N65-27473
(ACCESSION NUMBER)
115
(PAGES)
(NASA CR OR TMX OR AD NUMBER)
(THRU)
3
(CODE)
19
(CATEGORY)
FACILITY FORM 602

ORNL-3743
UC-32 – Mathematics and Computers
TID-4500 (40th ed.)

UTILIZATION OF A PRIORI INFORMATION BY MEANS OF MATHEMATICAL PROGRAMMING IN THE STATISTICAL INTERPRETATION OF MEASURED DISTRIBUTIONS

(THESIS)

W. R. Burrus

GPO PRICE $
S PRICE(S) $
rd copy (HC) 4.00
rofiche (MF) .75

N65-27473
(ACCESSION NUMBER)
115
(PAGES)
CR 63442
(NASA CR OR TMX OR AD NUMBER)
(THRU)
1
(CATEGORY)
19
FACILITY FORM 602

OAK RIDGE NATIONAL LABORATORY
operated by
UNION CARBIDE CORPORATION
for the
U.S. ATOMIC ENERGY COMMISSION

~~"Available to U.S. Government Agencies and U.S. Government Contractors Only"~~

Printed in USA. Price $4.00. Available from the Clearinghouse for Federal Scientific and Technical Information, National Bureau of Standards, U.S. Department of Commerce, Springfield, Virginia

LEGAL NOTICE

This report was prepared as an account of Government sponsored work. Neither the United States, nor the Commission, nor any person acting on behalf of the Commission:

A. Makes any warranty or representation, expressed or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or

B. Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used in the above, "person acting on behalf of the Commission" includes any employee or contractor of the Commission, or employee of such contractor, to the extent that such employee or contractor of the Commission, or employee of such contractor prepares, disseminates, or provides access to, any information pursuant to his employment or contract with the Commission, or his employment with such contractor.

CASE FILE COPY

ORNL-3743

Contract No. W-7405-eng-26

Neutron Physics Division

UTILIZATION OF A PRIORI INFORMATION BY MEANS OF MATHEMATICAL PROGRAMMING IN THE STATISTICAL INTERPRETATION OF MEASURED DISTRIBUTIONS*

W. R. Burrus

*Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate School of the Ohio State University. This work was partially supported by the National Aeronautics and Space Administration under Order R-104.

JUNE 1965

OAK RIDGE NATIONAL LABORATORY
Oak Ridge, Tennessee
operated by
UNION CARBIDE CORPORATION
for the
U.S. ATOMIC ENERGY COMMISSION

"Available to U.S. Government Agencies and
U. S. Government Contractors Only"

TABLE OF CONTENTS

		Page No.
1.	Introduction	3
	Statement of problem	3
	Statistical considerations	4
	Specific applications	4
	Inversion or unfolding methods	6
	Solution by statistical estimation	7
	Nonuniqueness of solution	8
	A priori information	10
	Constrained estimation	12
2.	Ordinary Estimation	17
	Linear estimation	17
	Confidence intervals	20
	Histogram approximation	25
	Window functions	27
	Errors in the matrix	32
3.	Constrained Estimation	33
	Inequality method	34
	A priori information	36
	Narrowest possible confidence limits	37
4.	Reduction to Finite Dimension	43
5.	Programming Formulation	53
6.	Linear Programming Approximation	59
	Efficiency of computational method	63
7.	Duality Relations	65
8.	Errors in the Response Matrix	77
9.	Modified Least-Squares Approximation	81
	Primal formulation	83
	Dual formulation	89
	Numerical calculation	93

TABLE OF CONTENTS (continued)

Page No.

10. Miscellaneous Applications ---------- 97

Solution of linear equations ---------- 97

Unfolding gamma spectra ---------- 98

Other applications ---------- 103

11. Additional Constraints ---------- 105

12. Conclusions ---------- 111

List of References ---------- 112

UTILIZATION OF *A PRIORI* INFORMATION BY MEANS OF MATHEMATICAL PROGRAMMING IN THE STATISTICAL INTERPRETATION OF MEASURED DISTRIBUTIONS

ABSTRACT

27473

A statistical approach is developed for unfolding instrument measurements which are related to an unknown function by a linear integral transformation. If statistical uncertainties are present in the measured function and in the kernel of the transformation, then the use of *a priori* information is shown to be necessary for a nontrivial solution. Without such information, an infinite-width confidence interval can be found for any nontrivial function of the unknown function. Suitable *a priori* information, however, nearly always exists in physically motivated problems. The weakest form of *a priori* information considered is simple nonnegativity. Stronger *a priori* information, such as smooth or monotonic behavior, can be put in the same form as simple nonnegativity by suitable transformation. The unfolding problem requires two steps: selecting a set of functions of the unknown function to adequately describe the function and then numerically obtaining the width of the confidence interval for each function. Several computational techniques based upon linear and quadratic programming are described. Finally, the method is illustrated by unfolding gamma-ray and neutron spectra from pulse-height distributions obtained with scintillation spectrometers and by solving a poorly conditioned set of equations.

Author

1. INTRODUCTION

Statement of problem

A problem encountered by many experimentalists in different fields is created by experimental measurements of a distribution $x(s)$ yielding a result $b(t)$ which is different from the one actually desired due to distortions introduced by the instrument. Some familiar examples of distortions are those occurring

1. in a neutron or gamma-ray spectrum in the pulse-height distribution of a multichannel scintillation spectrometer,
2. when scanning with a telescope or directional antenna, due to the finite width of the response pattern,
3. in the measurement of optical or infrared spectra, due to the infinite width "line" of the spectrometer.

More generally, in this type of problem the operation of the measuring instrument may be symbolically denoted by the operator O. Then O "operates" upon the desired distribution, or spectrum, $x(s)$ and yields the observed distribution $b(t)$:

$$O\, x(s) = b(t) \; . \tag{1.1}$$

Frequently, many interesting measurements may be characterized to a very good approximation by a linear operator. Equation (1.1) can then be written as the linear integral equation

$$\int K(t,s)\; x(s)\; ds = b(t) \; , \tag{1.2}$$

where $K(t,s)$ is the response function (or Green's function) of the measuring apparatus, $x(s)$ is the distribution function, or spectrum, we desire to measure, and $b(t)$ is the experimental distribution. The variable s may be continuous or may be discrete if the spectrum consists of components at certain values of s. The variable t may also be continuous, as when scanning an antenna through a continuous angle, or discrete as obtained with a multichannel pulse-height analyzer. Figure 1 is a sketch of the function $K(t,s)$ for a gamma-ray scintillation spectrometer, in which s is continuous, corresponding to the energy E of the incident particle, and t is discrete, corresponding to the channel number i of a multichannel analyzer. In the appropriate notation for this problem, the kernel is written as $K_i(E)$ and is interpreted as the probability that a gamma ray of energy E will result in a count in channel i.

Statistical considerations

In practice, the operator O does not always yield the same distribution $b(t)$ when applied to the same spectrum, or distribution, $x(s)$ because of noise in the measuring instrument, because of the statistics associated with the particular nature of $x(s)$, or because O itself may be stochastic. Thus a sequence $b^1(t)$, $b^2(t)$, ... results from the experiment being repeated under identical physical conditions. In any actual measurement, $b(t)$ will not be known, but only a member of the ensemble of possible distributions is measured.

Specific applications

As already mentioned, one of the most frequent problems is correcting for the effects of nonperfect line shape in various kinds of

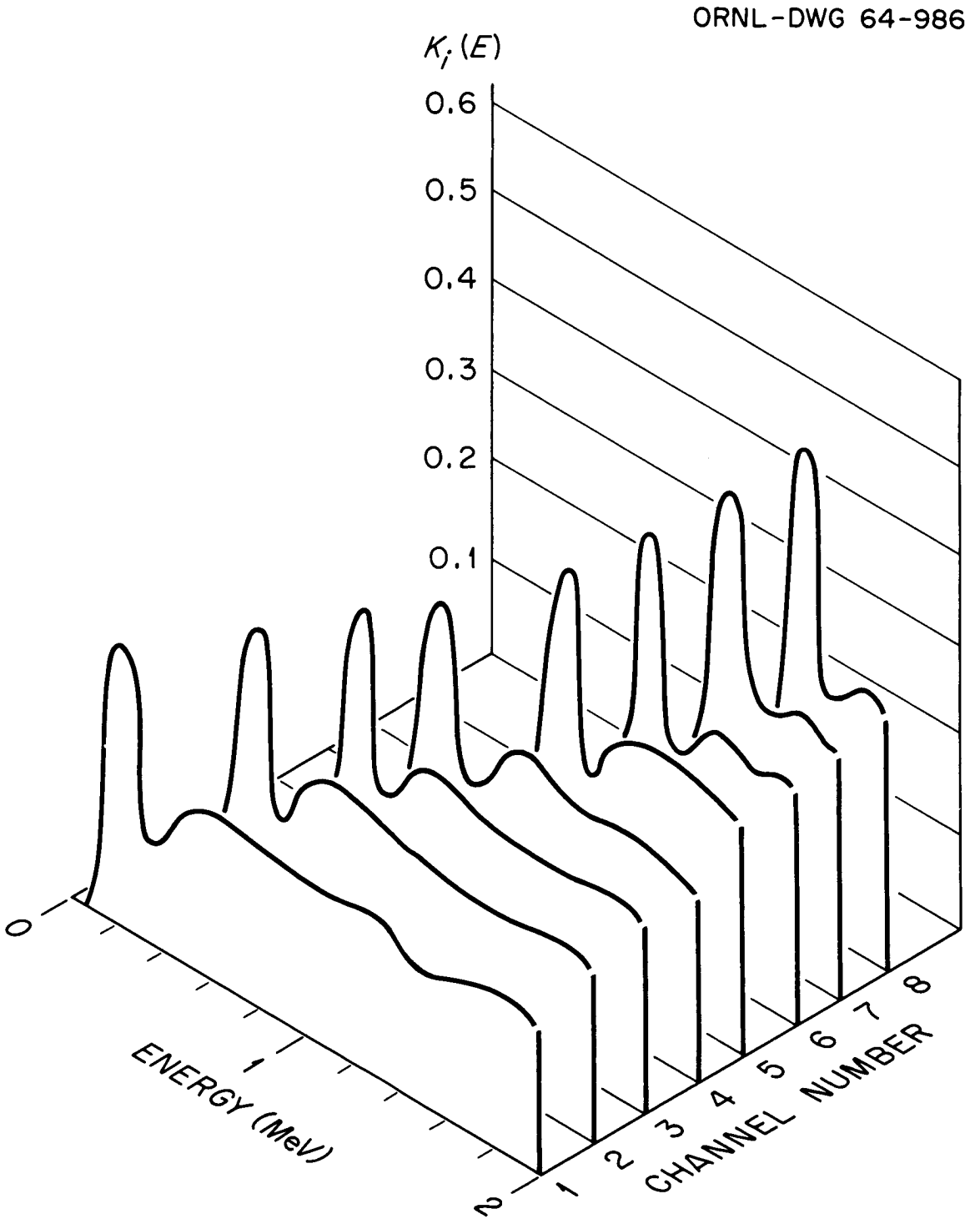


Fig. 1. A sketch of the kernel K(t,s) for a gamma-ray scintillation spectrometer.

spectrometers (1—8).[1] The response of electrical networks can be expressed as a convolution of the network "transient response" with the input signal, so that the network problem is a special case of the more general problem (9,10). Problems in the analysis of scintillation spectrometers (11—16) are of considerable interest. Other applications occur in meterology (17), in analysis of diffraction data (18), and in crystallography (19). Among the earlier nuclear spectroscopy correction methods, a method of determining the spectrum of x radiation from the shape of an absorption curve was published in 1932 by Silberstein (20). In addition, the purely numerical aspects of inverting the general Fredholm integral equation are of interest (21—25). Of a very similar type are certain fundamental problems in wave mechanics (26).

Inversion or unfolding methods

Ignoring for the moment the uncertainty in b(t), the determination of x(s) from b(t) is sometimes referred to as inversion or "unfolding." The latter term stems from the German usage of "Faltung" (folding) for the integral in Eq. (1.2). Several different methods are known for obtaining the formal solution (27), but, as was mentioned, the formal methods are not applicable if the function b(t) is known only as a curve or as a set of tabulated points (at discrete t). In place of a formal solution, various numerical techniques have been employed (28—36). Many of these numerical techniques make use of the equivalence between

[1]See "List of References" at end of paper.

the integral equation and a system of linear equations of sufficiently high order. Then the set of equations is solved by standard numerical techniques. Some other techniques consider the problem in its continuous form and make use of numerical integration formulas to reduce it to a finite form for computation. Among the earliest numerical formulas was one developed by Eddington in 1913 (37) for correcting astronomical observations for the smearing effect of a known standard error of observation. More recent work on this problem is summarized by Trumpler and Weaver (38).

Solution by statistical estimation

When the right-hand side, $b(t)$, of Eq. (1.2) is a random sample from an ensemble of possible distributions, statistical estimation may be used to find the solution. The simplest case occurs when both s and t are discrete variables with values of $(s_1, s_2, \ldots s_n)$ and $(t_1, t_2, \ldots t_m)$. Then the integral equation may be written as a matrix equation:

$$Kx = b \ , \tag{1.3}$$

where $K_{ij} = K(t_i, s_j)$, and x and b are vectors. The determination of the elements of x (or of any function of the elements) is a standard problem in linear estimation (39—41). If rank $K = n$, then a solution with a finite-width confidence interval can be found by ordinary least-squares techniques. But if the equation is overdetermined (with $n > m$) or if $x(s)$ is continuous instead of discrete, then special techniques have to be used. If the kernel $K(s,t)$ is known exactly, it is still possible to find estimates of certain combinations of the elements, even if it is impossible to estimate $x(s)$ itself.

Nonuniqueness of solution

If there are some solutions $x(s)^h$ to the homogeneous equation

$$\int K(t,s)\ x(s)\ ds = 0\ , \qquad (1.4)$$

then they may be added to any solution $x(s)$ of Eq. (1.2), and $x(s) + x(s)^h$ will still be a solution. Thus the solution $x(s)$ is not unique if any solutions to the homogeneous equation are possible. Experimentally this may be interpreted by noticing that there are "invisible" components which the spectrometer cannot see. For example, if a rectangular or triangular "line shape" with a full width at half maximum of W is convoluted with a sinusoidal function, then no output $b(t)$ will result if the period of the sinusoid is an exact submultiple of W. For the more realistic Gaussian line shape, a sinusoidal $x(s)$ is attenuated quite strongly as the period becomes much smaller than W, as shown in Fig. 2. In fact, for <u>any</u> kernel which is merely integrable, the attenuation of a sinusoidal distribution increases without limit as the period of the sinusoid decreases. Thus the solution to Eq. (1.2) when s is continuous must always be nonunique in the sense that an arbitrarily large amount of a sinusoidal component can be added to the solution and still satisfy the equation to within any prescribed nonzero amount.

In addition, when $b(t)$ is uncertain by a small amount, there may be components which nearly satisfy the homogeneous equation (1.4) such that the residual is smaller than the uncertainty in $b(t)$. Then these "nearly invisible" components must also be accepted as valid solutions to (1.2) since they satisfy the equation to within the uncertainty in $b(t)$. Moreover, if $x(s)$ is continuous, it will be shown that not only

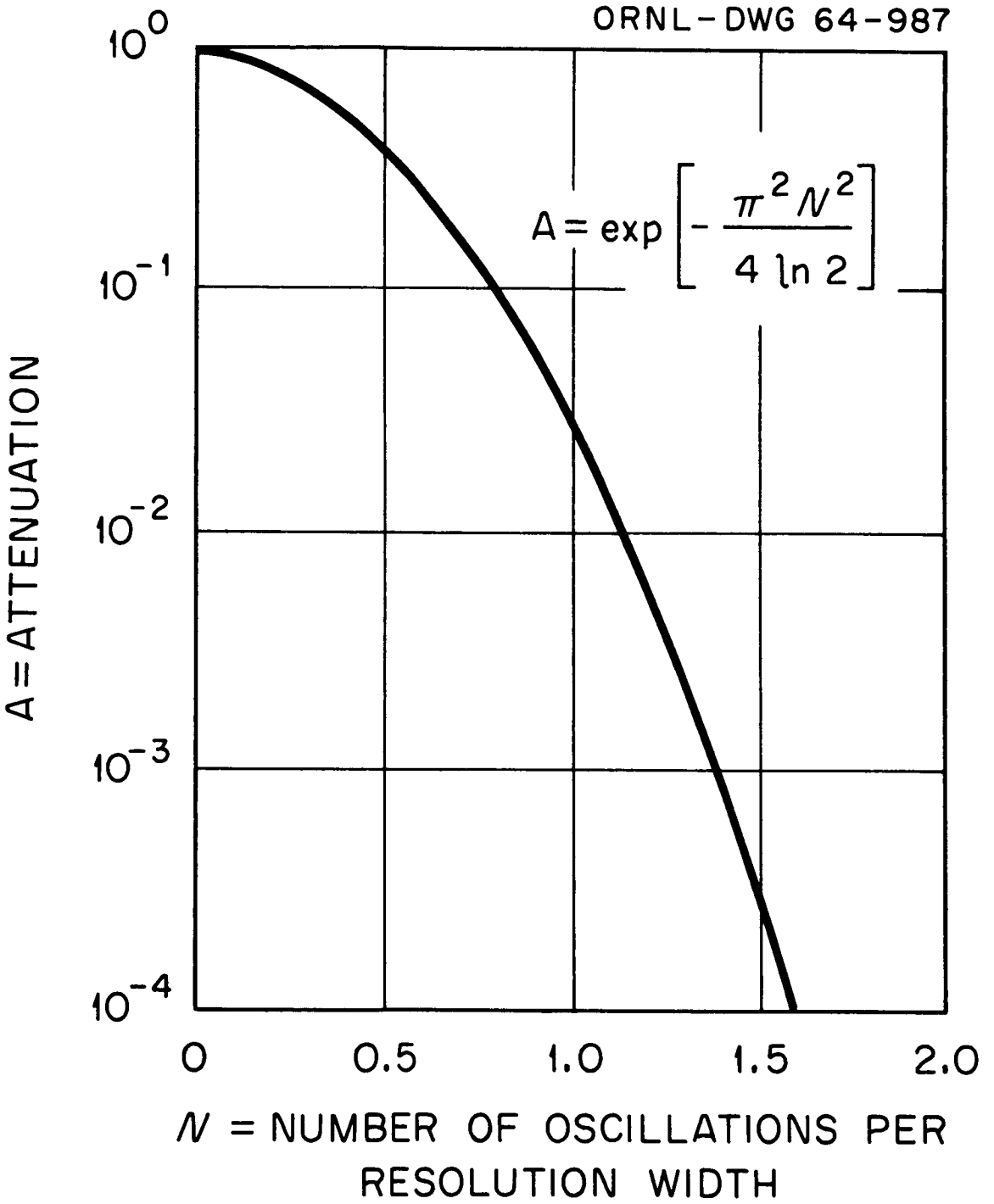


Fig. 2. The attenuation of a sinusoidal oscillatory component as a function of the number of oscillations per resolution width.

is the solution nonunique but also that any arbitrarily prescribed linear function of the solution is nonunique if both $K(t,s)$ and $b(t)$ are uncertain by a nonzero amount.

If it is impossible to unfold Eq. (1.2) in the presence of small uncertainties, then how can the previously successful results of the papers cited be explained? It is my contention that the success of those methods is based on the subtle use of certain information that is usually not recognized as such by the experimenter. In devising a workable scheme for interpreting the data, intuition apparently guides us to choose a mathematical process that includes some additional information. But however plausible these results seem, it is still desirable to uncover the underlying assumptions and to give error bounds that are as sharp as possible. Thus it is my purpose to attempt to implement Laplace's famous remark that statistics is common sense reduced to mathematics.

Although the general subject of unfolding is not yet mature in the sense that the underlying principles are commonly agreed upon, a rapidly increasing amount of work is being devoted to practical aspects of the problem. Thus it seems appropriate to review the basic ideas behind the statistical treatment and to try to put the current methods into an overall perspective. The unifying concept that will be used is the idea of constrained estimation by means of mathematical programming.

A priori information

By *a priori* information is meant that information about $x(s)$ which the experimenter has before he begins the experiment. For example, some information that the experimenter might know in advance is:

1. $x(s)$ is slowly varying with a few discrete peaks superimposed,
2. $x(s)$ cannot be negative,
3. $x(s)$ is bounded, with $x^{lo}(s) \leqq x(s) \leqq x^{up}(s)$.

A priori information such as item 2 is inherent in certain types of experiments. In the spectroscopy of neutrons or gamma rays, for example, the number of incident particles per unit energy cannot physically be negative. Almost always the experimenter has some information about the results which he knows with certainty. Thus, although a crystallographer may not know where in a molecule a certain atom is, he may know the maximum possible size for the molecule and thus know the location of a given atom to within, say, a fraction of a centimeter.

Remarkably, the simplest inequality constraint -- nonnegativity -- is often sufficient to constrain the arbitrary components in the solution to small values. In addition, it turns out that the nonnegativity can easily be extended to more general types of inequality constraints. For example, in a discrete space of n dimensions, several inequality conditions take one of the following forms:

nonnegativity:	$x_i \geqq 0, \ (i = 1, 2, \ldots n)$,
monotonic decreasing:	$x_1 \geqq x_2 \geqq x_3 \geqq \ldots \geqq x_n$,
bounded:	$lo_i \geqq x_i \geqq up_i \ (i = 1, 2, \ldots n)$.

These constraint regions are shown for a 2-dimensional space in Fig. 3. Of course, there may be a great deal of other information which the experimenter has about the nature of the solution. The practical problem is to include as much of this information as is necessary to yield a suitable result. Although in this paper the a priori conditions that

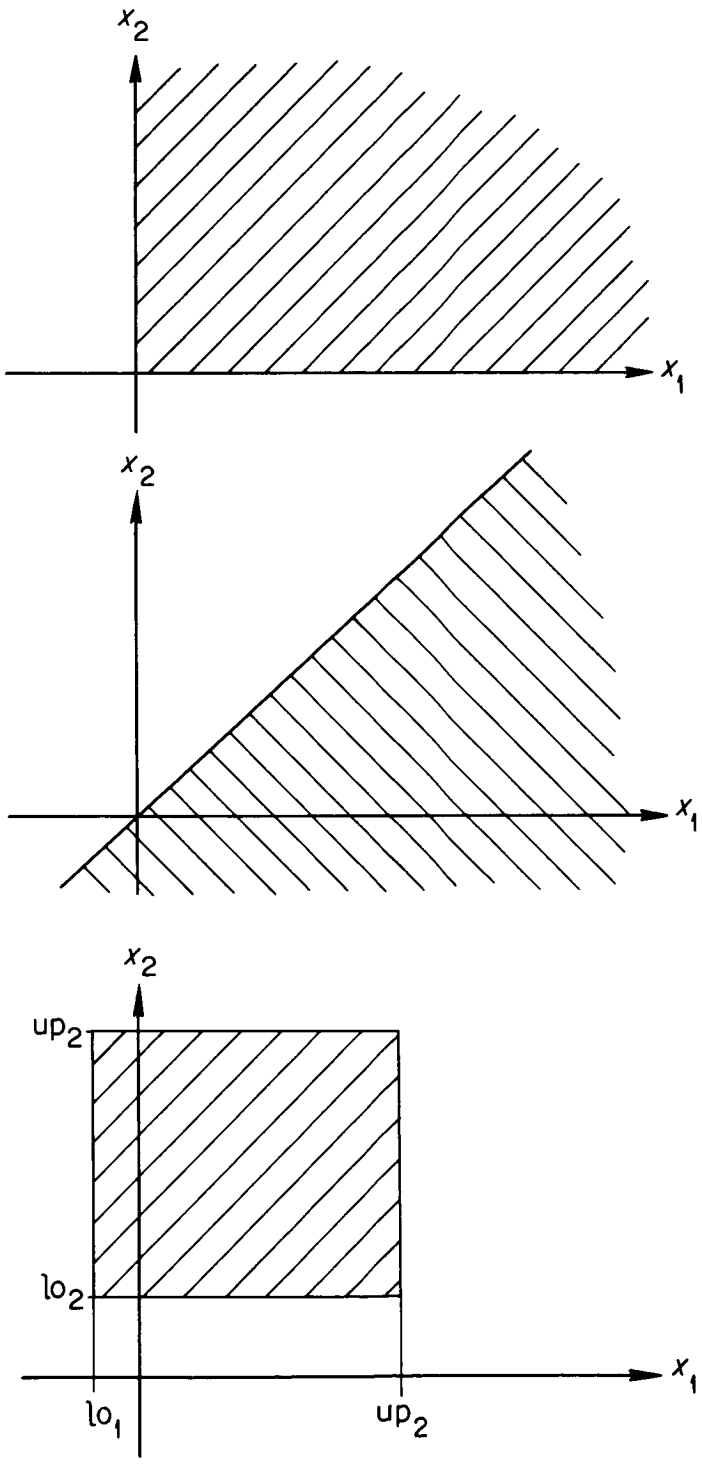
can be expressed as linear inequalities are stressed, many other kinds are possible. For example, it may be known from theoretical considerations that there must be a certain minimum spacing between discrete components in x(s). Such a case arises when analyzing the time decay spectrum (42) in the decay of a neutron population in a subcritical assembly.

Constrained estimation

By "unscrambling" is meant an extension of the ordinary ideas of estimation so that any *a priori* information that may be known can be considered in addition to the usual statistical information. In particular, it will be shown that confidence intervals for functions of x(s) can be found from the intersection of two sets. One is analogous to the statistical confidence region for ordinary regression, and the other to the *a priori* constraint set as shown in Fig. 3. Then unscrambling consists of two steps: (1) a series of questions about x(s) which must be posed in suitable mathematical form leading to a numerical value, and (2) the confidence interval which must be determined for each numerical value which takes into account all the available statistical information and all available *a priori* information.

In order to find the confidence interval, two extremal problems will be solved, one for each edge of the interval, by the techniques of mathematical programming. Mathematical programming is the maximization of a prescribed function with certain inequality constraints. In the present case, the constraints will be the ordinary statistical confidence region and the *a priori* constraint set, and the functions to be maximized will be determined by the questions posed concerning x(s).

ORNL-DWG 64-988



(*a*) NONNEGATIVE

(*b*) MONOTONIC DECREASING

(*c*) BOUNDED

Fig. 3. Two-dimensional regions corresponding to certain a priori inequality constraints.

Mathematical programming is relatively unfamiliar in the physical sciences because solutions generally have to be determined numerically and because the numerical calculations necessary have become economically feasible only since the development of large-scale computers.

Organization of report

Ordinary linear estimation theory is reviewed briefly in Chapter 2 and is extended to constrained estimation in Chapter 3, where it is shown that the constrained estimation problem can be posed as an extremal problem, soluble by the methods of mathematical programming, which reduces to ordinary estimation in the absence of any *a priori* information. How a problem with a continuous variable s may be reduced to discrete form is shown in Chapter 4; how it is put into a suitable form for solution by existing quadratic programming algorithms is shown in Chapter 5. Since the present quadratic programming algorithms are relatively inefficient for computation, Chapter 6 introduces an approximation which allows the problem to be solved by a linear programming algorithm.

In order to gain insight into the linear programming method and to develop some background necessary for an analysis of errors due to uncertainty in the kernel, some ideas concerning primal-dual relations in mathematical programming are introduced in Chapter 7. This material is the basis for the overall error analysis problem treated in Chapter 8.

In Chapter 9 another simplification is made which leads to a simple program but does not give as sharp results as the previous methods. Since any method must ultimately be evaluated on the basis of its success

in practical applications, several applications of the theory to problems in nuclear spectroscopy and applied mathematics are summarized in Chapter 10. Some extensions to the method which allow it to take into account other types of *a priori* conditions, such as smoothness, are given in Chapter 11. Conclusions from the present work and some predicted extrapolations to future use are given in the concluding chapter.

2. ORDINARY ESTIMATION

Linear estimation

A frequent approach to inverting (or unfolding) Eq. (1.2) for $x(s)$ is to replace the integral equation with a set of simultaneous linear equations and to determine the solution $x = (x_1, x_2, \ldots x_n)$ and variance of the solution, $\sigma^2(x_1)$, $\sigma^2(x_2)$, ... $\sigma^2(x_n)$, by least-squares analysis. This is the standard problem of "linear estimation" in statistics (39,40). In order to obtain a finite set of equations, a set of $s_j = (s_1, s_2, \ldots s_n)$ can be selected. Then

$$\begin{aligned} K_1(s_1)\, x_1 + K_1(s_2)\, x_2 + \ldots + K_1(s_n)\, x_n &= b_1 \\ \ldots \qquad \ldots \qquad \ldots \qquad \ldots \qquad &\ldots \\ K_m(s_1)\, x_1 + K_m(s_2)\, x_2 + \ldots + K_m(s_n)\, x_n &= b_m \end{aligned} \tag{2.1.1}$$

or, in matrix form,

$$Kx = b \ , \tag{2.1.2}$$

where

$$K_{ij} = K_i(s_j) \qquad \begin{aligned} i &= 1, 2, \ldots m \ , \\ j &= 1, 2, \ldots n \ . \end{aligned} \tag{2.1.3}$$

Equations (2.1) are indeterminate if $n > m$, can be solved by ordinary matrix inversion if $n = m$, or can be solved by least-squares analysis if $n \leq m$. If $n < m$, it may not be possible to find a solution $x = (x_1, x_2, \ldots x_n)$ which satisfies all m equations when $\hat{b}$ is used as

the right-hand side, because of the statistical errors in $\hat{b}$.[1] But a solution x can always be determined so that the sum of squares of the residuals is minimum [with the square of the residuals weighted by $(1/\sigma_i^2)$ in order to take proper account of the varying accuracy of the b_i]. In matrix form[2] the weighted sum of the square of the residuals

$$\sum_{i=1}^{m} [(Kx - \hat{b})_i/\sigma_i]^2$$

may be written

$$\epsilon^2 = (Kx - \hat{b})\ S^{-1}\ (Kx - \hat{b})\ , \qquad (2.2)$$

where S is the variance matrix of b. The diagonal elements S_{ii} are the square of the standard deviation of b_i, and the off-diagonal elements S_{ij} are the covariance between b_i and b_j. By the definition of the variance matrix, $S_{ij} = S_{ji}$ and thus S is symmetric. In many cases, the elements of b are statistically independent so that the covariance components are zero and S reduces to a diagonal matrix. The present results, however, hold in general.

To find the value of $x = \hat{x}$ which minimizes ϵ^2, the derivative of ϵ^2 with respect to each component of x is set equal to zero. Then

$$\frac{d\epsilon^2}{dx_j} = 2[K^T\ S^{-1}\ (Kx - \hat{b})]_j = 0 \quad \text{for } j = 1, 2, \ldots n\ . \qquad (2.3)$$

---

[1]The notation ^ over a letter indicates an estimate of the random vector.

[2]The subscript i denotes the ith element of a vector.

The easily verified rule that $d(y^T Qy)/dy_j = 2\ (Qy)_j$ is used for matrix differentiation. Rearranging Eq. (2.3) gives the "normal equations" of least-squares analysis

$$(K^T\ S^{-1}\ K)\ \hat{x} = K^T S^{-1}\ \hat{b} \tag{2.4}$$

which has the formal solution

$$\hat{x} = (K^T\ S^{-1}\ K)^{-1}\ K^T\ S^{-1}\ \hat{b} \tag{2.5}$$

if $(K^T\ S^{-1}\ K)$ is nonsingular.

The variance matrix $V(x)$ of x has diagonal elements equal to

$$E(\hat{x} - x)_i^2\ ,$$

and off-diagonal elements equal to

$$E(\hat{x} - x)_i\ (\hat{x} - x)_j\ ,$$

where the operator E gives the average over the statistical distribution of its operand.

The variance matrix of x is easily found in terms of S:

$$\begin{aligned} V(\hat{x}) &= E[(K^T\ S^{-1}\ K)^{-1}\ K^T\ S^{-1}\ (\hat{b} - b)][(K^T\ S^{-1}\ K)^{-1}\ K^T\ S^{-1}\ (\hat{b} - b)]^T \\ &= (K^T\ S^{-1}\ K)^{-1}\ K^T\ S^{-1}\ [E(\hat{b} - b)(\hat{b} - b)^T]\ S^{-1}\ K\ (K^T\ S^{-1}\ K)^{-1} \\ &= (K^T\ S^{-1}\ K)^{-1}\ . \end{aligned} \tag{2.6}$$

The fact that $E[(\hat{b} - b)(\hat{b} - b)^T] = S$ and that $(K^T S^{-1} K) = (K^T S^{-1} K)^T$ is used in the above equations. Thus the variance of $\hat{x}_j$ is given by the diagonal elements of $(K^T S^{-1} K)^{-1}$, and, more generally, the variance of any linear combination

$$\hat{\phi} = c_1\hat{x}_1 + c_2\hat{x}_2 + \ldots + c_n\hat{x}_n = c^T\hat{x} \tag{2.7}$$

is given by

$$\sigma^2(\hat{\varphi}) = E[(\hat{\varphi} - \varphi)^2] = c^T V(\hat{x})\, c = c^T (K^T S^{-1} K)^{-1} c \quad . \tag{2.8}$$

Note that estimation of the linear combination φ includes estimation of a single x_j if $c_j = 1.0$ and all other c's = 0.

Confidence intervals

If the spectrum $x(s)$ is known in advance to be discrete, with component only at $s_1, s_2, \ldots s_n$, then the solutions $\dot{x}_j$ to the least-squares problem are the intensities of the discrete (delta function) components, and confidence intervals can be constructed which will give a measure of the trust that can be put in the estimate. The confidence interval for a linear function, φ, is given by

$$\Pr[\hat{\varphi} - \kappa\, \sigma(\hat{\varphi}) \le \varphi \le \hat{\varphi} + \kappa\, \sigma(\hat{\varphi})] = \alpha \quad , \tag{2.9}$$

where κ depends upon the confidence level coefficient α. The probability interpretation of the confidence interval $[\hat{\varphi}^{lo}, \hat{\varphi}^{up}]$ is that on repeated measurements under the same conditions a fraction α of the confidence intervals will include the true value φ. If the $\hat{b}_i$ are independently normally distributed with known variance, then as is well known (40) the confidence coefficient is related to κ by:

$$\alpha = \frac{1}{\sqrt{2\pi}} \int_{-\kappa}^{\kappa} e^{-\frac{1}{2}\xi^2}\, d\xi \quad . \tag{2.10}$$

A few specific values of interest are given below:

α, confidence level	κ
0.6827	1.0
0.9545	2.0
0.9973	3.0
0.9999	4.0

Of great importance is the question of the proper interpretation to give to the components x_j of the linear estimation problem when the spectrum $x(s)$ is not known to be discrete, but may be continuous or have continuous parts. Equations (2.5) and (2.7) show that the estimate of any combination of the components x_j given by $c^T x$ may be expressed in terms of the experimental values of b_i by

$$\hat{\varphi} = c^T \hat{x} = \left[c^T (K^T S^{-1} K)^{-1} K^T S^{-1} \right] \hat{b} \quad . \tag{2.11}$$

The dependence of φ upon $x(s)$ can now be seen by noting that each component b_i is related to the spectrum $x(s)$ by $K_i(s)$. Thus the combination of b's given by Eq. (2.11) is related to $x(s)$ by

$$\varphi = \int L(s)\, x(s)\, ds \quad , \tag{2.12}$$

where

$$L(s) = \sum_{i=1}^{m} u_i \, K_i(s) \quad , \tag{2.14}$$

with

$$u^T = c^T (K^T S^{-1} K)^{-1} K^T S^{-1} \quad . \tag{2.15}$$

Because of their special importance, these coefficients u_i, which appear in the brackets of Eq. (2.11), are called "coefficients of combination" and the function $L(s)$ is called the "sensitivity function."

My overall point of view toward estimation is based strongly on the idea of seeking a combination of the experimental data points, $\hat{b}_i$, such that the combination estimates some function of $x(s)$ which is of interest to the experimenter. Generally, the experimenter will be interested in many such functions, and a different set of u_i will be determined for each one. The linear estimation procedure can be thought of as a method for determining a set of coefficients which synthesize the natural response functions $K_i(s)$ of the measuring instrument into the new combination $L(s)$, which more closely approximates the desired functions. The sensitivity function concept is discussed in many statistics books in a more general context, and has recently been emphasized in the analysis of photonuclear cross sections by Penfold and Leiss (43).

Figure 4 is a sketch of the sensitivity function corresponding to $\phi = x_4 + x_5$ for typical least-squares analysis with $m \gg n$. Note that the sensitivity function is 1.0 at s_4 and s_5 and is 0.0 at all other values of s_j. However, it is <u>not</u> zero at all s, but only at those that were selected when the equations were set up. Thus, as required, the least-squares analysis gives the correct result if the only possible components of $x(s)$ are known in advance to be located at $s = (s_1, s_2, \ldots s_n)$. In statistics, this requirement that the sensitivity function agree with the coefficients c in Eq. (2.7) at the values of s_j is known as the "unbiased condition." In terms of the coefficients of combination u, it takes the algebraic form

$$c^T - u^T K = 0 , \tag{2.18}$$

where the matrix elements are evaluated according to Eq. (2.1.3) at $s = (s_1, s_2, \ldots s_n)$.

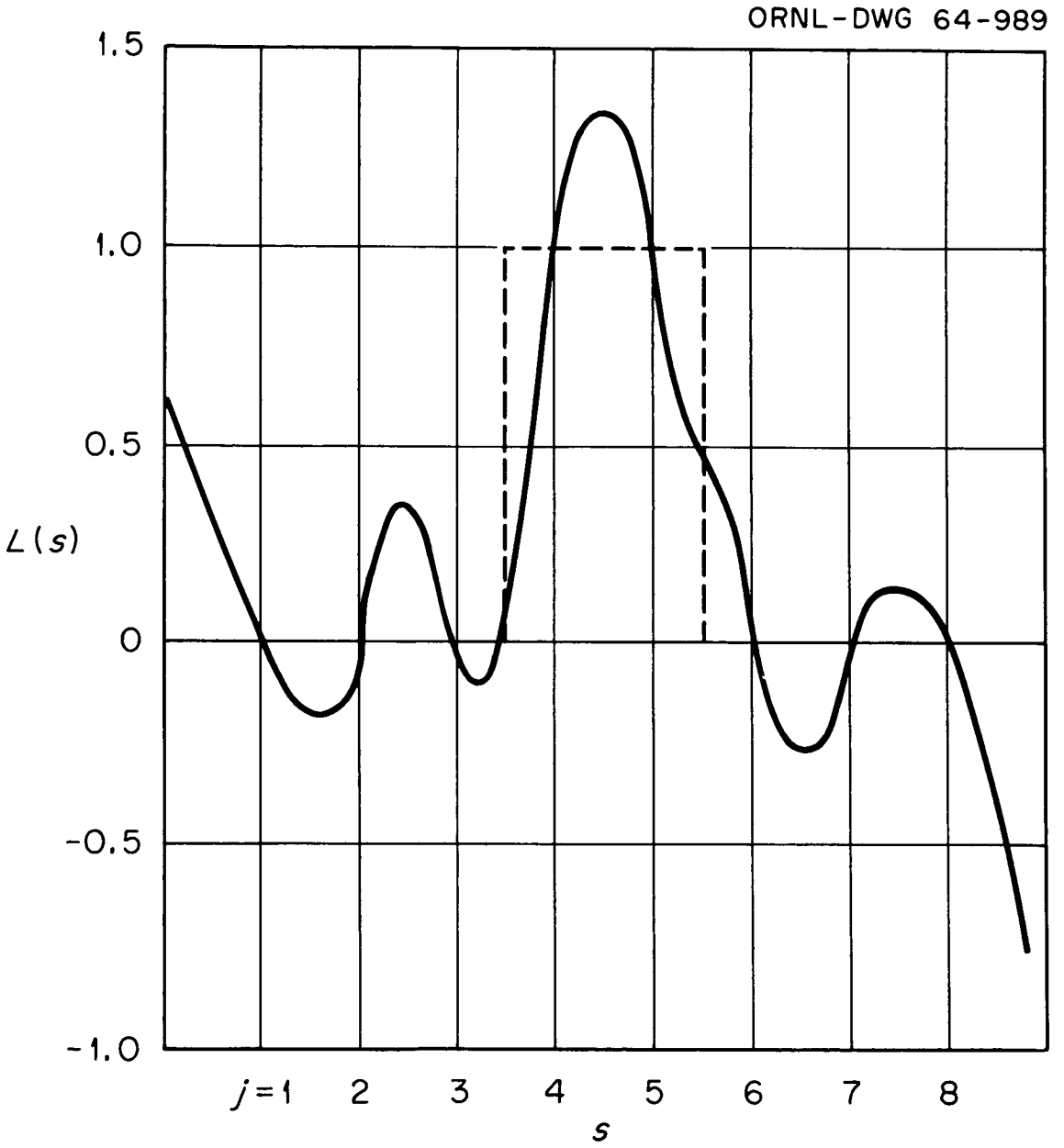


Fig. 4. Sensitivity function L(s) for an estimate of $\Phi = (1\ 1)\begin{pmatrix} x_4 \\ x_5 \end{pmatrix} = x_4 + x_5$.

The famous Markov dual formulation of the least-squares problem states that the linear estimation may be approached as follows:

$$\left\{ \begin{array}{l} \text{Select coefficients of combination } u = (u_1, u_2, \ldots u_m) \text{ so that} \\ \sigma^2(\hat{\phi}) = u^T S u \text{ is minimized with the constraint that } u^T K - c^T = 0. \end{array} \right\} \tag{2.19}$$

This minimization problem can easily be solved by use of Lagrange multipliers and leads to the following equations:

$$\ell = u^T S u - (u^T K - c^T) p \ , \tag{2.20.1}$$

$$\frac{\partial \ell}{\partial u_i} = 0, \qquad i = 1, 2, \ldots m \tag{2.20.2}$$

$$\frac{\partial \ell}{\partial p_k} = 0, \qquad k = 1, 2, \ldots n, \tag{2.20.3}$$

where $p = (p_1, p_2, \ldots p_n)$ is a vector of Lagrange multipliers.

By solving Eq. (2.20.2) for u and eliminating the Lagrange multipliers between the result and Eq. (2.20.3),

$$u^T = c^T (K^T S^{-1} K)^{-1} K^T S^{-1} \ , \tag{2.21}$$

which is the same result obtained when the weighted sum of squares of the residuals is minimized. This dual approach shows that using $1/\sigma_i^2$ weights is superior to using all other possible weights (such as equal weighting, etc.) if the objective is a small variance in the resulting estimates $\hat{\phi}$.

A heuristic interpretation of the Gauss-Markov theorem is that m rows of K are sufficient to satisfy the constraints $u^T K = c^T$, but if there are more than sufficient rows, there are many possible ways of

selecting $u = (u_1, u_2, \ldots u_m)$ so that the constraint holds. The least-squares method (according to the Markov theorem) selects that particular combination which results in the least variance.

Histogram approximation

The least-squares method has sometimes been applied in the expectation that the estimates $x_1, x_2, \ldots x_n$ would form a histogram approximation to the spectrum $x(s)$, with

$$x_j = \int_{\frac{1}{2}(s_j + s_{j-1})}^{\frac{1}{2}(s_j + s_{j+1})} x(s)\, ds \quad . \tag{2.22}$$

This is true only to the extent that the sensitivity functions for each x_j have the rectangular step form

$$L(s) = \begin{cases} 1 & \text{if} \quad s_i \le s \le s_{i+1} \\ 0 & \text{otherwise} \end{cases} \quad . \tag{2.23}$$

A comparison of this ideal histogram function with the sensitivity function resulting from a typical least-squares analysis is shown in Fig. 5. In this case, the histogram approximation is quite poor, since least squares ignores all points between the selected points $s_j = (s_1, s_2, \ldots s_n)$. An improvement up to a point in the histogram approximation may be obtained by increasing the number of points to be matched, as in Fig. 4. But when the number of points is increased, the variance begins to increase, although the sensitivity function may more nearly

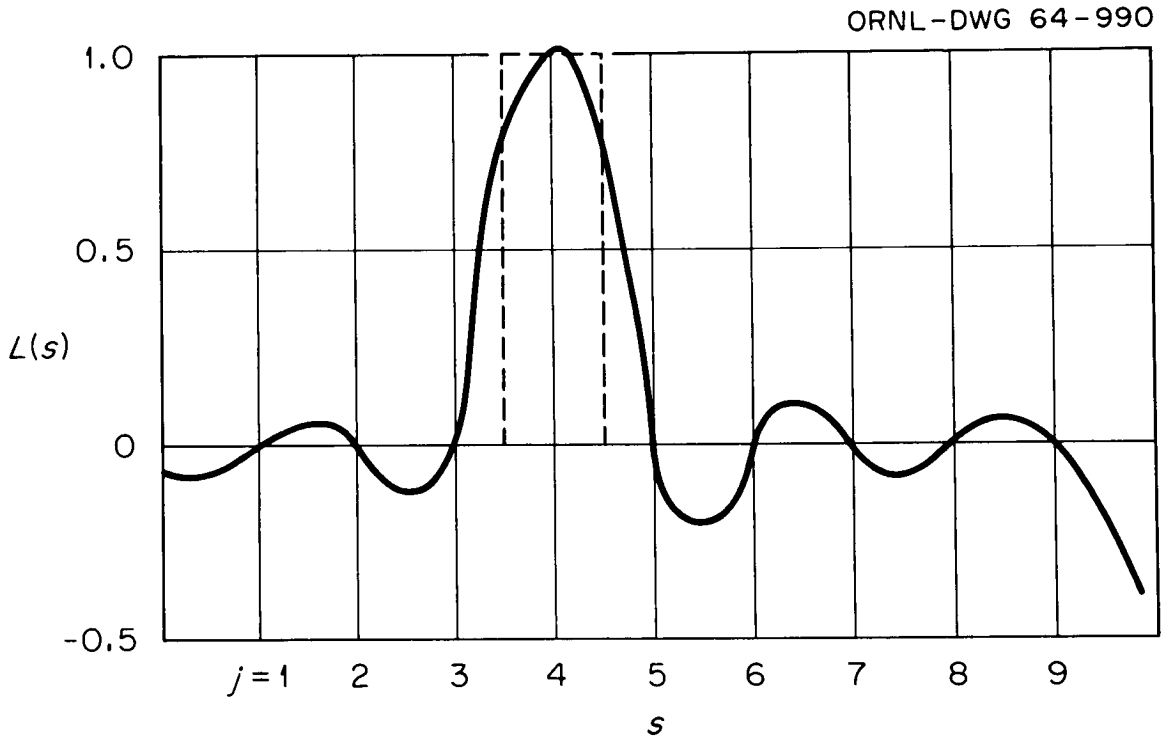


Fig. 5. Comparison of ideal histogram function with edges at $1/2(s_3 + s_4)$ and $1/2(s_4 + s_5)$ and the sensitivity function L(s) resulting from least-squares analysis.

agree with the histogram function. Practically, it becomes increasingly difficult to pass the combination $\sum_{i=1}^{m} u_i K_i(s)$ through n specified points $c = (c_1, c_2, \ldots c_n)$ at $s = (s_1, s_2, \ldots s_n)$ as the number of points increases, unless the coefficient vector c happens to be an exact combination of only a few of the rows of K. Similarly, as n increases, the least-squares matrix $K^T S^{-1} K$ becomes more poorly conditioned until it finally becomes singular when $n > m$. The increase in variance is due to an increase in the magnitude of coefficients of combination u_i given by Eq. (2.21) which are necessary to achieve the unbiased condition at the match points $s = (s_1, s_2, \ldots s_n)$.

A conceptual barrier to obtaining an insight into the minimum variance form of least-squares estimation is the tendency by experimenters to think of $K_i(s)$ exclusively as a function of i since $K_i(s)$ vs i is the pulse-height distribution obtained for a monoenergetic test spectrum. But the functions $K_i(s)$ vs s may equally well be interpreted as the counting efficiency of the ith count bin of a multichannel scintillation spectrometer for a unit intensity source of continuously adjustable energy s. Using the example of an organic scintillator neutron spectrometer, graphs of $K_i(s)$ vs i for several values of s and vs s for several values of i are shown in Figs. 6 and 7, respectively.

Window functions

It has been shown that ordinary linear estimation leads to estimates of quantities of the form

$$\Phi = \int \left[\sum_{i=1}^{m} u_i K_i(s) \right] x(s)\, ds \,, \qquad (2.24)$$

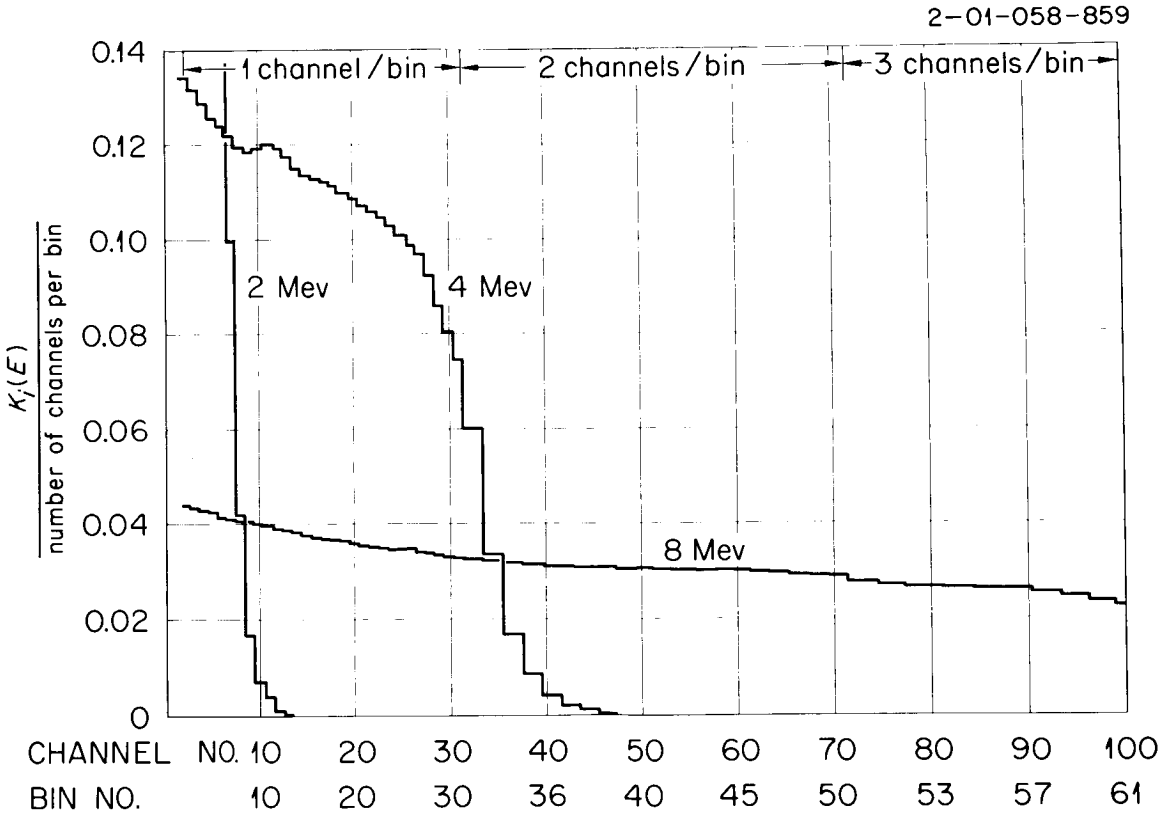


Fig. 6. Pulse-height distributions for monoenergetic neutrons incident on side of 2-in.-diam, 2-in.-long cylindrical NE-213 scintillator.

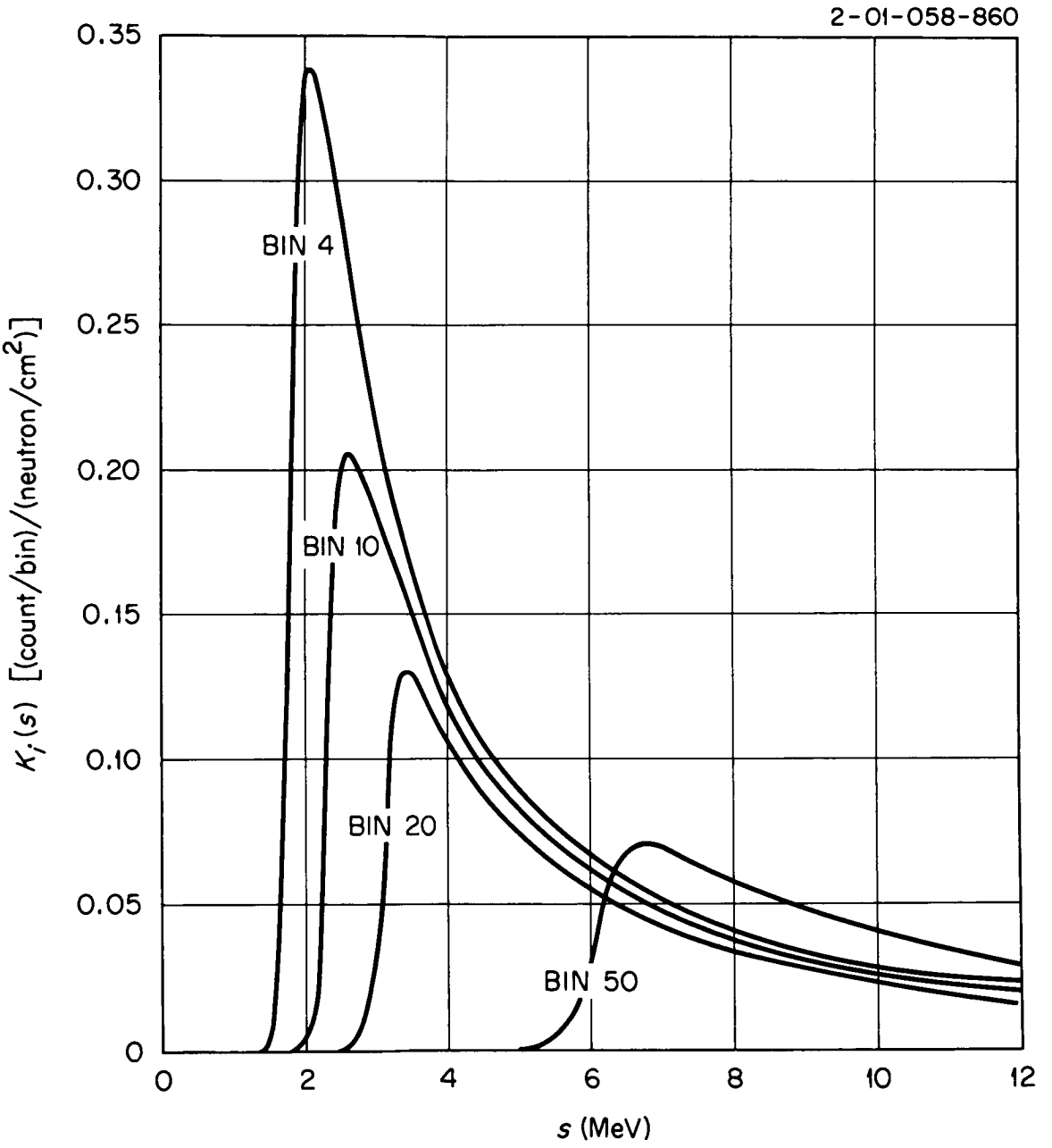


Fig. 7. Response functions $K_i(s)$ vs s with the bin No. i as a parameter. Bins with $i > 40$ are twice as wide as bins with $i \leq 40$.

where the sensitivity function $\sum_{i=1}^{m} u_i K_i(s)$ is given by a linear combination of the response functions. Also, it is not possible to obtain a rigorous histogram form of the spectrum from ordinary linear estimation because the required corners on the histogram function cannot be fitted by the smoothly varying $K_i(s)$.

An alternative to obtaining a histogram fit is to consider a more general approach of obtaining a set of estimates corresponding to "window functions." These window functions are denoted by w(s) and are the functions for which it is desired to obtain an estimate of

$$\phi = \int w(s)\ x(s)\ ds\ . \tag{2.25}$$

In other words, ϕ is the response of an ideal detector with a response window of "transmission" w(s). Usually, experimental considerations will suggest desirable forms for the window functions. For example, a set of conceivable window functions is shown in Fig. 8. In this set, there are several windows of the band-pass type. In addition, there is a window equal to a constant and a window proportional to s which will yield, respectively, the zero and first moments of x(s) More generally any function of interest could be included, such as a flux-to-dose conversion, an activation cross section, etc.

From the minimum variance interpretation of ordinary estimation, the problem is to obtain unbiased estimates to each of these windows at selected match points $s = (s_1, s_2, \ldots s_n)$ such that the variance in the resulting estimates is minimized. But generally the variance is infinite if $n > m$, unless the window function turns out to be an exact

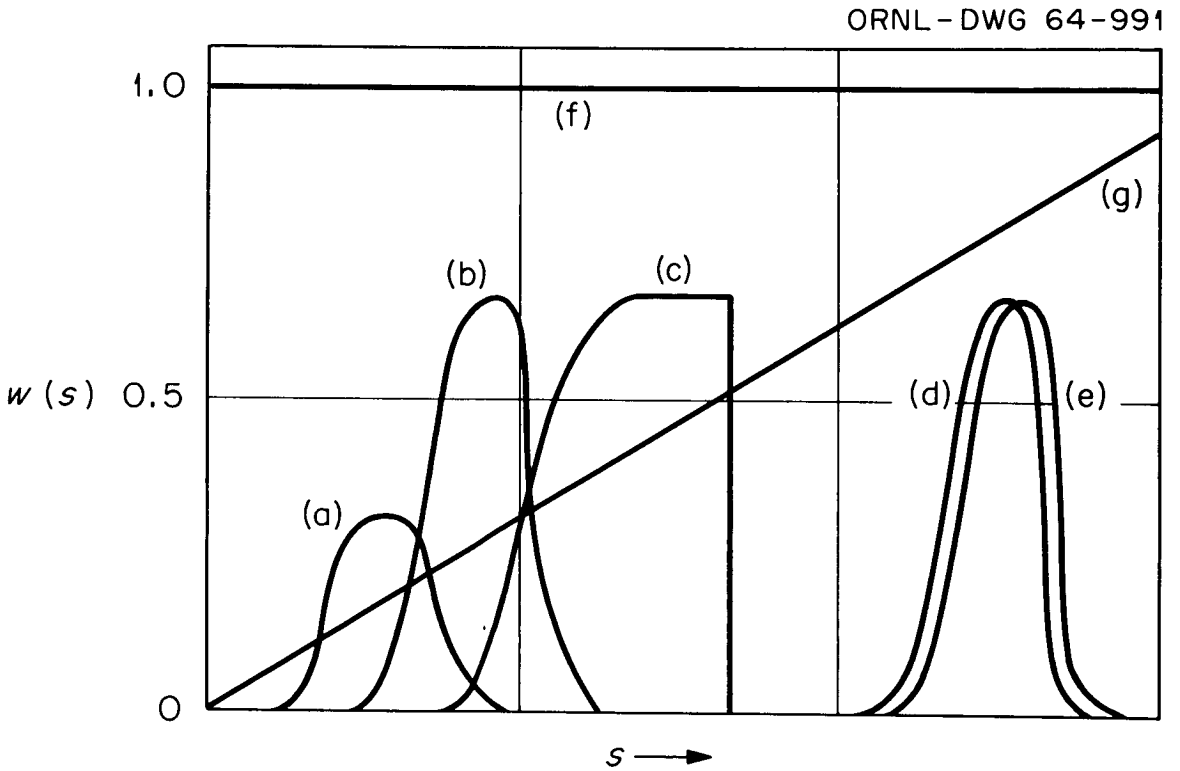


Fig. 8. A set of window functions: (a) - (e) yield band-pass filter type of estimates, $\Phi = \int w(s)\, x(s)\, ds$; (f) is a constant which yields an estimate of the zero moment, $\Phi = \int x(s)\, ds$; and (g) yields an estimate proportional to the first moment, $\Phi = \int s\, x(s)\, ds$.

combination of the response functions $K_i(s)$ so that a match at n points ensures a match at all points. Thus the function chosen for a window cannot be just any function, but must be an estimable function. Usually, an estimable function cannot be guessed before the analysis; instead, a set of plausible window functions must be proposed, values $w_j = (w_1, w_2, \ldots w_n)$ must be picked at the match points, and then the sensitivity functions found that correspond to each window. If the analysis is to be meaningful, the sensitivity functions should be close to the desired window functions at all values of s. If they are not, either new window functions or a new set of match points must be selected. Heuristically $w(s)$ may be thought of as the response function of a conceptual "non-existent" instrument that the experimenter wishes to estimate the response for, based on the actual instrument response $\grave{b}$. $L(s)$ is the "best" approximation which can be formed from the actual instrument response functions $K_i(s)$.

Errors in the matrix

In ordinary least-squares analysis, errors in the elements of the matrix K can be included in the estimates of the confidence interval if the problem is not underdetermined by expanding the matrix about the given values and carrying only first-order terms in the error. But when the problem is underdetermined, even if the windows are an exact combination of q of the rows of the matrix, any error in one of the q rows will yield an infinite error for the estimate.

3. CONSTRAINED ESTIMATION

We have seen that the confidence intervals obtained by the usual least-squares approach to unfolding instrument response have a straightforward interpretation if the unknown spectrum is known in advance to be discrete. But if there is a possibility that the spectrum is continuous, or partially continuous, then only certain special functions can be estimated of the form

$$\phi = \int \left[\sum_{i=1}^{m} u_i K_i(s) \right] x(s)\, ds ,$$

where $\sum_{i=1}^{m} u_i K_i(s)$ is a linear combination of the response functions and is called the "sensitivity function." If estimation of some particular function is desired, for example, obtaining a histogram representation of the spectrum or obtaining the integral of $x(s)$ weighted with some desired function $w(s)$, a sensitivity function cannot be obtained which agrees with the desired function at more than r points (where r is the rank of the estimation matrix), unless the window function is an exact linear combination of some of the rows of K. Generally, however, if agreement at r points is forced, the variance in the resulting interval is very large, and a smaller number of points of agreement must be selected.

Experimentalists have long felt that this type of problem <u>ought</u> to have a sensible solution, and various ad hoc techniques have been devised. The resulting methods have frequently embroiled statisticians and experimentalists in lively controversy. Usually the experimentalist

has been content to publish reasonable results as a point estimate without any assigned error or to give the error based on the propagation of counting errors only, ignoring the shape of the sensitivity function [and, with a very few exceptions, (43), never calculating it or realizing that it existed].

Inequality method

The key to these problems in the case of the scintillation spectrometer and similar situations is to drop the unbiased condition, which is for continuous s of the form

$$w(s) = \sum_{i=1}^{m} u_i K_i(s)$$

and replace it with the pair of inequalities

$$\sum_{i=1}^{m} u_i^{lo} K_i(s) \leqq w(s) \leqq \sum_{i=1}^{m} u_i^{up} K_i(s) . \qquad (3.1)$$

This new requirement is shown graphically in Fig. 9, where the single sensitivity function has been replaced by upper and lower sensitivity functions. Then the estimate corresponding to the desired window function lies in the interval between the estimates corresponding to the upper and lower sensitivity function. Thus the desired (possibly inestimable) function is bracketed between two estimable functions: the upper sensitivity function for obtaining the upper confidence limit and the lower sensitivity function for obtaining the lower confidence limit. Thus

$$\Pr\left[\hat{\phi}^{lo} - \kappa\sigma\,(\hat{\phi}^{lo}) \leqq \phi \leqq \hat{\phi}^{up} + \kappa\sigma\,(\hat{\phi}^{up})\right] \geqq \alpha , \qquad (3.2)$$

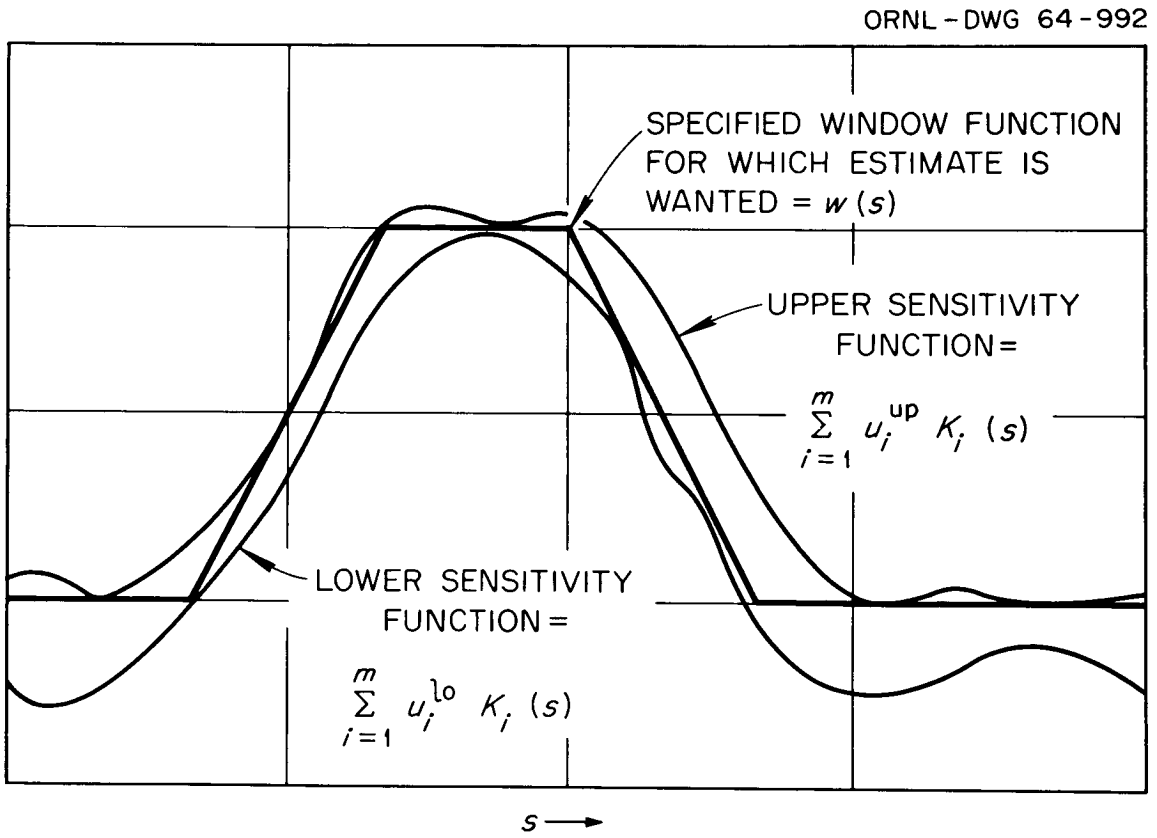


Fig. 9. Bracketing of an inestimable window function between two estimable functions.

where the constant κ depends upon the confidence level desired. Consideration of the exact value of κ to attain a specified confidence level is deferred until Chapter 7, where the relation of the inequality method to ordinary estimation is shown.

With the inequality method, the window functions can be chosen arbitrarily. Of course, very narrow or very abrupt functions will not be bracketed as closely as smoother broader functions, but the difference is now one of degree and there is no longer any fundamental distinction between obtaining a confidence interval for an estimable function and an inestimable function. Notice, however, that the equality in Eq. (2.9) has been replaced by an inequality in Eq. (3.2), since any number of different upper and lower sensitivity functions can be obtained which yield different estimates.

A priori information

At first sight, it does not seem that any new information has been used in the inequality method. However, in order for the upper sensitivity function to yield an upper estimate for Φ, the spectral function must be nonnegative, $x(s) \geq 0$. Otherwise, the estimate based on the upper sensitivity function might be smaller than that based on the window function. Thus the nonnegativity of the function $x(s)$ underlies the inequality method, and the improvement over ordinary estimation is due to the extra freedom this information provides. That this extension to ordinary least squares is capable of yielding useful results is perhaps most easily illustrated by unfolding a pulse-height distribution obtained with the organic neutron scintillation spectrometer whose response is

shown in Figs. 6 and 7. The window functions used were Gaussian with a relative full width at half maximum (FWHM) of 20% in energy. The pulse-height distribution (after combining channels into larger bins) is shown in Fig. 10, and the unfolded "spectrum" is shown in Fig. 11. The interpretation of the intervals on Fig. 11 is that they are confidence intervals for the response of Gaussian-shaped band-pass filters (or windows) centered at the indicated energy.

Narrowest possible confidence limits

It has been shown that if $x(s) \geqq 0$, a confidence interval can be obtained for any quantity of the form

$$\phi = \int w(s)\ x(s)\ ds \tag{3.3}$$

provided that a pair of estimable functions can be found which brackets $w(s)$. The estimable functions must satisfy

$$\sum_{i=1}^{m} u_i^{up}\ K_i(s) \geqq w(s) \tag{3.4.1}$$

and

$$\sum_{i=1}^{m} u_i^{lo}\ K_i(s) \leqq w(s)\ . \tag{3.4.2}$$

Although a set of upper and lower confidence limits for different window functions is of interest, for brevity just one such window function $w(s)$ and just the upper limit of the confidence interval will be considered. In order to find the lower limit of the confidence intervals, the upper limit of the interval is found for $-w(s)$, which will be the lower

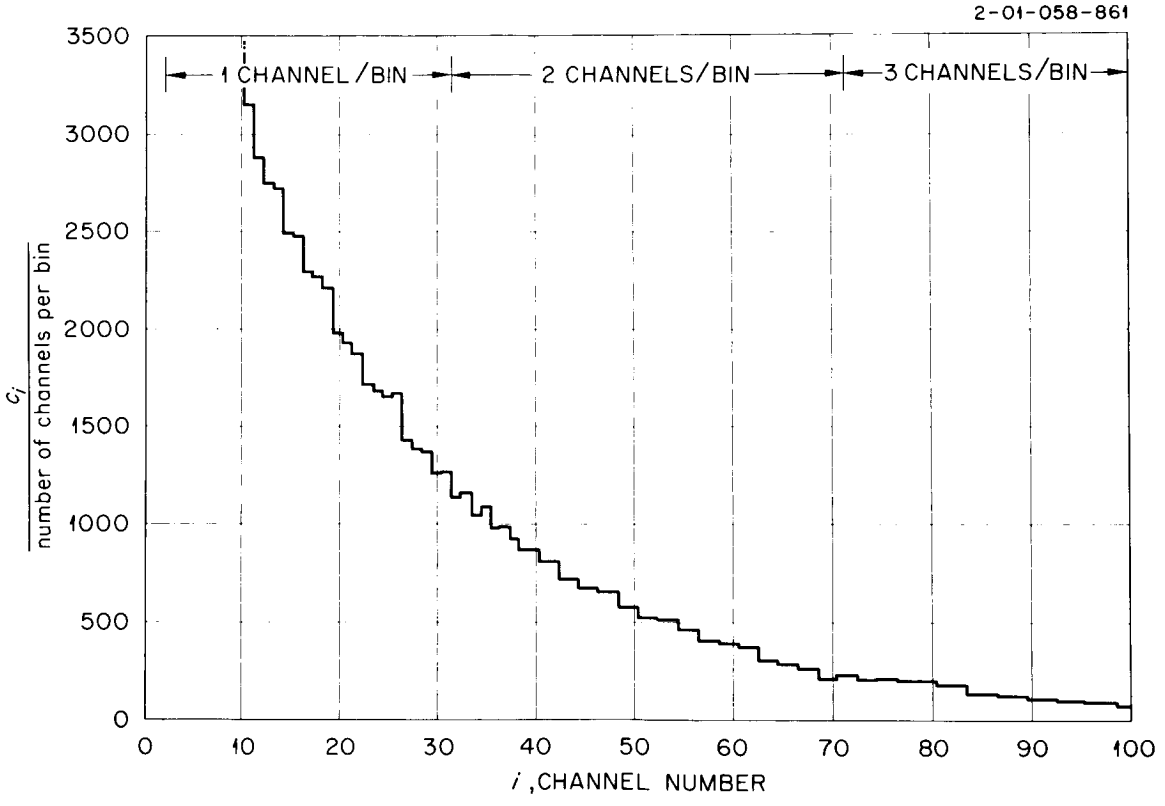


Fig. 10. Pulse-height distribution for Po-Be neutrons incident on NE-213 organic scintillation detector.

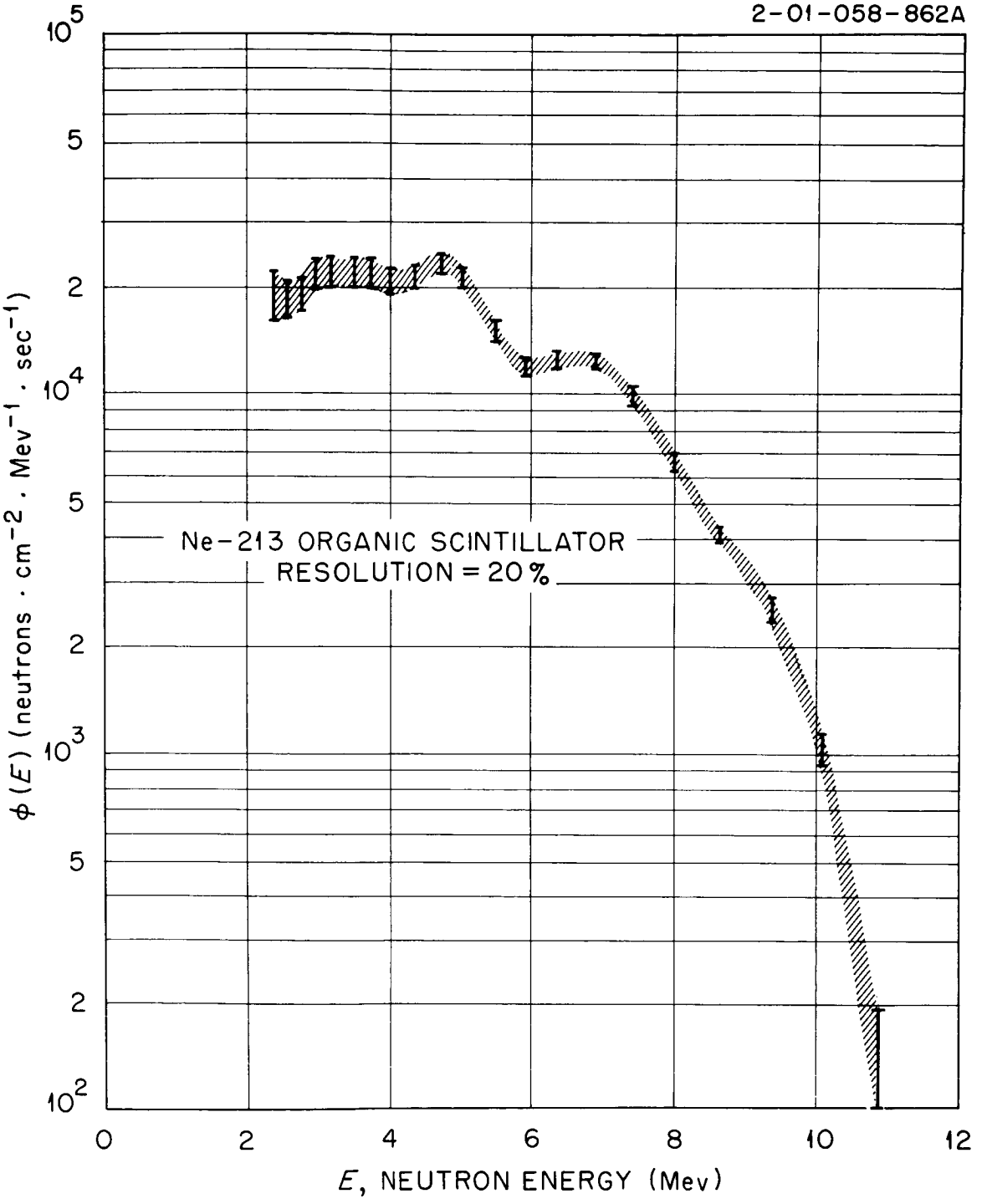


Fig. 11. Unfolded spectra obtained with Gaussian window functions of 20% FWHM.

limit for w(s) with reversed sign. Thus the following methods could be applied to a double set of window functions, one set taken with the original sign, and the other with reversed sign.

From Eqs. (3.2), (2.10), (2.13), and (2.19), the upper limit of the confidence interval given by the upper sensitivity function is given by

$$\hat{\Phi} = u^T\hat{b} + \kappa \sqrt{u^T S u}\,, \tag{3.5.1}$$

with

$$\sum_{i=1}^{m} u_i K_i(s) \geqq w(s) \;. \tag{3.5.2}$$

However, since more than one coefficient of combination vector u can generally be found which will satisfy the constraint, among all the possible coefficient vectors $u = (u_1, u_2, \ldots u_m)$ the u^* which results in the least upper-confidence-interval limit will be be chosen. In this selection process a compromise must be made between two conflicting desires: to approximate the window w(s) as closely as possible by the sensitivity function and at the same time keep the coefficients small, since then $\sum_{i=1}^{m} u_i^2 \sigma_i^2$ will be small. The selection problem may be rephrased as the following extremal problem:

Find a coefficient vector $u^* = (u_1^*, u_2^*, \ldots u_m^*)$ which minimizes

$$\sum_{i=1}^{m} u_i \hat{b}_i + \kappa \sqrt{\sum_{i=1}^{m} u_i^2 \sigma_i^2} \quad \text{with} \quad \sum_{i=1}^{m} u_i K_i(s) \geqq w(s) \;. \tag{3.6.1}$$

More compactly, this problem may be written as

$$\hat{\phi}^{up} = \min_{u} \left\{ u^T \hat{b} + \kappa \sqrt{u^T S u} \;\middle|\; \sum_{i=1}^{m} u_i \, K_i(s) \geq w(s) \right\} . \qquad (3.6.2)$$

In order to get this into a convenient form for numerical solution, it must first be reduced to a finite dimensional problem. Fortunately, the inequality method provides a natural solution to this problem, which is discussed in the following section.

4. REDUCTION TO FINITE DIMENSION

In the preceding section the constrained estimation problem was formulated as an extremal problem in terms of the m-dimensional vectors u and b. Only the constraint relation

$$\sum_{i=1}^{m} u_i K_i(s) \geqq w(s) \tag{4.1}$$

involved the continuous variable s. Before the edges of the confidence interval can be computed, a way must be found to keep the continuous constraint satisfied by only a finite computational method.

The simplest solution is to replace the window function w(s) by an upper linear approximation with

$$w(s) \leqq w(s)^{up} , \tag{4.2}$$

as shown in Fig. 12, and to replace each of the response functions $K_i(s)$, for i = 1, 2, ... m, by a pair of linear approximations which bracket it from above and below:

$$K_i(s)^{lo} \leqq K_i(s) \leqq K_i(s)^{up} , \tag{4.3}$$

as shown in Fig. 13. Then an inequality can be established in terms of these linear approximations which need only hold at the end points s_1, s_2, ... s_n of the linear segments to establish <u>a fortiori</u> that Eq. (4.1) holds for the original continuous functions at all s. First it is observed that

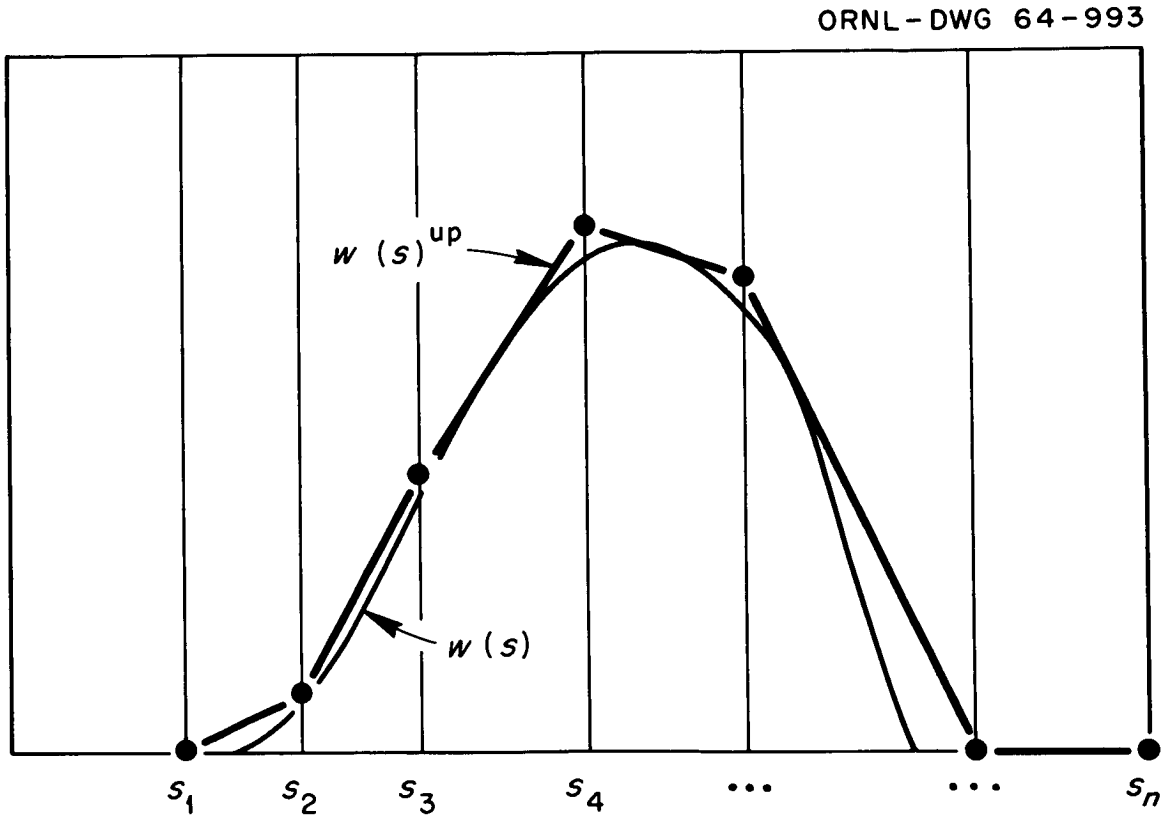


Fig. 12. Replacement of the continuous window function w(s) by an upper approximation $w(s)^{up}$ consisting of a finite number of linear segments with end points s_1, s_2, ... s_n.

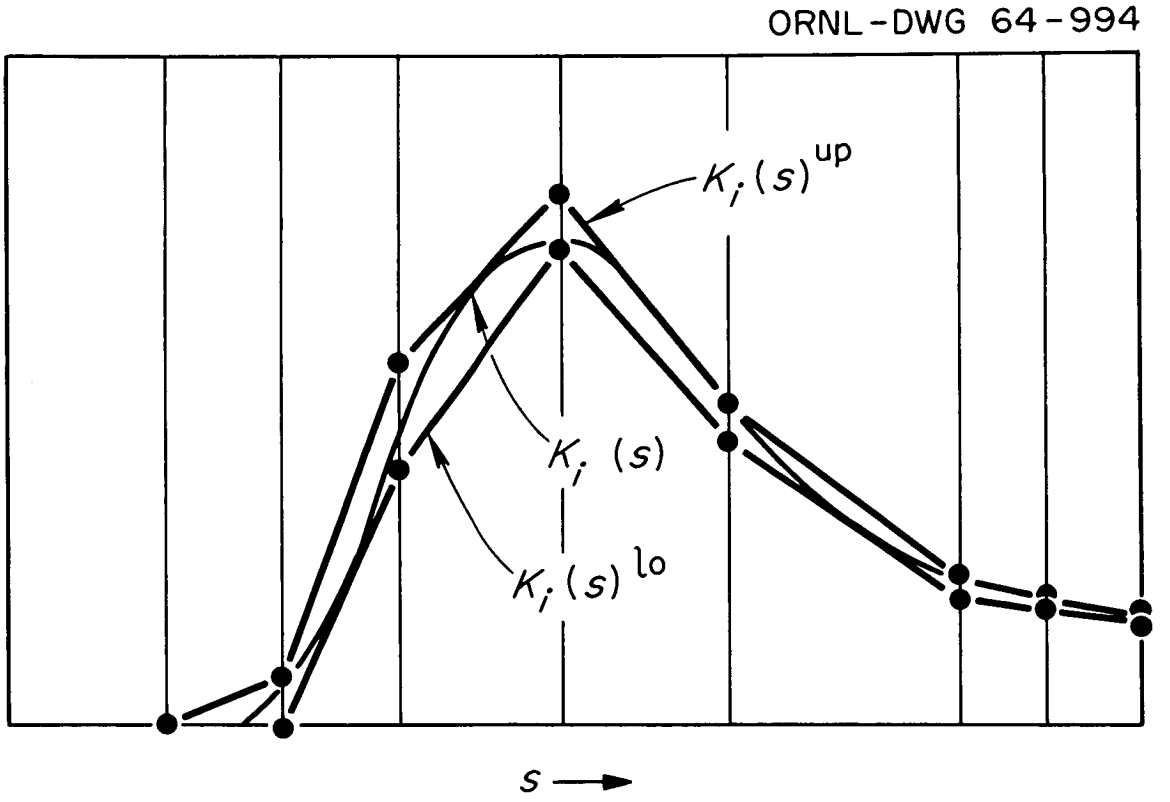


Fig. 13. Replacement of one of the continuous response functions $K_i(s)$ by an upper and lower approximation $K_i(s)^{lo}$ and $K_i(s)^{up}$ consisting of a finite number of straight-line segments with end points $s_1, s_2, \ldots s_n$.

$$\sum_{i=1}^{m} u_i \begin{bmatrix} K_i(s)^{lo} \\ K_i(s)^{up} \end{bmatrix} \leqq \sum_{i=1}^{m} u_i K_i(s) \tag{4.4}$$

where the upper term in brackets is to be summed if u_i is +, and the lower term if u_i is -. Then with the additional requirement that

$$\sum_{i=1}^{m} u_i \begin{bmatrix} K_i(s)^{lo} \\ K_i(s)^{up} \end{bmatrix} \geqq w(s)^{up} \tag{4.5}$$

the following series of inequalities (illustrated in Fig. 14) is implied:

$$\sum_{i=1}^{m} u_i K_i(s) \geqq \sum_{i=1}^{m} u_i \begin{bmatrix} K_i(s)^{lo} \\ K_i(s)^{up} \end{bmatrix} \geqq w(s)^{up} \geqq w(s) \ . \tag{4.6}$$

Equation (4.5) can be written in more convenient form if the coefficient vector u is decomposed into two parts: u^+ with + sign and u^- with - sign, so that $u = u^+ - u^-$ with $u^+ \geqq 0$ and $u^- \geqq 0$. Then Eq. (4.5) becomes

$$\sum_{i=1}^{m} \left[u_i^+ K_i(s)^{lo} - u_i^- K_i(s)^{up} \right] \geqq w(s)^{up} \ . \tag{4.7}$$

Now, letting

$$K_{ij}^{lo} = K_i(s_j)^{lo} \ ,$$

$$K_{ij}^{up} = K_i(s_j)^{up} \ ,$$

$$w_j^{up} = w(s_j)^{up} \ ,$$

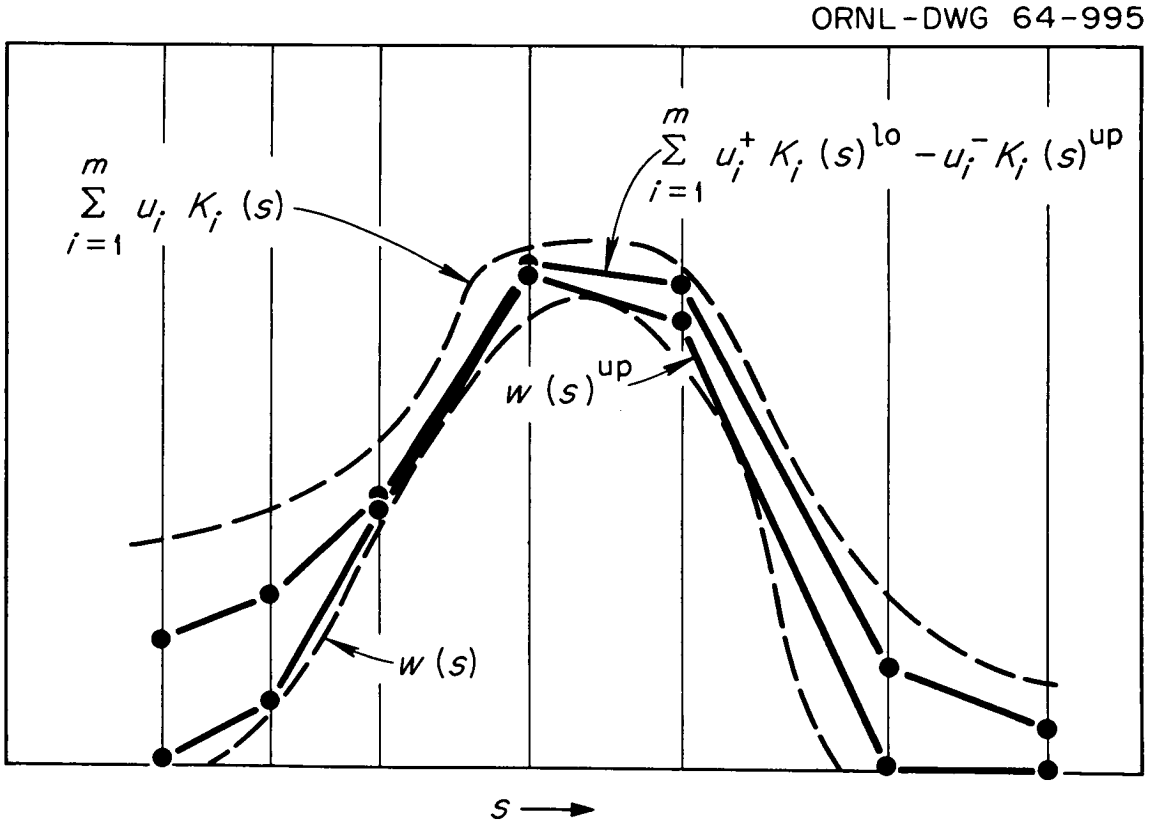


Fig. 14. The inequalities implied by Eqs. (4.2) and (4.3).

the finite-dimensional constraint can be written in matrix notation as

$$u^{+T} K^{lo} - u^{-T} K^{up} \geq w^{up^T} . \tag{4.8}$$

In computation, the required functions $w(s)^{up}$, $K_i(s)^{lo}$, and $K_i(s)^{up}$ need to be constructed for $i = 1, 2, \ldots m$. Then it is required only that the inequality Eq. (4.7) hold at the ends of the linear segments where $s = s_1, s_2, \ldots s_n$. This inequality is weaker than Eq. (4.1) and its use means that the least upper-confidence-interval edge may be somewhat greater than would result from use of the continuous inequality. But since a wider confidence interval will result, this error is conservative and the confidence intervals can still be guaranteed. The number of linear segments that must be used depends upon the shape of the functions, the uncertainty of $b = (b_1, b_2, \ldots b_m)$, and the amount of inefficiency one is willing to tolerate in order to reduce the number of comparison points. For example, if the functions were exactly piecewise linear, then the continuous constraints can be exactly replaced by the finite constraints without any loss in efficiency. Also, there is not much point in making the overall confidence interval very efficient if most of the width is due to the standard deviation term in Eqs. (3.6) and the linear approximation causes only a small increase in width relative to the continuous constraint. A practical criterion for the number of segments is to choose a number that will make the increase in error due to the finite approximation about half as large as the standard deviation part of the width. This number is easy to determine for a particular class of problems by running a few test cases with different choices.

A more efficient scheme is to use a piecewise parabolic approximation instead of a piecewise linear approximation. The comparison points can be taken much further apart if the functions $w(s)$ and $K_i(s)$ are to be approximated within a specified amount. But the inequality condition between the parabolic approximations must be formulated in such a way that in addition to there being inequality at the end points of each segment there is also inequality between them.

A sufficient condition to maintain inequality between parabolas $p_1(s)$ and $p_2(s)$ over the interval (s', s'') is to require the inequalities

$$p_1(s') \geqq p_2(s') , \tag{4.9.1}$$

$$p_1(s'') \geqq p_2(s'') , \tag{4.9.2}$$

and

$$\frac{d^2}{ds^2} p_1(s) \leqq \frac{d^2}{ds^2} p_2(s) . \tag{4.9.3}$$

Then the difference, $p_1(s) - p_2(s)$, cannot have a zero crossing between s' and s''.

The parabolic approximation inequality can be put in the same form as Eq. (4.8) by the use of the central difference formula for the second derivative:

$$\frac{d^2}{ds^2} p(s) = \frac{2}{s'' - s'} \left[p(s') - 2p\left(\frac{s' + s''}{2}\right) + p(s'') \right] . \tag{4.10}$$

The points of comparison $s_1, s_3, \ldots s_n$ are selected at the end points of the parabolic segments, and $s_2, s_4, \ldots s_{n-1}$ at the midpoints. Then the required inequality has the form

$$u^{+T} K^{lo} D - u^{-T} K^{up} D \geqq w^{up^T} D , \tag{4.11}$$

where D is a matrix whose odd-numbered columns correspond to the end points of the approximation segments and whose even-numbered columns are proportional to the finite operator for the second derivative within each segment:

$$D = \begin{bmatrix} 1 & 1 & & & & & & & \\ & -2 & & & & & & & \\ & 1 & 1 & 1 & & & & & \\ & & & -2 & & & & & \\ & & & 1 & 1 & 1 & & & \\ & & & & & -2 & & & \\ & & & & & 1 & \cdot & & \\ & & & & & & & \cdot & \\ & & & & & & & & \cdot \end{bmatrix} .$$

Thus in terms of the modified matrices $(K^{lo}D)$ and $(K^{up}D)$ and the modified vector $(w^{up^T} D)$, the inequality Eq. (4.11) has the same form as Eq. (4.8) for the piecewise linear approximation. However, it is more difficult to obtain the modified forms, since first the required piecewise parabolic functions $w(s)^{up}$, $K_i(s)^{lo}$, and $K_i(s)^{up}$ must be constructed, then be evaluated at s_1, s_2, ... s_n, and finally be transformed by postmultiplying by D.

In conclusion, there is no requirement for the end points of the approximation segments to be evenly spaced or for two contiguous segments to share the same end point. In fact, it is often convenient to take some of the spacings as zero to handle abrupt discontinuities in the response functions (as might result from x-ray absorption edges). Figure 15 illustrates an upper linear and an upper parabolic approximation for two different functions.

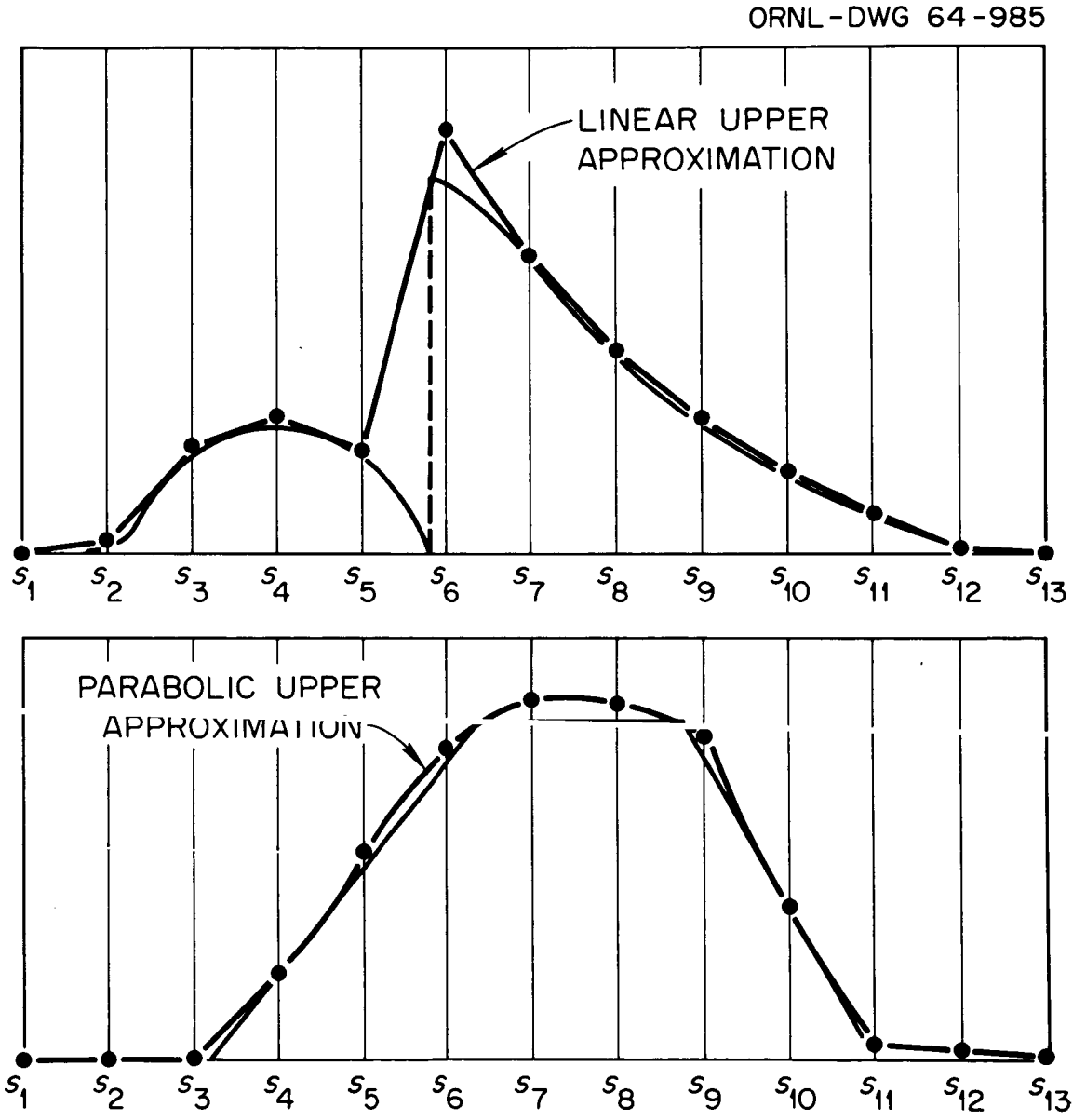


Fig. 15. A piecewise linear upper approximation and a piecewise parabolic upper approximation to two typical functions.

5. PROGRAMMING FORMULATION

In Chapter 2 the unfolding problem was posed in the following form:

given:

$$\left.\begin{array}{l} \int K_i(s)\, x(s)\, ds = \hat{b}_i + e_i \qquad\qquad i = 1, 2, \ldots m \\ \mathrm{var}(\hat{b}) = S = \begin{bmatrix} \sigma_1^2 & & & & & \\ & \sigma_2^2 & & & & \\ & & \cdot & & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & & & \cdot \end{bmatrix} \\ \text{wanted:} \\ \text{confidence interval for } \varphi = \int w(s)\, x(s)\, ds \end{array}\right\}, \tag{5.1}$$

where $K_i(s)$ is the response function relating the response b_i to the unknown spectrum $x(s)$, and e_i is the random error in the experimentally observed response vector $\hat{b}$. It is assumed that the variance-covariance matrix S of $\hat{b}$ is known. The experimenter may specify the window function $w(s)$. Generally, a "solution" consists of a large number of confidence intervals for different window functions.

In Chapter 3 it was shown that if $x(s) \geq 0$ the confidence interval $[\hat{\varphi}^{lo}, \hat{\varphi}^{up}]$ could be determined provided that $w(s)$ could be bracketed between two linear combinations of the response functions. Then the determination of the narrowest possible confidence interval was posed as a pair of extremal problems, with the upper confidence limit given by

$$\hat{\varphi}^{up} = \min_u \left\{ u^T\hat{b} + \kappa \sqrt{u^T S u} \;\middle|\; \sum_{i=1}^{m} u_i K_i(s) \geq w(s) \right\} . \tag{5.2}$$

Finally, in Chapter 4 it was shown that the continuous inequality constraint can be replaced by a finite constraint:

$$u^{+T} K^{lo} - u^{-T} K^{up} \geqq w^{up^T} , \tag{5.3}$$

where K^{lo} and K^{up} are n-by-m matrices and w^{up} is an n-dimensional vector. These are found by evaluating upper and lower piecewise approximations to $K_i(s)$ and $w(s)$ at points of comparison, s_1, s_2, ... s_n. Thus the finite extremal problem for determination of the upper-confidence limit is

$$\hat{\phi}^{up} = \min_u \left\{ u^T\hat{b} + \kappa \sqrt{u^T S u} \;\middle|\; u^{+T} K^{lo} - u^{-T} K^{up} \geqq w^{up^T} \right\} . \tag{5.4}$$

As was mentioned before, the determination of the lower-confidence limit $\hat{\phi}^{lo}$ has the same form as the determination of $\hat{\phi}^{up}$ if it is observed that $-\hat{\phi}^{lo}$ is the same as the upper limit of the function $\phi = \int -w(s)\, x(s)\, ds$. Thus to obtain $\hat{\phi}^{lo}$, w^{up} is replaced with $-w^{lo}$, and $\hat{\phi}^{up}$ with $-\hat{\phi}^{lo}$.

The solution for $\hat{\phi}^{up}$ of Eq. (5.4) is a problem in mathematical programming, the minimization of a given function under certain inequality constraints. Equation (5.4) as it stands is nonlinear in the function to be minimized, with a linear inequality constraint. No analytic solution to Eq. (5.4) can be given, but several computational methods are known for solving this problem to any required degree of accuracy (44). However, a simple observation allows the problem to be expressed as a "quadratic programming problem" with a quadratic form to be minimized, and many methods are known for its exact solution (apart from arithmetic errors) in a finite number of computational steps (45).

If $\hat{\phi}^{up}$ from Eq. (5.4) is considered to be a function of κ, then as κ is varied, the point (γ,ξ) will move along the boundary of the set of values of $\gamma = u^T\hat{b}$ and $\xi = \sqrt{u^TSu}$, which are feasible with respect to the constraints, as illustrated in Fig. 16. If the $\kappa\sqrt{u^TSu}$ term were to be replaced with $\mu(u^TSu)$, the problem would be one of standard quadratic programming. If the path of the point (γ,ξ^2) in the quadratic problem is considered, it is seen that there is a 1-to-1 correspondence between extreme points on the set of feasible (γ,ξ) for the two problems. Thus if the quadratic programming problem is solved for all values of μ, the solution to Eq. (5.4) is given simply by

$$\hat{\phi}^{up}(\kappa) = \min_{\mu}\left[\gamma(\mu) + \kappa\xi(\mu)\right]. \tag{5.5}$$

Although other methods for obtaining the desired confidence interval are discussed in later sections, the entire process of obtaining confidence intervals using constrained estimation can now be outlined. A flow diagram of the numerical calculation is shown in Fig. 17.

The computations shown in Fig. 17 could be carried out by an existing quadratic programming code (45), but a rather inefficient code would result, because approximately $5\cdot q\cdot n\cdot(n+m)^2$ floating-point operations (mainly multiplication and subtraction) are required in order to find both upper- and lower-confidence-interval edges for q windows. Thus in a problem of moderate magnitude, such as the one shown in Figs. 8 and 9, where $q = 80$, $n = 40$, and $m = 70$, about 10^9 operations are required, necessitating several hours of time on a modern computer. The speed can be improved considerably by using the result from the previous edge as a starting point for the next one, instead of solving a

ORNL-DWG 64-996

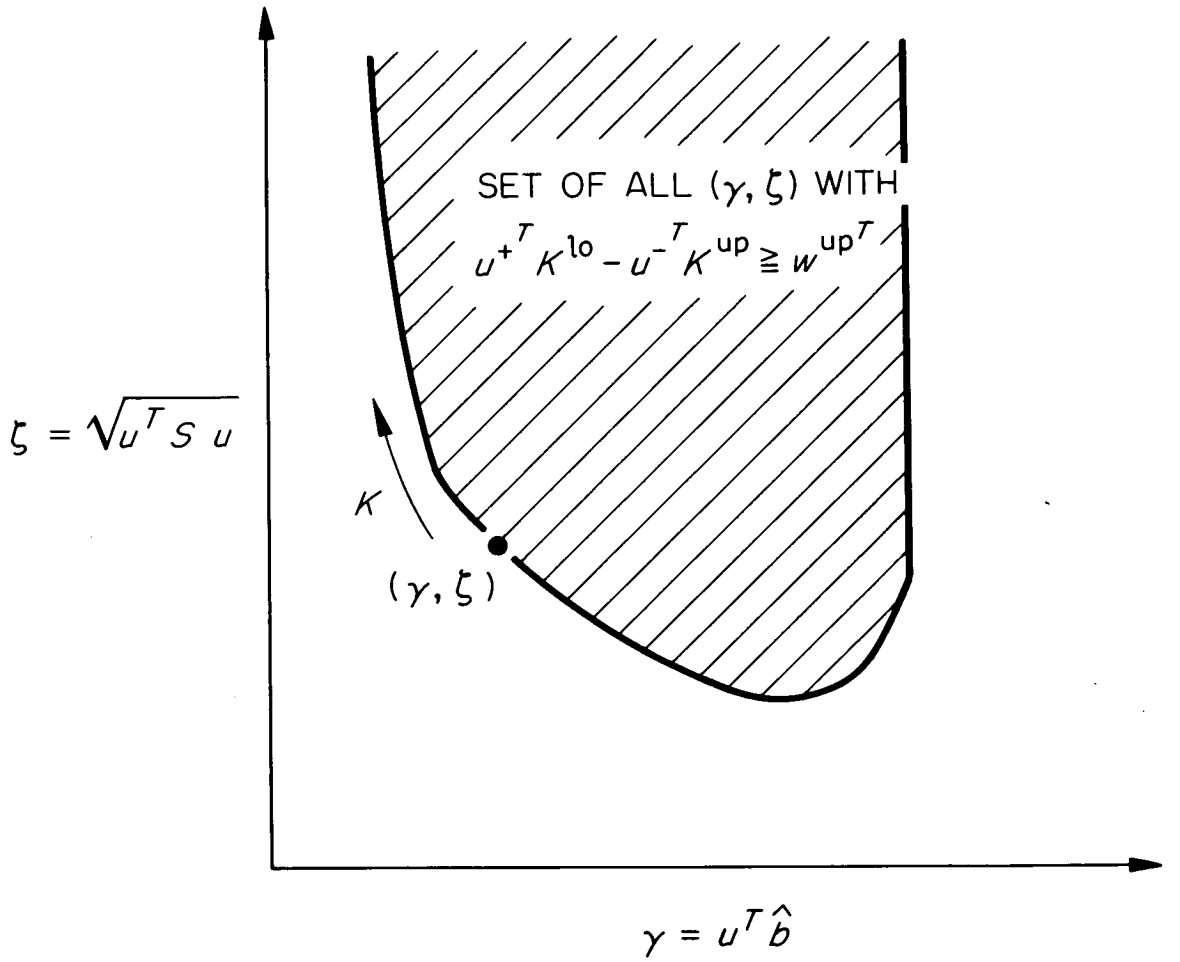


Fig. 16. The solution to both Eq. (5.4) and Eq. (5.5) trace out the path of (γ, ξ) on the boundary of the feasible set as κ or μ is varied.

ORNL-DWG 64-1003

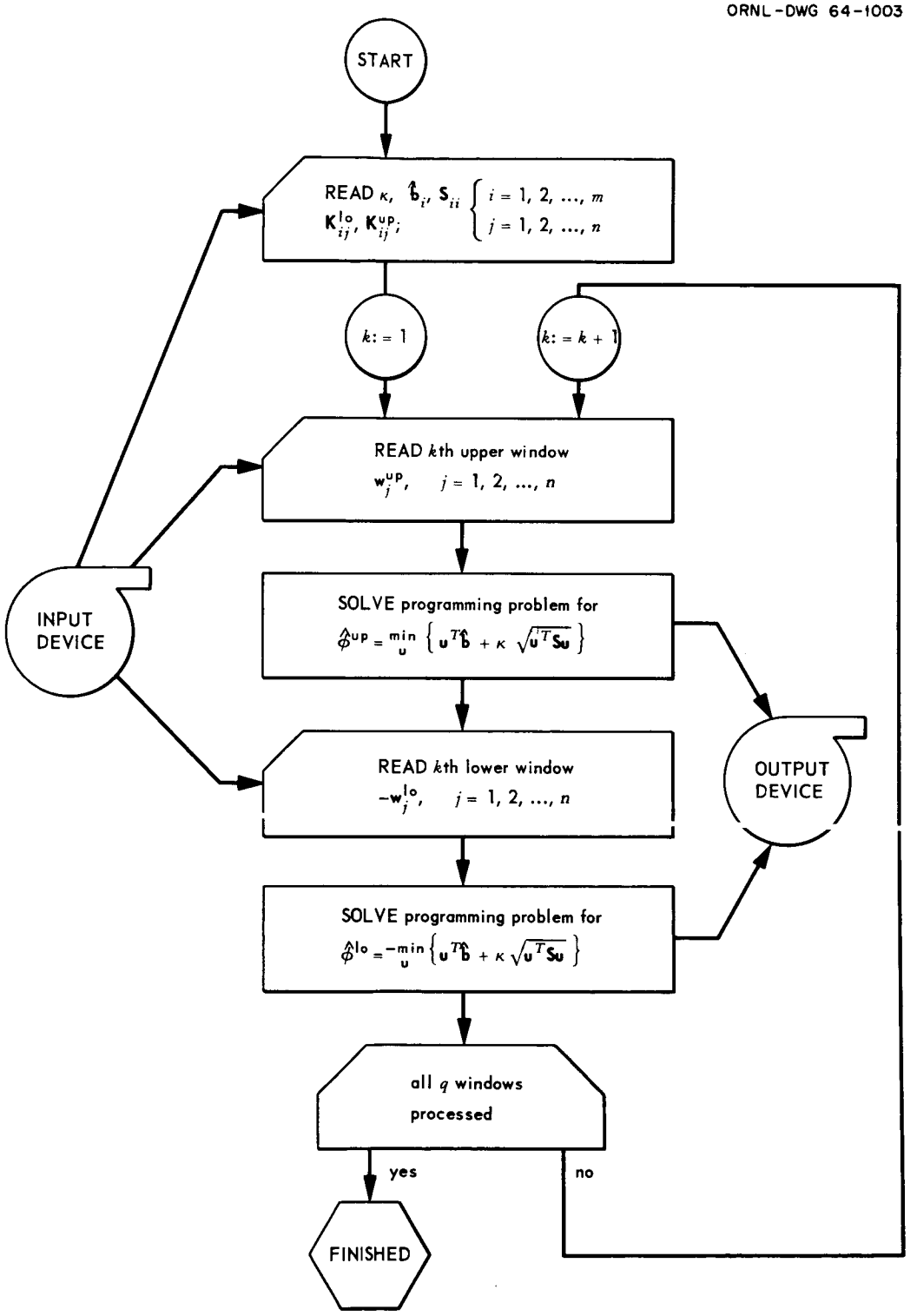


Fig. 17. Steps in the numerical calculation of the constrained confidence intervals $(\hat{\phi}^{lo}, \hat{\phi}^{up})$.

separate programming problem for each confidence interval edge. However, rather than pursue the exact quadratic programming formulation, we have developed two different approximations, which require only about $20 \cdot q \cdot n \cdot m$ floating-point operations and thus enable the problem mentioned above to be solved in a few minutes.

6. LINEAR PROGRAMMING APPROXIMATION

In order to reduce the amount of computation required in solving the unfolding problem, the nonlinear programming problem can be replaced by a linear one. This results in a simpler and faster computer code but imposes the penalty of the confidence intervals not being as narrow as possible. The efficiency of the linear programming code is discussed later.

The standard deviation term[1]

$$\sqrt{u^T S u} = \sqrt{\sum_{i=1}^{m} u_i^2 \sigma_i^2}$$

in the constrained estimation nonlinear programming formulation can be interpreted as the Euclidean norm (length) of the coefficient vector u in the metric of S. In order to obtain a linear programming problem in u^+ and u^-, the Euclidean norm can be replaced with the sum of absolute value norm. These two are related by the inequality

$$\sum_{i=1}^{m} |u_i| \, \sigma_i \geq \sqrt{\sum_{i=1}^{m} u_i^2 \sigma_i^2} \quad , \tag{6.1}$$

where the equality sign holds if and only if just one component of the summation is nonzero.

---

[1]Here the variance matrix S is assumed to be diagonal, with $S_{ii} = \sigma_i^2$ and $S_{ij} = 0$ $(i \neq j)$. By a suitable linear orthogonal transformation, an arbitrary S may be put in this form.

Now Eq. (5.4) may be put into the linear form:

$$\hat{\phi}^{up} = \min_{u} \left\{ u^T \hat{b} + \kappa \, |u|^T \sigma \;\middle|\; u^{+T} K^{lo} - u^{-T} K^{up} \geq w^{up^T} \right\} . \qquad (6.2)$$

In order to get Eq. (6.2) into a more convenient form for computation, the objective function is written explicitly in terms of u^+ and u^- and the inequality in the constraint is changed into an equality by subtracting a nonnegative "slack" vector from the left-hand side of the constraint:

$$\hat{\phi}^{up} = \min \left\{ (u^+ - u^-)^T \hat{b} + \kappa (u^+ + u^-)^T \sigma \;\middle|\; u^{+T} K^{lo} - u^{-T} K^{up} - v^T = w^{up^T} \right\} , \qquad (6.3)$$

with $u^+ \geq 0$, $u^- \geq 0$, and $v \geq 0$. The elements of the slack vector v may be interpreted as the difference between the piecewise-approximation sensitivity function and window function at the comparison points.

If the three vectors u^+, u^-, and v are collected into one combined vector $(u^+ \; u^- \; v)$, Eq. (6.3) can be rewritten as

$$\hat{\phi}^{up} = \min \left\{ (u^+ \; u^- \; v)^T \begin{pmatrix} \hat{b} + \kappa\sigma \\ -\hat{b} + \kappa\sigma \\ 0 \end{pmatrix} \;\middle|\; (u^+ \; u^- \; v)^T \begin{pmatrix} K^{lo} \\ -K^{up} \\ -I \end{pmatrix} = w^{up^T} \right\} , \qquad (6.4)$$

with $\begin{pmatrix} u^+ \\ u^- \\ v \end{pmatrix} \geq 0$.

Equation (6.4) is now in the correct form for solution by linear programming. Since there are many adequate references to linear programming methods, the reader is referred to these and just the details of

the Simplex algorithm (44,46–48), the best known of the linear programming methods, are sketched. Computationally, the Simplex method examines a vertex of the set of those feasible (u^+ u^- v) which will satisfy the linear inequality constraints. If this point does not yield the minimum value desired, the algorithm proceeds to an adjacent vertex at which the function is smaller (or the same if no adjacent vertex has a smaller value). Eventually the Simplex method will terminate with a solution in a finite number of steps or will show that the solution is unbounded. The process by which the numerical computation is done is very similar to the ordinary Gauss elimination commonly used in matrix inversion. The difference consists in the order in which the columns to be eliminated are selected.

Cycling is a phenomenon that occurs in linear programming problems when at some stage all adjacent vertices have the same value. It is then possible to move from one vertex to another without decreasing the desired function, and eventually it is possible to come back to the starting vertex. It was necessary to use one of several methods (44, 46–48) known for obtaining a solution in these cases. Linear programming codes have been written for most computers. Gass has reviewed the codes which were available up to about 1960 (49). A flow diagram for the linear programming solution to the unfolding problem is given in Fig. 18, which shows that a separate linear programming problem is solved for each upper and lower edge of the confidence intervals for each window. Typically, from n to 3n Gauss eliminations are required to solve just one minimization problem. Each Gauss elimination involves approximately $n(2n + m)$ floating-point operations; thus a more economical procedure

ORNL - DWG 64 - 1002

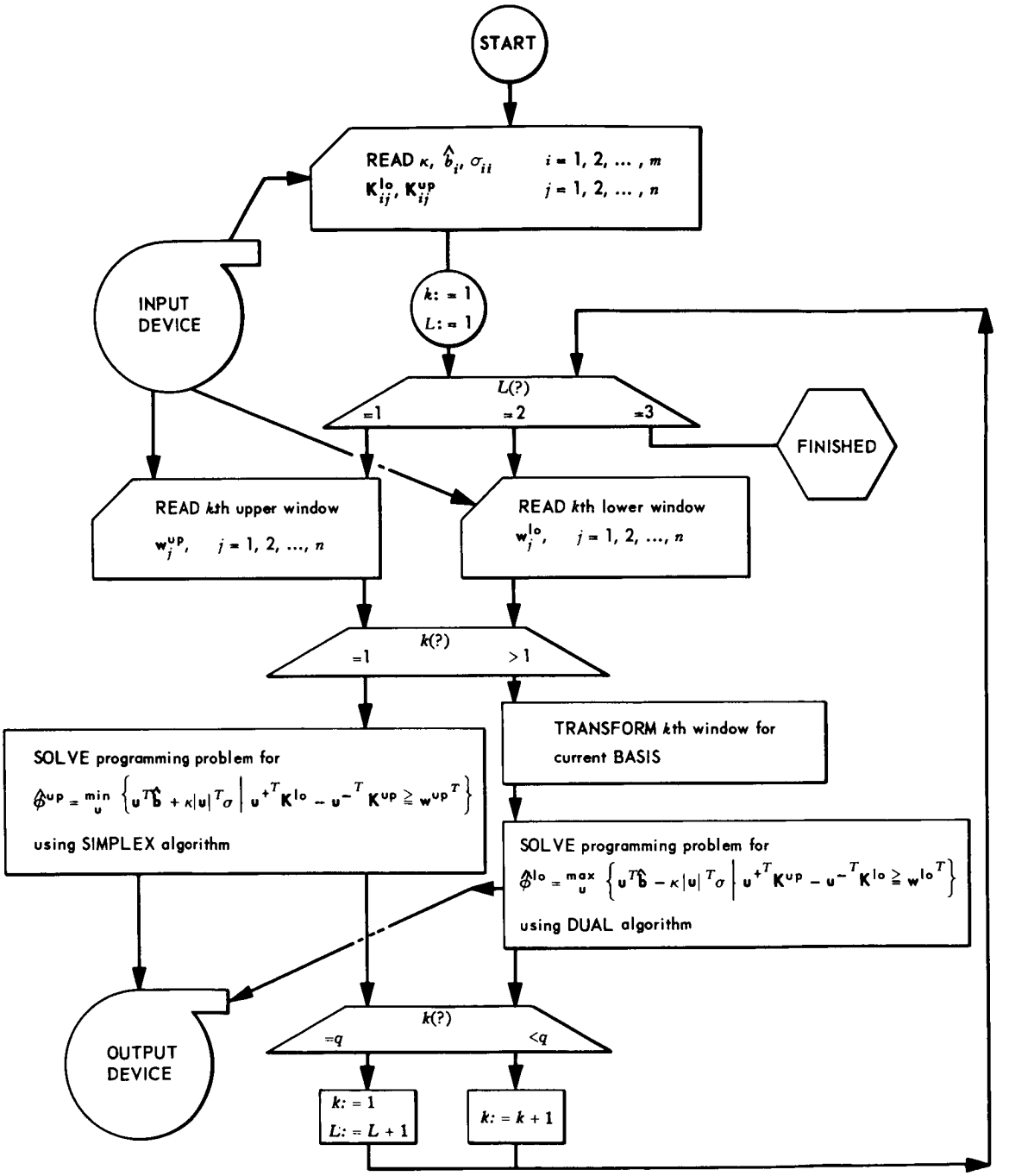


Fig. 18. Flow diagram for solution of unfolding problem by linear programming.

is desirable. A considerable reduction in the number of eliminations is possible if each new problem is started at the terminal vertex of the previous problem. Some existing linear programming codes have this facility.

Efficiency of computational method

Let the efficiency be defined

$$\eta = \frac{\text{minimum possible width of confidence interval}}{\text{width of computed confidence interval}} .$$

If a sufficiently large number of comparison points is taken, the nonlinear programming formulation of Eq. (5.4) would yield an efficiency of 1. But because of the replacement of the Euclidean norm by the sum of absolute value norm, the efficiency of the linear programming method may be less. It is difficult to give an estimate of the efficiency because it varies considerably, depending upon the problem. In the limit of small error, the efficiency is 1.0, where any remaining width is due to the underdetermination of the problem and not to the statistical errors. The worst possible case gives an efficiency of $1/\sqrt{m}$, and leaves much to be desired.

Two techniques are used to increase the efficiency. With one, although the sum of absolute value norm is used in the minimization, the final confidence interval edge can be computed by using the Euclidean norm, with the coefficient vector u which resulted from the approximate formulation. The other technique involves taking a Taylor expansion about the approximately optimal u, and continuing the minimization using the linearized "cost function."

7. DUALITY RELATIONS

Dualism occurs in many different forms. A well-known primal-dual relationship occurs in the formulation of electric network equations in terms of voltages around loops, or in terms of currents into node points. Either formulation leads to the same physical solution, and the results of either can be translated into the other. Yet the two physically equivalent approaches often differ in their computational difficulty. Two other familiar dualisms are the particle-wave dualism of quantum mechanics and the time-domain, frequency-domain dualism in the behavior of systems governed by linear differential equations. Aside from the practical (computational) aspects, consideration of these primal-dual relations often yields insight into the problems.

In mathematical programming, a dualism exists which allows reformulation of the standard minimization problem (called the "primal") with n-unknowns $z = (z_1, z_2, \ldots z_n)$ and m constraint relations in terms of a maximization problem (called the "dual") with m-unknowns $u = (u_1, u_2, \ldots u_m)$ and n constraints. The variable u is said to be the dual of z. In many cases the primal-dual relationship is symmetric, and either problem can be considered the dual of the other; in other words, z is also the dual of u.

The primal-dual relations of mathematical programming take an especially simple form for linear programming (where the function to be minimized is a linear function of the unknowns and the constraint

equations are linear inequalities). Then the primal and dual are the following:

$$\text{primal:} \quad \phi_p = \min_{z \geq 0} \left\{ c^T z \;\middle|\; Az \geq b \right\} \quad ; \tag{7.1.1}$$

$$\text{dual:} \quad \phi_d = \max_{u \geq 0} \left\{ u^T b \;\middle|\; u^T A \leq c^T \right\} \quad . \tag{7.1.2}$$

The primal-dual theorem of linear programming (closely related to the famous minimax theorem of Von Neumann in game theory) states that (6–9)

$$\phi_p = \phi_d \quad . \tag{7.2}$$

As a corollary,

$$c^T z \geq c^T z^* = \phi_p = u^{*T} Az^* = \phi_d = u^{*T} b \geq u^T b \quad , \tag{7.3}$$

where the z^* and u^* are optimal values which yield the minimum and maximum of Eqs. (7.1.1) and (7.1.2), and u and z are any feasible solutions which satisfy the constraints of Eqs. (7.1.1) and (7.1.2). An additional type of relationship between the primal and dual variables is expressed by

$$u_i = \frac{\partial \phi}{\partial b_i} \tag{7.4.1}$$

and

$$z_j = \frac{\partial \phi}{\partial c_j} \quad . \tag{7.4.2}$$

In other words, u_i gives the change in the optimum solution of Eq. (7.1.1)

or (7.1.2) as b_i is varied an infinitesimal amount. If some inequality of Eq. (7.1.1) is "slack" (i.e., the inequality holds), then the corrresponding component of u is zero, since then b_i could be varied a small amount without affecting the value of the optimal ϕ.

The two basic primal-dual relationships (7.1.1) and (7.1.2) can be algebraically manipulated to yield a large number of equivalent primal-dual relations. For example, the primal may be changed from minimization to maximization by minimizing $-c^T z$ instead of $c^T z$. In addition, the nonnegativity constraint can be removed, leaving only the linear inequality constraint, by the device discussed in Chapter 4 of expressing the unrestricted variable (with $-\infty \leq z \leq +\infty$) in terms of two nonnegative variables, z^+ and z^-, so that $z = z^+ - z^-$. The primal is then expressed in terms of the vector (z^+, z^-) of double the dimension of z. Many of the equivalent forms of the primal-dual relation are summarized in a recent publication (50).

Of particular interest for the unfolding problem are the following special forms:

$$\text{primal:} \quad \phi^{up} = \min_{u} \left\{ u^T b + |u|^T \sigma \;\middle|\; u^T A \geq c^T \right\} \; ; \tag{7.5.1}$$

$$\text{dual:} \quad \phi^{up} = \max_{z \geq 0} \left\{ c^T z \;\middle|\; |Az - b| \leq \sigma \right\} \; . \tag{7.5.2}$$

The primal form (7.5.1) of these two relations is identical to the linear programming formulation of the unfolding problem in the special case that $A = K^{up} = K^{lo}$ and $c = w^{up} = w^{lo}$. Without this simplifying assumption,

the primal-dual relation is the following:

$$\text{primal:} \quad \phi^{up} = \min_{u} \left\{ u^T b + |u|^T \sigma \;\middle|\; u^{+T} K^{lo} - u^{-T} K^{up} \geqq w^{up^T} \right\} ; \qquad (7.6.1)$$

$$\text{dual:} \quad \phi^{up} = \max_{z \geqq 0} \left\{ c^T z \;\middle|\; K^{lo} z - b \leqq \sigma, \quad -\sigma \leqq K^{up} z - b \right\} . \qquad (7.6.2)$$

The simplified forms (7.5.1) and (7.5.2) have a simple interpretation. Equation (7.5.1) has already been discussed in terms of finding a minimum-variance upper estimate, with the nonnegativity of z being a necessary condition, although the variable z does not occur anywhere in Eq. (7.5.1). Equation (7.5.2), however, explicitly involves the nonnegativity of z. The upper-confidence-interval edge is given by Eq. (7.5.2) as the largest possible value that $\phi = c^T z$ can attain for a nonnegative z which is consistent with the experimental data such that the magnitude of each residual $(Az - b)_i$ is as small as σ_i. Similarly, the lower-confidence-interval edge leads to a problem such that ϕ takes the smallest possible value consistent with the small residuals and with the nonnegativity of z.

In the general mathematical programming problem (where both the function to be minimized and the constraints may be nonlinear), the primal-dual relation is more complicated. Wolfe (51) has given the following form:

$$\text{primal:} \quad \text{minimize } f(z) \text{ subject to } g_i(z) \geqq 0, \quad i = 1,2,\ldots,m \quad ; \qquad (7.7.1)$$

$$\text{dual:} \begin{cases} \text{maximize} \left[f(z) - \sum_{i=1}^{m} u_i g_i(z) \right] \\ \\ \text{subject to } \nabla f(z) = \sum_{i=1}^{m} u_i \nabla g_i(z), \text{ and } u \geqq 0 \quad , \end{cases} \tag{7.7.2}$$

where $\nabla f(z)$ denotes the gradient of $f(z)$. Wolfe's principal theorem is the following: <u>If z^* solves the primal problem, then there exists u^* so that (z^*, u^*) solves the dual problem, and the extrema are equal</u>. This immediately reduces to Eq. (7.1) for the linear programming case. Of more interest is that Wolfe's theorem can be used to find the primal-dual relation for the nonlinear programming formulation of the unfolding problem. With $A = K^{lo} = K^{up}$, the primal-dual relation is:

$$\text{primal:} \quad \hat{\phi}^{up} = \min_{u} \left\{ u^T \hat{b} + \kappa' \sqrt{u^T S u} \;\middle|\; u^T A \geqq w^T \right\} \quad ; \tag{7.8.1}$$

$$\text{dual:} \quad \hat{\phi}^{up} = \max_{z \geqq 0} \left\{ w^T z \;\middle|\; (Az - \hat{b}) S^{-1} (Az - \hat{b}) \leqq \kappa'^2 \right\} \quad . \tag{7.8.2}$$

Thus it is seen that the upper-confidence-interval edge is given by the maximum possible value that $\phi = w^T z$ can attain for a nonnegative z and a specified value of the error sum of squares. The connection between ordinary least-squares estimation and constrained estimation can now be shown; this will determine the value of κ which yields a specified confidence level.

In Chapter 2 the confidence interval $[\hat{\phi}^{lo}, \hat{\phi}^{up}]$ for $\phi = w^T z$ is given by

$$\hat{\phi}^{lo} = w^T\hat{z} - \kappa \sqrt{w^T V(z)\ w} \quad , \tag{7.9.1}$$

$$\hat{\phi}^{up} = w^T\hat{z} + \kappa \sqrt{w^T V(z)\ w} \quad , \tag{7.9.2}$$

where κ is related to the confidence level according to Eq. (2.10). The same confidence interval edges can be obtained as the solution to the extremal problem:

$$\hat{\phi}^{up} = \max_z \left\{ w^T z \;\middle|\; (Az - \hat{b})^T S^{-1} (Az - \hat{b}) \leq \kappa^2 + \mu^2 \right\} \quad , \tag{7.10.1}$$

$$\hat{\phi}^{lo} = \min_z \left\{ w^T z \;\middle|\; (Az - \hat{b})^T S^{-1} (Az - \hat{b}) \leq \kappa^2 + \mu^2 \right\} \quad , \tag{7.10.2}$$

where μ^2 is the $\min_z [(Az - \hat{b})^T S^{-1} (Az - \hat{b})]$. This extremal problem can be interpreted geometrically in terms of the confidence ellipsoid $(Az - \hat{b})^T S^{-1} (Az - \hat{b})$ shown in Fig. 19. The extremal values of ϕ correspond to the points of contact with the two planes of support orthogonal to the direction of the window vector w. The values of z at these two points of contact are the extremal vectors z^{up} and z^{lo}, given by

$$\hat{z}^{up} = \hat{z} + \frac{\kappa}{\sqrt{w^T V(\hat{z})\ w}} V(z)\ w \tag{7.11.1}$$

and

$$\hat{z}^{lo} = \hat{z} - \frac{\kappa}{\sqrt{w^T V(\hat{z})\ w}} V(z)\ w \tag{7.11.2}$$

ORNL-DWG 64-997

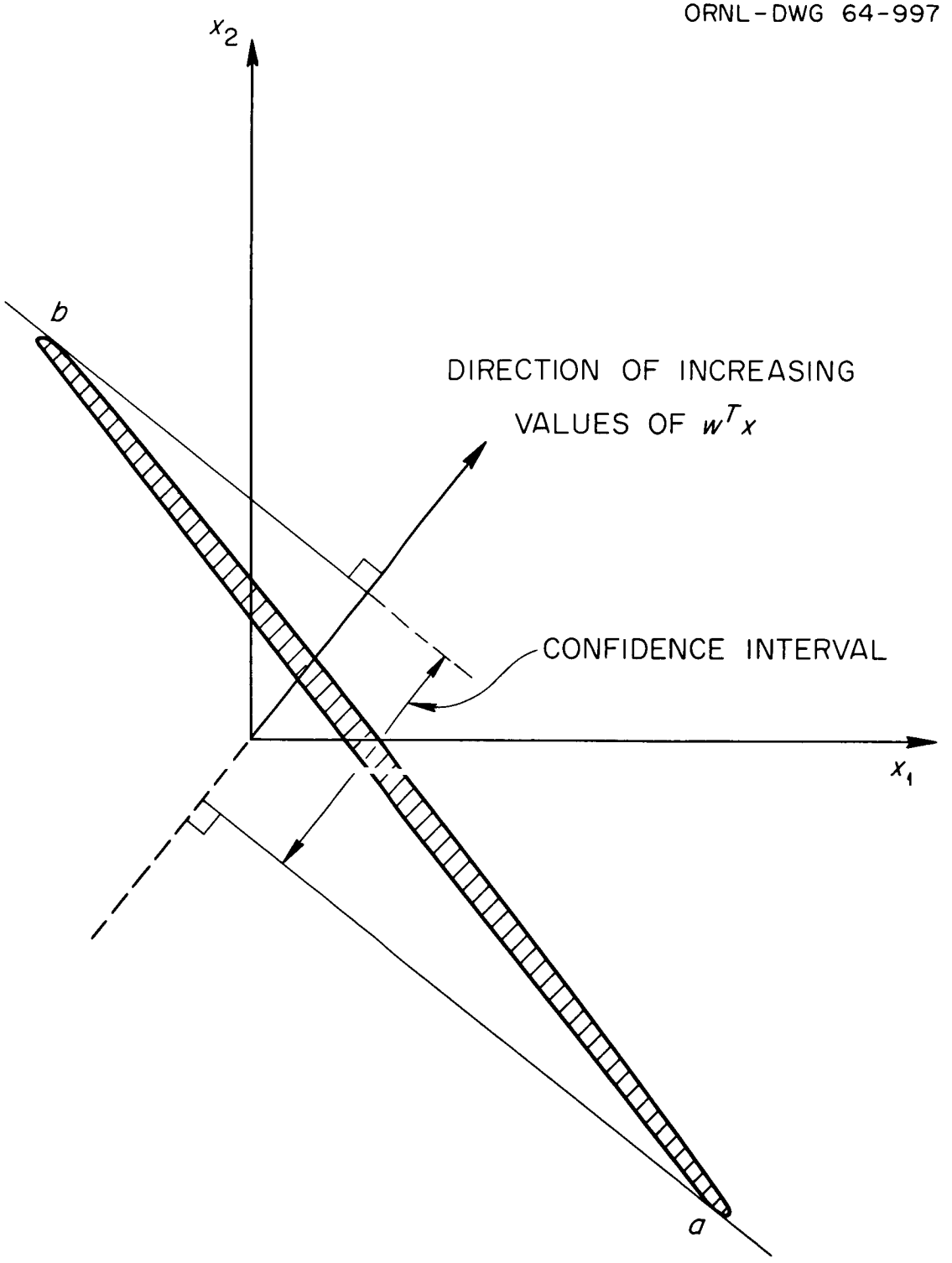


Fig. 19. The confidence ellipsoid for ordinary linear estimation (shown in two dimensions). The edges of the confidence interval for a function $\Phi = w^T x$ are given by the value of Φ at points a and b.

and are the two values of z which give the largest and smallest values of $w^T z$ consistent with $(Az - \hat{b})^T S^{-1} (Az - \hat{b}) \leq \kappa^2 + \mu^2$. That this extremal problem does indeed give the correct confidence interval can be verified by direct solution of the extremal problem by means of Lagrange multipliers. A dual extremal problem can be formulated which yields the same confidence interval as Eqs. (7.10):

$$\hat{\phi}^{up} = \min_u \left\{ u^T\hat{b} + \sqrt{\kappa^2 + \mu^2}\ \sqrt{u^T S u} \;\middle|\; u^T A = w^T \right\} ; \qquad (7.12.1)$$

$$\hat{\phi}^{lo} = \max_u \left\{ u^T\hat{b} - \sqrt{\kappa^2 + \mu^2}\ \sqrt{u^T S u} \;\middle|\; u^T A = w^T \right\} . \qquad (7.12.2)$$

In this classical dual, the constraint is the unbiased condition, and the "objective function" is the upper estimate $\pm\sqrt{\kappa^2 + \mu^2}$ times the standard deviation of the estimate.

With two exceptions these two classical extremal formulations of ordinary estimation are identical to the inequality primal-dual formulation for constrained estimation given in Eqs. (7.8.1) and (7.8.2): the primal problem is not restricted to nonnegative z, and the dual is an equality rather than an inequality. Thus for this problem (as well as for linear programming) relaxing the nonnegativity constraint in the primal corresponds to replacing the inequality in the unrestricted dual by an equality.

The desired connection with ordinary estimation can now be established so that the value of κ in Eqs. (7.8) neccessary to give a specified confidence level can be specified.

If in the discrete least-squares problem it is known a priori that $z \geqq 0$, then the confidence level resulting from solution of the extremal problems (7.10.1) and (7.10.2) will not be reduced if only $z \geqq 0$ is considered. Thus solution of

$$\hat{\phi}^{up} = \max_{z \geqq 0} \left\{ w^T z \;\middle|\; (Az - \hat{b})^T S^{-1} (Az - \hat{b}) \leqq \kappa^2 + \mu^2 \right\} \tag{7.13.1}$$

and

$$\hat{\phi}^{lo} = \min_{z \geqq 0} \left\{ w^T z \;\middle|\; (Az - \hat{b})^T S^{-1} (Az - \hat{b}) \leqq \kappa^2 + \mu^2 \right\} \tag{7.13.2}$$

will yield a confidence interval $[\hat{\phi}^{lo}, \hat{\phi}^{up}]$ such that

$$\Pr(\hat{\phi}^{lo} \leqq \phi \leqq \hat{\phi}^{up}) = \alpha \quad . \tag{7.14}$$

The geometrical interpretation of the constrained estimation is shown in Fig. 20. The confidence interval now is given by the points of contact of the two support planes of the intersection of the ordinary confidence ellipsoid with the positive region.

Comparison of Fig. 19 with Fig. 20 will show why the nonnegativity restriction yields an improved confidence interval. Furthermore, the ordinary estimation method yields a trivial (infinite width) confidence interval if the confidence ellipse is degenerate (singular least-squares matrix, or $n > m$) unless w is orthogonal to all degenerate axes. But with constrained estimation, a nontrivial confidence interval is obtained even if the ellipsoid is degenerate, provided that the intersection of the ellipsoid with the positive orthant (generalization of quadrant and octant to an arbitrary number of dimensions) is finite.

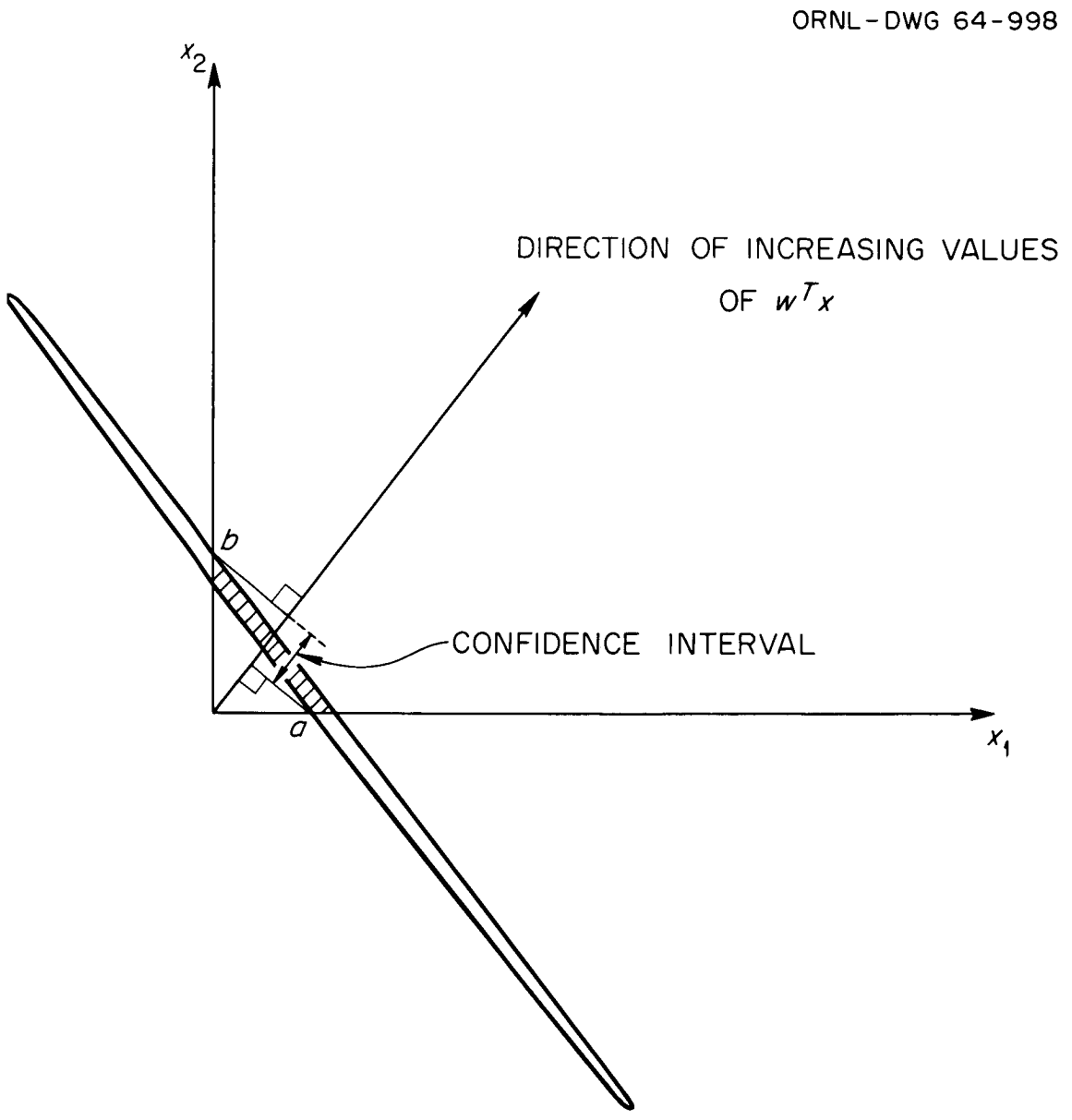


Fig. 20. The confidence region for constrained estimation (shown in two dimensions). The edges of the confidence interval for a function $\phi = w^T x$ are given by the value of ϕ at points a and b. The shaded area is the intersection of the ellipsoid and the positive region.

The proper interpretation of κ' in Eqs. (7.8) necessary to yield a confidence interval is

$$\kappa'^2 = \kappa^2 + \mu^2 = \kappa^2 + \min_z \left[(Az - \hat{b})^T S^{-1} (Az - \hat{b}) \right] . \tag{7.15}$$

Note that the minimization is taken over all z and not just nonnegative z. Note also that if $n \geq m$, then μ is always zero. This has a peculiar consequence: it is possible sometimes that a confidence interval does not exist. Geometrically, this situation arises when the ordinary confidence ellipse does not have any point in common with the positive orthant. This is possible statistically, but of course with a probability which is less than $(1 - \alpha)$. If α is increased sufficiently, the ordinary confidence ellipse will eventually intersect the positive orthant, but it is not "fair" to alter the confidence level in order to obtain a solution. If this is done, the probability interpretation of the confidence interval breaks down. Of course, if the observation of the response vector b were repeated, a valid interval would be obtained for any prespecified α in a fraction of the experiments at least as large as α in the long run.

This possible inconsistency is a serious conceptual problem since some experiments are so costly in time and equipment that it would be impracticable to repeat the experiment if no confidence interval is obtained on the first try. One solution to this dilemma is to replace the constraint $(Az - \hat{b})^T S^{-1} (Az - \hat{b}) \leq \kappa^2$ by the weaker constraint $(Az - \hat{b})^T S^{-1} (Az - \hat{b}) \leq \kappa^2 + \mu'^2$, where

$$\mu'^2 = \min_{z \geq 0} \left[(Az - \hat{b})^T S^{-1} (Az - \hat{b}) \right] . \tag{7.16}$$

Then the resulting ellipsoid will always intersect the positive orthant. The confidence interval, however, will be weaker than that of Eq. (7.14), so that

$$\Pr(\hat{\phi}^{lo} \leq \phi \leq \hat{\phi}^{up}) \geq \alpha \tag{7.17}$$

holds with inequality rather than with equality. In many cases, however, a weak confidence interval which always exists is preferable to a stronger confidence interval which may occasionally fail to exist. If a stronger confidence interval is demanded, then it must be expected that some experiments will be rejected [not exceeding $(1 - \alpha)$ of the total number in the long run].

8. ERRORS IN THE RESPONSE MATRIX

Thus far it has been assumed that the only statistical errors are the errors in observing the response vector $\hat{b} = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_m)$. However, in most experimental problems the response functions $K_i(s)$ are known only with some statistical uncertainty. Sometimes, the functional dependence of $K_i(s)$ on i and s is known from physical considerations, and certain parameters are estimated by comparison of experimental results with known calibration spectra. In other cases it may be possible to experimentally measure $K_i(s)$ for certain values of s for which there are available sources of discrete s.

Since it was shown in Chapter 4 that the continuous problem can be reduced to a finite problem (with a slight weakening of the confidence interval), it suffices to discuss the discrete problem in terms of the response matrices $\hat{K}^{up}$ and $\hat{K}^{lo}$ where the elements of these matrices may be subject to error. In ordinary least-squares estimation, where $A = K^{up} = K^{lo}$, it is fairly easy to evaluate the effect of errors in the matrix elements on the confidence interval for a function $\phi = w^T x$. If the normal equations are not singular, the matrix $(A^T S^{-1} A)^{-1}$ can be expanded in terms of the error dA by means of the approximation

$$(A - B)^{-1} = \left[I + (A^{-1}B) + (A^{-1}B)^2 + \ldots \right] A^{-1}$$

$$\cong A^{-1} + A^{-1}B\,A^{-1} \quad \text{if } B \ll A \quad . \tag{8.1}$$

But if the normal equations are underdetermined, the required inverse matrix does not exist. This means that any estimable function will have

an infinite variance if any column which enters into the combination for the sensitivity function has an error.

Thus our inequality approach for the unfolding problem, which is generally underdetermined, must be extended. If the error in $\hat{K}_{ij}$ is known (with certainty) to be less in magnitude than Σ_{ij}, it can be taken into account by enlarging the constraint set of (7.6.1) to

$$u^{+T}(\hat{K}^{lo} - \Sigma) - u^{-T}(\hat{K}^{up} + \Sigma) \geq w^{up^T} \quad . \tag{8.2}$$

This in effect provides a safety margin in the fit of the window function by a combination of response functions, so that even if each response function is in error by the maximum amount, the upper sensitivity function will be greater than w (and similarly for the lower sensitivity function).

If the errors in $\hat{K}^{up}$ and $\hat{K}^{lo}$ have a random distribution, the constraint set as above could still be enlarged by $\kappa\Sigma$, where Σ is now a matrix whose elements are $\sigma(K_{ij})$, and κ depends upon the confidence level desired. But this would possibly lead to a serious overestimate of the error due to the matrix, since any random cancellation of errors would be ignored.

A more satisfactory method for statistical errors in K is to make the approximation that the errors are much smaller than the matrix elements and to calculate the change in ϕ due to a small change in K. Then the overall confidence error can be obtained by averaging over the statistical distribution of the errors in K and $\hat{b}$. However, if the error calculation is deferred until the programming problems have been solved, a confidence interval which is too large will be obtained, since the

coefficients of combination, u, will not be optimized with respect to the combined matrix and response vector errors. Therefore the procedure wanted is one which incorporates the matrix errors into the programming problem before the numerical solution that will yield the optimal confidence interval.

The key to this problem is the interpretation of the dual variable z as the change in the objective function for small variations in b_i given by Eq. (7.4.1). More generally, the dual variable z_j is the change in objective function $\hat{\phi}$ which results from the jth constraint becoming unsatisfied by a small amount. Since the matrix $\hat{K}$ always enters the dual problem in combination with the response vector b in the form $(\hat{K}z - \hat{b})$, the constraints may become unsatisfied if b_i changes or if $(\hat{K}z)_i$ changes. The net change in the objective function is

$$d\phi = u^T(dKz - db) \quad , \tag{8.3}$$

where u is the vector of coefficients of combination in the unfolding problem and z is the dual of u.

Chapter 3 shows how to obtain the optimal confidence interval if only b is in error. Now Eq. (8.3) implies that the matrix errors dKz can be included if its contribution in the metric is included:

$$S = E(\hat{b} - b)(\hat{b} - b)^T + E(\hat{K} - K)\ zz^T(\hat{K} - K)^T \quad . \tag{8.4}$$

Then if the distribution of errors in $\hat{b}$ and $\hat{K}$ is known, the metric S can be calculated.

Two problems still remain. First the dual solution vector z is not generally known until after the problem is solved, since it emerges

as part of the computation. Second, since z is different for both the upper and lower edge of every confidence interval, the unfolding problem should now be solved anew in a different metric, twice for each confidence interval. Thus the computation would be impracticable except for small or very important problems.

If the final confidence interval is computed from Eq. (8.3), however, a safe confidence interval will still be achieved. At first, z can be assigned arbitrarily or set equal to zero. The entire calculation could be repeated by taking a new z for the metric as the average of all the extremal dual vectors obtained on the first calculation. Usually, however, some physical considerations will suggest a first approximation for z (for example, see Chapter 9). These procedures appear adequate except for serious matrix errors, where the small error assumption is not valid.

9. MODIFIED LEAST-SQUARES APPROXIMATION

As was shown, the continuous unfolding problem could be formulated as an extremal problem in a finite number of unknowns, $u = (u_1, \ldots u_m)$, with a continuous constraint inequality. Then by means of upper and lower piecewise approximations to the response function and window function, it was possible to obtain a finite constraint inequality. Since the true spectrum $x(s)$ never had to appear in the resulting programming formulation, it was possible to proceed directly from the experimental data to the desired results without any consideration of $x(s)$. This fortunate circumstance depended, however, upon a priori knowledge that $x(s) \geqq 0$. It was shown also that there is a nonnegative vector $z = (z_1, \ldots z_n)$ which is dual to u. The unfolding problem can be formulated in terms of this vector, which serves in analyzing errors due to the response matrix and, in general, also serves in lieu of an estimate of the spectrum itself whenever $x(s)$ is needed in the conventional unconstrained estimation.

The starting point for the modified least-squares problem is, as always, the continuous problem:

given:

$$\int K_i(s)\ x(s)\ ds = \hat{b}_i + e_i, \quad x(s) \geqq 0 \quad ,$$

e is normally distributed with known variance matrix S; (9.1)

wanted:

$$\text{confidence interval } \left[\hat{\phi}^{lo}, \hat{\phi}^{up}\right] \text{ for } \phi = \int w(s)\ x(s)\ ds \quad .$$

First, it is assumed that $K_i(s)$ and $w(s)$ can be replaced by a piecewise linear or parabolic approximation $K_i(s)^{approx}$ and $w(s)^{approx}$. Then it suffices to consider only values of these approximate functions at the comparison points $s_1, s_2, \ldots s_n$. These values constitute the matrix $K_{ij} = K_i(s_j)$ and the vector $w_j = w(s_j)$. So far the only difference in assumptions between the modified least-squares method and the general method is that instead of an upper and lower piecewise approximation being taken for $K_i(s)$ and $w(s)$ a single piecewise approximation is assumed to be sufficient. In practice, a discrepancy of the order of 1% between the piecewise approximations and the actual functions is often unimportant if the response vector components have comparable errors.

The dual finite programming problem which yields the confidence interval is then

$$\hat{\phi}^{up} = \max_{z \geq 0} \left\{ w^T z \;\middle|\; (Kz - \hat{b})^T S^{-1} (Kz - \hat{b}) \leq \kappa^2 \right\} ,$$

$$\hat{\phi}^{lo} = \min_{z \geq 0} \left\{ w^T z \;\middle|\; (Kz - \hat{b})^T S^{-1} (Kz - \hat{b}) \leq \kappa^2 \right\} . \qquad (9.2)$$

Now the solution to this dual problem is the solution to the following constrained estimation problem which serves as the point of departure for the modified least-squares method:

given:

$$\left.\begin{array}{l} Kz = \hat{b} + e, \\ z \geqq 0, \\ e \text{ is normally distributed with known variance matrix } S; \\ \text{wanted:} \\ \text{confidence interval } \left[\hat{\phi}^{lo}, \hat{\phi}^{up}\right] \text{ for } \phi = w^T z \quad . \end{array}\right\} \quad (9.3)$$

An effort has been made up to now to avoid the misleading (though valid) suggestion that z is a discrete approximation to the continuous spectrum x(s). But this implication emerges strongly now from Eq. (9.3). It is more desirable to think of Eq. (9.3) as equivalent to the continuous problem with response functions $K_i(s)^{approx}$ and window function $w(s)^{approx}$. This is a different motivation from the usual one of replacing an integral equation by a set of approximating equations. The assumptions which underlie Eq. (9.3) apply to the response function and window function and <u>not</u> to the spectrum x(s), which is completely arbitrary except for the nonnegativity requirement that $x(s) \geqq 0$.

<u>Primal formulation</u>

As in the minimum variance approach to ordinary estimation, a combination of response functions, $u^T K$, will be found which approximates the desired window w^T, but it will not be required that the combination $u^T K$ agree exactly with the window for each component. Then the overall confidence interval must take the discrepancy (or "slop") between $u^T K$ and w^T into account. In ordinary estimation, it is impossible to evaluate the error due to this discrepancy (hence the necessity for the

unbiased condition), but in constrained estimation, the a priori non-negativity information can be used to find an interval $[zmin_j, zmax_j]$ for z_j. Then an upper bound for the error can be found.

A lower bound, $zmin_j$, of zero for z_j is given immediately from $z \geqq 0$. But if $K \geqq 0$, an upper bound can also be obtained by the following simple method. First notice that the constraint inequality of Eq. (9.2) implies the weaker inequality

$$\max_i \left\{ \frac{1}{\sigma_i} \left| Kz - \hat{b} \right|_i \right\} \leqq \kappa \tag{9.4.1}$$

or

$$(Kz)_i \leqq \kappa\sigma_i + b_i \quad \text{for all } i \quad . \tag{9.4.2}$$

Then

$$K_{ij}\, z_j \leqq \kappa\sigma_i + b_i \qquad \begin{array}{l} i = 1, 2, \ldots m \\ j = 1, 2, \ldots n \end{array} \tag{9.5}$$

since only nonnegative terms have been dropped. Since Eq. (9.5) must hold for any i, that i which gives the least upper bound is picked:

$$zmax_j = \min_i \left[(b_i + \kappa\sigma_i)/K_{ij} \right], \quad i = 1, \ldots m \quad . \tag{9.6}$$

The geometrical interpretation of this bound is illustrated in Fig. 21. Equation (9.2) implies that z lies in an elliptical region (possibly degenerate). The inequality of Eqs. (9.4) means that the elliptical region has been replaced with a box-shaped region which encloses it.

ORNL-DWG 64-999

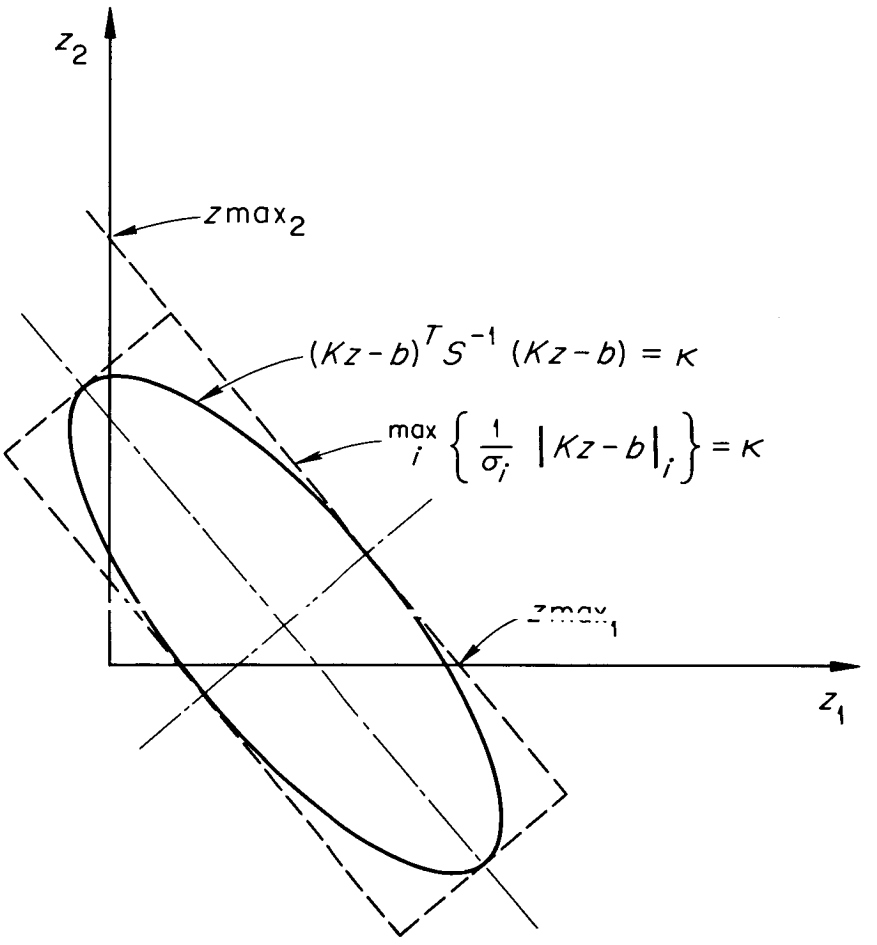


Fig. 21. A geometric interpretation of the weak upper bound, xmax, for z.

Then the bound implied by Eq. (9.6) is given by the smallest intersection of the edges of this diamond with z_j axes.

The weakness of this interval can be partly overcome by including in the modified least-squares procedure the calculation of a confidence interval $[zmin_j, zmax_j]$ in addition to the confidence intervals for ϕ. Then the improved bounds may be used for a second iteration.

In order to utilize both the upper and lower bound for z_j (in the anticipation that a nonzero lower bound for z_j may be obtained on the first iteration), the estimation problem is expressed in terms of $(z - d)$, where

$$d = 0.5\ (zmax + zmin) \quad , \tag{9.7}$$

so that

$$-\ 0.5\ (zmax - zmin) \leqq (z - d) \leqq 0.5\ (zmax - zmin) \quad . \tag{9.8}$$

Then the maximum uncertainty in $(z - d)$ is only half that in z since the expansion is about the center of the interval.

Now ϕ can be expressed as

$$\phi = w^T(z - d) + w^T d \quad . \tag{9.9}$$

When the unbiased condition $w^T = u^T K$ is replaced by

$$w^T = u^T K + (w^T - u^T K) \quad , \tag{9.10}$$

where the quantity in parentheses is the discrepancy between the combination of response functions and the desired window vector, Eq. (9.9) becomes

$$\phi = u^T K(z - d) + (w^T - u^T K)(z - d) + w^T d \quad ; \tag{9.11}$$

and since $Kz = b$,

$$\phi = w^T d + u^T(b - Kd) + (w^T - u^T K)(z - d) \quad . \tag{9.12}$$

Now, because there are bounds for $(z - d)$ from Eq. (9.8),

$$\min_{z-d} \left[w^T d + u^T(b - Kd) + (w^T - u^T K)(z - d) \right]$$

$$\leqq \phi$$

$$\leqq \max_{z-d} \left[w^T d + u^T(b - Kd) + (w^T - u^T K)(z - d) \right] \quad , \tag{9.13}$$

where $-0.5\ (zmax - zmin) \leqq (z - d) \leqq 0.5\ (zmax - zmin)$.

Since the maximization and minization is only over $(z - d)$, which is bounded by Eq. (9.8),

$$w^T d + u^T(b - Kd) + \left[w^T - u^T K \right]_{-}^{+} \begin{matrix} (z - d)min \\ (z - d)max \end{matrix}$$

$$\leqq \phi$$

$$\leqq w^T d + u^T(b - Kd) + \left[w^T - u^T K \right]_{-}^{+} \begin{matrix} (z - d)max \\ (z - d)min \end{matrix} \quad , \tag{9.14}$$

where again the convention is used that the upper elements are to be used if the sign of the bracket is +, and the lower elements if the sign of the bracket is - .

Finally, the confidence interval $[\hat{\phi}^{lo}, \hat{\phi}^{up}]$ for ϕ is obtained from Eq. (9.14) by taking into account the variance of $\hat{\phi}$ due to the errors in $\hat{b}$. Since $\mathrm{var}(u^T\hat{b}) = u^T S u$,

$$\hat{\phi}^{up} = w^T d + u^T(\hat{b} - Kd) + \left[w^T - u^T K \right]_{-(z-d)\min}^{+(z-d)\max} + \kappa \sqrt{u^T S u}$$

$$\hat{\phi}^{lo} = w^T d + u^T(\hat{b} - Kd) + \left[w^T - u^T K \right]_{-(z-d)\max}^{+(z-d)\min} - \kappa \sqrt{u^T S u} \quad , \qquad (9.15)$$

where κ depends upon the confidence level desired. The computational problem to be solved is the determination of the coefficient vector u so that the confidence interval will be as narrow as possible. Thus it will be desired to minimize the variance term and at the same time keep the errors due to the discrepancy $[w^T - u^T K]$ between the combination of response functions and the desired window as small as possible.

In order to obtain a simple solution, the quantity to be minimized is taken to be

$$\ell = \tau\, u^T S u + (w^T - u^T K)\, D^2 (w^T - u^T K)^T \quad , \qquad (9.16)$$

where D is a diagonal matrix with elements $D_{jj} = 0.5\ (\mathrm{zmax}_j - \mathrm{zmin}_j)$. The first term in Eq. (9.16) is just the variance of the estimate. The relation of the second term to the estimators in Eq. (9.15) depends upon the following inequalities:

$$(w^T - u^T K)\, D^2 (w^T - u^T K)^T \leq \left[\sum_{j=1}^{n} |(w^T - u^T K)_j| \quad D_{jj} \right]^2$$

$$\leq n\ (w^T - u^T K)\, D^2 (w^T - u^T K)^T \quad . \qquad (9.17)$$

The first equality holds if and only if just one component of $(w^T - u^T K)$ is nonzero. The second equality holds if all components of $(w^T - u^T K)$ are equal. Although the value of τ used in Eq. (9.16) could be selected to yield the minimum-width confidence interval, it is fairly complicated to find the optimal value, and we consider τ an empirical constant. Fortunately, the width of the confidence interval is not very sensitive to τ. A value of τ from about $1/n$ to about 1 seemed to be the best in most cases tried.

In order to find u which minimizes Eq. (9.16), $d\ell/du_j$ is set equal to 0 for each component and the resulting set of equations is solved. This results in

$$u^T = w^T D^2 K^T (\tau S + K D^2 K^T)^{-1} \quad . \tag{9.18}$$

Dual formulation

As is always the case, the method for finding u also has a dual. And as is almost always the case, the dual has an interesting algebriac and geometric interpretation which clarifies the problem. The dual form is

$$\hat{\phi}^{up} = \max_z \left\{ w^T z \;\middle|\; (Kz - \hat{b})^T S^{-1} (Kz - \hat{b}) + \tau (z - d)^T D^{-2} (z - d) = \kappa'^2 \right\} , \tag{9.19}$$

where

$$\kappa'^2 = \min_z \left[(Kz - \hat{b})^T S^{-1} (Kz - \hat{b}) \right] + \kappa^2 \tag{9.20}$$

and κ is given by Eq. (2.10). In this form, the problem is identical

to ordinary least squares except for the extra term $(z - d)^T D^{-2}(z - d)$. The behavior of this term is illustrated in Fig. 22. It is seen that the addition of this term tends to constrain z to the a priori region $z_{min} \leq z \leq z_{max}$. In fact, if the extremal problem could be solved for a p sufficiently larger than 2 (as illustrated in Fig. 22 by the solid curve), then this would be a exact solution to the quadratic programming problem. From this point of view, the modified least-squares method considers that the nonnegativity constraint $z \geq 0$ is equivalent to $0 \geq z \geq z_{max}$, where z_{max} is so large that the constraint is effective only near $z = 0$. Then in the modified least-squares method, the "box"-shaped region $0 \leq z \leq z_{max}$ is replaced by an elliptical-shaped region which approximates the box. Then the confidence interval is determined by the combination of the ordinary confidence ellipsoid (possibly degenerate) and the new constraint ellipsoid.

To see that the solution of the dual in Eq. (9.19) is the same as the primal from Eq. (9.18), it is noted that

$$z_o = (K^T S^{-1} K + \tau D^2)^{-1} (K^T S^{-1} b + D^2 d)$$

is the center of the composite ellipsoid corresponding to the quadratic form

$$(Kz - \hat{b})^T S^{-1}(Kz - \hat{b}) + \tau(z - d)^T D^{-2}(z - d) \quad ,$$

whose characteristic matrix is

$$V = (K^T S^{-1} K + \tau D^2)^{-1} \quad .$$

ORNL-DWG 64-1000

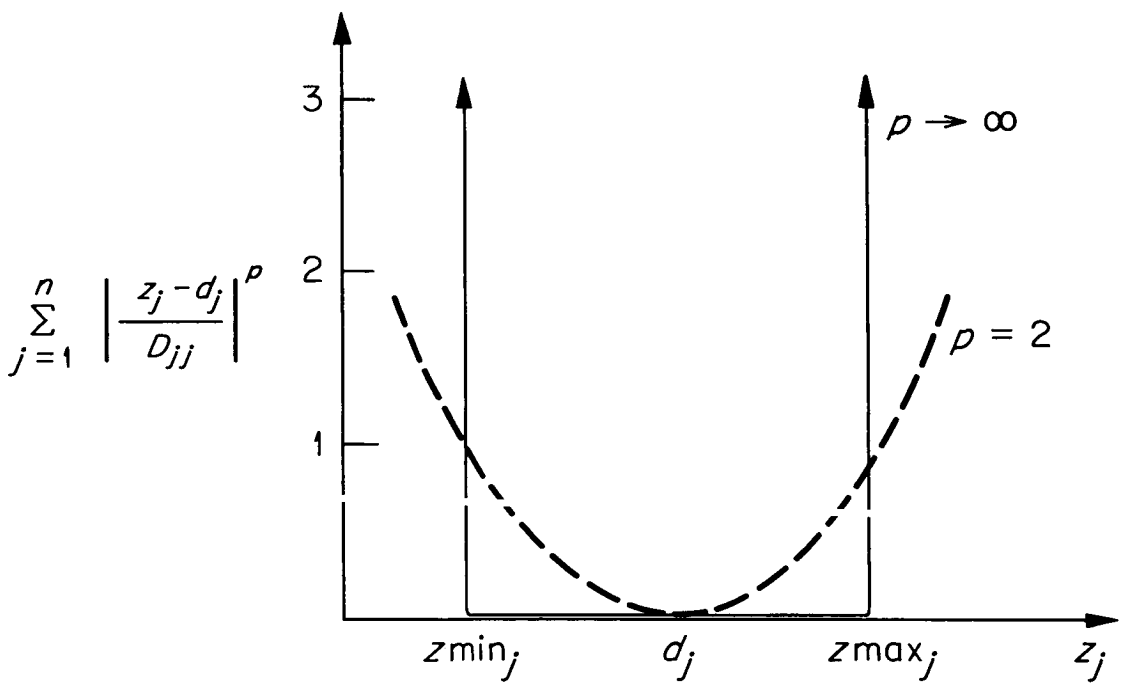


Fig. 22. Behavior of $(z - d)^T D^{-2}(z - d) = |(z_j - d_j)/(D_{jj})|^2$, illustrated in one dimension. Also shown is the limiting form when the absolute value term is raised to a higher power than 2.

Then the solution to the estimation problem is

$$\hat{\phi}^{up} = w^T \hat{z}^{up} = w^T \hat{z}_o + \tau \sqrt{w^T V w} \quad .$$

For comparison, the primal solution from Eq. (9.18) and the dual solution from Eq. (9.19) are given below:

$$\hat{\phi}_{primal} = w^T \left[D^2 K^T (\tau\ S + K D^2 K^T)^{-1} (\hat{b} - Kd) + d \right] \quad ,$$

$$\hat{\phi}_{dual} = w^T (K^T S^{-1} K + \tau\ D^{-2})^{-1} (K^T S^{-1} \hat{b} + \tau\ D^{-2} d) \quad .$$

To prove that these are the same, first notice that

$$(K^T S^{-1} K + \tau\ D^{-2})^{-1} (K^T S^{-1} \hat{b} + \tau\ D^{-2} d)$$

$$= (K^T S^{-1} K + \tau\ D^{-2})^{-1} K^T S^{-1} (\hat{b} - Kd) + d \quad ,$$

since multiplying both by $(K^T S^{-1} K + \tau\ D^{-2})$ gives

$$K^T S^{-1} \hat{b} + \tau\ D^{-2} d = K^T S^{-1} (\hat{b} - Kd) + (K^T S^{-1} K + \tau\ D^{-2})\ d \quad .$$

Thus

$$\hat{\phi}_{dual} = w^T \left[(K^T S^{-1} K + \tau\ D^{-2})^{-1}\ K^T S^{-1} (b - Kd) + d \right]$$

Finally it can be shown that

$$D^2 K^T (\tau\ S + K D^2 K^T)^{-1} = (K^T S^{-1} K + \tau\ D^{-2})^{-1}\ K^T S^{-1}$$

by multiplying both sides on the left by $(K^T S^{-1} K + \tau D^{-2})$ and on the right by $(\tau S + KD^2K^T)$, and collecting terms.

Numerical calculation

Although the confidence intervals could be calculated from the formulas given, it is better not to use either the primal or the dual results but to use an orthonormalizing calculation to solve the following augmented least-squares problem:

$$\begin{pmatrix} K \\ I \end{pmatrix} z = \begin{pmatrix} \hat{b} \\ \hat{d} \end{pmatrix}$$

with

$$\mathrm{var}\begin{pmatrix} \hat{b} \\ \hat{d} \end{pmatrix} = \begin{pmatrix} S & 0 \\ 0 & 1/\tau \cdot D^2 \end{pmatrix} .$$

It is readily seen that the solution is the same as given before. By using an orthonormalizing routine (52) rather than forming $(K^T S^{-1} K + \tau D^{-2})^{-1}$ or $(\tau S + KD^2K^T)^{-1}$, approximately half as many figures are lost due to arithmetic round-off during the computation. The use of an orthonormalizing method avoids the formation of a symmetric product $K^T S^{-1} K$ or KD^2K^T and the accumulation of round-off errors due to this step.

A flow chart of the computation is given in Fig. 23. The computation is organized so that the minimization is performed before any of the windows are read in. Steps 5 and 6 determine a better inequality for the dual vector z. After each recalculation of zmin and zmax, another estimate of bias_{lo}^{up} is made. This iterative "bias" loop usually converges in several cycles; M = 5 is a typical number of

ORNL-DWG 64-1004

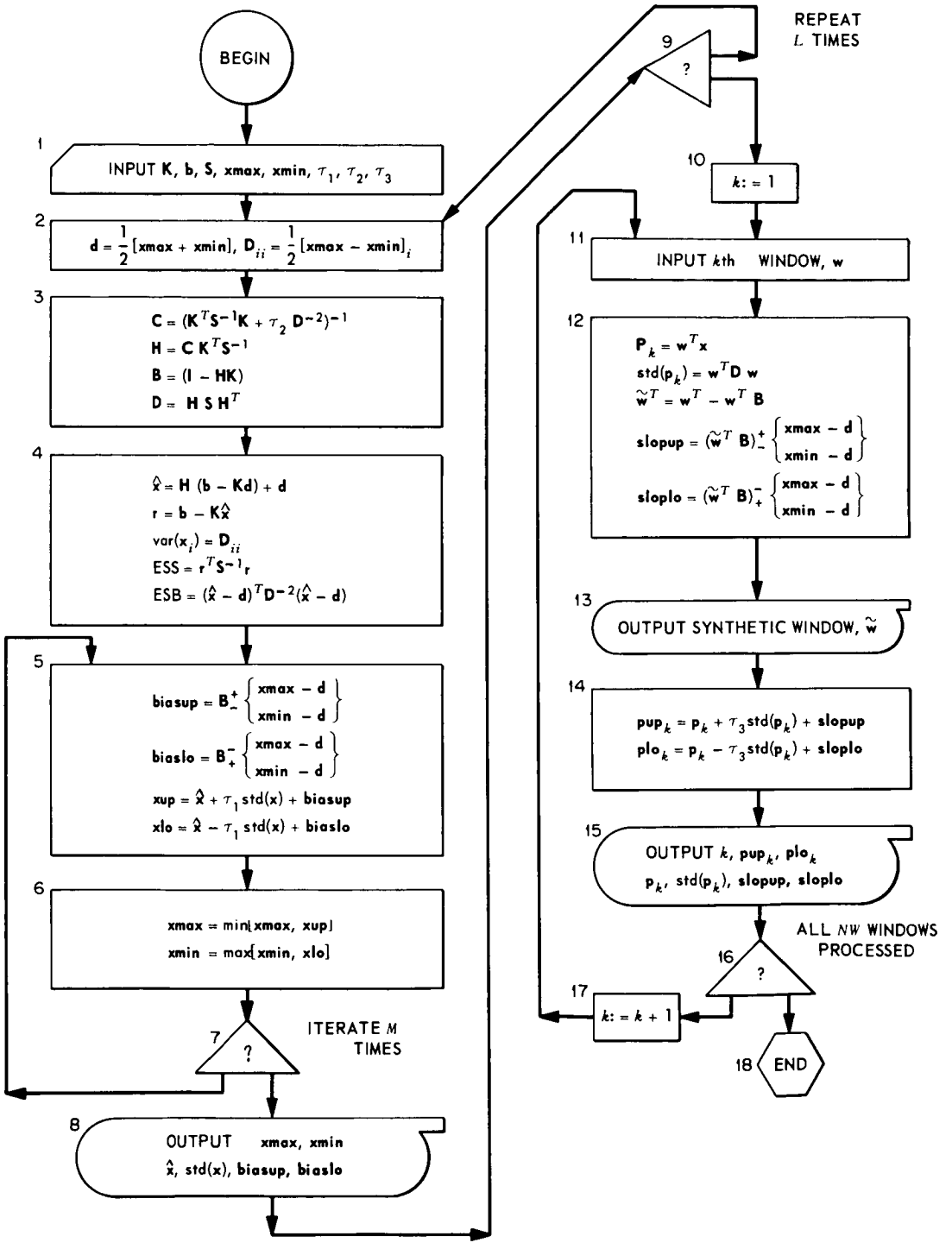


Fig. 23. Flow diagram for the modified least-squares calculation.

iterations. If desired, the entire least-squares procedure can be repeated by using the improved inequality for z. This reduces any small negative components of z which remain. Because almost all the computation must be repeated, L is usually taken as 1 or 2 in moderate problems, and perhaps as large as 3 in small problems. Usually no significant improvement is noticed after the second iteration.

In the second stage of the computation, the window vectors are read in one at a time, and the confidence interval for ϕ is calculated and published. As many windows as desired may be processed, since only one needs to be loaded in the computer memory at a time. A rough estimate of the computation time can be obtained based on the number of floating-point operations that must be executed:

$$\text{NO OPS} \cong 4 \cdot L \cdot [NC\ (NC + M) + 2 \cdot NW] \cdot (NH + M) \quad ,$$

where in a typical problem (the neutron unfolding example of Chapter 3),

number of windows	NW = 40
number of comparison energies	NH = 36
number of count bins	NC = 87
number of "bias" loops	M = 5
number of iterations	L = 1

For the IBM-7090 computer, the execution time for this problem was about 1 min, including input and output. The reason that the modified least-squares method is faster at present than the linear programming code is that the form of the solution for u is independent of the window function w. Therefore most of the arithmetical problems can be solved

for all windows at once. The price paid for the speed of the modified least-squares method is, of course, that the final confidence intervals are not as narrow as are theoretically possible.

10. MISCELLANEOUS APPLICATIONS

Solution of linear equations

In addition to unfolding instrument response, the general concept of constrained estimation is useful for many other applications. For example, the solution to simultaneous linear equations is often greatly improved by the use of *a priori* information. As a dramatic numerical example, the system of equations given below has been solved:

$$\begin{bmatrix} 1 & 1/2 & 1/3 & 1/4 & 1/5 & \dots & 1/20 \\ 1/2 & 1/3 & 1/4 & 1/5 & 1/6 & \dots & 1/21 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1/20 & 1/21 & 1/22 & 1/23 & 1/24 & \dots & 1/39 \end{bmatrix} \begin{bmatrix} \\ x \\ \\ \end{bmatrix} = \begin{bmatrix} 1 \\ 1/2 \\ \dots \\ 1/20 \end{bmatrix} . \quad (10.1)$$

where the matrix of the system is the famous Hilbert matrix of order 20. The solution of this particular set of equations by conventional methods is extremely difficult, owing to the large magnitude of the elements of the inverse Hilbert matrix. For the 20 by 20 matrix H,

$$H^{-1} = \begin{bmatrix} 4.000\dots\ 10^{2} & -7.980\dots\ 10^{4} & 5.266\dots\ 10^{6} & \dots & -1.378\dots\ 10^{12} \\ -7.980\dots\ 10^{4} & 2.122\dots\ 10^{7} & -1.576\dots\ 10^{9} & \dots & 5.238\dots\ 10^{14} \\ \dots & \dots & \dots & \dots & \dots \\ -1.378\dots\ 10^{12} & 5.238\dots\ 10^{14} & -4.950\dots\ 10^{16} & \dots & 4.872\dots\ 10^{22} \end{bmatrix}$$

If just one element of the right-hand side has an error of 1 part in 10^8, then the solution will be in error by a factor of about 10^{10}! Thus in order to obtain just one decimal figure of accuracy, it is necessary to carry approximately 30 decimal places of precision in the calculation, or else resort to special techniques in which the intermediate steps are

expressed as rational fractions. But if it is known in advance that the solution x must be nonnegative, then the constrained estimation method can be used. The example of Eq. (10.1) was constructed so that the solution was just x = (1, 0, 0, ... 0). Then x was solved for by the linear programming method and by the modified least-squares method. In order to take into account the effect of machine round-off on the solution, it was assumed that each component of b had a random error of 10^{-8}, since the word length of the IBM-7090 computer corresponds to about eight decimal digits. The calculated confidence intervals are shown in Table 1 for the two methods at the 95% confidence level. In this case, it is seen that the use of the a priori information reduced the uncertainty in the solution by about 20 orders of magnitude. The linear programming computation required about 30 sec, and the modified least-squares computation required about 5 sec.

Unfolding gamma spectra

The interpretation of measurements obtained with gamma scintillation spectrometers is complicated by the presence of "tails," "edges," and escape peaks and varying efficiency of the scintillator response. Some pulse-height distributions for a 9- by 12-in. total absorption NaI spectrometer are shown in Fig. 24 for several unit intensity monoenergetic gamma sources. The modified least-squares method was used to unfold the gamma spectrum due to capture of neutrons by the 27-keV resonance of fluorine (measured by J. R. Bird, on loan to ORNL from Atomic Energy Research Establishment, Harwell, England). Figure 25 gives the upper portion of the pulse-height distribution. The pulse-height channels were of equal width. Figure 26 gives the 90%

TABLE 1

The 95% Confidence Interval for the Solution to Eq. (10.1)

		Linear Programming Method		Modified Least Squares	
i	True Solution	Lower Edge	Upper Edge	Lower Edge	Upper Edge
1	1.000000	0.999990	1.000009	0.999993	1.000005
2	0.0	0.0	0.0003	-0.001	0.001
3	0.0	0.0	0.0005	-0.003	0.003
4	0.0	0.0	0.002	-0.009	0.009
5	0.0	0.0	0.002	-0.007	0.007
6	0.0	0.0	0.006	-0.010	0.010
7	0.0	0.0	0.009	-0.021	0.021
8	0.0	0.0	0.014	-0.022	0.022
9	0.0	0.0	0.015	-0.024	0.024
10	0.0	0.0	0.017	-0.025	0.025
11	0.0	0.0	0.024	-0.041	0.041
12	0.0	0.0	0.038	-0.074	0.074
13	0.0	0.0	0.032	-0.052	0.052
14	0.0	0.0	0.030	-0.049	0.049
15	0.0	0.0	0.026	-0.036	0.036
16	0.0	0.0	0.023	-0.058	0.058
17	0.0	0.0	0.023	-0.059	0.059
18	0.0	0.0	0.017	-0.027	0.027
19	0.0	0.0	0.004	-0.009	0.009
20	0.0	0.0	0.008	-0.014	0.014

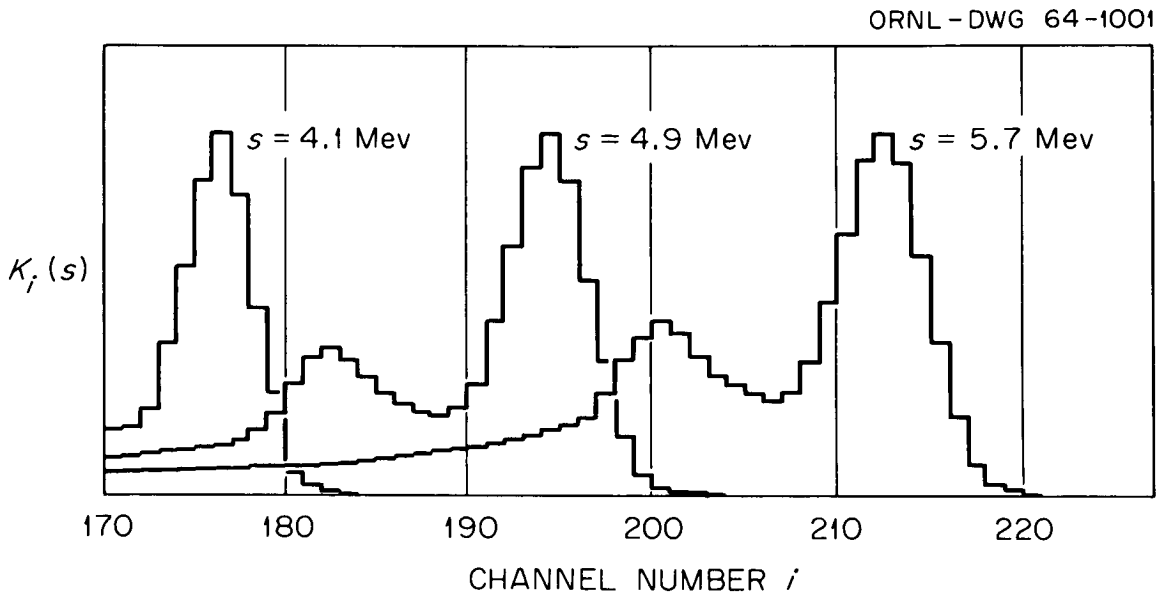


Fig. 24. Pulse-height distribution obtained with 9 by 12 in. total absorption NaI(Tl) scintillation spectrometer for monoenergetic gamma radiation.

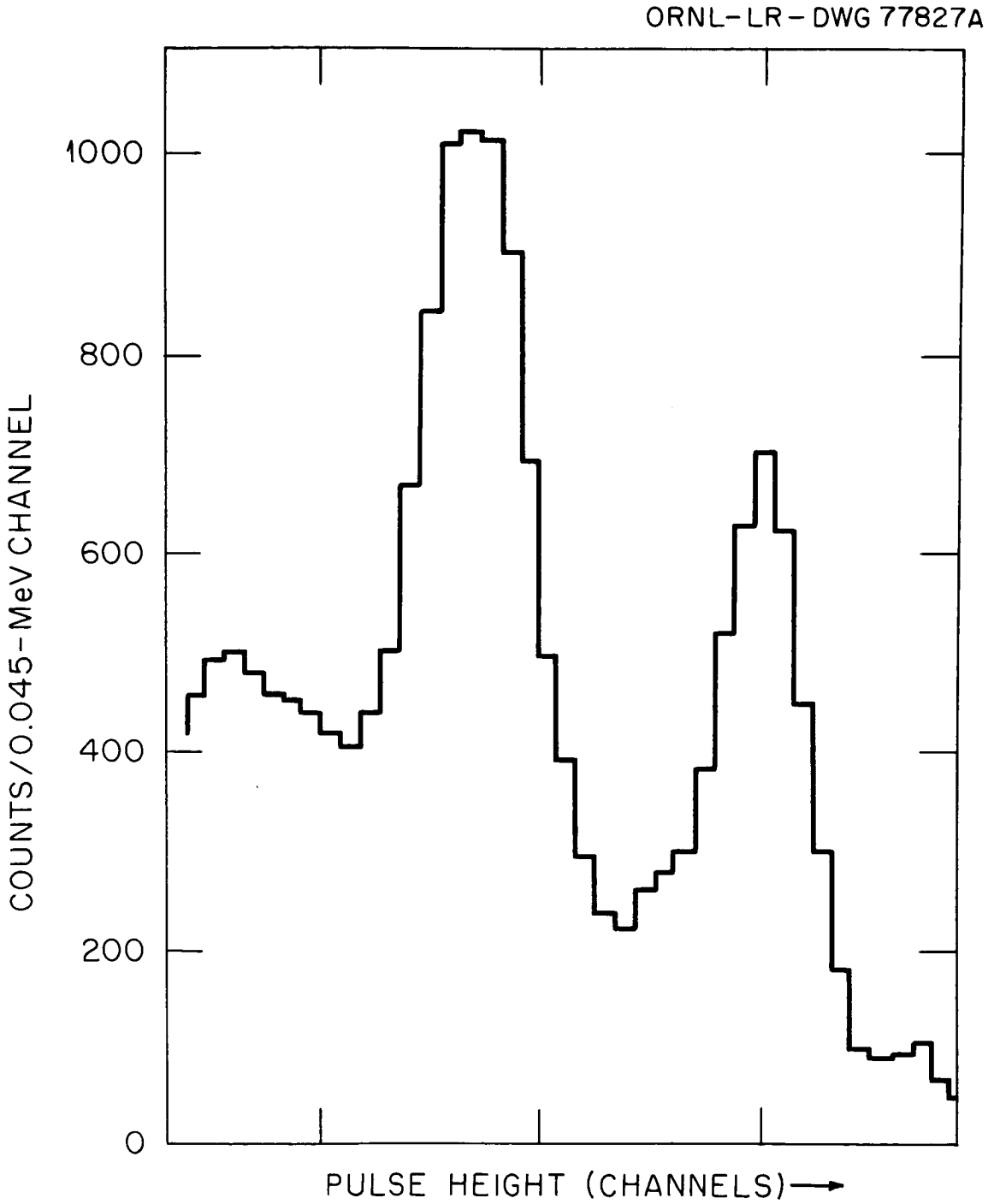


Fig. 25. Upper portion of pulse-height distribution for gamma rays resulting from neutron capture by the 27-keV resonance of fluorine.

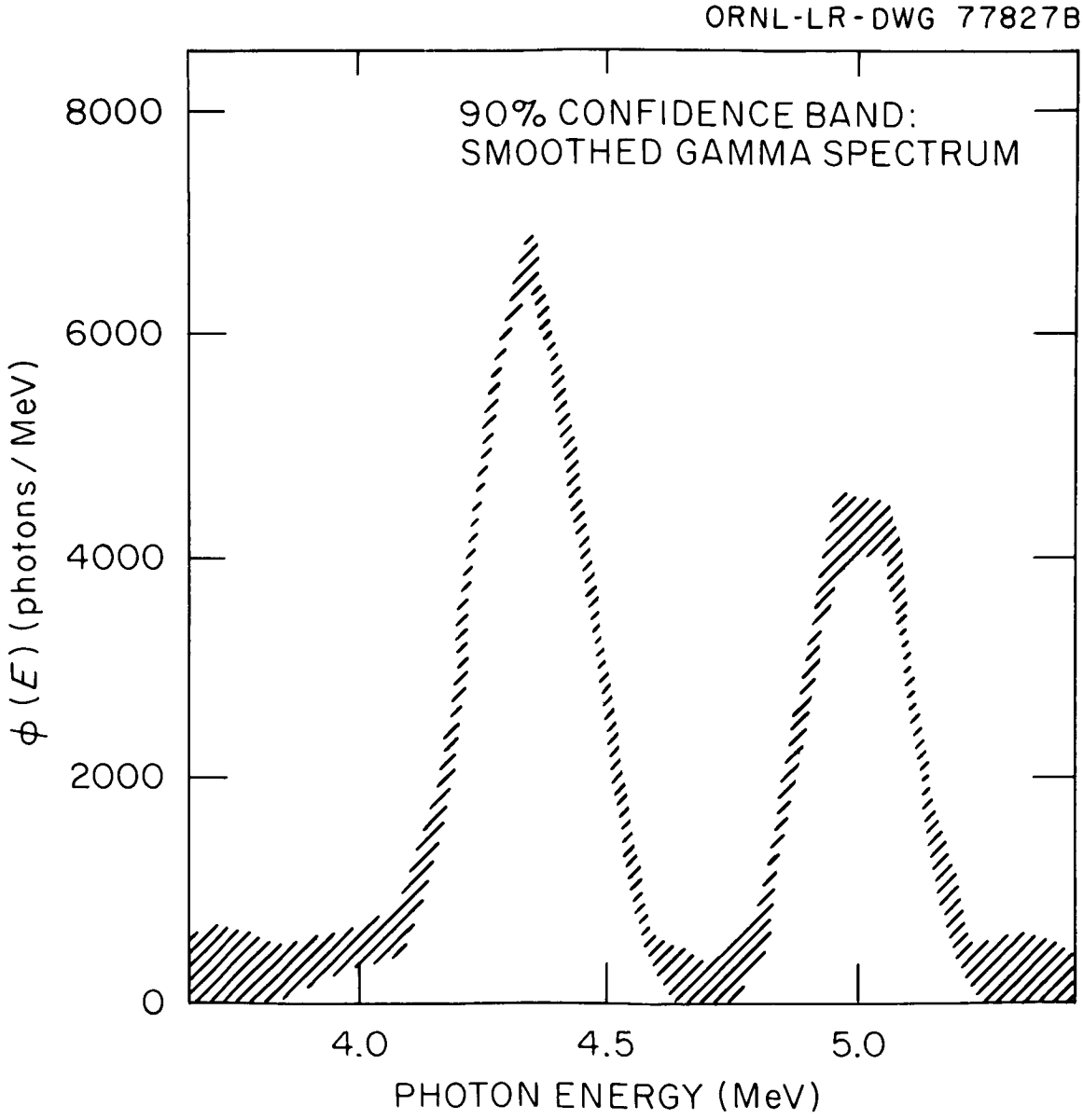


Fig. 26. Unfolded gamma spectrum. Since the window functions used in the unscrambling had a full width at half maximum of 300 keV, the resulting spectrum is a smoothed approximation to the true spectrum.

confidence intervals for a number of Gaussian window functions with a full width at half maximum of 300 keV. The modified least-squares code was used for this calculation with

number of windows	NW = 52
number of comparison energy	NH = 38
number of count bins	NC = 85
number of "bias" loops	M = 5
number of iterations	L = 2

and a computation time of about 2 min.

Note that only the higher energy portion of the spectrum was unfolded. This is possible because of the one-sided character of the response functions. It would not be possible, however, to obtain just the low-energy portion of the spectrum since it is contaminated by the pulses from the higher energy portion.

Other applications

Another possible application is the numerical inversion of a Fredholm integral equation of the form

$$f(x) = \int_a^b K(x,y)\ g(y)\ dy$$

when the unknown function is known to be nonnegative. First, the continuous variable is replaced by a finite set (x_1, x_2, ... x_n), and then

the resulting problem is solved by the constrained estimation method. Some advantages over the ordinary methods are the following:

1. Discontinuities in the kernel $K(x,y)$ can be easily handled by the techniques discussed in Chapter 4.

2. A singular kernel offers no special difficulties.

3. The result is in the form of an interval containing the true solution. The width of this interval includes the effect of replacing the continuous variable with a discrete set and the effect of all other computational devices. Thus there is no question as to the reliability of the numerical solution (apart from round-off).

Another interesting possibility is the use of the constrained estimation method for the solution of differential equations. A common method involves replacing the continuous differential operator with a discrete operator and solving the resulting system of linear equations. It is well known (53) that the discrete operators resulting from self-adjoint differential equations have nonnegative inverses. Thus the non-negativity condition arises naturally for a large class of differential equations. In this connection, the interesting paper by Young (54) on the use of linear programming for the solution of differential equations from a different point of view should be noted.

11. ADDITIONAL CONSTRAINTS

Although the nonnegativity condition often arises in a natural way in the solution of physically motivated unfolding problems, occasionally stronger *a priori* information is available. For example, it may be known that the unknown spectrum $x(s)$ has a certain amount of smoothness. A traditional way to incorporate such information is by means of the Wiener-Kolmogrov smoothing theory (55). But the existence of a suitable *a priori* probability distribution as required by the theory is sometimes difficult to justify. The constrained estimation method provides an alternative approach which is free from *a priori* probability considerations.

In many cases the additional information on $x(s)$ is based upon the knowledge that $x(s)$ is the result of folding a smoothing-type kernel $R(s,s_o)$ with another more fundamental spectrum, $y(s_o)$. For example, in neutron fission the gamma rays which are emitted from the fission fragments have a continuous distribution because of the random Doppler shifts from the moving particles. This shift is statistically distributed with a standard deviation of about 30 keV for an energy of 1 MeV. Thus, if there were a monoenergetic decay in the moving fission-fragment system, a broadened distribution of photon energies would result in the laboratory system in which the measurements are made. As another example, multiply scattered neutrons have a smoother spectrum than the source spectrum because of the continuous character of the scattering kernel. Thermal broadening and lifetime broadening are also familiar examples of this kind of process.

In such cases it has been found desirable to introduce this information into the estimation problem. Let it be assumed that

$$x(s) = \int R(s,s_o)\ y(s_o)\ ds_o \quad , \tag{11.1.1}$$

$$y(s_o) \geqq 0 \quad . \tag{11.1.2}$$

In other words, the desired spectrum $x(s)$ is the transformation of a nonnegative function $y(s_o)$ with a kernel $R(s,s_o)$. Then under certain conditions, $x(s)$ will be smoother than $y(s_o)$ (56). A nonnegativity condition still applies, but to $y(s_o)$ instead of to the unknown spectrum $x(s)$. If $y(s_o)$ could be restrained to be nonnegative, then $x(s)$ would be at least as smooth as the smoothing kernel $R(s,s_o)$.

Taking into account $y(s_o)$, the estimation problem is:

given:

$$\int K_i(s)\ x(s)\ ds = \hat{b} + e \quad , \tag{11.2.1}$$

e is normally distributed with known variance matrix S , (11.2.2)

$$x(s) = \int R(s,s_o)\ y(s_o)\ ds_o \quad , \tag{11.2.3}$$

$$y(s_o) \geqq 0 \quad ; \tag{11.2.4}$$

wanted:

confidence interval $\left[\hat{\phi}^{lo},\ \hat{\phi}^{up}\right]$ for $\phi = \int w(s)\ x(s)\ ds$. (11.2.5)

Note that the function ϕ of Eq. (11.2.5) is still a function of $x(s)$, instead of the underlying spectrum $y(s_o)$. It is my opinion that the purpose of the instrument response unfolding problem is to answer questions about the input to the instrument, $x(s)$, having observed the instrument output $\hat{b} = (\hat{b}_1, \hat{b}_2, \ldots \hat{b}_m)$. Nevertheless, the experimenter's main interest often lies in the underlying spectrum $y(s_o)$, and one can substitute

$$\phi' = \int w'(s_o)\ y(s_o)\ ds_o \tag{11.3}$$

for Eq. (11.2.5) if desired. Another reason I prefer to use Eq. (11.2.5) is that then the smoothing kernel $R(s, s_o)$ need be known only approximately. If it is known that the spectrum $x(s)$ should have a certain smoothness, it leads to a conservative error to assume a broadening kernel less smooth than theory indicates. Then the solution $x(s)$ will not be constrained as strongly as it could be, and the confidence interval will be valid. In any event, whether it is decided to include the interpretation of the underlying spectrum $y(s_o)$ as a part of the unfolding problem, or to deal with this part as a separate problem, the resulting estimation problem is of essentially the same form.

Equations (11.2) can be put into exactly the same form as for the previous case in Chapter 3, where $x(s) \geqq 0$, if the expression for

$x(s)$ is introduced in terms of $y(s_o)$ into the integrals. Then the problem is:

$$\left.\begin{array}{l}
\text{given:} \\
\quad \int K_i(s) \left[\int R(s,s_o)\ y(s_o)\ ds_o \right] ds = \hat{b}_i + e_i \quad , \\
\quad e \text{ is normally distributed with known variance matrix } S \quad , \\
\quad y(s_o) \geqq 0 \quad ; \\
\text{wanted:} \\
\quad \text{confidence interval } \left[\hat{\phi}^{lo}, \hat{\phi}^{up} \right] \text{ for} \\
\qquad\qquad \phi = \int w(s) \left[\int R(s,s_o)\ y(s_o)\ ds_o \right] ds \quad .
\end{array}\right\} \quad (11.4)$$

This may be simplified to:

$$\left.\begin{array}{l}
\text{given:} \\
\quad \int \tilde{K}_i(s_o)\ y(s_o)\ ds_o = \hat{b}_i + e_i \quad , \\
\quad e \text{ is normally distributed with known variance matrix } S \quad , \\
\quad y(s_o) \geqq 0 \quad ; \\
\text{wanted:} \\
\quad \text{confidence interval } \left[\hat{\phi}^{lo}, \hat{\phi}^{up} \right] \text{ for } \phi = \int \tilde{w}(s_o)\ y(s_o)\ ds_o \quad ,
\end{array}\right\} \quad (11.5)$$

where

$$\tilde{K}_i(s_o) = \int K_i(s)\ R(s,s_o)\ ds \quad , \tag{11.6}$$

and

$$\tilde{w}(s_o) = \int w(s)\ R(s,s_o)\ ds \quad . \tag{11.7}$$

Equation (11.5) is of exactly the same form as that previously obtained and can be solved by the same methods. Thus the incorporation of the more generalized constraints requires only modifying the response functions according to Eq. (11.6) and the window function according to Eq. (11.7). The reduction to discrete form should be done after the modification rather than before, since the modified functions are smoother and hence require fewer comparison points for a satisfactory piecewise approximation.

Besides choosing a physically motivated smoothing kernel $R(s,s_o)$ to impose regularity on the solution, kernels can be devised which will constrain the solutions to be monotonically decreasing, concave, etc. In addition, a series of smoothing kernels can be applied in cascade. Several other possible ways of introducing information into the formulation of the problem are possible, but perhaps are best discussed with respect to a particular application.

12. CONCLUSIONS

The basic motivation of the unfolding methods which have been discussed is the desire to use as much information as is known about the solution, consistent with the cost of obtaining a solution. The incorporation of nonnegativity of a spectral function into the unfolding problem is conceptually simple. The unfolding problem is merely posed in such a way to allow use of the intersection of the conventional solution for the spectrum with the positive region of the solution.

It has been shown by example that several typical practical problems, which are difficult to approach from the classical methods, yield easily to the new method. In addition, the primal-dual structure of constrained extremal problems provides a means of justifying the replacement of a continuous problem by a discrete one and allows errors in the kernel (in the approximation of small errors) to be properly taken into account. Computationally, it is no longer necessary to distinguish between a singular and a nonsingular kernel, since the mathematical formulation and solution of the problem are identical.

The two codes described here are presently in use. They are limited to problems that can be reduced to a finite form of order about 100 by 100. With the development of newer computers and the development of more efficient algorithms, this size limitation will surely be removed.

LIST OF REFERENCES

1. C. Eckart, Phys. Rev. 51, 735-738 (1937).

2. M. S. Freedman et al., Rev. Sci. Instr. 27, 716-719 (1956).

3. R. N. Bracewell and J. A. Roberts, Australian J. Phys. 7, 615-640 (1954).

4. J. O. Porteus and L. G. Parratt, An Optimized Method for Correcting for Smearing Aberrations (Cornell University Report AFOSR TN-59-754, September 1959).

5. F. D. Kahn, Proc. Cambridge Phil. Soc. 51, 519-525 (1955).

6. R. O. Lane, N. F. Morehouse, Jr., and D. L. Phillips, Nucl. Instr. Methods 9, 87-91 (1960).

7. G. E. Owens and H. Primakoff, Phys. Rev. 74, 1406-1412 (1948).

8. C. P. Flynn and E. F. W. Seymour, J. Sci. Instr. 39, 352-356 (1962).

9. D. Middleton, An Introduction to Statistical Communication Theory (McGraw-Hill Book Co., New York, 1960).

10. R. E. Kalman, "New Methods and Results in Linear Prediction and Filtering Theory," Proceedings of the Symposium of Engineering Applications of Random Function Theory and Probability, Purdue University, November 1960 (John Wiley, New York, 1963).

11. N. Starfelt and W. H. Kock, Phys. Rev. 102, 1598-1612 (1956).

12. K. Linden and N. Starfelt, Arkiv Fysik 7, 427-457 (1954).

13. W. R. Dixon and J. H. Aitken, Can. J. Phys. 36, 1624-1633 (1958).

14. L. Salmon, Analysis of Gamma-Ray Scintillation Spectra by the Method of Least Squares (United Kingdom Atomic Energy Authority Research Group Report AERE-R-3640, April 1961).

15. A. J. Ferguson, A Program for the Analysis of Gamma-Ray Scintillation Spectra Using the Method of Least Squares (Atomic Energy of Canada, Limited, Report CRP-1055, November 1961).

16. J. I. Trombka, Least-Squares Analysis of Gamma-Ray Pulse-Height Spectra (California Institute of Technology Report JPL-32-373, December 1962).

17. J. L. Holloway, Jr., "Smoothing and Filtering of Time Series and Space Fields," in Advances in Geophysics (edited by H. E. Landsberg and J. V. Mieghem, Academic Press, New York, 1958).

18. J. Waser and V. Schomaker, Rev. Mod. Phys. 25, 671-690 (1953).

19. A. I. Kitaigorodskii, The Theory of Crystal Structure Analysis (Consultants Bureau, New York, 1961).

20. L. Silberstein, Phil. Mag. 15, 379 (1933).

21. L. Fox and E. T. Goodwin, Phil. Trans. Roy. Soc. (London) A245, 501-534 (1953).

22. F. B. Hildebrand, Methods of Applied Mathematics (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1953).

23. H. J. Longley, Numerical Solutions and Applications of the Fold Integral (Los Alamos Scientific Laboratory Report LA-2729, October 1962).

24. S. Twomey, J. Assoc. Computing Machinery 10(1), 97-101 (1963).

25. D. L. Phillips, J. Assoc. Computing Machinery 9(1), 84-97 (1962).

26. M. Verde, Nucl. Phys. 9, 255-266 (1958/1959).

27. W. V. Lovitt, Linear Integral Equations (Dover Publications, New York, 1950).

28. L. D. Skarsgard, H. E. Johns, and L. E. S. Green, Radiation Res. 14, No. 3 (1961).

29. H. I. West, Jr., and B. Johnston, IRE (Inst. Radio Engrs.) Trans. Nucl. Sci. NS-7, 111-115 (1960).

30. N. E. Scofield, A Technique for Unfolding Gamma-Ray Scintillation Spectrometer Pulse-Height Distributions (U.S. Naval Radiological Defense Laboratory Technical Report USNRDL-TR-447, June 1960).

31. L. Katz and A. G. W. Cameron, Can. J. Phys. 29, 518-544 (1951).

32. J. C. Villforth et al., Comparison of Theoretical and Experimental Filtered X-Ray Spectra (Oak Ridge National Laboratory Report ORNL-2529, July 1958).

33. E. G. Rawson and D. V. Cormack, Nucleonics 16(10), 92-97 (1958).

34. W. B. Strickfaden and R. M. Kloepper, IBM 704 Programs for Unfolding Complex Gamma-Ray Spectra (Los Alamos Scientific Laboratory Report LA-2461, March 1961).

35. P. A. Read and F. M. Tomnovec, Computer-Aided Analysis of Gamma-Ray Spectrometer Data (U.S. Naval Radiological Defense Laboratory Report USNRDL-TR-592, October 1962).

36. J. F. Mollenauer, A Computer Analysis for Complex Sodium Iodide Gamma Spectra (University of California Radiation Laboratory Report UCRL-9748, August 1961).

37. A. S. Eddington, Monthly Notices Roy. Astron. Soc. 73, 359-360 (1913).

38. R. J. Trumpler and H. F. Weaver, Statistical Astronomy (University of California Press, Berkeley and Los Angeles, 1953).

39. T. W. Anderson, An Introduction to Multivariate Statistical Analysis (John Wiley, New York, 1958).

40. H. Scheffe, The Analysis of Variance (John Wiley, New York, 1959).

41. W. Feller, An Introduction to Probability Theory and Its Applications, Vol 1 (John Wiley, New York, 1950).

42. D. G. Gardner, "Analysis of Multicomponent Decay Curves by Use of Fourier Transforms," Proceedings of the Symposium on Applications of Computers to Nuclear Radiochemistry, October 17-19, 1962, Gatlinburg, Tenn. (edited by G. D. O'Kelley, National Academy of Sciences Report No. NAS-NS-3107, 1963).

43. A. S. Penfold and J. E. Leiss, Phys. Rev. 114(5), 1332-1337 (1959).

44. P. Wolfe, Recent Developments in Nonlinear Programming (Rand Corporation Report R-401-PR, May 1962).

45. P. Wolfe, Econometrica 27(No. 3), 382-398 (1959).

46. S. I. Gass, Linear Programming (Methods and Applications)(McGraw-Hill Book Co., New York, 1958).

47. S. Vajda, Mathematical Programming (Addison-Wesley, Reading, Mass., 1961).

48. W. W. Gavin, Introduction to Linear Programming (McGraw-Hill Book Co., New York, 1960).

49. S. I. Gass, "Recent Developments in Linear Programming," pp 296-377 in Advances in Computers (edited by F. L. Alt, Academic Press, New York, 1961).

50. S. Dreyfus and M. Freimer, Appendix II of Applied Dynamic Programming, R. E. Bellman and S. E. Dreyfus (Princeton University Press, Princeton, 1962).

51. P. Wolfe, Quart. Appl. Math. 19, 239-244 (October 1961).

52. D. Bogert and W. R. Burrus, "Comparison of Least-Squares Algorithms," pp 173-176 in Neutron Physics Division Annual Progress Report for Period Ending August 1, 1963 (Oak Ridge National Laboratory Report ORNL-3499, Vol. 1, October 1963).

53. R. S. Varga, Iterative Analysis (Prentice-Hall, Inc., Englewood Cliffs, N. J., 1962).

54. J. D. Young, Linear Programming Applied to Linear Differential Equations (University of California Radiation Laboratory Report UCRL-9498, January 1961).

55. M. Foster, J. Soc. Indust. Appl. Math. 9(3), 387-392 (1961).

56. I. I. Hirscham, Jr., Bull. Am. Math Soc. 66, 40-42 (1960).

ORNL-3743
UC-32 — Mathematics and Computers
TID-4500 (40th ed.)

INTERNAL DISTRIBUTION

1. Biology Library
2-4. Central Research Library
5. Reactor Division Library
6-7. ORNL — Y-12 Technical Library Document Reference Section
8-172. Laboratory Records Department
173. Laboratory Records, ORNL R.C.
174. J. A. Auxier
175. S. J. Ball
176. C. D. Baumann
177. R. E. Biggers
178. T. V. Blalock
179. E. P. Blizard
180. W. E. Browning
181. R. Burrus
182-216. W. R. Burrus
217. W. R. Busing
218. R. R. Coveyou
219. G. Czjzek
220. J. W. T. Dabbs
221. G. deSaussure
222. J. K. Dickens
223. F. F. Dyer
224. R. L. Ferguson
225. B. R. Fish
226. R. M. Freestone
227. D. A. Gardiner
228. W. A. Gibson
229. M. A. Halbert
230. J. A. Harvey
231. F. F. Haywood
232. W. H. Jordan
233. W. E. Kinney
234. C. E. Larson
235-236. M. Leimdorfer
237. H. G. MacPherson
238. F. C. Maienschein
239. B. F. Maskewitz
240. C. E. Miller, Jr.
241. W. F. Mruk
242. E. A. Nephew
243. C. W. Nestor, Jr.
244. Charles Nowlin
245. Jerry Olson
246. C. A. Preskitt
247. R. H. Ritchie
248. R. T. Santoro
249. M. J. Skinner
250. G. G. Slaughter
251. W. S. Snyder
252. J. H. Thorngate
253. V. R. R. Uppuluri
254. D. R. Vondy
255. J. W. Wachter
256. A. M. Weinberg
257. Robin Williams
258. W. Zobel
259. G. Dessauer (consultant)
260. M. L. Goldberger (consultant)
261. R. F. Taschek (consultant)

EXTERNAL DISTRIBUTION

262. Research and Development Division, AEC, ORO
263-831. Given distribution as shown in TID-4500 (40th ed.) under Mathematics and Computers category (75 copies —CFSTI)