - a_{13}x_3^{(i)} - \ldots - a_{1N}x_N^{(i)})$$

$$x_2^{(i+1)} = \frac{1}{a_{22}} (b_2 - a_{21}x_1^{(i+1)} - a_{23}x_3^{(i)} - \ldots - a_{2N}x_N^{(i)})$$

$$\vdots$$

$$x_N^{(i+1)} = \frac{1}{a_{NN}} (b_N - a_{N1}x_1^{(i+1)} - a_{N2}x_2^{(i+1)} - \ldots - a_{N,N-1}x_{N-1}^{(i+1)}) \tag{33}$$

Equations (33) define the Gauss-Seidel technique. These iterative methods, which will be referred to as the Jacobi iteration for equations (32) and the Gauss-Seidel iteration for equations (33), may or may not converge depending upon the form of the coefficient matrix A in equation (18). A general criterion for convergence is that a matrix in which the main diagonal dominates; i.e., one in which the largest term in any row is the term on the main diagonal, has a better probability for convergence. However, a more explicit definition of the convergence criteria is desirable. If one rewrites equations (32) as shown in equation (34)

$$x_j^{(i+1)} = \frac{1}{a_{jj}} (b_j - \sum_{k=1}^{j-1} a_{jk}x_k^{(i)} - \sum_{k=j+1}^{N} a_{jk}x_k^{(i)}) \tag{34}$$

and equation (33) as shown in equation (35)

$$x_j^{(i+1)} = \frac{1}{a_{jj}} (b_j - \sum_{k=1}^{j-1} a_{jk}x_k^{(i+1)} - \sum_{k=j+1}^{N} a_{jk}x_k^{(i)}) \qquad (35)$$

it becomes immediately obvious that equations (34) and (35) can be rewritten in matrix form as shown in equations (36) and (37) respectively[22].

$$x^{(i+1)} = D^{-1}(b-Lx^{(i)} - Ux^{(i)}) \qquad (36)$$

Jacobi

$$x^{(i+1)} = D^{-1}(b-Lx^{(i+1)} - Ux^{(i)}) \qquad (37)$$

Gauss-Seidel

The elements of the matrices D,L, and U are defined in equations (38); and this decomposition of the matrix A is shown schematically in Figure 3.

$$d_{ij} = \begin{matrix} 0 & ; & i \neq j \\ a_{ii}; & & i = j \end{matrix}$$

$$\ell_{ij} = \begin{matrix} 0 & ; & i \geq j \\ a_{ij}; & & i < j \end{matrix}$$

$$u_{ij} = \begin{matrix} 0 & ; & i \leq j \\ a_{ij}; & & i > j \end{matrix} \qquad (38)$$

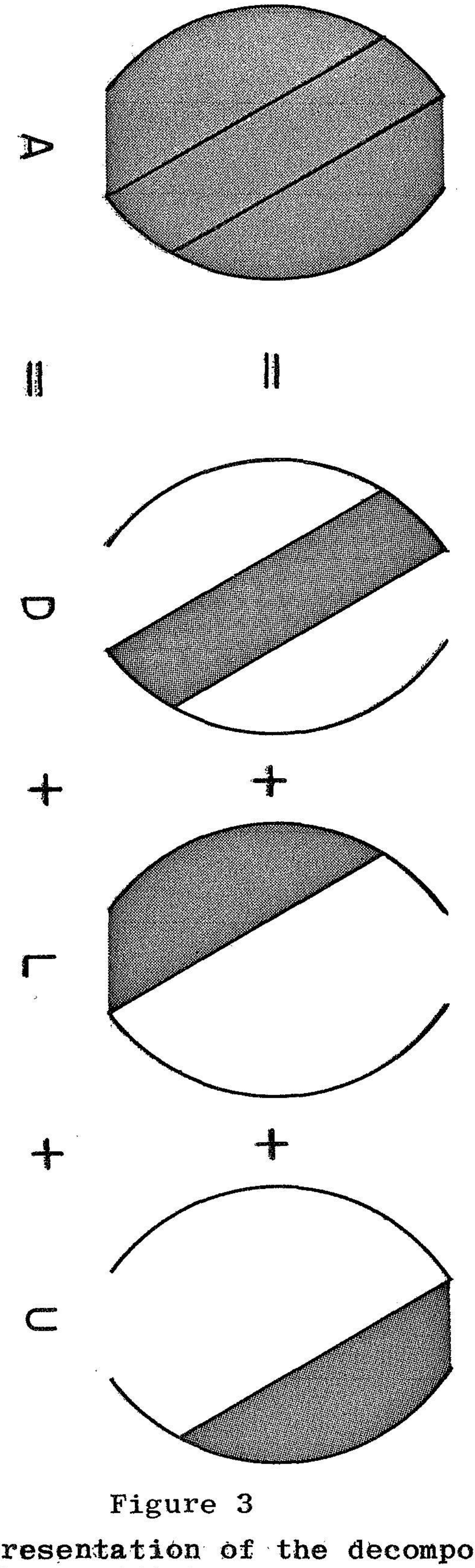


Figure 3

Graphic representation of the decomposition of the coefficient matrix A

Solving equations (36) and (37) for $x^{(i+1)}$, one obtains equations (39) and (40)[22].

$$x^{(i+1)} = D^{-1}(b - Ux^{(i)} - Lx^{(i)}) \tag{39}$$

Jacobi

$$x^{(i+1)} = (D + L)^{-1} (b - Ux^{(i)}) \tag{40}$$

Gauss-Seidel

Now as one further lets $x^{(0)} = 0$ as is the general practice with these iterative techniques, unless one has a good approximate solution with which to begin, equations (39) and (40) can be used to express $x^{(i+1)}$ in terms of a power series for the Jacobi and Gauss-Seidel iterations respectively.

$$x^{(i+1)} = \left[\sum_{n=0}^{i} (-D^{-1} \left[L + U\right])^{n} \right] D^{-1}b \tag{41}$$

Jacobi

$$x^{(i+1)} = \left[\sum_{n=0}^{i} (-\left[D + L\right]^{-1}U)^{n} \right] (D + L)^{-1}b \tag{42}$$

Gauss-Seidel

Both of these series will converge if and only if the true norm, as defined in appendix I, of the matrix term which is raised to the power, in each of them, is < 1 in analogy to an algebraic power series. A sufficient but not necessary criterion for convergence is that the Euclidian norm, as defined by equation (43), for any matrix E

$$N_E(E) \equiv \left| \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{N} e_{ij}^2} \right| \tag{43}$$

be < 1[56]. One can also use the Euclidian norm of a matrix to put bounds on the true norm[56] as shown in expression (44).

$$N_E(E) \geq N_T(E) \geq \frac{1}{\sqrt{M}} N_E(E) \tag{44}$$

where M is the order of the matrix. However, it can be shown that the true norm of a matrix is equal to the magnitude of its largest eigenvalue (see Appendix I). Further if the Jacobi-iteration converges for a system of equations, then the Gauss-Seidel iteration usually will also converge and will do so more rapidly[22]. Hence the Jacobi iteration is rarely used in favor of the Gauss-Seidel iteration.

The Gauss-Seidel iteration will always converge for a system of equations if the coefficient matrix A is positive definite[26]. A matrix A, is said to be positive definite if it satisfies the condition expressed by equations (45)[26]:

$$A = B^T B$$

or

$$\left|a_{11}\right|, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \quad \ldots \quad , \begin{vmatrix} a_{11} & \ldots & a_{iN} \\ \vdots & & \\ a_{N1} & \ldots & a_{NN} \end{vmatrix} \text{ all} > 0 \tag{45}$$

If one is working with a system of equations as defined by equation (18) in which the coefficient matrix is not positive definite, the system can be transformed by multiplying equation (18) by A^T from the left, as shown in equation (46):

$$A^T Ax = A^T b \tag{46}$$

Equation (46) can be written as

$$A'x = b' \tag{47}$$

where $A' \equiv A^T A$, and is positive definite by definition, and $b' = A^T b$. Then one can proceed to use the Gauss-Seidel iteration to solve equation (47) for x. It is interesting to note that a coefficient matrix A which satisfies equation (45) corresponds to an apparatus function which can be expressed as the convolution of some function B(x) with itself; i.e.,

$$A(x) = \int_{-\infty}^{\infty} B(x-x') \, B(x') \, dx' \qquad (48)$$

Another way of stating this is that A(x) is the autocorrelation function of some other function. Some of the common apparatus functions which fall into this category are the Gaussian, Lorentzian, and triangular functions.

6.2 Relaxed Versions of the Jacobi and the Gauss-Seidel Methods

The Jacobi iteration can be over-relaxed to hasten convergence. The new iteration is called the von Mises' iteration[23]. If one rewrites equation (36) in the following form

$$x^{(i+1)} = x^{(i)} + \left[D^{-1}(b - Lx^{(i)} - Ux^{(i)}) - x^{(i)}\right] \quad (49),$$

von Mises' iteration is a simple modification as shown in equation (50):

$$x^{(i+1)} = x^{(i)} + \beta \left[D^{-1}(b - Lx^{(i)} - Ux^{(i)}) - x^{(i)}\right] \qquad (50)$$

where β can be adjusted to guarantee convergence for a positive definite matrix[23]. The SOR method[27] is an over relaxed version of the Gauss-Seidel technique and is derived from the Gauss-Seidel formula in the same manner as above, resulting in equation (51)

$$x^{(i+1)} = x^{(i)} + \beta \left[D^{-1}(b - Lx^{(i+1)} - Ux^{(i)}) - x^{(i)} \right] \tag{51}$$

The SOR method will converge for a certain range of values of β when A is a positive definite matrix, and the rate of convergence will be a maximum for some value of β[23].

6.3 Steepest Descent Methods

The simplest method of steepest descent[11-14] or method of successive approximations for solving equation (18) where A is a positive definite coefficient matrix, uses the formula :

$$x_{n+1} = x_n + \alpha r_n \tag{52}$$

where

$$r_n = b - Ax_n \tag{53}$$

and α is a constant which must satisfy the inequality

$$0 < \alpha < \frac{2}{\lambda_{max}} \tag{54}$$

where λ_{max} is the largest eigenvalue of the matrix A, in order to guarantee convergence[11].

The standard steepest descent method[28,29] uses the formula :

$$x_{n+1} = x_n + \alpha_n r_n \tag{55}$$

where

$$r_n = b - Ax_n \tag{56}$$

and

$$\alpha_n = \frac{r_n^T r_n}{r_n^T A r_n} \tag{57}$$

For both of the above methods it can be shown that the convergence is fastest in the directions of the eigenvectors corresponding to the largest eigenvalues of the matrix A, which has the effect of damping out the oscillatory parts of a solution[11]. In general the methods of steepest descent converge very slowly.

It should be noted that the standard steepest descent method is derived from the minimization of the quadratic functional Q[22]:

$$Q(x) = \frac{1}{2} x^T Ax - x^T b \tag{58}$$

when A is positive definite. This form has a unique minimum when x is the solution of equation (18).[22] To derive the standard steepest descent method, consider the gradient of the functional Q evaluated at $x = x_i$:

$$\nabla_x Q \Big|_{x=x_i} = Ax_i - b = -r_i \tag{59}$$

Next an α_i is calculated such that $Q(x_i-\alpha_i r_i)$ is a minimum:

$$Q(x_i-\alpha_i r_i) = -\frac{1}{2} x_i^T r_i + \alpha_i r_i^T r_i + \frac{1}{2} \alpha_i^2 r_i^T A r_i - \frac{1}{2} x_i^T b \tag{60}$$

$$\frac{\partial Q}{\partial \alpha_i} = r_i^T r_i + \alpha_i r_i^T A r_i = 0 \tag{61}$$

$$\alpha_i = \frac{-r_i^T r_i}{r_i^T A r_i} \tag{62}$$

which yields the formulae for the standard steepest descent method.

An accelerated version of the steepest descent[30] method uses the following formulae

$$x_{n+1} = x_n + \alpha_1 r_n + \alpha_2 A r_n + \ldots + \alpha_p A^{p-1} r_n \tag{63}$$

where

$$r_n = b - Ax_n \tag{64}$$

and the $\{\alpha_i\}$ are determined by requiring that the functional $(x^T Ax - x^T x)$ is a minimum; i.e., by solving the following p equations for the α_i's:

$$r_n{}^T A^{j-1} r_n + \sum_{k=1}^{p} \alpha_k r_n{}^T A^{j+k-1} r_n = 0 \tag{65}$$

for j=1,2,...,p

Tal[11] found that large values of p speeded up convergence so much that no smooth approximate solution could be obtained, but that an optimum value of p could be determined experimentally for a particular type of problem.

6.4 Conjugate Gradients Method

The method of conjugate gradients[31] is similar to the method of steepest descent with the exception that each successive correction vector is calculated with the additional requirement that the residuals $\{r_i\}$ will form an orthogonal set of N vectors, or some $r_i=0$ for $i < N$, in which case the solution has been obtained. The formulae are:

$$x_{i+1} = x_i + \alpha_i V_i \tag{66}$$

$$\alpha_i = V_i{}^T r_i \Big/ V_i{}^T A V_i \tag{67}$$

$$V_{i+1} = r_{i+1} + \beta_i V_i \tag{68}$$

$$\beta_i = -V_i^T A r_{i+1} \Big/ V_i^T A V_i \qquad (69)$$

$$r_{i+1} = r_i - \alpha_i A V_i \qquad (70)$$

In practice, one generally lets $x_o=0$ and then $v_o=r_o=b-x_o=b$. The sequence then becomes:

a. Calculate α_1 from equation (67)

b. Calculate x_1 from equation (66)

c. Calculate r_1 from equation (70)

d. Calculate β_1 from equation (69)

e. Calculate V_1 from equation (68)

f. Calculate α_1 from equation (67)

g. Calculate x_2 from equation (66)

h. Repeat steps c. thru g. increasing all subscripts by one.

Since the r_i's form an orthogonal set, N of them will completely span the N-dimensional space in which x is represented; and, hence this technique in theory will converge to the proper solution in a finite number (N) of iterations. In practice, due to round-off error, more than N iterations are usually needed to obtain a sufficiently converged solution[11].

CHAPTER VII

THE EIGENVALUE APPROACH

Another method which can be used to solve equation (10) for T(x') utilizes the properties of eigenvalues and eigenfunctions. If one solves the characteristic equation

$$\varphi_i(x) = \frac{1}{\lambda_i} \int_a^b A(x-x') \; \varphi_i(x')dx' \qquad (71)$$

for the eigenvalues $\{\lambda_i\}$ and the corresponding complete orthonormal set of eigenfunctions $\left\{\varphi_i(x)\right\}$ on the interval (a,b), which will result for a positive definite A(x-x') which is symmetric about x = x', then one can expand functions on that interval in terms of the set $\left\{\varphi_i(x)\right\}$. If one assumes that both T(x) and F(x) of equation (10) can be represented by:

$$T(x') = \sum_i t_i\varphi_i(x') \qquad (72)$$

$$F(x) = \sum_i f_i\varphi_i(x) \qquad (73)$$

then the problem is one of solving for the relationships between the t_i's and f_i's. Firstly, the values of f_i can be found by multiplying equation (73) by $\varphi_j(x)$ and integrating with respect to x over the interval (a,b). By using the properties of orthonormal functions it can be shown that f_i is given by equation (74).

$$f_i = \int_a^b F(x)\varphi_i(x)dx \qquad (74)$$

Then by substituting from equations (72) and (73) into equation (10) and using equation (71),

$$\sum_i f_i\varphi_i(x) = \int_a^b A(x-x')\left[\sum_j t_j\varphi_j(x')\right] dx'$$

$$\sum_i f_i\varphi_i(x) = \sum_j t_j\lambda_j \left[\frac{1}{\lambda_j}\int_a^b A(x-x')\varphi_j(x')dx'\right]$$

the result is equation (75),

$$\sum_i f_i\varphi_i(x) = \sum_j t_j\lambda_j\varphi_j(x) \qquad (75)$$

from which one deduces that for all values of i,

$$t_i = \frac{f_i}{\lambda_i} \qquad (76)$$

After finding the values of t_i, T(x') is constructed using equation (72).

To actually implement this technique numerically, one must first formulate the matrix equation, of the form shown in equation (18), and then solve equation (77)

$$Ay_i = \lambda_i y_i \qquad (77)$$

for N eigenvalues, $\{\lambda_i\}$, and their corresponding eigenvectors, $\{y_i\}$. Since A will always be assumed to be a positive definite N x N matrix, the eigenvalues will all be real, distinct, and positive. However, due to the ill-condition of A, there will be many eigenvalues near zero. Further, the eigenvectors corresponding to these small eigenvalues will be highly oscillatory in nature[11]. Using a completely analogous development, one can finally show that x in equation (18) is given by equation (78).

$$x = \sum_{i=1}^{N} \frac{b^T y_i}{\lambda_i} y_i \qquad (78)$$

Further, if one indexes the eigenvalues in descending order $(\lambda_1 > \lambda_2 > . . . > \lambda_n)$, then by truncating the series of equation (78), an approximate solution of equation (18) can be obtained. This truncation effectively filters x or cuts out the higher spatial frequency components.

CHAPTER VIII

THEORETICAL ANALYSIS OF ERRORS IN THE DECONVOLUTION PROCESS

In most practical deconvolution problems, the function F(x) in equation (10) is known only within experimental errors and may also contain noise due to any number of sources. If one assumes for the moment that the apparatus function A(x-x') is known exactly, equation (10) can be modified to include these effects as shown in equation (79)[54].

$$F(x) + N(x) = \int_a^b A(x-x')T'(x')dx' \qquad (79)$$

N(x) is the function which accounts for all of the errors in the observed output of the instrument; i.e.,

$$F_{obs}(x) = F(x) + N(x) \qquad (80)$$

and T'(x) is the solution one will obtain when the observed instrument output $F_{obs}(x)$ is deconvoluted. The difference between T'(x) and T(x) then will be the error in the obtained solution due to the error term N(x) in the observed output of the instrument F(x). By using the techniques of section VII these errors can be analyzed[54].

If one lets F(x) be represented by the expansion in equation (73), and T'(x') and N(x) be represented by the expansions as defined by equations (81) and (82);

$$T'(x') = \sum_i t_i' \, \varphi_i(x') \tag{81}$$

$$N(x) = \sum_i n_i \, \varphi_i(x) \tag{82}$$

then, by substituting these expansions into equation (79) and using the properties of eigenfunctions, the following result is obtained[54]:

$$t_i' = \frac{f_i}{\lambda_i} + \frac{n_i}{\lambda_i} \tag{83}$$

Using equation (76), equation (83) can be rewritten

$$t_i' = t_i + \frac{n_i}{\lambda_i} \tag{84}$$

where t_i is defined by equation (72). This result can be stated in another way:

$$T'(x) = T(x) + \eta(x) \tag{85}$$

where $\eta(x)$ is defined in equation (86),

$$\eta(x) = \sum_i \frac{n_i}{\lambda_i} \varphi_i(x) \tag{86}$$

and represents the resulting error in the solution.

The same problem can also be analyzed in terms of the Fourier transform method as described by Rautian[32]. If equation (79) is Fourier transformed, the resulting equation is:

$$f(\omega) + n(\omega) = \sqrt{2\pi}\, a(\omega)t'(\omega) \qquad (87)$$

where f(ω), n(ω), a(ω), and t'(ω) represent the Fourier transforms of the functions F(x), N(x), A(x), and T'(x) respectively. Using equation (21) to substitute for f(ω) in equation (87), one obtains:

$$t'(\omega) = t(\omega) + \frac{n(\omega)}{\sqrt{2\pi}\, a(\omega)} \qquad (88)$$

By performing an inverse Fourier transform on equation (88) the result is

$$T'(x) = T(x) + \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{n(\omega)}{a(\omega)} e^{-i\omega x}\, d\omega \qquad (89)$$

in analogy to equation (85). This provides an equivalent definition of $\eta(x)$ as shown in equation (90).

$$\eta(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{n(\omega)}{a(\omega)} e^{-i\omega x}\, d\omega \qquad (90)$$

When dealing with errors of a random nature whose mean is zero, one characterizes the errors by their mean squared values. Rushforth and Harris[54] demonstrated that the total mean square error in the solution, which is obtained by deconvoluting a noisy problem, is given in the eigenvalue-eigenfunction representation as

$$\overline{e^2} = \sum_i \frac{\overline{n_i^2}}{\lambda_i^2} \tag{91}$$

where

$$\overline{n_i n_j} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \varphi_i(x) \; \varphi_j(u) \; R_N(x,u)dudx \tag{92}$$

and $R_N(x,u)$ is the autocorrelation of $N(x)$. Rautian[32] analyzed the total mean squared error in an approximate solution obtained by putting finite limits on the integral when performing the inverse Fourier transform on equation (88). This amounts to frequency limiting or band pass filtering the restored solution. The reason for this approximate approach is found in the fact that it can be shown that $\eta(x)$ is unbounded[32,54] unless one truncates the series in equation (86) or, equivalently, one limits the range of ω for the integration in equation (90). However, limiting the range of ω in order to bound the noise introduces another error, since this prevents exact reconstruction of the true solution; or, in effect, this limits the amount of resolution enhancement which is possible. The following expression describes the total mean squared error[32].

$$\overline{e^2} = \left| \frac{1}{2\pi} \int_{\omega_o}^{\infty} \frac{f(\omega)}{a(\omega)} \, e^{-i\omega x} \, d\omega + \frac{1}{2\pi} \int_{-\infty}^{-\omega_o} \frac{f(\omega)}{a(\omega)} \, e^{-i\omega x} \, d\omega \right|^2$$

$$+ \frac{1}{2\pi} \int_{-\omega_o}^{\omega_o} \frac{S_n(\omega)}{|a(\omega)|^2} \, d\omega \qquad (93)$$

The first term on the right hand side of equation (93) represents the error due to filtering the true solution and the second term represents the error due to noise. $S_N(\omega)$ is the noise power spectrum and is defined as the Fourier transform of $R_N(x,x)$. Further, in general, the first term is a monotonically decreasing function of ω, and the second is a monotonically increasing function of ω so that the effects are competing and a minimum value of $\overline{e^2}$ can be found, for some value of ω_o.

CHAPTER IX

The RM-5 Analog Device

9.1 Description of Device and its Operation

The RM-5 analog device is basically the same instrument as that reported previously by Zabielski[91]. It is an iterative analog device. A block diagram of the instrument is shown in Figure 4. The instrument is automatically self-correcting through the use of a feedback loop. During the deconvolution process, each time the wiper at point D in Figure 4 moves to the next memory channel, the error signal at point A - which is the difference between the convoluted trial solution on the capacitive memory and the problem to be deconvoluted - is amplified, inverted, and fed back to that memory channel until the error signal at point A is reduced to zero (for infinite amplifier gain); i.e., it is essentially a system with 100% negative feedback. This is the basic operating principle. One iteration consists of sweeping through the entire memory adjusting each channel in turn. The iterative procedure is continued until the potential distribution on the memory no longer changes from one iteration to the next, at which time a stable solution has been reached.

While the above discussion gives a general description of operation of the RM-5 , a more detailed mathematical description is desirable. Assume for the moment that

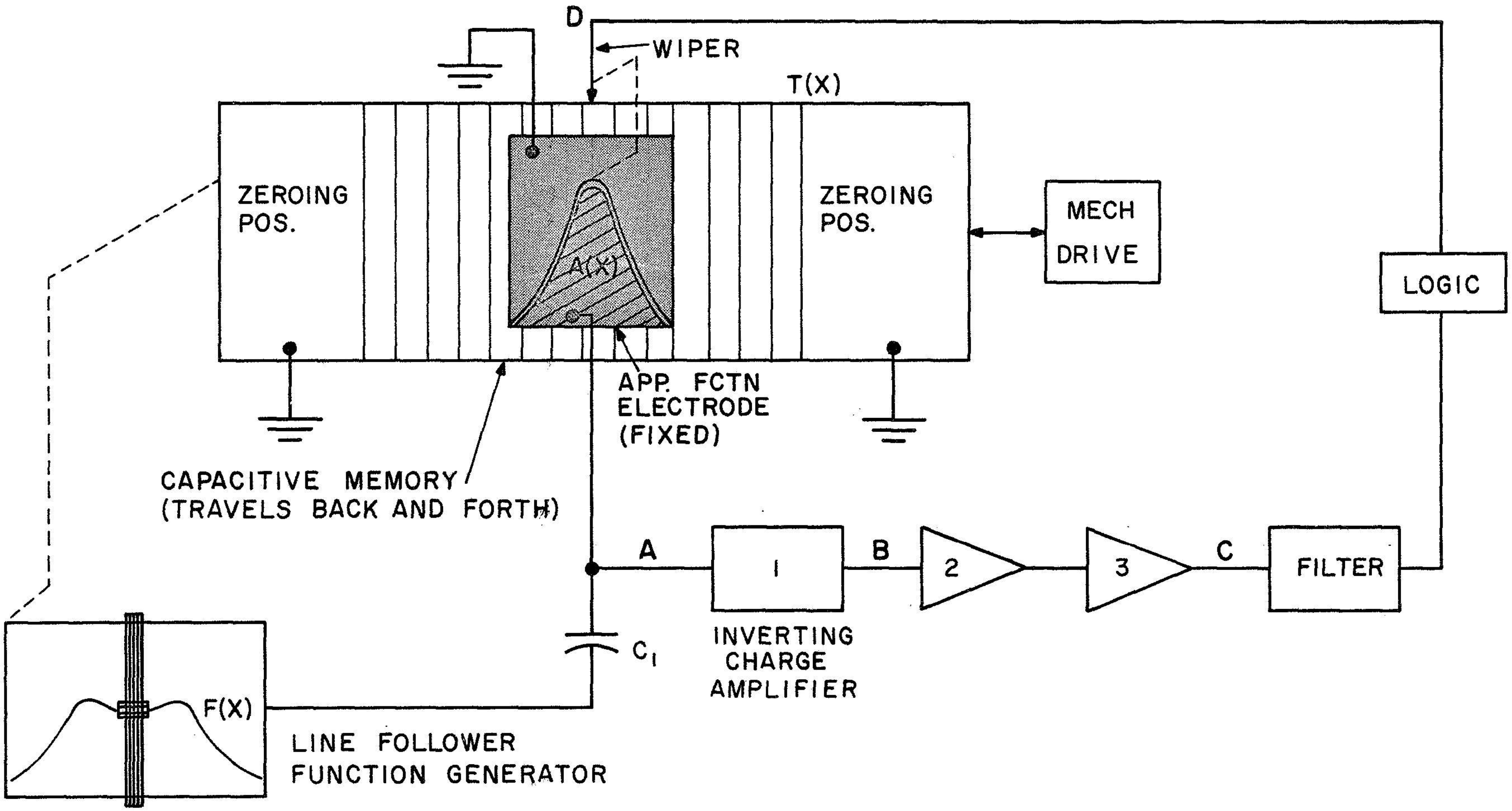

D
WIPER
T(X)
ZEROING POS.
A(X)
ZEROING POS.
MECH DRIVE
LOGIC
APP. FCTN ELECTRODE (FIXED)
CAPACITIVE MEMORY (TRAVELS BACK AND FORTH)
A
1
B
2
3
C
FILTER
C_1
INVERTING CHARGE AMPLIFIER
F(X)
LINE FOLLOWER FUNCTION GENERATOR

FIGURE 4

Block diagram of the RM-5 analog device.

the wiper is contacting the jth memory channel (indexing from left to right). Then the charge induced on the shaped electrode which represents the apparatus function A(x) can be represented by[91]

$$\begin{matrix}\text{charge induced}\\ \text{on electrode}\end{matrix} = K \sum_{i=1}^{N} a_{ji}x_i \qquad (94)$$

where

$$a_{ji} = A(\left[j-i\right] \Delta\xi), \qquad (95)$$

$\Delta\xi$ is the spacing between the centers of any two adjacent memory channels and is a constant, x_i is the voltage on the i th memory channel, and N is the number of channels. The constant K in equation (94) is a scale factor which is determined by the capacitance between the shaped electrode and the capacitive memory, and the spacing between memory channels, $\Delta\xi$; both of which are constants. Hence K will be assumed equal to one as was done previously. This will not affect the validity of the following derivation. Now if one further assumes that this is the (n+1)th time the potential distribution on the memory is being adjusted and that the memory is being cycled from left to right (i.e., the memory is moving from right to left); then using superscripts to denote iteration number, equation (94) can be rewritten (letting K=1).

$$\begin{matrix}\text{charge induced}\\ \text{on electrode}\end{matrix} = \sum_{i=1}^{j} a_{ji} x_i^{(n+1)} + \sum_{i=j+1}^{N} a_{ji} x_i^{(n)} \qquad (96)$$

Now similarly the charge induced at point A in Figure 4 due to the line follower function generator, feeding through c_1, can be written

$$\begin{matrix}\text{charge induced}\\ \text{from line follower}\end{matrix} = -k\ b_j \qquad (97)$$

where

$$b_j = F(j \cdot \Delta\xi) \qquad (98)$$

and F(x) represents the problem to be deconvoluted. Again the factor k in equation (97) is simply a scale factor and will be assumed equal to one to simplify the mathematics. The total charge induced at point A is then:

$$\begin{matrix}\text{total charge}\\ \text{electrode}\end{matrix} = \text{charge induced on} + \begin{matrix}\text{charge induced from}\\ \text{line follower}\end{matrix} \qquad (99)$$

or

$$\text{total charge} = \sum_{i=1}^{j} a_{ji} x_i^{(n+1} + \sum_{i=j+1}^{N} a_{ji} x_i^{(n)} - b_j \ . \qquad (100)$$

At point C in Figure 4, the voltage can be given by

$$V_c = G\,(b_j - \sum_{i=1}^{j} a_{ji}x_i^{(n+1)} - \sum_{i=j+1}^{N} a_{ji}x_i^{(n)}) \qquad (101)$$

since all of the amplifiers are inverting. G is the total loop gain of the system. Now if one assumes that the filter and logic do nothing in the simplest case, then one can say (c.f. Figure 4)

$$V_c = V_D \qquad (102)$$

But V is just $x_j^{(n+1)}$, so that the loop equation for this simple system becomes

$$x_j^{(n+1)} = G(b_j - \sum_{i=1}^{j} a_{ji}x_i^{(n+1)} - \sum_{i=j+1}^{N} a_{ji}x_i^{(n)}) \qquad (103)$$

Equation (103) can be solved for $x_j^{(n+1)}$ resulting in equation (104).

$$x_j^{(n+1)} = \frac{G}{1+a_{jj}G}\,(b_j - \sum_{i=1}^{j-1} a_{ji}x_i^{(n+1)} - \sum_{i=j+1}^{N} a_{ji}x_i^{(n)}) \qquad (104)$$

It can immediately be seen that the limit of equation (104) as G goes to infinity is simply equation (35)

$$x_j^{(n+1)} = \frac{1}{a_{jj}} \left(b_j - \sum_{i=1}^{j-1} a_{ji} x_i^{(n+1)} - \sum_{i=j+1}^{N} a_{ji} x_i^{(n)}\right) \qquad (35)$$

which describes the Gauss-Seidel iteration (c.f. section 6.1). This demonstrates that the Gauss-Seidel iteration is the basis for the operation of the RM-5. Now equation (104) can be further modified to more accurately describe the actual operation of the RM-5.

First, since the capacitive memory moves in different directions during alternate iterations, the iteration becomes a symmetric iteration which can be represented by

$$x_j^{(n+1)} = \begin{cases} \dfrac{G}{1+a_{jj}G} \left(b_j - \displaystyle\sum_{i=1}^{j-1} a_{ji} x_i^{(n+1)} - \sum_{i=j+1}^{N} a_{ji} x_i^{(n)}\right); & \text{n odd} \\ \dfrac{G}{1+a_{jj}G} \left(b_j - \displaystyle\sum_{i=1}^{j-1} a_{ji} x_i^{(n)} - \sum_{i=j+1}^{N} a_{ji} x_i^{(n+1)}\right); & \text{n even} \end{cases} \qquad (105)$$

For G going to infinity, this iteration reduces to the symmetric form of the Gauss-Seidel iteration[25], which converges for a positive definite coefficient matrix[23,25].

The next necessary modification is one which will describe the action of a low pass filter in the feedback loop (between points C and D in Figure 4). Since the filters in the RM-5 circuit are actually several low pass R-C sections, a very simple mathematical model can be used as a first approximation. This simple model is based upon the following argument. Assume that the wiper at point D in Figure 4 is contacting the (j-1)th channel and is at a voltage x_{j-1}. Now assume that the memory is moved so that the wiper is contacting the jth channel. The amplifiers now see a different error signal and begin to adjust the voltage x_j to its equilibrium value. However due to the R-C filtering action in the feedback loop, this adjustment of the voltage x_j will take a certain amount of time, determined by the R-C time constant of the filter. But at the same time, the memory is moving at a steady speed and the wiper will only contact the jth channel for a fixed time period before it moves to the (j+1)th position. Hence, during that time, the voltage x_j will not quite reach its equilibrium value; but will reach some fraction (less than 1) of it, which will be determined by the time constant of the R-C filter. This model can be represented very simply in mathematical form by

$$x_j^{(n+1)} = \begin{cases} \alpha(y - x_{j-1}^{(n+1)}) + x_{j-1}^{(n+1)}; & n \text{ odd} \\ \alpha(y' - x_{j+1}^{(n+1)}) + x_{j+1}^{(n+1)}; & n \text{ even} \end{cases} \qquad (106)$$

where

$$y = \frac{G}{1+a_{jj}G}\left(b_j - \sum_{i=1}^{j-1} a_{ji}x_i^{(n+1)} - \sum_{i=j+1}^{N} a_{ji}x_i^{(n)}\right) \qquad (107)$$

and

$$y' = \frac{G}{1+a_{jj}G} \ (b - \sum_{i=1}^{j-1} a_{ji}x_i^{(n)} - \sum_{i=j+1}^{N} a_{ji}x_i^{(n+1)}) \qquad (108)$$

which are the equilibrium values of $x_j^{(n+1)}$, as would be calculated from equation (105) in the absence of filtering. The parameter α in equation (106) determines the amount of filtering, and is related to the R-C time constant in the following way:

$$\alpha = e^{-\tau/\Delta t} \qquad (109)$$

where Δt is the time which the wiper at point D in Figure 4 spends contacting each channel in the memory. For $\tau = 0$, which is equivalent to no filtering, $\alpha=1$; and equation (106) reduces to equation (105). For $\tau > 0$, $\alpha<1$ and hence the smaller is α, the larger is the filtering action.

The final modification which is required to complete the mathematical description is one which will describe the action of the logic between points C and D in Figure 4. This logic is simply a diode clipping circuit which passes voltages of only one polarity. The reason for the inclusion of this circuit in the feedback loop is related to the type of problem which is to be deconvoluted. In most types of spectral measurements, the quantity being measured (light intensity, rf absorption, mass abundance, etc.) is positive by definition (i.e. ≥ 0); and a solution which has negative values is physically unreal. For this reason the logic was included. This "negative rejection" principle can be added to the mathematical description in a simple manner as shown in equation (110).

$$x_j^{(n+1)} = \begin{cases} Z\ H(Z)\ ; & n \text{ odd} \\ Z'H(Z')\ ; & n \text{ even} \end{cases} \qquad (110)$$

where

$$Z = \alpha(y - x_{j-1}^{(n+1)}) + x_{j-1}^{(n+1)} \qquad (111)$$

$$Z' = \alpha(y' - x_{j+1}^{(n+1)}) + x_{j+1}^{(n+1)} \qquad (112)$$

and H(Z) is the Heaviside step function defined by

$$H(Z) = \begin{matrix} 0;\ Z < 0 \\ 1;\ Z \geq 0 \end{matrix} \qquad (113)$$

Equation (110) represents the mathematical description of the operation of the RM-5 analog device.

9.2 Digital Simulation of the RM-5

Using equation (110), the RM-5 was digitally simulated by the use of a Fortran IV program, which was run on an IBM 360 computer, model 67. The basic program accepted as input data the following:

1. The problem to be deconvoluted, F(x), in a one dimensional array.

2. The apparatus function, A(x), in a one dimensional array.
3. The gain, G.
4. The parameter α, which determines the amount of filtering.
5. The number of iterations to be performed.

The output of the basic program consists of the following items:

1. The input data
2. The solution after the required number of iterations have been performed.

9.3 Experimental Results from the RM-5

Figure 5 shows a Gaussian curve which was used as the apparatus function, A(x), and the problem F(x), for a series of experiments with the RM-5. The true solution for this deconvolution problem, is a δ function singlet. Figure 6 shows the solution which was obtained from the RM-5 for varying amounts of filtering in the feedback loop. Curve a in Figure 6 was the solution obtained with the least amount of filtering, and curves b, c, and d, each, were obtained with more filtering than the previous curve. Examination of Figure 6 shows that there exists an optimum value of filtering corresponding to curve b, which gives the best approximation to a δ function singlet. The resolution enhancement obtained in curve b is ~ 8, using the ratio of the width at half height of Figure 5 to the

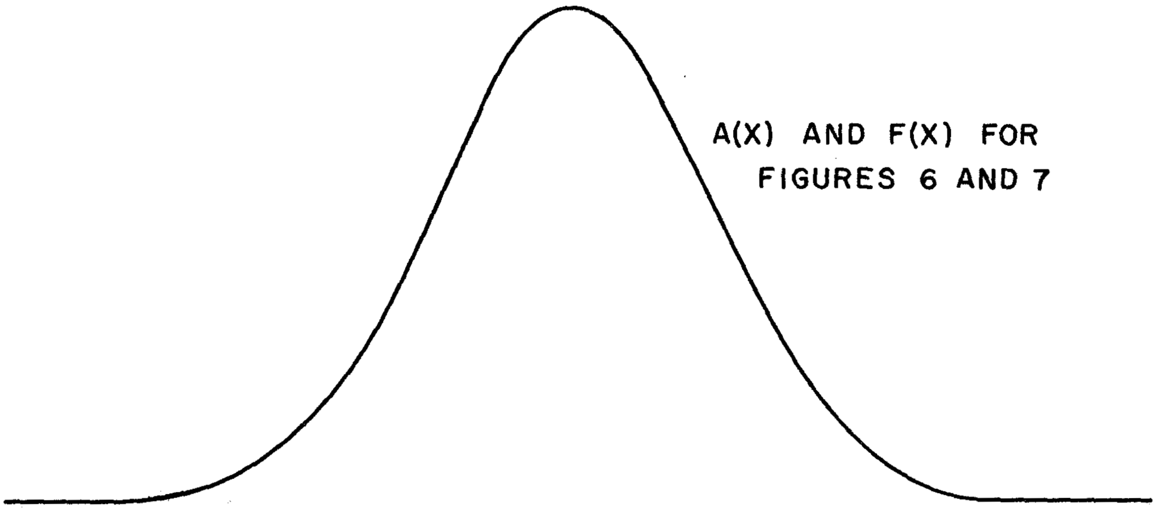


Figure 5

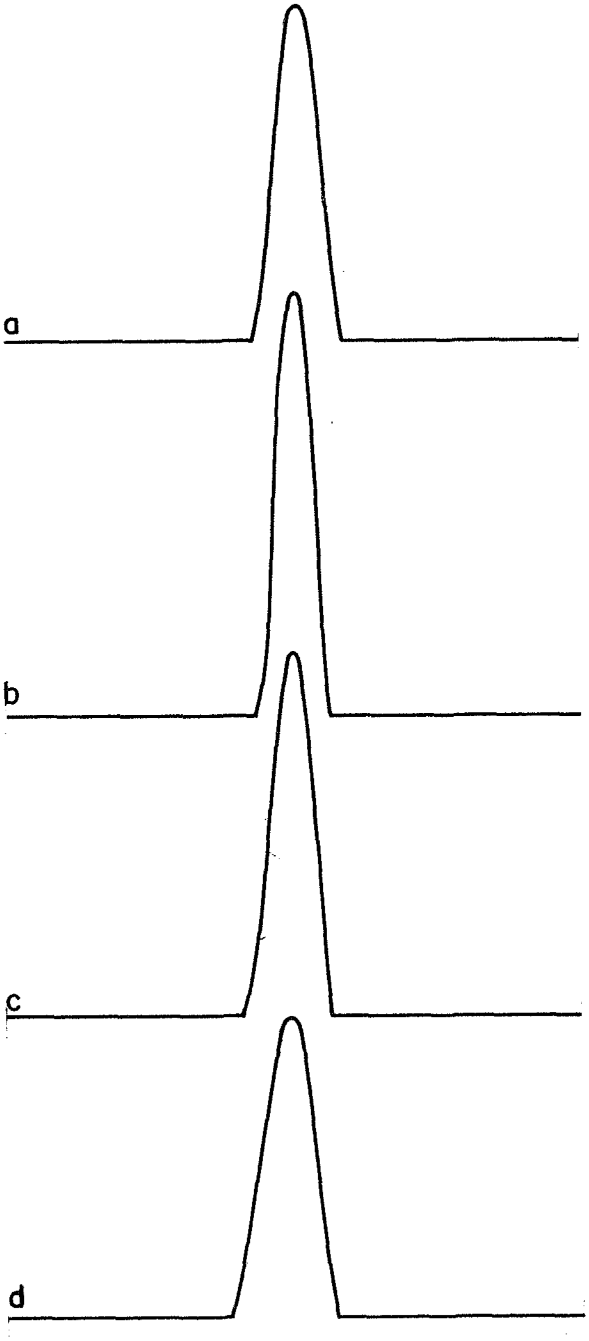


Figure 6

Approximate solutions [T approx (x)] obtained from the RM-5 with varying amounts of filtering for problem shown in Figure 5 width at half height of Figure 6b. The gain, G, as defined in section 9.1 was 4.2 for all of the curves in Figure 6. Figure 7 shows the results obtained from the RM-5 for various values of gain, G, when deconvoluting the problem in Figure 5. The amount of filtering is the same as that corresponding to Figure 6b for all of the curves in Figure 7. Figure 7 clearly demonstrates that it is desirable to have the gain as large as possible, as is generally true of any null seeking analog device. Although $G = 4.2$ is not a large value of gain, it is limited in the RM-5 device by the gain of the first amplifier in Figure 4. This amplifier can be considered a charge amplifier if one assumes that the input is the shaped electrode; but, if one considers the input to be the voltage on the memory channel which the wiper is contacting, then this amplifier can be considered to be a voltage amplifier, and the voltage gain which is calculated in this manner is the number which is needed (along with the voltage gain of the second and third amplifiers in Figure 4) in order to calculate the total loop gain G as discussed in section 9.1. When the voltage gain of the first amplifier is calculated in this manner, it is found to be $\sim 1/50$, depending upon the air gap between the shaped electrode and the capacitive memory, as shown in Figure 4. Hence, even though large voltage gains can be obtained with amplifiers 2 and 3 in Figure 4, the total loop gain is somewhat limited, since any noise which appears on the output of amplifier 1, is amplified by amplifiers 2 and 3.

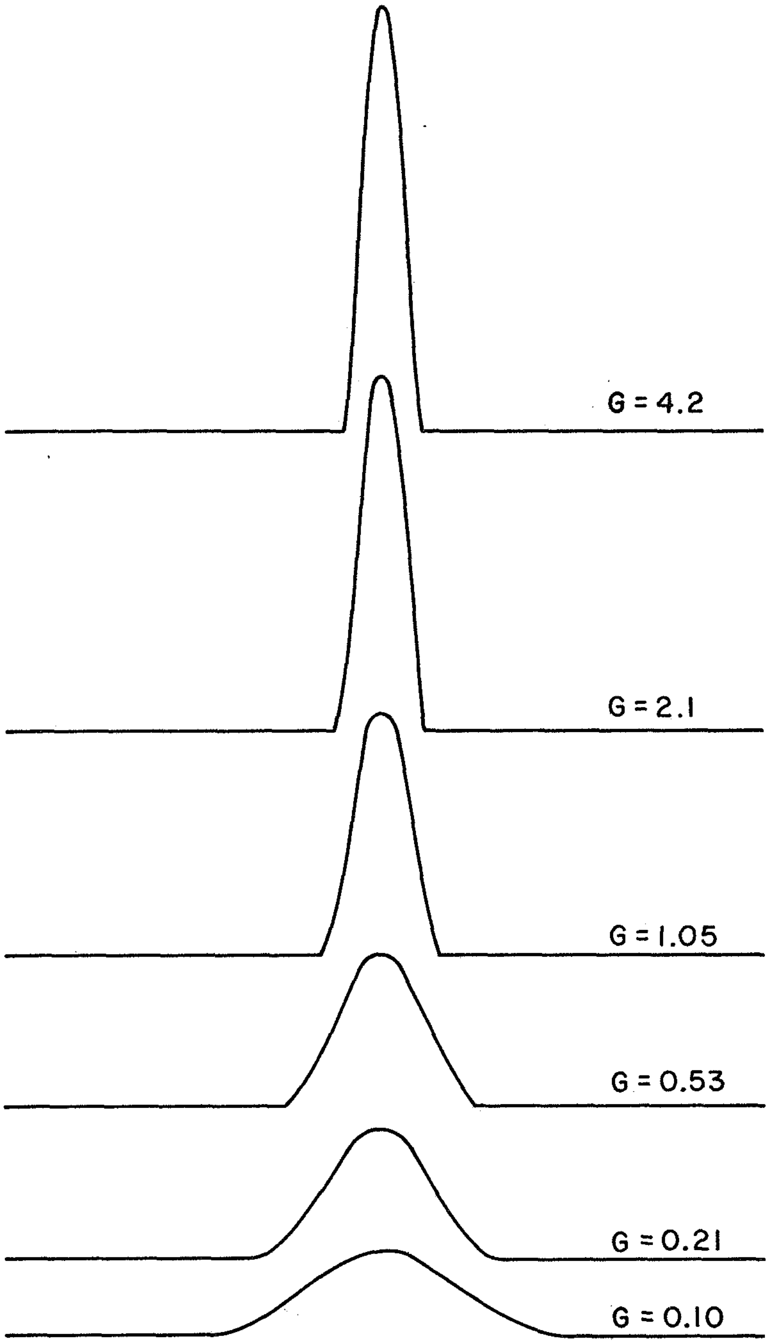


Figure 7

Approximate solutions [T approx (x)] obtained from the RM-5 with varying values of gain for the problem shown in Figure 5

Figures 8 through 12 show some typical results from the RM-5 for various types of problems. Figure 8 shows the apparatus function A(x) and its deconvolution to a δ function singlet. This apparatus function is not defined in closed form, but rather was drawn by hand in an attempt to construct an arbitrary apparatus function. Figure 9 shows the deconvolution of a problem whose true solution, T(x), is a δ function doublet with equal amplitudes and a spacing of 1.86 inches. Figure 10 shows the deconvolution of a problem whose true solution is a δ function doublet with equal amplitudes and a spacing of 1.57 inches. Figure 11 shows the result of deconvoluting a problem whose true solution is a δ function doublet with amplitudes having a ratio 1:2 and a spacing of 1.57 inches; and, Figure 12 is the result of deconvoluting a problem whose true solution is a δ function triplet with all amplitudes equal and equal spacings of 1.30 inches.

9.4 Results from the Digital Simulation of the RM-5

In order to check whether equation (110) is an adequate mathematical description of the operation of the RM-5 analog device, a series of problems were deconvoluted numerically using an IBM 360 computer and the Fortran IV program briefly described in section 9.2. The effects of varying the parameters G and α, as defined in section 9.1, were also investigated. Figure 13 shows a Gaussian doublet problem which was used to investigate the effect of varying G in equation (110). A(x) and F(x) were calculated exactly

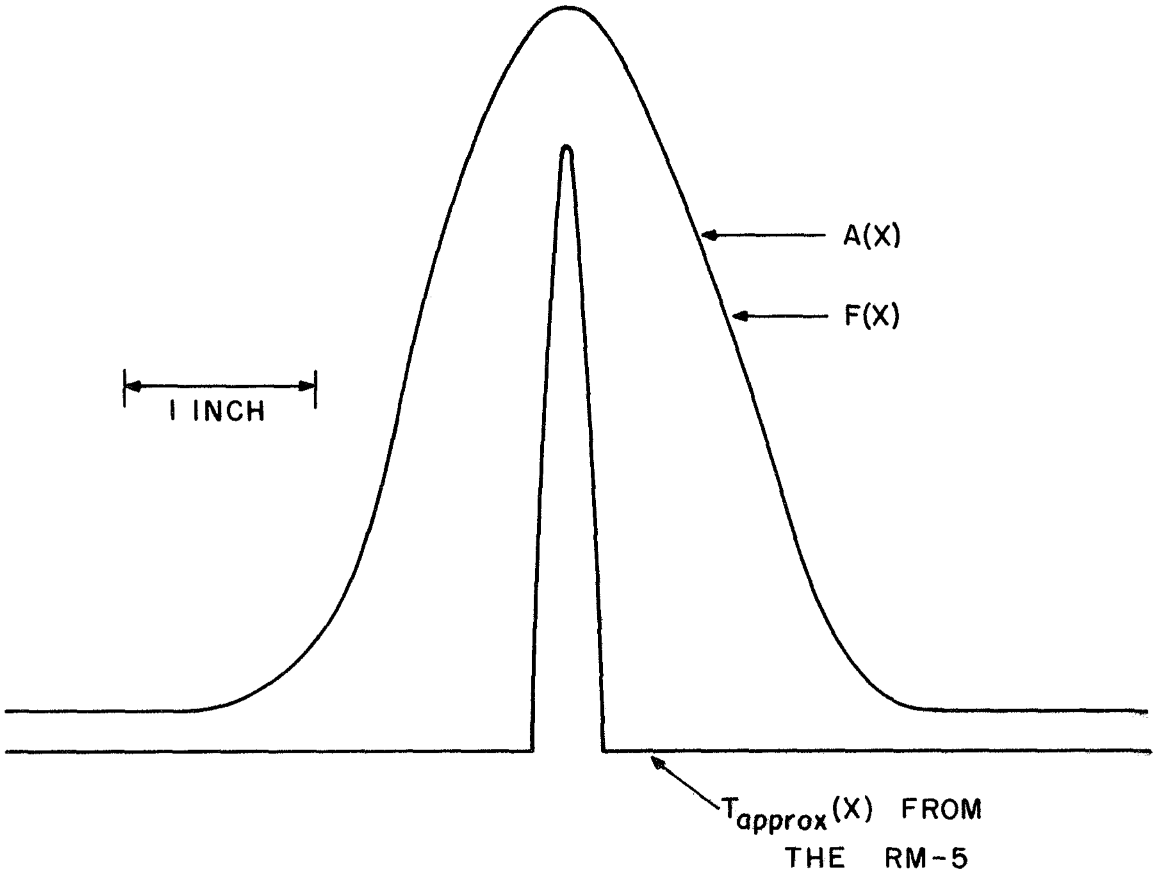


Figure 8

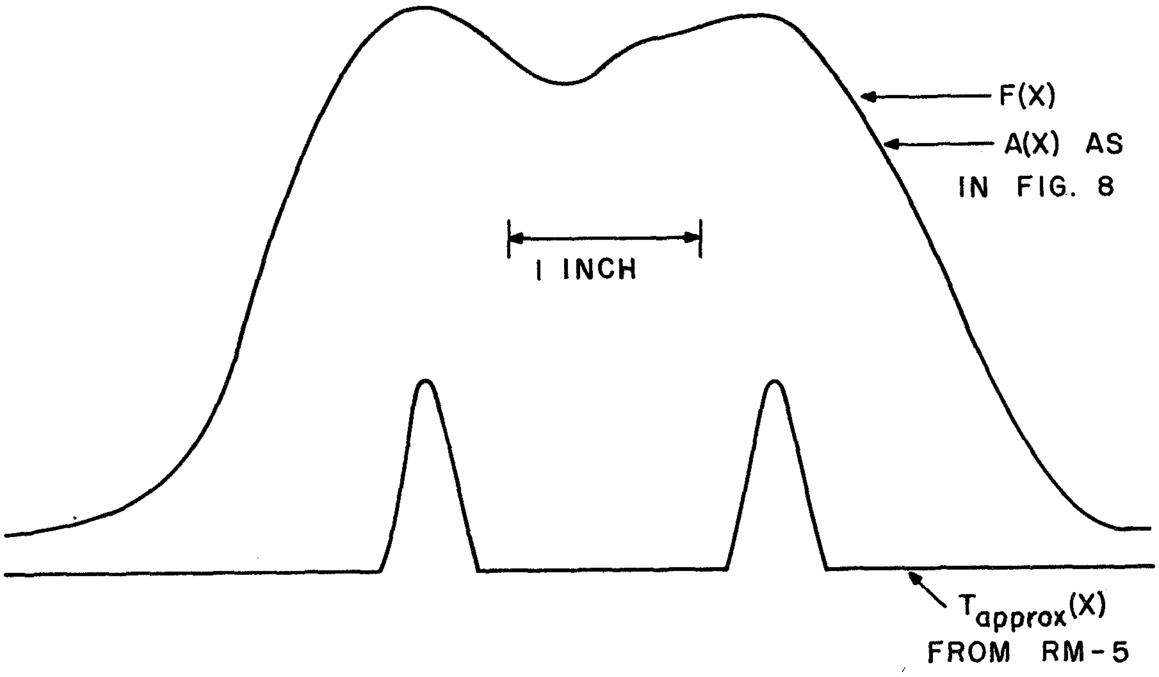


Figure 9

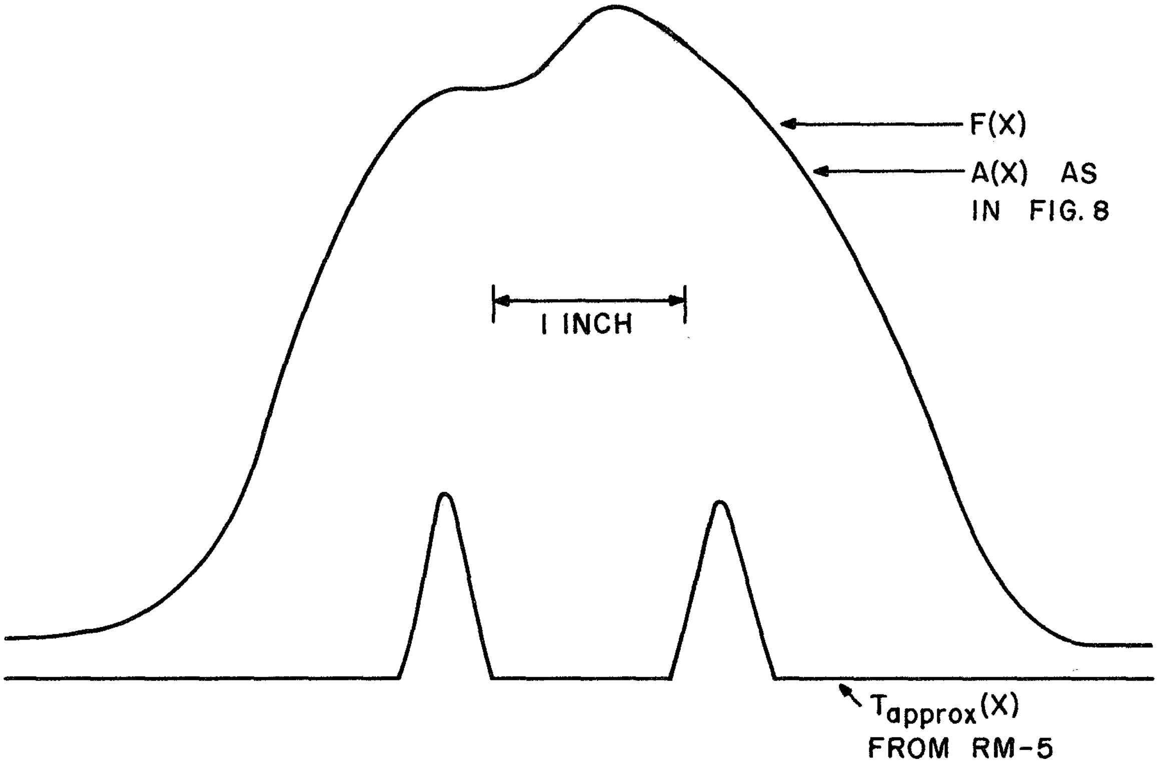


Figure 10

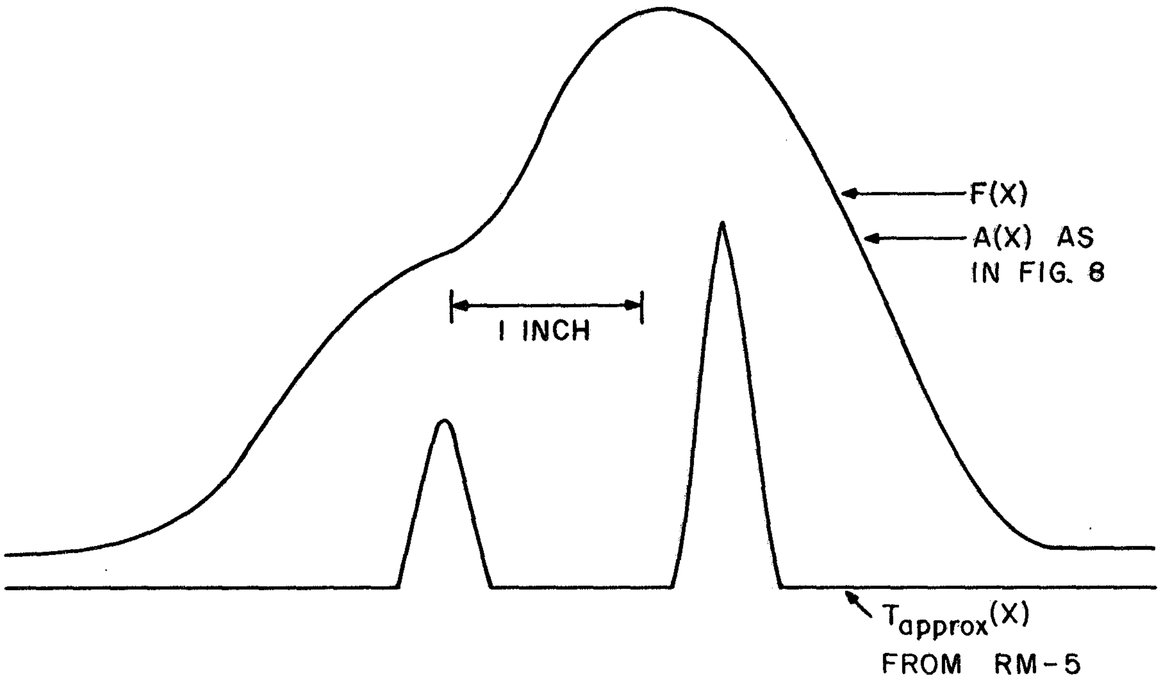


Figure 11

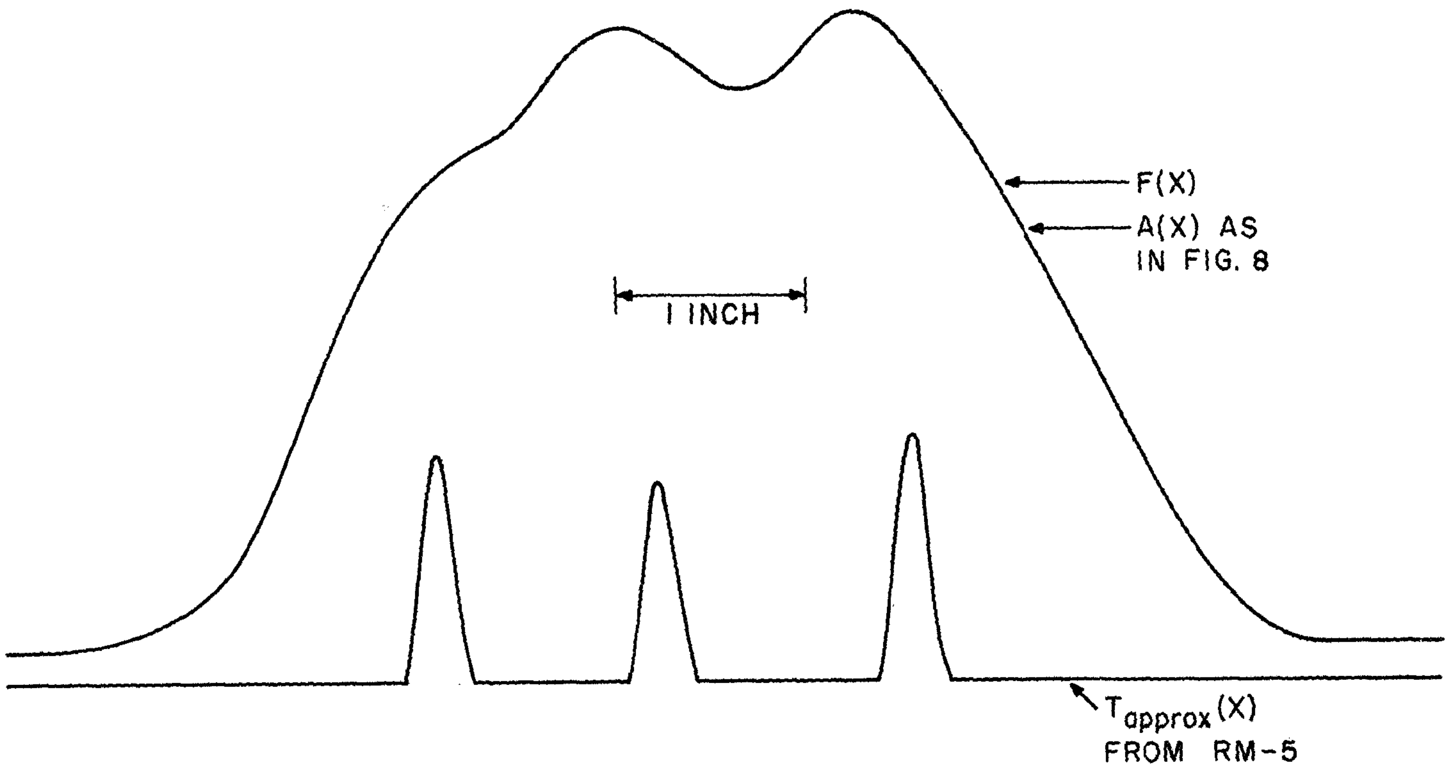


Figure 12

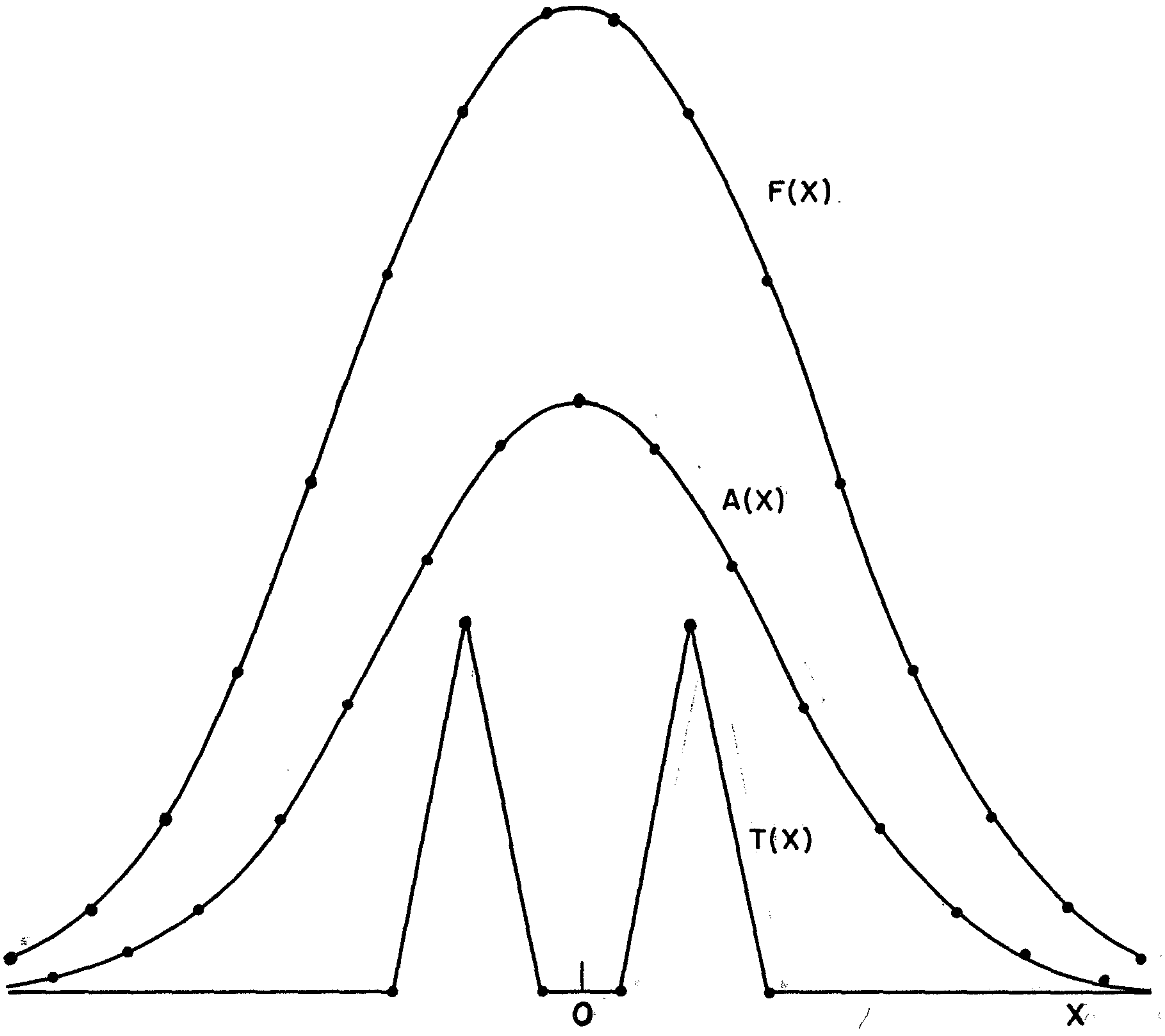


Figure 13

at the points shown in Figure 13 and these numbers were used as input data for the digital program. The points on T(x) in Figure 13 represent the exact numerical solution which should result from this calculation. The problem was then solved using equation (110), with α=1 and G being varied in steps from 1 to 1000, and carrying out a fixed number of iterations in each case. The results are shown in Figure 14. It can be seen that the doublet is resolved for G=10 and the form of the solution changes very little for gains greater than 20. Figure 15 shows a Gaussian doublet problem which was used to check the effect of varying α in equation (110). It was constructed in the same manner as was Figure 13 with the exception that many more points were used in an effort to provide a more realistic simulation of the RM-5. Figure 16 shows the solution obtained after 60 iterations with G=500 for various values of α. Figure 16 clearly demonstrates the value of filtering.

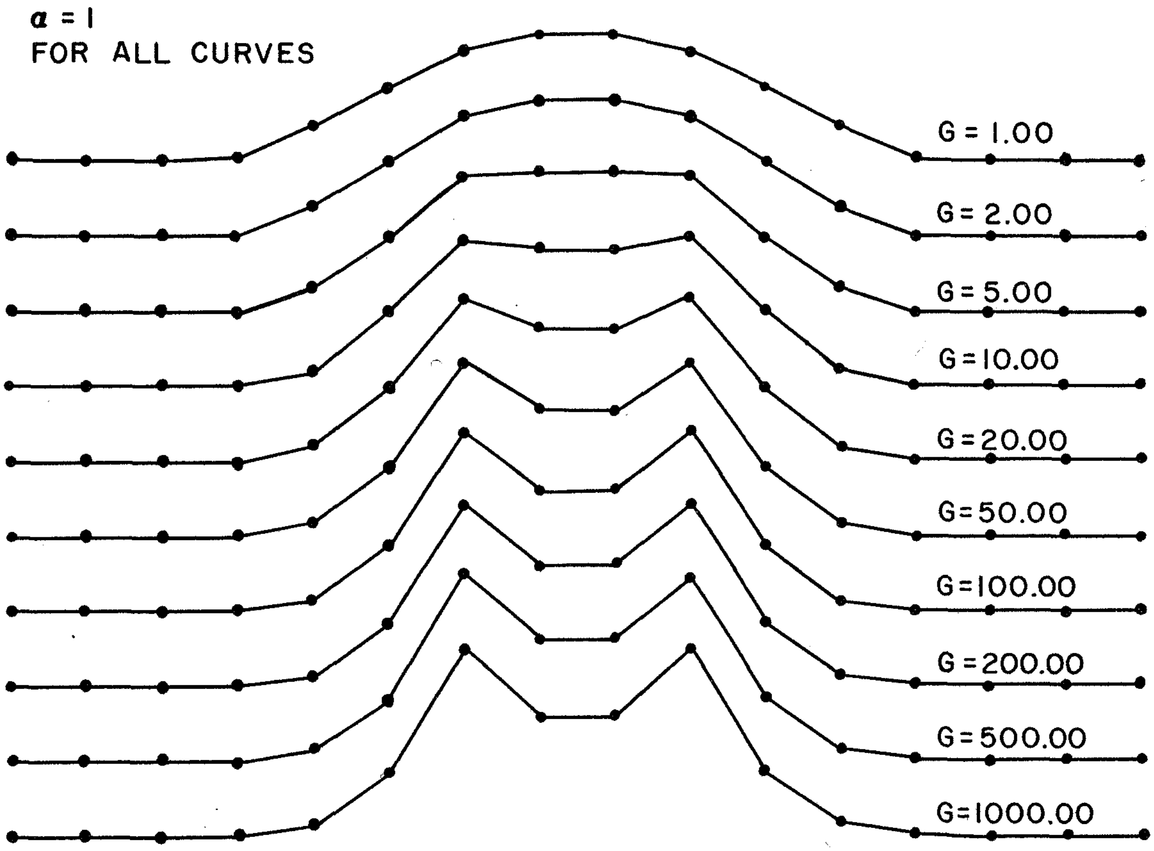


Figure 14

Solution obtained after 30 iterations for problem shown in Figure 13 using equation 110 on digital computer

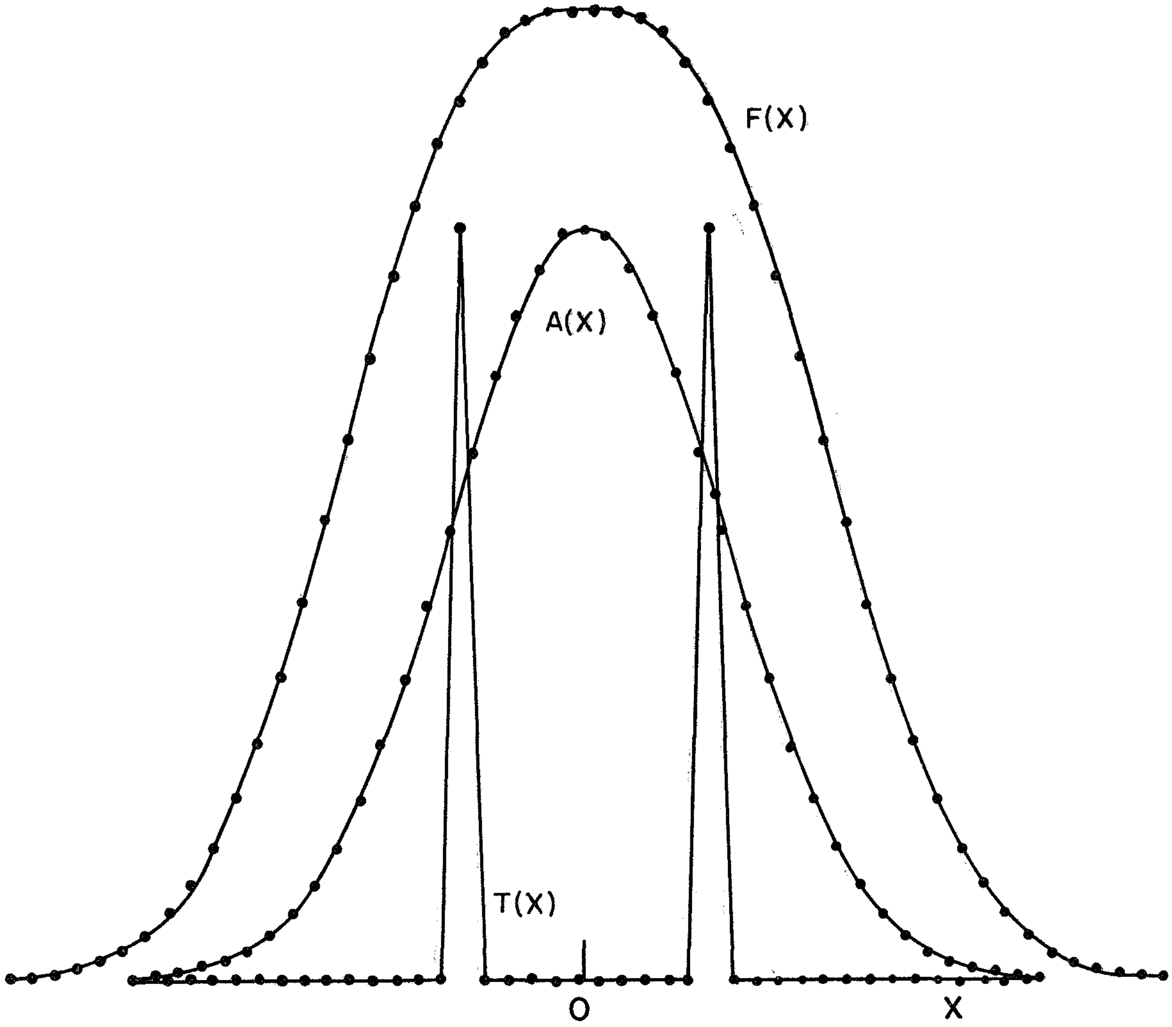


Figure 15

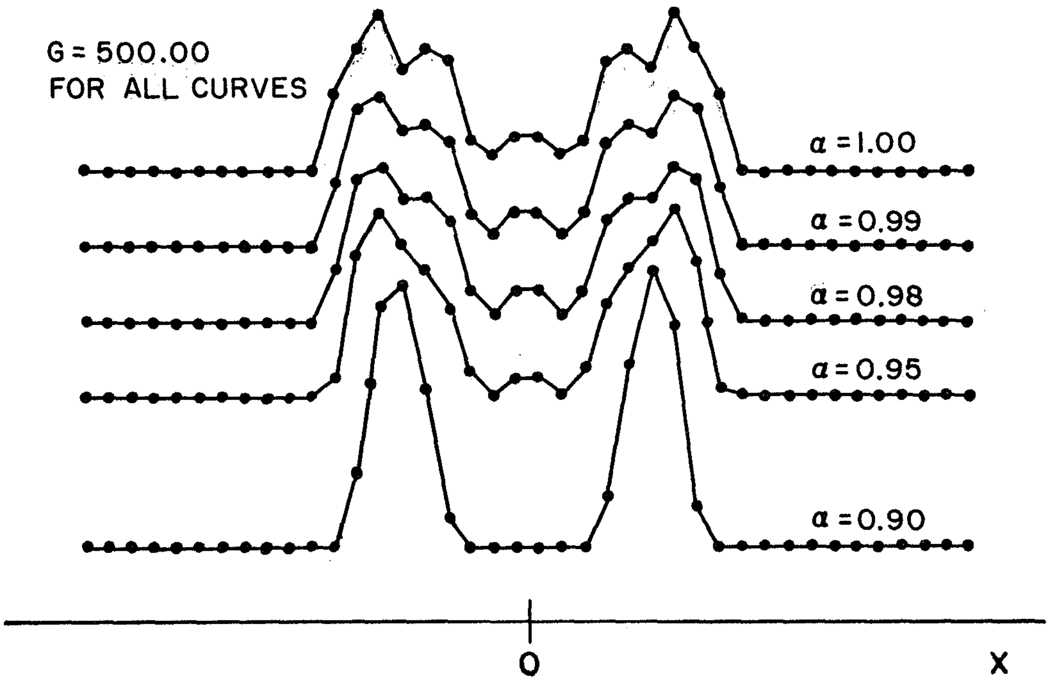


Figure 16

Solution after 60 iterations for problem shown in Figure 15 using equation 110 on digital computer

CHAPTER X

RESULTS FROM OTHER NUMERICAL COMPUTATIONS

The relatively simple Gaussian doublet problem shown in Figure 13 was also solved using the accelerated version of steepest descent as defined by equations 63, 64, and 65. Figure 17 shows the resulting solution after 10 iterations for different integer values of p. Figure 18 shows the resulting solution after 10 iterations for different values of p with the additional modification that after each iteration, any negative values in the trial solution were set equal to zero. This in effect is equivalent to the "negative rejection" feature of the RM-5 device. Examination of Figure 18 clearly shows that this "negative rejection" attenuates the high frequency noise in the solution and for values of p greater than 5, noticably accelerates convergence. The same problem was also solved using the method of conjugate gradients described in section 6.4. The solution obtained by this method after 5, 10, 15, and 20 iterations is shown in Figure 19.

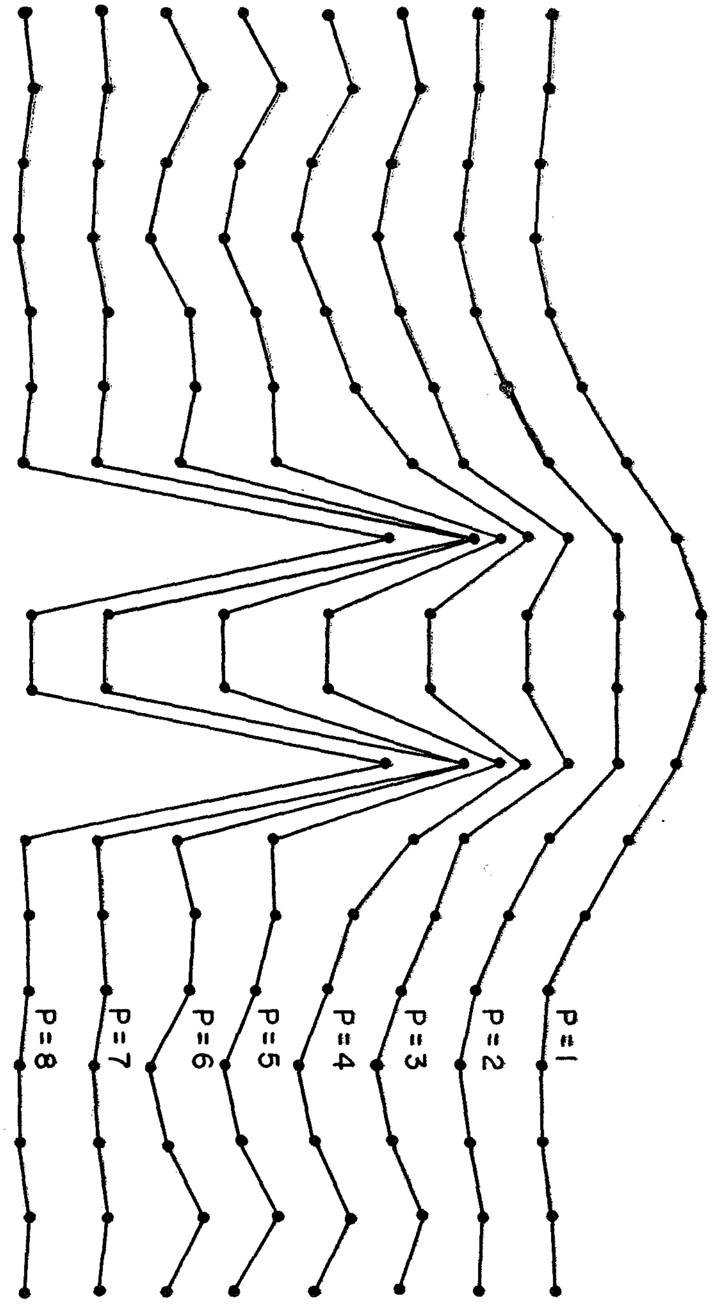


Figure 17

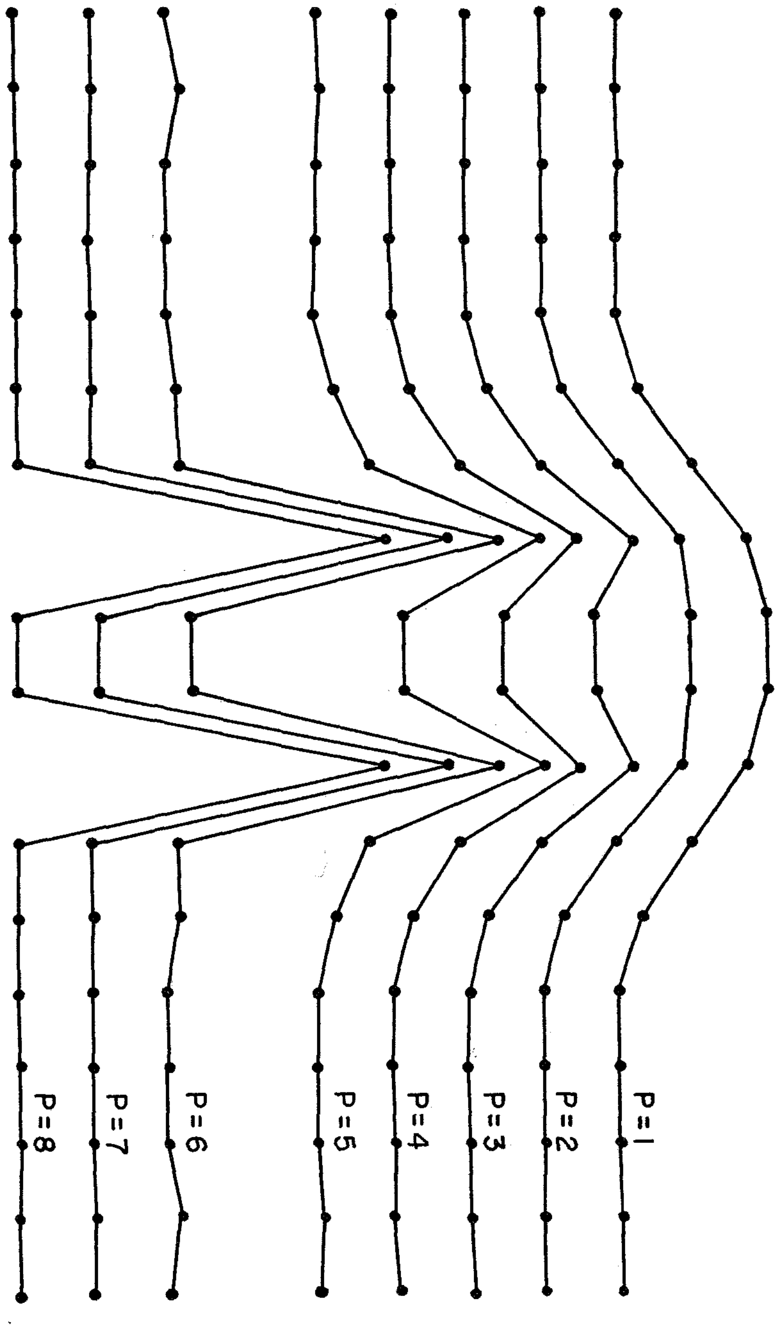


Figure 18

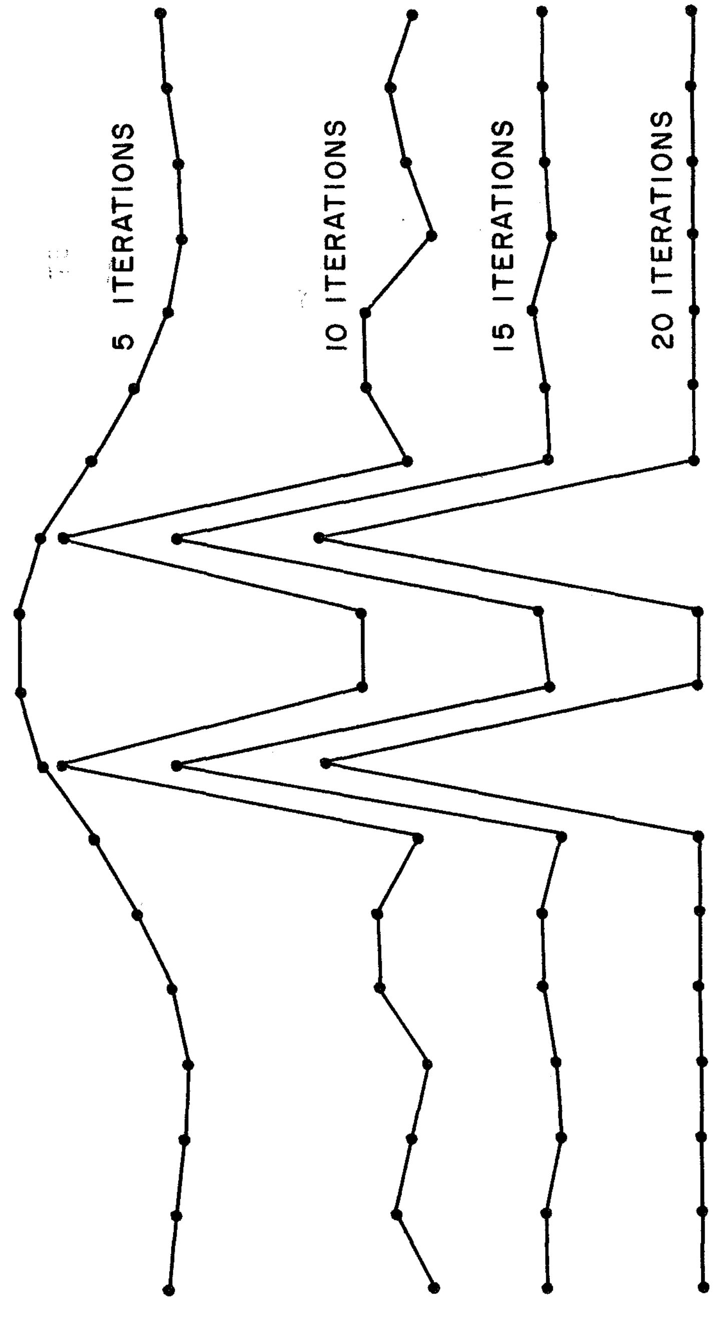


Figure 19

Results from Conjugate Gradients Method

CHAPTER XI

BRIEF REMARKS ABOUT TWO-DIMENSIONAL DECONVOLUTION

The areas of image enhancement and pattern recognition are typical 2-dimensional deconvolution problems. Because of the ease with which one can Fourier transform 2-dimensional functions with optical techniques, a new field of optical computers is evolving (typical of this method is the work of Stroke [93]). The present limitations on optical methods are the input-output problems, since the two-dimensional data is usually handled in the form of a photographic transparency. The technique used is just the two-dimensional analogy of the one-dimensional method as described in section 5.1.

CHAPTER XII

SUMMARY AND CONCLUSIONS

12.1 Discussion of Results

One of the general conclusions which can be drawn from the material presented in this thesis is that there are several different iterative techniques which may be used to solve the convolution integral equation. Perhaps a more important specific conclusion which is suggested by the results presented in sections IX and X, is the fact that the incorporation of filtering and "negative rejection" in a numerical iterative technique is highly desirable for the deconvolution of spectral information. Both of these properties are incorporated in the RM-5 analog device.

12.2 Implications for Further Work

It would be desirable to investigate the convergence criteria for the modified Gauss-Seidel iteration, which describes the operation of the RM-5 device (c.f. equation 110), in some detail, in order to theoretically determine the effects of G and α. Another study which could be carried out theoretically, and/or experimentally, would be the effects of random noise on the deconvolution process, again with special emphasis upon the effects of varying G and α.

REFERENCES

1. Emslie, A. G. and King, G. W., J. Opt. Soc. Amer., 43, No. 8, 657, (1953)

2. Frei, K. and Gunthard, Hs. H., J. Opt. Soc. Amer., 51, No. 1, 83, (1961)

3. Roseller, A., Infrared Physics, 5, 51, (1965)

4. Schrack, R. A., Nuc. Inst. Meth., 45, 319, (1966)

5. Rice, R. B., Geophysics, XXVII, No. 1, 4, (1962)

6. Rollet, J. S., and Higgs, L. A., Proc. Phys. Soc., 79, Part 1, No. 507, 87, (1962)

7. Wightman, Thirteenth Nat. Vacuum Symposium of the Amer. Vacuum Soc., San Francisco, 25-28, (1966)

8. von Neumann, J. and Goldstine, H. H., Bul. Amer. Math. Soc., 53, 1021, (1947)

9. Goldstine, H. H., and von Neumann, J., Proc. Amer. Math. Soc., 2, 188, (1951)

10. Turing, A. M., Q. J. Mech. Appl. Math., 1, 287, (1948)

11. Tal. A. A., U. of Md. Sci. Rep. #1, WBG-8D, Jan. (1967)

12. van Cittert, P. H., Z. Phys., 69, 304, (1931)

13. Burger, H. C., and van Cittert, P. H., Z. Phys., 79, 724, (1932)

14. Burger, H. C., and van Cittert, P.H., Z. Phys., 81, 428, (1933)

15. Bertolini, G., Cappellani, F., and Rota, A., Nucl. Inst. Meth., 9, 107, (1960)

16. Skarsgard, L. D., Johns, H. E., and Green, L. E. S., Radiation Research, 14, No. 3, 261, (1961)

17. Bracewell, R. N., Proc. Phys. Soc., 79, 1298, (1962)

18. Wortman, D. E. and Cramer, J. G., Jr., Nucl., Inst. Meth., 26, 257, (1964)

19. Slavinskas, D. D., Kennett, T. J., and Prestwich, W. V., Nucl. Inst. Meth., 37, 36, (1965)

20. Ioup, G. E., and Thomas, B. S., J. Chem. Phys., 46, No. 10, 3959, (1967)

21. Girard, A. and Lemartre, M., Optica Acta, 14, No. 4, 329, (1967)

22. Ralston, A., A First Course in Numerical Analysis, McGraw-Hill, New York, (1965)

23. Martin, D. W. and Tee, G. J., Comput. J., 4, No. 3, 242, (1961)

24. Seidel, L., Obhandlungen der Bayerischen Akademie, 11, Dritte Abteilung, 81, (1873)

25. Aitken, A. C., Proc. Roy. Soc. Edinburgh, 63, 52, (1950)

26. Reich, E., Ann. Math. Statistics, 20, 448, (1949)

27. Young, D., Trans. Amer. Math. Soc., 76, 92, (1954)

28. Temple, G., Proc. Roy. Soc. A, 169, 476, (1939)

29. Stiefel, E., Z. Angew. Math. Phys., 3, 1, (1952)

30. Kantorovich, L. V., and Akilov, G. P., Functional Analysis in Normed Spaces, MacMillan, New York, (1964)

31. Hestenes, M. R. and Stiefel, E., J. Res. Nat. Bur. Standards, 409, (1952)

32. Rautian, S. G., Soviet Physics (Uspekhi), 1, (66), No. 2, 245, (1958)

33. Lane, R. O., Morehouse, N. F. Jr., and Phillips, D. L., Nucl. Inst. Meth., 9, 87, (1960)

34. Sachenko, V. P., Bul. Acad. Sci. USSR., 25, 1043, (1961)

35. George, C. F., Smith, H. W., and Bostick, F. X., Proc. of I.R.E., Nov., 2313, (1962)

36. Morrison, J. D., J. Chem. Phys., 39, No. 1, 200, (1963)

37. Keller, H. and Segmuller, A., Rev. Sci. Instr., 34, No. 6, 684, (1963)

38. Mori, M. and Doi, K., Jap. J. Appl. Phys., 3, No. 2, 112, (1964)

39. Ritchie, R. H., and Anderson, V. E., Nucl. Inst. Meth., 45, 277, (1966)

40. Larson, H. P. and Andrew, K. L., Appl. Optics, 6, No. 10, 1701, (1967)

41. Oppenheim, A. V., Schafer, R. W., and Stockham, G., Jr., Proc. I.E.E.E., Aug. 1, 1264, (1968)

42. Sachenko, V. P., Bul. Acad. Sci. USSR, 25, 1052, (1961)

43. Flynn, C. P., Proc. Phys. Soc., LXXVIII, 1546, (1961)

44. Flynn, C. P., and Seymour, E. F. W., J. Sci. Instr., 39, 352, (1962)

45. Jones, A. F. and Misell, D. L., Brit. J. Appl. Phys., 18, 1479, (1967)

46. Zorner, K. H., Z. F. Angew Phys., 22, 239, (1967)

47. Stone, H., J. Opt. Soc. Amer., 52, 998, (1962)

48. Venkataraghavan, R., McLafferty, F. W., and Amy, J. W., Anal. Chem., 39, No. 2, 179, (1967)

49. Grissom, J. T., Koehler, D. R. and Gibbs, B. G., Nucl. Inst. Meth., 45, 190, (1966)

50. Luenberger, D. G., and Dennis, V., Analytical Chem., 38, No. 6, 715, (1966)

51. Su, Y. S., Nucl. Inst. Meth., 54, 109, (1967)

52. Jauch, J. M., and Misra, B., H. P. A., VI, 30, (1964)

53. Baker, C. T. H., Fox, L., Mayers, D. F., and Wright, K., Computer J., 1, 141

54. Rushforth, C. K., and Harris, R. W., J. Opt. Soc. Amer., 58, 539, (1968)

55. Hildebrand, F. B., Introduction to Numerical Analysis, McGraw-Hill, New York, (1956)

56. John, F., Lectures on Advanced Numerical Analysis, Gordon and Breach, London, (1967)

57. Fox, L. and Goodwin, E. J., Phil. Trans. Roy. Soc. (London), 245, 501, (1953)

58. Bracewell, R. N., Aust. J. Phys., 8, 200, (1955)

59. Genkin, Ya. E., and Rumyantsev, I. A., Bul. Acad. Sci. USSR, 25, 1057, (1961)

60. Nikiforov, I. Ya., Sachenko, V. P., and Blokhin, M. A., Bul. Acad. Sci. USSR, 25, 1057, (1961)

61. Allen, L. C., Nature, 196, No. 4855, 663, (1962)

62. Dotti, D., Nucl. Inst. Meth., 54, 125, (1967)

63. Mikusinski, J., Operational Calculus, Pergamon, New York, (1959)

64. Berg, L., Introduction to the Operational Calculus, North Holland Publishing Co., Amsterdam, (1967)

65. Phillips, D. L., J. Amer. Comp. Mach., 9, 84, (1962)

66. Tihonov, A. N., Soviet Math. Dokl., 4, No. 4, 1035, (1963)

67. Wang Lau, L., Amer. J. Phys., 36, No. 9, 842, (1968)

68. King, G. W. and Emslie, A. G., J. Opt. Soc. Amer., 43, No. 8, 664, (1953)

69. King, G. W., Blanton, E. H., and Frawley, J., J. Opt. Soc. Amer., 44, No. 5, 397, (1954)

70. Duffieux, P. M., Revue Opt. Theor. Instrum., 39, 491, (1960)

71. French, C. S., et. al., Rev. Sci. Instr., 25, No. 8, 765, (1954)

72. Noble, F. W., Hayes, J. E., and Eden, M., Proc. I.R.E., 47, 1952, (1959)

73. Von Profos, P., Regelungstechnik, 11, 491, (1964)

74. Diamantides, N. D., Electronics, April 13, p65, (1962)

75. Kindlemann, P. J. and Hooper, E. B., Jr., Rev. Sci. Instr., 39, No. 6, 864, (1968)

76. Zverev, V. A. and Orlof, E. F., Instr. and Exper. Tech., 1, 54, (1960)

77. Breton, C. and Hirschberg, J. G., Appl. Optics, 3, No. 6, 731, (1964)

78. Goldberg, E. A., R.C.A. Review, 9, 394, (1948)

79. Dolby, R. M. Proc. Phys. Soc., LXXIII, 1, 81, (1959)

80. Dolby, R. M., J. Sci. Instr., 40, 345, (1963)

81. Dolby, R. M. and Cosslett, V. E., Proc. of Stockholm Symposium on X-ray Microscopy and Microanalysis, 357, (1959)

82. Allen, L. C., Gladney, H. M., and Glarum, S. H., J. Chem. Phys., 40, No. 11, 3135, (1965)

83. Glarum, S. H., Rev. Sci. Instr., 36, No. 6, 771, (1965)

84. Krishnamurty, E. V., J. Sci. Instr., 37, 419, (1960)

85. Korsunskii, M. I. and Genkin, Ya. E., Bul. Acad. Sci. USSR, 25, 1020, (1961)

86. Kendall, B. R. F., Rev. Sci. Instr., 32, No. 6, 758, (1961)

87. Kendall, B. R. F., Rev. Sci. Instr., 33, No. 1, 30, (1962)

88. Kendall, B. R. F., U.S. Patent 3,154,747 (Oct. 27, 1964).

89. Kendall, B. R. F., J. Sci. Instr., 39, 267, (1962)

90. Kendall, B. R. F., J. Sci. Instr., 43, 215, (1966)

91. Zabielski, M. F., P.S.U. Thesis, M. S. in Physics, (1966)

92. Kendall, B. R. F. and Zabielski, M. F., Proc. of 15th Annual Conference on Mass Spectrometry and Allied Topics (ASTM E-14), (1967)

93. Stroke, G. W., Phys. Letters, 26A, No. 9, 443, (1968)

APPENDIX I

A definition of the true norm of a square N X N matrix can be given by[56]

$$N_T(A) = \max \frac{|AX|}{|X|} \qquad (I.1)$$

where X is any N X 1 vector and $|X|$ is defined by

$$|x| = \left| \sqrt{\sum_{i=1}^{N} x_i^2} \right| \qquad (I.2)$$

In order to find the true norm of a matrix A, one first lets the arbitrary vector X be expanded in terms of the N eigenvectors of A:

$$X = \sum_{i=1}^{N} C_i y_i \qquad (I.3)$$

where the y_i are eigenvectors of the matrix A. Then one can also write

$$AX = \sum_{i=1}^{N} C_i \lambda_i y_i \qquad (I.4)$$

where λ_i is the eigenvalue of A which is associated with the eigenvector y_i. Using equations (I.2), and (I.3) and (I.4) substituted into equation (I.1), one can write

$$N_T(A) = \max \sqrt{\sum_{i=1}^{N} (C_i \lambda_i)^2 \Big/ \sum_{i=1}^{N} C_i^2} \qquad (I.5)$$

Now for convenience, let the eigenvalues be indexed in descending order; i.e., $\lambda_1 > \lambda_2 > \lambda_3 . . . . > \lambda_n$. Then equation (I.5) can be rewritten

$$N_T(A) = \max \; |\lambda_1| \sqrt{\sum_{i=1}^{N} (C_i \frac{\lambda_i}{\lambda_1})^2 \Big/ \sum_{i=1}^{N} C_i^2} \qquad (I.6)$$

Now since $\lambda_i/\lambda_1 \leq 1$, with the equality holding only for $i = 1$; the term inside the radical is ≤ 1 with the equality holding only for the following conditions:

$$C_i = \begin{cases} C; & i=1 \\ 0; & i>1 \end{cases} \qquad (I.7)$$

Under these conditions also, the right hand side of equation (I.6) is maximized resulting in $|\lambda_1|$, which is the magnitude of the largest eigenvalue of the matrix A. Hence the result

$$N_T(A) = |\lambda_1| \qquad (I.8)$$