ISTITUTO NAZIONALE DI FISICA NUCLEARE

Sezione di Genova

INFN/TC-88/2

19 Gennaio 1988

M. Bertero:

LINEAR INVERSE AND ILL-POSED PROBLEMS

Servizio Documentazione
dei Laboratori Nazionali di Frascati

INFN - Istituto Nazionale di Fisica Nucleare
Sezione di Genova

INFN/TC-88/2
19 Gennaio 1988

LINEAR INVERSE AND ILL-POSED PROBLEMS

M. Bertero
Dipartimento di Fisica dell'Università and Istituto Nazionale di Fisica Nucleare, Via Dodecaneso 33, I-16146 Genova , Italy

CONTENTS

I. INTRODUCTION

In a paper published in 1902 on the boundary value problems for partial differential equations and their physical meaning, Jacques Hadamard introduces the basic concept of *well-posed problem* (Hadamard, 1902). In this first formulation he calls a problem well-posed when its solution is unique and exists for arbitrary (non-analytic) boundary values. Then, from the investigation of several examples related to elliptic and hyperbolic equations, he derives the conclusion that only the problems motivated by physical reality are well-posed. In particular, he shows that the Dirichlet problem for the Laplace equation and the Cauchy problem for the wave equation with boundary values at t=0 are well-posed, while the Cauchy problem for the Laplace equation and the Cauchy problem for the 3D wave equation, with boundary values at x=0, are not well-posed.

In subsequent work Hadamard emphasizes also the requirement of continuous dependence of the solution on the data (Hadamard, 1923). He claims that a solution which varies considerably for a small variation of the data is not actually a solution in the sense of physics. Physical data are never known exactly but only with a certain degree of accuracy and this should imply that the solution is not known at all. He provides a striking example of this fact in the case of the Cauchy problem for the Laplace equation in two variables

$$\Delta u \equiv u_{xx} + u_{yy} = 0 \qquad (1)$$

If we consider the following Cauchy data at y=0

$$u(x,0) = 0 \quad , u_y(x,0) = n^{-1} \sin(nx) \qquad (2)$$

then there exists a unique solution of Eq.(1) satisfying these conditions, which is given by

$$u(x,y) = n^{-2} \sin(nx) \sinh(ny) \quad . \qquad (3)$$

The factor sin(nx) produces a fluting of the surface which represents the solution of the problem and this fluting, however imperceptible near y=0, becomes enormous at any finite distance from the x-axis, provided the fluting be taken sufficiently small by taking n sufficiently large. Notice that when $n \longrightarrow \infty$, the amplitude of the oscillating data tends to zero while the frequency tends to infinity. This is now a classical example used for illustrating the effects which arise when the dependence of the solution on the data is not continuous.

In honour of this contribution a problem is now called *well-posed in the sense of Hadamard* if it has the property of continuous dependence of the solution on the data, even if the complete definition, containing the three requirements of uniqueness, existence and continuity, was firstly given by Courant and Hilbert (Courant and Hilbert, 1964, pg.227; Hadamard, 1964, pg.28).

The ideas outlined above are certainly related to a deep thougth of Hadamard. "He liked to work with the rigor of a mathematician and the practical sense of a physicist, and liked to repeat Poincaré 's words: *La Physique ne nous donne pas seulement l'occasion de résoudre des problémes,. . . , elle nous fait pressentir la solution"* (Mandelbrojt and Schwartz, 1965). The physics of Hadamard was, however, the physics of the nineteenth century. In fact, the requirements of existence, uniqueness and continuity of the solution were "deeply inherent in the ideal of a unique, complete, and stable determination of physical events. . . Laplace's vision of the possibility of calculating the whole future of the physical world from complete data of the present state is an extreme expression of this attitude" (Courant and Hilbert, 1962, pg.230).

The idea of well-posedness was extremely useful in the theory of partial differential equations and in functional analysis. A negative consequence was that, for years, the problems which are not well-posed,

and which now are called ill-posed or incorrectly posed, were just considered as mathematical anomalies and, for this reason, they were not seriously investigated. Recent developments in physics and especially in applied physics have shown that ill-posed problems can also be related to extremely important physical situations. For example, the solution of the Cauchy problem for elliptic equations may have interesting applications in electrocardiography (Colli Franzone *et al*, 1977), namely in the reconstruction of the epicardial potential from body surface maps. We want to emphasize, however, another ill-posed problem which has revolutionized diagnostic radiology.

In 1971 the first clinical machine for the detection of head tumors, based on a new X-ray technique called computer-assisted tomography or also *computerized tomography* was installed at the Atkinson Morley's Hospital, Wimbledon. In 1979 Allan M. Cormack and Godfrey N. Hounsfield awarded the Nobel prize in Medicine for the invention of this technique. As everybody knows, computerized tomography provides images of cross sections of the human body by measuring the attenuation of the X-rays along a large number of lines through the cross section. Then the processing of the data requires the reconstruction of a function of two variables from the knowledge of its line integrals. The solution of this mathematical problem was already contained in a paper of Johann Radon (Radon, 1917). To be precise, the result of Radon was more general, since he proved formulas for the reconstruction of a function on $\mathbb{R}^n$ from the knowledge of its integrals over all the hyperplanes of $\mathbb{R}^n$. In honour of this contribution, the mapping which transforms a function into the set of its integrals over hyperplanes is now called *Radon transform* and therefore the problem of tomography is just a special case of Radon transform inversion. Moreover, Radon transform inversion is a beautiful example of an ill-posed problem since the solution does not exist for arbitrary data and the dependence of the solution on the data, in general, is not continuous. As a consequence, the effect of the noise on the solution is amplified in a way similar to that discussed in the case of the Cauchy problem for the Laplace equation.

Even if Cormack and Hounsfield were not aware of the work of Radon (Hounsfield has also been quoted as saying "I find I've got other tools of thinking than math." (Di Chiro and Brooks, 1979)), there is no doubt that mathematics has given important contributions to the development and improvement of computerized tomography. On the other hand computerized tomography has certainly stimulated the fast growth of the theory of ill-posed problems in the last twenty years.

Radon transform inversion is an example of an *inverse problem*. A precise mathematical definition of inverse problems however, without reference to physics, is quite difficult, perhaps impossible. For a mathematician the distinction between direct and inverse problems is quite arbitrary. "We call two problems *inverses* of one another if the formulation of each involves all or part of the solution of the other. Often, for historical reasons, one of the two problems has been studied extensively for some time, while the other is never and not so well understood. In such cases, the former is called the *direct problem*, while the latter is called the *inverse problem* " (Keller, 1976). Another quote from the same paper shows a route for a characterization of inverse problems based on physics rather than on mathematics. " The main sources of inverse problems are science and engineering. Often these problems concern the determination of the properties of some inaccessible region from observations on the boundary of the region"(Keller, 1976).

In fact inverse problems are related to indirect measurements, remote sensing and so on and one must specify for each domain of physics the definition of the direct problem and the definition of the corresponding inverse problem. For example, in classical mechanics, the direct problem is the determination of the trajectories of the particles from the knowledge of the forces. Then the inverse problem is the determination of the forces from the knowledge of the trajectories (in this sense Isaac Newton has perhaps solved the first inverse problem). In potential theory, the direct problem is the determination of the potential generated by a known mass or charge or current distribution while the inverse problem is the determination of the mass or charge or current distribution from the

measured values of the potential. In the theory of wave propagation, the direct problem is the determination of fields distributions in time and space from given constitutions of sources or scatterers. Then the inverse problem is the determination of the characteristics of the sources or scatterers from the observation of the fields (computerized tomography is an inverse problem in this sense since it consists in the determination of the space distribution of the X-ray absorption coefficient from the observation of the attenuation of the X-rays passing through the probe). In instrumental physics the direct problem is the computation of the output of a given instrument, with a known impulse response function, from the knowledge of the input. Then the inverse problem is the determination of the input from the knowledge of the output and so on.

Having these examples in mind one can accept a definition of direct and inverse problems claiming that direct problems are problems oriented along a cause-effect sequence or, in other words, problems which consist in providing the consequences of given causes, while inverse problems are those associated with the reversal of the chain of casually related effects and therefore these problems consist in finding the unknown causes of known consequences (Turchin *et al*, 1971). This definition is not misleading if one never forgets that the formulation of any specific problem must be based on well-established physical laws and that physics must specify what is a cause and what is an effect as well as provide the equations relating the effects to the causes. Moreover, a merit of this definition is the attempt of justifying the fact that direct problems are, in general, well-posed while the corresponding inverse problems are, in general, ill-posed.

Nowadays inverse problems are fundamental in several domains of applied science: medical diagnostic, atmospheric sounding, radar and sonar target estimation, seismology, radioastronomy, microscopy and so on. Everyday applications are in use not only in X-ray or MR tomography, but also in seismic data processing for geophysical exploration and in radiometric data processing for meteorological forecasts and monitoring.

Moreover, in the last years, several books have been published on inverse problems related to various areas of applied physics: optics (Baltes,1979;1980), astronomy (Craig and Brown,1986), electromagnetics (Boerner *et al*,1983), atmospheric sounding (Zuev and Nats,1983), computerized tomography (Herman,1979;1980; Herman and Natterer,1981; Natterer,1986 a). Miscellaneous examples and mathematical results are contained in (Colin, 1973; Sabatier, 1978; 1987; Talenti, 1986; Cannon and Hornung, 1986). Ill-posed problems for partial differential equations are investigated in (Payne, 1975; Carasso and Stone, 1975).

Inverse problems, in general, are nonlinear. Two nonlinear problems have been the object of beautiful mathematical investigations, namely the *inverse Sturm-Liouville problem* and the *inverse scattering problem*. The first one is perhaps introduced by Lord Rayleigh (Lord Rayleigh, 1877) which, in describing the vibrations of strings of variable density, briefly discusses the possibility of deriving the density distribution from the frequencies of vibration. A modern version and generalizations are given by Kac (Kac, 1966). Roughly speaking the mathematical problem is the determination of the coefficients of a differential operator from the knowledge of its eigenvalue spectrum. Important contributions to this problem were given by outstanding mathematicians such as Levinson, Marchenko and Krein among others. A short survey is given in (Barcilon, 1986).

The inverse scattering problem was originally investigated in connection with Schroedinger equation and later extended to other equations such as Helmholtz equation, impedance equation and so on. In short, the problem is the determination of the potential (or of the refraction index, of the impedance, etc.) from the knowledge of quantities related to the scattering amplitude such as phase shifts, reflection and transmission coefficients, etc. A complete account of the main results is given in (Chadan and Sabatier, 1977).

The inverse problems above are nonlinear. For example, in the case of the vibrating string, the direct problem consists in solving the linear eigenvalue problem

$$- u''_k(x) = \omega_k^2 \rho(x) u_k(x) \; ; \qquad k = 0,1, \ldots \tag{4}$$

on the interval (0,L), for a given density function $\rho(x)$ and suitable boundary conditions at the end points 0 and L. But, since the eigenvalues ω_k^2 are nonlinear functionals of the density function $\rho(x)$, the inverse problem, consisting in the determination of $\rho(x)$ from given values of the ω_k^2, implies the inversion of a nonlinear mapping. However, when an estimate of $\rho(x)$, say $\rho_0(x)$, is known and the difference $f(x) = \rho(x) - \rho_0(x)$ is small, it is possible to linearize the functionals ω_k^2 using, for instance, perturbation theory. In this way the original nonlinear inverse problem is approximated by means of a linear one, the unknown function being now f(x). Similar results can be obtained in inverse scattering using Born or Rytov approximation, geometrical optics approximation and so on. Moreover some inverse problems related to inverse diffraction are rigorously linear problems.

In this paper we consider only *linear inverse problems* which have the following general structure. The first step is the definition of the direct problem, which must be linear. Then the solution of the direct problem defines a linear mapping L from the space X of all the functions characterizing the properties of the physical sample (such as the density function in the case of a vibrating string or the refraction index in the case of a semi-transparent object etc.) into the space Y of all the corresponding measurable quantities (such as sequences of eigenvalues, scattering amplitudes and so on). Of course in the direct problem the data are elements of X while the solutions are the elements of Y. In the corresponding inverse problem there is an exchange between data and

solutions. If we assume that the operator L is known (and this implies the need of solving the direct problem), then the inverse problem can be formulated as follows: given $g \in Y$ and given the linear operator $L : X \longrightarrow Y$, find $f \in X$ such that

$$g = Lf \quad . \tag{5}$$

An element of X will be called an *object* (or a *solution*) while an element of Y will be called an *image* (or a *data function*, a *data vector*, etc.). Accordingly X will be called the *object space* (or the *solution space*) and Y will be called the *image space* (or the *data space*).

In this paper, as in most of the mathematical papers on these problems, it is assumed that both X and Y are Hilbert spaces and that the linear operator $L : X \longrightarrow Y$ is continous. Continuity means that, given any sequence $\{f_n\}$ converging to zero in the sense of the norm of X, the corresponding sequence $\{Lf_n\}$ converges to zero in the sense of the norm of Y. In other words, it is assumed that the direct problem is well-posed in the sense of Hadamard.

In order to illustrate the difficulties of solving linear inverse problems, we consider a particular example of Eq.(5), namely the case of a *first kind Fredholm integral equations* of the following type

$$g(x) = \int_a^b K(x,y)\, f(y)\, dy \quad , \quad c \leq x \leq d \quad . \tag{6}$$

This can be written in the general form (5) if we introduce the integral operator

$$(Lf)(x) = \int_a^b K(x,y)\, f(y)\, dy \quad , \quad c \leq x \leq d \quad , \tag{7}$$

which transforms functions on [a,b] into functions on [c,d]. Moreover we consider the simple case where the solution and data space are spaces of square integrable functions, i.e. $X = L^2(a,b)$ and $Y = L^2(c,d)$. Then the operator L is continuous when the integral kernel K(x,y) is square integrable. Notice that this is only a sufficient condition for continuity. In the case of a convolution operator, i.e. $K(x,y) = K(x-y)$ and also $a=c=-\infty$, $b=d=+\infty$, the operator L is continuous in $L^2(\mathbb{R})$ when the impulse response function K(x) is integrable and therefore the kernel is not integrable nor square integrable as a function of the two variables x,y.

Assume now, for example, that the integral kernel K(x,y) is an analytic function of x for any $y \in [a,b]$. Then, given an arbitrary object $f \in L^2(a,b)$, the corresponding image g(x), computed by means of Eq.(5), is also analytic. It follows that the inverse problem does not have a solution for an arbitrary $g \in L^2(c,d)$ but only for functions g in a subset of analytic functions. The problem is ill-posed in the original sense of Hadamard.

Also continuous dependence of the solution on the data does not hold true. If the interval [a,b] is bounded and the kernel is square integrable, then the kernel is also integrable and, from the Riemann-Lebesgue theorem (Titchmarsh, 1948) it follows that

$$\int_a^b K(x,y)\ \cos(ny)\, dy \longrightarrow 0\ , \qquad n \longrightarrow \infty\ . \tag{8}$$

In this way it has been found a sequence of functions which does not tend to zero in $L^2(a.b)$ while the sequence of the corresponding images tends to zero. This example is similar to the example of Hadamard in the case of the Cauchy problem for the Laplace equation.

The previous example indicates that it is necessary to reconsider the validity of the mathematical model provided by Eq.(5) for the description of a physical experiment. When we apply the operator L to all the functions of X we obtain a set of images which can be called the set of the *exact images* or the set of the *noise free images* . In the example discussed above these functions are much more regular (analytic) than the

corresponding functions f and this situation is rather common in most inverse problems, i.e. the operator L, which solves the direct problem, has rather strong smoothing properties. But measurement errors or noise can destroy the smoothness of the exact image. The measured image is not related to f by Eq.(5) but by the following one

$$g = L f + h \tag{9}$$

where h is a stochastic function which represents the random noise. Such a function is, in general, much more irregular than any exact image Lf. For example, Lf is band-limited while h is not band-limited or has a band much broader than the band of Lf.

The remarks above indicate that the space Y must be broad enough to contain both the exact and measured images and that, in general, the set of exact images is a subset of Y. Moreover, if we assume that Y is a Hilbert space, then we must also assume that it is equipped with a scalar product such that the norm of the experimental error h is small with respect to the norm of the exact image Lf (for this reason a space of square integrable functions is the most convenient in several practical problems). As a consequence of these properties of the image space Y, the solution of Eq.(5) does not exist for arbitrary g and the problem is ill-posed. In the case of inverse problems with discrete data, however, the solution may exist for arbitrary g (even if it is not unique - see Chapt.III). It also depends continuously on the data since the space Y is a finite dimensional euclidean space and any operator is continuous in a finite dimensional space. One must never forget, however, that continuous dependence of the solution on the data is a necessary but not sufficient condition for the numerical stability (robustness) of the solution. Finite dimensional problems obtained by discretizing ill-posed problems are usually ill-conditioned, even extremely ill-conditioned, so that error propagation from the data to the solution can deprive the solution of any physical meaning.

If we come back now to a linear inverse problem formulated in

Hilbert spaces, we have a puzzling situation: or we use Eq.(5), but in this case solution may not exist, or we use Eq.(9), but in this case we have only one equation and two unknown functions, namely f and h. Also the idea of looking for approximate solutions of Eq.(5), i.e. for objects f such that Lf is approximately equal to g, is not successful due to the non-continuous dependence of the solution on the data. In fact, as clearly illustrated by the example of Hadamard, the set of the approximate solutions can contain wildly oscillating and completely unphysical functions. More precisely the set of all the functions f, such that the distance between Lf and g is not greater than a prescribed small quantity ϵ, is an unbounded set of the object space X.

Then the basic idea in the treatment of ill-posed problem is the use of *a priori* information about the unknown object in order to restrict the class of approximate solutions. This means that we need some additional informations, namely informations which cannot be deduced from Eq.(5) or Eq.(9), about properties of the unknown object f and that these informations must be inserted in the algorithm in order to produce a *physically meaningful* approximate solution. The additional informations can consist of upper bounds on the solution and/or its derivatives, of regularity properties of the solution (existence of derivatives up to a certain order, analyticity, etc.), of localization properties of the solutions (restrictions on its support, behaviour at infinity etc.), of lower bounds on the solution and/or its derivative (positivity of the solution and/or of its first derivative), and so on.

The idea of using prescribed bounds in order to produce approximate stable solutions was introduced by Pucci in the case of the Cauchy problem for the Laplace equation (Pucci, 1955), while the constraint of positivity was used by John in the solution of the heat equation for preceding times (John, 1955), another classical example of ill-posed problem. A general version of similar ideas was independently formulated by Ivanov in the case of Eq.(5) (Ivanov, 1962). This method and the method of Phillips for first kind Fredholm integral equations (Phillips, 1962) were the first examples of the *regularization theory for ill-posed problems*, which was

a little bit later formulated and developed by Tikhonov (Tikhonov, 1963 a; 1963 b; 1964). Expositions of this theory can now be found in a series of monographs (Tikhonov and Arsenine, 1977; Groetsch, 1984; Bertero, 1982; Morozov, 1984). A summary of this theory will be given in Chapt.V. This will be preceded by a short Chapter on the theory of generalized solutions which is the basis of the general formulation of regularization theory.

Regularization theory is essentially deterministic in the sense that statistical properties of the noise or of the objects are not used. Probabilistic methods for the solution of ill-posed problems have also been developed but they will not be considered in this paper. Some of them are *ad hoc* methods and have not yet been formulated in a sound mathematical framework. We only mention here the so-called method of *Wiener filters* which has been developed as a general method for solving ill-posed problems (Strand and Westwater, 1968; Franklin, 1970; Bertero and Viano, 1978) and which can be considered as a probabilistic version of a particular regularization algorithm (Bertero *et al.* , 1980). This method, of course, is very well established and it can be used whenever a knowledge of the statistical properties not only of the noise but also of the object is available. The last kind of information plays, in a sense, the same role as the prescribed bounds on the solution in the case of regularization theory and therefore it is the required *a priori* information in this formulation (Turchin *et al*, 1971). The difficulty in the use of Wiener filters is that, in several practical inverse problems, the required correlation functions are not known. When these can be determined, the method can be very useful as demonstrated by the applications to atmospheric remote sensing (Askne and Westwater,1986).

II. LINEAR INVERSE PROBLEMS

According to the definition introduced by Courant and Hilbert, the problem of solving the functional equation (5) is *well-posed* in the sense of Hadamard if the following conditions are satisfied:

i) the solution f is unique in X;

ii) the solution $f \in X$ exists for any $g \in Y$;

iii) the inverse mapping $g \longrightarrow f$ is continuous.

Conditions i) implies that the operator $L : X \longrightarrow Y$ admits an inverse operator $L^{-1} : Y \longrightarrow X$ while condition ii) means that L^{-1} is defined everywhere on Y. Then, since the continuous operator L is linear, from a corollary of Banach *open mapping theorem* (Yosida, 1966, pg.77) it follows that L^{-1} is also continuous and therefore, in the linear case, conditions i) and ii) imply condition iii).

We emphasize that the requirement of continuous dependence of the solution on the data is a necessary but not sufficient condition for the stability (robustness) of the solution against noise. In the case of a well-posed problem, the propagation of relative errors from the data to the solution is controlled by the *condition number* . If δg is a small variation of g and δf the corresponding variation of $f = L^{-1}g$, then

$$\| \delta f \|_X \, / \| f \|_X \leq \text{cond}\,(L) \, \| \delta g \|_Y \, / \| g \|_Y \qquad (10)$$

where cond (L) is the condition number given by

$$\text{cond}\,(L) = \| L \| \, \| L^{-1} \| \geq 1 \qquad (11)$$

Here $\| L \|$ and $\| L^{-1} \|$ denote the norms of the continuous operators L and L^{-1} respectively.

When cond(L) is not too large, the problem (5) is said to be *well-conditioned* and the solution is stable with respect to small variations of the data. On the other hand, when cond(L) is very large the problem is said to be *ill-conditioned* and a small variation of the data can produce a completely different solution. It is clear that the separation between well-conditioned and ill-conditioned problems is not very sharp and that the concept of well-conditioned problem is more vague than the concept of well-posed problem.

As we have announced in the Introduction, one of the conditions i)-iii) may not be satisfied when Eq.(5) is the mathematical model of a linear inverse problem. In such a case the problem is said to be *ill-posed* . According to this definition inverse problems with discrete data are always ill-posed because uniqueness never holds true (see Chapt. III). In the recent mathematical literature however, the term ill-posed is used in a more restrictive sense which is related to the theory of generalized solutions. Since the generalized solution, when it exists, is always unique (see Chapt. IV), the problem of solving Eq.(5) is called ill-posed only when the generalized solution does not exist for arbitrary g or, equivalently, it does not depend continuously on the data. It follows that only problems formulated in infinite dimensional spaces can be ill-posed in this sense, while problems with discrete data are always well-posed (but, may be, ill-conditioned).

In the next Sections we first investigate some general properties of Eq.(5) in the ill-posed case and successively we present some examples of linear inverse problems which are the most significant in our opinion.

A. General Properties

The ill-posedness of a problem is a property of the triple {L, X, Y}: the problem is ill-posed because, for instance, the space Y is too broad. As already discussed in the Introduction, however, the space Y cannot be

modified since it must be broad enough to contain both exact and measured (noisy) data. Therefore if we know that in an inverse problem the set of measured data does not coincide with the set of exact data, condition ii) is not satisfied.

As concerns the object space X, in many practical circumstances it is quite natural to assume that it is a space of square integrable functions. It may be convenient, however, to adapt the structure of X to the available *a priori* information about properties of the solution such as regularity properties (existence of derivatives of f up to a certain order, analyticity, etc.) and/or localization properties (approximate size of the support of f, asymptotic behaviour at infinity, etc.). This means that in the case of functions f defined on some interval [a,b], the appropriate Hilbert space can be a *weighted Sobolev space* whose scalar product is defined by

$$(f,\phi)_X = \sum_{k=0}^{m} \int_a^b p_i(x)\, f^{(i)}(x)\, \phi^{(i)}(x)^*\, dx \qquad (12)$$

Here the star denotes complex conjugation, the p_i are given functions which are continuous and positive and $f^{(i)}$ means $d^i f/dx^i$.

A more general description of the process which consists in restricting the object space by the use of *a priori* knowledge can be the following. Let X be the original Hilbert space where the problem (5) is formulated (for example a space of square integrable functions) and let us assume that the *a priori* information about the specific solutions we are considering can be described as follows: the solutions of the problem belong to the domain of an operator C from the Hilbert space X into another Hilbert space Z (the operator C is called the *constraint operator* and the space Z is called the *constraint space*) with the following properties :

a) the operator C has a domain D(C) dense in X and is closed (the importance of this property is based on the fact that, as a rule, all

differential operators are closed; for the definition see, for example, (Balakrishnan, 1976))

b) the operator C has a continuous (bounded) inverse C^{-1}.

Then it is possible to introduce a new space $X_C \subset X$, which is the domain of the operator C equipped with the scalar product

$$(f,\phi)_C = (Cf, C\phi)_Z \quad . \tag{13}$$

It is easy to prove that, when conditions a) and b) are satisfied, X_C is also a Hilbert space. Moreover the operator $L : X_C \longrightarrow Y$ is also continuous because the topology of X_C is stronger than the topology of X. It follows that the inverse problem can be reformulated by taking X_C as the new object space. We just notice that the scalar product (12) is a particular case of the scalar product (13).

As stated in the Introduction, we will consider Eq. (5) only in the case where L is a linear and continuous operators. Therefore, for the sake of completeness, we summarize here a few general properties of these operators which will be frequently used in the following. We also indicate their physical interpretation.

The *null space* of L, denoted by N(L), is the set of all the functions f which annihilate L, i.e.

$$N(L) = \{ f \in X \mid Lf = 0 \} \quad . \tag{14}$$

When the linear operator L is continuous, N(L) is a closed linear subspace of X. Moreover, N(L) is not trivial if and only if the inverse operator does not exist. The null space of the operator L will also be called the subspace of the *invisible objects* (Rust and Burrus, 1972), since the image of any

element of N(L) is exactly zero. This also means that the experiment, or the instrument, described by the operator L is unable to detect the objects which belong to N(L). On the other hand the orthogonal complement of N(L), $N(L)^{\perp}$, will be called the subspace of the *visible objects* , since the objects which belong to this subspace can be recovered, in principle, from exact data. Thanks to the orthogonal projection theorem (Yosida, 1966; pg. 82), an arbitrary object f can be uniquely represented as the sum of a visible plus an invisible component. It is evident that only the visible component can be at most recovered from the image in the absence of further *a priori* information about the object, .

The *range* of the operator L, denoted by R(L), is the set into which L maps X

$$R(L) = \{g \in Y | g = Lf, f \in X\} \tag{15}$$

and therefore R(L) is the linear subspace of the exact or noise free images (data). The distinction, discussed in the Introduction, between exact and measured images makes clear that R(L) in general does not coincide with Y. Moreover, in the case of inverse problems formulated in infinite dimensional spaces, R(L) may not be a closed subspace of Y. In some cases (for example in the case where L is the Laplace transformation) R(L) is dense in the image space Y.

The adjoint operator L^* is uniquely defined by the following relation

$$(Lf, g)_Y = (f, L^*g)_X \tag{16}$$

which holds for any $f \in X$ and $g \in Y$. The operator L^* is also linear and continuous and it has the same norm as L : $\|L^*\| = \|L\|$. The null space and the range of L^* will be denoted by $N(L^*)$ and $R(L^*)$ respectively.

For example, in the case of the integral operator (7), if we take X = $L^2(a,b)$ and Y = $L^2(c,d)$, the adjoiont operator is given by the following

equation

$$(L^{*}g)(y) = \int_{c}^{d} K(x,y)^{*}\, g(x)\, dx \quad , \qquad a \leq y \leq b \quad . \tag{17}$$

Finally we recall two important relations between the null spaces and the ranges of the operators L and L^{*}, and precisely

$$\overline{R(L)} = N(L^{*})^{\perp} \quad , \qquad \overline{R(L^{*})} = N(L)^{\perp} \tag{18}$$

where $\overline{R(L)}$ denotes the closure of R(L). In other words the investigation of $R(L^{*})$ can allow the determination of the subspace of the visible objects while the investigation of $N(L^{*})$ provides informations about the subspace of the exact images. These properties will be used in some of the examples of inverse problems discussed in the next Sections. The relations (18) imply the following decompositions of the spaces X and Y

$$X = N(L) \oplus \overline{R(L^{*})} \quad , \qquad Y = N(L^{*}) \oplus \overline{R(L)} \quad . \tag{19}$$

We shortly discuss now two important examples of linear operators whose range is not closed. The first is that of a compact operator (Balakrishnan, 1976) whose range is not finite-dimensional (i.e. we exclude the case of the so-called finite rank operators). An example of a compact operator is provided by the integral operator (7) when the kernel K(x,y) is a square integrable functions of the two variables x,y. More precisely, in such a case, L is a Hilbert-Schmidt integral operator from $L^{2}(a,b)$ into $L^{2}(c,d)$.

We consider the general case where the solution and data space do not coincide (for example, in Eq.(7) the intervals [a,b] and [c,d] do not coincide) and we introduce the so called *singular value decomposition* of

the operator L. We then indicate the relationship with the well known *spectral representation* of a compact, self-adjoint operator.

The operators L^*L and LL^* are compact, nonnegative operators in X and Y respectively. Moreover $N(L^*L) = N(L)$ and $N(LL^*) = N(L^*)$. It follows (Balakrishnan, 1976) that both operators admit a countably infinite set of positive eigenvalues. It is also easy to prove that L^*L and LL^* have exactly the same positive eigenvalues with the same finite multiplicity (Lanczos, 1961 ; Kato, 1966). If we denote these eigenvalues by σ_k^2 and if we count each eigenvalue as many times as required by its multiplicity, the σ_k^2 can be ordered in such a way as to form a non-decreasing sequence : $\sigma_0^2 \geq \sigma_1^2 \geq \sigma_2^2 \geq \ldots$ The compactness of L implies that

$$\lim_{k\to\infty} \sigma_k^2 = 0 \quad . \tag{20}$$

Now let u_k and v_k be the eigenfunctions of L^*L and LL^* respectively, associated with the same eigenvalue σ_k^2

$$L^*Lu_k = \sigma_k^2 u_k \ , \ LL^* v_k = \sigma_k^2 v_k \quad . \tag{21}$$

Then the u_k form an orthonormal basis in $N(L)^{\perp}$, i.e. the subspace of the visible objects, while the v_k form an orthonormal basis in $N(L^*)^{\perp}$, i.e. the closure of the subspace of the exact images.

As it is easy to verify, it is always possible to choose the pair $[u_k,v_k]$ in such a way that it is a solution of the following shifted eigenvalue problem

$$Lu_k = \sigma_k v_k \ , \ L^*v_k = \sigma_k u_k \quad . \tag{22}$$

The positive numbers σ_k are called the *singular values* of the operator L and the functions u_k, v_k the corresponding *singular functions*. The set of the triples $\{\sigma_k; u_k, v_k\}$ is the *singular system* of L. Then one can prove the following representation (Kato, 1966)

$$Lf = \sum_{k=0}^{+\infty} \sigma_k (f, u_k)_X v_k \qquad (23)$$

which is called the *singular value decomposition* of the compact operator L. A similar representation holds true for the operator L^*, which is obtained from Eq.(23) just by exchanging the role of the singular functions u_k and v_k.

The singular value decomposition (23) implies that the visible components of the object corresponding to small singular values give a small contribution to the corresponding components of the exact image. Therefore, in the case of a noisy image, these components can also be invisible in practice. Moreover, Eq. (23) implies that an image g is an exact image, i.e. $g \in R(L)$, if and only if the following conditions are satisfied

$$g \in N(L^*)^{\perp} \quad , \quad \sum_{k=0}^{+\infty} \sigma_k^{-2} |(g, v_k)_Y|^2 < +\infty \quad . \qquad (24)$$

These conditions are also called *Picard's conditions* (Nashed, 1976) for the existence of the solution of Eq.(5), since they were first derived by Picard in the case of first kind Fredholm integral equations (Picard, 1910). They show that R(L) is not closed, since an arbitrary function satisfying the first of the conditions (24) does not satisfy necessarily the second one. It is evident, however, that R(L) is dense in $N(L^*)^{\perp}$.

When the operator L is self-adjoint and non-negative (the integral operator (7) is self-adjoint when [a, b] = [c, d] and also $K(x, y)^* = K(y, x)$),

let us denote by λ_k its positive eigenvalues and by u_k the corresponding eigenfunctions. Then we have $\sigma_k = \lambda_k$, $v_k = u_k$, and Eq. (23) becomes the *spectral representation* of the operator L.

The second example is that of a convolution operator in $L^2(\mathbb{R}^n)$

$$(Lf)(\mathbf{x}) = \int_{\mathbb{R}^n} K(\mathbf{x} - \mathbf{y})\, f(\mathbf{y})\, d\mathbf{y} \quad , \quad \mathbf{x} \in \mathbb{R}^n \tag{25}$$

where $\mathbf{x} = \{ x_1, x_2, \ldots, x_n \}$ and $d\mathbf{y} = dy_1 dy_2 \ldots dy_n$. If the impulse response function $K(\mathbf{x})$ is integrable, then its Fourier transform $\hat{K}(\xi)$ is bounded and continuous and from the Riemann- Lebesgue theorem (Titchmarsh, 1948) it also follows that $|\hat{K}(\xi)| \longrightarrow 0$ when $|\xi| \longrightarrow \infty$. As a consequence, if Ω is the set of the points where $|\hat{K}(\xi)| \neq 0$ (the closure of Ω is the support of $\hat{K}(\xi)$), we have

$$m = \inf \{ |\hat{K}(\xi)| \mid \xi \in \Omega \} = 0 \tag{26}$$

When Ω is a bounded subset of $\mathbb{R}^n$, the function $K(\mathbf{x})$ is called band-limited. In such a case the set of the invisible objects is the set of the functions whose Fourier transform is zero over Ω. Moreover, the set of the exact images is contained in the subspace of the functions whose Fourier transform is zero out of Ω. More precisely, if g is an exact image, using the well-known convolution theorem of the Fourier transform we get

$$\hat{g}(\xi) = \hat{K}(\xi)\, \hat{f}(\xi) \tag{27}$$

and therefore $g \in R(L)$ if and only if

$$\hat{g}(\xi) = 0, \quad \xi \notin \Omega \quad ; \qquad \int_{\Omega} |\hat{K}(\xi)|^{-2} |\hat{g}(\xi)|^2 \, d\xi < +\infty \, . \tag{28}$$

These conditions, which are analogous to the conditions (24), combined with the property (26) of the kernel, imply that R(L) is not closed. Notice that this is not true when the quantity m, Eq.(26), is strictly positive. An example is provided by the band-limiting operator which will be discussed in the following. In this case, $K(\mathbf{x})$ is not integrable.

B. Inverse Source Problems

An inverse source problem can be defined, in general, as the problem of determining the constitution of a source from the measured values of the emitted radiation. A specific example is the determination of the current distribution of an antenna from the knowledge of its radiation pattern. Here we illustrate the main features of these problems by discussing the most simple case, namely that of a scalar source generating a scalar field. The more general problem of the determination of a charge-current distribution is investigated in (Devaney and Wolf, 1973; Bleinstein and Cohen, 1977; Hoenders, 1978) with a particular attention to the question of uniqueness.

The basic equation is the inhomogeneous Helmholtz equation:

$$\Delta u + k^2 u = -4\pi f \tag{29}$$

where $k = 2\pi/\lambda$ is the wave number of the emitted radiation. The function $f(\mathbf{r})$, the source density function, can be assumed to be zero outside some bounded region D. Without loss of generality we can also assume that D is a sphere whose radius a is known. Then the *direct problem* is the determination in the whole three-dimensional space of a field amplitude $u(\mathbf{r})$, satisfying Eq.(29) and also Sommerfeld radiation condition

$$\lim_{r\to\infty} r\,(\partial u/\partial r - iku) = 0 \quad , \qquad r = |\mathbf{r}| \quad , \tag{30}$$

i.e. $u(\mathbf{r})$ has to represent at infinity an outgoing spherical wave. As it is well known, if $f \in L^2(D)$, then there exists a unique continuous solution of this problem, given by

$$u(\mathbf{r}) = \int_D G(\mathbf{r}-\mathbf{r}')\, f(\mathbf{r}')\, d\mathbf{r}' \tag{31}$$

where

$$G(\mathbf{r}) = e^{ikr}/r \quad . \tag{32}$$

Moreover, the behaviour at infinity of this solution is

$$u(\mathbf{r}) = g(\mathbf{s})\, G(\mathbf{r}) \left[1 + O(r^{-1})\right] \quad . \tag{33}$$

Here $\mathbf{s} = \mathbf{r}/r = \{\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta\}$ and the function

$$g(\mathbf{s}) = \int_D e^{-ik(\mathbf{s},\mathbf{r}')}\, f(\mathbf{r}')\, d\mathbf{r}' \tag{34}$$

is the so-called *radiation pattern* .

The inverse problem is now the determination of the source distribution $f(\mathbf{r})$ given the radiation pattern $g(\mathbf{s})$ and this implies the solution of the first kind Fredholm integral equation (34) which has the same structure as Eq.(5) if we introduce the integral operator

$$(Lf)(\mathbf{s}) = \int_D e^{-ik(\mathbf{s},\mathbf{r}')}\, f(\mathbf{r}')\, d\mathbf{r}' \quad . \tag{35}$$

The operator L transforms a function $f \in L^2(D)$ into a function defined on the unit sphere $S^2 \subset \mathbb{R}^3$. Let us assume that the data space Y is the space $L^2(S^2)$ of the functions which are square integrable with respect to the Lebesgue measure on S^2, $d\mathbf{s} = \sin\theta\, d\theta\, d\phi$. Then it is easy to show that the operator $L : L^2(D) \longrightarrow L^2(S^2)$ is compact, and that its adjoint is given by

$$(L^* g)(\mathbf{r}) = \int_{S^2} e^{ik(\mathbf{r},\mathbf{s})}\, g(\mathbf{s})\, d\mathbf{s} \quad . \tag{36}$$

We first notice that $R(L) \subset L^2(S^2)$ is a subspace of analytic functions since, as follows from Eq. (34), any exact image $g(\mathbf{s})$ is the restriction, to the surface of the sphere of radius k (Ewald sphere), of the Fourier transform of a function with a bounded support. Moreover the null space N(L) can also be easily characterized. The radiation pattern is identically zero if and only if the Fourier transform of f is zero on the surface of the Ewald sphere. Since it is possible to prove (Devaney and Wolf, 1974; Hoenders, 1978) that the emitted field outside D is uniquely determined by the radiation pattern (this problem will be considered in the next Section), it follows that any $f \in N(L)$ produces a radiation field which is identically zero outside the source region D. In other words N(L), the subspace of the invisible objects, is also the subspace of the so-called *non-radiating sources*.

Given an arbitrary source distribution f, this can be uniquely decomposed into a component onto N(L) (the invisible component) and a component orthogonal to N(L) (the visible component). Using the second of the relations (18) the component orthogonal to N(L), i.e. the component which contributes to the radiation pattern, can be characterized by investigating the range of the operator L^*, as given by Eq.(36) (Bertero and De Mol, 1981 a).

If we use the well-known expansion of a plane wave into spherical harmonics

$$e^{ikr(\mathbf{s},\mathbf{s}')} = 4\pi \sum_{l=0}^{\infty} \sum_{m=-l}^{l} i^l j_l(kr)\, Y_{lm}(\mathbf{s})\, Y_{lm}(\mathbf{s}')^* \qquad (37)$$

and if we introduce the functions

$$u_{lm}(\mathbf{r}) = (4\pi/\sigma_l)\, i^l\, j_l(r)\, Y_{lm}(\mathbf{s}) \quad , \quad v_{lm}(\mathbf{s}) = Y_{lm}(\mathbf{s}) \quad , \qquad (38)$$

where

$$\sigma_l = 4\pi \left(\int_0^a r^2 j_l^2(kr)\, dr \right)^{1/2} \quad , \qquad (39)$$

from Eq.(36) we obtain

$$(L^* g)(\mathbf{r}) = \sum_{l=0}^{+\infty} \sum_{m=-l}^{l} \sigma_l \left(\int_{S^2} g(\mathbf{s}')\, Y_{lm}(\mathbf{s}')^* d\mathbf{s}' \right) u_{lm}(\mathbf{r}) \quad . \qquad (40)$$

In conclusion, the subspace of the visible sources is the closed subspace spanned by the functions $u_{lm}(\mathbf{r})$, Eq.(38). Moreover the set of the triples $\{\sigma_l;\, u_{lm},\, v_{lm}\}$ is just the singular system of the compact operator L, as follows from the relations

$$L u_{lm} = \sigma_l v_{lm} \quad , \quad L^* v_{lm} = \sigma_l u_{lm} \qquad (41)$$

which are easily obtained from Eqs. (34), (36) and (37) using the orthogonality properties of the spherical harmonics.

We finally remark that the singular values σ_l tend to zero exponentially fast when $l > ka$, indicating a strong ill-posedness of the

problem of determining the visible component of the source distribution.

C. Inverse Diffraction Problems

The problem of inverse diffraction can be defined as the problem of determining the field distribution on a boundary surface from the knowledge of the field distribution on a surface situated within the domain where the wave propagates (Shewell and Wolf, 1968). More precisely, consider a scalar field u in a region exterior to a surface Σ_0 which is the boundary of the region where the sources or the scatterers are situated. Then u is a solution, in the exterior region, of the homogeneous Helmholtz equation

$$\Delta u + k^2 u = 0 \qquad (42)$$

and it also satisfies Sommerfeld radiation condition (30) at infinity (in the case of a scattering problem this condition is satisfied by the scattered wave). The problem of inverse diffraction is now the problem of determining the field distribution on Σ_0 from the knowledge of the field distribution on another surface Σ_1 which belongs to the exterior region. In particular, when the surface Σ_1 is a sphere of large radius sorrounding Σ_0, the data consists of the radiation pattern or of the scattering amplitude, and the problem is called *inverse diffraction from far-field data*.

A general discussion of the uniqueness of the solution of the problem of inverse diffraction is given in (Hoenders, 1978) both in the scalar and in the vector case. The uniqueness of the problem of inverse diffraction from far-field data is proved in (Devaney and Wolf, 1974) in the case of the electromagnetic field.

The scalar problem formulated above can be easily solved when the

surfaces Σ_0 and Σ_1 are circular cylinders with the same centre (Cabayan *et al.*, 1973) or spheres with the same centre (Hoenders, 1978) or parallel planes (Shewell and Wolf, 1968). Here we sketch the case of inverse diffraction from far-field data and the case of inverse diffraction from plane to plane.

1. *Inverse Diffraction from Far-Field Data*

In the case of the problem of inverse diffraction from far-field data we can assume, without loss of generality, that Σ_0 is the surface of a sphere of radius a (for example the sphere, containing the source distribution, introduced in the discussion of the problem of Section B). Then we denote by $f(\mathbf{s})$ the field distribution on Σ_0. The direct problem is the determination of the solution $u(r,\mathbf{s})$ of Eq.(42) in the region $r > a$, satisfying Sommerfeld condition (30) at infinity and also the boundary condition $u(a,\mathbf{s}) = f(\mathbf{s})$. This is a typical exterior problem for the homogeneous Helmoltz equation. If we represent $f(\mathbf{s})$ in terms of spherical harmonics

$$f(\mathbf{s}) = \sum_{l=0}^{+\infty} \sum_{m=-l}^{l} c_{lm} Y_{lm}(\mathbf{s}) \tag{43}$$

where

$$c_{lm} = \int_{S^2} f(\mathbf{s})\, Y_{lm}(\mathbf{s})^* d\mathbf{s}\,, \tag{44}$$

then the solution $u(r,\mathbf{s})$ of the direct problem is given by

$$u(r,\mathbf{s}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} c_{lm} \left[h_l^{(1)}(kr)/h_l^{(1)}(ka) \right] Y_{lm}(\mathbf{s}) \qquad (45)$$

where the $h_l^{(1)}(kr)$ are the spherical Hankel functions of first kind. Notice that $|h_l^{(1)}(x)|$ is never zero for real values of x and therefore the solution of the direct problem (for a fixed r) always exists, is unique and depends continuously on f in the norm of $L^2(S^2)$.

Using now the well-known asymptotic behaviour of the spherical Hankel functions

$$h_l^{(1)}(kr) = (-i)^{l+1} \left(e^{ikr}/r \right) [1 + O(r^{-1})] \quad , \qquad (46)$$

from Eq.(33), which defines the radiation (diffraction) pattern $g(\mathbf{s})$, we derive the following integral relationship between $g(\mathbf{s})$ and $f(\mathbf{s})$

$$g(\mathbf{s}) = \int_{S^2} K(\mathbf{s},\mathbf{s}') f(\mathbf{s}')d\mathbf{s} \, , \qquad (47)$$

where

$$K(\mathbf{s},\mathbf{s}') = k^{-1} \sum_{l=0}^{\infty} \sum_{m=-l}^{l} (-i)^{l+1} \left[h_l^{(1)}(ka) \right]^{-1} Y_{lm}(\mathbf{s}) \, Y_{lm}(\mathbf{s}')^* \quad . \qquad (48)$$

It follows that the solution of the inverse diffraction problem from far-field data has been reduced to the solution of a first kind Fredholm integral equation. It is also easy to recognize that Eq.(48) is the spectral representation of the kernel. As a consequence the corresponding integral operator L is compact in $L^2(S^2)$, its eigenvalues are given by $\lambda_l = k^{-1}(-i)^{l+1}/h_l^{(1)}(ka)$ and the corresponding eigenfunctions are just the spherical harmonics. Since the latter form a basis inf $L^2(S^2)$ and all the

eigenvalues λ_l are different from zero, it follows that N(L) = {0} and therefore the solution of Eq.(47) is unique. This is the uniqueness result already mentioned in the previous Section. However, since $\lambda_l \longrightarrow 0$ when $l \longrightarrow +\infty$, the solution does not depend continuously on the data. As a consequence, even if uniqueness holds true, the information content of the far field is poorer than the information content of the near field. This result is obvious on the basis of elementary physical considerations.

We notice also that, when the source or the scatterer is interior to the sphere of radius a, it is possible to determine the field to the surface of the source or of the scatterer by means of suitable Bessel function expansions which provide a sort of analytic continuation of the field. This technique is described, for instance, in (Bertero *et al*, 1980 a) and it has been applied to the problem of determining the shape of perfectly conducting bodies (Imbriale and Mittra, 1970).

2. *Inverse Diffraction from Plane to Plane*

We consider now the inverse diffraction from plane to plane (Shewell and Wolf, 1968). Stability results for this problem are given in (Bertero and De Mol, 1981 b; Magnanini and Papi, 1985). An application of this problem is investigated in (Sondhi, 1969). The direct problem is the following : find, in the half-space $z > 0$, a solution $u(\mathbf{r}) = u(x, y, z)$ of Eq.(42), satisfying Sommerfeld radiation condition (30) at infinity and taking given boundary values on the plane $z = 0$. If we denote by $f(\boldsymbol{\rho})$, $\boldsymbol{\rho} = (x,y)$, the boundary field distribution and by $g_a(\boldsymbol{\rho}) = u(x,y,a)$ the field distribution on the plane $z = a$, then the solution of the direct problem is given by (Luneburg, 1964)

$$g_a(\boldsymbol{\rho}) = \int_{\mathbb{R}^2} G_1(\boldsymbol{\rho} - \boldsymbol{\rho}', a)\, f(\boldsymbol{\rho}')\, d\boldsymbol{\rho}' , \qquad (49)$$

where

$$G_1(\boldsymbol{\rho}, z) = -\frac{1}{2\pi} \frac{\partial}{\partial z} G(\mathbf{r}) \qquad (50)$$

and $G(\mathbf{r})$ is defined in Eq.(32).

The inverse diffraction problem is now the determination of the boundary distribution $f(\boldsymbol{\rho})$ from the knowledge of the field distribution $g_a(\boldsymbol{\rho})$ on the plane $z = a > 0$. According to Eq.(49) this corresponds to the inversion of a convolution operator such as the operator (25), a problem which has been treated by means of the Fourier transform. Notice that the representation of the field on the plane $z = a$ in terms of its Fourier transform is also called its representation in the form of an *angular spectrum of plane waves* (Shewell and Wolf, 1968).

The 2D Fourier transform of the kernel (50) is given by

$$\hat{G}_1(\boldsymbol{\xi}, z) = e^{i z\, m(\boldsymbol{\xi})} \qquad (51)$$

where

$$m(\boldsymbol{\xi}) = (k^2 - |\boldsymbol{\xi}|^2)^{1/2} \quad , \; |\boldsymbol{\xi}| \leq k \qquad (52)$$
$$m(\boldsymbol{\xi}) = i\,(|\boldsymbol{\xi}|^2 - k^2)^{1/2} \quad , \; |\boldsymbol{\xi}| > k \; ,$$

and therefore, in this case, $\Omega = \mathbb{R}^2$. The part of the spectrum $|\boldsymbol{\xi}| \leq k$ corresponds to the so-called *homogeneous waves* while the part $|\boldsymbol{\xi}| > k$ corresponds to the so-called *inhomogeneous* or *evanescent waves*.

Since $|\hat{G}_1(\boldsymbol{\xi}, z)| \neq 0$ for any finite value of $|\boldsymbol{\xi}|$ we conclude that the solution of Eq.(49) is unique in $L^2(\mathbb{R}^2)$. The solution however does not exist for arbitrary data and it does not depend continously on the data due

to the fast decay of $|\hat{G}_1(\xi,z)|$ when $|\xi| \longrightarrow +\infty$ or , in other words, due to the existence of evanescent waves. We also notice that, when z is very large, $|\hat{G}_1(\xi,z)|$ is practically zero for $|\xi| > k$. In such a case one can extract from the data only the values of $\hat{f}(\xi)$ in the disc $|\xi| \leq k$ and therefore this inverse diffraction problem provides a first example of the general problem of Fourier transform inversion with limited data (see Section F).

D. Linear Inverse Scattering Problems

The inverse scattering problems are in general nonlinear. They have stimulated beautiful mathematical researches which have been briefly mentioned in the Introduction and which are not the subject of this paper. Under some circumstances, however, it is possible to introduce physical approximations which allow a linearization of the nonlinear problem.

A well-known case is that of a weak scatterer, in which case it is possible to use Born approximation. Another kind of approximation which also leads to a linear problem is the so-called Rytov approximation (Chernov, 1967) which is valid when the scale at which the properties of the scatterer fluctuate is large compared to the wavelength λ of the incident radiation.

A third example is the investigation of dispersed systems by light scattering experiments. The typical case is that of a suspension of spherical particles which have the same physical properties (for instance, refraction index) but different sizes. The aim of the experiment is the determination of the distribution function of the particle sizes. In the case of dilute systems one can neglect multiple scattering and the problem becomes linear.

1. *Semi-Transparent Objects*

The use of Born approximation for reconstructing the refraction index of weakly scattering semi-transparent objects has been widely investigated (Wolf, 1969; Devaney, 1978; Hoenders, 1978) and also applied to experimental data processing (Carter, 1970; Carter and Ho, 1974; Fercher *et al.*, 1979).

Let $n(\mathbf{r})$ be the refraction index of the object and $f(\mathbf{r}) = 1 - n^2(\mathbf{r})$. Since we consider a bounded object, situated, for instance, inside the sphere of radius a, $f(\mathbf{r})$ is zero for $r > a$. Then the total field u (incident plus scattered) is a solution of the wave equation

$$\Delta u + k^2 u = k^2 f(\mathbf{r}) u \tag{53}$$

where $k = 2\pi/\lambda$ is the wave number (in the free space) of the incident radiation.

When the incident radiation is a plane wave, $u_0(\mathbf{r}) = \exp[ik(\mathbf{s}_0,\mathbf{r})]$, the first Born approximation to $u(\mathbf{r}) = u_0(\mathbf{r}) + u_S(\mathbf{r})$ (incident + scattered radiation) is

$$u^{(B)}(\mathbf{r}) = e^{ik(\mathbf{s}_0,\mathbf{r})} - (k^2/4\pi) \int G(\mathbf{r}-\mathbf{r}') e^{ik(\mathbf{s}_0,\mathbf{r}')} f(\mathbf{r}')\, d\mathbf{r}' \tag{54}$$

Now, let $u_S^{(B)}(\mathbf{r})$ be the second term on the r.h.s. of Eq.(54) (Born approximation of the scattered wave). The 2D Fourier transform of $u_S^{(B)}(\boldsymbol{\rho},z)$ ($\boldsymbol{\rho} = \{x,y\}$), over any plane of constant z which does not intersect the region containing the object, is given by (Wolf, 1969)

$$\hat{u}_S^{(B)}(\boldsymbol{\xi},z) = (ik/s_z) \exp[iks_z z]\, \hat{f}[k(\mathbf{s}-\mathbf{s}_0)] \tag{55}$$

where, in the r.h.s., $\hat{f}$ denotes the 3D Fourier transform of f and $k\mathbf{s} = \{\xi_x, \xi_y, (k^2 - \xi_x^2 - \xi_y^2)^{1/2}\}$. If follows from Eq.(55) that the knowledge of $\hat{u}_S^{(B)}(\boldsymbol{\xi}, z)$ is equivalent to the knowledge of $\hat{f}$ on the surface of the Ewald sphere with centre $k\mathbf{s}_0$ and radius k and therefore, from the mathematical point of view, this problem is similar to the inverse source problem discussed in Section B. Moreover, by varying the direction of incidence $\mathbf{s}_0$, a (theoretically infinite) number of experiments would allow one to determine $\hat{f}$ within the sphere with centre the origin and radius 2k (limiting Ewald sphere). In such a case the uniqueness of the reconstruction of f is assured, since $\hat{f}$ is an analytic function due to the bounded support of f. We notice that we have again the problem of recontructing a function from a limited knowledge of its Fourier transform.

The same mathematical problem must be solved if one applies Rytov approximation (Devaney, 1981). The basic point is the introduction of the complex phase function

$$\psi(\mathbf{r}) = \ln u(\mathbf{r}) - ik(\mathbf{s}_0, \mathbf{r}) \quad . \tag{56}$$

Then the Rytov approximation for the complex phase function is given by the following equation (Chernov, 1967)

$$\psi^{(R)}(\mathbf{r}) = e^{-ik(\mathbf{s}_0, \mathbf{r})}\, u_S^{(B)}(\mathbf{r}) \quad . \tag{57}$$

It is obvious now that if one takes the 2D Fourier transform of $\exp[ik(\mathbf{s}_0, \mathbf{r})]\, \psi^{(R)}(\mathbf{r})$, over any plane of constant z which does not intersect the scatterer, one gets again the values of $\hat{f}$ on the surface of the Ewald sphere. A short discussion of the limits of validity of the two approximations is given in (Devaney, 1981).

In the case of the scattering of an electromagnetic wave by a perfectly conducting body the Born approximation (also known as Kirchhoff or physical optics approximation) consists in taking for the magnetic field, at the surface of the body, the values of the incoming field (Bojarski, 1966; Lewis, 1969; Hoenders, 1978). This approximation is valid when the wavelength λ of the incoming field is small with respect to the smallest geometrical details of the surface of the body.

Consider an incident plane wave with electric field $\mathbf{E}_i(\mathbf{r}) = \exp[ik(\mathbf{s}_0,\mathbf{r})]\,\mathbf{E}_0$ and assume that the target is a smooth, convex and bounded body which occupies the domain D. The backscattered field $\mathbf{E}_b(\mathbf{r})$, i.e. the field observed in the direction $\mathbf{s}_b = -\mathbf{s}_0$, at a large distance from the scatterer is given by

$$\mathbf{E}_b(\mathbf{r}) = \rho(\mathbf{k})\left(e^{ikr}/(2\sqrt{\pi}\,r)\right)\mathbf{E}_0 + O(r^{-2}) \qquad (58)$$

where $\mathbf{k} = k\mathbf{s}_0$. Then in the papers mentioned above it is proved, using Born approximation, that

$$\Gamma(\mathbf{k}) \equiv 2\sqrt{\pi}\,k^{-2}\,[\,\rho(\mathbf{k}) + \rho^*(-\mathbf{k})\,] = \int \chi(\mathbf{r})\,e^{i(\mathbf{k},\mathbf{r})}\,d\mathbf{r} \qquad (59)$$

where $\chi(\mathbf{r})$ is the characteristic function of the target, i.e. the function which is 1 when $\mathbf{r}$ is in D and zero elsewhere.

Since, in practice $\Gamma(\mathbf{k})$ can be measured only for values of $\mathbf{k}$ in a restricted domain (in particular, in radar applications, only for points interior to the annular region $m \leq |\mathbf{k}| \leq M$, where m and M are related to the minimum and maximum value of the usable frequency band), we have

again a problem which consists in inverting the Fourier transform with limited data.

A special feature of this problem is the lack of information at low frequencies. For this reason it has been suggested of reconstructing the directional derivative of the characteristic function rather than the characteristic function itself (Mager and Bleinstein, 1978). In this way one simultaneously attenuates low-frequency data while enhancing the effect of high-frequency data.

3. *Dispersed Systems*

Consider a dilute suspension or a dilute aerosol consisting of spherical particles having the same physical properties (refraction index etc.) but different sizes. We denote by f(r) the probability density of the particle size distribution. Moreover we denote by Q(p,r) the measured scattering pattern in the case where all the particles have the same radius r (monodisperse system). This pattern, in general, is a function of the radius r and of the measured scattering variable p. For example Q(p,r) is the correlation function of the scattered field in photon correlation spectroscopy (Cummins and Pike, 1974) - here p is the correlation time - or the small angle differential cross-section in Fraunhofer diffraction (Van de Hulst, 1981) - here p is the scattering angle - or the absorption coefficient as a function of the wavelength of the incident radiation (Van de Hulst, 1981).

Then, by neglecting multiple scattering, in the case of a polydispersed system with particle size probability f(r), the scattering pattern is (Shifrin and Perelman, 1965)

$$g(p) = \int_0^{+\infty} Q(p,r)\, f(r)\, dr \qquad . \qquad (60)$$

In several important cases Q(p,r) depends only on the product of the two variables, i.e. Q(p,r) = K(pr) and the first kind Fredholm integral equation (60) becomes

$$g(p) = \int_0^{+\infty} K(pr)\, f(r)\, dr \quad . \tag{61}$$

For example in polydispersity analysis by photon correlation spectroscopy (Cummins and Pike, 1974) we have K(x) = exp(-x) and the solution of Eq.(61) becomes the inversion of the Laplace transform, a problem considered in Section G (in such a case, however, the variable r is not the radius of the particles but their translational diffusion coefficient, which is proportional to the inverse of the radius of the particles). The integral equation (61) is also a satisfactory approximation both in the case of Fraunhofer diffraction and in the case of extinction experiments, when using anomalous diffraction approximation (Shifrin and Perelman, 1965).

The integral equation (61) has the form (5) with

$$(Lf)(p) = \int_0^{+\infty} K(pr)\, f(r)\, dr \quad . \tag{62}$$

We assume, for simplicity that the data g(p) are measured for all the values of p in (0, +∞) and we consider L as a linear operator in $L^2(0,+\infty)$. Then the integral equation (61) can be investigated using the spectral representation of L in terms of generalized eigenvalues and eigenfunctions (McWhirter and Pike, 1978).

An equivalent approach is based on the use of the Mellin transform which, for a square integrable function is defined by (Titchmars, 1948).

$$(\mathfrak{M}f)(\xi) = \int_0^{+\infty} f(r)\, r^{-\frac{1}{2}+i\xi}\, dr \quad . \tag{63}$$

The transform $\mathfrak{M}$ is an isometry from $L^2(0, +\infty)$ into $L^2(-\infty, +\infty)$

$$\int_0^{+\infty} |f(r)|^2 \, dr = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |(\mathfrak{M}f)(\xi)|^2 \, d\xi \qquad (64)$$

and the following inversion formula holds true

$$f(r) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} (\mathfrak{M}f)(\xi) \, r^{-(\frac{1}{2}+i\xi)} \, d\xi \quad . \qquad (65)$$

Now, by taking the Mellin transform of both sides of Eq.(61) one gets

$$(\mathfrak{M}g)(\xi) = (\mathfrak{M}K)(\xi)(\mathfrak{M}f)(-\xi) \quad . \qquad (66)$$

We find that the operator L is a continuous, self-adjoint operator in $L^2(0,+\infty)$ if K(x) is real and satisfies the condition

$$\int_0^{+\infty} x^{-\frac{1}{2}} |K(x)| \, dx \; < \; +\infty \quad . \qquad (67)$$

Moreover the operator L is invertible if and only if the support of $(\mathfrak{M}K)(\xi)$ coincides with $(-\infty,+\infty)$. In such a case from Eqs.(65) and (66) we can derive that

$$(L^{-1}g)(r) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{(\mathfrak{M}g)(-\xi)}{(\mathfrak{M}K)(-\xi)} \, r^{-(\frac{1}{2}+i\xi)} \, d\xi \quad . \qquad (68)$$

The condition (67) however, implies that $|(\mathfrak{M}K)(\xi)| \longrightarrow 0$ when $|\xi| \longrightarrow +\infty$

(this is again a consequence of the Riemann-Lebesgue theorem) and therefore the operator L^{-1} is not continuous in $L^2(0,+\infty)$.

E. Radon Transform Inversion and Tomography

In a paper devoted to inverse problems it is impossible to forget the problem of Radon transform inversion which is basic in several fundamental areas, such as diagnostic radiology (*computerized tomography*), radio astronomy and electron microscopy. On the other hand it is impossible to give even a short account of the explosive development of this field where the most spectacular applications of the theory of inverse problems have been obtained. Therefore we limit ourselves to show the connection between Radon transform inversion and the general formulation given above. The interested reader is deferred to the books which have been published both on the mathematical aspects (Herman and Natterer, 1981; Natterer, 1986 a) and on the applications as well as on the computational methods (Herman, 1979; Herman, 1980).

The problem of Radon transform inversion, which is also called *object reconstruction from projections*, and the related problem of Abel transform inversion are two examples of linear inverse scattering problems which arise when the variations of the dynamical functions over a given wavelength are so small that diffraction can be neglected. In such a case a geometrical-optics description of the process is possible and, in several instances, it is also possible to assume straight line ray propagation. These approximations are adeguate especially at X-ray wavelengths.

The n-dimensional Radon transform R maps a function $f(\mathbf{x})$ defined on $\mathbb{R}^n$ into the set of its integrals over the hyperplanes of $\mathbb{R}^n$. Therefore, if S^{n-1} is the surface of the unit sphere in $\mathbb{R}^n$ and $\boldsymbol{\theta} \in S^{n-1}$, $s \in \mathbb{R}^1$, the integrals of f over the hyperplanes perpendicular to $\boldsymbol{\theta}$, with signed

distance from the origin s, are given by

$$(Rf)(\boldsymbol{\theta},s) = \int_{(\mathbf{x},\boldsymbol{\theta})=s} f(\mathbf{x})\, d\mathbf{x} \qquad (69)$$

where $(\mathbf{x},\boldsymbol{\theta})$ is the scalar product in $\mathbb{R}^n$. The operator R defines the *Radon transformation* and the inverse problem is the solution of the equation g=Rf for a given g. A terse discussion of related transforms such as the *X-ray transform,* the *divergent beam transform* and the *attenuated Radon transform* is contained in (Louis and Natterer, 1983). Another related problem is *diffraction tomography* which is derived from the scattering problem of the wave equation by the use of Rytov approximation (Mueller *et al.*, 1979; Devaney, 1982, 1984; Natterer, 1986 a).

We focus now on the 2D case, in which case Radon transform and X-ray transform coincide since the integrals over hyperplanes are simply integrals over straight lines. Eq.(69) can also be written in the following form, where $\boldsymbol{\theta}^{\perp}$ is the unit vector orthogonal to $\boldsymbol{\theta}$

$$(Rf)(\boldsymbol{\theta},s) = \int_{-\infty}^{+\infty} f(s\boldsymbol{\theta}+t\,\boldsymbol{\theta}^{\perp})\, dt \qquad . \qquad (70)$$

Another notation which is often used is the following

$$(Rf)(\boldsymbol{\theta},s) = (R_{\boldsymbol{\theta}}f)(s) \qquad (71)$$

and the function $R_{\boldsymbol{\theta}}f$, with fixed $\boldsymbol{\theta}$, is called a projection of f. This is the origin of the name object reconstruction from projections which is often used as a synonymous of Radon transform inversion.

One of the basic properties of the Radon transform, which clarifies its information content about the function f, is the so-called *projection slice theorem* (or *projection theorem* or also *Fourier slice theorem* - for

a proof see, for example, (Natterer, 1986 a)) - whose content is the following relation between Fourier transforms

$$(R_\theta f)\hat{}\,(\xi) = \hat{f}(\xi\theta) \tag{72}$$

where $(R_\theta f)\hat{}$ is the 1D Fourier transform of $R_\theta f$ while $\hat{f}$ is the 2D Fourier transform of f. The meaning of this result is clear : the knowledge of $R_\theta f$ is equivalent to the knowledge of $\hat{f}$ along the straight line parallel to θ and passing through the origin.

Several results concerning the uniqueness of the solution can be immediately deduced from Eq.(72). If $R_\theta f$ is known for all the values of θ, then $\hat{f}$ is known everywhere and therefore f is uniquely determined. On the other hand, if $R_\theta f$ is known only for values of θ in a subset of the half-circle - this is the problem of *limited angle tomography* - the solution of the problem in general is not unique. But, when the function f has a bounded support, its Fourier transform is an analytic function and therefore it can be uniquely recovered from its values in a finite sector. We have again a problem of restoring a function f from limited values of its Fourier transform.

An explicit inversion formula for the transform (69) was already obtained by Radon (Radon, 1917), as we recalled in the Introduction. Here we sketch an approach which is the basis of the algorithms which are currently used in the practical applications. If we introduce the formal adjoint $R^\#$ of the Radon transform, also called *back projection operator*

$$(R^\# g)(\mathbf{x}) = \int_{S^1} g(\theta, (\mathbf{x}, \theta))\, d\theta \tag{73}$$

then the equation g = Rf can be replaced by $R^\# g = R^\# R f$. By taking the 2D

Fourier transform of both sides of this equation we have (Natterer,1986)

$$(R^{\#}R\, f)\hat{}\,(\xi) = 4\pi\, |\xi|^{-1}\, \hat{f}\,(\xi) \qquad . \qquad (74)$$

Then, if we introduce the operator Λ defined by

$$(\Lambda\, f)\hat{}\,(\xi) = |\xi|\ \hat{f}\,(\xi) \qquad (75)$$

we obtain $(\Lambda\, R^{\#}R\, f)\hat{}\, = 4\pi\, \hat{f}$ and this equation, combined with the original equation $Rf = g$, provides the following inversion formula

$$f = (4\pi)^{-1} \Lambda\, R^{\#} g \quad . \qquad (76)$$

This formula is the basis of the filtered backprojection method which is the algorithm most frequently used in practice.

In the previous computations we have not specified the object space X and the image space Y. As concerns the unknown function f it is always convenient, and also reasonable in practice, to assume that f has a bounded support. Moreover, in several papers it has been considered the case where both X and Y are Sobolev spaces. Then it is possible to prove that with an appropriate choice of these Sobolev spaces the operator R, Eq.(69), is continuous and has a continuous inverse (Natterer, 1980; Louis and Natterer, 1983). The continuity of R^{-1}, however is only obtained when the functions of the data space are smoother than the functions of the solution space. As we have already discussed in the Introduction, this choice may not be reasonable in practice due to the effect of the noise on the data. Therefore Radon transform inversion is an ill-posed problem when the data space is, for example, a space of square integrable functions. The ill-posedness is essentially related to the fact that the operator Λ, Eq.(75), is not continuous in L^2. We also mention that, in the case of functions f with bounded support, the Radon transform R defines a

compact operator if X and Y are suitable weighted L^2-spaces. Then the singular system of R has been explicitely determined for arbitrary values of n (Davison, 1981; Louis, 1984) while for the limited angle problem the singular system has been determined only in the case n=2 (Louis, 1986).

F. Fourier Transform Inversion with Limited Data

We have seen in the previous Sections that several inverse problems can be reduced to the same basic problem: determine a function $f(\mathbf{x})$ defined on $\mathbb{R}^n$ from the knowledge of its Fourier transform $\hat{f}(\xi)$ on a bounded domain $\Omega \subset \mathbb{R}^n$. Therefore if we denote by $g(\xi)$ the known (in general noisy) values of $\hat{f}$, we must solve the following problem

$$g(\xi) = \int_{\mathbb{R}^n} e^{-i(\xi,\mathbf{x})} f(\mathbf{x})\, d\mathbf{x} \quad , \qquad \xi \in \Omega \quad . \tag{77}$$

A quite natural framework for dealing with this problem is to take $X = L^2(\mathbb{R}^n)$ and $Y = L^2(\Omega)$. Then it is obvious that the solution of the problem is not unique. The null space of the operator is the set of all the functions f whose Fourier transform is zero over Ω. Uniqueness, however, holds true when the function f has a given bounded support $D \subset \mathbb{R}^n$. In such a case one can take $X = L^2(D)$. Then the Fourier transform of any $f \in L^2(D)$ is analytic and therefore it can be uniquely determined from its values on a suitable infinite set of points. The problem however is ill-posed because existence and continuous dependence of the solution on the data in general do not hold true.

In this Section we treat essentially the one-dimensional case using the basic results of Slepian and coworkers (Slepian and Pollak, 1961; Landau and Pollak, 1961; 1962). The extension to many dimensions can be

done along the lines indicated in (Slepian, 1964).

We assume that the values of the Fourier transform of a function $f \in L^2(-\infty,+\infty)$ are given in the interval $[-c,c]$ and we notice that, in such a case, Eq. (77) has the form (5) if the operator L is defined as follows

$$(Lf)(\xi) = \int_{-\infty}^{+\infty} e^{-i\xi x} f(x)\, dx\,, \qquad |\xi| \leq c \quad . \tag{78}$$

We assume also that the data space Y is $L^2(-c,c)$ and we define the norm of g as its L^2- norm divided by 2π. Then the adjoint operator is given by

$$(L^* g)(x) = \frac{1}{2\pi} \int_{-c}^{c} e^{ix\xi} g(\xi)\, d\xi \quad . \tag{79}$$

In this particular case the null space of L is the set of all the functions $f \in L^2(-\infty,+\infty)$ whose Fourier transform is zero on the interval $[-c,c]$, and therefore, as we have already pointed out, the solution of the problem is not unique. We also notice that the operator LL^* is just the identity operator in $L^2(-c,c)$, while the operator L^*L is the so-called band-limiting operator B_c, given by

$$(B_c f)(x) = \int_{-\infty}^{+\infty} (\pi/c)\ \mathrm{sinc}\,[c(x-y)/\pi]\, f(y)\, dy \tag{80}$$

where the following standard notation has been used

$$\mathrm{sinc}(x) = \sin(\pi x) / (\pi x) \quad . \tag{81}$$

In fact the operator B_c transforms a function $f \in L^2(-\infty,+\infty)$ into a function whose Fourier transform coincides with the Fourier transform of

f over the interval [-c,c] and is zero elsewhere. It follows that B_c is the projection operator onto the subspace of the band-limited functions with bandwidth c. Such a subspace of entire function, which is the subspace of the visible objects, is closed with respect to the L^2-norm and therefore it is itself a Hilbert space, also called Paley-Wiener space and denoted by PW_c. We recall that, for any function $f \in PW_c$ the following Whittaker-Shannon expansion (sampling theorem) holds true

$$f(x) = \sum_{n=-\infty}^{+\infty} f(x_n)\ \mathrm{sinc}[c(x-x_n)/\pi] \quad . \tag{82}$$

where $x_n = n\pi/c$, $n = 0, \pm 1, \ldots$ The distance π/c between adjacent sampling points is usually called the *Nyquist sampling distance*.

As we have already remarked, uniqueness holds true when the function f has a bounded support, say interior to the interval [-1,1]. Then, if we take $X = L^2(-1,1)$, the operator (78) is replaced by the following one

$$(Lf)(\xi) = \int_{-1}^{1} e^{-i\xi x} f(x)\, dx, \qquad |\xi| \leq c \quad . \tag{83}$$

It is easy to recognize that this is a compact operator from $L^2(-1,1)$ into $L^2(-c, c)$. The determination of its singular system can be reduced to the solution of the eigenvalue problem of the operator L^*L which is given by

$$(L^*Lf)(x) = \int_{-1}^{1} (\pi/c)\ \mathrm{sinc}[c(x-y)/\pi]\, f(y)\, dy, \qquad |x| \leq 1 \ . \tag{84}$$

This is the integral operator investigated by Slepian and coworkers. Its eigenfunctions, denoted by $\psi_k(c,x)$, $k = 0,1,2,\ldots$, are called *prolate*

spheroidal wave functions (PSWF). The corresponding eigenvalues λ_k, ordered to form a decreasing sequence have a typical step behaviour : they are approximately equal to one for values of the index less than $2c/\pi$ and then fall off to zero exponentially. Notice that $2c/\pi$ is just the number of the sampling points appearing in Eq.(82) and interior to the interval [-1,1].

The main properties of the PSWF, which are usually normalized to one with respect to the norm of $L^2(-\infty, +\infty)$, are the following:

a) the norm of $\psi_k(c,x)$ in $L^2(-1,1)$ is $\sqrt{\lambda_k}$;

b) $\psi_k(c,x)$ and $\psi_j(c,x)$, with $k \neq j$, are doubly orthogonal, i.e. they are orthogonal both with respect to the scalar product of $L^2(-1,1)$ and with respect to the scalar product of $L^2(-\infty, +\infty)$;

c) $\psi_k(c,x)$ has exactly k zeros interior to the interval [-1,1];

d) the set of the $\psi_k(c,x)$ forms an orthonormal basis in PW_c while the set of the $\lambda_k^{-\frac{1}{2}} \psi_k(c,x)$ forms an orthonormal basis in $L^2(-1,1)$;

e) The $\psi_k(c,x)$ are also eigenfunctions of the differential operator

$$(Df)(x) = -[(1-x^2) f'(x)]' + c^2x^2f(x) \qquad (85)$$

which is a self-adjoint operator in $L^2(-1,1)$, with boundary conditions defined by the requirement that the eigenfunctions must be bounded at the points ± 1.

The last property is quite important in practice for the computation of the PSWF and it is also important from a theoretical point of view since it provides one of the few examples of a differential operator commuting with an integral operator with an analytic kernel (Grünbaum, 1986). A second one will be shown in the next Section.

Finally, by investigating the operator LL^* it is easy to conclude

that the singular system of the operator (83) is given by

$$\sigma_k = \lambda_k^{1/2}, \qquad u_k(x) = \lambda_k^{-1/2}\,\psi_k(c,x) \tag{86}$$

$$v_k(\xi) = (c\lambda_k)^{-1/2}\,\psi_k(c,\xi/c) \quad .$$

As a concluding remark we point out that a related problem is the inversion of the following integral operator

$$(Lf)(x) = \int_{-1}^{1} (\pi/c)\ \mathrm{sinc}[c(x-y)/\pi]\, f(y)\, dy, \qquad -\infty < x < +\infty \ . \tag{87}$$

which is relevant for the investigation of optical systems (Bertero and Pike, 1982). The difference with respect to the Slepian operator (84) is obvious: while in that case the range of the variable x is restricted to the interval [-1,1], in this case it is the full real line. As a consequence the operator (87) is not self-adjoint, but it is still possible to show that it is a compact operator from $L^2(-1,1)$ into $L^2(-\infty,+\infty)$. Moreover its singular system can be given again in terms of the PSWF. More precisely the singular values σ_k and the singular functions u_k are given again by Eq.(86) while the singular functions v_k are now given by $v_k(x) = \psi_k(c,x)$. The operator (87) is also basic in the solution of the problem of bandwidth extrapolation (Viano, 1976;Bertero etal.,1980 a).

6. Laplace Transform Inversion

In several domains of experimental science, the experimenter is concerned with the problem of recovering and resolving exponential

relaxation rates. Many examples come to mind : nuclear magnetic resonance in chemistry and, more recently, in medical imaging, photon correlation spectroscopy, fluorescence, sedimentation equilibrium and, in general, relaxation kinetics. In all such cases, the basic problem is the inversion of the Laplace transform

$$g(p) = (Lf)(p) = \int_0^{+\infty} e^{-pt} f(t)\, dt \qquad (88)$$

It is well known that, in general, g(p) is analytic in a half-plane $Re(p) > p_0$ and that an inversion formula can be given using contour integration in the complex plane. This formula, however, is useless in practice because the available data will be *noisy* values of g(p) for a *finite* number of values of p. Here we assume that the values of g(p) are known for all the values of p in $(0, +\infty)$ and we defer the case of discrete data to the next Chapter.

The linear mapping L, defined by Eq.(88), is continuous and injective in $L^2(0, +\infty)$. In fact, from Eq.(66), recalling that the Mellin transform of the exponential is the gamma function, we have

$$(\mathcal{M}g)(\xi) = \Gamma\left(\tfrac{1}{2} + i\xi\right) (\mathcal{M}f)(-\xi) \qquad (89)$$

Then, using the bound

$$|\Gamma(\tfrac{1}{2} + i\xi)|^2 = \frac{\pi}{\cosh(\pi x)} \leq \pi \qquad (90)$$

and Eq.(64), with f replaced by Lf, we obtain

$$\int_0^{+\infty} |(Lf)(p)|^2 \, dp \leq \pi \int_0^{+\infty} |f(t)|^2 \, dt \quad . \qquad (91)$$

Moreover, from Eq.(89) and Eq.(68) we derive the following inversion formula for the Laplace transform

$$f(t) = (L^{-1}g)(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{(\mathfrak{M}g)(-\xi)}{\Gamma(\frac{1}{2} - i\xi)} \, t^{-(\frac{1}{2}+i\xi)} \, d\xi \quad . \qquad (92)$$

We conclude that L is a continuous, self-adjoint operator in $L^2(0,+\infty)$, with $\| L \| = \sqrt{\pi}$ and that L^{-1} exists but is not continuous.

An important case which has been investigated only recently (Bertero *et al.*, 1982) is the inversion of the *finite Laplace transform*, i.e. the inversion of the Laplace transform of a function with a bounded support interior to a given interval, say [a,b], $0 < a < b < +\infty$. The dual problem is the inversion of the Laplace transform with limited data, i.e. the case where the Laplace transform is known only on the interval [a,b] and no restriction is introduced on the support of the unknown function.

Thanks to the scaling properties of the Laplace transformation, it is not restrictive to assume that the support of f is interior to the interval [1,γ], so that the finite Laplace transformation is defined as follows

$$(Lf)(p) = \int_1^{\gamma} e^{-pt} \, f(t) \, dt, \qquad 0 \leq p \leq \infty \quad . \qquad (93)$$

It is easy to recognize that L is a compact operator from $L^2(1,\gamma)$ into $L^2(0,+\infty)$ and that its inverse operator L^{-1} exists. Analogously, its adjoint operator

$$(L^*g)(t) = \int_0^{+\infty} e^{-tp}\, g(p)\, dp, \qquad 1 \le t \le \gamma \tag{94}$$

is also compact and invertible, so that the range of L is dense in $L^2(0,+\infty)$. It follows that if we introduce the singular system of L, i.e. $\{\sigma_k; u_k, v_k\}$, then the singular functions u_k form an orthonormal basis in $L^2(1,\gamma)$ while the singular functions v_k form an orthonormal basis in $L^2(0,+\infty)$. From the general results on compact operators presented in Section A, we know that the u_k are the eigenfunctions of the operator L*L which is given by

$$(L^*Lf)(t) = \int_1^{\gamma} \frac{1}{t+s} f(s)\, ds \quad , \quad 1 \le t \le \gamma \tag{95}$$

and can be called the *finite Stieltjes transformation*. Analogously the singular functions v_k are the eigenfunctions of the operator L^*L which is given by

$$(LL^*g)(p) = \int_0^{+\infty} \frac{1}{p+q} \left[e^{-(p+q)} - e^{-\gamma(p+q)} \right] g(q)\, dq \; . \tag{96}$$

The singular values σ_k have been computed numerically in (Bertero *et al.*, 1982) for a few values of γ. If they are ordered to form a decreasing sequence, then it is possible to prove that, for a fixed k, σ_k is an increasing function of γ. As concerns the dependence on k, for fixed γ, the σ_k tend to zero very rapidly so that only a few singular values are significantly large. For example, in the case γ=5 only 5 singular values are

greater than 10^{-3} : $\sigma_0 = 0.8751$, $\sigma_1 = 0.1935$, $\sigma_2 = 0.03827$, $\sigma_3 = 0.007434$, $\sigma_4 = 0.001435$.

A remarkable property, similar to the basic property of the PSWF, is that it is possible to find differential operators commuting with the integral operators L^*L and LL^* (Bertero and Grünbaum, 1985). This result also provides a tool for the computation of the singular functions (Bertero *et al.*, 1986). More precisely the singular functions u_k are the eigenfunctions of the second order differential operator

$$(D^{(1)}f)(t) = -\left[(t^2-1)\,(\gamma^2-t^2)\,f'(t)\right]' + 2(t^2-1)\,f(t) \qquad (97)$$

which is a self-adjoint operator in $L^2(1,\gamma)$, with boundary conditions defined by the requirement that the eigenfunctions must be bounded at the points $1,\gamma$. On the other hand, the singular functions v_k are the eigenfunctions of the fourth order differential operator

$$(D^{(2)}g)(p) = \left[p^2\,g''(p)\right]'' - (\gamma^2+1)\left[p^2\,g'(p)\right]' + (\gamma^2\,p^2-2)\,g(p) \qquad (98)$$

which is also self-adjoint in $L^2(0,+\infty)$ if one looks for eigenfunctions which are bounded at the origin and square integrable at infinity.

Using these results it is possible to prove (Bertero and Grünbaum, 1985) that:

a) all the singular values have multiplicity 1;

b) the singular function u_k has exactly k zeros interior to $[1,\gamma]$; the points 1 and γ can never be zeros of u_k .

In the limit $\gamma \longrightarrow 1$ it has also been found an interestng relationship

between the singular functions u_k and the Legendre polynomials (Bertero *et al.*, 1986 a).

H. Generalized Moment Problems

A *generalized moment problem* can be defined as follows: let ϕ_1, ϕ_2, ϕ_3,. . . be a sequence of linearly independent functions in the Hilbert space X; then find a function $f \epsilon X$ such that

$$(f, \phi_n)_X = g_n ; \qquad n = 1,2,3,. . . \tag{99}$$

where g_1, g_2, g_3,. . . is a sequence of given numbers. This problem has the general form (5) if we introduce a linear operator L which transforms a function $f \epsilon X$ into a sequence of numbers according to Eq.(99). Moreover we will assume that the data space Y is l^2, i.e. the space of square summable sequences, with a norm defined by

$$\| g \|_Y^2 = \sum_{n=1}^{\infty} | g_n |^2 \tag{100}$$

where we have denoted by g the sequence g_1, g_2, g_3,. . .

If the ϕ_n form an orthonormal basis in X, then the problem is trivially well-posed and there exists a unique solution, for any $g \epsilon Y$, given by the expansion

$$f = \sum_{n=1}^{\infty} g_n \phi_n \quad . \tag{101}$$

The problem is also well-posed when the set of the functions ϕ_n is nearly an orthonormal basis in the following sense : there exists an orthonormal basis $\{\psi_n\}$ and a positive number $\theta < 1$, such that, for any sequence $g \in Y$

$$\| \sum_{n=1}^{\infty} g_n (\phi_n - \psi_n) \|_X \leq \theta \| g \|_Y \quad . \tag{102}$$

In such a case one can prove (Riesz and Nagy, 1972) that there exists the dual basis $\{ \phi^n \}$, i.e. the set of the functions satisfying the conditions

$$(\phi_n, \phi^m)_X = \delta_{nm} \; ; \quad n,m = 1,2,.... \tag{103}$$

and that the sets $\{ \phi_n \}$ and $\{ \phi^n \}$ form a biorthogonal basis in X. This result implies that there exists a unique solution of the problem (99), which is given by

$$f = \sum_{n=1}^{\infty} g_n \phi^n \quad . \tag{104}$$

Moreover the mapping $g \longrightarrow f$ is continuous since, as follows from inequalities proved in (Riesz and Nagy, 1972), we have

$$\| f \|_X \leq (1 + \theta) \| g \|_Y \quad . \tag{105}$$

Applications of this results to the theory of non-harmonic Fourier series are outlined in (Riesz and Nagy, 1972).

When the functions ϕ_n do not satisfy the previous conditions and, in particular, the angle between ϕ_n and ϕ_{n+1} tends to zero for $n \longrightarrow \infty$, the problem can be ill-posed. It is easy however to give a general condition for the uniqueness of the solution or, in other words, for the existence of

the inverse of the operator L : the solution is unique if and only if the span of the ϕ_n is dense in X. Otherwise the null space of L is just the orthogonal complement of the span of the functions ϕ_n.

A classical example which satisfies the requirement of uniqueness is the *Hausdhorff moment problem* : find a function f defined on (0,1) from the values of its moments

$$g_n = \int_0^1 x^{n-1} f(x)\, dx \; ; \qquad n = 1, 2, 3, \ldots \tag{106}$$

This problem has the form (99) if $X = L^2(0,1)$, in which case the span of the functions $\phi_n(x) = x^{n-1}$ is dense in X. The operator $L : L^2(0,1) \longrightarrow l^2$, defined by Eq.(106) is continuous and $\| L \| = \sqrt{\pi}$. The range of L, however, is only dense in l^2 and therefore L^{-1} is not continuous. In fact, the characterization of the necessary and sufficient conditions for the numbers $g_1, g_2, g_3, \ldots$ to be the moments of a function $f \in L^p(0,1)$, $p > 1$, has been the subject of several beautiful mathematical investigations (Widder, 1946). A terse discussion of the main results can also be found in (Talenti, 1987).

It is important to point out that the Hausdorff moment problem is related to Laplace transform inversion when the Laplace transform is given at the points $p_n = n - \frac{1}{2}$, $n = 1, 2, \ldots$ In fact if we consider the generalized moment problem

$$g_n = \int_0^{+\infty} e^{-(n-\frac{1}{2})t}\, u(t)\, dt \; ; \qquad n = 1, 2, \ldots \tag{107}$$

using the change of variables $x = e^{-t}$ and introducing the function $f(x) =$

$x^{-\frac{1}{2}}\, u(-\ln x)$, we transform this problem into the problem (106). Moreover if the function u(t) belongs to $L^2(0,+\infty)$, then the function f(x) belongs to $L^2(0,1)$.

Other examples are the *Stieltjes moment problem*, i.e. the problem of determining a function f(x) defined on $(0,+\infty)$ from the knowledge of its moments

$$g_n = \int_0^{+\infty} x^{n-1} f(x)\, dx \; ; \qquad n = 1, 2, 3, \ldots \tag{108}$$

or the *Hamburger moment problem*, which is the problem of determining a function defined on $(-\infty,+\infty)$ always from the knowledge of its moments. It is obvious that the functionals (108) are not continuous in $L^2(0,+\infty)$ and therefore the problem must be considered in some suitable weighted L^2-space or, roughly speaking, one must consider a Hilbert space of functions which tend to zero at infinity more rapidly than any inverse power of x. The uniqueness of the solution however is not assured in general (it depends on the choice of the space) since, as it was already shown by Stiltjes, all the moments of the following function

$$f(x) = \exp(-x^{1/4}) \sin(x^{1/4}) \tag{109}$$

are zero (Widder, 1947).

Another problem is *Poisson transform* inversion (Saleh, 1978; Bertero and Pike, 1986), a problem which is related to the inversion of photon counting distributions to obtain distributions of classical light intensity fluctuations

$$g_n = \frac{1}{(n-1)!} \int_0^{+\infty} x^{n-1}\, e^{-x} f(x)\, dx \; ; \qquad n = 1, 2, 3, \ldots \tag{110}$$

The solution of this problem is unique in $L^2(0,+\infty)$ as easily follows from the completeness of Laguerre polynomials. Notice that the inversion of the Poisson transform is equivalent to the solution of the Stieltjes moment problem in a suitable weighted space with an exponential weight. Moreover, if the problem (110) is also formulated in a weighted space with an exponential weight, i.e. if we define the norm of f(x) as follows

$$\|f\|_X^2 = \int_0^{+\infty} e^{2\beta x} |f(x)|^2 \, dx \quad , \qquad (111)$$

with $\beta > 0$, then, as shown in (Bertero and Pike, 1986), the linear operator $L : X \longrightarrow l^2$, defined by Eq.(110), is compact and one can use its singular system for the investigation of the Poisson transform inversion.

III. LINEAR INVERSE PROBLEMS WITH DISCRETE DATA

In most of the examples of the previous Chapter it is assumed that the data are known everywhere in some domain of the measured variable. For example it is assumed that the scattering amplitude or the diffraction pattern is known for all the values of the scattering angle in a given interval. Analogously it is assumed that the Fourier transform of the unknown function is known for all the frequencies in a given interval and so on. Such an assumption, however, does not provide a satisfactory model of real experimental situations. In practice one has only a finite number of detectors which can only measure the data function in a finite number of points. Therefore the output of an experiment is a set of (real or complex) numbers g_1 , g_2 , ..., g_N . These numbers can be viewed as the components of a vector which will be called the *data vector* and denoted by $\mathbf{g}$.

A. General Formulation

In the case of linear problems we can assume that the g_n are the values of prescribed linear functionals of the unknown solution (Bertero *et al* ., 1985 a). Consider, for example, the case where the physical quantity g(x) measured by the detectors is related by an integral operator to the unknown function f(y), so that the inverse problem is the solution of a first kind Fredholm integral equation, Eq.(6). Then, if the responses of the detectors are linear, their outputs g_n are proportional to the values of g at some points x_n. By neglecting the constant related to the efficency of the instrument we can write

$$g_n = g(x_n) = \int K(x_n,y)\, f(y)\, dy) \qquad (112)$$

Such an equation does not take into account that any detector integrates over some region in the domain of the physical variable x . When this effect cannot be neglected, Eq.(112) must be replaced by the following one

$$g_n = \int P_n(x) \left(\int K(x,y)\, f(y)\, dy \right) dx = \qquad (113)$$

$$= \int \left(\int P_n(x)\, K(x,y)\, dx \right) f(y)\, dy$$

Here $P_n(x)$ is an averaging function which describes the integration effect of the n-th detector and which has typically a peak centred near the experimental point $x = x_n$.

The r.h.s. of Eq.(112) or (113) is a scalar product in L^2 and therefore it defines a linear and continuous functional in this space. More generally, we can assume that the data g_n depend continuously on the object f and that the object space X is a Hilbert space (as concerns the choice of this space we recall the considerations developed in Chapt. II, Section A). Then, since according to Riesz representation theorem (Balakrishnan, 1976) any linear continuous functional on a Hilbert space X can be represented as a scalar product, we can summarize the remarks above as follows:

A) The design of an experiment for the indirect determination of a physical quantity f consists in specifying a finite set of linear, continuous functionals F_n , n=1,2,...,N. The output of the experiment is the set of the values g_n of these functionals F_n.

B) Given the object space X, if the functionals F_n are continuous on X, then

to any functional F_n is associated a function ϕ_n such that

$$F_n(f) = (f, \phi_n)_X \quad ; \quad n = 1,..., N \quad . \tag{114}$$

C) In the case where experimental errors or noise are neglected, then the *linear inverse problem with discrete data* corresponding to the experiment specified above consists in determining a function $f \in X$ satisfying the equations

$$g_n = (f, \phi_n)_X \quad ; \quad n = 1,..., N \quad . \tag{115}$$

The previous scheme applies also to problems where the data are intrinsically discrete, in the sense that, even in the ideal case, they do not depend on a continuous variable. A very simple example is provided by the Hausdorff moment problem already discussed in Chapter II, Section H. Another important example is the determination of the physical properties of a vibrating system (for instance the density of a vibrating string) from the knowledge of its eigenfrequencies. Such a problem, shortly discussed in the Introduction, has also interesting applications to the investigation of the structure of the Earth in the large. In such a case, a successful approach (Backus and Gilbert, 1968; 1970) consists in postulating an Earth model about which linearize the nonlinear inverse problem. The resulting linear problem is consistent with the general definition given above.

If we remember the definition of a generalized moment problem given in Chapt. II, Section H, we conclude that a linear inverse problem with discrete data is always a finite section of a generalized moment problem.

It is obvious that the information about the physical quantity f which can be extracted from Eq.(115) is incomplete. If we denote by X_N the

linear, finite-dimensional subspace spanned by the functions ϕ_n, then the data vector **g** depends only on the orthogonal projection of f onto X_N. Any function f which is orthogonal to X_N produces a zero data vector and therefore it cannot be recovered by means of the experiment specified by the functions ϕ_n. According to the general definitions given in Chapter II, Section A, the component of f orthogonal to X_N can be called the *invisible component* of f (Rust and Burrus, 1972; Bertero *et al.*, 1985 a) since this component cannot be detected by the experiment. In a mathematical language this means that *the solution of Eq.(115) is never unique.*

When the ϕ_n are linearly independent, in which case the dimension of the subspace X_N is exactly N, the component of f onto X_N, which is the *visible component* of f and which will be denoted by f^+, must be a linear combination of the ϕ_n and therefore it can be written in the following form

$$f^+ = \sum_{m=1}^{N} a_m \phi_m \qquad (116)$$

Substitution into Eq.(115) shows that the coefficients a_n must solve the linear system

$$\sum_{m=1}^{N} G_{mn} a_m = g_n \; ; \qquad n = 1, \ldots, N \qquad (117)$$

where the quantities

$$G_{mn} = (\phi_m, \phi_n)_X = G_{nm}^{*} \qquad (118)$$

are the elements of the so-called *Gram matrix*. If we denote by G^{mn} the elements of the inverse of the Gram matrix (remember that the Gram matrix is invertible if and only if the functions ϕ_n are linearly independent) which satisfy the relations

$$\sum_{l=1}^{N} G^{ml}\, G_{ln} = \sum_{l=1}^{N} G_{ml}\, G^{ln} = \delta_{mn} \qquad (119)$$

and we also introduce in X_N the dual basis, given by

$$\phi^n = \sum_{m=1}^{N} G^{nm}\, \phi_m \;\; ; \;\; n = 1, \ldots, N \qquad (120)$$

so that the sets $\{ \phi^n \}$ and $\{ \phi_n \}$ form a biorthogonal basis in X_N, then we have

$$f^+ = \sum_{n=1}^{N} g_n\, \phi^n \qquad , \qquad (121)$$

which is just the finite dimensional version of the solution (104) of a well-posed generalized moment problem. This representation clearly shows that f^+ depends continuously on the data. If $\delta\mathbf{g}$ is a small variation of $\mathbf{g}$ and δf^+ is the corresponding variation of f^+, then $\|\delta f^+\|_X$ tends to zero when $\delta\mathbf{g}$ tends to zero.

Continuous dependence of the solution on the data, however, does not imply numerical stability which is related to more deep properties of the Gram matrix of the functions ϕ_n. Even if the functions ϕ_n are linearly

independent, some (or many) of the ϕ_n can be *nearly parallel*. In such a case the problem of determining the component of f onto X_N exhibits numerical instability. In fact this problem is equivalent to invert the Gram matrix of the functions ϕ_n and, as it is known, the Gram matrix indicates how much the functions (vectors) ϕ_n depart from an orthogonal system : it becomes ill-conditioned when the vectors are close to a linearly dependent system. We recall that, in the finite dimensional case, ill-conditioning means essentially that the smallest eigenvalues cluster near zero while the others spread elsewhere.

In order to quantify the stability of the problem of determining the visible component of f it is necessary to introduce a measure of the errors on the data or, in other words, to introduce a metric in the data space Y. We assume that Y is an euclidean space with a scalar product defined by

$$(\mathbf{g},\mathbf{h})_Y = \sum_{n,m=1}^{N} W_{nm}\, g_m\, h_n^{*} \quad , \tag{122}$$

the weights W_{nm} being the matrix elements of a given positive matrix **W**.

The simplest choice of the weights is obviously $W_{nm} = \delta_{nm}$. On the other hand, in the case of least squares problems it is quite natural to relate **W** to the covariance matrix **C** of the errors on the components of the data vector. In linear regression theory the relation is

$$\mathbf{W} = \mathbf{C}^{-1} \tag{123}$$

and in such a case the choice $W_{nm} = \delta_{nm}$ corresponds to white noise. Another possible choice will be discussed in Section E in connection with the problem of moment discretization.

We can define now a linear operator L, from X into Y, which transforms a function of X into a vector of Y according to the rule

$$(Lf)_n = (f,\phi_n)_X \quad ; \qquad n = 1, \ldots, N \ . \tag{124}$$

The mapping is onto when the ϕ_n are linearly independent, otherwise the range of L is a subspace of dimension $N' < N$, if N' is the number of linearly independent ϕ_n. Moreover, in terms of the operator L, Eq. (115) can be written in the following way

$$\mathbf{g} = Lf \tag{125}$$

which has precisely the general form (5).

We consider, for simplicity, the case where the ϕ_n are linearly independent. Since L is a finite rank operator, we can always introduce its singular system $\{\sigma_k; u_k; \mathbf{v}_k\}$ which is the set of the solution of the shifted eigenvalue problem

$$Lu_k = \sigma_k \mathbf{v}_k \quad , \quad L^*\mathbf{v}_k = \sigma_k \ u_k \tag{126}$$
$$k = 0,1, \ldots ,N-1 \ .$$

The adjoint operator, defined by Eq.(16), transforms a vector of Y into a function of X. Its explicit expression is the following one

$$L^*\mathbf{g} = \sum_{n=1}^{N} \ (\sum_{m=1}^{N} W_{nm} \ g_m) \ \phi_n \ . \tag{127}$$

A few remarks about the computation of the singular system of L.

The singular vectors $\mathbf{v}_k$ are the eigenvectors, associated with the eigenvalues σ_k^2 , of the operator LL^*. This is an operator in Y and therefore it can be characterized by means of a matrix which will be denoted by $\hat{\mathbf{L}}$. Combining Eq.(124) with Eq.(127) it follows that

$$\hat{\mathbf{L}} = \mathbf{G}^T \mathbf{W} \tag{128}$$

where $\mathbf{G}^T$ denotes the transpose of the Gram matrix (notice that the rank of $\hat{\mathbf{L}}$ coincides with the rank of $\mathbf{G}$ because $\mathbf{W}$ is positive definite). It follows that the computation of the singular values and singular vectors is a standard eigenvalue problem. When this has been solved, the corresponding singular functions u_k can be obtained by means of the second of the equations (126) which, using Eq.(127), can be written explicitely as follows

$$u_k = \sigma_k^{-1} \sum_{n=1}^{N} \Big(\sum_{m=1}^{N} W_{nm} (\mathbf{v}_k)_m \Big) \phi_n \tag{129}$$

where $(\mathbf{v}_k)_m$ is the m-th component of the vector $\mathbf{v}_k$.

Since the singular vectors $\mathbf{v}_k$ form an orthonormal basis in Y while the singular functions u_k form an orthonormal basis in X_N, it is easy to obtain the following representation of the visible component of f

$$f^+ = \sum_{k=0}^{N-1} \sigma_k^{-1} (\mathbf{g}, \mathbf{v}_k)_Y \, u_k \quad . \tag{130}$$

As we have already remarked in the Introduction, the determination of f^+

is a well-posed problem and therefore the propagation of relative errors from the data to the solution is controlled by the condition number. If $\delta\mathbf{g}$ is a small variation of $\mathbf{g}$ and δf^+ the corresponding variation of f^+, then Eq.(10) holds true with

$$\text{cond}\,(L) = \sigma_0/\sigma_{N-1} \qquad (131)$$

This inequality is precise in the sense that equality can hold true. In spite of this fact, however, the condition number may be a rather pessimistic estimate of error propagation. Equality holds in a very special case which, in general, is not satisfied in practice. For this reason the condition number can be called the "worst magnification" of relative errors (Twomey, 1974). A more realistic estimate is given by the "average magnification" of relative errors which will be denoted by $\langle \text{cond}(L) \rangle$. Its expression is the following (Twomey, 1974)

$$\langle \text{cond}(L) \rangle = N^{-1} \Big(\sum_{k=0}^{N-1} \sigma_k^{\,2} \Big)^{1/2} \Big(\sum_{k=0}^{N-1} \sigma_k^{\,-2} \Big)^{1/2} \qquad (132)$$

It is easily verified that this quantity is always smaller than cond(L) but greater than cond(L)/N.

It is important to point out that the ill-conditioning of an inverse problem with discrete data usually derives from the fact that it is the discrete version of an ill-posed inverse problem. A few relevant examples will be discussed in the next Sections. From these examples it clearly appears that, when the number of data points increases, the ill-conditioning of the problem also increases. The reason is that, by increasing the number of data points, one obtains a sequence of discrete problems providing better and better approximations of the ill-posed,

infinite dimensional inverse problem, whose condition number is infinite. As a consequence, an increase of the data points produces an increase of the instability in the computation of the visible component of f, without a significant increase of the information content of the data, because the new functions ϕ_n are nearly parallel to the subspace spanned by the previous ones.

On the other hand, when the number of data points is sufficiently small, an inverse problem with discrete data can be well-posed. Since the number of data points cannot be too small (in such a case the information content of the data is too poor) it follows that there exists an optimum number of data points, which corresponds to a compromise between stability and information content. In other words there exists an optimum experiment for the determination of the desired physical quantity. To our knowledge, however, such a problem has not yet been solved both from the theoretical and from the practical point of view. Some interesting results in this direction have been obtained in the case of the finite Hausdorff moment problem (Talenti, 1987).

B. Fourier Transform Inversion with Discrete Data

As a first example of an inverse problem with discrete data we consider the problem of determining a function of bounded support when its Fourier transform is given in a finite number of points. As in Chapt. II, Section F, let us assume that the support of the function f is interior to the interval [-c,c]. When the points where $\hat{f}$ is known are $x_1, x_2, \ldots, x_n$, we have the following problem

$$g_n = \int_{-c}^{c} \exp(- i x_n y) f(y)\, dy \; ; \qquad n = 1, \ldots, N \; . \tag{133}$$

If we assume that $f \in L^2(-c,c)$ and that the data space Y is the usual euclidean space, i.e. the scalar product is defined by Eq. (122) with $W_{nm} = \delta_{nm}$, then the operator $\hat{\mathbf{L}}$, Eq. (128), coincides with the Gram matrix (which is symmetric) and the latter is given by

$$G_{nm} = 2c \, \mathrm{sinc}[c(x_n - x_m)/\pi] \quad . \tag{134}$$

The singular values σ_k of the problem are the square roots of the eigenvalues of $\mathbf{G}$ and the singular vectors $\mathbf{v}_k$ are the corresponding eigenvectors of $\mathbf{G}$. Then the singular functions u_k are given by

$$u_k(y) = \sigma_k^{-1} \sum_{n=1}^{N} (\mathbf{v}_k)_n \exp(-i x_n y) \quad . \tag{135}$$

Consider now the important case of uniformly spaced sampling points

$$x_n = x_1 + d(n-1) \; ; \; n = 1, \ldots, N \; , \tag{136}$$

where d is the sampling distance.

When $d = \pi/c$, the Fourier transform is sampled at the Nyquist rate. In such a case the functions $\phi_n(y) = \exp(-i x_n y)$ are orthogonal and the Gram matrix is a multiple of the unit matrix, $G_{nm} = 2 c \, \delta_{nm}$. The problem is well-conditioned and the solution (121) is a truncated Fourier series

$$f^+(y) = \sum_{n=1}^{N} g_n \exp(-i x_n y) \quad . \tag{137}$$

When $d < \pi/c$ the Fourier transform is over-sampled and the problem becomes ill-conditioned. A qualitative argument for explaining this fact is the following. If we fix the interval where the data are given, the optimum number of data points is obtained when the data are sampled at the Nyquist rate, as shown by the orthogonality of the functions ϕ_n in such a case. If we increase the sampling rate, we add more and more points which are less and less linearly independent from the previous ones and which therefore do not contain significant new information about the function f.

A quantitative analysis of the problem follows from the properties of the matrix

$$S_{nm} = 2W \, \mathrm{sinc}[2W(n-m)] \quad , \quad W < 1/2 \tag{138}$$

which can be obtained from the Gram matrix (134) (with x_n given by Eq.(136)) if we put $W = c\,d/2\pi$ and if we multiply G_{nm} by $d/2\pi$.

The matrix $\mathbf{S}$ has been studied by Slepian (Slepian, 1978) who denotes its eigenvalues by $\lambda_k(N,W)$, with $\lambda_0(N,W) > \lambda_1(N,W) > \ldots > \lambda_{N-1}(N,W)$, and defines the *discrete prolate spheroidal sequences* (DPSS) as the real solutions, for $k = 0, 1, \ldots, N-1$, of the system of equations

$$\sum_{m=1}^{N} S_{nm} \, v_m^{(k)}(N,W) = \lambda_k(N,W) \, v_n^{(k)}(N,W); \; n = 0, \pm 1, \pm 2, \ldots \tag{139}$$

Therefore the eigenvectors of $\mathbf{S}$ are obtained by index-limiting the DPSS to (1, N).

It follows that the singular values of the problem (133) are proportional to the square roots of the eigenvalues of $\mathbf{S}$

$$\sigma_k = [\, 2\pi\, \lambda_k(N,W)\, /d]^{1/2}\,; \qquad k = 0\,,\,\ldots,\,N-1 \tag{140}$$

while the singular vectors are just obtained by index-limiting the DPSS

$$(\mathbf{v_k})_n = v_n^{(k)}\,(N,W)\,; \qquad k = 0\,,\,\ldots,\,N-1;\, n = 1,\,\ldots,\,N \tag{141}$$

since, according to the definition given by Slepian, the euclidean norm of these vectors is one. Finally the singular functions $u_k(y)$ are related to the *discrete prolate spheroidal wave functions* (DPSWF) which are defined by Slepian as follows

$$U_k(N,W;\,x) = \epsilon_k \sum_{n=1}^{N} v_n^{(k)}(N,W)\, e^{-i\,\pi(N\,+\,1\,-\,2n)x} \tag{142}$$

where $\epsilon_k = 1$, when k is even, and $\epsilon_k = i$, when k is odd. In fact, by comparing Eq.(142) with Eq.(135) (the x_n being defined in Eq. (136)), one finds that

$$u_k(y) = (\sigma_k \epsilon_k)^{-1}\, \exp\{i[x_1+(N-1)d/2]\}\;\; U_k(N,W;dy/2\pi) \tag{143}$$

As proved in (Slepian, 1978) the DPSWF are simultaneous eigenfunctions of an integral and of a differential operator (a property analogous to a basic property of the PSWF); they are doubly orthogonal in the sense that they are orthogonal both with respect to the scalar product of $L^2(-W,\, W)$ and with respect to the scalar product of $L^2(-\frac{1}{2},\, \frac{1}{2})$; moreover $U_k(N,\, W;\, x)$ is an even or odd function of f according to the parity of k, has exactly k zeros in the open interval (-W, W) and has exactly N-1 zeros in the interval $(-\frac{1}{2},\, \frac{1}{2})$.

In Fig.1 we give a plot of the DPSWF in the case N=5 and W = 0.2.

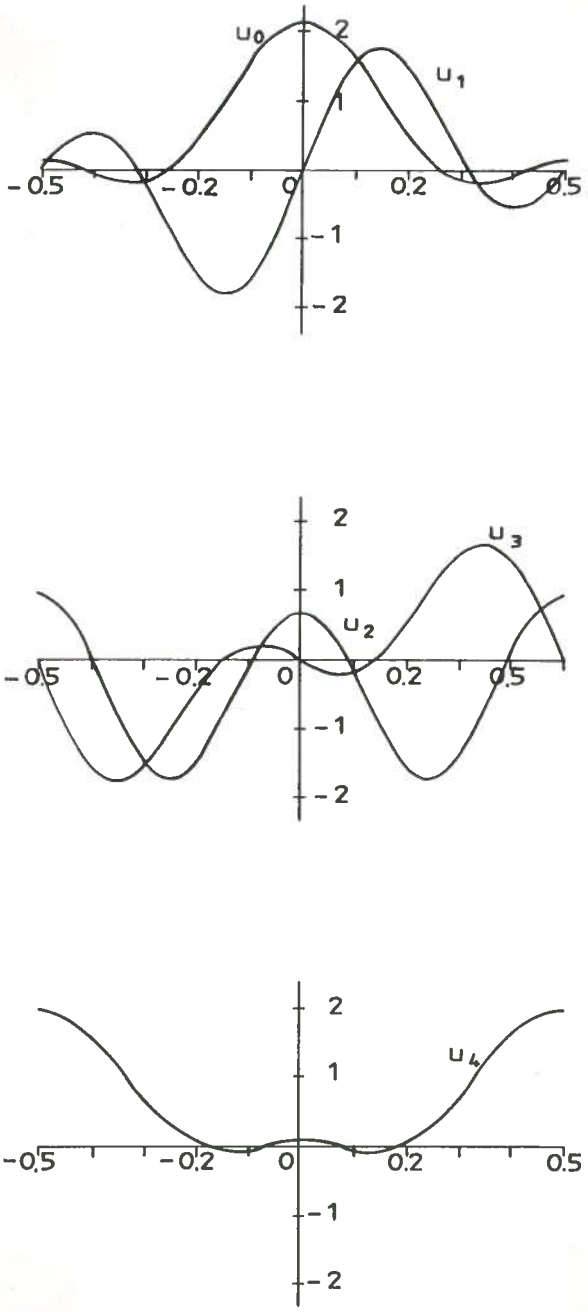

Fig.1 Plot of the discrete prolate spheroidal wave functions (DPSWF) in the case N=5 and W=0.2.

The corresponding singular values (square roots of the $\lambda_k(N,W)$) are: σ_0 = 0.993, σ_1 = 0.875, σ_2 = 0.481, σ_3 = 0.128, σ_4 = 0.0165. Asymptotics of the eigenvalues and eigenfunctions are given in (Slepian, 1978). In

particular it is proved that, when $W \longrightarrow 0$ and $N \longrightarrow \infty$, in such a way that $\pi NW \longrightarrow c$, then

$$\lambda_k(N,W) \longrightarrow \lambda_k \quad , \qquad \sqrt{W}\, U_k(N,W;Wx) \longrightarrow \psi_k(c,x) \tag{144}$$

where $\psi_k(c,x)$ is the PSWF of order k and λ_k the corresponding eigenvalue. If we recall the behaviour of the eigenvalues of the PSWF, this property shows that, when the number of sampling points in a fixed interval is large, the number of singular values nearly equal to one is approximately equal to the number of sampling points corresponding to the Nyquist rate.

Finally, when $d > \pi/c$, the Fourier transform is sampled below the Nyquist rate and the problem becomes well-conditioned since each sampling point contains an independent piece of information. For example, in the case N=5 and W=0.8, the square roots of the eigenvalues of the matrix (138) are: σ_0= 1.414, σ_1= 1.408, σ_2 = 1.330, σ_3 = 1.111, σ_4 = 1.007, with a condition number equal to 1.404.

C. Interpolation and Numerical Derivation

Interpolation is the problem of determining a function f when its values are given in a finite number of points. Analogously numerical derivation is the problem of determining f' from the same data. These are classical problems of numerical analysis and have recently found interesting applications to inverse problems in computational vision (Grimson, 1982; Torre and Poggio, 1986; Bertero *et al.*, 1986 b).

We show the relationship between these problems and the general formulation of an inverse problem with discrete data in the simple case of functions depending on one variable. Then the problem is the determination of a function f(x), defined on the interval [a,b], which takes prescribed

values g_n at the N points $x_1, x_2, \ldots, x_N$ with $a \leq x_1 < x_2 < \ldots < x_N \leq b$:

$$g_n = f(x_n) \; ; \qquad n = 1, \ldots, N \tag{145}$$

When we have a differentiable solution f(x) of this problem then a solution of the corresponding problem of numerical derivation is just f'(x). It must be pointed out, however, that the problem of numerical derivation can be formulated independently of the corresponding problem of interpolation. If we put $f'(x) = h(x)$ and we assume that $x_1 = a$ then the problem is

$$g_n - g_1 = \int_a^{x_n} h(x)\, dx \; ; \qquad n = 2, \ldots, N \tag{146}$$

which is already in the form (115), at least when $X = L^2(a,b)$.

The interpolation problem (145) can be formulated in the form (115) if the space X of the solutions is a *reproducing kernel Hilbert space* (RKHS), i.e. a Hilbert space of continuous functions such that all the evaluation functionals are continuous. In fact, if $f \in X$ then, from the Riesz representation theorem (Balakrishnan, 1976) it follows that, for a given $x \in [a,b]$, there exists a function $Q_x \in X$ such that

$$f(x) = (f, Q_x)_X \tag{147}$$

and this shows that the problem (145) has the form (115), the functions ϕ_n being the functions Q_x associated with the points x_n. The symmetric kernel

$$Q(x,x') = (Q_x, Q_{x'})_X = Q_x(x') = Q_{x'}(x) \tag{148}$$

is called the *reproducing kernel* of X (Aronszajn, 1950). Once the RKHS has

been choosen, the Gram matrix of the interpolation problem is given by

$$G_{nm} = Q(x_n, x_m) \quad . \tag{149}$$

We consider now two examples.

1. *Interpolation of Band-Limited Functions*

The first example is the interpolation of a band-limited function. Let X be the Paley-Wiener space PW_c already discussed in Chapter II, Section F. Then, if B_c is the band-limiting operator (80), for any $f \in X$ we have $B_c f = f$ and this shows that PW_c is a RKHS with the reproducing kernel

$$Q(x,x') = (\pi/c)\, \mathrm{sinc}[c(x-x')] \quad . \tag{150}$$

This interpolation problem is strictly related to the problem (133) of Fourier transform inversion with discrete data, since one can solve the latter just by determining a band-limited function which interpolates the data values g_n and then by taking the inverse Fourier transform of the result. As a consequence the two problems have the same Gram matrix, except for a multiplicative factor, as follows from Eqs.(149) and (150). Moreover, the singular values of this interpolation problem are the singular values of the Fourier inversion problem, multiplied by $(2\pi)^{-1/2}$, the singular vectors are the same and the singular functions are given by

$$u_k(y) = \sigma_k^{-1} \sum_{n=1}^{N} (\mathbf{v}_k)_n\, (\pi/c)\, \mathrm{sinc}[c(x-x_n)] \quad . \tag{151}$$

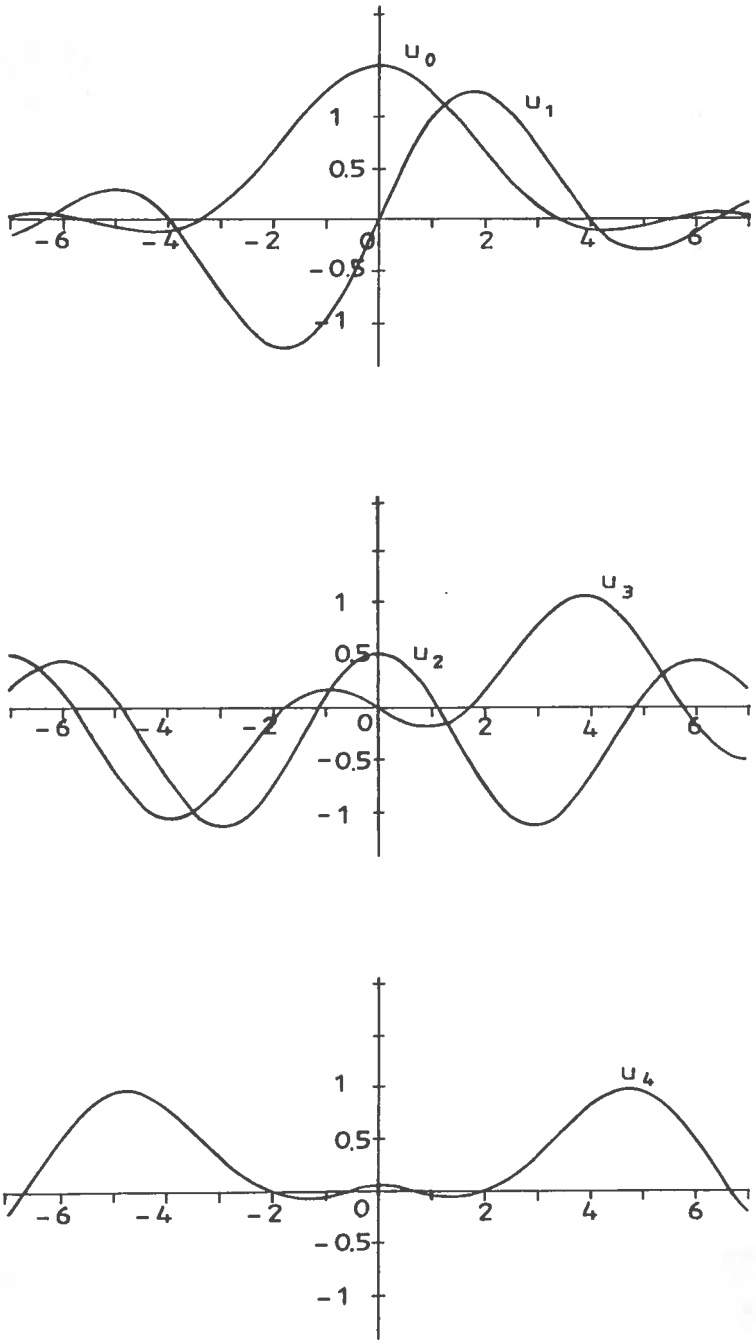

Fig.2 Plot of the singular functions for the interpolation of a band-limited function, with bandwidth $c=2\pi W$, $W=0.2$, given at the points $0, \pm 1, \pm 2$.

Notice that these singular functions are just the inverse Fourier transforms of the singular functions (135) (multiplied by a suitable

normalization factor).

In the case of equidistant points, Eq.(136), the analysis runs parallel to that of the Fourier transform inversion. In particular, if $d=\pi/c$, the solution (121) is the interpolating function obtained by truncating the Whittaker-Shannon expansion (82) and it is exactly the inverse Fourier transform of the solution (137).

When $d < \pi/c$ the singular system can be expressed again in terms of the eigenvalues and eigenvectors of the Slepian matrix (138). As follows from Eqs.(151) and (141) the singular functions are the interpolating functions of the index-limited DPSS or also the inverse Fourier transforms of the singular functions (143). Notice that the singular functions of the interpolation problem are even or odd, according to the parity of k, only when the set of the data points is symmetric, i.e. $x_1 = -d(N-1)/2$. In Fig.2 we give a plot of these singular functions in the case N=5, W=0.2, where $W=cd/2\pi$. We notice the similarity of these functions and of the DPSWF of Fig.1. This is explained by the limiting property (144) if one recalls that the Fourier transform of a PSWF is still a PSWF.

As the problem of Fourier transform inversion, the interpolation problem is also ill-conditioned when $d < \pi/c$ and this is related to the fact that the sampling process is made as efficient as possible when data are sampled at the Nyquist rate. It must be observed however that efficient reconstructions of oversampled functions can be obtained using generalized sinc-series (Campbell, 1968; Natterer, 1986 b). These series allow the evaluation of f(x) using a small number of sampling points in the neighborhood of x and therefore they provide a nearly local method. It may be that these generalized sinc-series are a sort of regularization of the ill-conditioned interpolation problem, but they have not yet been investigated from this point of view.

Finally the case $d > \pi/c$ has the same properties of the corresponding case in the problem of Fourier transform inversion.

The second example is the interpolation of a function, defined on the interval [0,1], which has square integrable derivatives up to the order k. For simplicity we consider only the cases k=1 and k=2. The case of arbitrary k is discussed in (Bertero *et al.*, 1985 a).

Let X be the space of the continuous functions with a square integrable first derivative - this space is usually denoted by $H^1(0,1)$ - and let us introduce in X the scalar product

$$(f,\phi)_X = f(0)\phi(0) + \int_0^1 f'(x)\,\phi'(x)\,dx \quad . \tag{152}$$

Then, using Taylor formula, it is easy to show that the reproducing kernel is given by

$$Q(x, x') = 1 + x - (x - x')_+ = 1 + \min[x, x'] \tag{153}$$

where x_+ denotes the function which is zero when $x < 0$ and is equal to x when $x > 0$. Since $\phi_n(x) = Q(x,x_n)$, it is obvious that, in such a case, the solution (121) is just the linear interpolation of the data values g_n.

Let X be now the space of the continuous functions with a square integrable second derivative - this space is usually denoted by $H^2(0,1)$ - and let us introduce in X the scalar product

$$(f,\phi)_X = f(0)\phi(0) + f'(0)\phi'(0) + \int_0^1 f''(x)\,\phi''(x)\,dx \quad . \tag{154}$$

Then X is a RKHS and, using again Taylor formula for a function $f \in X$, it is

easy to show that the reproducing kernel is

$$Q(x,x') = 1 + xx' + x^2x'/2 - x^3/6 + (x-x')_+^3/6 \qquad (155)$$

If we recall that the class $S_m(x_1, x_2,\ldots, x_N)$ of the *spline functions* of degree m, having the knots $x_1, x_2, \ldots, x_N$, is the set of all the functions s(x) which have the following representation

$$s(x) = p(x) + \sum_{n=1}^{N} c_n (x - x_n)_+^m \qquad (156)$$

where p(x) is an arbitrary polynomial of degree $\leq$ m (Greville, 1969), we see that the functions $\phi_n(x) = Q(x, x_n)$ are spline functions of degree 3 (cubic splines) and that the subspace X_N spanned by the functions $\phi_n(x)$ is a linear subspace of $S_m(x_1, x_2,\ldots, x_N)$. Therefore the interpolation provided by the solution (121) or (130) is just an interpolation in terms of cubic splines. Interpolation in terms of natural cubic splines (Greville, 1969) is obtained by minimizing the L^2-norm of the second derivative of f. Since this is a seminorm, this kind of problem will be discussed in the next Chapter, Section B.

Finally, if we consider the problem (146) in the space (152), the corresponding generalized solution is obtained by interpolating the data values g_n in terms of cubic splines and then by differentiating the result (Bertero *et al.*, 1986 a).

Interpolation of functions of 2 variables in RKHS are investigated in (Duchon, 1976; Wahba and Wendelberger, 1980).

D. Finite Hausdorff Moment Problem

We consider a finite section of the Hausdorff moment problem and therefore we want to determine a function f(x) defined on the interval [0,1] from the knowledge of its first N moments:

$$g_n = \int_0^1 x^{n-1} f(x)\, dx ; \qquad n = 1, \ldots, N . \tag{157}$$

Let X be $L^2(0,1)$ and Y the usual N-dimensional euclidean space. Then the problem has the form (115) with $\phi_n(x) = x^{n-1}$. It follows that X_N is the subspace of the polynomials of degree N-1 and that the solution (121) is also a polynomial of degree N -1.

A well-known procedure which can be used for the computation of this solution consists in representing $f^+(x)$ in terms of shifted Legendre polynomials $L_n(x)$ (Papoulis, 1956)

$$f^+(x) = \sum_{m=1}^{N} c_m \, L_{m-1}(x) \qquad . \tag{158}$$

We recall that the shifted Legendre polynomials are uniquely defined (except for the sign) by the following properties: a) the degree of $L_n(x)$ is exactly n; b) $(L_n, L_m)_X = \delta_{nm}$ (n,m = 0,1,2. . .), the scalar product being that of $L^2(0,1)$. The relation between the shifted Legendre polynomial $L_n(x)$ and the usual Legendre polynomial $P_n(x)$ is: $L_n(x) = (2n+1)^{\frac{1}{2}} P_n(2x - 1)$.

Now, by substituting the representation (158) into Eq.(157) and by taking into account the properties of the $L_n(x)$ one finds that

$$g_n = \sum_{m=1}^{n} \beta_{nm}\, c_m\,; \qquad n = 1, \ldots, N \tag{159}$$

where

$$\beta_{nm} = \int_0^1 x^{n-1}\, L_{m-1}(x)\, dx = \tag{160}$$

$$= (2m-1)^{1/2}\, (-1)^{m-1}\, [(n-1)!]^2 \,/\, [(n-m)!\,(n+m-1)!]$$

Since in Eq.(159) the summation index goes from 1 to n, it is possible to obtain recursively the coefficients c_n from the moments g_n and the coefficient c_n depends only on the moments of order $\leq n$. Therefore it is not necessary to change the algorithm when the number N of given moments is changed.

The previous algorithm is very simple but also very ill-conditioned even when N is moderately large (for instance, N=10). This can be shown by introducing the singular system of the problem. The Gram matrix is now given by

$$G_{nm} = (n + m - 1)^{-1}\,; \qquad n, m = 1, \ldots, N \tag{161}$$

and this is a well-known example of an ill-conditioned matrix. It is called *Hilbert matrix*, it is denoted by $\mathbf{H}_N(-1)$ and it is often used for testing numerical algorithms (Gregory and Karney, 1969). A remarkable property of this matrix is that it commutes with a tridiagonal matrix (Grünbaum, 1982) so that its eigenvectors can be easily computed.

It follows that the singular values of the finite moment problem are just the square roots of the eigenvalues of $\mathbf{H}_N(-1)$, the singular vectors are the corresponding eigenvectors of $\mathbf{H}_N(-1)$ and the singular functions

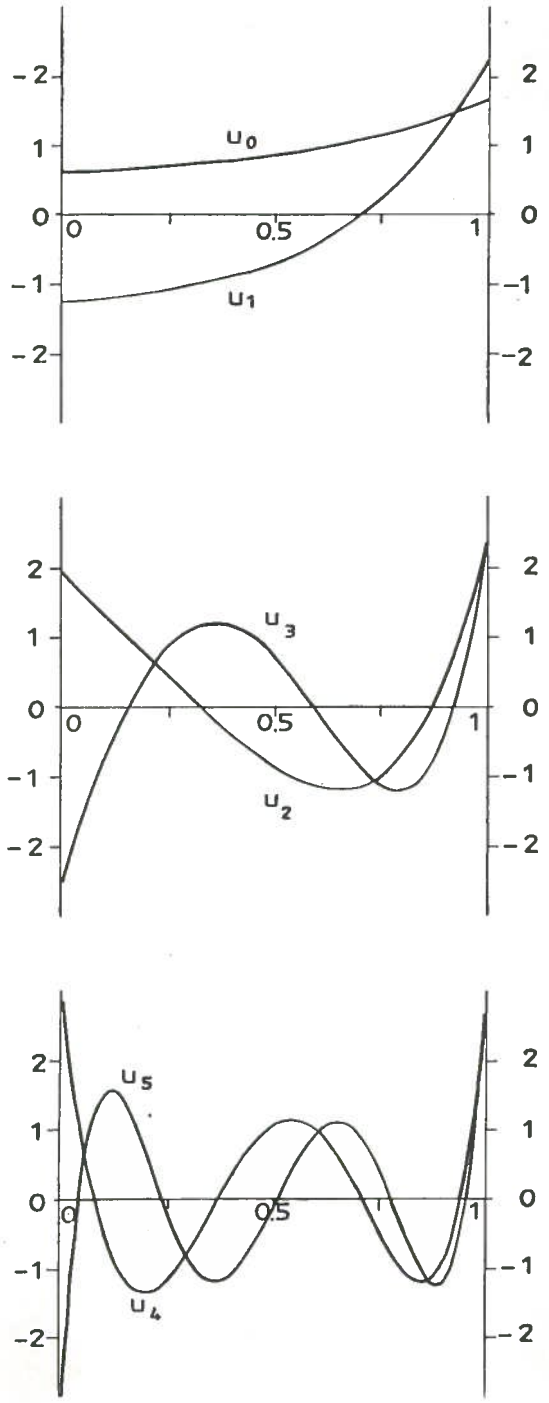

Fig.3 Plot of the singular functions of the finite Hausdorff moment problem in the case N=6.

are orthogonal polynomials of degree N-1, obtained by means of Eq.(129).

In order to give a numerical example of the ill-conditioning of the

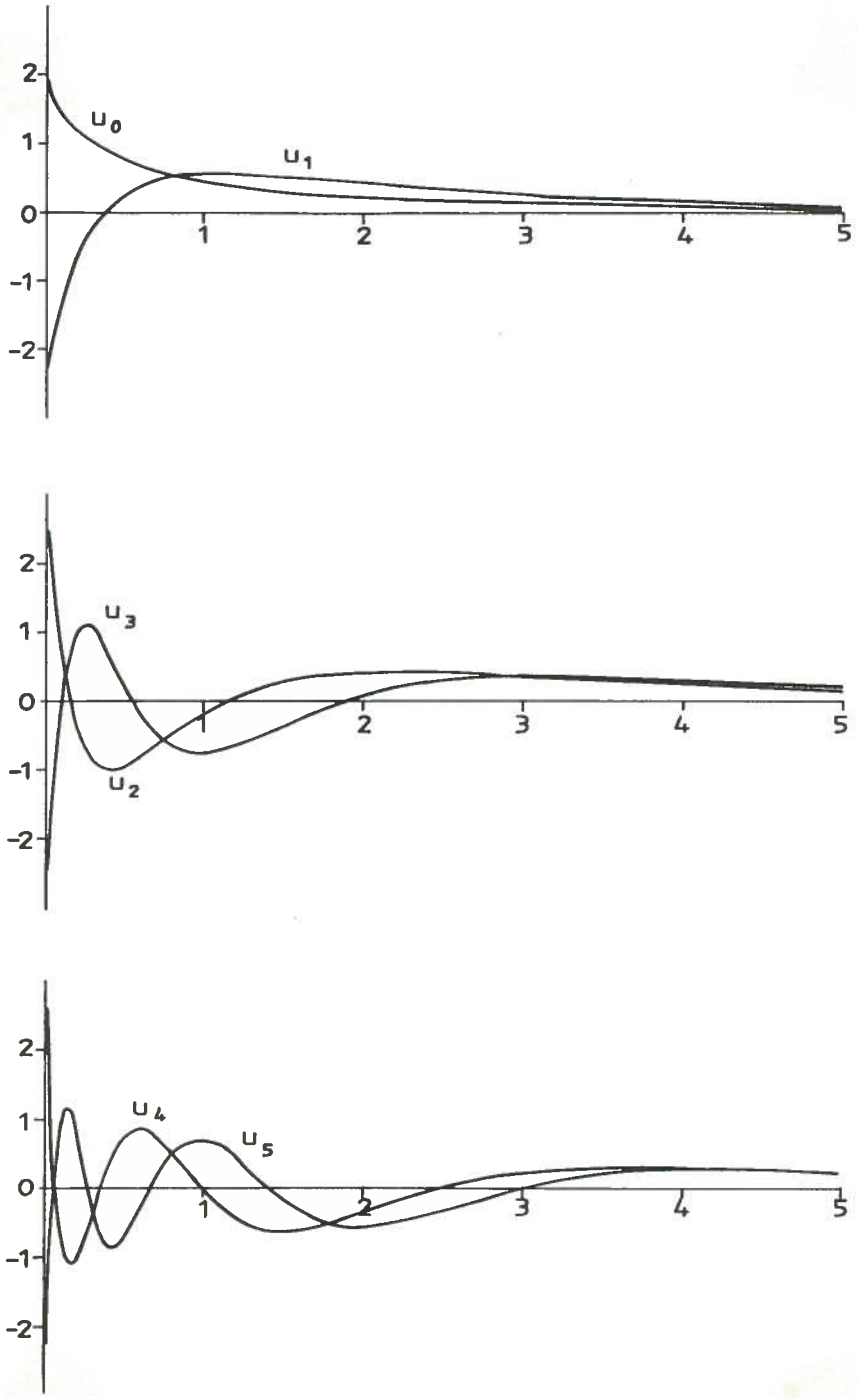

Fig.4 Plot of the singular functions of the inversion of the Laplace transform, given at the points $p_n = n - 1/2$, with $n = 1,2,...,6$.

moment problem we give the singular values in the case N=6 : σ_0= 1.2724, σ_1=0.4923, σ_2=0.1277, σ_3=0.2481 x 10^{-1}, σ_4=0.3545 x 10^{-2}, σ_5=0.3291 x

10^{-3} with a condition number cond(L) = 3866. The plot of the corresponding singular functions is given in Fig.3. For large N, the asymptotic estimate of the condition number is (Gregory and Karney, 1969)

$$\text{cond}(L) = e^{1.75\,N} \tag{162}$$

and therefore it increases very rapidly for increasing N. We have here a very clear example of the relation between the ill-conditioning and the number of data points, as discussed at the end of Section A.

Another problem related to the moment problem is the inversion of the Laplace transform when this is given at the points $p_n = n - 1/2$ with n= 1,2,. ., N. This relation has been already discussed in Chapter II, Section H. In Fig.4 we give the singular functions of this problem, always in the case N=6. These singular functions can also be obtained from the singular functions of Fig.3 by means of the change of variables which transforms Eq.(107) into Eq.(106). For a more general discussion of the Laplace and of the finite Laplace transform inversion in the case of equidistant points and of geometrically distributed points see (Bertero *et al.* , 1985 b).

E. Moment-Discretization of First Kind Fredholm Integral Equations

Consider a first kind Fredholm integral equation, such as that given in Eq.(6), with a continuous kernel K(x,y) and a continuous data function g(x). Assume also that the intervals [a,b] and [c,d] are bounded. Then we denote by L_0 the integral operator associated with the kernel K(x,y), Eq.(7). As we know, this is a compact operator from $L^2(a,b)$ into $L^2(c,d)$. We denote its singular system by $\{\sigma_{0,k}; u_{0,k}, v_{0,k}\}$.

The method of *moment-discretization* , which provides a natural way for approximating a solution of Eq.(6) is as follows (Nashed, 1976 b;

Nashed and Wahba, 1974; Groetsch, 1984): given a finite set of points $a \leq x_1 < x_2 < \ldots < x_N \leq b$, seek for a function f(y) which satisfies the equations

$$g(x_n) = \int_a^b K(x_n,y)\, f(y)\, dy \; ; \qquad n = 1,\ldots,N \; . \tag{163}$$

This is a problem in the form (115), with $\phi_n(y) = K(x_n,y)$, if $X = L^2(a,b)$.

In order to investigate the relationship between the singular system of the integral operator L_0 and the singular system of the operator L associated with the problem (163), as defined by Eq.(124), we assume that the points x_n are the knots of some quadrature formula. We denote by w_1, $w_2, \ldots, w_N$ the corresponding weights and we introduce in the data space Y a scalar product defined as in Eq.(122) with $W_{nm} = w_n\, \delta_{nm}$.

We recall now that the singular values of L_0 are the square roots of the eigenvalues of the integral operator $L_0^* L_0$ whose kernel is given by

$$T_0(y,y') = \int_c^d K(x,y)\, K(x,y')\, dx \tag{164}$$

and that the singular functions $u_{0,k}$ are the corresponding eigenfunctions. Analogously the singular values of L are the square roots of the eigenvalues of the finite rank integral operator $L^* L$ whose kernel is

$$T(y,y') = \sum_{n=1}^{N} w_n\, K(x_n,y)\, K(x_n,y') \tag{165}$$

as follows from Eqs.(124) and (127). Again the singular functions u_k are the corresponding eigenfunctions. Clearly $T(y,y')$ is just the approximation of $T_0(y,y')$ provided by the quadrature formula corresponding to the knots x_n and the weights w_n.

Since $K(x,y)$ is continuous, the kernel $T(y,y')$ converges to the kernel $T_0(y,y')$ when $N \longrightarrow \infty$ and the maximum distance between adjacent knots tends to zero. Using this result and well-known perturbative lemmas - more precisely the Weyl-Courant lemma (Riesz and Nagy,1972)- it is possible to prove (Bertero, 1986) that the singular values σ_k of the problem (163) converge to the singular values $\sigma_{0,k}$ of the integral operator L_0. Analogously the singular functions u_k converge, in the norm of $L^2(a,b)$, to the corresponding singular functions $u_{0,k}$. Since the singular values $\sigma_{0,k}$ tend to zero when $k \longrightarrow \infty$, the previous result implies that the ill-conditioning of the problem (163) increases when the number of the data points increases. We have another example of the property discussed at the end of Section A.

A special kind of moment-discretization can also be used for the approximate solution of first kind integral equations associated with operators of the following type

$$(L_0 f)(x) = \int_{-\infty}^{+\infty} K(x-y)\, P(y)\, f(y)\, dy \quad , \qquad -\infty < x < +\infty \tag{166}$$

where $K(x)$ is a band-limited function with bandwidth c. This operator is compact whenever the functions $K(x)$ and $P(y)$ are square integrable. An example is provided by the operator (87), in which case $K(x) = (\pi/c)\mathrm{sinc}(cx/\pi)$ and $P(y)$ is the characteristic function of the interval $[-1,1]$. Another important example is an operator whose inversion is related

to the problem of data processing in confocal scanning microscopy and which is obtained from Eq.(166) by taking $K(x) = P(x) = \mathrm{sinc}(x)$ (Bertero *et al.*, 1984).

Since the functions in the range of L_0 are band limited, it is quite natural to sample the data at the Nyquist rate, so that we can consider the problem

$$g(x_n) = \int_{-\infty}^{+\infty} K(x_n - y)\, P(y)\, f(y)\, dy \qquad (167)$$

$$x_n = n(\pi/c) \;\; ; \;\; n = 0, \pm 1, \ldots, \pm N$$

and take $w_n = \pi/c$. Then the kernel $T_0(y,y')$ and $T(y,y')$ are given respectively by

$$T_0(y,y') = P(y) \left(\int_{-\infty}^{+\infty} K(x-y)\, K(x-y')\, dx \right) P(y') \qquad (168)$$

and

$$T(y,y') = P(y) \left(\sum_{n=-N}^{N} (\pi/c)\, K(x_n - y)\, K(x_n - y') \right) P(y') \qquad (169)$$

Using the fact that the trapezoidal rule is exact for bandlimited functions in the case $N=\infty$, it is possible to prove that $T(y,y')$ converges to $T_0(y,y')$ in the L^2-norm when $N \longrightarrow \infty$ (with fixed sampling distance). It follows again that the singular values of the problem with discrete data converge to the singular values of the integral operator.

It is important to point out that, in some cases, it is possible to obtain very good approximations of the first singular values using a very

small number of data points. As an example we give a numerical result obtained in the case of the integral operator mentioned above and related to confocal scanning microscopy. In such a case it is possible to derive analytic expressions of the singular values and singular functions (Gori and Guattari, 1985). From this result it follows that the first five singular values are given by $\sigma_{0,0}$= 0.821898, $\sigma_{0,1}$=0.450158, $\sigma_{0,2}$= 0.206417, $\sigma_{0,3}$=0.150053, $\sigma_{0,4}$= 0.109845. On the other hand, in (Bertero *et al.*, 1987) it is shown that, using five sampling points, the corresponding five singular values are: σ_0 = 0.821807, σ_1 =0.413815, σ_2 =0.206327, σ_3 = 0.136873, σ_4 =0.109751. As concerns the even singular values the error is at most one unit on the fourth digit !

and which now are called ill-posed or incorrectly posed, were just considered as mathematical anomalies and, for this reason, they were not seriously investigated. Recent developments in physics and especially in applied physics have shown that ill-posed problems can also be related to extremely important physical situations. For example, the solution of the Cauchy problem for elliptic equations may have interesting applications in electrocardiography (Colli Franzone *et al*, 1977), namely in the reconstruction of the epicardial potential from body surface maps. We want to emphasize, however, another ill-posed problem which has revolutionized diagnostic radiology.

In 1971 the first clinical machine for the detection of head tumors, based on a new X-ray technique called computer-assisted tomography or also *computerized tomography* was installed at the Atkinson Morley's Hospital, Wimbledon. In 1979 Allan M. Cormack and Godfrey N. Hounsfield awarded the Nobel prize in Medicine for the invention of this technique. As everybody knows, computerized tomography provides images of cross sections of the human body by measuring the attenuation of the X-rays along a large number of lines through the cross section. Then the processing of the data requires the reconstruction of a function of two variables from the knowledge of its line integrals. The solution of this mathematical problem was already contained in a paper of Johann Radon (Radon, 1917). The result of Radon was even more general, since he proved formulas for the reconstruction of a function on $\mathbb{R}^n$ from the knowledge of its integrals over all the hyperplanes of $\mathbb{R}^n$. In honour of this contribution, the mapping which transforms a function into the set of its integrals over hyperplanes is now called *Radon transform* and therefore the problem of tomography is just a special case of Radon transform inversion. Moreover, Radon transform inversion is a beautiful example of an ill-posed problem since the solution does not exist for arbitrary data and the dependence of the solution on the data, in general, is not continuous. As a consequence, the effect of the noise on the solution is amplified in a way similar to that discussed in the case of the Cauchy problem for the Laplace equation.

IV. GENERALIZED SOLUTIONS

As it is known, the study of Eq.(5) is a rather intricate problem even in the case where this equation is an n x m linear system, i.e. the operator L is a matrix with n rows and m columns. Existence and uniqueness of the solution depend on the number n of rows, on the number m of columns and on the rank p of the matrix (Lanczos, 1961). A great simplification is obtained by introducing the so-called *Moore-Penrose generalized inverse* of a matrix, which is deeply related to the problem of looking for a least squares solution of minimal norm of the original linear system. Such a solution, which is also called the generalized solution of the linear system, always exists and is unique, independently of the number of rows, of the number of columns and of the rank of the matrix. A very important fact for the treatement of linear inverse problems is that the concept of generalized (Moore-Penrose) inverse can also be extended to the case of linear continuous operators in Hilbert spaces (Nashed, 1976 a; Groetsch, 1977). In such a case, however, the generalized inverse is not always continuous and therefore the problem of determining the generalized solution may be ill-posed.

The essential result is that the generalized solution is unique, exists for arbitrary data g and depends continuously on g when the range of the operator L is closed, even if the requirements of existence and uniqueness of the solution are not satisfied by the original equation (5). On the other hand, when the range of L is not closed (examples are provided by the compact operators and by the convolution operators discussed in Chapt. II, Section A) the generalized solution is unique but it does not exist for arbitrary data and it does not depend continuously on the data. In some mathematical literature the term ill-posed is deserved to this case. According to this definition, inverse problems with discrete data are always well-posed (but, of course, they can be very ill-conditioned) and therefore the concept of ill-posedness is restricted to the case of problems formulated in infinite dimensional spaces.

In this Section we sketch the procedure which leads to the introduction of the Moore-Penrose generalized inverse for a linear continuous operator and we also provide a physical interpretation of this procedure in terms of the concepts of visible and invisible components of the object, introduced in Chapt. II, Section A. We also introduce an extension of the Moore-Penrose generalized inverse which is obtained when the criterion for selecting a least squares solutions is not the minimization of a norm, but the minimization of a seminorm. Finally we sketch the Backus-Gilbert method for linear inverse problems with discrete data in order to clarify the analogies and differences between this method and the method of generalized solutions.

A. Moore-Penrose Generalized Inverse

The solution of Eq.(5) may not exist for arbitrary data g because R(L) is a subspace of Y. In the case of the band-limiting operator B_C, Eq.(80), for example, Y is the space of the square integrable functions and R(L) is the closed subspace of the band-limited functions with a fixed bandwidth. Similarly, in the case of a problem with discrete data, R(L) is a subspace of Y if the functions ϕ_n are not linearly independent.

When R(L) is a subspace of Y, the measured data may have a component orthogonal to R(L), as an effect of the noise contribution h in Eq.(9). For example, in the case of the bandlimiting operator, the noise may have Fourier components out of the band of the ideal low-pass filter described by the operator B_C. Under these circumstances, the solution of Eq.(5), with g given by Eq.(9), does not exist. Then a quite natural procedure is to look for the function (or the functions) u such that Lu is as close as possible to g.

The projection operator onto R(L) will be denoted by P, so that,

recalling the relations (18), we conclude that $Q = I - P$ is the projection operator onto $N(L^*)$. Moreover the following relations are obvious

$$PL = L\,, \qquad QL = 0\,. \tag{170}$$

We give now the following definition : a function $u \in X$ is said to be a *least squares solution* (or *pseudosolution*) of Eq.(5) if it minimizes the distance between Lf and g:

$$\|Lu - g\|_Y = \inf \{\|Lf - g\|_Y \mid f \in X\}\,. \tag{171}$$

By considering the first variation of the functional $\|Lf - g\|_Y$, it is quite easy to see that any least squares solution must satisfy the orthogonality condition

$$(Lu - g, L\phi)_Y = 0 \tag{172}$$

for any $\phi \in X$. It follows that u must be a solution of the Euler equation

$$L^*Lu = L^*g\,. \tag{173}$$

On the other hand, if we use the relations (170) we can write

$$\|Lf - g\|_Y^2 = \|Lf - Pg\|_Y^2 + \|Qg\|_Y^2 \tag{174}$$

and therefore we see that a least squares solution exists if and only if there exists a function which annihilates the first term of the r.h.s. of this equation. We conclude that Eq.(173) has a solution if and only if the equation

$$Lu = Pg \tag{175}$$

has also a solution. In other terms, the process which consists in replacing the usual solutions with least squares solutions is equivalent to project the measured data onto R(L) and then to solve Eq.(5) with g replaced by the projected data Pg.

The advantage of Eq.(173) with respect to Eq.(175) is that, in practice, it can be difficult to determine the projection operator P while, in general, it is quite easy to determine the adjoint operator L^*.

From the previous remarks it follows that: *a least squares solution of Eq.(5) exists if and only if Pg ∈ R(L).*

As a corollary we have that, when R(L) is closed, a least squares solution exists for any measured data g. This result applies, for example, to the case of inverse problems with discrete data and also to the case of the band-limiting operator (80).

As concerns uniqueness, the solution of Eq.(175) is unique if and only if N(L) = {0}. When N(L) is not trivial, let us denote by S(g) the set of all the least squares solutions associated with g. If $u^{(0)}$ is one of these least squares solutions, then S(g) is the closed affine subspace given by

$$S(g) = \{u \in X \mid u = u^{(0)} + \phi, L\phi = 0\} \tag{176}$$

i.e. S(g) is a translation of N(L), $S(g) = u^{(0)} + N(L)$.

For example, in the case of the band-limiting operator (80), S(g) is the set of all the square integrable functions whose Fourier transform coincides with the Fourier transform of g over the band [-c,c] and is arbitrary elsewhere. In a similar way, in the case of an inverse problem with discrete data, S(g) is the set of all the functions whose component in X_N is given by Eq.(121), while the component orthogonal to X_N is arbitrary.

Now, for any g such that Pg ∈ R(L), S(g) is not empty and it is a closed and convex set of X. Then, from a general theorem of functional analysis (Balakrishnan, 1977) it follows that there exists a unique solution of minimal norm. This is called the *generalized solution* (or also the

normal pseudosolution) and is denoted by f^+:

$$\|f^+\|_X = \inf \{\|u\|_X \,|\, u \in S(g)\} \quad . \tag{177}$$

It is easy to see that f^+ is the unique least squares solution which is orthogonal to N(L) (see Fig.5) and therefore this procedure is equivalent to

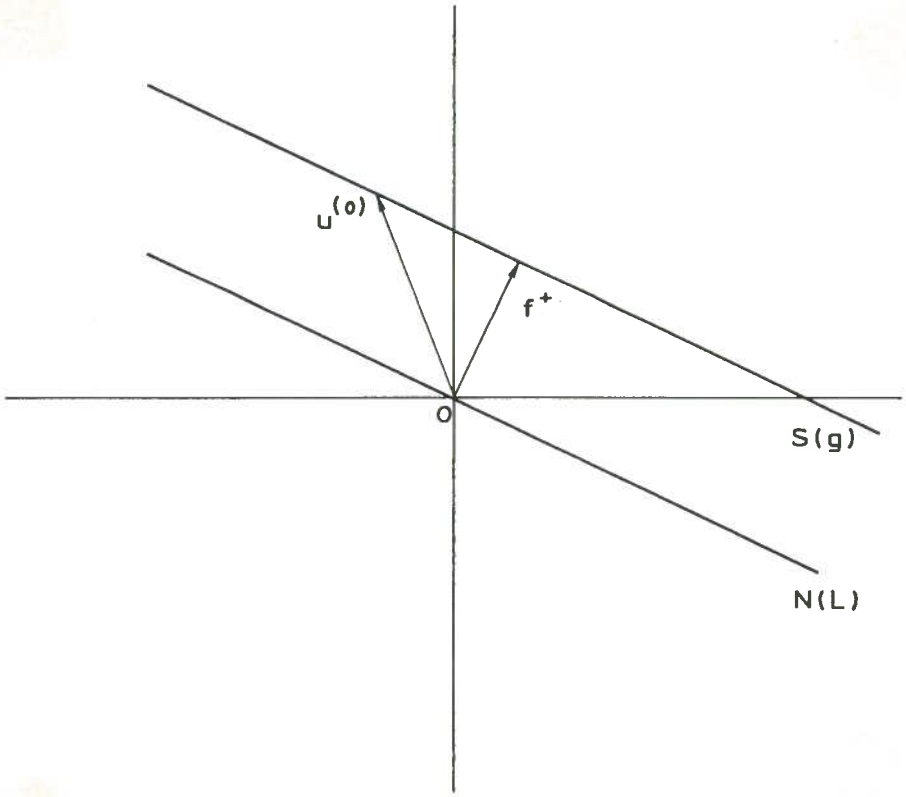

Fig.5 Two-dimensional geometric representation of the generalized solution. The null space N(L) is a straight line passing through the origin and the set of least squares solutions S(g) is a straight line parallel to N(L). Then f^+ is the element of S(g) orthogonal to N(L).

the restriction of the operator L to the orthogonal complement of N(L). In other words the generalized solution is the unique least squares solution whose invisible component, i.e. the component not transmitted by the instrument described by the operator L, is exactly zero.

We point out that, *when R(L) is closed, there exists a unique generalized solution for any* $g \in Y$. Since it is also possible to prove that, in such a case, the mapping $g \longrightarrow f^+$ is continuous (Groetsch, 1977), it follows that *when R(L) is closed, the problem of the determination of* f^+ *is well-posed*. Again, this result applies both to the case of the bandlimiting operator (80) and to the case of inverse problems with discrete data. In the latter case, when the functions ϕ_n are linearly independent, the generalized solution f^+ just coincides with the visible component of f as given by Eq.(121) or also by Eq.(130). When not all the ϕ_n are linearly independent, then f^+ is given again by Eq.(130) with a number of terms which is not N but the number N' of the linearly independent ϕ_n.

We consider now the two examples of linear operators with non-closed range already discussed in Chapter II.

In the case of a compact operator, using the representation (23), we have

$$\|Lf - g\|_Y^2 = \sum_{k=0}^{\infty} |\sigma_k (f,u_k)_X - (g,v_k)_Y|^2 + \|Qg\|_Y^2 \qquad (178)$$

and therefore the generalized solution exists if and only if Pg satisfies the second of the conditions (24) (the first condition just means Qg = 0). Then the generalized solution is given by

$$f^+ = \sum_{k=0}^{\infty} \sigma_k^{-1} (g,v_k)_Y u_k \qquad (179)$$

In a similar way, when L is a convolution operator, using Parseval equality we obtain

$$\|Lf - g\|_Y^2 = \frac{1}{(2\pi)^n} \int_{\Omega} |\hat{K}(\xi)\hat{f}(\xi) - \hat{g}(\xi)|^2 d\xi + \int_{\mathbb{R}^n/\Omega} |\hat{g}(\xi)|^2 d\xi \qquad (180)$$

and it follows again that the generalized solution exists if and only if the second of the conditions (28) is satisfied, in which case it is given by

$$f^+(x) = \frac{1}{(2\pi)^n} \int_{\Omega} \frac{\hat{g}(\xi)}{\hat{K}(\xi)} e^{i(x,\,\xi)} d\xi \qquad (181)$$

We come back now to the general case and we notice that both in the case of an operator with a closed range and in the case of an operator with a non-closed range, the mapping $g \longrightarrow f^+$ defines a linear operator $L^+ : Y \longrightarrow X$ as follows

$$f^+ = L^+ g \qquad (182)$$

L^+ is called the *generalized inverse* or also the *Moore-Penrose generalized inverse* of the linear operator L. The latter name comes from the fact that, as we have already remarked, L^+ is the natural extension, to the case of a linear continuous operator, of the Moore-Penrose inverse on a matrix.

It is easy to prove (Groetsch, 1977) that, when L is linear and continuous, L^+ is closed. Now, when R(L) is closed, f^+ exists and is unique for any $g \in Y$ and therefore, if we denote by $D(L^+)$ the domain of L^+, we have $D(L^+) = Y$. From the closed graph theorem (Balakrishnan, 1976) it follows that L^+ is continuous or, as we have already pointed out, that the problem of the determination of f^+ is well-posed. Its stability is

controlled by the condition number

$$\text{cond}(L) = \|L\| \, \|L^+\| \tag{183}$$

which is just a generalization of the condition number (11). For a problem with discrete data Eq.(183) coincides with Eq.(131). A generalization of this expression to arbitrary linear continuous operators is the following: if λ_{min} and λ_{max} denote respectively the lower and upper bound of the positive part of the spectrum of the operator LL^*, then $\|L\| = \sqrt{\lambda_{max}}$ and $\|L^+\| = \sqrt{\lambda_{min}}$, so that

$$\text{cond}(L) = (\lambda_{max} / \lambda_{min})^{1/2} \tag{184}$$

The expression for the norm of L^+ comes from the following relationship

$$L^+ = (L^*L)^+ \, L^* = L^* (LL^*)^+ \tag{185}$$

which can be easily proved using Eq.(173), and from the spectral representation of LL^*.

When R(L) is not closed, then L^+ is not defined everywhere on Y, but $D(L^+) = R(L) \oplus R(L)^{\perp}$. As a consequence, the determination of the generalized solution is an ill-posed problem. In fact it does not exist for arbitrary data and it does not depend continuously on the data. In such a case, as well as in the case of well-posed but ill-conditioned problems, one must use the regularization techniques which will be discussed in the next Chapter.

B. C-Generalized Inverses

In some problems one is looking for a least squares solution minimizing not the norm of X but some suitable seminorm defined on a subset of X. Examples come from the theory of interpolation by means of natural splines (Greville, 1969), from numerical methods for the solution of first kind Fredholm integral equations (Phillips, 1962) and from certain problems of computational vision (Hildreth, 1984; Bertero *et al*, 1986 b).

We will consider a norm or a seminorm of the following type

$$p(f) = \|Cf\|_Z \tag{186}$$

where $C : X \longrightarrow Z$ is a constraint operator, as defined in Chapt. II, Section A, and we will look for a generalized solution f_C^+ obtained by solving the following variational problem

$$\|Cf_C^+\|_Z = \inf \{ \|Cu\|_Z \mid u \in S(g) \} \tag{187}$$

When this problem has a unique solution, this will be called the *C-generalized solution* of Eq.(5). This is also called by Morozov the *solution of the basic problem* (Morozov, 1984).

We first consider the case where the constraint operator C satisfies the conditions a), b) of Chapter II, Section A. In such a case, as we know, Eq.(186) defines a norm and not a seminorm and the solution of the problem (187) can be reduced to the solution of the problem (177) just by redefining the space X, as explained in Chapter II. It follows that there exists a unique solution of the problem (187) if and only if $Pg \in LD(C)$ (image of the domain of the operator C under the action of the operator L). The mapping $g \longrightarrow f_C^+$ defines a linear operator L_C^+

$$f_C^+ = L_C^+ g \tag{188}$$

which will be called the *C-generalized inverse* of L.

It may be interesting to give in this simple case, an explicit representation of L_C^+ in terms of a suitable Moore-Penrose generalized inverse. If we put $f = C^{-1}\phi$, then, by means of simple computations, we find that (Bertero, 1986)

$$L_C^+ = C^{-1} (LC^{-1})^+ \tag{189}$$

Notice that L_C^+ may not be continuous, even when R(L) is closed, because LD(C) may not be closed. Examples are given in (Hildreth, 1984; Bertero *et al*, 1986 b). In these cases a well-posed problem is transformed into an ill-posed one but this transformation can be required by the introduction of physical constraints. In other words, the Moore-Penrose generalized solution is not meaningfull from the physical point of view and it must be replaced by other generalized solutions.

We consider now the case where the constraint operator C does not satisfy condition b) of Chapter II, Section A. In other words C is a closed operator with a closed range but it may have a non trivial null space N(C). A very simple example is provided by

$$\|Cf\|_Z^2 = \int |f^{(k)}(x)|^2 \, dx \tag{190}$$

where $f^{(k)}$ denotes the derivative of order k. Then N(C) is the subspace of the polynomials of degree $\leq k - 1$.

We assume now that condition b) of Chapter II, Section A, is replaced by the following one

b') the range of C is closed, more precisely R(C) = Z, and the unique

solution of the set of equations

$$Lf = 0 \quad , Cf = 0 \tag{191}$$

is $f = 0$, i.e. $N(L) \cap N(C) = \{0\}$.

When this condition is satisfied, it is possible to prove (Groetsch, 1986) that there exists a constant $m > 0$ such that

$$\|Lf\|_Y^2 + \|Cf\|_Z^2 \geq m \|f\|_X^2 \tag{192}$$

for any $f \in D(C)$. Such a condition is called by Morozov the *completion condition* (Morozov, 1984) and it is taken as the basic assumption for the solution of the problem (187). The inequality (192) is also proved in (Bertero, 1986) using not only conditions a), b') as in (Groetsch, 1986), but also the assumption that N(C) is finite dimensional.

By means of the inequality (192) it is easy to show that D(C), endowed with the scalar product

$$(f, \phi)_C = (Lf, L\phi)_Y + (Cf, C\phi)_Z \tag{193}$$

is a Hilbert space, which will be denoted by X_C. Analogously, the restriction of L to X_C will be denoted by L_C.

Assume now that, for a given g, there exist least squares solutions $u \in D(C)$. It is evident that this is true when $Pg \in LD(C)$. Then Eq.(175) implies that, for those least squares solutions, we have : $\|u\|_C^2 = \|Pg\|_Y^2 + \|Cu\|_Z^2$. Since, for fixed g, $\|Pg\|_Y^2$ is a constant, the solution of the problem (187) is equivalent to the solution of the following one

$$\|f_C^+\|_C = \inf \{ \|u\|_C \mid u \in S(g) \cap D(C) \} \quad . \qquad (194)$$

It follows that the *C-generalized inverse of the operator L is just the Moore-Penrose generalized inverse of the operator* L_C .

As an example we consider the case where $L : X \longrightarrow Y$ is compact. If we notice that, as follows from the inequality (192), any bounded set in X_C is also a bounded set in X, we conclude that the operator L_C is also compact. We can then introduce its singular system $\{\sigma_{C,k};\ u_{C,k},\ v_{C,k}\}$ which is the set of the solutions of the shifted eigenvalue problem

$$L_C\, u_{C,k} = \sigma_{C,k}\, v_{C,k} \ , \qquad L_C^*\, v_{C,k} = \sigma_{C,k}\, u_{C,k} \quad . \qquad (195)$$

When this problem has been solved then f_C^+ is given by Eq.(179) with $\{\sigma_k;\ u_k,\ v_k\}$ replaced by $\{\sigma_{C,k};\ u_{C,k},\ v_{C,k}\}$.

We want to show now that the solution of the shifted eigenvalue problem (195) can be reduced to the determination of the set of the solutions $\{\omega_k^2;\psi_k\}$ of the following generalized eigenvalue problem

$$L^*L\, \psi_k = \omega_k^2\, C^*C\, \psi_k \quad . \qquad (196)$$

We notice that this problem is analogous to the problem encountered in the investigation of the small oscillations of a mechanical system. In that case L^*L is related to the potential energy, while C^*C is related to the kinetic energy of the mechanical system.

The first step is the determination of L_C^* in terms of L^*. From the relation

$$(f, L_C^* g)_C = (Lf, L\, L_C^* g)_Y + (Cf, C\, L_C^* g)_Z = \qquad (197)$$

$$= (f, (L^*L + C^*C)\, L_C^*\, g)_X = (f, L^*g)_X$$

which holds true for any $f \in D(C)$ - dense in X - and any $g \in Y$, we obtain that

$$L_C^* = (L^*L + C^*C)^{-1}\, L^* \qquad . \qquad (198)$$

Therefore the second of the Eqs.(195) can be written as follows

$$L^*\, v_{C,k} = \sigma_{C,k}\, (L^*L + C^*C)\, u_{C,k} \qquad . \qquad (199)$$

Finally, if we apply the operator L^* to both sides of the first of the Eqs.(195) and if we use Eq.(199) in order to eliminate $v_{C,k}$, we obtain

$$(1 - \sigma^2_{C,k})\, L^*L\, u_{C,k} = \sigma^2_{C,k}\, C^*C\, u_{C,k} \qquad , \qquad (200)$$

and this equation can be identified with Eq.(196) by putting

$$\sigma_{C,k} = \omega_k\, (1 + \omega_k^2)^{-\frac{1}{2}} \qquad , \qquad u_{C,k} = \psi_k \qquad . \qquad (201)$$

Notice that, as a byproduct of this procedure, we have obtained that all the singular values $\sigma_{C,k}$ are smaller than one.

Similar results apply to inverse problems with discrete data. Since in this case L is a finite rank operator, the C-generalized inverse L_C^+ is always continuous but it may be ill-conditioned, the condition number being given again by Eq.(131), with σ_0 and σ_{N-1} replaced respectively by $\sigma_{C,0}$ and $\sigma_{C,N-1}$, respectively.

As a concluding remark we point out that, in the case of the interpolation problem in RKHS, as discussed in Chapter III, Section C, if

we look for C-generalized solutions associated with the functional (190), then the result is the interpolation in terms of natural splines (Greville, 1969). We just notice that condition b') is satisfied whenever $k \leq N$, where k is the order of the derivative in the functional (190) and N is the number of points. For a discussion of the interpolation and derivation problems in terms of C-generalized solutions see (Bertero *et al.*, 1985 a).

C- The Backus-Gilbert Method for Problems with Discrete Data

The Backus-Gilbert method (Backus and Gilbert, 1968; 1970) has been proposed for the solution of an inverse problem which consists in the determination of the structure of the Earth, using data related to properties of the whole Earth such as mass, moment of inertia and frequencies of elastic-gravitational normal modes. It has also been applied to the Fourier transform inversion (Oldenburg, 1976), to the inverse scattering problem (Colton, 1984) and to the Laplace transform inversion (Haario and Somersalo, 1985). A relationship between the Backus-Gilbert method and the Fejer theory of Fourier series expansions is discussed in (Bertero *et al.*, 1988 a).

The Backus-Gilbert method can only be used in the case of an inverse problem with discrete data, when the object space X is a space of square integrable functions. In such a case the problem (115) takes the form

$$g_n = \int f(x)\, \phi_n(x)^* \, dx \; ; \qquad n = 1, \ldots, N \quad . \tag{202}$$

It must also be pointed out that the method does not provide an exact but only an approximate solution of these equations. It is discussed in this Chapter because it shows some analogies with the method of generalized solutions.

For introducing the basic idea of the method, let us reconsider for a moment the Moore-Penrose generalized solution, or a C-generalized solution, of the problem (202). When the ϕ_n are linearly independent and the data values g_n are exact, by combining Eq.(202) with Eq.(121) we obtain

$$f^+(x) = \int A^+(x,x')\, f(x')\, dx' \tag{203}$$

where

$$A^+(x,x') = \sum_{n=1}^{N} \phi^n(x)\, \phi_n(x')^* \tag{204}$$

and therefore, at any point x, the generalized solution $f^+(x)$ is an average of the true solution f(x). Moreover this averaged function lies in the subspace spanned by the functions ϕ_n.

A similar result holds true for any C-generalized solution f_C^+ if we notice that f_C^+ belongs to the subspace spanned by the functions ψ_n which satisfy the relations

$$(f, \phi_n)_X = (f, \psi_n)_C \; ; \qquad n = 1, \ldots, N \tag{205}$$

for any $f \in D(C)$. When the C-scalar product is defined as in Eq.(13), the functions ψ_n are obtained by solving the equations

$$C^* C\, \psi_n = \phi_n \; ; \qquad n = 1, \ldots, N \tag{206}$$

while, when the scalar product is defined as in Eq.(193), the functions ψ_n are obtained by solving the equations

$$(L^{*}L + C^{*}C)\,\psi_n = \phi_n\,; \qquad n = 1, \ldots, N \qquad . \qquad (207)$$

Then, by introducing the Gram matrix of the functions ψ_n and the dual basis ψ^n , one finds for $f_C{}^{+}$ an expression similar to Eq.(121) with ϕ^n replaced by ψ^n. By combining this equation with Eq.(202) we find again

$$f_C{}^{+}(x) = \int A_C{}^{+}(x,x')\, f(x')\, dx' \qquad (208)$$

where

$$A_C{}^{+}(x,x') = \sum_{n=1}^{N} \psi^n(x)\, \phi_n(x')^{*} \quad . \qquad (209)$$

The functions f^{+} and $f_C{}^{+}$ are solutions of Eq.(202). The Backus-Gilbert method consists in looking for an approximate solution of these equations, let us denote it by $\tilde{f}_{BG}(x)$, which is also an average of the true solution f(x)

$$\tilde{f}_{BG}(x) = \int A(x,x')\, f(x')\, dx' \qquad (210)$$

and which linearly depends on the exact values of the functionals (202). This condition implies that the kernel A(x,x'), called *averaging kernel* by Backus and Gilbert, must have the following expression

$$A(x,x') = \sum_{n=1}^{N} a_n(x)\, \phi_n(x')^{*} \quad , \qquad (211)$$

with functions $a_n(x)$ which must be determined. By combining Eqs.(210),

(211) and (202) we have

$$\tilde{f}_{BG}(x) = \sum_{n=1}^{N} g_n a_n(x) \qquad (212)$$

and therefore $\tilde{f}_{BG}(x)$ is a function in the subspace spanned by the functions $a_n(x)$. It is obvious that the generalized solutions discussed in the previous Sections have a similar structure. In that case the functions $a_n(x)$ are determined by requiring that the solution satisfies Eq.(202) and by adding a variational principle for the solution (smallest norm, etc.) in order to ensure uniqueness. Backus and Gilbert follow a different approach and, in particular, they do not require that the function (212) satisfies Eq.(202). In fact they introduce a variational principle for the averaging kernel itself, since they require that it must be the sharpest in a sense to be specified. For this purpose, let J(x,x') be a function which vanishes when x = x' and increases monotonically as x goes away from x'. An example of such a function is

$$J(x, x') = (x - x')^2 \qquad . \qquad (213)$$

Then the unknown funtions $a_n(x)$ in Eq.(211) are determined by solving the following minimization problem

$$\delta^2(x) = \int J(x,x') |A(x,x')|^2 \, dx' = \text{minimum} \qquad (214)$$

with the constraint

$$\int A(x,x') \, dx' = 1 \qquad . \qquad (215)$$

In other words one looks for a kernel of the form (211) which gives a good approximation of the delta distribution $\delta(x - x')$.

If we introduce now the quantities

$$b_n = \int \phi_n(x)\, dx, \qquad S_{nm}(x) = \int J(x,x')\, \phi_n(x')\, \phi_m(x')^*\, dx' \tag{216}$$

we find that the solution of the problem (214), (215) implies the minimization of the following quadratic functional

$$\delta^2(x) = \sum_{n,m=1}^{N} S_{nm}(x)\, a_m(x)\, a_n(x)^* = \text{minimum}\ , \tag{217}$$

with the linear constraint

$$\sum_{n=1}^{N} b_n^*\, a_n(x) = 1\ . \tag{218}$$

This problem can be solved in a standard way using the method of Lagrange multipliers. If $\mathbf{S}(x)$ is the matrix with elements $S_{nm}(x)$, if this matrix is not singular and if we denote by $S^{nm}(x)$ the elements of $[\mathbf{S}(x)]^{-1}$, then the solution of the problem (217), (218) is

$$a_n(x) = \lambda(x) \sum_{m=1}^{N} S^{nm}(x)\, b_m \quad ; \quad n = 1, \ldots, N \tag{219}$$

the Lagrange multiplier $\lambda(x)$ being given by

$$\lambda(x) = \left(\sum_{n,m=1}^{N} S^{nm}(x)\, b_m\, b_n^{*} \right)^{-1} . \qquad (220)$$

The functions $a_n(x)$ depend on the choice of the function $J(x, x')$.

As concerns the stability of the Backus-Gilbert method it should be obvious that, for those problems where the generalized solution is extremely ill-conditioned, the solution $\tilde{f}_{BG}(x)$ must also be unstable. Numerical results obtained in the case of the Fourier series summation (Backus and Gilbert, 1968) indicate however that $\tilde{f}_{BG}(x)$ is more stable than $f^{+}(x)$. This follows from the fact that the kernel (204) is more narrow than the kernel (211) and therefore, using $f^{+}(x)$, one requires higher resolution. The connection between resolution and stability will be discussed in Chapter VI.

As a concluding remark, we point out that a convergence result has been recently proved for the Backus-Gilbert method (Schomburg and Berendt, 1987). In this paper it is assumed that the problem (202) is a finite section of a generalized moment problem (Chapter II, Section H) satisfying the requirement of uniqueness, i.e. the span of the functions ϕ_n, with $n=1,2,3,\ldots\ldots,\infty$, is dense in X. Moreover it is assumed that for a certain set of values g_n of the generalized moments there exists a solution f of the problem which is real valued and Lipschitz continuous.

Then let $\tilde{f}_N$ be the approximate solution provided by the Backus-Gilbert method, using the exact values of the first N generalized moments of f, the function $J(x,x')$ being given by Eq.(213). The result is that $\tilde{f}_N$ converges everywhere to f in the case of functions depending on one or two variables while $\tilde{f}_N$ in general does not converge to f in the case of functions depending on more than two variables.

It is evident that this convergence result can only be proved by

assuming that the values g_n of the generalized moments are not affected by experimental errors. It is interesting, however, to know that in this case the method can provide an approximation of the exact solution. It should also be important to investigate the convergence (or non-convergence) of the Backus- Gilbert method when applied to problems such as the moment discretization of a first kind Fredholm integral equation or the Fourier transform inversion with limited data.

V. REGULARIZATION THEORY FOR ILL-POSED PROBLEMS

The method of the generalized solutions, discussed in the previous Chapter, provides a satisfactory answer to the questions of existence and uniqueness for Eq.(5) only when the generalized inverse is continuous and well-conditioned. As we know, this means that the range of the operator L is closed and that the condition number (183) is not much greater than one. The method is not adequate when the generalized inverse is not continuous or when it is continuous but the condition number (183) is too large. In the first case the generalized solution may not exist because the data are contaminated by experimental errors; in the second case the generalized solution always exists but it may be deprived of any physical meaning as a consequence of the dramatic error propagation from the data to the solution. In both cases one must introduce methods for obtaining physically meaningful approximations of the generalized solutions. As already discussed in the Introduction, the basic idea is to constrain in some way the solution in order to avoid the wild oscillations generated by noise propagation. Several methods related to various kinds of constraints have been introduced and most of them have been unified in a general theory which is known now as *regularization theory* or also *Tikhonov regularization theory* . In this Chapter we sketch the basic ideas of the theory and we provide the main references of the mathematical literature which is grown very fast in the last years. Obviously our presentation will be strongly biased by our personal experience in this domain.

A. The Ivanov-Phillips-Tikhonov Regularization Method

The so-called *Tikhonov regularization method* was introduced independently by several authors at the beginning of the sixties. The first

versions of the method were published in 1962 (Ivanov, 1962; Phillips, 1962) and a more general, unifying formulation - restricted however to the case of first kind Fredholm integral equations - was later proposed by Tikhonov (Tikhonov, 1963 a; 1963 b). Other important contributions are due to Morozov (Morozov, 1966; 1968) and Miller (Miller, 1970).

In our presentation we first sketch the methods of Ivanov, Phillips and Miller and successively we show their relation with Tikhonov regularization theory.

For the purpose of providing a concise outline of these methods we first give a few results concerning the minimization of the following functional

$$\Phi_\alpha [f] = \|Lf - g\|_Y^2 + \alpha \|f\|_X^2 \qquad (221)$$

where α can be any positive number. Let f_α denote the function which minimizes $\Phi_\alpha [f]$. Then, by annihilating the first variation of $\Phi_\alpha [f]$, it is easy to show that f_α must satisfy the following orthogonality condition

$$(Lf_\alpha - g , L\phi)_Y + \alpha(f_\alpha , \phi)_X = 0 \qquad (222)$$

for any $\phi \in X$. It follows that f_α is a solution of the Euler equation

$$(L^*L + \alpha I)f_\alpha = L^*g \qquad (223)$$

Then the inequality

$$\| (L^*L + \alpha I) f\|_X \geq \alpha \|f\|_X \qquad (224)$$

which holds true for any $\alpha > 0$, implies that there exists a unique

solution of Eq.(223) which can be written in the following form

$$f_\alpha = R_\alpha g \tag{225}$$

where

$$R_\alpha = (L^* L + \alpha I)^{-1} L^* \tag{226}$$

By simple algebraic manipulation one can also show that

$$R_\alpha = L^* (LL^* + \alpha I)^{-1} \tag{227}$$

This representation of R_α implies that f_α belongs to the range of L^* and therefore that it is orthogonal to the null space of L, as follows from Eq.(18). An important consequence of this property is that f_α converges, in the limit $\alpha = 0$, to the generalized solution f^+ associated with g , provided that $Pg \in R(L)$. Moreover, using the spectral representation of the self-adjoint, positive semi-definite operator LL^*, it is easy to show that the function

$$\epsilon_\alpha^2 = \| Lf_\alpha - g \|_Y^2 = \alpha^2 \| (LL^* + \alpha I)^{-1} Pg \|_Y^2 + \| Qg \|_Y^2 \tag{228}$$

is a strictly increasing function of α, whose values at $\alpha = 0$ and $\alpha = \infty$ are respectively $\| Qg \|_Y^2$ and $\| g \|_Y^2$. In a similar way one can show that the function

$$E_\alpha^2 = \| f_\alpha \|_X^2 = \| L^* (LL^* + \alpha I)^{-1} g \|_X^2 \tag{229}$$

is a strictly decreasing function of α, whose values at $\alpha = 0$ and $\alpha = \infty$

are respectively $\|f^+\|_X^2$ (∞ ,when the generalized solution does not exist) and zero.

For the sake of clarity we give a more explicit representation of f_α and of the functions ϵ_α^2 and E_α^2 in the two particular cases already discussed in the previous Chapters, namely compact operators and convolution operators.

When L is compact, from Eq.(225) and the singular value decomposition of L and L^*, one easily derives that f_α admits the following expansion

$$f_\alpha = \sum_{k=0}^{+\infty} \frac{\sigma_k}{\sigma_k^2 + \alpha} (g, v_k)_Y u_k \tag{230}$$

which clearly shows that $f_\alpha \longrightarrow f^+$ (if f^+ exists, as given by Eq.(179)) when $\alpha \longrightarrow 0$. Moreover we get

$$\epsilon_\alpha^2 = \sum_{k=0}^{+\infty} \frac{\alpha^2}{(\sigma_k^2 + \alpha)^2} |(g, v_k)_Y|^2 + \|Qg\|_Y^2 \tag{231}$$

and also

$$E_\alpha^2 = \sum_{k=0}^{+\infty} \frac{\sigma_k^2}{(\sigma_k^2 + \alpha)^2} |(g, v_k)_Y|^2 \tag{232}$$

and the properties, stated above, of the functions ϵ_α, E_α are evident.

Analogously, in the case of a convolution operator, the function f_α takes the following form

$$f_\alpha(\mathbf{x}) = \frac{1}{(2\pi)^n} \int_\Omega \frac{\hat{K}^*(\xi)}{|\hat{K}(\xi)|^2 + \alpha} \hat{g}(\xi)\, e^{i(\mathbf{x},\, \xi)} d\xi \qquad (233)$$

where Ω is the support of $\hat{K}(\xi)$. Then, from Parseval equality, one gets

$$\epsilon_\alpha^2 = \frac{1}{(2\pi)^n} \int_\Omega \frac{\alpha^2}{(|\hat{K}(\xi)|^2 + \alpha)^2} |\hat{g}(\xi)|^2 d\xi + \qquad (234)$$

$$+ \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n/\Omega} |\hat{g}(\xi)|^2 d\xi$$

and also

$$E_\alpha^2 = \frac{1}{(2\pi)^n} \int_\Omega \frac{|\hat{K}(\xi)|^2}{(|\hat{K}(\xi)|^2 + \alpha)^2} |\hat{g}(\xi)|^2 d\xi \qquad (235)$$

and the properties of the functions ϵ_α, E_α are evident again. Similar formulas apply also to problems which can be diagonalized by means of the Mellin transform such as the Abel or the Laplace transform inversion-see Chapt. II, Section D.3 and Section G.

1. Ivanov Method (Constrained Least Squares Solutions)

The basic idea of the Ivanov method (Ivanov, 1962) consists in restricting the approximate solutions to some suitable subset defined by physical constraints. Here we consider only the case where the subset is a sphere of radius E in X

$$S_E = \{ f \in X \mid \| f \|_X \leq E \} \quad . \qquad (236)$$

This choice has a precise physical motivation when $|f|^2$ is an energy density and X is a space of square integrable functions. Then the knowledge of an upper bound E^2 on the total energy of the signal implies that $f \in S_E$.

When the constant E is given, it is quite natural to look for the function (or the functions) $\tilde{f}^{(E)} \in S_E$ such that $L\,\tilde{f}^{(E)}$ has the minimal distance from g and this is equivalent to solve the following constrained least squares problem

$$\| L\,\tilde{f}^{(E)} - g \|_Y = \inf \{ \| Lf - g \|_Y \mid \| f \|_X \leq E \} \quad . \qquad (237)$$

Any solution of this problem will be obviously called a *constrained least squares solution*.

We must consider separately two cases.

a) The generalized solution exists and satisfies the constraint : $\| f^+ \|_X \leq E$. In such a case the solution of problem (237) is not unique (except when $\| f^+ \|_X = E$). The set of the constrained least squares solutions is the intersection of the set S(g) of the unconstrained least squares solutions , Eq.(176), with the sphere S_E (see Fig.6). Then there exists a unique constrained least squares solution of minimal norm and this obviously coincides with f^+.

b) The generalized solution does not exist or, if it exists, it does not satisfy the constraint, i.e. $\| f^+ \|_X > E$. This case is the most likely when data are noisy. Then the intersection of S_E with the set of the unconstrained least squares solutions is empty (see Fig.6). Under these

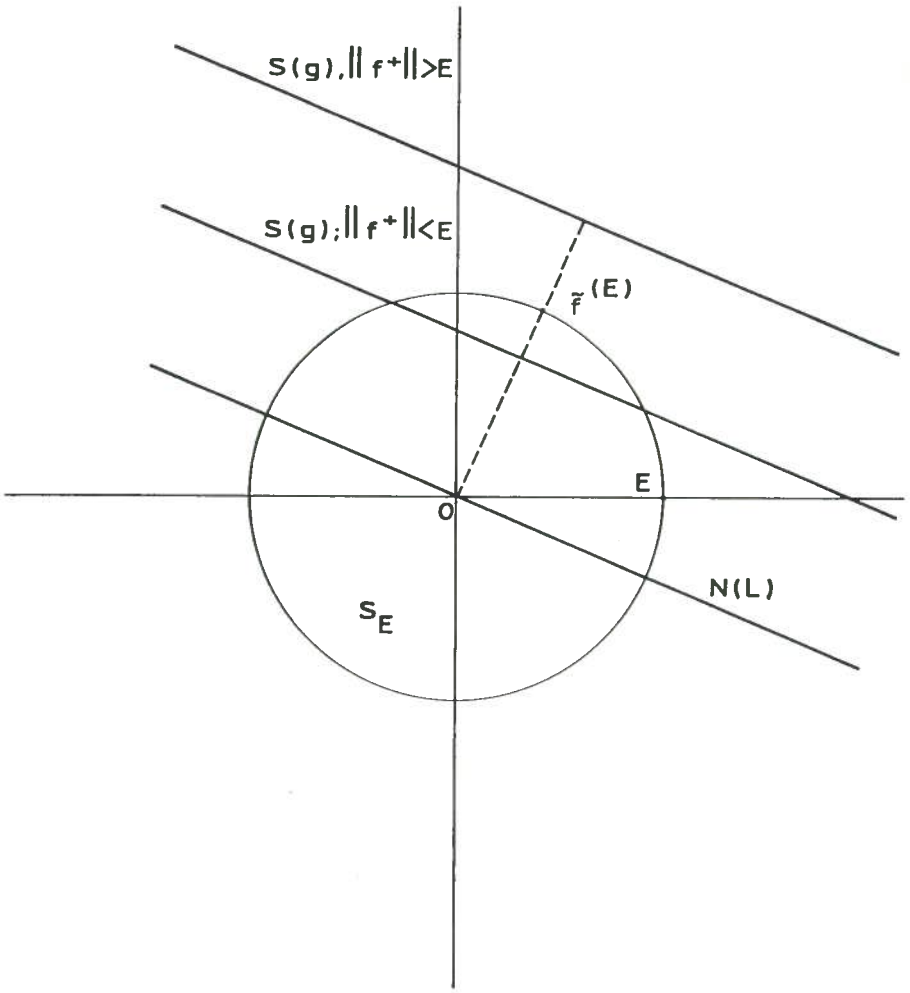

Fig.6 Two-dimensional geometric representation of the constrained least squares solution. In the case of data g such that $\|f^+\| < E$, S(g) intersects S_E and therefore $\tilde{f}^{(E)}$ coincides with f^+(see Fig.5). In the case of data g such that $\|f^+\| > E$, S(g) does not intersect S_E and therefore $\tilde{f}^{(E)}$ lies on the circle of radius E and is orthogonal to N(L).

circumstances it is obvious that the constrained minimum points of the functional $\|Lf - g\|_Y$ cannot be interior to S_E but must lie on the surface of this sphere. Since these points satisfy the condition $\|f\|_X = E$, one can use the method of Lagrange multipliers for determining the solution of

problem (237). This method consists of the following steps:

1) for any $\alpha>0$ minimize the functional (221);

2) since, for any α, there exists a unique minimum point f_α of this functional, then search for a value of α such that

$$\| f_\alpha \|_X = E \qquad (238)$$

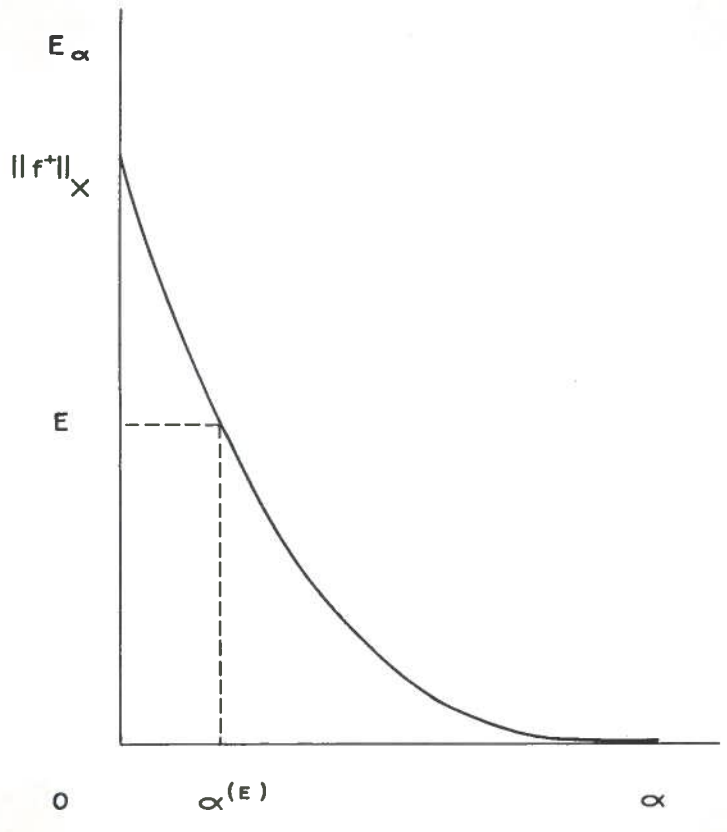

Fig.7 Graph of the function E_α, in the case $E < \|f^+\|_X < +\infty$, illustrating the determination of $\alpha^{(E)}$.

From the properties of the function E_α, Eq.(229), stated above and illustrated in Fig. 7, it follows that there exists a unique value of α, say $\alpha^{(E)}$, which solves Eq.(238). The corresponding solution f_α is just $\tilde{f}^{(E)}$ and *this is the unique solution of problem (237).*

We conclude with a few results about the convergence properties of the constrained least squares solution $\tilde{f}^{(E)}$ in the ideal case of experimental errors tending to zero. For this purpose we assume that a family $\{g_\epsilon\}_{\epsilon>0}$ of noisy data functions is given and that, when $\epsilon \longrightarrow 0$, g_ϵ converges to a noise free data function g, namely a function in the range of L. Let $\tilde{f}_\epsilon^{(E)}$ be the constrained least squares solution associated with g_ϵ and let $f^+ = L^+ g$ be the generalized solution associated with g. Moreover, let us assume that, for any ϵ, the case b) applies. Then the following results hold true (Bertero, 1986) :

i) if $\|L^+ g\|_X < E$, i.e. the prescribed constant is overestimated, then $\tilde{f}_\epsilon^{(E)}$ weakly converges to f^+ when $\epsilon \longrightarrow 0$ (for the definition of weak convergence, see (Balakrishnan, 1976)) ;

ii) if $\|L^+ g\|_X = E$, i.e. the prescribed constant is precise, then $\tilde{f}_\epsilon^{(E)}$ strongly converges to f^+ when $\epsilon \longrightarrow 0$;

iii) if $\|L^+ g\|_X > E$, i.e. the prescribed constant is underestimated, then $\tilde{f}_\epsilon^{(E)}$ strongly converges to $\tilde{f}^{(E)}$, the constrained least squares solution associated with the noise free data g.

For problems with discrete data the case i) must be modified since, for sufficiently small ϵ, there is necessarily a transition from the case b) to the case a) and therefore the constrained least squares solution

corresponding to noisy data is not unique. In such a case one can identify, by definition, the constrained least square solution with the generalized solution and therefore strong convergence applies also to this case. This result must be evident since weak convergence and strong convergence coincide in finite dimensional spaces.

2. Phillips Method

This method was firstly proposed for the approximate solution of first kind Fredholm integral equation (Phillips, 1962). A more general formulation was given by Ivanov and Morozov (Ivanov, 1966; Morozov, 1966, 1968) while Reinsch (Reinsch, 1967) applied independently the same method to the *smoothing* problem, a problem which replaces strict interpolation when the values of the function are only approximately given.

The starting point is the assumption that an upper bound ϵ on the error is known. We denote by $J_\epsilon(g)$ the set of all the elements of X which are compatible with the data g within the error ϵ

$$J_\epsilon(g) = \{f \in X \mid \|Lf - g\|_Y \leq \epsilon\} \quad . \tag{239}$$

This set is always unbounded when the problem is ill-posed. In the case of a problem with discrete data, for example, it is a cylinder whose basis is an ellipsoid in X_N (see Chapt.III, Section A). This cylinder is not bounded in the directions orthogonal to X_N because the solution of the problem is not unique. On the other hand, in the case of an operator whose inverse is not continuous $J_\epsilon(g)$ is unbounded for the following reason : it is always possible to find a sequence $\{f_n\}$ such that $\|Lf_n\|_X \longrightarrow 0$ while $\|f_n\|_Y \longrightarrow \infty$. Notice that the example of Hadamard, discussed in the

Introduction, is just a particular case of this general result. It is easy to prove, however, using the continuity and linearity of L, that $J_\epsilon(g)$ is always a closed and convex set.

Since $J_\epsilon(g)$ contains wildly oscillating and completely unphysical approximate solutions, it is quite natural to look for the smoothest element of $J_\epsilon(g)$, i.e. the element of minimal norm ,which will be denoted by $\tilde{f}^{(\epsilon)}$. This leads to solve the following problem

$$\| \tilde{f}^{(\epsilon)} \|_X = \inf \{ \|f\|_X \mid \|Lf - g\|_Y \leq \epsilon \} \quad . \qquad (240)$$

As remarked by Ivanov (Ivanov, 1966) this problem is just the dual of the problem (237).

Since the set $J_\epsilon(g)$ is closed and convex, from the general theorem of functional analysis (Balakrishnan, 1977) already used for proving the existence and uniqueness of the generalized solution, it follows that there exists a unique solution of the problem (240). This solution is not the null element of X provided that the data g satisfies the inequality $\|g\|_Y > \epsilon$. This inequality is quite reasonable since it implies that the norm of the data is greater then the norm of the noise. If it is not satisfied, it means that the data function (vector) consists only of noise and that it does not contain any information about the unknown object f. Since we exclude this case and we also assume that, if a nonzero component of g orthogonal to the range of L exists, then this component can only be an effect of the noise, we concludee that the following inequalities must hold true

$$\|Qg\|_Y < \epsilon < \|g\|_Y \quad . \qquad (241)$$

A representation of the solution of problem (240), analogous to the representation of the constrained least squares solution, can be obtained

if we notice that $\tilde{f}^{(\epsilon)}$ must satisfy the condition $\|L\,\tilde{f}^{(\epsilon)} - g\|_Y = \epsilon$. Then we can use again the method of Lagrange multipliers in order to determine $\tilde{f}^{(\epsilon)}$. We must minimize the functional (221) for any $\alpha > 0$ and then search for a minimum point f_α such that

$$\|Lf_\alpha - g\|_Y = \epsilon \quad . \tag{242}$$

From the properties of the function ϵ_α, Eq.(228), and the conditions (241) it follows that there exists a unique value of α, say $\alpha^{(\epsilon)}$, which solves Eq.(242) - see Fig.8. The corresponding solution f_α is just $\tilde{f}^{(\epsilon)}$, i.e. the

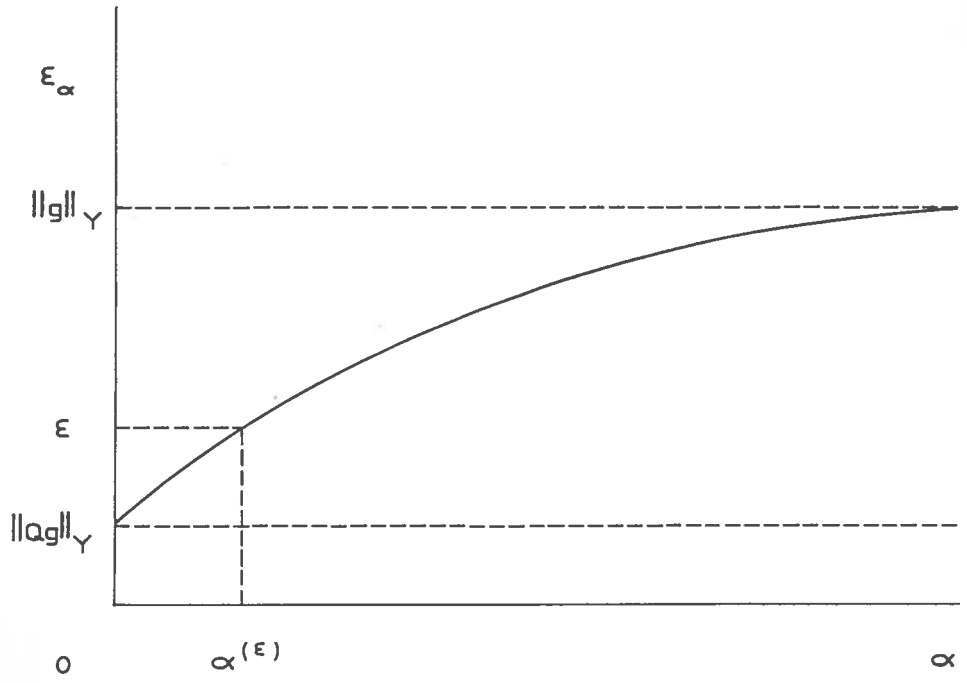

Fig.8 Graph of the function ϵ_α, in the case $\|Qg\|_Y < \epsilon < \|g\|_Y$, illustrating the determination of $\alpha^{(\epsilon)}$.

unique solution of problem (240). We see that the difference between the solution of problem (237) and the solution of problem (240) consists only in a different choice of the Lagrange multiplier.

As concerns the convergence of $\tilde{f}^{(\epsilon)}$, in the case where the experimental errors tend to zero, the situation is much more simple than in the case of the constrained least squares solutions. If $\{g_\epsilon\}_{\epsilon>0}$ is the family of noisy data functions introduced in Section A.1, if $\tilde{f}_\epsilon^{(\epsilon)}$ is the solution of problem (240) with g replaced by g_ϵ and if $f^+ \in J_\epsilon(g_\epsilon)$, for any ϵ, then it is possible to prove (Ivanov, 1966; Groetsch, 1984; Bertero, 1986) that $\tilde{f}_\epsilon^{(\epsilon)}$ strongly converges to f^+ when $\epsilon \longrightarrow 0$.

3. Miller Method

The method of Section A.1 requires a prescribed bound on the solution while the method of Section A.2 requires a prescribed bound on the error. In a paper of Miller (Miller, 1970) the case where both bounds are known is also considered. Results similar to those of Miller were also obtained by Franklin (Franklin, 1974).

Let us assume that two constants ϵ, E are given and that one wants to look for functions f such that

$$\|Lf - g\|_Y \leq \epsilon , \qquad \|f\|_X \leq E \quad . \tag{243}$$

The set K of all the functions satisfying these conditions is just the intersection of the set S_E, Eq.(236), and of the set $J_\epsilon(g)$, Eq.(239) : $K = S_E \cap J_\epsilon(g)$. Any function $f \in K$ can be called an *admissible approximate*

solution.

In such an approach we must consider two problems: the first is to assure that the set K is not empty, in which case we say that the pair (ϵ,E) is *permissible*; the second is to extract an element of K in order to produce one specific approximate solution.

As concerns the first question, it is easy to characterize the set of permissible pairs. We first notice that *the set K is not empty if and only if both* $\tilde{f}^{(E)}$ *and* $\tilde{f}^{(\epsilon)}$ *belong to K.* The "if" part is trivial. The "only if" part follows from the variational properties of $\tilde{f}^{(E)}$ and $\tilde{f}^{(\epsilon)}$. For example, the condition that K is not empty implies that there exist elements of $J_\epsilon(g)$ whose norm is less than E. Since $\tilde{f}^{(\epsilon)}$ is the element of minimal norm of $J_\epsilon(g)$, it follows that the norm of $\tilde{f}^{(\epsilon)}$ must also be less than E. Moreover $\tilde{f}^{(\epsilon)}$ satisfies, by definition, the first of the conditions (243), and therefore $\tilde{f}^{(\epsilon)} \in K$. Similar arguments apply to $\tilde{f}^{(E)}$.

The remarks used for proving the previous result also imply that, when K is not empty, then

$$\| \tilde{f}^{(\epsilon)} \|_X \leq \| \tilde{f}^{(E)} \|_X \tag{244}$$

and

$$\| L\tilde{f}^{(E)} - g \|_Y \leq \| L\tilde{f}^{(\epsilon)} - g \|_Y \quad . \tag{245}$$

Moreover, if we recall that E_α is a decreasing function of α (or that ϵ_α is an increasing function of α) we also have

$$\alpha^{(E)} \leq \alpha^{(\epsilon)} \quad . \tag{246}$$

Finally, since the previous results imply that, when K is not empty,

then $\| \tilde{f}^{(\epsilon)} \|_X \leq E$ and $\| L \tilde{f}^{(E)} - g \|_Y \leq \epsilon$, it follows that the set of the permissible pairs in the plane $\{ \epsilon, E \}$ is just the set of all the pairs which are to the right and above the curve described by $\{ \epsilon_\alpha, E_\alpha \}_{\alpha>0}$ (see Fig.9).

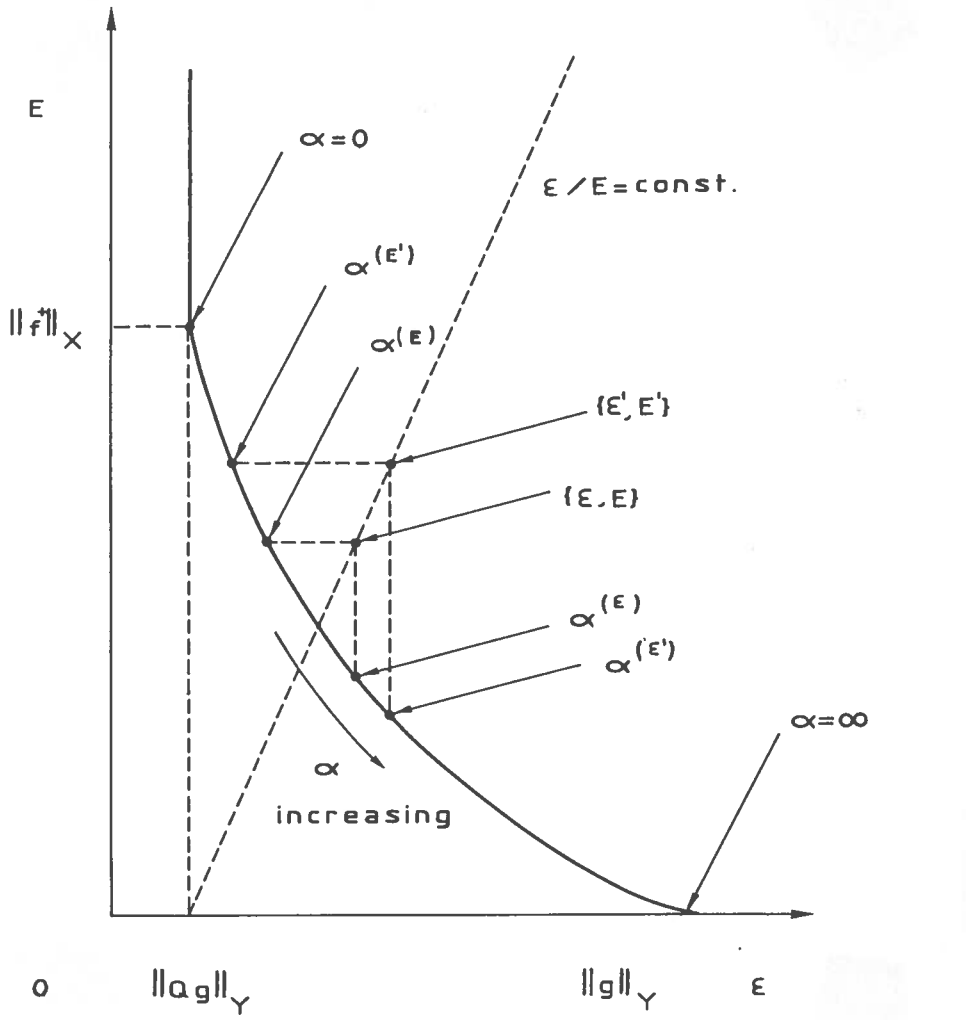

Fig.9 Representation of the set of the permissible pairs (ϵ,E) in the case where the generalized solution f^+ exists. When f^+ does not exist, the line $\epsilon = \|Qg\|_Y$ is an asymptote of the boundary curve $(\epsilon_\alpha, E_\alpha)$.

We also conclude that all the functions f_α with α between $\alpha^{(E)}$ and $\alpha^{(\epsilon)}$ belong to K.

As concerns the problem of determining an element of K, it is obvious that both $\tilde{f}^{(E)}$ and $\tilde{f}^{(\epsilon)}$ provide an answer to this question. Another possibility, however, is to take the function f_α with $\alpha = (\epsilon/E)^2$. This function will be denoted by $\tilde{f}^{(0)}$. In fact it has been proved by Miller (Miller, 1970) that, when K is not empty, $\tilde{f}^{(0)}$ satisfies the inequalities (243) with $[\epsilon,E]$ replaced by $[\sqrt{2}\,\epsilon, \sqrt{2}\,E]$. It is easy to find, however, a sufficient condition which assures that $\tilde{f}^{(0)} \in K$. Let us denote by $\Phi^{(0)}[f]$ the functional (221) with $\alpha = (\epsilon/E)^2$ and by $K^{(0)}$ the following subset of X

$$K^{(0)} = \left\{ f \in K \,\middle|\, \Phi^{(0)}[f] \leq \epsilon^2 \right\} \tag{247}$$

Then, since $\tilde{f}^{(0)}$ minimizes the functional $\Phi^{(0)}[f]$, it is obvious that $K^{(0)}$ is not empty if and only if the following condition is satisfied

$$\Phi^{(0)}\left[\tilde{f}^{(0)}\right] \leq \epsilon^2 \tag{248}$$

(notice that this condition can be easily verified in a numerical application of the method). Finally, using the inclusion

$$K^{(0)} \subset K \tag{249}$$

which follows from the remark that any element of $K^{(0)}$ satisfies the conditions (243), we conclude that when condition (248) is satisfied, the set K is not empty and $\tilde{f}^{(0)} \in K$.

Moreover the following inequalities are a trivial consequence of the inequalities (244) - (246) and of the fact that $f_\alpha \in K$ if and only if α

belongs to the interval $[\alpha^{(E)}, \alpha^{(\epsilon)}]$

$$\| \tilde{f}^{(\epsilon)} \|_X \leq \| \tilde{f}^{(0)} \|_X \leq \| \tilde{f}^{(E)} \|_X \tag{250}$$

$$\| L \tilde{f}^{(E)} - g \|_Y \leq \| L f^{(0)} - g \|_Y \leq \| L \tilde{f}^{(\epsilon)} - g \|_Y \tag{251}$$

$$\alpha^{(E)} \leq (\epsilon/E)^2 \leq \alpha^{(\epsilon)} \tag{252}$$

We see therefore that the Miller solution $\tilde{f}^{(0)}$ has a degree of smoothness which is intermediate between that of $\tilde{f}^{(E)}$ and that of $\tilde{f}^{(\epsilon)}$.

Finally, as concerns the convergence of $\tilde{f}^{(0)}$ to the true generalized solution f^+ when the error of the data tends to zero, this approximate solution has properties similar to those of the constrained least squares solution $\tilde{f}^{(E)}$ (Bertero, 1986).

4. Extensions and Comments

The method of contrained least squares solutions can be considered as a generalization of the pioneering works on ill-posed problems for partial differential equations (Pucci, 1955; John, 1955; Fox and Pucci, 1958; John, 1960). In these papers, already mentioned in the Introduction, one looks for approximate solutions satisfying a prescribed bound and this condition is replaced, in the method of Section A.1 , by the condition of looking for approximate solutions in the sphere (236). An extension of this condition is provided in the same paper of Ivanov (Ivanov, 1962) and is based on a topological lemma, due to Tikhonov (Lavrentiev, 1967; Tikhonov and Arsenine, 1977) which in our context can be formulated as follows: *Let H be a compact subset of the Hilbert space X and let assume that the*

linear, continuous operator L: X ⟶ Y, when restricted to H, has an inverse. Then the inverse operator is continuous.

The result of Ivanov can now be formulated as follows: *If H is a compact and convex set of X and if the restriction of L to H admits an inverse operator, then, for any g ∈ Y, there exists in H a unique least squares solution* $\tilde{f}$ *of Eq.(5). Moreover the mapping g ⟶* $\tilde{f}$ *is continuous.*

It is obvious that, in this way, one is not obliged to restrict solutions to a sphere (in fact, a sphere is not a compact set). An important example of an application of the Theorem above is the case where the function to be restored is the distribution function of a random variable, and therefore it is an increasing function with values in the interval [0,1]. Then, according to Helly theorem (Titchmarsh, 1958, pg.342), a set of increasing and uniformly bounded functions, defined on a bounded interval [a,b], is compact in $L^2(a,b)$ and therefore Ivanov theorem can be used whenever the inverse operator L^{-1} exists.

As a second comment we point out that the methods outlined in the previous Sections provide approximations of the Moore-Penrose generalized solution. Then one can also look for approximations of the *C-generalized solutions* introduced in Chapt.IV, Section B. In fact the method of Phillips (Phillips, 1962) applies to this case, since it provides an approximate solution of a first kind Fredholm integral equation such that the L^2-norm of its second derivative is as small as possible. Also the smoothing method of Reinsch (Reinsch, 1967) corresponds to this case.

Approximations of the C-generalized solutions can be obtained if we replace the functional (221) with the following one

$$\Phi_{C,\alpha}[f] = \|Lf - g\|_Y^2 + \alpha \|Cf\|_Z^2 \qquad (253)$$

where the constraint operator C satisfies the properties assumed in Chapt.IV, Section B and, in particular, property b'). Then one can prove (Groetsch, 1984; Morozov, 1984; Bertero, 1986) that, for any $\alpha > 0$, there exists a unique function $f_{C,\alpha} \in X$ which minimizes the functional (253)

and which can be obtained by solving the functional equation

$$(L^*L + \alpha C^*C)\, f_{C,\alpha} = L^*g \quad . \tag{254}$$

Moreover, the mapping $g \longrightarrow f_{C,\alpha}$, given by

$$f_{C,\alpha} = R_{C,\alpha}\, g \tag{255}$$

with

$$R_{C,\alpha} = (L^*L + \alpha C^*C)^{-1}\, L^* \tag{256}$$

is continuous. This operator can be written in the standard form (226) by introducing the adjoint of the operator L_C with respect to the scalar product (193) (Groetsch, 1986). Then, using Eq.(198), by some simple algebraic manipulations, from Eq.(256) one obtains

$$R_{C,\alpha} = \beta\, (L_C^*L_C + \beta\alpha I)^{-1}\, L_C^* \tag{257}$$

where $\beta = (1 - \alpha)^{-1}$. In particular, when L is a compact operator (or a finite rank operator as in the case of inverse problems with discrete data), it follows that $f_{C,\alpha}$ has a representation, in terms of the singular values $\sigma_{C,\alpha}$, Eq. (195), analogous to the representation (230) of f_α.

We also notice that, when L is an integral operator and C is a differential operator such that the scalar product (13) is given by Eq.(12), then the solution of Eq.(254) implies the solution of a boundary value problem for an integrodifferential equation (Tikhonov, 1963 a). The latter is equivalent to the solution of the functional equation (223) in the Sobolev space defined by the scalar product (12) (Groetsch, 1984).

If one introduces now the functions

$$\epsilon_{C,\alpha} = \| L\, f_{C,\alpha} - g \|_Y \quad , E_{C,\alpha} = \| C\, f_{C,\alpha} \|_Z \tag{258}$$

one can easily prove that $\epsilon_{C,\alpha}$ is an increasing function of α while $E_{C,\alpha}$ is a decreasing function of α. As a consequence, all the results proved in the previous Sections and concerning the operator (226) can be extended to the present case. More precisely, there exist a unique value of α minimizing $\epsilon_{C,\alpha}$ with the constraint $E_{C,\alpha} \leq E$ and conversely there exists a unique value of α minimizing $E_{C,\alpha}$ with the constraint $\epsilon_{C,\alpha} \leq \epsilon$.

A final comment about the *Backus-Gilbert method*, which can also be affected by numerical instability. No regularization method has been developed for this algorithm. Backus and Gilbert, however, have introduced two methods in order to improve stability (Backus and Gilbert, 1970), even if rigorous results have not yet been proved for these methods.

If the covariance matrix $\mathbf{C}$ of the noise is known, then from Eq.(212) one easily derives that, at any given point x, the variance of the error induced by the noise on $\tilde{f}_{BG}(x)$ is

$$\sigma^2(x) = \sum_{n,m=1}^{N} C_{nm}\, a_m(x)\, a_n(x)^* \quad . \tag{259}$$

Then the two methods introduced by Backus and Gilbert are the following:

1) Minimize the functional $\sigma^2(x)$ with the constraint (218) and also the constraint $\delta^2(x) \leq E^2$ ($\delta^2(x)$ is defined in Eq. (217)). The latter constraint prescribes an upper limit on the desired resolution.

2) Minimize the functional $\delta^2(x)$, Eq.(217), with the constraint (218) and also the constraint $\sigma^2(x) \leq \epsilon^2$. The latter constraint prescribes an upper limit on the desired error affecting the reconstructed solution.

Both problems can also be solved by means of the method of Lagrange multipliers and we do not give the details here. Let us just remark that these two problems are one the dual of the other and that they are similar respectively to the Ivanov and to the Phillips method for the regularization of the Moore-Penrose generalized solution.

B. General Formulation of Regularization Methods

The common feature of the methods presented in Section A is that they provide different criteria for selecting a specific element from the same family of approximate solutions, namely $f_\alpha = R_\alpha g$, defined by Eqs.(225) and (226) – or (227). This family describes a trajectory in the Hilbert space X and in order to have a clear picture of this trajectory we need some further properties of the operators R_α. More precisely we want to show that :

i) for any $\alpha > 0$, $R_\alpha : Y \longrightarrow X$ is a linear, continuous operator whose norm is bounded by

$$\|R_\alpha\| \leq \frac{1}{\sqrt{\alpha}} \quad ; \tag{260}$$

ii) if g belongs to the range of L and f^+ is the generalized solution associated with g, then

$$\lim_{\alpha \downarrow 0} \|R_\alpha g - f^+\|_X = 0 \quad . \tag{261}$$

The proof of i) follows from Eqs.(229) and (224). In fact Eq.(229)

implies that

$$\|R_\alpha g\|^2 = (LL^*(LL^* + \alpha I)^{-1} g,\ (LL^* + \alpha I)^{-1} g)_Y \quad . \tag{262}$$

Then, if we notice that the norm of the operator $LL^*(LL^* + \alpha I)^{-1}$ is smaller than 1 and that, according to the inequality (224), the norm of the operator $(LL^* + \alpha I)^{-1}$ is smaller than α^{-1}, we obtain the inequality (260).

As concerns ii), if $g \in R(L)$, then $g = Lf^+$ and from Eq.(226) we get

$$\|R_\alpha g - f^+\|_X = \|R_\alpha L f^+ - f^+\|_X = \alpha \|(LL^* + \alpha I)^{-1} f^+\|_X \quad . \tag{263}$$

Then property ii) follows from the spectral representation of the self-adjoint, positive semi-definite operator L^*L and from the dominated convergence theorem (Groetsch, 1984; Bertero, 1986).

Properties i), ii) can be verified in an elementary way in the case of a compact operator or of a convolution operator, using respectively Eq.(230) or Eq.(233).

In the case of noisy data, say g_ϵ where ϵ is an estimate of the norm of the error, i.e. of the distance between g_ϵ and the exact data $g \in R(L)$

$$\|g_\epsilon - g\|_Y \leq \epsilon \quad , \tag{264}$$

the approximate solution $R_\alpha g_\epsilon$ may have no limit when $\alpha \longrightarrow 0$ or, as in the case of problems with discrete data, the limit is the generalized solution f_ϵ^+ and the distance between f_ϵ^+ and $f^+ = L^+ g$, $\|f_\epsilon^+ - f^+\|_X$, can be extremely large. There exists however, a value of α, say $\alpha^{(opt)}$, such that the distance between $R_\alpha g_\epsilon$ and f^+ is minimum. If we write

$$R_\alpha g_\epsilon - f^+ = (R_\alpha g - f^+) + R_\alpha (g_\epsilon - g) \qquad (265)$$

from Eq.(260) and the triangular inequality we have

$$\|R_\alpha g_\epsilon - f^+\|_X \leq \|R_\alpha g - f^+\|_X + \frac{\epsilon}{\sqrt{\alpha}} \qquad (266)$$

Then, since the first term in the r.h.s. is an increasing function of α, as follows from Eq.(263), while the second term is a decreasing one, there exists a unique value of α which minimizes the r.h.s. of Eq.(266). It is obvious that this optimum value of α, $\alpha^{(opt)}$, cannot be determined in practice because its determination requires the knowledge of the true solution f^+. It is important however to know that such an optimum value certainly exists.

We can now describe the trajectory of the approximate solutions $f_{\alpha,\epsilon} = R_\alpha g_\epsilon$ in the case where the set K, defined by the inequalities (243), is not empty - see Fig.10. The trajectory starts at the origin (null element) of X when $\alpha = \infty$ and for large values of α it lies inside the sphere S_E, Eq.(236). Then, for $\alpha = \alpha^{(\epsilon)}$, the trajectory crosses the surface of the ellipsoid $J_\epsilon(g)$, Eq.(239) and for values of α between $\alpha^{(\epsilon)}$ and $\alpha^{(E)}$ it goes through the set K. In this part of the trajectory we have the point corresponding to the optimum value of α discussed above and also the point corresponding to $\alpha = (\epsilon/E)^2$ (at least when condition (248) is satisfied). Finally, for all the values of α smaller than $\alpha^{(E)}$ the trajectory always lies inside the ellipsoid $J_\epsilon(g)$ and its end point, for $\alpha = 0$, will be the centre of $J_\epsilon(g)$, i.e. f_ϵ^+, when there exists the generalized solution associated with g_ϵ. Otherwise it goes to infinity.

These comments on the methods of the previous Section justify the general definition of a *regularizing algorithm* (in the sense of Tikhonov)

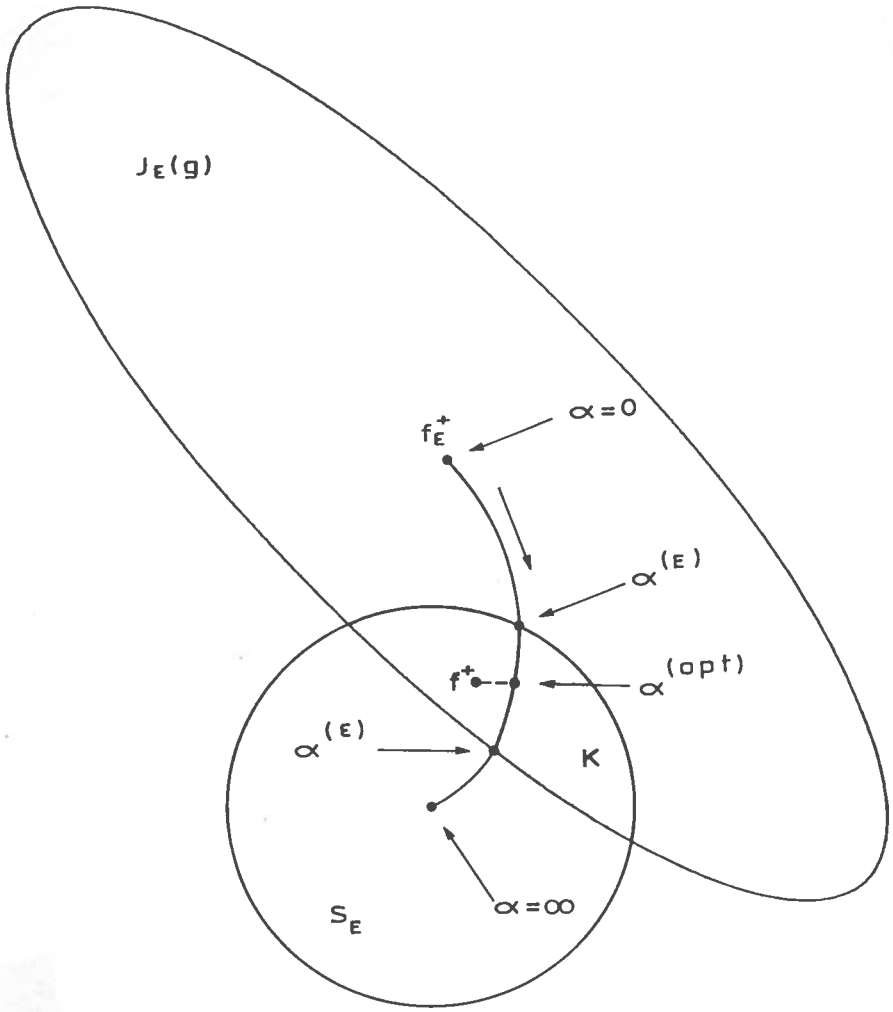

Fig.10 Two-dimensional representation of the trajectory described by $R_\alpha g_\epsilon$ as α increases from 0 to ∞. It is assumed that there exists a generalized solution associated with noisy data. Otherwise the trajectory goes to infinity when $\alpha \longrightarrow 0$.

which will now be given and discussed.

We say that a one-parameter family of operators $\{R_\alpha\}_{\alpha>0}$ defines

a regularizing algorithm for the approximate determination of the generalized inverse L^+ of the linear operator L if:

i') for any $\alpha>0$, $R_\alpha : Y \longrightarrow X$ is a continuous operator;

ii') for any $g \in R(L)$

$$\lim_{\alpha \downarrow 0} \| R_\alpha g - f^+ \|_X = 0 \quad . \tag{267}$$

When the operators R_α are linear, then we have a *linear regularizing algorithm*. It is possible however to introduce nonlinear regularizing algorithms for the solution of linear problems. We will give a few examples of such algorithms in Section D.

The parameter α is usually called the *regularization parameter* and, in general, it is a positive real number. In some cases, however, it may be convenient to introduce a discrete variable which take only integer numbers. In such a case we have a sequence $\{R^{(n)}\}$ of regularizing operators and the limit $\alpha \longrightarrow 0$ is replaced by the limit $n \longrightarrow \infty$. If we want a unified notation then we can put $R_\alpha = R^{(n)}$ when $[\alpha^{-1}]$= (integer part of α^{-1}) = n.

The meaning of condition ii') is obvious. It implies that, when $g \in R(L)$, it is possible to obtain arbitrarily accurate approximations of f^+ by means of continuous operators. Moreover, in the case of noisy data g_ϵ , satisfying condition (264), we still have an inequality analogous to (266) and precisely

$$\| R_\alpha g_\epsilon - f^+ \|_X \leq \| R_\alpha g - f^+ \|_X + \epsilon \| R_\alpha \| \quad . \tag{268}$$

The first term in the r.h.s. can be called the "approximation error", which is introduced when the non-continuous (or ill-conditioned) operator L^+, acting on exact data, is replaced by the continuous (or well-conditioned)

operator R_α. This "approximation error" tends to zero when α tends to zero. The second term in the r.h.s. represents "error propagation" from the data to the solution and it becomes exceedingly large when α tends to zero. Therefore it is clear that the choice of α will be based, in general, on a compromise between approximation error and noise propagation.

As in the case of the specific example provided by Eq.(226), given a noisy data function g_ϵ and given a regularizing algorithm $\{R_\alpha\}_{\alpha>0}$, the family of approximate solutions $f_{\alpha,\epsilon} = R_\alpha g_\epsilon$ will describe a trajectory in the Hilbert space X and, in general, there will be a point on this trajectory which has a minimum distance from the true solution f^+.

A similar definition can be introduced for the regularization of C-generalized inverses. Condition i) is not modified, while condition ii) is modified as follows:

ii") for any $g \in LD(C)$

$$\lim_{\alpha \downarrow 0} \| R_\alpha g - f_C^+ \|_X = 0 \qquad (269)$$

where f_C^+ is the C-generalized solution associated with g.

It is easy to show that the family of linear continuous operators $\{R_{C,\alpha}\}_{\alpha>0}$, defined by Eq.(256), is a regularization algorithm for L_C^+. In fact, the representation (257) of these operators coincides with Eq.(226), except for the factor β. Then, since $\beta \longrightarrow 1$ when $\alpha \longrightarrow 0$, Eq.(269) can be proved in the same way as Eq.(261).

Finally, a few words as concerns linear inverse problems with discrete data. In this case the definition of a regularization algorithm needs some modifications (Bertero *et al.*, 1988 b). In the case of an ill-posed problem, indeed, a regularization algorithm is a family of continuous (bounded) operators which approximate an operator which is

not bounded. But, for a problem with discrete data, the generalized inverse is always continuous since it is a linear operator on a finite dimensional space. The problem must be regularized when the norm of L^+ is much greater than $1/\|L\|$ (ill-conditioning) and therefore a regularization algorithm must provide an approximation of L^+ which has a norm smaller than the norm of L^+ . For these reasons we say that a one-parameter family of operators $\{R_\alpha\}_{\alpha>0}$ is a regularization algorithm for an inverse problems with discrete data when:

i) for any $\alpha > 0$, the range of R_α is contained in X_N, the subspace spanned by the functions ϕ_n;

ii) for any $\alpha > 0$, the norm of R_α is smaller than the norm of L^+, i.e.

$$\|R_\alpha\| \leq \|L^+\| = 1/\sigma_{N-1} \quad ; \tag{270}$$

iii) the following limit holds true in the sense of the norm of bounded operators

$$\lim_{\alpha \downarrow 0} \|R_\alpha - L^+\| = 0 \quad . \tag{271}$$

A similar definition can be given of a regularization algorithm for a C-generalized inverse (Bertero *et al*, 1988 b). In condition i) the subspace X_N is replaced by $R(L_C^*)$, which is the orthogonal complement of the null space of L_C and coincides with the subspace spanned by the functions ψ_n, Eq.(205), while in condition ii) the norm of L^+ is obviously replaced by the norm of L_C^+.

C. Spectral Windows

For a fixed $\alpha > 0$, consider the function of λ, $F_\alpha(\lambda) = (\lambda + \alpha)^{-1}$, defined on $(0,+\infty)$. Then the Tikhonov regularizer (226) can also be written in the form $R_\alpha = F_\alpha(L^*L)L^*$, where the operator $F_\alpha(L^*L)$ is obtained from $F_\alpha(\lambda)$ using the spectral representation (Yosida, 1966) of the self-adjoint, non-negative operator L^*L. Moreover, if the operator R_α is applied to a noise free image $g = Lf^+$, one obtains

$$f_\alpha = R_\alpha g = W_\alpha(L^*L)f^+ \tag{272}$$

where $W_\alpha(\lambda) = \lambda(\lambda + \alpha)^{-1}$. This function is small in the neighborhood of the spectral point $\lambda = 0$ (notice that this point belongs to the spectrum of L^*L if and only if the problem is ill-posed) and therefore the effect of the regularizing algorithm is a windowing (or filtering) of the spectral components of f^+ related to the ill-posedness of the problem. For example, in the case of a compact operator, using Eq.(230) and the relation $(g,v_k)_Y = (Lf^+,v_k)_Y = (f^+,L^*v_k)_X = \sigma_k(f^+,u_k)_X$, we obtain

$$f_\alpha = \sum_{k=0}^{+\infty} \frac{\lambda_k}{\lambda_k + \alpha} (f^+, u_k)_X u_k \tag{273}$$

where we have introduced the notation $\lambda_k = \sigma_k{}^2$. Analogously, in the case of a convolution operator, from Eq.(233) we obtain

$$f_\alpha(\mathbf{x}) = \frac{1}{(2\pi)^n} \int_\Omega \frac{|\hat{K}(\xi)|^2}{|\hat{K}(\xi)|^2 + \alpha} f^+(\xi)\, e^{i(\mathbf{x},\, \xi)} d\xi \quad . \tag{274}$$

In this case the spectrum of L^*L is the set of the values of the function $\lambda(\xi) = |\hat{K}(\xi)|^2$, $\xi \in \Omega$.

The previous remark suggests the idea of looking for regularization algorithms of the following form

$$R_\alpha = F_\alpha(L^*L)L^* \tag{275}$$

where now $\{F_\alpha(\lambda)\}_{\alpha>0}$ denotes a suitable family of functions defined on $(0,+\infty)$ and again $F_\alpha(L^*L)$ is given in terms of $F_\alpha(\lambda)$ using the spectral representation of L^*L (Bakushinskii, 1965; Groetsch, 1980; 1984). The problem is now to find sufficient conditions on $\{F_\alpha(\lambda)\}_{\alpha>0}$ assuring that $\{R_\alpha\}_{\alpha>0}$ is a regularization algorithm. These conditions can be given on the *window functions* $W_\alpha(\lambda) = \lambda F_\alpha(\lambda)$.

For example, it is not difficult to prove (Groetsch, 1980; Bertero, 1986) that, if $\{W_\alpha(\lambda)\}_{\alpha>0}$ is a family of real valued; piecewise continuous functions defined on $(0, +\infty)$ and satisfying the following conditions:

i) for any $\alpha > 0$, $\quad 0 \leq W_\alpha(\lambda) \leq 1$;

ii) for any $\lambda > 0$,

$$\lim_{\alpha \downarrow 0} W_\alpha(\lambda) = 1 \ ; \tag{276}$$

iii) for any $\alpha > 0$, there exists a constant c_α such that

$$W_\alpha(\lambda) \leq c_\alpha \lambda \ ; \tag{277}$$

then, the family of operators defined by Eq.(275), with $F_\alpha(\lambda) = \lambda^{-1} W_\alpha(\lambda)$,

is a regularization algorithm.

Conditions i)-iii) are satisfied by the Tikhonov window $W_\alpha(\lambda) = \lambda(\lambda+\alpha)^{-1}$. Another important example is the following one

$$W_\alpha(\lambda) = 0, 0 \leq \lambda < \alpha \; ; \quad W_\alpha(\lambda) = 1, \; \lambda \geq \alpha \tag{278}$$

which corresponds to a truncation of the spectral representation of L^*L. This regularization algorithm is very important both in the case of compact operators and in the case of convolution operators.

In the case of compact operators, however, it is more convenient to define spectral windows in terms of singular function expansions as follows

$$R_\alpha \, g = \sum_{k=0}^{\infty} \sigma_k^{-1} \, W_{\alpha,k} \, (g,v_k)_Y \, u_k \tag{279}$$

and therefore one must introduce a family of window sequences rather than a family of window functions. The conditions i),ii) above are unchanged (the variable λ is replaced by the index k), while condition iii) must be replaced by the requirement that, for any $\alpha > 0$, there exists a constant c_α such that: $W_{\alpha,k} \leq c_\alpha \sigma_k$. Then the use of the window function (278) is equivalent to the use of the following window sequence

$$W_{\alpha,k} = 1 \; , \quad k \leq [\alpha^{-1}] \quad ; \quad W_{\alpha,k} = 0 \; , \quad k > [\alpha^{-1}] \; . \tag{280}$$

The corresponding regularization algorithm is the well-known method of *truncated singular function expansions* (Twomey, 1965; Miller, 1970; Groetsch, 1984), which is also known as *numerical filtering*.

Analogously, in the case of a convolution operator it is convenient to define spectral windows in terms of the Fourier transform as follows

$$(R_\alpha g)(\mathbf{x}) = \frac{1}{(2\pi)^n} \int_\Omega W_\alpha(\xi) \frac{\hat{g}(\xi)}{\hat{K}(\xi)} e^{i(\mathbf{x}, \xi)} d\xi \quad . \qquad (281)$$

Again, conditions i) and ii) above are unchanged (the variable λ is replacedby the variable ξ), while condition iii) is replaced by the requirement that, for any $\alpha > 0$, there exists a constant c_α such that $W_\alpha(\xi) \leq c_\alpha |\hat{K}(\xi)|$. Then the use of the window function (278) is equivalent to the use of the following one

$$W_\alpha(\xi) = 1 \ , \ |\xi| \leq \alpha^{-1} \ ; \ W_\alpha(\xi) = 0 \ , \ |\xi| > \alpha^{-1} \qquad (282)$$

or, in other words, it is equivalent to the use of a cut-off in the Fourier integral.

We point out that, in the case of a regularization algorithm defined as in Eq.(281), the operator $W_\alpha(L^*L)$ is a convolution operator given by

$$[W_\alpha(L^*L) f](\mathbf{x}) = \int_{\mathbb{R}^n} A_\alpha(\mathbf{x}-\mathbf{x}') f(\mathbf{x}') d\mathbf{x}' \qquad (283)$$

where $A_\alpha(\mathbf{x})$ is the inverse Fourier transform of $W_\alpha(\xi)$. For example, in the case of functions defined on $(-\infty, +\infty)$, and of the window functions (282) we have

$$A_\alpha(x) = (\pi\alpha)^{-1} \operatorname{sinc}(x/\pi\alpha) \quad . \qquad (284)$$

It follows that, in the case of noise free data, the regularized solution is an average of the true solution over a distance of the order of α.

Other interesting window functions for the inversion of convolution operators in one dimension are the following:

a) The *triangular window*

$$W_\alpha(\xi) = (1 - \alpha\,|\xi|)\,,\ |\xi| \leq \alpha^{-1}\ ;\ W_\alpha(\xi) = 0\,,\ |\xi| > \alpha^{-1} \qquad (285)$$

which is related to the approximation of Fourier integrals in the sense of (C,1)-summability (Titchmarsh, 1948). In this case the averaging function $A_\alpha(x)$ is given by

$$A_\alpha(x) = (2\pi\alpha)^{-1}\,\mathrm{sinc}^2(x/2\pi\alpha) \qquad (286)$$

Notice that this averaging function is positive and therefore, in the absence of noise, the corresponding regularization algorithm provides positive approximations of positive functions (Bertero *et al.*, 1988 a).

b) The *Hanning window*

$$W_\alpha(\xi) = \tfrac{1}{2}\left[1 + \cos(\pi\alpha\xi)\right]\,,\ |\xi| \leq \alpha^{-1}\ ;\ W_\alpha(\xi) = 0\,,\ |\xi| > \alpha^{-1} \qquad (287)$$

which is well-known in the theory of signal processing (Kunt, 1980). The corresponding averaging function is

$$A_\alpha(x)=(4\pi\alpha)^{-1}\left\{\mathrm{sinc}[(x-\pi\alpha)/\pi\alpha]+2\mathrm{sinc}(x/\pi\alpha)+\mathrm{sinc}[(x+\pi\alpha)/\pi\alpha]\right\} \qquad (288)$$

and this is not positive. The negative parts, however, are quite small and the side-lobes are smaller than the side-lobes of the function (286), so that the use of this window can be very convenient in practice.

c) The *gaussian window*

$$W_\alpha(\xi) = \exp(-\alpha\xi^2/2) \tag{289}$$

with the corresponding averaging kernel

$$A_\alpha(x) = (2\pi\alpha)^{-1/2} \exp(-x^2/2\alpha) \tag{290}$$

Notice that this kernel is also positive and that side-lobes are absent. The disadvantage is that it can be used only in the case where $\hat{K}(\xi)$ tends to zero at infinity less rapidly than any gaussian, while the band-limited windows introduced above can be used for regularizing the inversion of an arbitrary convalution operator.

The method of spectral windows, based on the use of Fourier integrals, includes also the method of *filtered backprojection* (Natterer, 1986) which is presently the most important reconstruction algorithm in computerized tomography.

As a final remark we point out that the methods outlined above can also be used for the inversion of Laplace transform, and more generally for the inversion of integral operators of the type (62), as well as for the solution of Abel equation (Bertero, 1986; Bertero *et al*., 1988 a). In these cases the Fourier transform is replaced by the Mellin transform. Then the analysis runs parallel to that performed in the case of convolution operators.

D. Iterative Methods

Iterative methods are frequently used for the solution of n x n linear systems. The Jacobi method and the Gauss-Seidel method, for

example, are well-known since a long time. Complete account of iterative algorithms can be found in any text-book on numerical analysis (Ralston, 1965; Marchuk, 1975).

The most simple iterative process can be obtained by writing the linear system $\mathbf{Ax} = \mathbf{y}$ in the form $\mathbf{x} = \mathbf{(I-A)x} + \mathbf{y}$ which suggests the iteration $\mathbf{x}_{n+1} = \mathbf{(I-A)x}_n + \mathbf{y}$. The latter can also be written in the form $\mathbf{x}_{n+1} = \mathbf{x}_n - (\mathbf{Ax}_n - \mathbf{y})$ and it must be, in general, modified as follows : $\mathbf{x}_{n+1} = \mathbf{x}_n - \tau\,(\mathbf{Ax}_n - \mathbf{y})$, in order to have convergence. The arbitrary parameter τ is called *relaxation parameter* and the vector $\mathbf{r}_n = \mathbf{Ax}_n - \mathbf{y}$ is called the *residual* of the iterative process (Marchuk, 1975). Moreover the iterative process is called *stationary* if the parameter τ does not depend on a particular iteration, while it is called *non-stationary* if $\tau = \tau_n$ is changing from one iteration to another. The methods of steepest descent and conjugate gradient are examples of nonstationary iterative algorithms.

The interesting feature of some of these methods is that they can be extended to functional equations such as Eq.(5) and that they have regularizing properties in the sense specified in Section B. In other words the approximate solution provided by a finite number of iterations is a stable approximate solution and the number of iterations (or more precisely, the inverse of the number of iterations) plays the role of a regularization parameter.

The extension of the simple stationary iteration method mentioned above to the solution of first kind Fredholm integral equations is due to Landweber (Landweber, 1951) who also proves the convergence of the algorithm in the case of noise free data, using however too strong restrictions on the kernel (or, equivalently, on the relaxation parameter). An extension of this method, obtained by replacing the relaxation parameter with a fixed linear operator, was proposed by Strand (Strand,1974).

The results of Landweber and Strand apply essentially to Eq.(5) or,

more precisely, to Eq.(173), in the case where L is a compact operator. The extension to the general case of a linear continuous operator is given by Bialy (Bialy, 1959) who also proves convergence for the correct range of values of the relaxation parameter. A survey of these results with applications to least-squares linear signal restoration is given in (Sanz and Huang, 1983). Finally we mention that also the methods of steepest descent and conjugate gradient have been extended to Eq.(5) or Eq.(173) (Kammerer and Nashed, 1971; 1972) and that also in this case convergence of the algorithm has been proved for certain classes of noise free data.

Iterative reconstruction of distorted signals has also received much attention in the engineering literature (Schafer *et al* ., 1981). Examples are the recovery of the input to a linear shift-invariant system from its output (deconvolution), the restoration of a multidimensional signal from its projections and the extrapolation of a signal from a finite segment of that signal. These problems can be classified as linear inverse problems and, in fact, their mathematical representation is given by Eq.(5). In particular, for the problem of extrapolating a band-limited signal, which is equivalent to the problem of Fourier transform inversion with limited data, a very attractive algorithm was proposed by Gerchberg and Papoulis (Gerchberg, 1974; Papoulis, 1975). The convergence of this method in the case of noise free data was proved by De Santis and Gori (De Santis and Gori, 1975), using expansions in terms of the prolate spheroidal wavefunctions. The main interest of the method is that it can be easily implemented on a computer and that it can achieve a rather good super-resolution in the case of noise free data. It was later recognized (Maitre, 1981; Sanz and Huang, 1983) that this algorithm is just a special case of the Landweber-Bialy iteration with $\tau = 1$.

In this Section we give the main results concerning Landweber-Bialy iteration, steepest descent and conjugate gradient and we indicate why they can be considered as regularization algorithms for the approximate determination of the generalized solution. Therefore the basic equation is not Eq.(5) but Eq.(173).

The approximation of f^+ given by the n-th iteration will be indicated

by f_n and the corresponding residual r_n will be defined by

$$r_n = L^* L f_n - L^* g \qquad (291)$$

It is evident that the approximation f_n and the residual r_n belong to the Hilbert space X. This is not convenient in the case of inverse problems with discrete data, since in this case one must essentially compute N-dimensional vectors. A very simple modification of the equations is however possible (Bertero et al., 1988 b) by putting

$$f_n = L^* \mathbf{f}_n \quad , \quad r_n = L^* \mathbf{r}_n \qquad (292)$$

so that (assuming that the ϕ_n are linearly independent) we have

$$\mathbf{r}_n = \hat{\mathbf{L}} \mathbf{f}_n - \mathbf{g} \quad , \qquad (293)$$

where $\hat{\mathbf{L}}$ is the matrix associated with the operator LL^* and related to the Gram matrix of the functions ϕ_n by Eq.(128). Then all the algorithms can be formulated in terms of the vectors $\mathbf{f}_n$, $\mathbf{r}_n$ and of the matrix $\hat{\mathbf{L}}$.

1. *Landweber-Bialy iteration*

The sequence of the approximations is given by

$$f_0 = 0 \quad , \quad f_{n+1} = f_n - \tau r_n \qquad (294)$$

where τ is a fixed value of the relaxation parameter. Then, using Eq.(291) it

is easy to show that

$$f_n = R^{(n)}g = \tau \sum_{k=0}^{n-1} (I - \tau L^*L)^k L^*g \qquad (295)$$

and therefore this algorithm has the general structure (275) with

$$W^{(n)}(\lambda) = \lambda F^{(n)}(\lambda) = \tau \lambda \sum_{k=0}^{n-1} (1 - \tau \lambda)^k = 1 - (1 - \tau\lambda)^n \quad . \qquad (296)$$

This window function satisfies the conditions i)-iii) of Section C for values of λ in the spectrum of L^*L (which is interior to the interval $[0, \|L\|^2]$) when the relaxation parameter τ satisfies the following conditions

$$0 < \tau < 2 \|L\|^{-2} \quad . \qquad (297)$$

These are precisely the conditions assuring that, if $Pg \in R(L)$, then the sequence $\{f_n\}$ converges to the generalized solution f^+ of Eq.(5) (Bialy, 1959). It follows that the sequence $\{R^{(n)}\}$ defines a regularization algorithm. Then the problem of choosing an "optimum value" of the regularization parameter is equivalent to the problem of choosing an "optimum number" of iterations. In fact, in the case of noisy data, the first iterations improve the accuracy of the solution but, after a certain critical value, the noise induces instability and the quality of the solution is rapidly degraded.

Finally, it is not difficult to prove that the following inequalities hold true

$$\|r_{n+1}\|_X \leq \|r_n\|_X \quad . \qquad (298)$$

and also

$$\| f_{n+1} \|_X \geq \| f_n \|_X \quad . \tag{299}$$

These properties are analogous to properties proved for the regularization algorithm (226) and precisely to the fact that ϵ_α, Eq.(228), is an increasing function of α while E_α, Eq.(229), is a decreasing function of α.

2. *Steepest descent*

In this case the iteration scheme is given by

$$f_0 = 0 \quad , \quad f_{n+1} = f_n - \tau_n r_n \tag{300}$$

where

$$\tau_n = \| r_n \|_X^2 / \| L r_n \|_Y^2 \quad . \tag{301}$$

It has been proved that, when $Pg \in R(LL^*)$ (this condition is obviously stronger than the condition $Pg \in R(L)$), then f_n converges to f^+ (Kammerer and Nashed, 1971). Therefore, if we put $f_n = R^{(n)}g$, the family of operators $\{R^{(n)}\}$ defines a regularization algorithm. Notice that the operator $R^{(n)}$ is continuous but it is not linear. Moreover the inequalities (297) and (298) hold true also in this case. We have mentioned this algorithm for completeness but we do not know applications of it to the solution of inverse problems.

3. *Conjugate gradient*

In this case the iteration scheme is given by

$$f_0 = 0 \quad , \quad f_{n+1} = f_n - \tau_n p_n \tag{302}$$

where

$$p_0 = r_0 = -L^*g \quad , \quad p_n = r_n + \sigma_{n-1} p_{n-1} \tag{303}$$

and also

$$\tau_n = (r_n, p_n)_X / \| Lp_n \|_Y^2 \, , \, \sigma_{n-1} = - (Lr_n, Lp_{n-1})_Y / \| Lp_{n-1} \|_Y^2 \, . \tag{304}$$

It is known that, for an N-dimensional problem, this method is a finite iterative method, in the sense that a theoretical convergence in N steps is guaranteed. This theoretical result holds true, for example, in the case of inverse problems with discrete data. In practice roundoff errors prevent the achievement of this theoretical convergence. Moreover, in the case of a functional equation in an infinite-dimensional space, the number of iterations required for convergence is infinite.

The convergence of f_n, as given by Eqs. (302)-(304), to f^+ has been proved when $Pg \in R(LL^*L)$ (Kammerer and Nashed, 1972). This condition is stronger than the condition required for the convergence of the steepest descent method. Again, if we put $f_n = R^{(n)}g$, the sequence $\{R^{(n)}\}$ defines a regularization algorithm. The operators $R^{(n)}$ are continuous and nonlinear.

A comparison of the Gerchberg-Papoulis (or Landweber-Bialy) method and of the conjugate gradient method has been performed in the case of the extrapolation of a signal of finite extent (Maitre, 1981). The

result obtained by means of numerical simulations is that the conjugate gradient produces the same accuracy of the Gerchberg-Papoulis algorithm but that the number of required iterations is much smaller. In some cases, using the conjugate gradient, a factor of 5000 in the number of iterations was gained in order to obtain the same accuracy as the Gerchberg-Papoulis algorithm.

In fact, both Landweber-Bialy iteration and conjugate gradient method compute first those parts of the solution which belong to the large singular values. The conjugate gradient however seems to be more efficient in this procedure as indicated by arguments developed in (Natterer, 1986 c). An impressive example has been found in the case of Laplace inversion in a weighted space (Bertero *et al*, 1986 c); the approximate solution given by the n-th iteration of the conjugate gradient method practically coincides with the approximate solution obtained using the first n terms in the singular function expansion, at least for small values of n.

E. Choice of the Regularization Parameter

A regularization algorithm provides a one-parameter family of approximations of the unknown generalized solution f^+. This family describes a trajectory in the Hilbert space X and, as follows from the inequality (268), there exists a unique point of this trajectory which has minimum distance from f^+. This implies the existence of an optimum value of the regularization parameter for a given noisy image g_ϵ. The determination of this optimum value, however, requires the knowledge of the unknown generalized solution and therefore it cannot be performed in practice. It follows that the solution of a practical problem involves two essential steps: the first is the choice of the regularization algorithm and the second is the choice of a criterion for selecting the regularization

parameter.

From this point of view, the methods presented in Sectio A can be considered as methods for selecting the regularization parameter in the case of the regularization algorithm (226). The typical feature of these criteria is that some additional information about the solution and/or the error is required. An extension of some of them to more general regularization algorithm can be performed as follows.

We introduce in the general case two functions of the regularization parameter which have been already introduced in the case (226) and, precisely, the norm of the regularized solution

$$E_\alpha = \|R_\alpha g\|_X \tag{305}$$

and the *discrepancy function*

$$\epsilon_\alpha = \|LR_\alpha g - g\|_Y \tag{306}$$

The latter is the distance between the data computed using the approximation $f_\alpha = R_\alpha g$ and the real data.

Then we assume that these functions have the properties proved in the case of the algorithm (226) and precisely:

(a) E_α is a strictly decreasing function of α whose values at $\alpha = 0$ and $\alpha = \infty$ are respectively $\|f^+\|_X^2$ (∞, when the generalized solution does not exist) and zero.

(b) ϵ_α is a strictly increasing function of α, whose values at $\alpha = 0$ and $\alpha = \infty$ are respectively $\|Qg\|_Y$ and $\|g\|_Y$.

One can easily check that these conditions are also satisfied by the examples of spectral windows given in Section C and by the iterative

algorithms of Section D.

We consider now two criteria for selecting the regularization parameter. The first criterion is based upon the assumption that a bound E for the norm of f is known, i.e. $f \in S_E$, Eq.(236). If the prescribed constant E is smaller than the norm of the generalized solution f^+, then property (a) implies that there exists a unique value of α, say $\alpha^{(E)}$, which solves the equation $E_\alpha = E$. For $\alpha > \alpha^{(E)}$ we have $E_\alpha < E$ and therefore all the corresponding regularized solutions belong to the sphere S_E. Moreover, from condition (b) it follows that, for $\alpha > \alpha^{(E)}$, the discrepancy ϵ_α is greater than the discrepancy corresponding to $\alpha = \alpha^{(E)}$. We conclude that *$\alpha = \alpha^{(E)}$ is the value of the regularization parameter providing a regularized solution which is compatible with the prescribed constraint and which minimizes the discrepancy between the computed and the measured data.* Then it is obvious that the method of constrained least squares solutions of Section A.1 gives a value of the discrepancy function which is smaller than the value provided by any other regularization algorithm, for a given value of the prescribed constant E.

The second criterion is based upon the assumption that a bound ϵ on the error is known. Then, if ϵ satisfies the inequalities (241), property (b) implies that there exists a unique value of α, say $\alpha = \alpha^{(\epsilon)}$ which solves the equation $\epsilon_\alpha = \epsilon$. For $\alpha < \alpha^{(\epsilon)}$ we have $\epsilon_\alpha < \epsilon$ and therefore all the corresponding regularized solutions are compatible with the data within the accuracy ϵ. On the other hand, property (a) implies that, for $\alpha < \alpha^{(\epsilon)}$, the norm of f_α increases. We conclude that *$\alpha = \alpha^{(\epsilon)}$ is the value of the regularization parameter providing a regularized solution which is compatible with the measured data and which has minimal norm*. Notice that the method of Section A.2 gives a solution whose norm is smaller than the norm of any other regularized solution, for a given value ϵ of the error estimate.

This second method is also known as the *discrepancy principle* (Morozov, 1966; 1968). From the results given in Sect.A.2, it follows that, in the case of the regularization algorithm (226) this method always provides a regularized solution which strongly converges to the true generalized solution f^+ when the error of the data tends to zero. Then the question arises whether the same property is true for other regularization algorithms and, in particular, for spectral windows and iterative methods. The answer is, in general, negative if the discrepancy principle is formulated as above. But if the discrepancy principle is slightly modified, in the sense that it is required to find a value of α such that

$$\| LR_\alpha g - g \|_Y = \mu \epsilon \quad , \tag{307}$$

where $\mu > 1$ is a given (but arbitrary) constant, then it is possible to prove the convergence result for a large class of regularization algorithms, including, for example, the method of truncated singular function expansions and the Landweber-Bialy iterative method (Vainikko, 1982; Defrise and De Mol, 1987).

In the case of truncated singular function expansions it is possible to introduce a method for the selection of the regularization parameter, or equivalently of the number of terms in the expansion, which is analogous to the method of Section A.3.

If we assume that the solution satisfies the constraints (243) and if we keep in the expansion (179) only those terms which correspond to singular values fulfilling the condition

$$\sigma_K \geq \epsilon / E \quad , \tag{308}$$

then the resulting truncated singular function expansion satisfied the constraints (243) except for a factor of $\sqrt{2}$ (Miller, 1970). Notice that the quantity controlling the truncation of the expansion is a kind of signal-to-noise ratio and therefore we have here an extension of the

method of numerical filtering (Twomey, 1965). The criterion given by Eq.(308) applies, of course to the case of compact operators but it can also be extended to the general case of a continuous operator, when the regularization algorithm is defined by the spectral window (278) (Miller, 1970).

In the case of an ill-conditioned problem with discrete data the regularized solution always converges to the true generalized solution when the error of the data tends to zero. For example, it is obvious that, when $\epsilon \longrightarrow 0$, both $\alpha^{(\epsilon)}$ and $\alpha^{(E)}$, as defined in this Section, tend to zero. Moreover one can always use the method of truncated singular function expansions and the criterion (308) for the choice of the optimum number of terms.

Finally it is important to mention a method which has been proposed for the regularization algorithm (226) and which can be used only in the case of problems with discrete data. This is the method of *cross-validation* which was essentially suggested in the context of smoothing spline functions (Wahba and Wold, 1975a; 1975b) and later extended to more general problems (Wahba, 1977). This method does not require any upper bound on the solution and/or data error and is based upon the idea of letting the data themselves choose the value of the regularization parameter. More precisely it is required that a good value of the regularization parameter should predict missing data value.

If we consider an inverse problem with discrete data, formulated as in Chapt.III, Section A, we denote by $f_{\alpha,k}$ the minimizer of the functional

$$\Phi_{\alpha,k}[f] = N^{-1} \sum_{n \neq k} |(Lf)_n - g_n|^2 + \alpha \| f \|_X^2 \qquad (309)$$

where $(Lf)_n$ is defined by Eq.(124). This functional is just the functional (221), in the case where the data space is the usual Euclidean space, with the k-th data missing. The extension to the case of a weighted norm in Y is

easy. Then the *cross-validation function* $V_0(\alpha)$ is defined by

$$V_0(\alpha) = N^{-1} \sum_{k=1}^{N} |(Lf_{\alpha,k})_k - g_k|^2 \qquad (310)$$

and the cross-validation method consists in determining the unique value of α which minimizes $V_0(\alpha)$. The computation of the minimum is based on the relation (Golub *et al*, 1979; Craven and Wahba, 1979)

$$V_0(\alpha) = N^{-1} \sum_{k=1}^{N} |1 - A_{kk}(\alpha)|^{-2} |(Lf_\alpha)_k - g_k|^2 \qquad (311)$$

where f_α is the minimizer of the functional

$$\Phi_\alpha [f] = N^{-1} \sum_{n=1}^{N} |(Lf)_n - g_n|^2 + \alpha \|f\|_X^2 \quad , \qquad (312)$$

which is a special case of the functional (221), and $A_{kk}(\alpha)$ is the kk-entry of the NxN matrix

$$\mathbf{A}(\alpha) = LL^* (LL^* + \alpha I)^{-1} \quad . \qquad (313)$$

Notice that LL^* is essentially the Gram matrix of the functions ϕ_n since, in this formulation, $\mathbf{W} = N^{-1}\mathbf{I}$.

It has been shown (Golub *et al.*, 1979; Craven and Wahba, 1979) that, from the point of view of minimizing the predictive mean square error, the minimization of $V_0(\alpha)$ must be replaced by the minimization of the *generalized cross-validation function*, defined by

$$V(\alpha) = \left(N^{-1}\,\mathrm{Tr}[\mathbf{I} - \mathbf{A}(\alpha)]\right)^{-2} \left(N^{-1}\,\|[\mathbf{I} - \mathbf{A}(\alpha)]\,\mathbf{g}\|^2\right) \qquad (314)$$

where the norm is the usual euclidean norm. An important property of $V(\alpha)$ is its invariance with respect to permutations and, more generally, with respect to rotations of the data values.

The problem of the choice of the regularization parameter has been the subject of several papers and it is impossible to give in this review a complete account of all the criteria which have been suggested and of their main mathematical and computational properties. The criteria we have outlined are the most general and most significant in our opinion. Our feeling, however, is that it is not possible to find a criterion which can work for any ill-posed problem. Therefore, given an ill-posed problem one must investigate the various algorithms and criteria which have been proposed and perhaps invent a new one in order to take full account of the specific characteristics of the problem.

VI. INVERSE PROBLEMS AND INFORMATION THEORY

The results presented in the last Chapter indicate that there are plenty of algorithms for solving ill-posed or ill-conditioned inverse problems. This situation can be confusing but it is unavoidable. After years of theoretical investigations and of computational work, a rather generally accepted point of view is that no general method exists and that, even in solving a specific problem, it is convenient to use different algorithms for different classes of solutions.

The basic reason of this difficulty is that, very often, the data contain rather poor information about the solution. In the case of a compact operator, for example, this is related to the fact that the singular values tend to zero so that the data components corresponding to small eigenvalues are completely contaminated by noise. In more general cases it is a consequence of the fact that the exact image can be much more smooth than the corresponding object. The problems discussed in Chapt.II provide several striking examples of this situation. It follows that two completely different objects can produce very similar smooth images. Then the noise contribution h to the real image, Eq.(9), hides the smoothness of the exact image and it becomes impossible to distinguish between the two different objects.

A first consequence is that, in designing an algorithm for solving an inverse problem, one must never forget a general principle formulated by Lanczos (Lanczos, 1961, pg.132). "...*a lack of information cannot be remedied by any mathematical trickery* ". In other words, clever algorithms cannot produce miracles. The first basic point is to understand the information content of the data and the role of the available *a priori* informations about the solution.

Several concepts introduced in the theory of inverse and ill-posed problems go in this direction: stability estimates, resolution limits, number of degrees of freedom and so on. In this Chapter we attempt a presentation of these ideas with the scope of showing the

various relationships between them. We do not think that the result of this effort is already a completely satisfactory theory. We hope however that the main features of a future, complete theory can emerge from all these ideas.

A. *Modulus of Continuity and Uncertainty of the Solution*

Given a regularization algorithm and given a criterion for the choice of the regularization parameter, one has a recipe for computing an approximate solution of an inverse problem. Then one can try to have an answer to the following questions: a) the stability, or robustness of the algorithm; b) the convergence of the approximate solution to the true solution when the noise tends to zero.

The first is a typical problem of numerical analysis which can be solved by looking at the condition number or at other estimations of numerical stability. The second seems to be a purely mathematical question since, in practice, the noise is never zero. However, when the convergence result holds true, one knows that, by reducing the noise, one can get a better solution and therefore the proof of the convergence of an algorithm is also interesting from the practical point of view. For example the method of Chapt.V, Section A.2, which corresponds to the choice of the regularization parameter given by the discrepancy principle, provides a stable approximate solution which converges to the true generalized solution when the error of the data tends to zero.

We must point out, however, that, even if convergence is assured, the convergence can be arbitrarily slow since no rate of convergence can be found for arbitrary solutions. In fact, if we want to have a rate of convergence we must restrict the class of admissible solutions by means of some kind of *a priori* information. Then one can introduce a *modulus of continuity* which, as we will show, is essentially a measure of the *uncertainty* of the solution. An upper bound for the modulus of continuity

is also called a *stability estimate* (John, 1960; Miller, 1964; 1970).

We will assume, for simplicity, that the inverse operator L^{-1} exists and that, in general, it is not continuous. Several results, however, can be easily extended to the case of the generalized inverse L^{+} just by restricting L to $N(L)^{\perp}$, i.e. by taking $N(L)^{\perp}$ as the new object space.

We first define a *convergence rate* of the regularization algorithm $\{R_\alpha\}_{\alpha>0}$ in the case of exact data associated with functions f of a prescribed set H

$$\omega_H(\alpha) = \sup \left\{ \|R_\alpha L f - f\|_X \;\middle|\; f \in H \right\} . \tag{314}$$

An estimate of $\omega_H(\alpha)$, combined with inequality (268), can be used for obtaining a value of the regularization parameter which is optimum for the set H (Groetsch, 1984). In fact one can look for the value of α which minimizes the function $\omega_H(\alpha) + \epsilon \|R_\alpha\|$.

Moreover, in the case of noisy data g_ϵ corresponding, within an error ϵ, to solutions in the set H, we define a *modulus of convergence* of the regularization algorithm $\{R_\alpha\}_{\alpha>0}$ as follows (Franklin, 1974)

$$\sigma_H(\epsilon,\alpha) = \sup \left\{ \|R_\alpha g_\epsilon - f\|_X \;\middle|\; f \in H,\ \|Lf - g_\epsilon\|_Y \leq \epsilon \right\} . \tag{315}$$

Then, the inequality (268) implies that

$$\sigma_H(\epsilon,\alpha) \leq \omega_H(\alpha) + \epsilon \|R_\alpha\| . \tag{316}$$

Finally we introduce the *modulus of continuity* of the operator L^{-1} when restricted to LH :

$$\mu_H(\epsilon) = \sup \{ \|f\|_X \mid f \in H, \ \|Lf\|_Y \leq \epsilon \} \ . \qquad (317)$$

If the set H contains a neighborhood of 0, then $\mu_H(\epsilon)$ *is a continuous, increasing function of* ϵ. Moreover, if the set H is compact, then the topological Lemma of Tikhonov, mentioned in Chapt.V, Section A.4, implies that $\mu_H(\epsilon) \longrightarrow 0$ when $\epsilon \longrightarrow 0$. We point out, however, that the compactness of H is only a sufficient condition for this result: it is easy to find examples of bounded sets H which are not compact and such that $\mu_H(\epsilon) \longrightarrow 0$ when $\epsilon \longrightarrow 0$. Examples are given, for instance in (Bertero, 1982; 1986)

The relationship between the modulus of convergence and the modulus of continuity is clarified by the following result (Franklin, 1974): *if the set H contains a neighbourhood of 0, then for any linear regularization algorithm* $[R_\alpha]_{\alpha > 0}$ *and any* α

$$\mu_H(\epsilon) \leq \sigma_H(\epsilon,\alpha) \ . \qquad (318)$$

The relevance of this result is obvious: given a set H of solutions, no regularization algorithm and no choice of the regularization parameter can provide a modulus of convergence which tends to zero more rapidly than the modulus of continuity and therefore the last one is the best possible convergence rate for the approximation of elements of the set H when data are noisy.

We point out that this optimum converge rate is obtained in the case of contrained least squares solutions, as defined in Chapt.V, Sect.A.4. In fact it is proved by Ivanov (Ivanov, 1962) that: *if the set H is closed, convex and symmetric with respect to 0 and if* $\mu_H(\epsilon) \longrightarrow 0$ *when* $\epsilon \longrightarrow 0$, *then, for sufficiently small* ϵ

$$\| \tilde{f} - \tilde{f}' \|_X \leq \mu_H(\epsilon) \quad . \tag{319}$$

where $\tilde{f}$ and $\tilde{f}'$ *are the constrained least squares solutions associated respectively with data g and g' satisfying the condition* $\| g - g' \|_Y \leq \epsilon$.

Very often the set H can be characterized as a level set of a functional of the kind (13), i.e.

$$H = \{ f \in D(C) \mid \| Cf \|_Z \leq E \} \quad . \tag{320}$$

Then, if the constraint operator C satisfies the conditions of Chapt. II, Section A the set H is closed, convex and symmetric with respect to 0. Moreover H is also compact when C has a compact inverse C^{-1} and in such a case all the assumptions of Ivanov theorem on the convergence rate of constrained lest squares solutions are satisfied.

Another approximate solution which has the optimum convergence rate is provided by Eq.(255) with $\alpha = (\epsilon/E)^2$. If we denote this solution by $\tilde{f}_0$, then it has been proved (Miller, 1970) that

$$\| \tilde{f}_0 - f \|_X \leq \sqrt{2}\, \mu_H(\epsilon) \quad , \tag{321}$$

for any arbitrary $f \in K$, the set K being defined by

$$K = \{ f \in X \mid f \in H , \ \| Lf - g \|_Y \leq \epsilon \} \quad . \tag{322}$$

We stress now another interpretation of the modulus of continuity (317). The set K defined in Eq.(322) is a generalization of the set introduced in Chapt.V, Sect.A.3. In that case H is the sphere of radius E in X. It is obvious that K is *the set of all the admissible approximate solutions compatible with the data function g within the error* ϵ. Any

element of K is an acceptable approximate solution of the problem and therefore *the diameter of K is a measure of the uncertainty of the solution* for a given *a priori* information (the set H) and a given noise level (the value of ϵ).

When the set H is convex and symmetric with respect to zero - conditions satisfied by the set (320) - the diameter of K can be easily estimated in terms of $\mu_H(\epsilon)$ and therefore the modulus of continuity is also an estimate of the uncertainty of the solution. In fact, if f, f' $\in$ K, then $f_1 = (f - f')/2$ belongs also to H. From the inequality : $\| Lf_1 \|_Y \leq \frac{1}{2}(\| Lf - g \|_Y + \| Lf' - g \|_Y) \leq \epsilon$, it follows $\| f_1 \| \leq \mu_H(\epsilon)$ and therefore

$$\text{diam}\,(K) = \sup \{ \| f - f' \|_X \,|\, f, f' \in K \} \leq 2\,\mu_H(\epsilon) \,. \qquad (323)$$

In other words, when $\mu_H(\epsilon) \longrightarrow 0$ for $\epsilon \longrightarrow 0$, the uncertainty of the solution tends to zero in the case of vanishing noise. We also notice that the modulus of continuity provides an estimate of the uncertainty which is independent of the data function g.

Finally we consider the problem of estimating the modulus of continuity $\mu_H(\epsilon)$, i.e. the problem of determining stability estimates. We restrict ourselves to the case of a set H defined as in Eq.(320) since only in this case it is possible to give rather general methods. A first remark is the following: if we introduce the quantity

$$\mu_C(\epsilon) = \sup \{ \| f \|_X \,|\, \| Cf \|_Z \leq 1 \,,\ \| Lf \|_Y \leq \epsilon \} \qquad (324)$$

then we have

$$\mu_H(\epsilon) = E\, \mu_C(\epsilon/E) \qquad , \qquad (325)$$

and therefore we can restrict the attention to the estimate of $\mu_C(\epsilon)$.

If we introduce now the quantity

$$\mu_C^{(0)}(\epsilon) = \sup \{ \| f \|_X \mid \| Lf \|_Y^2 + \epsilon^2 \| Cf \|_Z^2 \leq \epsilon^2 \} \tag{326}$$

and the sets

$$K^{(0)} = \{ f \in X \mid \| Lf \|_Y^2 + \epsilon^2 \| Cf \|_Z^2 \leq \epsilon^2 \} \tag{327}$$

$$K^{(1)} = \{ f \in X \mid \| Lf \|_Y^2 + \epsilon^2 \| Cf \|_Z^2 \leq 2\epsilon^2 \} \tag{328}$$

from the inclusions

$$K^{(0)} \subset K \subset K^{(1)}, \tag{329}$$

(here K is defined by Eqs.(322) and (320)), we derive the inequalities

$$\mu_C^{(0)}(\epsilon) \leq \mu_C(\epsilon) \leq \sqrt{2}\, \mu_C^{(0)}(\epsilon) \ . \tag{330}$$

We find that $\mu_C^{(0)}(\epsilon)$ is a good stability estimate of the modulus of continuity. Moreover $\mu_C^{(0)}(\epsilon)$ can be easily computed. In the case of a finite dimensional problem, $K^{(0)}$ is an ellipsoid and $\mu_C^{(0)}(\epsilon)$ is the maximum length of the semi-axes of $K^{(0)}$. Therefore it can be determined by solving an eigenvalue problem. This result holds true in general, i.e. the computation of $\mu_C^{(0)}(\epsilon)$ can always be reduced to the solution of a spectral problem for the operator $L^*L + \epsilon^2 C^*C$. This operator is positive

definite and therefore its spectrum has a positive lower bound $\gamma^2(\epsilon)$ which can be determined by solving the following variational problem

$$\gamma^2(\epsilon) = \inf \left\{ \|Lf\|_Y^2 + \epsilon^2 \|Cf\|_Z^2 \;\middle|\; \|f\|_X = 1 \right\} . \tag{331}$$

Then, by comparing this equation with Eq.(326), we have

$$\mu_C^{(0)}(\epsilon) = \epsilon/\gamma(\epsilon) \tag{332}$$

For more details see (Bertero *et al.* , 1980).

The relation (332) can be used for computing stability estimates in several important cases (Bertero *et al*, 1980; Bertero, 1982; 1986).

Finally we comment an important remark due to John (John, 1960). We say that we have *Hölder continuity* in the dependence of the solution on the data when there exist constants A, α with $0 < \alpha \leq 1$ such that

$$\mu_C^{(0)}(\epsilon) \leq A\epsilon^{\alpha} \tag{333}$$

while we say that we have *logarithmic continuity* when

$$\mu_C^{(0)}(\epsilon) \leq A\,|\ln \epsilon|^{-\alpha} \tag{334}$$

with α arbitrary.

When Hölder continuity holds true John calls the ill-posed problem *well-behaved* . In such a case only a fixed percentage of the significant digits is lost in determining f from g and therefore the uncertainty of the solution is not very severe. On the contrary, in the case of logarithmic continuity, even an improvement in the noise level of several orders of magnitude does not induce a significant reduction of the uncertainty of the solution. In other words, the information content of the

data is practically noise independent.

It is important to realize that the kind of continuity does not only depend on the problem, i.e. the operator L, but also on the set H. For the same problem we can have Hölder continuity for certain sets H and logarithmic continuity for others.

For certain ill-posed problems one can have Hölder continuity when one prescribes bounds on a finite number of derivatives of the unknown function f. In such a case the problem is also called *mildly ill-posed* . Examples are tomography, Abel transform inversion and numerical differentiation (Bertero *et al.*, 1980; Louis and Natterer, 1983).

On the other hand, when prescribed bounds on a finite number of derivatives only imply logarithmic continuity, the problem is also called *severely ill-posed*. Examples are the Laplace transform inversion (Bertero *et al.*, 1982), the problem of bandwidth extrapolation and, in general, the solution of a first kind Fredholm integral equation with an analytic kernel (Bertero *et al.* , 1980). As we already pointed out, however, one can have Hölder continuity also in the case of a severely ill-posed problem just by choosing an appropriate set H of admissible solutions. For example, in the case of Laplace transform inversion, we have Hölder continuity when the set H is a bounded set of functions having suitable analyticity properties (Bertero *et al.*, 1982).

B. *Functional Estimation and Resolution Limits*

In several applications one is not directly interested in estimating the solution of the problem but rather the value of some suitable functionals of the solution. These functionals can be, for example, a moment of a given order (the average radius or the average occupied volume in the problems of particle sizing described in Chapt.II, Section C.3) or, in general, a linear continuous functional, i.e. a generalized

moment. Several examples related to the applications of Abel equation are described in (Anderssen, 1986).

Stated in a precise form the problem is the following: given an element $\phi \in X$, estimate the value of the functional

$$F_{\phi}(f) = (f, \phi)_X \tag{335}$$

where f is a solution or a generalized solution of Eq.(5).

The important feature of these problems is that some of them are much more stable than the problem of determining the solution f itself. In fact for a certain class of these functionals the evaluation problem is well-posed. Since this class can be characterized for any given linear inverse problem we have here a precise answer to a question succintly expressed by Sabatier (Sabatier, 1984) as the need to identify and ask , within the framework of indirect measurements, well-posed questions about the phenomenon of interest.

In general, however, the problem of estimating the functional (335) is not well-posed and therefore it is necessary to use regularization theory. A modulus of continuity can also be defined for this problem and by considering special classes of functionals one can find along these lines a rather precise definition of the resolution limits achievable in a given inverse problem.

1. *Well-Posed Functional Estimations*

Consider a functional of the type (335) with $\phi \in R(L^*)$. Then there exists a function $\psi \in Y$ such that $\phi = L^*\psi$ and, by replacing this relation in Eq.(335), we find

$$F_\phi(f) = (f, L^* \psi)_X = (Lf, \psi)_Y \quad . \tag{336}$$

In such a case F_ϕ is a linear continuous functional of Lf and therefore it is obvious that, given the data function g, the estimation of the value of the functional is

$$a_\phi = (g, \psi)_Y \quad . \tag{337}$$

In other words these functionals can be estimated directly from the data without any need of solving for the unknown function f. It is also obvious that the dependence of the value of F_ϕ on g is continuous. If δg is a variation of g and δa_ϕ the corresponding variation of a_ϕ, then, from Eq.(337), using Schwarz inequality we get

$$|\delta a_\phi| \leq \|\psi\|_Y \cdot \|\delta g\|_Y \leq \epsilon \|\psi\|_Y \quad . \tag{338}$$

It is obvious, however, that the problem can be ill-conditioned because the error δa_ϕ can be exceedingly large when $\|\psi\|_Y$ is too large.

It is also important to notice that Eq.(336) characterizes all the functionals which depend continuously on Lf and therefore it also characterizes all the functionals which can be directly estimated in terms of g. Assuming again, for simplicity, that the inverse operator L^{-1} exists, this result can be proved as follows. If the functional F_ϕ has the property

$$|F_\phi(f)| \leq c \|Lf\|_Y \tag{339}$$

where c is a constant independent of f, then, given $f \in X$, it is always

possible to find $g \in R(L)$ such that $f = L^{-1}g$. It follows, using the inequality (339), that

$$\left|F_\phi(f)\right| = \left|(f,\phi)_X\right| = \left|(L^{-1}g,\phi)_X\right| \leq c\,\|g\|_Y \qquad (340)$$

Therefore $(L^{-1}g,\phi)_X$ is a linear and continuous functional on R(L) and it can be extended, by continuity, to a linear and continuous functional on R(L). This implies, thanks to Riesz representation theorem, that there exists an element $\psi \in R(L)$ such that

$$(L^{-1}g,\phi)_X = (g, \psi)_Y \quad . \qquad (341)$$

If we put now $g = Lf$, we have

$$(f,\phi)_X = (Lf, \psi)_Y = (f, L^*\psi)_X \qquad (342)$$

and therefore $\phi = L^*\psi$.

As an example, consider the case of the integral operator (87), whose adjoint is just the bandlimiting operator (80) when the values of x are restricted to the interval [-1,1]. It follows that, in the inversion of the integral operator (87), a linear and continuous functional can be estimated directly from the data whenever the function ϕ is the restriction to the interval [-1,1] of a band-limited function.

2. *Ill-Posed Functional Estimation*

When $\phi \notin R(L^*)$, the functional (335) cannot be estimated

directly in terms of the data g and therefore the use of regularization theory is required in such a case. It is obvious that if $\tilde{f}$ is some regularized solution of the problem, then the corresponding regularized value of the functional is

$$\tilde{a}_\phi = (\tilde{f}, \phi)_X \qquad (343)$$

Also for this problem one can define a modulus of continuity as follows

$$\mu_H(\epsilon;\phi) = \sup \{ |(f,\phi)_X| \mid f \in H, \ \|Lf\|_Y \leq \epsilon \} \qquad (344)$$

and the analysis runs parallel to that outlined in Section A. In particular, in the case where the set H is of the type (320), one can introduce the stability estimate

$$\mu_C^{(0)}(\epsilon;\phi) = \sup \{ |(f,\phi)_X| \mid \|Lf\|_Y^2 + \epsilon^2 \|Cf\|_Z^2 \leq \epsilon^2 \} \qquad (345)$$

and inequalities analogous to the inequalities (330) hold true also in this case. Moreover it is possible to compute $\mu_C^{(0)}(\epsilon;\phi)$. The result is (Miller, 1970; Bertero *et al.*, 1980)

$$\mu_C^{(0)}(\epsilon;\phi) = \epsilon \, ([L^*L + \epsilon^2 C^*C]^{-1} \phi, \phi)_X^{\frac{1}{2}} \qquad (346)$$

Using this equation it is not difficult to prove that $\mu_C^{(0)}(\epsilon;\phi) \longrightarrow 0$ for $\epsilon \longrightarrow 0$ and ϕ arbitrary, whenever the constraint operator C has a bounded inverse (Bertero, 1982).

As a conclusion we can state that regularized solutions provide also stable estimations of linear and continuopus functionals and that the corresponding stability estimates can be easily computed.

3. *Resolution limits*

We apply now the analysis of the previous Sections to the investigation of resolution limits. We restrict ourselves to the case where X is a space of square integrable functions. Then the functional (335) takes the form

$$(f,\phi)_X = \int f(x)\,\phi(x)\,dx \qquad (347)$$

(for simplicity we consider only the case of real functions of one variable). Moreover, let us assume that the function ϕ is positive and peaked upon the point x_0 (and, for example, symmetric with respect to x_0), that its integral is 1 and its second central moment is σ^2, i.e.

$$\int \phi(x)\,dx = 1 \quad , \qquad \int (x - x_0)^2\,\phi(x)\,dx = \sigma^2 \quad . \qquad (348)$$

Then the value of the functional (347) can be considered as a blurred value of f at the point $x = x_0$. In other words we want to estimate a local average of f over a resolving length σ. It is quite natural to predict that the estimation error (for a fixed noise level ϵ) will grow for decreasing values of σ, i.e. for increasing resolution and therefore the achievable resolution will be obtained by fixing the acceptable value of the estimation error.

Consider the case C = I, then, as follows from Eq.(346) the absolute error in the estimation of the functional (347) is bounded by

$$\mathcal{E}_{abs}(\epsilon;\phi) = \epsilon\,([L^*L + \epsilon^2 I]^{-1}\,\phi, \phi)_X^{1/2} \quad . \tag{349}$$

It is more interesting, however, to introduce relative errors. This approach is quite natural in the case of stochastic regularization (Wiener filters). Here we can proceed along the lines indicated in (Bertero *et al.*, 1980). The absolute error (349) is the maximum value of $|(f,\phi)_X|$ under the constraint $\|Lf\|_Y^2 + \epsilon^2 \|f\|_X^2 \leq \epsilon^2$. This *a posteriori* constraint is compatible with the *a priori* constraint $\|f\|_X \leq 1$. Then it is easy to see that the maximum value of $|(f,\phi)_X|$ under this *a priori* constraint is just $\|\phi\|_X$. Therefore we can define as an estimate of the relative error the ratio between the *a posteriori* and the *a priori* maximum value of $|(f,\phi)_X|$, i.e.

$$\mathcal{E}_{rel}(\epsilon;\phi) = \epsilon\,\frac{([L^*L + \epsilon^2 I]^{-1}\,\phi, \phi)_X^{1/2}}{\|\phi\|_X} \quad . \tag{350}$$

It is immediate to show that $\mathcal{E}_{rel}(\epsilon;\phi) \leq 1$.

We can consider now a family $\{\phi_\sigma\}_{\sigma>0}$ of functions having different values of the variance σ^2 (for instance, gaussians of variance σ^2). Then the relative error (350) is a function of ϵ and σ, let us say $\mathcal{E}_{rel}(\epsilon,\sigma)$. Since ϕ_σ is an approximation of the Dirac delta function, the L^2-norm of ϕ_σ tends to infinity when $\sigma \longrightarrow 0$. Without loss of generality we can assume that this norm is a decreasing function of σ. Then from

Eq.(350) one can derive the following properties of the relative error:

(a) For fixed σ, $\mathcal{E}_{rel}(\epsilon,\sigma)$ is an increasing function of ϵ;

(b) For fixed ϵ, $\mathcal{E}_{rel}(\epsilon,\sigma)$ is a decreasing function of σ and it tends to one (100% error) when $\sigma \longrightarrow 0$.

The typical behaviour of $\mathcal{E}_{rel}(\epsilon,\sigma)$ as a function of σ is indicated in Fig.11.

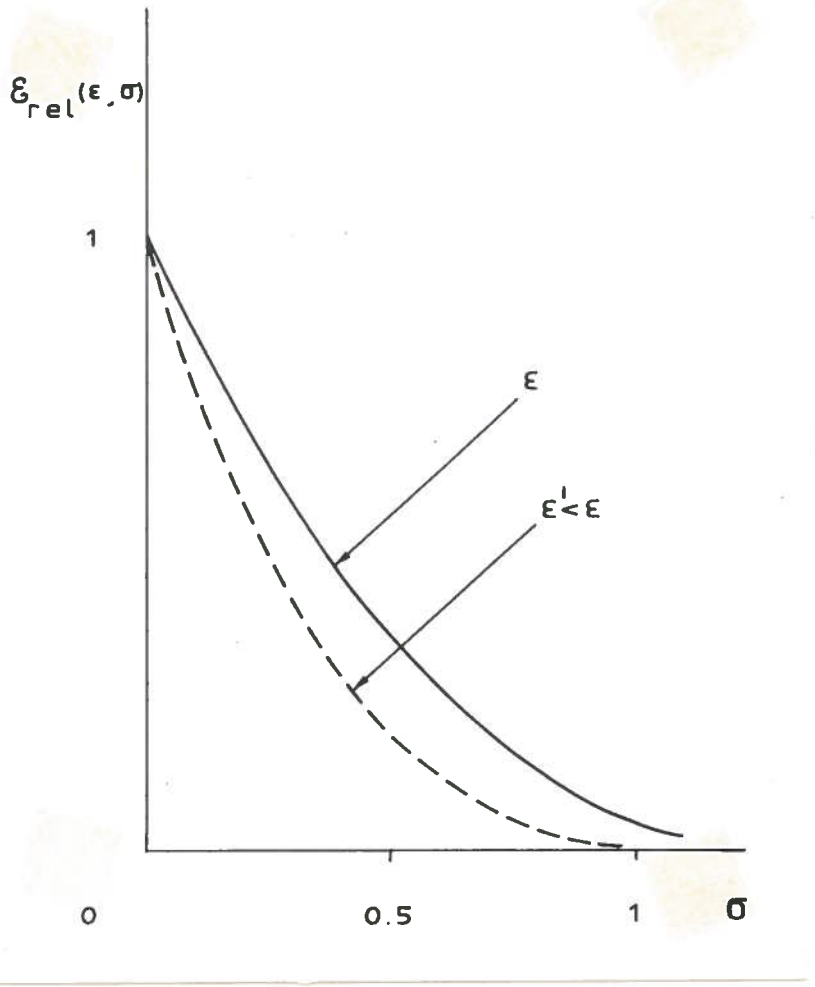

Fig.11 Illustration of the trade-off between relative error and resolution. It is assumed that the unit of resolution is some typical length related to the problem (for instance, the Rayleigh resolution distance in the case of an imaging system).

Numerical computations of these curves for various inverse problems are presented in (Bertero *et al.*, 1980 a; 1980 b; Abbiss *et al.*, 1983).

We comment these results. The plot of $\varepsilon_{rel}(\epsilon,\sigma)$ as a function of σ represents the trade-off between resolution and error. If we fix the acceptable error on the averaged solution, for instance 10%, then from the curve we can deduce the corresponding resolution and viceversa. As also indicated in Fig.11, if we reduce the error on the data, say $\epsilon' < \epsilon$, then we have an improvement in resolution. However, when the inverse problem we are considering is affected by logarithmic continuity, the change in the plot of $\varepsilon_{rel}(\epsilon,\sigma)$ is imperceptible even when we have a change in ϵ of several order of magnitude. This effect is clearly shown by the computations presented in (Bertero *et al.*, 1980 b) and concerning the inversion of the Slepian operator (84). In all the problems which exhibit this behaviour, we have a resolution limit which is practically noise independent and which represents a fundamental limitation in the possibility of recovering details of the unknown object. In the problem considered in (Bertero *et al*, 1980 b), for example, this limit is just the classical Rayleigh resolution distance. A possibility of going beyond this limit is indicated in (Bertero and Pike, 1982) when:

1) the full image is measured, so that the data contain more information about the solution;
2) the value of c is small and this is essentially a limitation on the size of the unknown object (which is an important *a priori* information).

In general, the resolution limit, as defined by the previous approach, depends on the point x_0 (see Eq.(348)) and therefore we do not have a uniform resolution over the domain of the variable x. In the inversion of a convolution operator, however, due to the translational invariance of the problem, the resolution does not depend on x.

The most simple example of a problem with non-uniform resolution

is provided by the inversion of an integral operator of the form (62). Since such an operator can be transformed into a convolution operator by taking as new variables the logarithm of the old ones, it is obvious that in this case we have a uniform resolution in the log-variable. This implies that the resolution distance increases for increasing values of the variable r which appears in Eq.(62). In fact it is more appropriate to introduce a *resolution ratio* δ_0 rather than a resolution distance (McWhirter and Pike, 1978). The meaning of this resolution ratio is the following : given two delta pulses at the positions r_1 and r_2 , it is impossible to resolve these pulses unless $r_2 \geq \delta_0 r_1$. A general method for the estimation of δ_0 is discussed in (McWhirter and Pike, 1978)

C. *Number of Degrees of Freedom*

The concept of number of degrees of freedom was first introduced in Optics in terms of the sampling expansion and successively clarified in terms of the basic properties of the PSWF (Toraldo di Francia, 1969 a). The typical step behaviour of the eigenvalues of the prolate spheroidal functions indicates that, while the object can have an arbitrary number of degrees of freedom, i.e. an arbitrary number of large components with respect to the PSWF, the image has always a finite one. This number, which is also called by Toraldo di Francia the *Shannon number*, is given by $S = c/\pi$ and it is proportional to the space-bandwidth product. In fact the Shannon number is approximately equal to the number of sampling points interior to the geometric image and it is essentially a characteristic parameter of the optical instrument, giving a measure of the information transmitted by the instrument itself. In subsequent work (Toraldo di Francia, 1969 b) it was recognized that this number is in fact noise-dependent. The dependence however is so weak that the original conclusion was correct in practice. An interpretation of this result was

later given in terms of the theory of ill-posed problems by showing that the problem of inverting the Slepian operator is affected by logarithmic continuity (Bertero *et al.*, 1980 a).

The concept of *number of degrees of freedom* or, more precisely, the concept of a noise dependent number of degrees of freedom, can be generalized to all inverse problems which can be treated in terms of singular systems, i.e. problems corresponding to the inversion of a compact operator and inverse problems with discrete data (Twomey, 1974; Bertero and De Mol, 1980 a; Pike *et al.*, 1984).

As discussed in Chapt.V, an acceptable regularized solution of these problems can be provided by a truncated singular function expansion. When the noise level ϵ is known and the *a priori* information on the solution is represented by a prescribed bound E on the norm of the solution, the truncation rule is given by Eq.(308). This condition has also a nice statistical interpretation. Assume that the data is represented by Eq.(9) and that both f and h are representatives of *zero-mean , uncorrelated random processes*. Moreover assume that the signal f is from a white noise process with power spectrum E^2 and that h is also from a white noise process with power spectrum ϵ^2. Then the variance of any given component of f with respect to the basis $\{u_k\}$ is : $< |(f, u_k)_X|^2 > = E^2$. Analogously the variance of any given component of h with respect to the basis $\{v_k\}$ is : $< |(h, v_k)_Y|^2 > = \epsilon^2$. From the relation, $(g, v_k)_Y = (Lf, v_k)_Y + (h, v_k)_Y = \sigma_k (f, u_k)_X + (h, v_k)_Y$, we get

$$< |(g, v_k)_Y|^2 > = \sigma_k^2 E^2 + \epsilon^2 \quad . \tag{351}$$

This equation shows that the variance of a given component of the data consists of two terms: the first is the contribution of the object while the second is the contribution of the noise. The first term tends to zero when $k \longrightarrow \infty$, or becomes very small for large k in the case of ill-conditioned inverse problems with discrete data; the second term is

constant. Moreover the first term is a decreasing function of k. Therefore for k greater than some critical value, the variance of the noise is greater than the variance of the object contribution and the corresponding data components do not contain any information about the object. Eq.(308) is just the requirement that the variance of the object contribution must be greater than the variance of the noise.

If we introduce the quantity

$$N(\epsilon/E) = \max \{k+1 \,|\, \sigma_k \geq \epsilon/E\} \quad , \tag{352}$$

then, in the case of the inversion of the Slepian operator, it is approximately equal to the Shannon number and therefore it is quite natural to call it the *number of degrees of freedom* (NDF) also in the case of a more general problem. The NDF is a function of the signal-to-noise ratio E/ϵ and it gives a measure of the *information content* of the data. It is a very useful parameter for estimating, in general, how many distinct elements one can resolve with the available data. From the examples discussed in Chapt.III it follows that for some important inverse problems with discrete data, such as the moment problem or Laplace transform inversion, the NDF can be quite small (of the order of 3 or 5, in any case less than 10). For other problems such as tomography the NDF can be very large. We find here another way for introducing a distinction between mildly and severely ill- posed problems.

In general, in the case of a mildly ill-posed problem, where it is possible to restore Hölder continuity by means of prescribed bounds on a small number of derivatives of the solution, the NDF depends rather strongly on the signal-to-noise ratio and it is possible to obtain a significant increase in the NDF by an increase of E/ϵ. On the contrary, in the case of a severely ill-posed problem, affected by logarithmic continuity, the NDF is nearly independent of the signal-to-noise ratio, at least in the case of reasonable values of this quantity. This usually happens if the singular values σ_k tends to zero exponentially fast when

$k \longrightarrow \infty$ (this is the case, for example, of the Slepian operator). Then, if N_0 is the value of $N(\epsilon/E)$ corresponding to some preassigned value of the signal-to-noise ratio, say E_0/ϵ_0, one can easily deduce that

$$N(\epsilon/E) = (N_0 - 1) + c \log_{10} (E\epsilon_0/\epsilon E_0) \qquad (353)$$

where c is a constant (Bertero and De Mol, 1981 a). Therefore, if N_0 is very large, even when improving the signal-to-noise ratio by many orders of magnitude, we do not get a significant improvement in the NDF. This is what happens in the problem discussed by Toraldo di Francia and mentioned at the beginning of this Section, where the NDF can be considered as practically noise independent and equal to N_0.

It is important to point out that the NDF can always be introduced (and computed) in the case of an inverse problem with discrete data, since in this case one can always use singular function expansions, as indicated in Chapt. III, Section A. Then the NDF can depend not only on the signal-to-noise ratio E/ϵ but also on the number N of data points. This is true, for example, in the case of the finite Hausdorff moment problem (Chapt. III, Section D). If we define the NDF as the number of singular values greater than 10^{-3} (we recall that the singular values are just the square roots of the eigenvalues of the Hilbert matrix), then we have NDF=4 for N=4, NDF=6 for N=10, NDF=8 for N=50, NDF=9 for N=100. As we see the NDF depends on the number N of given moments, even if the dependence is rather weak. This dependence, however, is related to the fact that the (infinite dimensional) Hausdorff moment problem does not correspond to the inversion of a compact operator and therefore the singular values and singular functions of the finite Hausdorff moment problem do not have a limit when $N \longrightarrow \infty$.

The situation is different when the problem with discrete data is a finite version of an infinite dimensional problem corresponding to the inversion of a compact operator. In such a case the NDF can be defined for

the infinite-dimensional problem and the NDF of any finite version of it cannot be greater than the limiting NDF. But, when the number of points of the finite version is sufficiently large, its first singular values are good approximations of the corresponding singular vbalues of the compact operator. It follows that, for N greater than a suitable number N_0 of data points, the NDF is independent of N and equal to the NDF of the compact operator. An illustration of this behaviour, in the case of the Poisson transform inversion, is given in (Bertero and Pike, 1986).

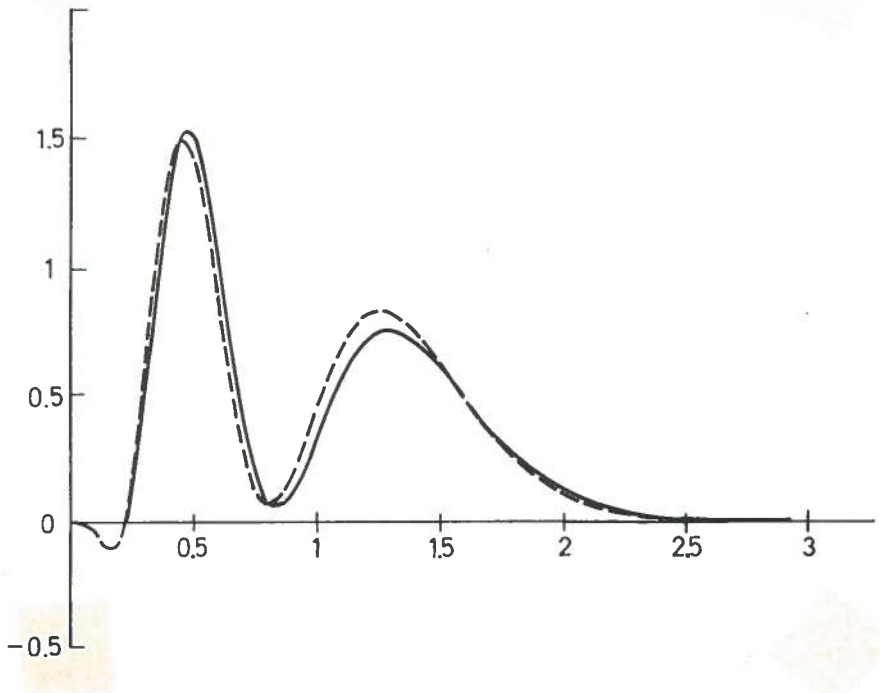

Fig.12 Reconstruction of two delta pulses in the case of Laplace transform inversion in a weighted space, using 32 equidistant points (dotted line) and 4 geometrically spaced points (full line).

We have also some indications that, in the case of the inversion of a compact operator, satisfactory results can be obtained using a finite version of the problem with a number of (suitably placed) data points which just coincides with the NDF of the infinite dimensional p[roblem. We give an example taken from the inversion of the Laplace transformation in a weighted space, using as a weight the gamma distribution (Bertero *et al.*, 1985 c). In such a case the Laplace transformation is compact and the NDF, defined as the number of singular values greater than 10^{-2}, is 4. In Fig. 12 we give the reconstruction of two delta pulses having the same mass, i.e. 0.5. As we see, using 4 geometrically spaced data points, it is possible to obtain a reconstruction which practically coincides with the reconstruction obtained using 32 uniformly spaced data points. Of course the position of the four points must be optimized (in such a case the criterion is the minimization of the condition number).

As a conclusion, we find that this example provides a strong indication of the fact that the NDF can coincide with the optimum number of data points. In other words one needs a number of data points just equal to the maximum number of pieces of information which can be extracted from complete (infinite dimensional) noisy data.

D. Impulse Response Function : Another Approach to Resolution Limits

Given a regularizing operator R_α, with α fixed, and given an exact (noise free) data $g = Lf$, the corresponding regularized solution is $f_\alpha = R_\alpha Lf$. Since f_α converges to f, when the operator L has an inverse L^{-1}, and to f^+, when L has a generalized inverse L^+, the operator

$$T_\alpha = R_\alpha L \tag{354}$$

is an approximation of the identity operator in the first case and of the

projection operator over $N(L)^{\perp}$ in the second case. This is the mathematical interpretation of the operator T_α, emphasizing the fact that, in the noise free case, the regularized solution f_α is just a (stable) approximation of the true solution f (or generalized solution f^+).

There exists, however, an interesting physical interpretation of the operator T_α. Since the operator L describes the transmission of the signal by the instrument in the absence of noise while the operator R_α describes the processing of the data by the computer in the absence of round-off errors, the operator T_α describes the total effect of both the transmitting instrument and the processing computer (in the absence of any kind of error).

If T_α is an integral operator

$$f_\alpha(x) = (T_\alpha f)(x) = \int A_\alpha(x, x') f(x')\, dx' \qquad (355)$$

the averaging kernel $A_\alpha(x, x')$ is the *impulse response function* of the system consisting of the instrument plus the computer. Moreover if, for a fixed x, $A_\alpha(x, x')$ as a function of x' has the form of a central lobe flanked by decreasing side-lobes, the width of the central lobe may be used as a measure of the resolution achievable at the point x by means of the algorithm R_α.

Notice that the form (355) for the approximate solution is just the starting point of the Backus-Gilbert method. Moreover, in the case of inverse problems with discrete data, assuming that X is a space of square integrable functions and that the regularized solutions are defined by means of windowed singular function expansions, the kernel $A_\alpha(x, x')$ is given by

$$A_\alpha(x,x') = \sum_{k=0}^{N-1} \sigma_k^{-1} W_k(\alpha)\, u_k(x)\, u_k(x') \qquad (356)$$

In this approach it is obvious that the resolution is determined by the choice of the regularization parameter. This choice, however, can be an important point in the case of mildly ill-posed problems, while in the case of severely ill-posed problems the choice of the regularization parameter depends very weakly on the noise and we find again a resolution limit which is practically noise independent.

It is obvious that the form and the width of the central lobe depend, in general, on the regularization algorithm. We think, however, that the width does not strongly depend on the algorithm since it is a measure of the achievable resolution and therefore a measure of the information which can be extracted from the data. We must never forget the principle formulated by Lanczos and emphasized at the beginning of this Chapter. Even if this question must still be carefully investigated, we give here an example in support of our statement.

The example is taken from the inversion of Fraunhofer diffraction data. This corresponds to the inversion of an integral operator of the form (62) with $K(x) = J_1(x)^2/x$. The corresponding first kind Fredholm integral equation can be conveniently approximated by means of the so-called *exponential sampling method* (Ostrowsky *etal*, 1981). When the problem has been discretized in such a way, one can compute the singular values and singular functions. For example, using 32 geometrically spaced data points and assuming that the support of the unknown function is interior to the interval [7,130] (the wavelength of the incident radiation is taken as a unit of the radius of the particles), one finds 18 singular values greater than 10^{-2}.

In Fig. 13 we give various reconstructions of a delta pulse of unit mass obtained using various regularization algorithms. The radius of the particles is represented on a logarithmic scale, in units of the wavewlength of the incident radiation. In all the reconstructions the

maximum number of used singular functions is 18. In the figure we have superimposed 10 different reconstructions corresponding to 10 different contaminations of the exact data by means of random errors of the order of 1 %. In this way a clear picture of the robustness of the solution is obtained.

In Fig. 13 a) we plot the reconstruction obtained by means of the

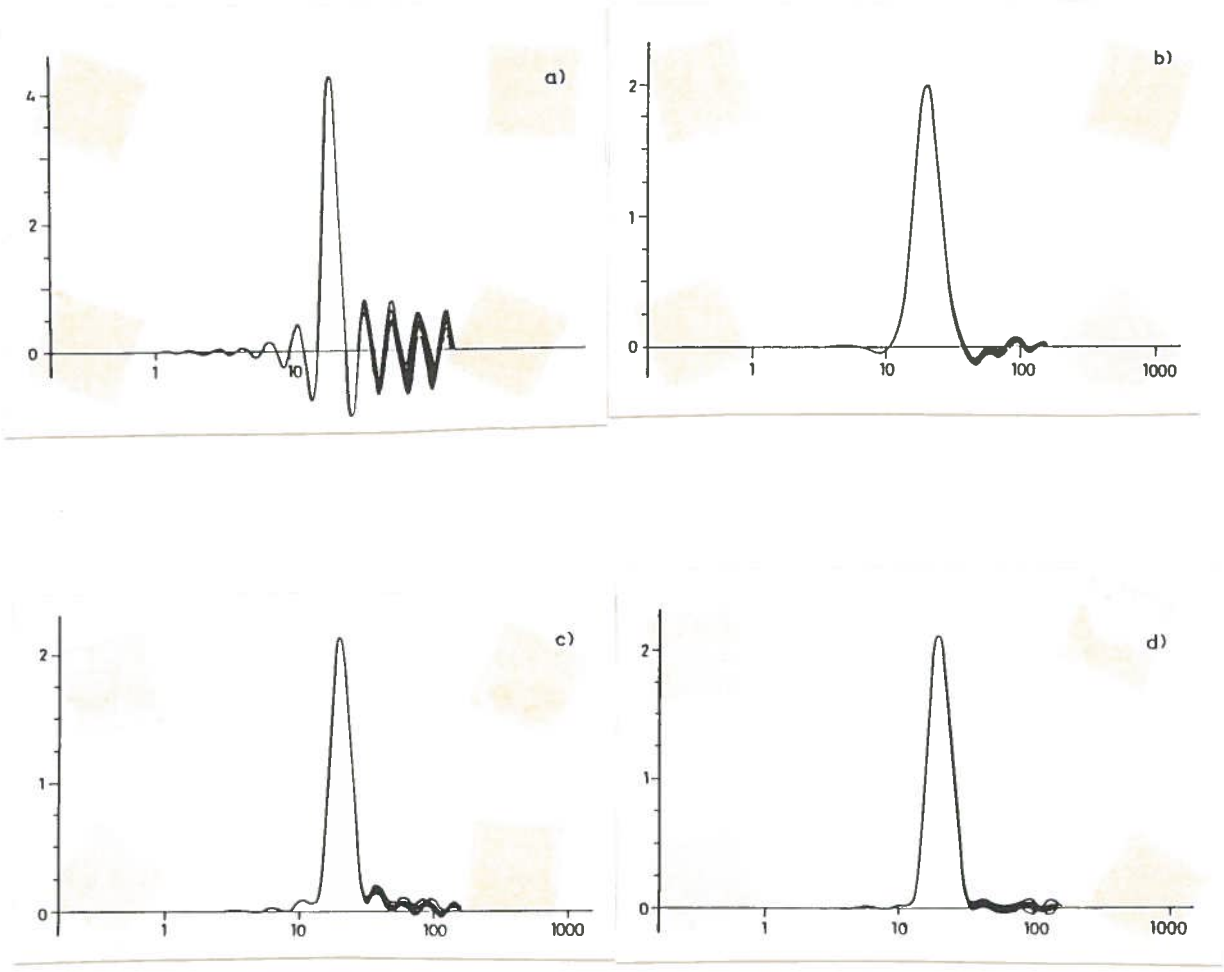

Fig. 13 Reconstruction of a delta pulse in the case of the inversion of Fraunhofer diffraction data. a) Truncated singular function (TSF) expansion. b) Tikhonov regularization with $\alpha = 10^{-3}$. c) TSF expansion with triangular window. d) TSF expansion with Hanning window.

truncated singular function expansion. This provides the maximum resolution achievable by means of 18 singular functions, but this resolution is obtained at the cost of very large side-lobes. In the other parts of the figure we plot the reconstructions obtained by means of various filterings of the previous truncated singular function expansion. In particular, Fig. 13 b) corresponds to the Tikhonov window, Fig. 13 c) corresponds to the triangular window and Fig. 13 d) corresponds to the Hanning window. In all these cases we have a loss in resolution of approximately a factor of 2 with respect to the reconstruction of Fig. 13 a) but the side-lobes are always much smaller than in Fig.13 a) and the reconstructions are nearly positive.

The previous example seems to indicate that we can obtain positivity at the cost of a loss in resolution, a result which is in conflict with a rather diffuse opinion about the beneficial effect of the constraint of positivity in the solution of inverse problems. This question, however, goes beyond the scope of this paper, which is essentially devoted to linear methods for linear problems. We just mention it here as an example of the many questions which are still open.

AKNOWLEDGEMENTS

It is my pleasure to thank Laura Opisso for the careful preparation of the manuscript. Her ability in correctly decoding and type-writing math. has greatly facilitated my work.

REFERENCES

Abbiss, J.B., Defrise, M., De Mol, C. and Dhadwal, H.S. (1983). Regularized iterative and noniterative procedures for object restoration in the presence of noise: an error analysis. *J. Opt. Soc. Am.* **73**, 1470.

Anderssen, R.S. (1986). The linear functional strategy for improperly posed problems. In "Inverse Problems", J.R. Cannon and U.Hornung, eds. Birkhauser, Basel.

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68**, 337.

Askne, J.I.H. and Westwater, E.R. (1986). A review of ground-based remote sensing of temperature and moisture by passive microwave radiometers. *IEEE Trans. Geosci. Remote Sensing* **GE-24**, 340.

Backus, G. and Gilbert, F. (1968). The resolving power of gross Earth data. *Geophys. J. R. Astron. Soc.* **16**, 169.

Backus, G. and Gilbert, F. (1970). Uniqueness in the inversion of inaccurate gross Earth data. *Phil. Trans. R. Soc.* **266**, 123.

Bakushinskii, A.B. (1965). A numerical method for solving Fredholm integral equations of the first kind. *USSR Comp. Math. Math. Phys.* **5** (No.4), 226.

Balakrishnan, A.V. (1976). "Applied Functional Analysis". Springer, Berlin.

Baltes, H.P., ed. (1978). "Inverse Source Problems in Optics". Topics in Current Physics, vol.9. Springer, Berlin.

Baltes, H.P., ed. (1980). "Inverse Scattering Problems in Optics". Topics in Current Physics, vol.20. Springer, Berlin.

Barcilon, V. (1986). Inverse eigenvalue problems. In "Inverse Problems", ed. G. Talenti. Lect.Notes in Math., Vol.1225. Springer, Berlin.

Bertero, M. (1982). "Problemi Lineari non ben posti e Metodi di Regolarizzazione". Pubbl. Istituto di Analisi Globale e Applicazioni, Firenze.

Bertero, M. (1986). Regularization methods for linear inverse problems. In

"Inverse Problems", ed. G. Talenti. Lect. Notes in Math., Vol.1225. Springer, Berlin.

Bertero, M. and De Mol, C. (1981a). Ill-posedness, regularization and number of degrees of freedom. *Atti Fond. G. Ronchi* **36**, 619.

Bertero, M. and De Mol, C. (1981b). Stability problems in inverse diffraction. *IEEE Trans. Antennas Propagat.* **AP-29**, 368.

Bertero, M. and Grünbaum, A.F. (1985). Commuting differential operators for the finite Laplace transform. *Inverse Problems* **1**, 181.

Bertero, M. and Pike, E.R. (1982). Resolution in diffraction-limited imaging, a singular value analysis - I. The case of coherent illumination. *Optica Acta* **29**, 727.

Bertero, M. and Pike, E.R. (1986). Intensity fluctuation distributions from photon counting distributions: a singular-system analysis of Poisson tranform inversion. *Inverse Problems* **2**, 259.

Bertero, M and Viano, G.A. (1978). On probabilistic methods for the solution of improperly posed problems. Bollettino U.M.I. **15-B**, 483.

Bertero, M., De Mol, C. and Viano, G.A. (1980 a). The stability of inverse problems. In "Inverse Scattering Problems in Optics", H.P.Baltes ed., Topics in Current Physics, vol.20, pg.161. Springer, Berlin.

Bertero, M., Viano, G.A. and De Mol, C. (1980 b). Resolution beyond the diffraction limit for regularized object restoration. *Optica Acta* **27**, 307.

Bertero, M., Boccacci, P. and Pike, E.R. (1982). On the recovery and resolution of exponential relaxation rates from experimental data: a singular value analysis of the Laplace transform inversion in the presence of noise. *Proc. R. Soc. Lond.* **A383**, 15.

Bertero, M., De Mol, C., Pike, E.R. and Walker, J.G. (1984). Resolution in diffraction limited imaging: IV - The case of uncertain localization or non-uniform illumination of the object. *Opt.Acta* **31**, 923.

Bertero, M., De Mol, C. and Pike, E.R. (1985 a). Linear inverse problems with discrete data. I: General formulation and singular system analysis, *Inverse Problems* **1**, 301.

Bertero, M., Brianzi, P. and Pike, E.R. (1985 b). On the recovery and

resolution of exponential relaxation rates from experimental data. III. The effect of sampling and truncation of data on the Laplace transform inversion. *Proc. R. Soc. Lond* **A398**, 23.

Bertero, M., Brianzi, P. and Pike, E.R. (1985 c). On the recovery and resolution of exponential relaxation rates from experimental data: Laplace transform inversions in weighted spaces. *Inverse Problems* **1**,1

Bertero, M., Grünbaum, A.F. and Rebolia, L. (1986 a). Spectral properties of a differential operator related to the inversion of the finite Laplace transform. *Inverse Problems* **2**, 131.

Bertero, M., Poggio, T. and Torre, V. (1986 b). Ill-posed problems in early vision. M.I.T. Memo 924.

Bertero, M., Brianzi, P., Defrise, M. and De Mol, C. (1986 c). Iterative inversion of experimental data in weighted spaces. In "Proc.U.R.S.I. Int. Symp. on Electr. Theory", Budapest, August 25-29, Part A, 315. Akademiai Kiado, Budapest.

Bertero, M., Brianzi, P. and Pike, E.R. (1987). Super-resolution in confocal scanning microscopy. *Inverse Problems* **3**, 195.

Bertero, M., Brianzi, P., Pike, E.R. and Rebolia, L. (1988 a). Linear regularizing algorithms for positive solutions of linear inverse problems. To be published in *Proc. R. Soc.Lond.*

Bertero, M., De Mol, C. and Pike, E.R. (1988 b). Linear inverse problems with discrete data. II: Stability and regularization. To be published in *Inverse Problems*

Bialy, H. (1959). Iterative behandlung linearer funktionalgleichungen. *Arch. Rot. Mech. Anal.* **4**, 166.

Bleinstein, N. and Cohen, J. (1977). Nonuniqueness in the inverse source problem in acoustics and electromagnetics. *J. Math. Phys.* **18**, 194.

Boerner, W.M., Brand, H., Cram, L.A., Gjessing, D.T., Jordan, A.K., Keydel, W., Schwierz, G. and Vogel, M., eds. (1983) "Inverse Methods in Electromagnetic Imaging". Part I and Part II. Reidel, Dordrecht.

Bojarski, N.N. (1966). A survey of electromagnetic inverse scattering. Syracuse Univ. Res. Corp., Special Projects Lab. Rpt., DDC # AD-813-851.

Cabayan, H.S., Murphy, R.C. and Pavlâsek, T.J.F. (1973). Numerical stability and near-field reconstruction. *IEEE Trans. Antennas Propagat.* **AP-21**,346.

Campbell, L.L. (1968). Sampling theorem for the Fourier transform of a distribution with bounded support. *SIAM J. Appl. Math.* **16**, 626.

Cannon, J.R. and Hornung, U., eds. (1986). "Inverse Problems". ISNM 77. Birkhäuser, Basel.

Carasso, A. and Stone, A.P., eds. (1975). "Improperly Posed Boundary Value Problems". Pitman, London.

Carter, W.H. (1970). Computational reconstruction of scattering objects from holograms. *J. Opt. Soc. Am.* **60**, 306.

Carter, W.H. and Ho, P.C. (1974). Reconstruction of inhomogeneous scattering objects from holograms. *Appl. Opt.* **13**, 162.

Chadan, K. and Sabatier, P.C. (1977). "Inverse Problems in Quantum Scattering Theory". Springer, Berlin.

Chernov, L.A. (1967). "Wave Propagation in a Random Medium". Dover, New York.

Colin, L., ed. (1972). "Mathematics of Profile Inversion". NASA Technical Memorandum, NASA TMX-62-150.

Colli Franzone, P. , Taccardi, B. and Viganotti, C. (1977). An approach to inverse calculation of epicardial potentials from body surface maps. *Adv. Cardiol.* **21**, 167.

Colton, D. (1984). The inverse scattering problem for time-harmonic acoustic waves. *SIAM Rev.* **26** (3), 323.

Courant R. and Hilbert, D. (1962). "Methods of Mathematical Physics". Vol.II. Interscience, New York.

Craig, I.J.D. and Brown, J.C. (1986). "Inverse Problems in Astronomy". Adam Hilger, Bristol.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation, *Num. Math.* **31**, 377.

Davison, M.E. (1981). A singular value decomposition for the Radon transform in n-dimensional euclidean space. *Numer. Funct. Anal. and Optimiz.* **3**, 321.

Defrise, M. and De Mol, C. (1987). A note on stopping rules for iterative regularization methods and filtered SVD. In " Inverse Problems : An Interdisciplinary Study ", P.C. Sabatier, ed., Advances in Electronic and Electron Physics, Supplement 19, pg. 261. Academic Press, New York.

De Santis, P. and Gori, F. (1975). On an iterative method for superresolution. *Optica Acta* **22**, 691.

Devaney, A.J. (1978). Nonuniqueness in the inverse scattering problem. *J. Math. Phys.* **19**, 1526.

Devaney, A.J. (1981). Inverse-scattering theory within the Rytov approximation. *Opt. Lett.* **6**, 374.

Devaney, A.J. (1982). A filtered backpropagation algorithm for diffraction tomography. *Ultrasonic Imaging* **4**, 336.

Devaney, A.J. (1984). Geophysical diffraction tomography. *IEEE Trans. Geoscience Remote Sensing* **GE-22**, 3.

Devaney, A.J. and Wolf, E. (1973). Radiating and nonradiating classical current distributions and the fields they generate. *Phys. Rev.* **D8**, 1044.

Devaney, A.J. and Wolf, E. (1974). Multipole expansions and plane wave representations of the electromagnetic field. *J. Math. Phys.* **15**, 234.

Di Chiro, G. and Brooks, R.A. (1979). The 1979 Nobel prize in Physiology or Medecine. *Science* **206**, 1060.

Duchon, J. (1976). Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces, *Analyse Numerique* **10**, 5.

Fercher, A.F., Bartelt, H., Becker, H. and Wiltschko, E. (1979). Image formation by inversion of scattered field data: experiments and computational simulation. *Appl. Opt.* **18**, 2427.

Fox, D. and Pucci, C. (1958). The Dirichlet problem for the wave equation. *Ann. Mat. Pura Appl.* **46**, 155.

Franklin, J.N. (1970). Well-posed stochastic extensions of ill-posed linear

problems. *J. Math. Anal. Appl.* **31**, 682.

Franklin, J.N. (1974). On Tikhonov's method for ill-posed problems. *Math. Comp.* **28**, 889.

Gerchberg, R.W. (1974). Super-resolution through error energy reduction. *Optica Acta* **21**, 709.

Golub, G.H., Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215.

Gori, F. and Guattari, G. (1985). Signal restoration for linear systems with weighted impulse. Singular value analysis for two cases of low-pass filtering. *Inverse Problems* **1**, 67.

Gregory, R.T. and Karney, D.L. (1969) "A collection of Matrices for Testing Computational Algorithms". Wiley-Interscience, New York.

Greville, T.N.E. (ed.) (1969). "Theory and Applications of Spline Functions". Academic Press, New York.

Grimson, W.E.L. (1982). A computational theory of visual surface interpolation. *Phil. Trans. R. Soc. Lond.* **B298**, 395.

Groetsch, C.W. (1977). "Generalized Inverses of Linear Operators". Dekker , New York.

Groetsch, C.W. (1980). On a class of regularization methods. *Boll. Un. Mat. Ital.* **17-B**, 1411.

Groetsch, C.W. (1984). "The theory of Tikhonov Regularization for Fredholm Equations of the First Kind". Research Notes in Math., Vol.105. Pitman, Boston.

Groetsch, C.W. (1986). Regularization with linear equality constraints. In "Inverse Problems", ed. G. Talenti. Lect. Notes in Math., Vol.1225. Springer, Berlin.

Grünbaum, F.A. (1982). A remark on Hilbert's matrix. *Linear Algebra Appl.* **43**, 119.

Grünbaum, F.A. (1986). Some mathematical problems motivated by medical imaging. In "Inverse Problems", ed. G.Talenti. Lect.Notes in Math., Vol.1225. Springer, Berlin.

Haario, H. and Somersalo, E. (1985). On the numerical implementation of the Backus-Gilbert method. *Cahiers Math. de Montpellier* **32** , 107.

Hadamard, J. (1902). Sur les problèmes aux dérivées partielles et leur signification physique. *Bull. Univ. Princeton* **13**, 49.

Hadamard, J. (1923). "Lectures on Cauchy's Problem in Linear Partial Differential Equations". Yale Univ. Press, New Haven.

Hadamard, J. (1964). "La Theorie des Equations aux Derivees Partielles". Editions Scientifiques, Pekin.

Herman, G.T. (ed.) (1979). "Image Reconstruction from Projections - Implementation and Applications", Topics in Applied Phys., Vol.32. Springer, Berlin.

Herman, G.T. (1980). "Image Reconstruction from Projections - The Fundamentals of Computerized Tomography". Academic Press, New York.

Herman, G.T. and Natterer, F. (eds.) (1981). "Mathematical Aspects of Computerized Tomography". Lect.Notes in Med.Inf., Vol.8. Springer, Berlin.

Hildreth, E.C. (1984). Computation of the velocity field. *Proc. R. Soc. Lond.* **B221**, 189.

Hoenders, B.J. (1978). The uniqueness of inverse problems. In "Inverse Source Problems in Optics", ed. H.P.Baltes. Topics in Current Physics, Vol.9. Springer, Berlin.

Imbriale, W.A. and Mittra, R. (1970). The two-dimensional inverse scattering problem. *IEEE Trans. Antennas Propagat.* **AP-18**, 633.

Ivanov, V.K. (1962). On linear problems which are not well-posed. *Soviet Math. Dokl.* **3**, 981.

Ivanov, V.K. (1966). The approximate solution of operator equations of the first kind. *U.S.S.R. Comp. Math. Math. Phys.* **6** (No.6), 197.

John, F. (1955). Numerical solution of the equation of heat conduction for preceding times. *Ann. Mat. Pura Appl.* **40**, 129.

John, F. (1960). Continuous dependence on data for solutions of partial differential equations with a prescribed bound. *Comm. Pure Appl.*

Math. **13**, 551.

Kac, M. (1966). Can one hear the shape of a drum? *Am. Math. Monthly* **73**, no.4, part II, 1.

Kammerer, W.J. and Nashed, M.Z. (1971). Steepest descent for singular linear operators with nonclosed range. *Appl. Anal.* **1**, 143.

Kammerer, W.J. and Nashed, M.Z. (1972). On the convergence of the conjugate gradient method for singular linear operator equations. *SIAM J. Numer. Anal.* **9**, 165.

Kato, T. (1966). "Perturbation Theory for Linear Operators". Springer, Berlin.

Keller, J.B. (1976). Inverse Problems. *Am. Math. Montly* **83**, 107.

Kunt, M. (1980). "Traitement Numerique des Signaux". Georgi, St-Saphorin.

Lanczos, C. (1961). "Linear Differential Operators". Van Nostrand, London.

Landau, H.J. and Pollak, H.O. (1961). Prolate spheroidal wave functions, Fourier analysis and uncertainty. II. *Bell System Tech.. J.* **40**, 65.

Landau, H.J. and Pollak, H.O. (1962). Prolate spheroidal wave functions, Fourier analysis and uncertainty. III: The dimension of the space of essentially time- and band-limited signals. *Bell System Tech. J.* **41**, 1295.

Landweber, L. (1951). An iteration formula for Fredholm integral equations of the first kind. *Amer. J. Math.* **73**, 615.

Lavrentiev, M.M. (1967). "Some Improperly Posed Problems of Mathematical Physics". Springer, Berlin.

Lewis, R.M. (1969). Physical optics inverse diffraction. *IEEE Trans. Antennas Propagat.* **AP-17**, 308.

Lord Rayleigh, J.W.S. (1877). "The Theory of Sound". Dover, New York 1945.

Louis, A.K. (1984). Orthogonal function series expansions and the null space of the Radon transform. *SIAM J. Math. Anal.* **15**, 621.

Louis, A.K. (1986). Incomplete data problems in X-ray computerized tomography. I. Singular value decomposition of the limited angle

transform. *Numer. Math.* **48**, 251.

Louis, A.K. and Natterer, F. (1983). Mathematical problems of computerized tomography. *Proc. IEEE.* **71**, 379.

Luneburg, R.K. (1964). "Mathematical Theory of Optics". Univ. of California Press, Berkeley.

Mager, R.D. and Bleinstein, N. (1978). An examination of the limited aperture problem of physical optics inverse scattering. *IEEE Trans. Antennas Propagat.* **AP-26**, 695.

Magnanini, R. and Papi, G. (1985). An inverse problem for the Helmholtz equation. *Inverse Problems* **1**, 357.

Maitre, H. (1981). Iterative superresolution. Some new fast methods. *Optica Acta* **28**, 973.

Mandelbrojt, S. and Schwartz, L. (1965). Jacques Hadamard (1865-1963). *Bull.Amer.Math. Soc.* **71**, 107.

Marchuk, G.I. (1975). "Methods of Numerical Mathematics". Springer, Berlin.

McWhirter, J.G. and Pike, E.R. (1978). On the numerical inversion of the Laplace transform and similar Fredholm integral equations of the first kind. *J. Phys.* **A11**, 1729.

Miller, K. (1964). Three circle theorems in partial differential equations and applications to improperly posed problems. *Arch. Rot. Mech. Anal.* **16**, 126.

Miller, K. (1970). Least squares methods for ill-posed problems with a prescribed bound. *SIAM J. Math. Anal.* **1**, 52.

Morozov, V.A. (1966). On the solution of functional equations by the method of regularization. *Soviet Math. Dokl.* **7**, 414.

Morozov, V.A. (1984). "Methods for solving Incorrectly Posed Problems". Springer, Berlin.

Mueller, R.K., Kaveh, M. and Wade, G. (1979). Reconstructive tomography and applications to ultrasonic. *Proc. IEEE* **67** , 567.

Nashed, M.Z. (ed.) (1976 a). "Generalized Inverses and Applications". Academic Press, New York.

Nashed, M.Z. (1976 b). On moment-discretization and least-squares solutions of linear integral equations of the first kind. *J. Math. Anal. Appl.* **53**, 359.

Nashed, M.Z. and Wahba, G. (1974). Convergence rates of approximate least squares solutions of linear integral and operator equations of the first kind. *Math. Comp.* **28**, 69.

Natterer, F. (1980). A Sobolev space analysis of picture reconstruction. *SIAM J. Appl. Math.* **39**, 402.

Natterer, F. (1986 a). "The Mathematics of Computerized Tomography". Teubner, Stuttgart.

Natterer, F. (1986 b). Efficient evaluation of oversampled functions. *J. Comp. Appl. Math.* **14**, 303.

Natterer, F. (1986 c). Numerical treatment of ill-posed problems. In "Inverse Problems", ed. G.Talenti. Lect.Notes in Math., Vol. 1225. Springer, Berlin.

Oldenburg, D.W. (1976). Calculation of Fourier transforms by the Backus-Gilbert method. *Geophys. J. R. Astron. Soc.* **44**, 413.

Ostrowsky, N., Sornette, D., Parker, P. and Pike, E.R. (1981). Exponential sampling method for ligth scattering polydispersity analysis. *Optica Acta* **28**, 1059.

Payne, L.E. (1975). "Improperly Posed Problems in Partial Differential Equations". SIAM Regional Conf.Series in Applied Math. SIAM, Philadelphia.

Papoulis, A. (1956). A new method of inversion of the Laplace transform. *Q.Appl. Math.* **14**, 405.

Papoulis, A. (1975). A new algorithm in spectral analysis and band-limited extrapolation. *IEEE Trans. Circuits Syst.* **CAS-22**, 735.

Phillips, D.L. (1962). A technique for the numerical solution of certain integral equations of the first kind. *J. Asso. Comput. Mach.* **9**, 84.

Picard, E. (1910). Sur un theoreme general relatif aux equations integrales de premiere espece et sur quelques problemes de physique mathematique. *R. C. Mat. Palermo.* **29**, 615.

Pike, E.R., McWhirter, J.G., Bertero, M. and De Mol, C. (1984). Generalized information theory for inverse problems in signal processing. *IEE Proc.* **131**, 660.

Pucci, C. (1955). Sui problemi di Cauchy non ben posti. *Atti Acc. Naz. Lince,* **18**, 473.

Radon, J. (1917). Uber die bestimmung von funktionen durch ihre integralwerte längs gervisser mannigfaltigkeiten. *Berichte Sächsische Akademie der Wissenschaften, Leipzig, Math.-Phys. Kl.* **69**, 262.

Ralston, A. (1965). "A First Course in Numerical Analysis". Mc-Graw-Hill, New York.

Reinsch, C.H. (1967). Smoothing by spline functions. *Numer. Math.* **10**, 177.

Riesz, F. and Sz. Nagy, B. (1972). "Lecons d'Analyse Fonctionnelle". Gauthier-Villars, Paris.

Rust, B.W. and Burrus, W.R. (1972). "Mathematical Programming and the Numerical Solution of Linear Equations". American Elsevir Publishing Company, New York.

Sabatier, P.C., ed. (1978). "Applied Inverse Problems". Lecture Notes in Phys., Vol.85. Springer, Berlin.

Sabatier, P.C. (1984). Well-posed questions and exploration of the space of parameters in linear and non-linear inversion. In "Proc. Conf. on Inverse Problems of Acoustical and Elastic Waves". F.Santora, Y.H.Pao and W.W. Symes, eds. SIAM, Philadelfia.

Sabatier, P. C. ed. (1987). "Inverse Problems : An Interdisciplinary Study". Advances in Electronics and Electron Physics, Supplement 19. Academic Press, New York.

Sanz, J.L.C. and Huang, T.S. (1983). Unified Hilbert space approach to iterative least-squares linear signal restoration. *J. Opt. Soc. Am.* **73**, 1455.

Schafer, R.W., Mersereau, R.M. and Richards, M.A. (1981). Constrained iterative restoration algorithms. *Proc. IEEE* **69**, 432.

Schomburg, B. and Berendt, G. (1987). On the convergence of the Backus-Gilbert algorithm. *Inverse Problems* **3**, 341.

Shewell, J.R. and Wolf, E. (1968). Inverse diffraction and a new reciprocity theorem. *J. Opt. Soc. Am.* **58**, 1596.

Shifrin, K.S. and Perelman, A.Y. (1965). Inversion of light scattering data for the determination of spherical particle spectrum. In "Proceedings, Second Interdisciplinary Conference on Electromagnetic Scattering, Massachusetts, June 1965". R.L. Rowell and R.S. Stein, eds. Gordon and Breach, New York.

Slepian, D. (1964). Prolate spheroidal wave functions, Fourier analysis and uncertainty - IV: Extensions to many dymensions; generalized prolate spheroidal functions. *Bell Syst. Tech. J.* **43**, 3009.

Slepian, D. (1978). Prolate spheroidal wave functions, Fourier analysis and uncertainty - V: the discrete case. *Bell Syst. Tech. J.* **57**, 1371.

Slepian, D. and Pollak, H.O. (1961). Prolate spheroidal wave functions, Fourier analysis and uncertainty - I. *Bell System Tech. J.* **40**, 43.

Sondhi, M.M. (1969). Reconstruction of objects from their sound-diffraction patterns. *J. Acoust. Soc. Am.* **46**, 1158.

Strand, O.N. (1974). Theory and methods related to the singular-function expansion and Landweber's iteration for integral equations of the first kind. *SIAM J. Numer. Anal.* **11**, 798.

Strand, O.N. and Westwater, E.R. (1968). Minimum-RMS estimation of the numerical solution of a Fredholm integral equation of the firstr kind. *SIAM J. Numer. Anal.* **5**, 287.

Talenti, G., ed. (1986). "Inverse Problems". Lecture Notes in Math., Vol.1225. Springer, Berlin.

Talenti, G., (1987). Recovering a function from a finite number of moments. *Inverse Problems* **3**, 501.

Tikhonov, A.N. (1963a). Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.* **4**, 1035.

Tikhonov, A.N. (1963b). Regularization of incorrectly posed problems. *Soviet Math. Dokl.* **4**, 1624.

Tikhonov, A.N. (1964). Solution of nonlinear integral equations of the first kind. *Soviet Math. Dokl.* **5**, 835.

Tikhonov, A.N. and Arsenin, V.Y. (1977). "Solutions of Ill-Posed Problems". Winston/Wiley, Washington.

Titchmarsh, E.C. (1948). "Introduction to the Theory of Fourier Integrals". Clarendon Press, Oxford.

Titchmarsh, E.C. (1958). "Eigenfunction Expansions Associated with Second Order Differential Equations". Vol II. Clarendon, Oxford.

Toraldo di Francia, G. (1969 a). Degrees of freedom of an image. *J. Opt. Soc. Am.* **59**, 799.

Toraldo di Francia, G. (1969 b). Some recent progress in classical optics. *Riv. Nuovo Cimento* **1**, 460.

Torre, V. and Poggio, T. (1986). On edge detection. *IEEE Trans. Pattern Anal. Mach. Intelligence* **PAMI-8**, 147.

Turchin, V.F., Kozlov, V.P. and Malkevich, M.S. (1971). The use of mathematical statistics methods in the solution of incorrectly posed problems. *Soviet Phys. Usp.* **13**, 681.

Twomey, S. (1965). The application of numerical filtering to the solution of integral equations encountered in indirect sensing measurements. *J. Franklin Inst.* **279**, 95.

Twomey, S. (1974). Information content in remote sensing. *Appl. Opt.* **13**, 942.

Vainikko, G.M. (1982). The discrepancy principle for a class of regularization methods. *USSR Comp. Math. Math. Phys.* **22** (No.3), 1.

Van de Hulst, H.C. (1981). "Light Scattering by Small Particles". Dover, New York.

Viano, G.A. (1976). On the extrapolation of optical image data. *J. Math. Phys.* **17**,1160.

Wahba, G. (1977). Practical approximate solutions to linear operator equations when the data are noisy. *SIAM J. Numer. Anal.* **14**, 651.

Wahba, G. and Wendelberger, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross validation. *Monthly Weather Review* **108**, 1122.

Wahba, G. and Wold, S. (1975 a). A completely automatic French curve: Fitting spline functions by cross-validation. *Comm. Stat.* **4**, 1.

Wahba, G. and Wold, S. (1975 b). Periodic splines for spectral density estimation: The use of cross-validation for determining the degree of smoothing, *Comm. Stat.* **4**, 125.

Widder, D.V. (1946). "The Laplace Transform". Princeton Univ. Press.

Wolf, E. (1969). Three-dimensional structure determination of semi-transparent objects from holographic data. *Opt. Commun.* **1**, 153.

Yosida, K. (1966). "Functional Analysis". Springer, Berlin.

Zuev, V.E. and Naats, I.E. (1983). "Inverse Problems of Lidar Sensing of the Atmosphere". Optical Sciences, vol.29. Springer, Berlin.