JOURNAL OF COMPUTATIONAL PHYSICS 18, 188–200 (1975)

Comparison of Constrained Linear Inversion and an Iterative Nonlinear Algorithm Applied to the Indirect Estimation of Particle Size Distributions

S. TWOMEY

*Institute of Atmospheric Physics, The University of Arizona, Tucson, Arizona 85721**

Received June 6, 1974; revision received February 18, 1975

Particle size distributions in the atmosphere are not smooth, and the application of the usual smoothing constraints to the inversion of indirect measurements of such distributions is unsuccessful for this reason, even though the method used is successful for hypothetical (but unreal) distributions. It is shown that nonlinear iterative procedures can be applied successfully in either situation.

INTRODUCTION

The inference of particle size distribution from measurements of the number of particles transmitted through filters with known filtration characteristics gives rise to a very typical inversion problem, wherein the kernels (the filter transmission as a function of particle size) are smooth and "band limited," in the sense that the kernels fall to zero with finality on each side. In these respects, there is a strong similarity to "optical" kernels, even though the physics and dimensions are very different in the two problems.

Particle distributions in the atmosphere are, however, far from smooth even when averaged, for they are known to exhibit behavior of the form $f(r) \equiv Ar^{-\alpha}$ with $\alpha \sim 4$ over a portion of the range, but fall toward zero below some size which is a priori unknown. If, for example, one plots the distribution

$$f(x) = Ar^{-3} \exp(-10^{-12}r^{-2}), \qquad (x = \ln r),$$

the result (Fig. 1) is a distribution which is probably close to that of many real atmospheric particle distributions. If, however, the result is plotted on a linear scale, it is far from smooth. Clearly the smoothness could be greatly enhanced by transformation to $r^3 f(x)$, but the inverse cubic form of the distribution holds only

* A portion of the research for this manuscript was performed at the Division of Cloud Physics, CSIRO, Sydney, Australia.

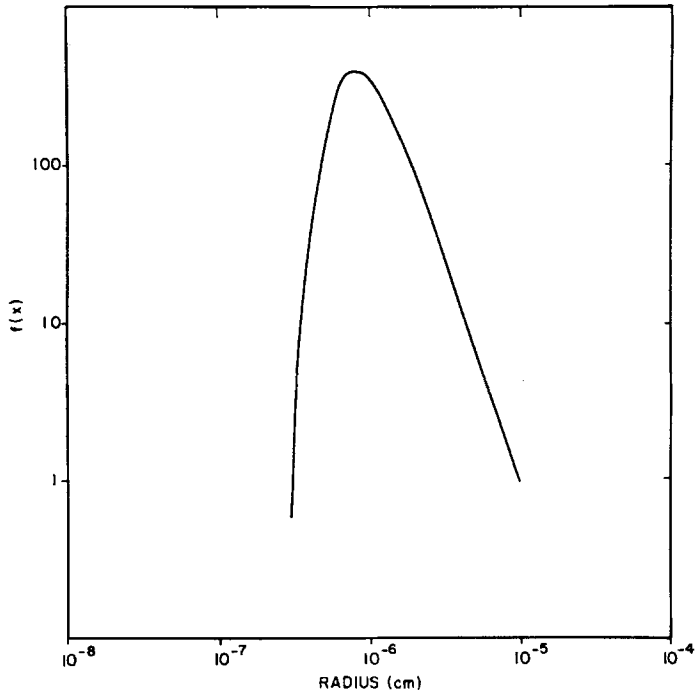

FIG. 1. A typical distribution function for the atmospheric aerosol.

approximately and on average; furthermore, the peak of the distribution obviously must occur where no inverse power law holds even approximately. Clearly, two of the most fundamental parameters which one might hope to derive from measurements are: (1) the location of the peak, and (2) the most appropriate value of the exponent (i.e., the slope of the distribution) in the region where an inverse power law tends to hold. Any method of inversion which places constraints on either or both of these aspects of the solution is therefore undesirable. In the following sections, we will give examples of the application of linear constrained inversion techniques to this problem and will also discuss a nonlinear iterative algorithm which gives results that seem to be, in general, superior and are, in a sense, less constrained than those given by the more familiar procedure. In each case, the goal is the retrieval (as far as possible) of the function $f(x)$ from measurements $g_1, g_2, \ldots, g_r$, where

$$g_i = \int_a^b K_i(x) f(x)\, dx + \text{noise}. \tag{1}$$

Figure 2 shows a set of kernel functions. They are, in fact, filter transmissions for

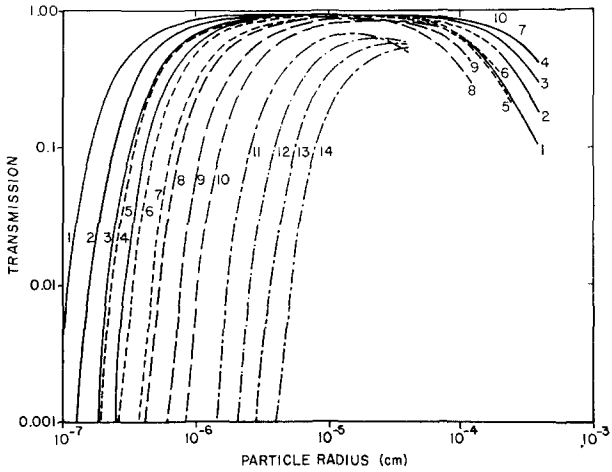

FIG. 2. Transmission curves (i.e., kernel functions) for a set of measurements using severa Nuclepore filters at various flow rates.

Nuclepore[1] filters at various air flows as a function of particle radius. In the context of these problems, it is more realistic and more useful to use the logarithm of radius, rather than radius, as the independent variable, and x in (1) above may be identified with $\log r$. Where possible, however, the discussion will be kept general, and there will, in fact, be little need to refer to the specific physics of particle filtration. The band-limited behavior arises from the removal of small particles by diffusion to the pore walls and the removal of large particles by impaction and simple mechanical sieve-type removal. The smooth nature of the kernels arises from the fact that the diffusion coefficient and the particle radius and mass, which are the relevant physical parameters, are themselves smooth, indeed monotonic, functions of size.

1. Application of Constrained Linear Inversion

Numerical Method

The principle of constrained linear inversion has been applied in a variety of physical problems. A fairly detailed discussion was given by the writer in an earlier paper [4]. For completeness, however, a short summary will be given at this point.

If $K_i(x)$ is smooth, then the change in g_i brought about by the addition of a

[1] General Electric Company, Pleasanton, CA.

periodic term to $f(x)$ in (1) decreases rapidly with increase of frequency of the periodic term. In other words,

$$\left| \int_a^b K_i(x) \begin{matrix} \cos \omega x \\ \sin \omega x \end{matrix} dx \right|$$

decreases with increase of ω. The rate of decrease will depend on the nature of the kernels; for many physical kernels, the decrease will be at least as fast as ω^{-2} or even ω^{-3}. There is thus an inherent insensitivity of g_i (for all i) to higher-frequency components and an instability on inversion, whereby small noise (error) components in measured or numerically computed g_i give rise to extremely large spurious oscillations in the "solution." The instability cannot be removed, but it can be avoided if one adopts a constrained method of solution wherein the smoothest possible solution is selected from all those $f(x)$ which give g_i lying within a prescribed distance from the measured g_i. The algebra of this process is straightforward:

Given

$$\int K_i(x) f(x)\, dx = g_i + \epsilon_i ,$$

this is transformed to a quadrature

$$\sum_{j=1}^{m} a_{ij} f_j = g_i + \epsilon_i , \tag{2}$$

or in matrix notation

$$A\mathbf{f} = \mathbf{g} + \boldsymbol{\epsilon}.$$

(Any quadrature error need only be incorporated into ϵ_i.) The problem is then to find the "smoothest" vector $(f_1, f_2, \ldots, f_m)$ from the manifold of vectors for which $\sum \epsilon_i^2$ is less than or equal to some prescribed overall error, e^2. One can adopt different measures of departure from smoothness, for example:

the variance, $\sum (f_j - \bar{f})^2$,

the sum of squares of first differences, $\sum (f_j - f_{j-1})^2$,

the sum of squares of second differences, $\sum (f_j - 2f_{j-1} + f_{j-2})^2$,

and so on, but all, including more complicated spectral formulations, can readily be shown to be quadratic forms which in vector–matrix notation may be written $\mathbf{f}^* H \mathbf{f}$, where the asterisk denotes transposition and H is a matrix which is determined by the measure of smoothess selected. For simple measures of smoothness,

H takes a simple form, e.g., it is the identity matrix if one uses as a measure the variance of $\mathbf{f}$. Constrained inversion is thus accomplished by finding that vector $\mathbf{f}$ which makes $\mathbf{f}^*H\mathbf{f}$ a minimum while

$$\sum \epsilon_i^2 = \boldsymbol{\epsilon}^*\boldsymbol{\epsilon} = (A\mathbf{f} - \mathbf{g})^* (A\mathbf{f} - \mathbf{g}) \tag{3}$$

is held fixed. This familiar mathematical problem is solved by the method of Lagrangian multipliers; in fact, one finds the extremum of

$$(A\mathbf{f} - \mathbf{g})^* (A\mathbf{f} - \mathbf{g}) + \gamma \mathbf{f}^* H\mathbf{f},$$

where γ is the undetermined Lagrangian multiplier. Differentiation with respect to $f_1, f_2, \ldots, f_m$ leads to a linear system of equations in $f_1, f_2, \ldots, f_m$, the solution of which is readily found as:

$$\mathbf{f} = (A^*A + \gamma H)^{-1} A^*\mathbf{g}. \tag{4}$$

The undetermined multiplier γ is uniquely determined by the prescribed error magnitude $\sum \epsilon_i^2$, but the relationship is a complex one, and it is far easier to obtain the solution for several values of γ and obtain $\sum \epsilon_i^2$ by substitution of the solution into (3).

Results

Starting with an assumed distribution and a set of kernel functions (each kernel being the transmission curve of a filter at a specified flow rate), a set of data was computed giving the fraction of the initial total particle number transmitted at each filter/flow rate combination. For convenience, each such combination will, in the future, be referred to as a "filter"; "filter function" or "filter transmission" will be understood in the same sense. The direct measurement of total particle number was also included as a filtered measurement made with a filter of unit transmission at all particle sizes.

For "reasonable" distributions, this method gives excellent results. (Unfortunately, real atmospheric particle distributions are not reasonable.) Figure 3 shows an assumed original distribution (bimodal but otherwise smooth) and solutions given by inversion according to (4) of the calculated transmissions through 17 filters (which incidentally were restricted to physically possible filters; unrealistically large or small flow rates, for example, were excluded). The smoothing constraint employed was that of minimum variance, so that the identity matrix I was employed for H.

The conclusion that filter measurements of the kind envisaged can give, by inversion, size distribution to the degree of agreement and resolution suggested by Fig. 4, is not however warranted unless the necessary experimental accuracy can be achieved. Table I shows the high accuracy needed.

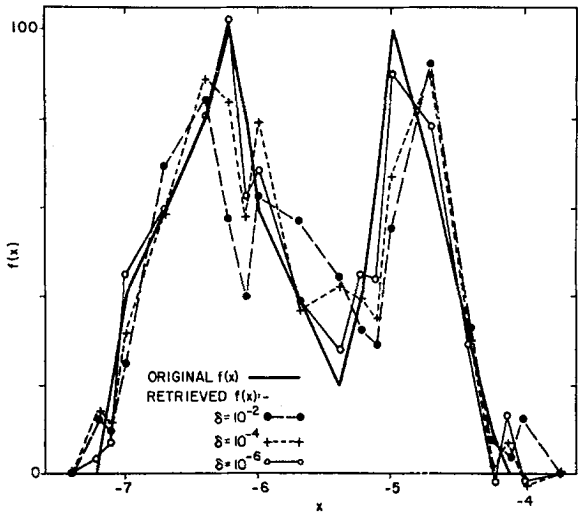

FIG. 3. An assumed bimodal distribution (heavy curve) and the result of inverting filter measurements (calculated) by constrained linear inversion with a smoothing parameter γ.

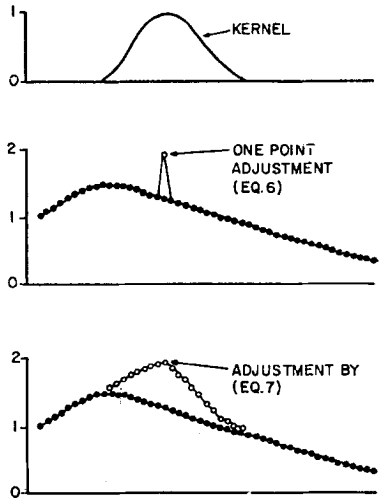

FIG. 4. Illustration of the iterative adjustment process applied by Eq. (7), as contrasted with procedures (e.g., Eq. (6)) in which only one point is adjusted at a time.

TABLE I

γ vs. Relative Error

γ	$(\Sigma \epsilon_i^2/\Sigma f_i^2)^{1/2}$
10^{-2}	1.5%
10^{-3}	0.3%
10^{-4}	0.09%
10^{-5}	0.03%
10^{-6}	0.007%

In field measurements of particle number, a relative accuracy of a few percent is achievable with some difficulty, but fractional percent accuracy is hardly possible. Thus the inversion for $\gamma = 10^{-2}$ represents about the best that could be achieved in practice with a well-behaved unknown distribution. The "measured" numbers ranged from about 20 to 100% of the total, so the measurement would not be unduly demanding from the point of view of the "dynamic range" of number over which the measurements would have to extend.

Turning now to a more realistic distribution, a Junge-type (inverse cube) distribution down to $x = -5.7$ (i.e., radius 2×10^{-6} cm), falling to zero at $x = -6.0$ (radius 10^{-6} cm), the results of applying the same procedure were extremely disappointing. The results are too poor to be effectively shown in a figure, but they show either (for smaller γ) large oscillations or (for very large γ) smoother curves which do not agree very well at all with the original function from which the data were generated. The difficulty evidently arises from the nature of $f(x)$, since the inversion procedure was exactly the same as for Fig. 3.

An inverse-cube distribution as mentioned earlier, is very far from smooth. If there are good grounds for expecting the solution always to behave in this way, some of the trouble can be resolved by looking upon $r^3 f(x)$ rather than $f(x)$ as the unknown and applying the smoothing constraint to it. If C is a diagonal matrix such that $C_{ii} = r_i^3$, then the fundamental equation

$$A\mathbf{f} = \mathbf{g} + \boldsymbol{\epsilon}$$

can be rewritten

$$AC^{-1}(C\mathbf{f}) = \mathbf{g} + \boldsymbol{\epsilon}.$$

If this equation is inverted by (4) with $C\mathbf{f}$ treated as the unknown, one obtains

$$\begin{aligned} C\mathbf{f} &= (C^{*-1}A^*AC^{-1} + \gamma H)^{-1}\, C^{*-1}A^*\mathbf{g} \\ &= C(A^*A + \gamma C^*HC)^{-1}\, A^*\mathbf{g} \\ \mathbf{f} &= (A^*A + \gamma C^*HC)^{-1}\, A^*\mathbf{g}. \end{aligned} \tag{5}$$

Thus the constraint can be modified so as to "see" an inverse-cube distribution (or for that matter, any other power) as smooth. If, for example, $H = I$, then the inverse-cube behavior is forced just by adding $r^6\gamma$ rather than γ to the diagonal elements of A^*A. However, this forcing is often undesirable and may force an inverse-cube behavior where it does not really exist.

2. The Nonlinear Iterative Inversion

Discussion

When the kernel function has a single maximum such that the ith kernel attains a maximum at say $x = \xi_i$, then the value of the object distribution near ξ_i has the greatest relative influence on the ith measurement. This has been the basis of iterative schemes used successfully by Chahine [1] and by Grassl [2]; iterative methods have also been applied successfully by Smith [3]. The basis of the Chahine method is to start with a first guess $f_0(x)$ and repeatedly proceed through each of the m measurements, adjusting at each step only the ordinate for the value of x for which the corresponding kernel attains a maximum. If $f_p(x)$ denotes the iterated function after the pth pass through the set of m measurements (in other words, the $[\{p-1\}m + i]$ th iteration), then $f_{p+1}(x)$ is obtained from $f_p(x)$ by:

$$f_{p+1}(\xi_i) = g_i \Big/ \left[\int K_i(x) f_p(x)\, dx\right] \cdot f_p(\xi_i). \tag{6}$$

This method obviously is applicable only when the number of tabular x values is equal to the number of measurements and when each measurement can be associated with one of the tabular x's. Furthermore, the larger the number of measurements, the higher the frequencies permitted by the spacing in x and the greater the opportunity for instabilities. The precise limit to the number of tabular points in x for stability to be maintained obviously depends on the spectral behavior of the kernels. If no kernel varies across an interval by more than the error component, then clearly conditions for instability exist.

The problem of instability can be looked at in the context of orthogonal function vectors and linear function space. If the object function is expanded in terms of orthogonal functions $\phi_1(x)$, $\phi_2(x)$,...,

$$f(x) = c_1\phi_1(x) + c_2\phi_2(x) + \cdots,$$

then obviously c_k cannot be inferred from measurements of $\mathbf{f} \cdot \mathbf{K}_i$ [using the

shorter dot-product notation for $\int_b^a K_i(x) f(x)\, dx$], if $\phi_k(x)$ happens to be orthogonal to each of the $K_i(x)$, for c_k can then have any value without altering any of the y_i. However, if some $\phi \cdot K$ is very small for all the K's the situation is just as bad, since large changes can be made in c_k with only a slight change resulting in any y_i.

There is therefore considerable value in approximations to $f(x)$ by means of sums of the kernel functions themselves:

$$f(x) = a_1 K_1(x) + a_2 K_2(x) + \cdots + a_m K_m(x).$$

This is algebraically equivalent to expansion of $f(x)$ in terms of orthogonal functions derived from the $K_i(x)$, but use of such functions had the disadvantage that, if there exist nontrivial combinations of the K's which almost vanish throughout the interval of interest, there will be some orthogonal functions which are represented only very weakly in the set $K_i(x)$, $i = 1, 2, \ldots, m$. Use of the nonorthogonal K's reduces such problems of near-singularity.

These considerations relate most directly to the question of information content in a set of measurements with nonorthogonal kernels, but they also suggest a modification of the nonlinear iterative procedure which increases its stability and allows the number and location of tabular points in x to be made independent of the number of measurements. The algorithm suggested differs from that of (6) in that, instead of modifying only the ordinate for $x = \xi_i$ when one is dealing with the measurement the kernel of which attains a maximum at $x = \xi_i$, one modifies the previous iterate over the entire region where the kernel is nonzero, with a weighting proportional to the value of the kernel. The procedure is as follows: If $r_p^{(i-1)}$ denotes the ratio of the observed g_i [i.e., $\int K(x) f(x)\, dx$] to $\int K(x) f_p^{(i-1)}(x)\, dx$, then the ith iteration is:

$$f_p^{(i)}(x) = [1 + (r_p^{(i-1)} - 1)\, K_i(x)]\, f_p^{(i-1)}(x). \tag{7}$$

The kernels involved here, being filter transmissions, necessarily have $|K_i(x)| < 1$ for all x and i. If kernels exceeding unity were involved, they should be scaled to ensure $|K_i(x)| \leqslant 1$; this is essential to ensure that $f_p^{(i)}(x)$ never can become negative. The procedures are illustrated in Fig. 4. Note that the shape of the alteration obtained with (7) is set mainly by the kernel; it will *not* become narrower if a greater number of tabular x's are employed. Furthermore, there is no necessity to assign for each of the kernel functions a nodal value of x, a highly artificial procedure when the kernels are the flat-topped functions shown in Fig. 2. It may be noticed that if the change effected in any one iteration is small, then the procedure of (7) tends to constrain the iterated function to remain in the function subspace spanned by $K_1(x), K_2(x), \ldots, K_m(x)$ if the previous iterate is within that subspace. In fact, if

the first guess were $f(x) \equiv 1$ and a sequence of small changes were made, the result would be

$$[1 + \beta_1 K_1(x)][1 + \beta_2 K_2(x)][1 + \beta_3 K_3(x)] \cdots,$$

where the β's represent the value of $[r_p^{(i-1)} - 1]$ at each iteration. To the extent that quadratic and higher combinations of the β's can be neglected, the iterates can be written

$$1 + \beta_1 K_1(x) + \beta_2 K_2(x) + \cdots.$$

The change at any iteration can, of course, be scaled down by any desired factor at the expense of computation time.

Results

Figure 5 shows the result of applying the iterative algorithm of (7) to the inversion problem where the constrained linear inversion procedure gave the solution shown in Fig. 3. The solution in Fig. 5 is seen to be of comparable quality to that in Fig. 3.

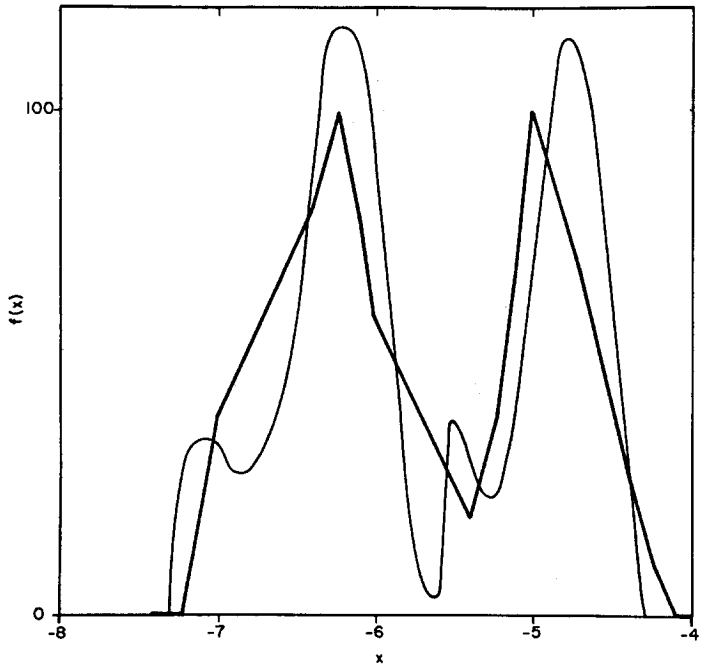

FIG. 5. The iterative procedure applied to the same data as used for Fig. 3.

In the case of the inverse-cube distribution where the constrained linear inversion failed to give a useful solution (except by the introduction of weights which artificially force an inverse-cube solution), the iterative algorithm was highly effective. Figure 6 shows the solution obtained for the distribution shown: Evidently the slope was recovered very well and the position of the peak in the solution was not far from that of the original.

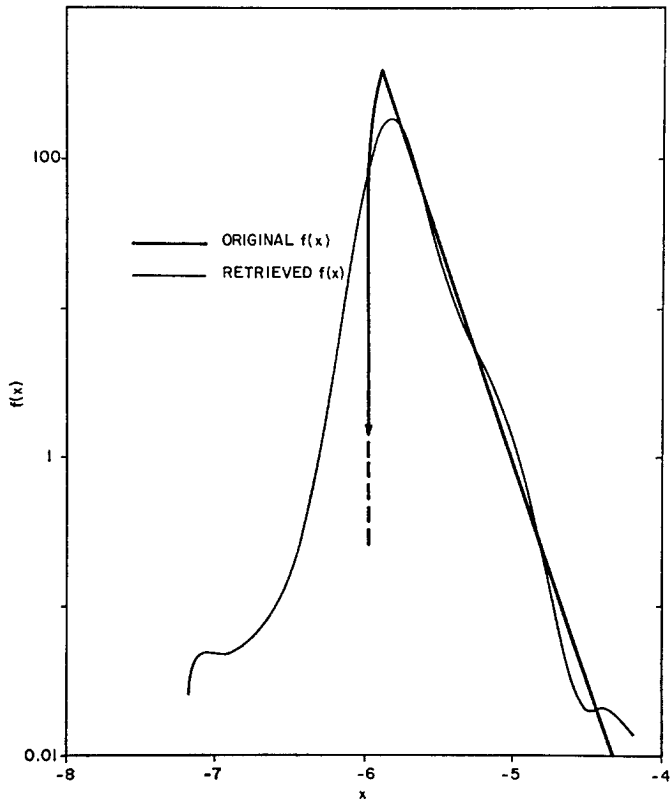

FIG. 6. The iterative procedure applied to filter data calculated for an object function similar to a real atmospheric aerosol distribution.

It is important to realize that a solution of the quality shown here, wherein $f(x)$ is accurately recovered over three decades in $f(x)$, was achieved with a similar range in magnitude of the g_i, which ranged from 140 down to 0.001. No existing instrument can measure particle concentrations accurately over five decades of concentration, and it is not being suggested that solutions of the quality of that shown in Fig. 6 are technologically feasible at present. In other fields, the dynamic

range of measurements is not so restricted and solutions of this accuracy and dynamic range may be feasible.

3. Effect of Errors

To assess the effect of errors on the solution obtained, we have repeated some inversions after a random error component was introduced into the g_i. The

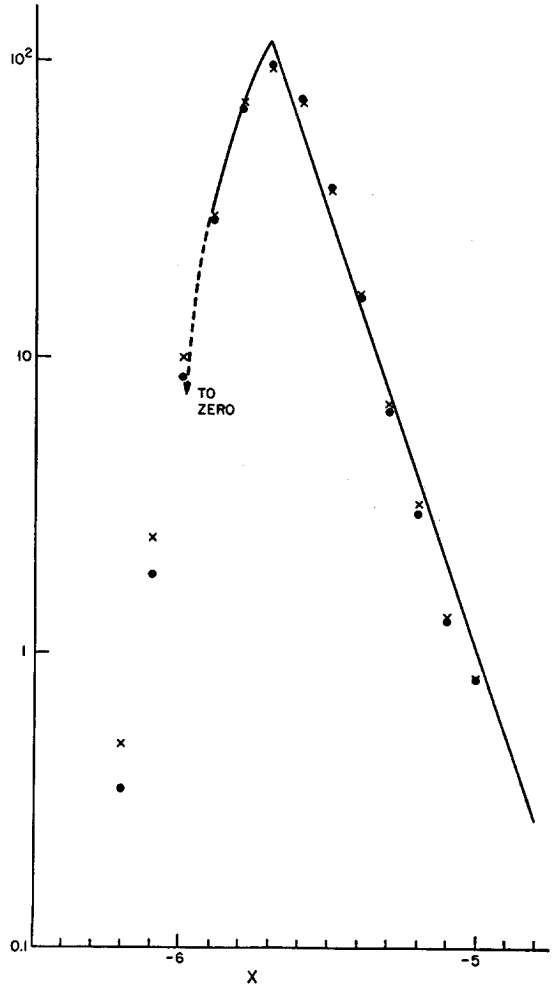

Fig. 7. The effect of introducing random errors in g is shown by the crosses. The dots represent the solution obtained when the error was not introduced.

inversions were found to be quite stable with respect to such errors. Figure 7 shows an initial hypothetical distribution, the inversion obtained after 100 passes through a set of 16 g_i using (7) (these results being indicated by dots), and the inversion obtained after the same number of iterations when g_i was adulterated by a random error of $\pm 4\%$ (r.m.s. value 2.3%). The effect of the errors is seen to be slight in comparison to the differences between either solution and the original hypothetical distribution, and they have not introduced any additional oscillatory instabilities in the solution.

Conclusions

The iterative nonlinear algorithm described by (7) possesses advantages over more familiar and more direct inversion methods when the object function and the measurements extend over a wide dynamic range. Stability is achieved by the process itself, since a smooth initial guess is multiplied by smooth adjusting functions at each step. The method is, however, considerably slower than constrained linear inversion, requiring about five times more computing time.

References

1. M. T. Chahine, *J. Opt. Soc. Am.* **58** (1968), 1643.
2. H. Grassl, *Appl. Opt.* **10** (1971), 2534–2538.
3. W. L. Smith, *Appl. Opt.* **9** (1970), 1994–1999.
4. S. Twomey, *J. Franklin Inst.* **279** (1965), 95–109.