JOURNAL OF COMPUTATIONAL PHYSICS 24, 416–444 (1977)

Exponential-Sum Fitting of Radiative Transmission Functions

W. J. WISCOMBE

National Center for Atmospheric Research, Boulder, Colorado 80303*

AND

J. W. EVANS

University of California, San Diego, California 92037

Received September 14, 1976; revised November, 1976.

Fitting transmission functions with exponential sums is the basis for a widely used approximation for calculating spectrally integrated radiative fluxes in planetary atmospheres, especially when both line absorption and scattering are important. The error in this method depends crucially on the accuracy of the fits, but unfortunately exponential-sum fitting is a classical ill-conditioned problem of numerical analysis. Previous techniques devised for exponential-sum fitting are often unsatisfactory in this application. We present a new method which sidesteps the ill conditioning, guarantees convergence to the unique best least-squares fit, gives positive coefficients, and produces fits orders of magnitude more accurate than any which have so far been published. The method is demonstrated to be capable of recovering an exponential sum, given data sampled from that sum and rounded to as few as two decimal places. Sample fits are given for the Goody and Malkmus random band models and the Yamamoto H_2O solar absorption data in order to illustrate the high accuracy of the method. The effect on the fits of errors in the transmission data is examined in some depth.

1. INTRODUCTION

In the past decade there have been tremendous advances in our computational tools for solving *monochromatic* radiative transfer problems in plane-parallel horizontally-homogeneous atmospheres with vertically inhomogeneous distributions of molecular absorbers, clouds, and aerosols. But in order to account for absorption line structure hundreds of thousands of such monochromatic solutions would have to be summed to obtain spectrally integrated solar and infrared fluxes. Even were a vast computing resource available for such a line-by-line calculation, the enterprise would be of dubious value owing to our ignorance of exact line strengths and line shapes. Therefore approximate methods of integrating the radiative flux across a spectral region containing many absorption lines must be sought.

One method for bypassing the line-by-line integration, the exponential-sum fitting

* The National Center for Atmospheric Research is sponsored by the National Science Foundation.

ISSN 0021-9991

of transmissions (ESFT) method, has been widely exploited since Hunt and Grant [1] applied it to an infrared absorbing-scattering problem involving cirrus cloud. The basis of the ESFT method is that the transmission function $T(u)$ for a given spectral interval is fit by a sum of exponentials $E(u)$, viz.

$$T(u) \cong E(u) \equiv \sum_{i=1}^{M} a_i e^{-k_i u}. \tag{1}$$

Then the fluxes from M monochromatic probems with absorption coefficients corresponding to k_i are summed up with weights a_i to give the spectrally integrated flux. Because of the physical interpretations placed on a_i and k_i, we must require $a_i > 0$ and $k_i \geqslant 0$. Subsequently the ESFT method was applied to clear atmospheres by Wiscombe and Freeman [2], Arking and Grossman [3], Lacis and Hansen [4], and Raschke *et al.* [5]; to aerosol-laden atmospheres by Sargent [6], Sargent and Beckman [7], Liou and Sasamori [8], and Pollack *et al.* [9]; and to cloudy atmospheres by Yamamoto *et al.* [10], Wiscombe [11], Lacis and Hansen [4], and Raschke *et al.* [5]. Only Wiscombe and Freeman [2], Wiscombe [11], and Pollack *et al.* [9] applied the method across the complete solar and infrared spectrums; the other applications were for either the solar or infrared separately.

While Hunt and Grant were the first to apply the ESFT method to a complex line absorption-scattering problem, the basic idea behind the method considerably antedates their work. It can be traced back at least to Chandrasekhar [12], who considered a two-absorption-coefficient (two-term exponential fit) model as the next step beyond the gray-gas assumption for a stellar atmosphere. Hottel used the idea in the 1930s to study radiant heat transfer in furnaces [13]. It was used in neutron transport theory under the name "picket fence model" [14], presumably because the histogram for the absorption coefficient distribution resembles a picket fence. Sargent [6] reviews the use of the idea under the name "sum-of-gray-gases approximation" in heat transfer theory. Kondrat'yev [15] also reviews some atmospheric applications of the ESFT idea dating back as far as 1939; some of these were simply in the nature of preband-model fits to transmission, while others were genuine computations of clear-sky infrared fluxes in the spirit of the ESFT method. It should be emphasized, however, that none of these earlier applications treated scattering in any sort of credible way, if at all. The really powerful techniques needed to handle multiple scattering (including reliable Mie scattering calculations) have only become available in the last 10 to 15 years.

Unfortunately for the applications (not only in radiative transfer but also in fields as diverse as biology, electrical engineering, and nuclear physics) the exponential-sum fitting problem is a classically ill-conditioned one [16]. It is customary to blame this ill conditioning on the nonorthogonality of the set of exponential functions, however our analysis below will illuminate the nature of this ill conditioning more precisely. A number of often unsatisfactory approaches have been made to exponential-sum fitting. These approaches fall into five broad categories, which are in order of increasing difficulty (and increasing computation time): (1) successive subtraction; (2) selecting

the k_i a priori; (3) Prony's method and variants; (4) inverse Laplace transforms; and (5) nonlinear least-squares techniques. We briefly discuss each of these in turn.

Successive subtraction is also known as the "graphical" technique because, when $T(u)$ is plotted on semilogarithmic graph paper, its asymptotic behavior for large u is often linear. This asymptotic line represents the component $a_1 e^{-k_1 u}$ with the smallest exponent k_1. If we then plot $T(u) - a_1 e^{-k_1 u}$ and repeat this process, we can "peel off" successive terms in the exponential sum, one by one. In practice, this method, whether implemented graphically or on a computer, tends to break down after two or three terms are recovered and generally is incapable of fitting to better than about 1%. The computer implementation of successive subtraction often produces negative coefficients a_i and/or negative exponents k_i when applied to transmission functions. Avrett and Hummer [17] devised a variant, giving at best 1% error, which seems to prevent these difficulties but which (a) gives a number of terms equal to the number of data points being fitted, and (b) requires the derivative of the function being fitted. The latter requirement precludes the direct use of measured transmission data, for which the derivative can seldom be estimated with any accuracy; however, band models can still be used. Liou and Sasamori [8] used Avrett and Hummer's algorithm with band models for their solar-spectrum aerosol study.

C. D. Rodgers (private communication) obtained the four-term and ten-term exponential-sum fits used by Hunt and Grant [1] by preselecting the k_i, based on knowledge of the range of line strengths, then obtaining the a_i from a standard linear least-squares fit. By fixing both the k_i and the number of terms, this approach loses the flexibility of the other methods, but in return it gains a certain freedom from the difficulties which plague them. However, the only way to increase the accuracy of this method is to increase the number of terms, and thus to achieve a given level of accuracy it may force many more monochromatic problems to be solved than would be required by an optimal fit.

Prony's method [16, p. 272; 18, p. 378] was put forward in 1795 and has been rather thoroughly analyzed since. Hudson's [19] method is essentially a variant. Prony's method first finds the k_i as roots of a polynomial, then the a_i from a linear least-squares solution. Large variations are produced in the a_i and k_i by small variations in the function being fitted, although this feature is found in all the methods (including ours) and is an intrinsic property of the problem rather than of Prony's method. Defects peculiar to Prony's method are that it often produces negative or complex k_i, negative a_i, and $\sum a_i \neq 1$. It requires preselection of the number M of terms, and for $M > 3$ the aforementioned difficulties almost invariably occur. Prony's method fits data sampled from $T(u)$ at equally spaced values of u, and so the portion of the $T(u)$ curve for small absorber amounts, where $T(u)$ plummets rapidly, is undersampled. Lanczos points out that even separating three exponential components with Prony's method often presents insurmountable difficulties, and that the method can fail even when the number of terms is known and good initial approximations for the k_i are available. Raschke and Stucke [20] adapted Prony's method to ESFT by introducing several empirical steps to force $a_i > 0$, $k_i \geqslant 0$, and $\sum a_i = 1$. They obtained fits of from two to six terms, with accuracy on the order of 1%, to a wide variety of H_2O,

CO_2, and O_3 transmission functions. Osborne [21] has also generalized Prony's method and overcome some of its numerical difficulties, but still cannot guarantee real k_i (much less $k_i \geqslant 0$).

The inverse Laplace transform or "k-distribution" technique relies on the observation that if we can determine the inverse Laplace transform of $T(u)$,

$$f(k) = \mathscr{L}^{-1}[T(u)]$$

then the inverse relationship

$$T(u) = \mathscr{L}[f(k)] = \int_0^\infty f(k)\, e^{-ku}\, dk \tag{2}$$

yields an exponential fit for $T(u)$ when the integral is approximated by a finite sum (a quadrature formula). This method is superficially attractive for several reasons: it guarantees $k_i \geqslant 0$, and if $f(k) \geqslant 0$ as one would expect on physical grounds, $a_i \geqslant 0$ as well. Furthermore the accuracy of the fit may be adjusted by taking more or less terms in the quadrature formula. But these nice features are predicated on the accurate computation of $f(k)$. The operation $\mathscr{L}^{-1}$ is ill conditioned numerically [22], and so, if $T(u)$ is known only in the form of data, $f(k)$ is recoverable at best with only poor accuracy. If $T(u)$ is given analytically, on the other hand, $f(k)$ may be determinable in closed form, thus bypassing the ill conditioning. Arking and Grossman [3] found $f(k)$ in closed form for several regular band models, including Elsasser's; interestingly enough, they did this not by taking $\mathscr{L}^{-1}[T(u)]$ but by regarding $f(k)$ as the distribution function for the band model absorption coefficient. Domoto [23] has commented on this dual interpretation of $f(k)$ and has furthermore given $f(k)$ in closed form for the Malkmus band model (apparently a closed form does not exist for the Goody model). Even when $f(k)$ is known in closed form, however, its behavior for large and small k often causes standard quadrature formulas to yield very poor exponential fits from Eq. (2) even if many terms are taken (this point is demonstrated in Section 3(d)).

Another variant of the transform method rests on the observation that the inverse Laplace transform of a sum of exponentials is a sum of δ-functions. Thus applying $\mathscr{L}^{-1}$ to a function which is close to a sum of exponentials should yield peaks at each k_i with height proportional to a_i. Gardner *et al.* [24], Gardner [25], Brownell and Callahan [26], and Papoulis [27] have all developed variations on this idea. The difficulties here are very similar to the well-known ones in the Fourier analysis of time series: spurious peaks, loss of peaks in the background noise, merging of peaks when the corresponding k_i are not well separated, etc. The peak-merging problem is especially severe for transmission functions—the random model $f(k)$'s shown by Arking and Grossman and by Domoto have only one or two broad peaks, while for regular models $f(k)$ is U-shaped and therefore has no peaks at all. For a portion of the CO_2 15μ band, Arking and Grossman show an $f(k)$ with hardly any discernible peak structure whatsoever. Thus Gardner-type methods are hopelessly ill suited to ESFT.

By far the most popular methods for exponential-sum fitting are nonlinear least-squares (NLLS) techniques, which are often available as standard subroutines in computer libraries (e.g. [28]). The reason for this popularity is that NLLS methods are very general and very sophisticated, and they guarantee improvement in the fit (as measured by the least-squares residual R) at each step. All such methods search the 2M-dimensional surface

$$R(a_1,\ldots,a_M,k_1,\ldots,k_M) = \sum_{n=0}^{N} w_n\left[T(u_n) - \sum_{i=1}^{M} a_i e^{-k_i u_n}\right]^2$$

(where the w_n are user-chosen weights) in order to find a minimum. It is beyond our scope to review the many applications of NLLS methods to exponential-sum fitting, but it is interesting to note that the fitting of even simple two- or three-term exponential sums is regarded as one of the severest tests of any new NLLS technique (see, for example, [29]). Box [30] gives an illustration of the R-surface for a very simple two-term exponential sum; it has a narrow, steep-sided, curved, and highly asymmetric valley. Since NLLS methods assume the R-surface becomes locally quadratic, in order to converge faster than the impossibly slow steepest descents method, it is not surprising that the decidedly nonquadratic R-surfaces associated with exponential sums confound them. NLLS techniques require a good initial guess for the a_i and k_i, which must be obtained from one of the flawed simpler methods above, or else convergence will be either nonexistent or to an unacceptably poor fit. Even with a good initial guess, it has been the authors' experience that NLLS methods creep with agonizing slowness, often stopping and claiming convergence when they are still far from the correct minimum [cf. Section 3(b)]. NLLS methods must furthermore be modified in some way to constrain $a_i > 0$, $k_i \geqslant 0$, and $\sum a_i = 1$ (for example by transformation of variables, cf. [30]). Sargent [6] and Sargent and Beckman [7] obtained two- and three-term NLLS fits with typical errors exceeding 1 %. Lacis and Hansen [4] and Pollack *et al.* [9] report NLLS fits with accuracy on the order of 0.1 %.

A natural question to ask, given the success of suitably modified NLLS methods in fitting transmission functions to measurement accuracy, is: Why pursue the subject further? There are several answers to this question. First let us note that, in fitting to 0.1 %, NLLS methods are generally working to their uttermost limit of accuracy; they will be blind to further improvements in measurement accuracy, which are sorely needed to reduce uncertainties in present calculations. Second, Raschke *et al.* [5] and Zdunkowski (private communication) have noted that insufficiently accurate fits produce spurious oscillations in vertical heating rate profiles computed by the ESFT method. This is because heating rates are related to the *derivative* of $T(u)$, which is usually fitted at least an order of magnitude worse than $T(u)$ itself. Even 0.1 % fits may produce serious errors in heating rate. Third, since the error in the ESFT method is partially dependent on the accuracy of the fits, one would like to have the best fit possible; knowing this best fit, one may then degrade its accuracy in steps to study the impact on computed fluxes and heating rates.

Previous fitting techniques are hard pressed to find a fit at all, much less the best fit, and are ill suited to systematic accuracy-degradation studies. Finally, from a numerical analysis standpoint, none of the previous techniques really faces squarely the essential difficulties of exponential fitting; they all succeed only to the extent that a patchwork of ad hoc fixes is successful in staving off disaster.

Below we present a fast numerical method for ESFT which we believe to be superior to all previous methods. It is based on some little-known work of Cantor and Evans [31] and avoids all the problems which plague the other methods, excepting only the requirement for equispaced data (which causes $T(u)$ to be poorly fitted for very small u, as in Prony's method). Our method

(1) is well conditioned;

(2) has guaranteed convergence to the best fit;

(3) automatically selects some optimal number of terms (determined by data quality, roundoff error, and user-specified parameters);

(4) constrains $a_i > 0$, $k_i \geqslant 0$, and $\sum a_i = 1$, and

(5) produces fits whose accuracy is regulated primarily by the accuracy of the data.

Sample fits to the Goody and Malkmus band models and to Yamamoto's H_2O solar absorption data are given. The effect of rounded data on the fits is studied. Other valuable applications of the method, such as the recovery of radioactive decay rates and chemical rate constants, are also touched upon.

2. The Exponential-Sum Fitting Method

The fundamental theorems guaranteeing existence, uniqueness, and convergence for our method, without a detailed discussion of the ill conditioning or the numerical procedures required, are contained in a terse and formidably abstract paper by Cantor and Evans [31], herein-after referred to as C/E. The present paper completes the development of a numerical technique based upon the C/E theorems. The style of C/E has apparently been a barrier to its assimilation by physical scientists and even by numerical analysts, many of whom continue, with little success, to apply various nonlinear least-squares techniques to exponential fitting. (Our perusal of the standard citation indexes shows but one citation of C/E since its publication (by Raschke and Stucke [20], who however chose to modify Prony's method).) Therefore it seemed essential to recapitulate those results of C/E which bear directly on the numerical procedure, omitting proofs and striving for understanding at the expense of rigor and generality. In what follows, we shall intersperse relevant C/E results with the presentation, referring to specific sections or theorems of C/E for proofs. (Note that we specialize the C/E norm to Euclidean and the C/E set S to the interval [0, 1], and we omit mathematical niceties which add nothing to the construction of the algorithm.)

(a) *The problem.* The problem is to find an exponential-sum fit $E(u)$ to a function $T(u)$, as in Eq. (1), with $a_i > 0$ and $k_i \geqslant 0$. Picking a set $u_n = n\,\Delta u$ $(n = 0,..., N)$ of equally spaced arguments, we measure the "distance" between $E(u)$ and $T(u)$ by the least-squares residual

$$R_0 \equiv \sum_{n=0}^{N} w_n[T(u_n) - E(u_n)]^2 \tag{3}$$

where $w_n \geqslant 0$ are least-squares weights and $E(u_n)$ is, from Eq. (1), a sum of powers

$$E(u_n) = \sum_{i=1}^{M} a_i \theta_i{}^n \tag{4a}$$

where

$$\theta_i \equiv e^{-k_i \Delta u}. \tag{4b}$$

Clearly $\theta_i \in [0, 1]$ when $k_i \geqslant 0$.

The "best" fit we define as the one which minimizes R_0 over all allowable values of M, a_i, and k_i.

(b) *Existence and uniqueness.* In brief, a best fit, as defined above, exists and is unique (C/E, Sections 1 and 2). Uniqueness may be lost if the data $T(u_n)$ are *exactly* representable by a finite sum of exponentials, but this is very rare with finite-precision arithmetic. In the past there has been some confusion between stability and uniqueness in this problem. Certainly there is instability: if we vary the data $T(u_n)$ by small amounts, there may be large changes in the best fit. But for any particular set of data, the best fit is unique.

It is useful to visualize the fitting process in an $(N + 1)$-dimensional vector space, illustrated schematically in Fig. 1. The data vector $[T(u_0),..., T(u_N)]$ is a point $\bar{T}$ in that space and lies outside the convex subspace C_N formed by all possible exponential-sum

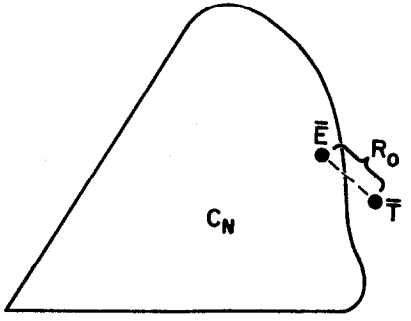

FIG. 1. Schematic illustration of exponential fitting. $\bar{T}$ is the data vector, C_N is the subspace of all exponential sums with positive coefficients, and $\bar{E}$ is the point on the boundary of C_N which is closest to $\bar{T}$. R_0 is the distance between $\bar{E}$ and $\bar{T}$.

vectors $[E(u_0),..., E(u_N)]$. Any point $\bar{E}$ in C_N which is closest to $\bar{T}$ (as measured by R_0, Eq. (3)) will obviously be on the boundary of C_N. That there is only one such closest point follows from the fact that C_N is convex, it "bulges outward" everywhere. Furthermore, points on the boundary of C_N correspond to unique $E(u)$'s (i.e., unique a_i, k_i, and M). Thus a closest point exists, is unique, and corresponds to a unique $E(u)$, which is our best fit.

(c) *The residual polynomial.* Presuming the θ_i in Eq. (4) to be known, the standard linear least-squares "normal" equations for $a_1,..., a_M$, are

$$\partial R_0/\partial a_i \equiv P(\theta_i) = 0 \qquad [i = 1,..., M] \tag{5}$$

where

$$P(\theta) \equiv 2 \sum_{n=0}^{N} p_n \theta^n \tag{6}$$

and

$$p_n \equiv w_n[E(u_n) - T(u_n)]. \tag{7}$$

The set of a_i which satisfy Eq. (5) clearly minimize R_0 for fixed θ_i. We call $P(\theta)$ the "residual polynomial" because its coefficients p_n are the (weighted) point-by-point differences between the fit and the data.

The residual polynomial is fundamental to our entire method because of the following theorem (C/E, Theorem 2.1): A best fit to the data (with $a_i > 0$ and $k_i \geqslant 0$) has been obtained if and only if the residual polynomial satisfies

$$\text{(a)} \quad P(\theta_i) = 0 \qquad \text{for} \quad i = 1,..., M;$$

$$\text{(b)} \quad P(\theta) \geqslant 0 \qquad \text{for} \quad \theta \in [0, 1].$$

These two conditions furnish a remarkably simple definition of a best fit (condition (a) is just Eqs. (5)). The essence of our method is that it iterates back and forth between solving condition (a) for the coefficients a_i and improving toward condition (b) by adding a new exponential factor θ_i. Complications arise because Eqs. (5) very quickly become ill conditioned, because their solution often yields some $a_i < 0$, and because a straightforward calculation of the residual R_0 is ill conditioned, but we have managed to deal effectively with each of these problems.

Note that conditions (a) and (b) taken together imply that each θ_i is a root of $P(\theta)$ of even multiplicity. In practice, the roots seem to be double since the numerical procedure tends to produce very close pairs of θ_i's. These close pairs of θ_i's are directly responsible for the ill conditioning we encounter in Eqs. (5), yet every attempt to coalesce these pairs during the iterative procedure and so eliminate the ill conditioning has caused the method to fail. It is apparently vital to the numerical procedure that the θ_i's be allowed to remain close without actually coalescing. Only

after convergence has it been found profitable to coalesce any such remaining pairs (see Section 2(i)).

(d) *Overview of method.* The logical flow of our method is illustrated schematically in Fig. 2; the parts labeled I through V are described in detail in Sections 2(e) through 2(i). There is a main iteration back and forth between the nonlinear part of the problem (I, finding a new θ_i) and the linear part (II, finding $a_1, \ldots, a_M$). As a by-product of the

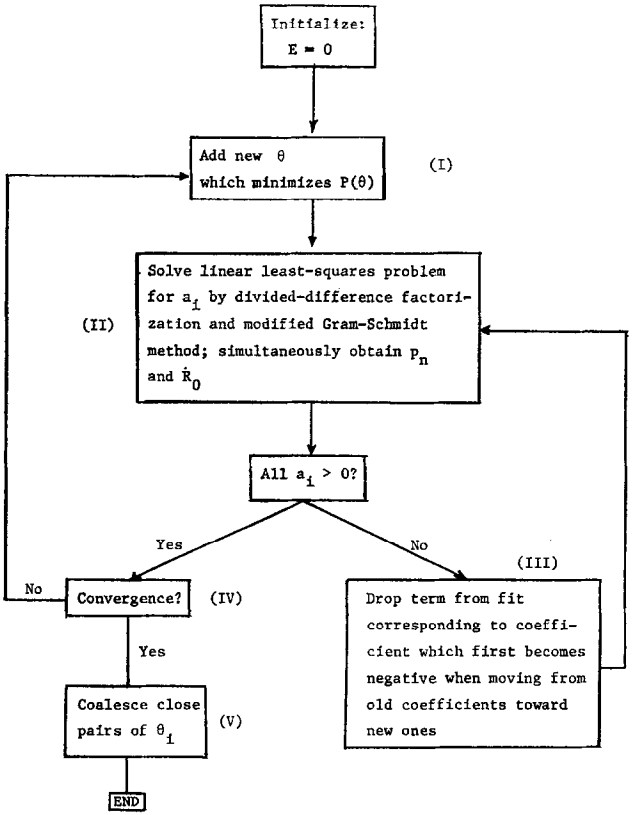

FIG. 2. Logical flow of exponential-sum fitting algorithm. Blocks labeled I–V are discussed in that order in the text.

linear part, we obtain the residual polynomial coefficients p_n (Eq. (7)) and from them the residual R_0. There is a minor iteration (III) to drop terms with negative a_i. If at any stage of the major iteration all a_i are positive and the convergence criteria (IV) are satisfied, then close pairs of θ_i are coalesced (V) and we are finished. The method is initialized by an identically-zero exponential sum, $E(u) \equiv 0$; therefore we experience none of the initialization difficulties faced by NLLS methods.

(e) *Nonlinear part* (I). Unless we have already achieved a best fit, $P(\theta)$ must be negative over at least part of the interval [0, 1]. Let $\theta = \theta_0$ be the point at which $P(\theta)$ is the most negative:

$$0 > P(\theta_0) = \min_{\theta \in [0,1]} P(\theta). \tag{8}$$

By adding θ_0 to the other θ_i's and then enforcing $P(\theta_0) = 0$ in the linear part, we have, intuitively speaking, effected the most improvement toward $P(\theta) \geqslant 0$ (condition (b) above). This procedure essentially "lifts up" the part of the residual polynomial which "droops" most below zero.

The choice for a new θ in Eq. (8) can be motivated more rigorously as follows. Suppose we have an approximant $E(u)$ (Eq. (4)) and we want to add a new term $a_0\theta_0{}^n$ in such a way that, as its coefficient a_0 increases from zero, the residual R_0 (Eq. (3)) decreases most rapidly. This is in the spirit of the steepest descents methods. The rate of change of R_0 with a_0 at $a_0 = 0$ is

$$\begin{aligned} \frac{\partial R_0}{\partial a_0}\bigg|_{a_0=0} &= \left[\frac{\partial}{\partial a_0} \sum_{n=0}^{N} w_n[E(u_n) + a_0\theta_0{}^n - T(u_n)]^2\right]_{a_0=0} \\ &= P(\theta_0) \end{aligned}$$

and making this rate of change as negative as possible leads again to Eq. (8) as the condition for picking θ_0.

The finding of the global minimum of $P(\theta)$, as required by Eq. (8), is straightforward but potentially time consuming, especially since the degree of $P(\theta)$ is equal to the number of data points N, and N may need to be large to properly resolve a transmission function. The examination of the roots of $P'(\theta)$ is not viable because polynomial root-finding algorithms are notoriously ill conditioned when some of the coefficients have considerably different magnitudes than others; this will certainly be the case for $P'(\theta)$, whose coefficients are proportional to the residuals. We first used a direct 1000-point search of [0, 1], with a refined mesh around each of the old θ_i in order to catch close pairs. This was perfectly satisfactory, but, since there are N multiplications in one evaluation of $P(\theta)$, roughly $1000N$ multiplications were required for a single search. Already at $N = 100$ this imposes a considerable burden, for the direct search must be performed perhaps 50 to 100 times before the major iteration loop in Fig. 2 converges satisfactorily.

In practice, we have found that most of the roots θ_i of $P(\theta)$ tend to cluster near $\theta = 1$, corresponding to a roughly equal spacing of $\ln\ln \theta_i$ (or $\ln k_i$). This, coupled with the fact that the minimum of $P(\theta)$ is often very near to one of its roots, makes direct-search minimization of $P(\theta)$ difficult in principle as well as computationally time consuming.

In order to circumvent the direct search, we devised a global minimization method with guaranteed convergence which employs quadratic bounds and interval elimination. This method requires only a fraction of the number of $P(\theta)$ evaluations as the direct search and at the same time yields a more accurate minimum. All the compu-

tations reported here used this improved method, which we are publishing separately. We emphasize, however, that such a sophisticated minimization procedure, while desirable from the viewpoint of computational speed, is in no way essential to the success of the fitting method.

(f) *Linear part* (II). Lawson and Hanson [32, p. 122], hereinafter abbreviated as L/H, emphasize that in practice one should never solve Eqs. (5) directly because the modern linear least-squares methods which they describe are much better behaved numerically. Once we have factored the ill conditioning out of Eqs. (5), they can be solved by a slight variant of one of these modern methods (the modified Gram–Schmidt).

Basically, the overdetermined linear system which we wish to solve for the a_i is

$$E(u_n) = \sum_{i=1}^{M} a_i \theta_i{}^n \cong T(u_n) \qquad (n = 0,..., N).$$

In matrix form, this is

$$Aa \cong b \tag{9a}$$

where

$$A = \begin{pmatrix} 1 & \cdots & 1 \\ \theta_1 & & \theta_M \\ \vdots & & \vdots \\ \theta_1{}^N & & \theta_M{}^N \end{pmatrix}, \qquad a = \begin{pmatrix} a_1 \\ \vdots \\ a_M \end{pmatrix}, \qquad b = \begin{pmatrix} T(u_0) \\ \vdots \\ T(u_N) \end{pmatrix}. \tag{9b}$$

It is easily verified that if $W \equiv \mathrm{diag}[w_i]$ is a diagonal matrix of the weights, then Eqs. (5) result from multiplying both sides of Eq. (9) by $A^T W$

$$Da = d \tag{10a}$$

where

$$D \equiv A^T W A, \qquad d \equiv A^T W b. \tag{10b}$$

Since we do not in fact solve Eq. (10), we shall introduce the weights by multiplying both sides of Eq. (9) by $W^{1/2}$, but it is computationally more economical to postpone this operation until after the ill conditioning is factored out of matrix A.

The ill conditioning of A arises when $\theta_i \to \theta_j$ for any i and j, for then two columns of A become identical and A has less than full rank. Matrix D (Eq. (10)) becomes singular when $\theta_i \to \theta_j$. Since close pairs of θ_i's are a fixture of our method, this ill conditioning is a serious and ever-present problem. To sidestep it, we postmultiply A by a matrix B carrying away all the ill conditioning and yielding a matrix A' which always has full rank,

$$A' = AB \tag{11}$$

B performs elementary column operations on A which resemble the Newton divided-difference technique.

As a simple example, consider a case with $N = 3$. Subtract column 1 of A (Eq. (9b)) from columns 2 and 3 (postmultiply by matrix operator E_1), then divide column 2 by $(\theta_2 - \theta_1)$ and column 3 by $(\theta_3 - \theta_1)$ (postmultiply by matrix operator D_1). Now subtract column 2 from column 3 (matrix operator E_2) and divide column 3 by $(\theta_3 - \theta_2)$ (matrix operator D_2), to yield

$$A' = AE_1D_1E_2D_2 = \begin{pmatrix} 1 & 0 & 0 \\ \theta_1 & 1 & 0 \\ \theta_1^2 & \theta_1 + \theta_2 & 1 \\ \theta_1^3 & \theta_1^2 + \theta_1\theta_2 + \theta_2^2 & \theta_1 + \theta_2 + \theta_3 \end{pmatrix}.$$

The last matrix has full rank even if $\theta_1 = \theta_2 = \theta_3$.

Generalizing this example, the transformation matrix B must be

$$B = \prod_{k=1}^{M-1} E_k D_k \tag{12}$$

where E_k subtracts column k from columns $k + 1$ through M and D_k divides column j by $(\theta_j - \theta_k)$ for $j = k + 1$ to M. It is elementary to show that

$$E_k = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & \bigcirc & & \\ & & 1 & -1 & \cdots & -1 & \\ & & & \ddots & & \bigcirc & \\ & \bigcirc & & & \ddots & & \\ & & & & & 1 & \end{bmatrix} k\text{th row}, \tag{13a}$$

$$D_k = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & \bigcirc & \\ & & 1 & & & & \\ & & & (\theta_{k+1} - \theta_k)^{-1} & & & \\ & \bigcirc & & & \ddots & & \\ & & & & & (\theta_M - \theta_k)^{-1} & \end{bmatrix} (k+1)\text{st row}. \tag{13b}$$

The matrix A' is complicated, as the example indicates, but it may be formed column-by-column with a simple recursion procedure. First set the diagonal elements of A' (to unity) and its first column equal to the first column of A. Then

$$A'_{ij} = \theta_j A'_{i-1,j} + A'_{i-1,j-1} \qquad (i = j + 1 \text{ to } N + 1; j = 2 \text{ to } M). \tag{14}$$

Using the above conditioning transformation (Eqs. (11) to (13)), the problem becomes

$$W^{1/2}A'a' \cong W^{1/2}b \tag{15}$$

where

$$a = Ba'. \tag{16}$$

The numerical least-squares solution of Eq. (15) for a' now presents no particular difficulties, and the reintroduction of the ill conditioning, through multiplication of a' by B to get a (Eq. (16)), is done explicitly and not by any numerical technique. Note that it is computationally most efficient to multiply a' by the factored form (12) of B. The recursion for so doing is

$$\left.\begin{aligned} a_j' &= a_j'/(\theta_j - \theta_{M-i}), \qquad j = M - i + 1 \text{ to } M \\ a'_{M-i} &= a'_{M-i} - \sum_{j=M-i+1}^{M} a_j', \end{aligned}\right\} \quad i = 1 \text{ to } M - 1.$$

Note that the vector a' continually overwrites itself during this recursion, and at the conclusion contains the desired vector a. The only use which is made of a in our method (until convergence) is to decide which terms to drop on account of negative components in a. For this purpose, no great accuracy in a is necessary. The most important output from solving Eq. (15) is not a, but the coefficients p_n (Eq. (7)) of the residual polynomial and the residual R_0 (Eq. (3)). One needs to know the p_n accurately in order that the nonlinear step be well behaved numerically, and one needs R_0 accurately in order to determine convergence.

To compute the p_n directly from their definition, Eq. (7), is hopelessly inaccurate because, the better the fit, the more significant digits are lost in subtracting $E(u_n)$ from $T(u_n)$; and furthermore $E(u_n)$ involves a, which is ill conditioned. But note that the vector of residuals may be written

$$r \equiv W^{1/2}(Aa - b) = W^{1/2}(A'a' - b) \tag{17}$$

and if r is computed from the second expression, involving A' and not A, its components r_n will not be subject to ill conditioning as $\theta_i \to \theta_j$. From these components we obtain

$$p_n = w_n^{1/2} r_n, \qquad R_0 = \sum_{n=0}^{N} r_n^2. \tag{18}$$

We have chosen the modified Gram–Schmidt (MGS) method [L/H, p. 129] for solving Eqs. (15) because it yields the residual vector, Eq. (17), as an automatic by-product of the computation; the more popular Householder technique (L/H, Chap. 11) requires us to obtain it from Eq. (17) as an additional step. L/H do not derive the

result, but it is a fairly simple matter to show that the $(M + 1)$st column of their matrix $\tilde{Q}$ [L/H, Eqs. (19.32–33)] is $-r$.

Two slight modifications of the MGS method conserve computation in our particular case. First, the matrix $W^{1/2}A'$ has a zero upper triangular part, and these zeros will be preserved through the Gram–Schmidt orthogonalization process (i.e., will occur in the same locations of $\tilde{Q}$) if we start at column M (the shortest one) and work backward. By explicitly noting these zeros in the formation of dot products of columns, computation is reduced. Second, as much information as possible should be saved in the MGS process (A', dot products, etc.) because whenever an exponential factor θ_i is dropped or added as described in Sections 2(e) and 2(g), all columns of A, and therefore of A', to the left of the column dropped or added are unchanged.

(g) *Term-dropping Step* (III). The linear step may, and often does, produce negative a_i (although by construction the coefficient of the term added in the nonlinear step will always be positive). This is a signal that the corresponding terms are of dubious desirability. We shall eliminate the "least desirable" of these terms by an adaptation of the L/H algorithm NNLS (p. 158). Our adaptation sacrifices NNLS's guaranteed satisfaction of the Kuhn–Tucker conditions, but gains in return a great increase in speed by initializing at the old a_i's (rather than zero) and by not allowing a term, once zeroed, to be unzeroed.

In order to understand our term-dropping procedure, certain basic facts from quadratic programming theory [33] are necessary. The residual (Eq. (3)) can be written (C/E Lemma 3.1)

$$R_0 = c_0 - 2d^T a + a^T D a \tag{19}$$

with c_0 a constant, and D, a, and d as in Eqs. (10). Since R_0 is a quadratic form, we may view the problem of finding $a \geqslant 0$ for fixed θ_i as a quadratic programming problem of a special sort, which is

$$\text{find} \qquad \min R_0(a) \qquad \text{with} \qquad a \geqslant 0. \tag{20}$$

The matrix D is positive definite, therefore the M-dimensional surface consisting of points $(a, R_0(a))$ is strictly convex, or "bowl-shaped," and has a unique global minimum. By keeping the bowl shape of the R_0-surface in mind, it will be easier to visualize how R_0 decreases even when we are dropping terms from the fit rather than adding them. In passing, we note that the ill conditioning as $\theta_i \to \theta_j$ is manifested by the R_0-surface becoming increasingly flat along a certain line through the minimum, making it difficult to distinguish the correct minimum with finite-precision arithmetic. This is because, if $\theta_i = \theta_j$, D becomes singular and there is a one-dimensional subspace of vectors a_S satisfying $Da_S = 0$; it then follows that $R_0(a) = R_0(a + a_S)$ so that the minimum can no longer be unique.

Our term-dropping iteration for solving Eq. (20) proceeds as follows. Suppose the

nonlinear step (Section 2(e)) has yielded a new θ which, for simplicity, is placed in the $(M + 1)$st position. The old coefficient vector,

$$a^{\text{old}} = (a_1^{\text{old}},..., a_M^{\text{old}}, 0), \tag{21}$$

is augmented by a zero in the $(M + 1)$st position to reflect this new term. The first M components of a^{old} are all positive on account of the successful completion of the prior term-dropping iteration (cf. Fig. 2). Suppose the new coefficient vector produced by Step II,

$$a^{\text{new}} = (a_1^{\text{new}},..., a_{M+1}^{\text{new}}),$$

has at least one negative entry (which will not be a_{M+1}^{new} because of the way θ_{M+1} was selected). As we move from a^{old} towards a^{new} along the straight line

$$a = (1 - \beta)\, a^{\text{old}} + \beta a^{\text{new}}, \qquad 0 \leqslant \beta \leqslant 1$$

we will cross one coordinate hyperplane for each negative entry in a^{new}. We stop when we encounter the first such hyperplane, $a_k \equiv 0$, where k is the value of i at which the minimum is assumed in

$$\beta_k = \min_{a_i^{\text{new}} < 0} \left(\frac{a_i^{\text{old}}}{a_i^{\text{old}} - a_i^{\text{new}}} \right).$$

Then we reset the "old" coefficient vector to

$$a^{\text{old}} = (1 - \beta_k)\, a^{\text{old}} + \beta_k a^{\text{new}} \tag{22}$$

which will contain all positive entries except for $a_k^{\text{old}} = 0$. Because we moved toward the minimum of the R_0-surface in going from the initial coefficients (Eq. (21)) toward the improved ones (Eq. (22)), the improved coefficients give a smaller value of R_0 even though a term has effectively been dropped from the fit.

One now returns to Step II, resolves the linear problem for a^{new} with the constraint $a_k^{\text{new}} = 0$ (i.e., with the kth column dropped from the matrix A) and repeats the procedure of the last paragraph if a^{new} contains any negative entries. One can view this process as restricting the search for the minimum to the intersection of the R_0-surface with the hyperplane $a_k = 0$. Once the search is so restricted, it is not allowed to become unrestricted again; that is, the kth term is not allowed to be

[redacted]

the exponential sum.

The Kuhn–Tucker conditions [33, p. 214] guarantee that the correct minimum in our quadratic programming problem (Eq. 20) has been found if

$$v \equiv Da - d \geqslant 0$$

and if v_i is only positive when a_i is zero, and vice versa ($a_i v_i = 0$ for all i). The procedure in this section never failed to meet these conditions in many thousands of test cases which were checked. However, the guaranteed convergence of our exponential fitting method does not depend on whether or not the Kuhn–Tucker conditions are satisfied; all that matters is that R_0 decreases on each pass through Step III.

(h) *Convergence* (IV). Suppose E_∞ is the best fit, and subscript j refers to the jth major iteration in Fig. 2. The value of the jth-step residual polynomial at the new θ, $P_j(\theta_0)$, furnishes the following lower bound on the decrease in the residual in one step (C/E Theorem 3.2):

$$(R_0)_j - (R_0)_{j+1} \geqslant \tfrac{1}{4}[P_j(\theta_0)]^2 \Big/ \sum_{n=0}^{N} w_n \theta_0^{2n}.$$

Clearly $P_j(\theta_0) \to 0$ as $j \to \infty$ since otherwise the residual would eventually become negative. $|P_j(\theta_0)|$ also bounds the departure of the jth residual from the best-fit residual (C/E Theorem 3.4):

$$0 < (R_0)_j - (R_0)_\infty \leqslant E_\infty(0) \mid P_j(\theta_0)|.$$

(In fitting transmission functions, $E_\infty(0) = 1$ for all practical purposes.) That E_j converges to E_∞, data point by data point, is ensured by (C/E Corollary 3.5)

$$\sum_{n=0}^{N} w_n[E_\infty(u_n) - E_j(u_n)]^2 \leqslant E_\infty(0) \mid P_j(\theta_0)|.$$

These inequalities prove the convergence of our method and demonstrate that the residual polynomial is a very sensitive indicator of the rate of that convergence.

Not only does $E_j \to E_\infty$ in the least-squares sense, but the actual coefficients a_i and exponential factors θ_i of E_j converge to those of E_∞. As $j \to \infty$, the exponential factors of E_j cluster in shrinkingly small neighborhoods around those of E_∞, and the sum of the coefficients of E_j in each such neighborhood approaches the corresponding coefficient of E_∞ (C/E Theorem 3.7). Extraneous exponential factors may occur in E_j which are not close to any in E_∞, but the coefficients of such extraneous terms become arbitrarily small as $j \to \infty$. In Section 3(a) we shall see examples of extraneous terms with small coefficients, which are not entirely eliminated because roundoff error precludes the passage to the limit $j \to \infty$.

After considerable experience with our method, we have developed the set of convergence criteria

$$\text{(i)} \quad \frac{(R_0)_{j-1} - (R_0)_j}{(R_0)_{j-1}} < \epsilon_1,$$

or

$$\text{(ii)} \quad P_j(\theta_0) \geqslant -\epsilon_2,$$

or

$$\text{(iii)} \quad |\theta_0 - \theta_i| < \epsilon_3 \qquad (i = 1 \text{ to } M),$$

or

$$\text{(iv)} \quad \text{number of major iterations} > I_{\max}.$$

The last criterion is merely to guard against computational runaway, and, with $I_{\max} = 150$, is rarely triggered. The method works successfully with $\epsilon_1 = \epsilon_2 = \epsilon_3 = 0$, in which case convergence is achieved only when the residual *increases* from the previous iteration (due to random fluctuation at the level of roundoff error) or when the nonlinear step picks a new exponential factor θ_0 equal to one of the old ones. The sample fits presented in Section 3 were generated with $\epsilon_1 = 10^{-20}$, $\epsilon_2 = 10^{-20}$, and $\epsilon_3 = 10^{-6}$. In general, making the convergence criteria less restrictive does not necessarily lead to fewer terms in the fit, since it may prevent pairs of θ_i's from coming close enough together to be coalesced.

(i) *Coalescence* (V). Suppose two exponential factors θ_k and θ_l are close, in a sense to be defined below. Then the kth and lth terms are replaced by a single term in order to shorten the fit. Some accuracy may be lost thereby, but this is usually quite tolerable.

If we define the data to be fitted as

$$\gamma_n \equiv T(u_n) - \sum_{\substack{i=1 \\ i \neq k,l}}^{M} a_i \theta_i{}^n \qquad (n = 0,\ldots, N)$$

then this data is best fitted by $a_k\theta_k{}^n + a_l\theta_l{}^n$. We choose instead to fit it by a single term $a\theta^n$, where a and θ are chosen to minimize

$$R_0 = \sum_{n=0}^{N} w_n(\gamma_n - a\theta^n)^2 \tag{23}$$

which leads to the pair of equations

$$\tfrac{1}{2}(\partial R_0/\partial a) = f(a, \theta) = \sum_{n=0}^{N} w_n(a\theta^n - \gamma_n)\,\theta^n = 0,$$

$$\tfrac{1}{2}(\partial R_0/\partial \theta) = g(a, \theta) = \sum_{n=1}^{N} w_n(a\theta^n - \gamma_n)\,na\theta^{n-1} = 0.$$

We solve these equations by a Newton's method iteration

$$a^{(i+1)} = a^{(i)} + \left(g\frac{\partial f}{\partial \theta} - f\frac{\partial g}{\partial \theta}\right)\Big/\left(\frac{\partial f}{\partial a}\frac{\partial g}{\partial \theta} - \frac{\partial f}{\partial \theta}\frac{\partial g}{\partial a}\right),$$

$$\theta^{(i+1)} = \theta^{(i)} + \left(f\frac{\partial g}{\partial a} - g\frac{\partial f}{\partial a}\right)\Big/\left(\frac{\partial f}{\partial a}\frac{\partial g}{\partial \theta} - \frac{\partial f}{\partial \theta}\frac{\partial g}{\partial a}\right),$$

where f, g, etc., are all evaluated at $a = a^{(i)}$ and $\theta = \theta^{(i)}$, and where the initial guess is

$$a^{(0)} = a_k + a_l,$$
$$\theta^{(0)} = \tfrac{1}{2}(\theta_k + \theta_l).$$

The convergence criterion is the first one from the previous section, as applied to the residual R_0 in Eq. (23). No more than six iterations are usually required, even with $\epsilon_1 = 10^{-20}$, due to the excellence of the initial guess. The final value of θ almost always lies between θ_k and θ_l, but in rare cases this is not so.

The procedure of the last paragraph is applied first to the closest pair, then the next closest pair, and so forth, until none of the remaining values of θ are "close." Any term resulting from coalescence should be barred from further coalescence.

A variety of different closeness criteria have been tried. Those based on the relative or absolute differences of θ_i's all fail because the θ_i's tend to cluster near $\theta = 1$. The relative difference criteria, if adjusted to cause the correct amount of coalescence near $\theta = 1$, do not cause enough near $\theta = 0$; the absolute criteria, if similarly adjusted, cause too much coalescence near $\theta = 0$. Based on our observation that $\ln k_i$ tends to be uniformly distributed once pairs are properly coalesced, the most sophisticated procedure would be to establish "major" and "minor" intervals of $\ln k_i$ by search, then coalesce all pairs separated only by a "minor" interval. But in practice we have been successful with coalescing terms whose values of k_i differ by less than 5 to 25%. The larger the value of this percentage, the more coalescence will be done. To illustrate this process, we show in Fig. 3 fits to the Goody [34] random band model with a line spacing to halfwidth ratio of 20. The 225 transmission data, including values from 0.001 to 1, were all rounded to eight decimal places. Each term in the fitted exponential sum is represented by a single spike, whose foot is located at the appropriate value of k_i and whose height represents a_i. Plots are shown for coalescence

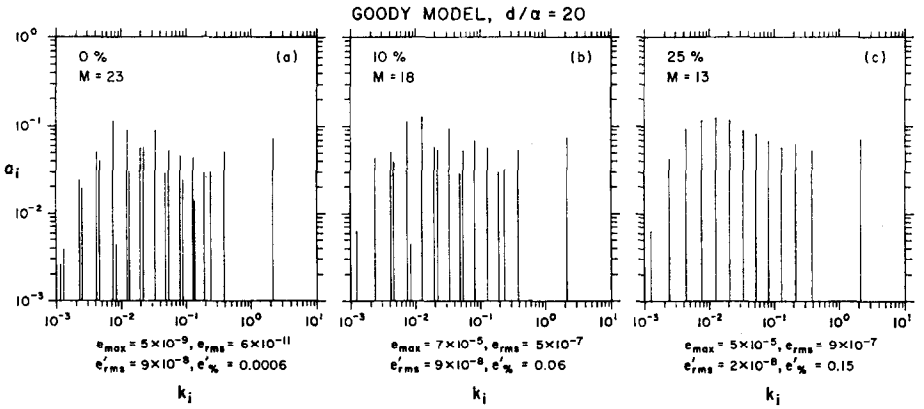

FIG. 3. Spike plots for exponential-sum fits to the Goddy random band model, with a mean line spacing (d) to mean halfwidth (α) ratio of 20, and for coalescence criteria of (a) 0%, (b) 10%, (c) 25%.

criteria of 0, 10, and 25%. When no coalescence is done, we see a 23-term fit in which all but three terms exist closely paired with another term. For the 10% case, the five closest pairs have been eliminated. For the 25% case, all 10 pairs have been eliminated and the spike plot has a much more uniform and pleasing appearance than for the 0 and 10% cases. One advantage of these spike plots is that close pairs are easily spotted; another is that one can observe how smooth the distribution of a_i and k_i becomes when pairs are eliminated. The smoother plots give a much better feeling for the distribution of absorbing power within a spectral interval than do the cluttered plots resulting from insufficient coalescence. (Note that the spike envelope appears to be double peaked in qualitative agreement with the k-distribution in Arking and Grossman's [3] Fig. 2 for the Goody model.)

The errors noted beneath each plot in Fig. 3 indicate the penalty one pays for coalescence. The quantities $e_{\max}$ and e_{rms} are the maximum absolute error and r.m.s. error in the fit $E(u)$; e'_{rms} is the r.m.s. error and $e'_{\%}$ the average percent error in $E'(u)$, excluding the data point $u_0 = 0$ at which errors in E' can be arbitrarily large (see Section 2(j)). These quantities are defined formally as

$$e_{\max} = \max_{0 \leqslant n \leqslant N} | T(u_n) - E(u_n)|, \tag{24a}$$

$$e_{\text{rms}} = (1/(N+1)) \left\{ \sum_{n=0}^{N} [T(u_n) - E(u_n)]^2 \right\}^{1/2}, \tag{24b}$$

$$e'_{\text{rms}} = (1/N) \left\{ \sum_{n=1}^{N} [T'(u_n) - E'(u_n)]^2 \right\}^{1/2}, \tag{24c}$$

$$e'_{\%} = (100/N) \sum_{n=1}^{N} | T'(u_n) - E'(u_n)|/| T'(u_n)|. \tag{24d}$$

When no coalescence is done (Fig. 3(a)) the error is never any worse than the rounding error (5×10^{-9}) in the data. This is typical when fitting to infinitely differentiable analytic functions such as band models. In going to a 10% coalescence criterion (Fig. 3(b)) both $e_{\max}$ and e_{rms} increase by four orders of magnitude but remain less than the error in transmission measurements. The fit error becomes no worse if the remaining pairs are coalesced by going up to a 25% criterion (Fig. 3(c)). This pattern is typical, although the details vary widely; the first few coalescences increase the error dramatically, but after that there is little penalty for further coalescence. Hence one might as well use a liberal criterion like 25% if one chooses to coalesce at all.

Coalescence is not as important when fitting real data as when fitting analytic functions, for in the former case many fewer terms tend to be produced, and therefore there are many fewer close pairs. For analytic functions, rounding the data before fitting has generally been found to be a more satisfactory way of reducing the number of terms, rather than relying exclusively on coalescence. A combination of rounding and coalescence is usually the most effective.

(j) *Practical operation.* The fitting procedure operates in two distinct stages. For the first 10 to 20% of the iterations, it adds terms, seldom dropping any; for the remainder of the iterations it usually drops one term each time it adds one, so that the number of terms remains constant. Occasionally (less than 5% of the time) two terms are dropped; almost never are more than two terms dropped. In the initial stage the θ_i remain fairly well separated, but in the later stages of the iteration close pairs of θ_i are very much in evidence, causing the linear problem to be extremely ill conditioned.

We do not constrain $E(0) = \sum a_i = 1$ because we inevitably find this relationship satisfied to many significant digits (at least four or five). If this were ever unsatisfactory, it would be possible to set the weight $w_0 \gg 1$ to force $\sum a_i = 1$ to greater precision, but the use of a very large weight may cause numerical difficulties (L/H, p. 148). The reason that $\sum a_i = 1$ is well satisfied is related to the fact that one of the θ's can, if necessary, assume a value so close to zero that the corresponding term in $E(u)$ is negligible for $u = u_1, \ldots, u_N$. At $u = u_0 = 0$ this term has a value equal to its coefficient, and this coefficient then plays the role of an adjustable parameter which can force $E(0) = 1$.

Since $\theta = 0$ corresponds to an infinite value of k, which is not permissible, we always restrict $\theta \geqslant \theta_{\min} > 0$; $\theta_{\min}$ corresponds to a certain maximum absorption coefficient $k_{\max}$. The fits in this paper were done with $\theta_{\min} = 10^{-8}$, which is rather conservative. In the interests of an accurate derivative fit (of $E'(u)$ to $T'(u)$) between $u_0 = 0$ and u_1, one should select $\theta_{\min}$ as large as possible, since if a $k_{\max}$ ($\theta_{\min}$) term is present in $E'(u) = -\sum a_i k_i \exp(-k_i u)$ it will dominate between $u = 0$ and $u = u_1$. This essentially arbitrary term clearly becomes more and more dominant and therefore undesirable as $k_{\max} \to \infty$ ($\theta_{\min} \to 0$), especially at $u = 0$, where we already see errors When of 100% and more in E' when $\theta_{\min} = 10^{-8}$. When the fitting procedure actually chooses a $\theta_{\min}$ term with a nonnegligible coefficient, it is inevitably because the function $T(u)$ is undersampled in the interval $[0, u_1]$, i.e., there is a large jump between $T(0) = 1$ and $T(u_1)$. In this case the exponentials which fit the data points $u_1, \ldots, u_N$ do not correctly extrapolate to $u_0 = 0$, and so a $\theta_{\min}$ term must make up the difference.

When the last data point $T(u_N)$ is not near zero, the fitting procedure will often select $\theta = 1$, corresponding to no absorption, $k = 0$. The coefficient of this constant term is, of course, smaller than $T(u_N)$ but is generally of the same order of magnitude; it represents that fraction of the given spectral interval which is effectively transparent for the range of absorber amounts under consideration.

With 14-significant-digit computations, we generally find about 150 data points to be a practical upper limit due to roundoff error in the computation of the residual polynomial. With 29-significant-digit computations we have had no difficulty using up to 250 data points. The method does not deteriorate catastrophically when too many data points are used, but there are telltale signs of numerical difficulty, like substandard accuracy in the fits.

(k) *Experience with transmission fitting.* We have fitted realistic H_2O, CO_2, and O_3 transmission data taken from the LOWTRAN model [35] and from band models.

These fits have been used by one of the authors [2, 11] for ESFT method calculations in the earth's atmosphere for 120- and 340-spectral-interval partitions of the solar and infrared spectrums. Thus the fitting method described above has been exercised for the widest variety of transmission functions, and has performed without failure in all cases. The fits to LOWTRAN transmission functions generally contain from three to eight terms and are accurate to better than 0.1 % at all data points, with 0.01 to 0.001 % being the typical error.

Generally the number of terms in the fit, the relative sizes of the a_i, and the range and distribution of the k_i are not much affected by the particular manner in which the transmission data is selected, for a fixed range of absorber amount. This indicates that the a_i and k_i are not merely numerical artifacts but contain legitimate information about the structure of the underlying absorption band. Our method may, as a consequence, have some utility in analyzing measured transmission data.

Based on considerations of L/H, and on the fact that transmission measurements tend to have a uniform absolute error, we have used weights $w_n = 1$ in fitting transmission functions. Other selections could be made to emphasize one part or another of the transmission curve.

3. Examples

There are very few published fits to which we can compare our method. Those of Sargent [6] and of Raschke and Stucke [20] are of roughly 1 % accuracy; we never see such large errors in our method. Lacis and Hansen [4] give an eight-term fit to Yamamoto's H_2O solar absorption data; since their method and their accuracy are about the best one can find in the literature, we have singled out their fit for special attention below. We also examine the problems of fitting to data actually sampled from an exponential sum and of fitting by the inverse Laplace transform technique with Laguerre quadrature.

In order to examine the effect of random errors in the data on the fits, we have chosen simply to round the data after varying numbers of decimal places. (The notation nD will hereinafter mean rounded to n decimal places.) The advantage of rounding is that it preserves monotonicity ($T(u_0) > \cdots > T(u_N)$) and it simulates the laboratory situation. Actual transmission measurements are presently of accuracy 2D to 3D, but hopefully better measurements will be forthcoming in the future.

(a) *Three-term exponential sum.* Great difficulty is experienced by all previously published methods in reconstructing an exponential sum, given data sampled from that sum (cf. [16, 30]). This problem is important, for example, in attempting to recover chemical rate constants or radioactive decay rates, as well as the number of species involved, from laboratory measurements. We demonstrate the excellent ability of our method to recover such a sum, even when the data are only accurate to 2D.

The sum we have selected simulates transmission data from a more strongly

absorbing spectral interval. The k_i range over two orders of magnitude, and the most rapidly decaying component has a substantial coefficient so that the curve is undersampled for small u. The sum is

$$T_{ES}(u) = 0.1e^{-0.01u} + 0.3e^{-0.1u} + 0.6e^{-u}.$$

We fit 100 data points $\{T_{ES}(u); n = 0, 1,..., 99\} = 1.0, 0.59,..., 0.04$. Table I shows the number of significant digits to which our method recovers each a_i and k_i, as a function of the number of decimal digits (5D,..., 2D) to which the data is rounded.

TABLE I

Number of Significant Digits to Which Coefficients a_i and Exponents k_i ($i = 1, 2, 3$) in $T_{ES}(u)$ are Recovered When the 100 Data Values are Rounded to 5D, 4D, 3D, and 2D. Also Corresponding Errors between Fit and Data, and Coefficients of Any Spurious Terms in the Fit

	5D	4D	3D	2D
$a_1(0.1)$	4	3	2	1
$k_1(0.01)$	4	3	3	1
$a_2(0.3)$	5	5	3	2
$k_2(0.1)$	5	3	2	2
$a_3(0.6)$	4	3	2	2
$k_3(1.0)$	4	4	2	2
e_{max}	5×10^{-6}	5×10^{-5}	5×10^{-4}	5×10^{-3}
e_{rms}	3×10^{-7}	3×10^{-6}	3×10^{-5}	3×10^{-4}
e'_{rms}	6×10^{-8}	6×10^{-8}	9×10^{-6}	1×10^{-5}
$e'_{\%}$	0.004	0.02	0.2	1.2
$a_{spurious}$	2×10^{-5}		4×10^{-4}	0.002
	2×10^{-5}	2×10^{-4}	3×10^{-3}	0.008

It is seen that our method recovers the a_i and k_i of the exponential sum to roughly n significant digits when the data are rounded to nD. Furthermore it finds the number of terms correctly, save for additional "spurious" terms (cf. Section 2(h)) with identifiably small coefficients as shown (normally the spurious terms impact only the last correct digit of the rounded data).

Note that this example is constructed so that no component falls below the level of rounding accuracy at all data points. If one of them did, it would of course be completely unrecoverable. This is an ever present caveat on the separation of exponentials; that is, there may always be unrecoverable terms with values down in the "noise" level. Their coefficients will almost always be less than $T(u_N)$, which

argues for as large a data range as possible; but nothing can be said about their exponents. A larger data range can furthermore compensate for lack of accuracy in the data; there is always a trade-off between these two factors in recovering exponential components.

The fitting errors $e_{\max}$, e_{rms}, and $e'_{\%}$ noted in Table I (cf. Section 2(i)) all increase by an order of magnitude each time a digit is rounded from the data, and in fact $e_{\max}$ is precisely the maximum rounding error at each step. e_{rms} is about an order of magnitude smaller than $e_{\max}$. These relationships are so general that $e_{\max}$ (or e_{rms}) would be an excellent indicator of the accuracy of the data if that were not known a priori.

A final point concerning $e_{\max}$ and e_{rms} is that they are based on differences between the fit and the *rounded* data. If residuals are computed from the *exact* data, $e_{\max}$ and e_{rms} fall by factors of 2 to 10, indicating that our method senses the smoothness of the underlying function and fits to it more than to the "rougher" function corresponding to the rounded data. This property has been found in all of our fitting experiments involving smooth underlying functions.

(b) *Yamamoto's* H_2O *solar absorption.* Lacis and Hansen [4] give the empirical fit to Yamamoto's solar-spectrum water vapor transmission data

$$T_Y(u) = 1 - \frac{2.9u}{(1 + 141.5u)^{0.635} + 5.925u}$$

where u is water vapor amount in precipitable centimeters. They then fit this curve with an eight-term exponential sum (their Table 1) good to within 0.1 % for $0.01 < u < 10$, using an NLLS method.

In Fig. 4 are plotted the fits generated by our method for $T_Y(u)$ for 225 data points

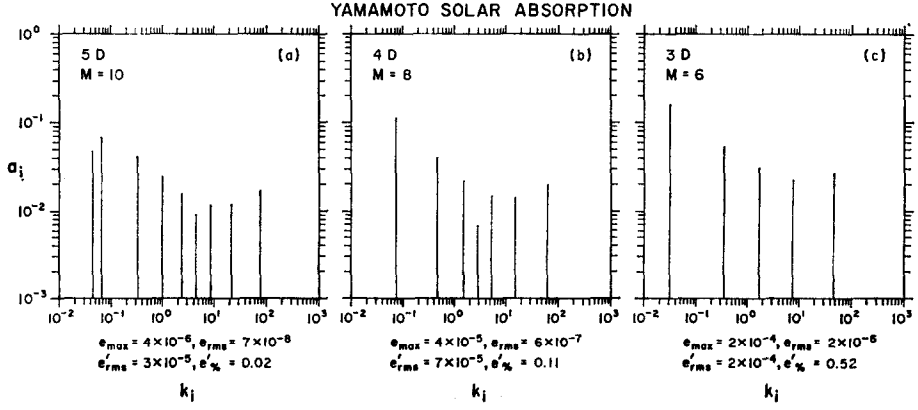

FIG. 4. Spike plots for exponential-sum fits to Lacis and Hansen's fit to the Yamamoto solar

rounded to 5D, 4D, and 3D and covering the range $u_0 = 0$ to $u_{224} = 10$. A $k = 0$ spike in each fit could not be included because of the logarithmic scale for k_i. The various errors, defined in Section 2(i) and given beneath each plot, refer to the exact rather than the rounded data. It is plain that rounding is an excellent artifice for reducing the number of terms in the fit if one is willing to sacrifice some accuracy. It is equivalent to removing a certain amount of differentiability or smoothness from the data, and the algorithm responds by fitting fewer terms, just as it does when fitting interpolated tabular data from LOWTRAN (cf. Section 2(k)).

Note that the maximum absolute error in our eight-term fit (Fig. 4(b)) is 0.004% and the r.m.s. absolute error is 0.00006%; compare this to the 0.1% error in the eight-term fit of Lacis and Hansen. Even our six-term fit (Fig. 4(c)) is significantly more accurate than their fit. Clearly the NLLS method of Lacis and Hansen falsely claimed to have converged to the best eight-term fit; the authors' own experiences with NLLS methods for exponential fitting are replete with such cases of false convergence.

(c) *Goody and Malkmus band models.* The Goody [34] random band model for transmission is

$$T_G(\tilde{u}) = \exp\left(-\frac{\pi\alpha}{d}\frac{\tilde{u}}{(1+\tilde{u})^{1/2}}\right)$$

where

$$\tilde{u} \equiv Su/\pi\alpha$$

and S is the mean line strength, d the mean line spacing, and α the mean line halfwidth. In the same notation, the Malkmus [36] random band model is

$$T_M(\tilde{u}) = \exp[-(\pi\alpha/2d)((1+4\tilde{u})^{1/2} - 1)].$$

We examine the behavior of fits for both the Goody and Malkmus models as d/α varies. Figure 5 shows spike plots for $d/\alpha = 2$, 20, and 200. (Rodgers and Walshaw [37] show d/α ratios varying from 7 to 60 for real infrared absorption bands.) All transmission data are rounded to 7D, and have a range 0.001–1. The fits shown in Fig. 5 have maximum errors ($e_{\max}$) between 10^{-4} and 10^{-5}; while we have derived fits of equivalent accuracy with many fewer terms, these particular ones were selected because the envelope of the spikes gives a better picture of the underlying k-distribution.

The logarithms of the exponents k_i exhibit a remarkably uniform spacing in Fig. 5 (the same thing can be seen in Fig. 4). Furthermore, each time d/α increases by an order of magnitude the major peak of the spike distribution shifts downward by two orders of magnitude, for both models. For $d/\alpha = 2$ the absorption lines tend to be very close together, therefore the windows are filled in, so that large values of k_i are emphasized. Our method reflects this by selecting a much smaller range of k_i values than when $d/\alpha = 20$ or 200. The shifting of the group of spikes to smaller values of k_i

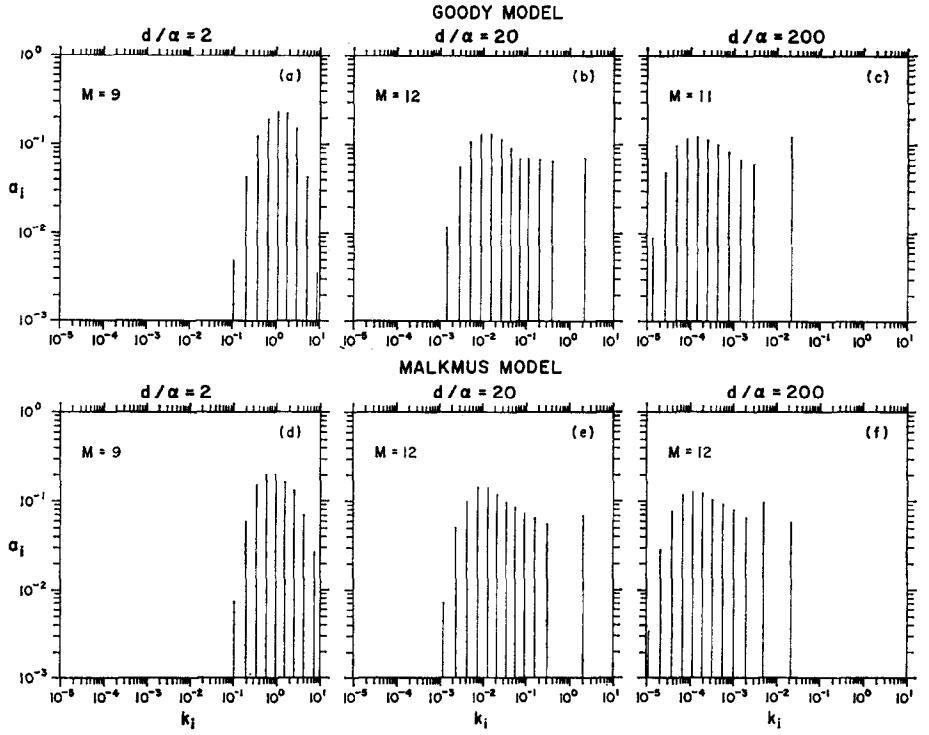

FIG. 5. Spike plots for exponential-sum fits to the Goody and Malkmus random band models, with line-spacing-to-halfwidth ratios of (a) and (d), 2; (b) and (e), 20; (c) and (f), 200.

as d/α increases reflects the increasing "gappiness" of the model spectrum with its increasing percentage of small absorption.

Another point about Fig. 5 is that the corresponding Goody and Malkmus fits appear very similar qualitatively, i.e., in the range and distribution of the k_i and the sizes of the a_i. This is not surprising since the two models have the same value and slope at $u = 0$, the same asymptotic form for large u, and do not differ significantly for intermediate values of u. But when their spike plots are overlaid, one finds systematic shifts of the Malkmus spikes toward smaller values of k_i; in this way the fitting method reflects the larger percentage of weak lines in the Malkmus model. The larger the value of k_i, the more the Malkmus spike is shifted to the left compared to the corresponding Goody spike (except when both models have coincident right-most spikes corresponding to $\theta_{\min}$). Corresponding Goody and Malkmus spikes have remarkably similar heights (values of a_i) except for the very right-most and left-most ones. Thus the Malkmus fits are very close to the Goody ones except that they are compressed toward smaller values of k_i.

(d) *Laguerre quadrature of the inverse Laplace transform.* Domoto [23] has pointed

out that the inverse Laplace transform (see the Introduction) of the Malkmus model takes the particularly simple form

$$f_M(k) = \frac{\beta e^{\beta}}{\pi^{1/2}} k^{-3/2} \exp\left(-\frac{k}{4} - \frac{\beta^2}{k}\right)$$

where

$$\beta = \pi\alpha/(2d).$$

Let us use this in a quadratured form of Eq. (2) and compare the exponential fits so generated with those from our method.

The Laguerre quadrature rule is the standard one for integrals like the Laplace transform. It may be written

$$\int_0^{\infty} F(x)\, e^{-x}\, dx \cong \sum_{i=1}^{M} A_i F(x_i)$$

or

$$\int_0^{\infty} G(x)\, dx \cong \sum_{i=1}^{M} A_i e^{x_i} G(x_i).$$

Applied to Eq. (2) for the Malkmus model, this leads to

$$T_M(u) = \int_0^{\infty} f_M(k)\, e^{-ku}\, dk \cong \sum_{i=1}^{M} A_i e^{k_i} f_M(k_i)\, e^{-k_i u}. \tag{25}$$

Values of A_i and k_i $(=x_i)$ were obtained using a computer program in Stroud and Secrest [38] for a wide variety of values of M. Selected fits generated by Eq. (25) for $M = 6$ and 30 are plotted as dotted lines in Fig. 6 for ratios $d/\alpha = 7$ and 20 ($d/\alpha = 7$ is the smallest value found by Rodgers and Walshaw [37]). $T_M(u)$ is plotted for comparison in each case. It would have been useless to present our Malkmus model fits in this form because they would have been indistinguishable from the exact curve. This indicates the vastly poorer accuracy from fitting according to Eq. (25).

The trends we observe in Fig. 6 are indicative. That is, for a fixed number of terms M, the fits continue to improve as d/α decreases below 7 and continue to grow worse as d/α increases above 20. They also improve as M increases, as we might expect, but only up to a point. These Laguerre-quadrature fits behave very much like asymptotic series—they are not infinitely improvable by increasing the number of terms. This is perhaps even more damning for such fits than their abysmally poor accuracy or their frequent prediction of transmission greater than one for small u.

The bad behavior of Laguerre quadrature in this example is of course related to the fact that it is designed to integrate polynomials multiplied by exponentials. The behavior of $f_M(k)$ is decidedly nonpolynomial, especially for $k \to 0$ where it has an essential singularity. But if one adapts a quadrature rule specifically to $f_M(k)$, one has

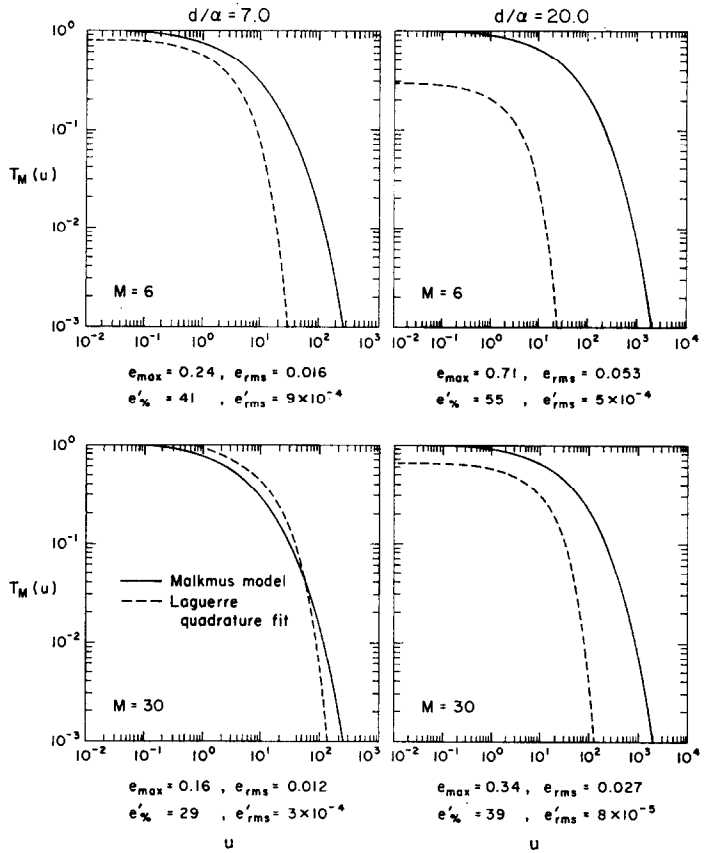

FIG. 6. Plots of the Malkmus random band model transmission $T_M(u)$ vs u, compared to exponential-sum fits to $T_M(u)$ generated by the Laguerre quadrature method and having $M = 6$ and $M = 30$ terms. The first column is for a line-spacing-to-halfwidth ratio $d/\alpha = 7$, and the second column for $d/\alpha = 20$.

entirely lost generality since the behaviors of $f_M(k)$ as $k \to 0$ and $k \to \infty$ are very special.

A deeper understanding of the whole quadrature-fit question would require results from Tauberian theory (cf. [22]) relating the asymptotic behaviors of $T(u)$ and $f(k)$ in Eq. (2). For example, the essentially singular behavior of $f_M(k)$ as $k \to 0$ profoundly influences the fit accuracy as $u \to \infty$. But it is beyond our scope to enter into such questions here; we merely wish to show that Eq. (2) does not yield a useful general procedure for exponential-sum fitting in realistic situations.

4. SUMMARY AND CONCLUSIONS

We have given a new method for obtaining the best fit, in the least squares sense, of a sum of exponentials to arbitrary data, with particular reference to the fitting of

transmission functions. The fitted sum consists of purely decaying exponentials with positive coefficients. Our method rests on a solid theoretical groundwork of existence, uniqueness, and convergence proofs laid by Cantor and Evans [31], and in addition to these important features it internally selects the correct number of terms consistent with the accuracy of the data and has procedures to sidestep the inherent ill conditioning of the problem. The fits which we produce are generally orders of magnitude more accurate than any which have been published heretofore.

The focus of this work has been the fitting of transmission functions, because of our long-standing interest in the ESFT method (reviewed in the Introduction) for performing broad band radiative flux calculations in a multiple scattering/line absorbing atmosphere. Example fits were given and discussed for the Goody and Malkmus random band model and the Yamamoto H_2O solar absorption data. We looked particularly at the effect of rounding the data (corresponding to measurement errors) on both the accuracy and number of terms in the fit; based on these studies, we pointed out the potential utility of our method in analyzing measured transmission data. Finally, we made some preliminary studies of the inverse Laplace transform method of generating exponential fits, and for the Malkmus model case with Laguerre quadrature found this approach to be totally unsatisfactory.

But exponential fitting has much broader application than to ESFT problems, for example to the identification of multiple radioactive species and to the analysis of human lung exhalations and chemical rate constants. This is because many physical processes are formulated mathematically as systems of linear differential equations with constant coefficients. Such systems have sums of exponentials for solutions, except in certain special cases, and therefore one may, with our method, work backward from the solutions to the constant coefficients (which are often rate constants of some sort). We have therefore given an example of the success of our method in analyzing data actually sampled from an exponential sum and rounded to as few as two decimal digits, which relates to these other sorts of applications.

A computer code implementing our exponential-sum fitting method is available from the authors on request.

ACKNOWLEDGMENTS

The authors wish to thank Professor C.D. Rodgers for helpful discussions on band models. We also acknowledge the support of the ARPA Climate Dynamics Program during part of this work and of NCAR and particularly, Dr. Stephen Schneider during the remainder.

REFERENCES

1. G. HUNT AND I. GRANT, *J. Atmospheric Sci.* **26** (1969), 963.
2. W. J. WISCOMBE AND B. E. FREEMAN, Preprint volume, First Conference on Atmospheric Radiation, Ft. Collins, Colo., American Meteorology Society, Boston, Mass., 1972.
3. A. ARKING AND K. GROSSMAN, *J. Atmospheric Sci.* **29** (1972), 937.

4. A. LACIS AND J. E. HANSEN, *J. Atmospheric Sci.* **31** (1974), 118.
5. E. RASCHKE, M. KERSCHGENS, U. PILZ, AND U. REUTER, Abstract Volume, Second Conference on Atmos. Radiation, Arlington, Va., American Meteorology Society, Boston, Mass., 1975.
6. S. SARGENT, "A Numerical Solution to the Transfer Equations for Infrared Radiation in a Non-Gray, Absorbing, Emitting, and Scattering Atmosphere," Ph.D. Thesis, Univ. of Wisconsin, 1971.
7. S. SARGENT AND W. BECKMAN, *J. Atmospheric Sci.* **30** (1973), 88.
8. K. LIOU AND T. SASAMORI, *J. Atmospheric Sci.* **32** (1975), 2166.
9. J. POLLACK, O. TOON, A. SUMMERS, W. VAN CAMP, AND B. BALDWIN, *J. Appl. Meteorol.* **15** (1976), 247.
10. G. YAMAMOTO, M. TANAKA, AND S. ASANO, *J. Quant. Spectrosc. Radiat. Transfer* **11** (1971), 697.
11. W. J. WISCOMBE, Solar radiation calculations for arctic summer stratus conditions, *in* "Climate of the Arctic" (G. Weller and S. Bowling, Eds.), Geophysical Institute, Univ. of Alaska Press, Fairbanks, Alaska, 1973.
12. S. CHANDRASEKHAR, *Mon. Notic. Roy. Astron. Soc.* **96** (1935), 21.
13. W. MACADAMS, "Heat Transmission," pp. 111–112, McGraw-Hill, New York, 1954.
14. G. BOND AND C. SIEWERT, *J. Quant. Spectrosc. Radiat. Transfer* **10** (1970), 865.
15. K. YA. KONDRAT'YEV, "Radiation in the Atmosphere," Academic Press, New York, 1969.
16. C. LANCZOS, "Applied Analysis," Prentice-Hall, Englewood Cliffs, N.J., 1956.
17. E. AVRETT AND D. HUMMER, *Mon. Notic. Roy. Astron. Soc.* **130** (1965), 295.
18. F. HILDEBRAND, "Introduction to Numerical Analysis," McGraw-Hill, New York, 1956.
19. G. HUDSON, *Amer. J. Phys.* **21** (1953), 362.
20. E. RASCHKE AND U. STUCKE, *Beitr. Physik Atmos.* **46** (1973), 203.
21. M. OSBORNE, *SIAM J. Numer. Anal.* **12** (1975), 571.
22. R. BELLMAN, R. KALABA, AND J. LOCKETT, "The Numerical Inversion of the Laplace Transform," American Elsevier, New York, 1966.
23. G. DOMOTO, *J. Quant. Spectrosc. Radiat. Transfer* **14** (1974), 935.
24. D. GARDNER, J. GARDNER, G. LAUSH, AND W. MEINKE, *J. Chem. Phys.* **31** (1959), 978.
25. D. GARDNER, *Ann. N.Y. Acad. Sci.* **108** (1963), 195.
26. G. BROWNELL AND A. CALLAHAN, *Ann. N.Y. Acad. Sci.* **108** (1963), 172.
27. A. PAPOULIS, *SIAM J. Control* **11** (1973), 466.
28. G. WESTLEY AND J. WATTS, Eds., "The Computing Technology Center Numerical Analysis Library," Oak Ridge National Lab, Report CTC-39, Oak Ridge, Tenn., 1970.
29. J. LANG AND R. MÜLLER, *Comput. Phys. Commun.* **2** (1971), 79.
30. M. BOX, *Comput. J.* **9** (1966), 67.
31. D. G. CANTOR AND J. W. EVANS, *SIAM J. Appl. Math.* **18** (1970), 380.
32. C. LAWSON AND R. HANSON, "Solving Least Squares Problems," Prentice-Hall, Englewood Cliffs, N.J., 1974.
33. G. HADLEY, "Non-Linear and Dynamic Programming," Addison-Wesley, Reading, Mass., 1964.
34. R. GOODY, *Quart. J. Roy. Meteorol. Soc.* **78** (1952), 165.
35. J. E. A. SELBY AND R. MCCLATCHEY, "Atmospheric Transmittance from 0.25 to 28.5 μm: Computer Code LOWTRAN 2," Air Force Cambridge Research Labs, Report AFCRL-72-0745, Bedford, Mass., 1972.
36. W. MALKMUS, *J. Opt. Soc. Amer.* **57** (1967), 323.
37. C. D. RODGERS AND C. D. WALSHAW, *Quart. J. Roy. Meteorol. Soc.* **92** (1966), 67.
38. A. STROUD AND D. SECREST, "Gaussian Quadrature Formulas," Prentice-Hall, Englewood Cliffs, N. J., 1966.