Iterative Nonlinear Inversion Methods for Tomographic Problems

S. TWOMEY

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

(Manuscript received 4 March 1987, in final form 16 June 1987)

ABSTRACT

The application of nonlinear iterative algorithms to two-dimensional tomographic reconstructions is discussed and a number of numerical examples are given, using as an illustrative basis reconstruction of the spatial distribution of liquid water in clouds from measurements of microwave attenuation. (The method, however, is not restricted to that specific problem and appears to be especially suitable for inversions involving a large number of unknowns.)

1. Introduction

The advantages of tomographic approaches to atmospheric remote sensing problems were discussed in a general context by Fleming (1982). Warner et al. (1985) discussed tomographic reconstruction of spatial distributions of liquid water in clouds from measurements of microwave radiance, and Warner et al. (1986) showed some experimental results from that technique. In this paper, a hypothetical problem, very similar in essence to that of Warner et al. (1986) is examined in order to test the usefulness of quite a different iterative algorithm for such problems. The algorithm, although originally developed for conventional one-dimensional inverse problems involving smooth kernel functions, appeared to have potential advantages for atmospheric tomographic problems.

The distinction between "tomographic" and other problems is simple, although not in every case clearcut. In conventional satelliteborne infrared temperature sounders, for example, an instrument measures radiance at several wavelengths. These radiances arise from thermal emission in a localized atmospheric column (defined by the field-of-view of the instrument), and a vertical temperature profile in that column is obtained from those measured radiances. With an instrument looking vertically downward, as the satellite moves in orbit, sequential sets of radiances can be processed to obtain temperature soundings at each of many latitudes and longitudes; i.e., a three-dimensional temperature field $T(x, y, z)$ is described by sets of vertical temperature profiles $T_i(z)$, which are obtained as the satellite moves horizontally from (x_1, y_1) to (x_2, y_2) to (x_3, y_3), etc.

That is not a tomographic approach (as the word is normally applied), since there are no intersecting "rays" and each of the profiles, $T_1(z)$, $T_2(z)$, etc., is independent of all the others. If, however, while located at (x_2, y_2), an instrument also looked back along a slant path which intersected the vertical column viewed previously, then there would be at (x_1, y_1) an interval in z which contributed to both measurements. While making for a greater complexity, tomographic measurements have, nevertheless, distinct advantages with respect to vertical resolution. For example, in the slant measurement envisaged above, intersection of the two columns implies, in the z-direction, a weighting of zero to all vertical levels at (x_1, y_1) which lie *outside* the intersection. This contrasts with a conventional remote sensing measurement sequence, in which *all* levels in z make *some* contribution to each of the (spectral) measurements made at (x_i, y_i). In the tomography case, the weighting functions contain higher-frequency components, which can only improve vertical resolution on inversion.

Clearly, there are also disadvantages in the tomographic approach. Instead of solving for perhaps 10–20 unknowns representing a vertical profile $T(x_i, y_i, z)$ at (x_i, y_i) and then moving on to (x_{i+1}, y_{i+1}), solving for a similar number of unknowns representing $T(x_{i+1}, y_{i+1}, z)$, and so on, we encounter in tomography measurements to which several horizontal positions may have contributed. Thus a sequence of comparatively compact systems of equations in the conventional approach are essentially replaced by a much larger system wherein hundreds or thousands of unknowns appear.

Such a large system can, however, be handled, and the following discussion will describe one method which has been applied successfully to reconstructions involving up to 40 000 unknowns. While phrased in terms of the specific tomographic problem of microwave remote sensing for two-dimensional spatial distribution of liquid water, the physical characteristics of that problem do not play a crucial role, so the results may be interpreted in a more general context. We propose to extend them in later papers to include emission and passive sounding measurements.

The hypothetical, somewhat idealized, experiment

described in the next section is an "active" measurement, in which there is a source of radiation the position of which is changed to produce a different set of "rays" through the object being studied; it is therefore closely analogous, for example, to tomography procedures in medicine and other fields.

2. Idealized experiment

A hypothetical experiment may be described as follows: An aircraft or satellite, carrying a radiation source with a known nondirectional output, flies over a portion of the atmosphere within which the two-dimensional distribution of an absorbing material, $f(y, z)$ is to be inferred from measurements of the radiance received from that airborne source at several ground-based receivers, for several separate positions of the source. If attenuation of the radiation follows an exponential law, with absorption coefficient k m^{-1} per volume concentration of absorber, then a given measured intensity, I_m, implies $-k^{-1}\ln(I_m/I_0)$ total units of absorber along the entire ray path between source and receiver when I_0 is the intensity which would prevail in the absence of any absorption. If the quantity $-k^{-1}\ln(I_m/I_0)$ is defined as g_m, the result is the same equation that arises in many inversion problems:

$$g_m = \int_c^d \int_a^b K_m(z, y) f(z, y) dz dy,$$

$$m = 1, 2, \cdots M. \tag{1}$$

In practice, $f(z, y)$ may be changing with time, but it is assumed throughout that the M measurements are completed before any appreciable temporal change has occurred.

Since (1) represents a linear integral transform over a finite interval in z and y, the two-dimensional character of K_i and f is incidental. If (1) is expressed in terms of summation over a finite set of locations (z_1, y_1), (z_2, y_2), (z_3, y_3), etc., the resulting equation is essentially one-dimensional (and even more familiar):

$$g_i = \sum_j A_{ij} f_j. \tag{2}$$

Even though there is a complete algebraic analogy with atmospheric inversion problems, such as temperature sounding from infrared or microwave radiance (Kaplan, 1959; Houghton, 1961; Wark and Fleming, 1966; Twomey, 1966), or ozone distribution from backscattered ultraviolet (Singer and Wentworth, 1957; Twomey, 1961), there are nevertheless significant differences. First, a much larger number of unknowns is involved in the tomographic version, which, being two dimensional, requires N^2 unknowns for a resolution comparable to that given by N unknowns in a one-dimensional problem. Second, in the tomographic problem a given measurement is influenced only by $f(z, y)$ at locations along the relevant ray, so, for any m, there will be a large number of zeros in the set $A_{m,1}$, $A_{m,2} \cdots A_{m,N^2}$; In fact, the number of nonzero entries will be of order N—out of a total of N^2 elements. (This also has the consequence that the kernels contain more higher-frequency components than those for conventional one-dimensional sounding problems, in which $K_i(z)$, having the physical character of a transmission, or the derivative of a transmission, is necessarily monotonic and smooth in the variable z).

In the tomographic approach, the matrix **A** is very large, but quite sparse. Even keeping such a large matrix in storage could be difficult, while its direct inversion (or even multiplication) would be awkward and costly. That problem is avoided if the measurements are considered one at a time (then only one row of **A** is needed at a time) and used to improve an estimate of the unknown function f. Clearly, such a procedure must be iterative, since there is no one-to-one correspondence between any single measurement and any single horizontal or vertical location.

The following section will describe an iterative procedure (which has already proved its worth in one-dimensional inversion problems) and illustrate its application to tomographic problems. The discussion will be couched mainly in terms of a continuous variable x, which has a one-to-one relationship with spatial position (y, z). This has been done in the belief that the presentation is thereby made clearer; in practice, of course, x becomes replaced by a discrete subscript, and integrated quantities such as $\int K_i(x) f(x) dx$, being evaluated by quadrature, become discrete sums of the form $\sum_j w_{ij} K_{ij} f_j$, where j indicates one of N discrete spatial locations $(y_1, z_1), (y_2, z_2) \ldots (y_N, z_N)$, f_j represents the value taken there by the unknown $f(y, z)$, and w_{ij} is a quadrature coefficient.

3. Iterative nonlinear adjustment

As in any iterative procedure, we start with a first guess for the unknown distribution and adjust it measurement-by-measurement to improve the agreement between measured values and computed values. In a one-dimensional problem, where M "measured quantities" $g_1, g_2, \ldots g_M$ represent $\int K_1(x) f(x) dx$, $\int K_2(x) \times f(x) dx$, etc., the adjustment from iterate n to iterate $(n + 1)$ is produced (Twomey, 1977) as follows:

$$f^{(n+1)}(x) = f^{(n)}(x)[1 + \zeta K_n(x)]. \tag{3}$$

The choice of ζ depends on g_n and will be discussed in more detail later. The objective is to bring $\int K_n(x) f^{(n+1)}(x) dx$ closer to g_n; n can be as large as is necessary, since, after using all of the M measurements, the sequence $i = 1 \cdots M$ can be repeated, so that $K_n(x)$ for $n > M$ represents the kernel with subscript n modulo M. If we pass to a two-dimensional unknown $f(x, y)$, the generalization is trivial,

$$f^{(n+1)}(y, z) = f^{(n)}(y, z)[1 + \zeta K_n(y, z)], \tag{4}$$

and, in discrete notation, there is no difference between one- and two-dimensional unknowns since the N^2 elements of f are described by a one-dimensional array, f_j.

The equation

$$f_j^{(n+1)} = f_j^{(n)}[1 + \zeta A_{n,j}] \tag{5}$$

(in which for each successive n one runs through all values of j) describes the adjustment procedure, regardless of geometric dimension. Since negative concentrations are physically nonsense, a positive-definite constraint is desirable, and *some* kind of constraint is essential for stability. For their tomographic reconstruction, Warner et al. (1986) used the nonnegative least-squares algorithm of Lawson and Hanson (1974). Nonnegativity—an exigent, physically based requirement—can easily be imposed implicitly in iterations described by (3), (4) and (5). The nature of f and $\mathbf{A}$ in those equations assures the continued generation of nonnegative $f^{(n)}$, provided only that the first guess, $f^{(0)}$, contains only positive elements (a zero element in f would necessarily remain zero throughout all subsequent iterations) and that negative values of ζ are suitably limited (there can be no *negative* contributions to a measured g_m, so no element of $\mathbf{A}$ can be negative). With an iterative process, there is no uniquely "correct" or even "optimum" value for ζ so long as the iteration produces eventually a solution which, on insertion into (1) or (2), gives a set of g which is acceptably close to the measured set of g and does not vary dramatically in successive iterations. It is not even necessary that the errors $|\mathbf{g} - \mathbf{g}'|$ diminish monotonically at *every* step [here and elsewhere the prime symbol indicates quantities computed by inserting an iterate or solution for f in Eq. (1)].

If all ζ are sufficiently small, iterated solutions are simply linear combinations of the M kernels, and therefore contain no components orthogonal to all kernels. That property is advantageous for stability, since if a solution contained a component that was orthogonal to all the M kernels, that component would contribute nothing to any of the M measurements, and so could be made arbitrarily large or arbitrarily small (and of either sign) without altering the vector of measurements, $\mathbf{g}$—in other words, instability would result from such components. They must therefore either be eliminated totally or very heavily attenuated for a solution process to be stable. While such orthogonal components can result sooner when ζ is not restricted to very small values, the solution, although potentially less stable, will be acquired more rapidly, since larger adjustments are then permitted. Since higher frequencies are introduced with increasing iteration, it is possible, perhaps even likely, that after a sufficiently large number of iterations, a cyclic behavior can set in, causing successive iterates to differ seriously from each other. This has not been observed in the present study, but clearly, in practice, it is important to terminate iteration as soon as possible, i.e., before $|\mathbf{g} - \mathbf{g}'|$ becomes smaller than the uncertainty in $\mathbf{g}$ itself—from that point on, one would be attempting to invert the "noise."

It is possible to construct fairly complicated recipes for the choice of ζ (and for the iterative adjustment process), and some authors have taken that route. This writer is of the opinion that a simple adjustment process and a straightforward selection method for ζ is preferable, and here we restrict ourselves to such methods.

In work on one-dimensional problems, the following choice for ζ has often been used successfully (Twomey, 1975, 1977):

$$\zeta_n = \frac{g_i}{\int K_i(x) f^{(n)}(x)dx} - 1 \tag{6}$$

(the integration being, of course, approximated by a summation). [In other words: the estimate after n iterations, $f^{(n)}(x)$, is assessed by comparing $\int K_i(x) \times f^{(n)}(x)dx$ with g_i (which quantities should be equal for it to be a perfect solution). It is then updated by application of (3), which improves the comparison of g_i with $\int K_i(x) f^{(n)}(x)dx$, but does not necessarily make them exactly equal.] An alternative procedure is to select ζ_n so as to force $\int K_i(x) f^{(n+1)}(x)dx$ to equal g_i exactly. Since $f^{(n+1)}(x)$ is $[1 + \zeta_n K_i(x)] f(x)$, that would require:

$$\int K_i(x) f(x)dx + \zeta_n \int K_i^2(x) f(x)dx = g_i.$$

Hence,

$$\zeta_n = \left\{ g_i - \int K_i(x) f(x)dx \right\} \Big/ \int K_i^2(x) f(x)dx. \tag{7}$$

Application of (7) tends to diminish the difference vectors $\mathbf{g} - \mathbf{g}'$ faster than (6), but is more susceptible to instability and has not proved very useful in conventional sounding inversions. It was found to work well for the present tomographic problem, possibly because of the previously mentioned greater high-frequency content of tomographic kernels vis-à-vis one-dimensional kernels of smooth exponential character. To avoid generation of negative values, it was necessary when using (7) to check whether any iterated value $f_j^{(n+1)}$ became negative, and to correct such a condition whenever it occurred. With the first method [i.e., Eq. (6)], nonnegativity of $f(x)$ is assured, provided only that the first guess $f^{(0)}(x)$ is nonnegative and that the $K_i(x)$ are all nonnegative and scaled so that no $K_i(x)$ exceeds unity for any value of i or x. Then by (6), $\zeta = g_i / \int K_i(x) f^{(n)}(x)dx - 1$, so we have:

$$f^{(n+1)}(x) = f^{(n)}(x)[1 - K_i(x)] + g_i K_i(x) f^{(n)}(x) \left[\int K_i(x) f^{(n)}(x)dx \right]^{-1}.$$

Hence $f^{(n+1)}(x)$ cannot be negative if $f^{(n)}(x)[1 - K_i(x)]$

TABLE 1. Kernel values for $j = 1$ to N^2 if unknowns represent gridpoint values of $f(y, z)$ with trapezoidal distribution between grid points.

0.0333	0.0167	0	0	0	0	0	0	0	0
0.0167	0.0667	0.0167	0	0	0	0	0	0	0
0	0.0167	0.0667	0.0167	0	0	0	0	0	0
0	0	0.0167	0.0667	0.0167	0	0	0	0	0
0	0	0	0.0167	0.0667	0.0167	0	0	0	0
0	0	0	0	0.0167	0.0667	0.0167	0	0	0
0	0	0	0	0	0.0167	0.0667	0.0167	0	0
0	0	0	0	0	0	0.0167	0.0667	0.0167	0
0	0	0	0	0	0	0	0.0167	0.0667	0.0167
0	0	0	0	0	0	0	0	0.0167	0.0333

≥ 0, and so, for the iterated solutions by the first method (6) to remain nonnegative, it is sufficient that $K_i(x) < 1$ for all i and x. That condition can always be met by suitably scaling the problem.

As a first guess for $f(y, z)$, a spatially constant distribution was employed, with a value equal to a mean obtained by averaging the ray-path liquid-water totals $[-\mu_m k^{-1} \ln I_m/I_0]$, μ_m being the cosine of angle from the vertical for the mth ray. The iteration process was usually terminated after 50 cycles, i.e., $50 \times M$ individual steps; the number of measurements, M, was of the order of 100. In the case of the smoother assumed distributions, the r.m.s. magnitude of $\mathbf{g} - \mathbf{g}'$ decreased relatively rapidly to below one percent, and iteration could have been terminated earlier; however, continuing it to 50 cycles did not produce any harmful effects. In the case of the more discontinuous assumed distributions, $\mathbf{g} - \mathbf{g}'$ remained quite large (often of the order of ten percent) even after 50 cycles. In those cases, a stopping criterion based on $|\mathbf{g} - \mathbf{g}'|$ would have been wasteful in terms of computing time.

4. Sample inversion results

Since Warner et al. (1985) were concerned with tomographic reconstruction of liquid-water content of clouds, using microwave wavelengths at which absorption was predominantly due to liquid-water, the following examples will be discussed in those terms and so $f(y, z)$ will be envisaged to be the planar distribution of cloud liquid-water content with respect to height z and horizontal position y, in a slice through a three-dimensional cloud in the scanned (y, z) plane.

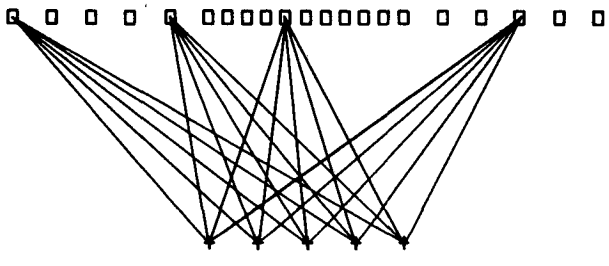

FIG. 1. One of several transmitter/receiver configurations used in the study. Five receivers (below) and 21 transmitter positions give a total of 105 rays. 20 of these have been drawn.

A numerical experiment proceeded as follows: A distribution $f(y, z)$ was concocted, and a set of ground-station locations and aircraft positions were selected. The integrated liquid-water amount along each of the rays connecting every possible pair of aircraft and ground locations were then computed. (The spatial distribution was envisaged to remain unchanged during the course of the measurements.) A typical set of ground and air positions is shown in Fig. 1, together with some—but not all—of the interconnecting rays (there were 105 in all). Other geometric configurations were also tested, but the results were not significantly different, provided the test region was sampled adequately. The kernels for the problem consist of sets of weights proportional to the contribution of a particular box or grid point to the ray. As an illustration: a ray slicing diagonally through a 10×10 area would pass through (1, 1), (2, 2), ... (10, 10) boxes. If the boxes were numbered sequentially from 1 to 100, that would mean zero weight for all boxes except those numbered 1, 12, 23, 34, 45, 56, 67, 78, 89, 100. If $f(y, z)$ was modeled by uniform distributions inside each box, the weights for each nonzero box would be 0.1414. On the other hand, if the unknowns represented gridpoint values of $f(y, z)$ and a trapezoidal distribution was assumed between grid points, the kernel takes values for $j = 1$ to N^2, as set out in Table 1.

Once the synthetic "measurements" $g_1, g_2 \ldots g_m$ had been calculated from the assumed distribution $f(x)$ and the kernels, a first guess was constructed and iteration according to (5) commenced. The first guess $f^{(0)}$ was a spatially constant liquid-water concentration obtained by dividing the sum of the g_i by the mean of the discrete kernel values A_{ij}. The following subsections

TABLE 2. Solution for a distribution with five concentric square shells each containing a uniform absorber concentration.

Region	f (correct)	Inversion			Standard deviation
		From	To	Mean	
1 (inner)	1000	970	1060	1020	45
2	800	700	870	790	64
3	600	460	670	550	66
4	400	310	470	376	4
5 (outer)	200	170	400	250	7

will discuss and briefly comment on three test distributions and their retrieval. Many other tests were run; those included here were neither best nor worst, but a reasonably representative sampling.

There are, of course, many ways of comparing nonidentical two-dimensional patterns, and assessments of degrees of similarity or difference often are highly subjective. Since any inversion procedure must ultimately be blind to high-enough frequencies (spatial frequencies in this case), reconstructions can be meaningful only in the sense of providing an averaged or smoothed version of the data field.

TABLE 3. Original asymmetric distribution in a square 10 × 10 array, showing two blocks with high liquid-water content in portions of the upper-left and lower-right quadrants; also, reconstructions in the inversions after 30 cycles as well as 50 cycles of iteration. The numbers give thousandths of a unit (1 g m^{-3}) and have been rounded to one thousandth.

(a) Original

100	200	200	500	500	500	500	200	200	100
200	200	3333	3333	3333	500	500	500	200	200
200	500	3333	3333	3333	800	500	500	500	200
200	500	3333	3333	3333	800	800	500	500	200
200	500	800	800	800	800	800	800	500	200
200	500	800	800	800	800	2222	2222	2222	200
200	500	500	800	800	800	2222	2222	2222	200
200	500	500	500	800	800	2222	2222	2222	200
200	200	500	500	500	500	500	500	200	200
100	200	200	500	500	500	500	200	200	100

(b) Inversion after 30 cycles

332	861	2013	2121	2095	542	590	381	281	219
234	765	2625	2402	2336	510	456	444	394	219
155	359	2956	3048	2926	414	405	457	641	316
112	131	2608	2805	2634	400	520	616	951	412
110	084	911	1668	1787	885	533	878	1120	495
174	176	404	425	838	806	1408	1360	2340	252
130	332	621	358	278	617	2795	1689	1852	097
174	461	433	513	940	185	1438	2467	926	136
172	245	874	500	1047	767	471	826	914	140
9	294	0	672	880	600	1204	335	572	137

(c) Inversion after 50 cycles

138	540	951	1840	2037	592	708	221	254	135
356	294	2763	2153	2060	630	463	532	181	176
217	303	3266	3330	3071	516	431	464	446	162
107	227	2849	2542	2917	678	684	501	593	240
96	451	1078	1468	1404	1101	905	861	544	198
134	577	969	882	731	681	2037	2084	2226	261
259	500	386	430	501	718	2400	2221	2026	134
202	414	246	372	745	478	1697	1924	1989	210
207	269	956	530	445	889	895	871	288	192
86	231	1	792	841	535	565	201	457	103

TABLE 4. Resolution testing using a 10 × 10 array, where alternate diagonal stripes are envisaged to contain either unity (1 g m^{-3}) or zero liquid-water content.* Numbers give thousandths of a unit and are rounded to one thousandth.

← •	•	•	•	•	•	•	•	•	•	• →
1075	(7)	854	(410)	480	(348)	485	(142)	1003	(3)	
(0)	979	(12)	400	(510)	410	(484)	758	(3)	999	
941	(0)	1026	(195)	804	(313)	918	(12)	969	(13)	
(0)	996	(0)	981	(370)	746	(49)	1026	(15)	966	
990	(0)	950	(0)	858	(236)	1104	(47)	1005	(7)	
(0)	1002	(0)	1044	(1)	961	(20)	1072	(4)	886	
961	(0)	1015	(0)	956	(1)	919	(4)	1081	(147)	
(35)	994	(0)	970	(0)	979	(0)	969	(81)	940	
871	(45)	982	(0)	1040	(0)	1044	(0)	868	(163)	
(119)	983	(32)	983	(0)	1021	(0)	978	(1)	890	
R		*R*		*R*			*R*		*R*	

* Original data: no parentheses, 1.0 gm^{-3}, parentheses, 0.0 gm^{-3}.

a. Case 1

This was a spatially smooth distribution resembling the "onion-skin" distribution of Warner et al. (1985). It consisted of concentric squares with 1.0 g m^{-3} (in the innermost), 0.8 g m^{-3} (in the region between that and the next), 0.6 g m^{-3} (in the next region), 0.4 g m^{-3} (in the next), and 0.2 g m^{-3} (in the outermost region). Inversion retrieved the central maximum values to better than 10%, values in the next region to within 13%, etc., out to the outermost (0.2 g m^{-3}) where the errors were largest, ranging from −12% to +105%. The mean values within each region and their standard deviations are tabulated in Table 2 (in this and following tables, values for f are rounded to three decimal places and written as a whole number of thousandths, e.g., 0.75 appears as "750", 1.005 as "1005", etc.). It is seen that both water-content levels and the spatial gradient inward were reproduced adequately by the inversion.

b. Case 2

This was an asymmetric distribution in which (as shown in Table 3a) there were two "blobs" with high liquid-water content, one of 3.3 g m^{-3} in a portion of the upper-left quadrant, the other of 2.2 g m^{-3} in a portion of the lower-right quadrant. Elsewhere the water content was 0.8, decreasing to 0.5 in the central part of the top and base and 0.2 at the sides and outer part of top and base. Qualitatively, these features were well reconstructed in the inversion, which is shown below after 30 cycles (Table 3b) and 50 cycles (Table 3c) of iteration. After 30 cycles, the qualitative features were clear, even though the rms size of the "error" $\mathbf{g} - \mathbf{g}'$ was still almost 10%; after a further 20 cycles of iteration, that error had been reduced to a little over 1/2% (i.e., what one would expect only in a very high-quality measurement).

c. Case 3

This was a fairly exacting test of resolution. In a square 10 × 10 array, alternate diagonal stripes were envisaged to contain either unity (1 g m^{-3}) or zero liquid-water content. The results are shown schematically in Table 4. The numbers again give thousandths of a unit and are rounded to one thousandth; parentheses indicate locations where the initial assumed pattern (i.e., the "correct result") contained zero. Overall, the pattern was recovered in a satisfactory manner. Of the 50 zeros in the original pattern, 18 values were returned that were very small, lying between zero and 0.001 g m^{-3}; 27 were ≤0.01, while values greater than 0.1 g m^{-3} were returned in 10 cases, the worst result for a correct value of zero being 0.51 g m^{-3}. For the 50 entries for which the correct result was 1.0 g m^{-3}, values between 0.9 and 1.1 were returned in 36 cases; the smallest value returned in place of a correct 1.0 was 0.40, and the largest 1.104.

5. Discussion

The results suggested that the nonlinear iterative algorithm could usefully be applied to the reconstruction of spatial patterns, such as were studied by Warner et al. (1985). A considerable advantage of the method, compared to more direct inversion methods, such as that used by Warner et al., is that computing time is (roughly) linear in the number of unknowns, whereas for matrix inversion time is proportional to between the square and the cube of that number. This is illus-

TABLE 5. Computing time tabulated against the number of unknowns for a standard matrix inversion and for the iterative procedure.

Number of unknowns	Computing time: Matrix inversion	Computing time: Iteration
20	0.017	0.20
50	0.145	0.39
100	1	1
150	3.15	1.32
200	—†	2.8

† Storage limitations did not permit inversion of a 200 × 200 matrix.

TABLE 6. Results obtained for hypothetical distribution case (*b*) when there were only 85 hypothetical measurements with 100 unknowns. The numbers give thousandths of a unit (1 g m^{-3}) and have been rounded to one thousandth.

127	373	712	1015	904	687	677	271	307	215
172	106	2489	2241	2584	252	480	518	301	83
129	358	3641	3735	3280	600	274	388	365	65
114	433	2990	2894	2976	897	490	372	442	199
602	605	1155	1147	1262	1200	1116	1168	704	177
589	452	706	620	605	775	2224	2156	2255	187
449	300	603	514	865	841	2205	2254	2092	464
126	571	526	576	805	733	2049	2235	1997	105
190	410	564	349	986	278	593	701	316	204
100	185	197	363	422	533	247	176	190	99

trated in Table 5, where computing time, normalized with respect to the 10 × 10 problem (100 unknowns), has been tabulated against the number of unknowns involved for a standard (IMSL) matrix inversion algorithm and for the iterative procedure described above. [Eqs. (5) and (6)].

The differences between the rates of increase of computing time for the two cases are rather striking. Furthermore, direct inversion could not be extended beyond 200 unknowns because of insufficient central memory (on a Cyber 76). This was not a problem with nonlinear iteration, which conserved memory in two ways. Firstly, only one row of the matrix was used at a time. Secondly, each row contained many zeros and only the nonzero elements were stored (together with pointers to their positions in the longer, original array), so that an abbreviated version of each row was stored in mass storage and recalled when needed.

a. Number of measurements vs number of unknowns

Warner et al. (1986, 1985) mentioned that the number of elements (i.e., unknowns) cannot exceed the number of measurements, since "obviously no retrieval is possible unless the number of elements is less than the number of beams." That restriction, however, would be necessary only if the unknowns were truly free to take any value. To stabilize solution processes in ill-conditioned inversion problems, it is always necessary to introduce constraints, and both our inversion methods and those of Warner et al. (1985) contain such constraints—explicit in the former, implicit in the latter. Such constraints represent an implicit smoothing, rejecting or strongly discriminating against higher frequencies. (The nonnegativity constraint used by Warner et al. (1985) cannot readily be described in terms of frequency; nevertheless, it serves to discriminate against high-frequency components and produces a degree of smoothing). To illustrate the practicality of inversion when $N^2 > M$, Table 6 shows results obtained for the hypothetical distribution (*b*), above, when there were only 85 hypothetical measurements for 100 unknowns.

The solution is comparable in quality to that shown earlier (Table 3) and clearly did not suffer from instability or blow-up because of algebraic singularity, even though there were more "unknowns" than measure-

TABLE 7. Results from an inversion applied to Table 6 distribution, first contaminated by adding a random ±2% noise.

120	360	727	1034	864	686	680	263	296	215
182	101	2407	2312	2572	255	465	512	300	84
138	362	3743	3717	3334	567	289	387	361	65
128	420	2982	2931	2990	971	492	359	450	222
574	623	1172	1146	1317	1160	1143	1208	742	174
562	454	702	588	671	767	2179	2131	2202	171
447	281	611	456	836	868	2222	2244	2084	468
118	581	535	582	809	746	2040	2303	2005	100
193	418	536	368	953	282	599	741	300	206
101	164	201	378	422	508	269	175	185	99

ments—the "unknowns" are in fact correlated spatially and so not truly independent.

b. *Measurement noise*

In earlier more conventional applications, the nonlinear iterative method has been found to be very robust in resisting spurious high-frequency components in the solution. (Such components often arise from inversion of "noise" in the measurement vector **g**.) A similar quality has been found in tomography. Table 7 shows, as an example, results from an inversion applied to the same vector **g** that was used for Table 6 contaminated by a random ±2% noise added before inversion. The results in Table 7 differ only trivially from those in Table 6 (or those in Table 2, which related to the same initial distribution). Even with ±50% noise added, solutions were obtained which, while very distorted, were stable and showed recognizable similarities to the initial distribution.

6. Conclusions

The nonlinear iterative method appears especially suitable for tomography problems, where there are likely to be several thousand unknowns, while the matrix and kernels for the problems are relatively sparse. Because of the nature of the process, it is never necessary to place the entire matrix of the problem in storage, and that greatly eases storage problems; furthermore, computation time can be significantly abbreviated by explicit recognition of the many zeros in the kernels. (The procedure would also seem to lend itself to parallel processing, but that aspect has not been examined in the present work.)

We have applied the method (in test problems, such as those discussed earlier) to solve for up to 40 000 unknowns without encountering any adverse effects in terms of stability, storage or computing time.

Acknowledgments. This work was supported by the Office of Naval Research under Grant N00014-85-K-0120. The final manuscript was edited by Margaret Sanderson Rae.

REFERENCES

Fleming, H. E., 1982: Satellite remote sensing by the technique of computed tomography. *J. Appl. Meteor.*, **21,** 1538–1549.

Houghton, J. T., 1961: Meteorological significance of remote measurements of infrared emission from atmospheric carbon dioxide. *Quart. J. Roy. Meteor. Soc.*, **87,** 102–104.

Kaplan, L. D., 1959: Inference of atmospheric structure from remote radiation measurements. *J. Opt. Soc. Amer.*, **49,** 1004–1007.

Lawson, C. L., and R. J. Hanson, 1974: *Solving Least-Squares Problems.* Prentice-Hall.

Singer, S., and R. C. Wentworth, 1957: A method for determination of vertical ozone distribution from a satellite. *J. Geophys. Res.*, **62,** 299–308.

Twomey, S., 1961: On the deduction of vertical distribution of ozone by ultraviolet spectral measurements from a satellite. *J. Geophys. Res.*, **66,** 2153–2162.

——, 1966: Indirect measurements of atmospheric temperature profiles from satellites. II: Mathematical aspects of the inversion problem. *Mon. Wea. Rev.*, **94,** 363–366.

——, 1975: Comparison of constrained linear inversion and an iterative nonlinear algorithm applied to the indirect estimation of particle size distributions. *J. Comput. Phys.*, **18,** 188–200.

——, 1977: *Introduction to the Mathematics of Inversion in Remote Sensing and Indirect Measurements.* Elsevier, Amsterdam.

Wark, D. E., and H. E. Fleming, 1966: Indirect measurements of atmospheric temperature profiles from satellites. I: Introduction. *Mon. Wea. Rev.*, **94,** 351–362.

Warner, J., J. F. Drake and P. R. Krehbiel, 1985: Determination of cloud liquid water distribution by inversion of radiometric data. *J. Atmos. Oceanic Technol.*, **2,** 293–303.

——, —— and J. B. Snider, 1986: Liquid water distributions obtained from coplanar scanning radiometers. *J. Atmos. Oceanic Technol.*, **3,** 542–546.