NUMERICAL METHODS FOR ILL-POSED, LINEAR PROBLEMS

Thesis by

Thomas Stevens

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1975

(Submitted August 9, 1974)

ACKNOWLEDGEMENTS

I am deeply indebted to Professor Joel N. Franklin, a very thoughtful and patient research advisor. When progress faltered, a visit with him always resulted in a better sense of direction and boosted morale. No advisor could have better filled my needs.

This topic was far removed from the endeavours of any other student but there were always interested kindred spirits with whom my work could be meaningfully discussed. Thanks go to my fellow graduate students in whom there was no lack of academic flexibility.

Elizabeth Fox, (card-sharp and secretary extraordinaire), has gamely wrestled for more than a month with this tedious and difficult manuscript; to her my special thanks. I have suggested she write a paper "On the Psychology of What the Student Really Means by . . . . .".

Financial support during my four years at Caltech was provided by CIT graduate teaching assistantships.

This has been much more than an academic experience. My fond memories include all those within and outside the Caltech community who have made my stay here what it has been. As great as any academic gifts bestowed upon me by this university is the privilege of being able to number many fine people among my friends.

ABSTRACT

A means of assessing the effectiveness of methods used in the numerical solution of various linear ill-posed problems is outlined. Two methods: Tikhonov's method of regularization and the quasireversibility method of Lattès and Lions are appraised from this point of view.

In the former method, Tikhonov provides a useful means for incorporating a constraint into numerical algorithms. The analysis suggests that the approach can be generalized to embody constraints other than those employed by Tikhonov. This is effected and the general "T-method" is the result.

A T-method is used on an extended version of the backwards heat equation with spatially variable coefficients. Numerical computations based upon it are performed.

The statistical method developed by Franklin is shown to have an interpretation as a T-method. This interpretation, although somewhat loose, does explain some empirical convergence properties which are difficult to pin down via a purely statistical argument.

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont'd)

TABLE OF CONTENTS (Cont'd)

INTRODUCTION

A problem is well posed in the sense of Hadamard if it satisfies the criteria of existence, uniqueness and stability of solution; that is, if it has a unique solution depending continuously upon the data.* If a problem lacks any one of these solution properties, it is said to be ill posed. Three categories of ill-posedness are immediately suggested. One encounters the modifiers "overdetermined," "underdetermined" and "unstable" in description of the cases of nonexistence, nonuniqueness and instability of solutions respectively.

A hasty appraisal of what is implied by ill-posedness might give one the impression that such problems are to be given a wide berth because of the difficulties inherent in their solution. In point of fact, virtually all scientific investigators will encounter ill-posed problems in their work and in most instances will come to terms with them rather easily. Consider the following hypothetical situations.

An experimenter, plotting his data on a graph, expects his points to lie on a straight line whose slope is of interest. What he, in fact, discovers is that because of errors in his measurements, the points are not quite colinear. Undismayed, he draws a line passing through some and quite close to the others.

A draftsman is asked to pass a smooth curve through a collection of isolated points. He does not vociferously protest that any number of such curves might be drawn. With the aid of a French curve, he simply makes an eminently reasonable choice of one.

----

*In any application, this statement would require further clarification.

A freshman differentiates cos x. He does not trouble himself with the fact that a very small perturbation of $\varepsilon \sin \frac{1}{\varepsilon^2} x$ to cos x will produce a huge perturbation of $\frac{1}{\varepsilon} \cos(\frac{1}{\varepsilon^2} x)$ in his answer.

In the three situations, problems arose which were (in order of appearance) overdetermined, underdetermined and unstable. In the first two cases, simple mental readjustments of intent were made and the ill-posedness was effectively exorcized. In the third case, data were exact and a means was available for exact solution in the abstract. Under these circumstances, instability will never be given the opportunity to make itself felt. However, where computer solution is envisioned, such instability will always be relevant. Inasmuch as numerical roundoff will always be a source of error, data can never be regarded as exactly given, even should no other source of error be present.

In coping with ill-posedness, we recognize that asking for a solution to the original problem per se is very naive. Either a solution does not exist, there is more than one, or a small but inescapable source of error in the data could lead us to arbitrarily erroneous answers. There is still hope, however, that useful information may be obtained via the solution of a related well-posed problem: a well-posed extension.

The ideas involved in such extension need not be profound. If our experimenter performed a least squares fit, his extended problem would have been to find that line coming as close as possible (in a well-defined mathematical sense) to all the points. But he may prefer to simply "eye-ball" the points and then to draw his line. In that instance, precise mathematical description of the extension would be quite impossible. He is, in effect, drawing that line which in passing close to the points optimizes his peace of mind. That will likely apply to the draftsman's curve, the only difference being that the choice is made from the class of solutions to the original problem rather than from a class of "near misses." This kind of vague thinking was perfectly adequate in the modest applications considered. But as the problems get harder, the thrust of results obtained by unsupported intuition becomes increasingly nebulous. The need for a sounder basis of operations becomes more keenly felt.

This work deals with linear inverse problems in which instability is the essential source of ill-posedness. Numerical methods for inverting equations of the form $Kf_0 = g_0$ (where K is a linear operator mapping from one Hilbert space into another) are discussed. In those examined, existence and uniqueness will be assumed so that were g_0 known exactly, a unique f_0 would exist satisfying the equation. However, g_0 will be assumed known only approximately. K will invariably be such that some information crucial to the description of f_0 can not be found from our approximate knowledge of g_0. Furthermore, the class of admissible f (consistent with our knowledge) includes pairs f_1 and f_2 vastly different from one another. The effect of K is to reduce their difference $f_1 - f_2$ so much that $K(f_1 - f_2)$ is hidden in the small uncertainty associated with g_0. A well-posed extension must provide us with a means of providing the information missing. Directly or indirectly, assumptions will be made about the solution f_0. The class of admissible f's will be made much smaller than that satisfying just $Kf \approx g$. In analyzing a method geared to numerical solution, a first step is to identify those additional assumptions. Ultimately, we will wish to know how accurately, in light of this additional information, a solution computed using the method approximates f_0.

In Chapter 1, an overview of the problems to be considered and the kinds of measures to be taken in their well-posed extension is provided. In Chapter 2 Tikhonov's method of regularization is introduced and its application to a few specific problems considered. This motivates the subsequent discussion in Chapter 4 of what is believed to be a new numerical approach to the backwards heat equation with spatially variable coefficients. This, in turn, motivates a generalization of Tikhonov's method to utilize solution set constraints other than those suggested by Tikhonov. This theory is used in Chapter 6 to explain certain convergence phenomena encountered in the apparently unrelated statistical method developed in Franklin [6]. Chapter 3 stands somewhat apart from the others. It is a critique of a class of techniques known collectively as quasi-reversibility methods illustrated with two which are perhaps the most famous.

CHAPTER 1

GENERAL DESCRIPTION VIA FUNCTIONAL ANALYSIS

The aim of this chapter is to provide an overview of unstable problems which will motivate subsequent treatment. A few basic results from functional analysis afford considerable insight into the instability phenomenon and the kinds of measures one takes in attempting to denature it.

The machinery needed for this discussion is developed in the references: Riesz-Sz.-Nagy [21], Simmons [23] and Taylor [27]. Although these will be freely quoted, the scope of their theorem statements will often be contracted. For clarity of focus, generality transcending our requirements will be sacrificed. That may cause some of the facts stated to appear rather weak to anyone acquainted with the full power of their sources.

For completeness, all relevant concepts standard to functional analysis will be defined. However, to avoid unwieldiness in this chapter, such definition will be relegated to Appendix A. Any underlined word will be appearing for the first time (in the main text) and will be defined in Appendix A.

1.1 Description

Let $B_1 \; \|\cdot\|_1$ and $B_2 \; \|\cdot\|_2$ be Banach spaces and K a linear operator $(K : B_1 \rightarrow B_2)$. Consider the problem of finding f in B_1 such that Kf = g for some g in the range of K (Ran K). Our attention will be restricted to those K's with the following properties:

a) K is one to one (injective)

b) K is bounded (continuous)

c) K's inverse (K^{-1} : Ran K $\rightarrow B_1$) is unbounded (1.1.1)

d) The range of K is dense in B_2.

1.2 Restriction of Data

The range of an operator K having the properties (1.1.1) can not be all of B_2. (K can not be an onto (surjective) mapping). There will be g in B_2 for which no f exists satisfying Kf = g. We get this from the following theorem:

Theorem 1.2.1

Let $B_1 \ \| \cdot \|_1$ and $B_2 \ \| \cdot \|_2$ be Banach spaces and K a bounded, one to one linear operator whose domain is B_1 and whose range is all of B_2. Then K^{-1} exists and is bounded.

Proof. Taylor [27] page 180; Simmons [23] page 236.

This considered, one can not admit properties 1.1.1 a) through c) and ontoness. The range of K must be a dense but proper subset of B_2.

1.3 Compactness

Many of the K's of interest to us will be compact operators. This is a stronger property than boundedness. Such K's map weakly convergent sequences in B_1 onto (strongly) convergent sequences in B_2. If the Banach space B_1 is norm-reflexive, that mapping of sequences characteristic can be used as an alternative means of defining compact operator (to that given in Appendix A). (Taylor page 287).

1.4 Well-posed Extension

Suppose there is some g_0 in the range of K whose inverse image f_0 is desired. Suppose too that our knowledge of g_0 is somewhat limited and, for one reason or another, we actually have g as data, knowing only that $\|g-g_0\|_2 \leqslant \varepsilon$ for some small number ε.* Since g may not even be in the range of K, "solving" Kf = g is immediately a dubious idea. By 1.1.1 d) our neighbourhood of g will contain an infinite class of points in the range of K. If all of these points have an equal claim on being identified as g_0, we will be in a most unhappy situation because the inverse image under K of those points will be an unbounded set in B_1. This unboundedness of K^{-1} is, of course, the central issue in all this work. It is a fact of life that in the absence of additional information, we are hopelessly stuck. Approximate knowledge of g_0 is not sufficient to yield for us approximate knowledge of f_0.

Suppose we can add the additional requirement that f_0 lie in some admissibility set A contained in B_1. The only elements in $N_\varepsilon(g)$ which will be considered "candidates for being g_0" will have inverse images in A. This procedure will define a well-posed extension if in view of the restriction to A, our knowledge of f_0 becomes better and better as our knowledge of g_0 improves. More precisely:

----

*The set of elements h in B_2 satisfying $\|g-h\|_2 \leqslant \varepsilon$ will be denoted $N_\varepsilon(g)$.

Definition 1.4.1

$$\text{Denote by } \eta_A(g,\varepsilon) = \sup_{\substack{f_1, f_2 \in A \\ Kf_1, Kf_2 \in N\varepsilon(g)}} \|f_1 - f_2\|_1 .$$

An extension by this restriction to A will be said to be well-posed (convergent) if $\eta_A(g,\varepsilon) \to 0$ as $\varepsilon \to 0$.

The following theorems will give us an idea of the kinds of extensions likely to be helpful.

Theorem 1.4.2

Let $K : B_1 \to B_2$ be linear, one to one and bounded and let A be a compact set in B_1. Then the restriction of K to A has a continuous inverse.

Proof:

Let $\{f_n\}$ be a sequence in A whose image under K $\{Kf_n\} \subset B_2$ converges to a limit g. So $Kf_n \to g$. Since A is compact, $\{f_n\}$ has a convergent subsequence $\{f_{n_p}\}$ whose limit shall be denoted by f. Suppose the entire sequence does not converge to f. Then there is a subsequence bounded away from f from which we can extract a further subsequence $\{f_{m_k}\}$ converging to a limit h necessarily different from f. But K is bounded so $\lim_{\ell \to \infty} f_{m_\ell} = h \Longrightarrow \lim_{k \to \infty} Kf_{m_\ell} = Kh$ and $\lim_{p \to \infty} f_{n_p} = f \Longrightarrow \lim_{p \to \infty} Kf_{n_p} = Kf$. Since $\{Kf_{m_\ell}\}$ and $\{Kf_{n_p}\}$ are both subsequences of $\{Kf_n\}$ whose limit is g, $Kh = g = Kf$ and by the one to oneness of K, $h = f$. This is a contradiction so the entire sequence f_n had to be convergent. □

Theorem 1.4.3

Let $K : B_1 \to B_2$ be linear, one to one and compact. Let A be a closed, bounded set in B_1 having the property that any sequence in A whose image under K is convergent is itself convergent. Then A is a compact set.

Proof:

Let $\{f_n\}$ be a sequence in A. The image of A under K has compact closure so $\{Kf_n\}$ has a convergent subsequence $\{Kf_{n_k}\}$. By hypothesis this would make $\{f_{n_k}\}$ convergent and hence every sequence in A has a convergent subsequence. So A is compact. □

We see then that restriction to a compact set will lead to a well-posed extension for all bounded operators K of interest to us and that if K is compact, our restriction must be to a compact set if we are to produce a convergent extension.

1.5 Spectral Theory

In many instances, B_1 and B_2 will be one and the same. K will be an injection from a Banach space B into itself. In the problems to be examined later, B will, in fact, be a Hilbert space (usually L^2).

Spectral decomposition of operators is often useful in analysis of methods and in ultimate numerical solution of the extended problem.

Definition 1.5.1

The spectrum of a bounded linear operator $K : B \rightarrow B$ (where $B \; \| \cdot \|$ is a Banach space) is the set of (complex) scalars λ for which $K-\lambda I$ does not have a bounded inverse. The spectrum is denoted $\sigma(K)$ and its complement (the resolvent) by $\rho(K)$.

----

*Here I refers to the identity mapping $Ix \equiv x$.

The spectrum is subdivided into three subsets.

Definition 1.5.2

a) The point spectrum $\sigma_p(K)$ is the set of λ for which $K-\lambda I$ fails to be one to one (has a non-trivial null space).

b) The continuous spectrum $\sigma_c(K)$ is that subset of $\sigma(K)$ for which $K-\lambda I$ is one to one and the range of $K-\lambda I$ is dense in B.

c) The residual spectrum consists of those λ in $\sigma(K)$ for which $K-\lambda I$ is one to one and the range of $K-\lambda I$ is not dense in B.

It is at once apparent that all operators K of interest here have 0 in their continuous spectra. We shall not encounter K having (non-empty) residual spectra. A few useful results about the spectra of particular classes of K (in our domain of interest) will now be listed. Their proofs are scattered through the references and will not be given here.

Theorem 1.5.3

The spectrum of a bounded operator K is contained in a closed circle of radius $r_\sigma(K) = \lim_{n \to \infty} \|K^n\|^{1/n}$.

When B is a Hilbert space H (,), it makes sense to talk about normal operators $KK^* = K^*K$ and self-adjoint operators $K = K^*$.

Theorem 1.5.4

If $K : H \to H$ is bounded and normal, its residual spectrum is empty.

For compact operators, we get the following results.

Theorem 1.5.6

The spectrum of a compact operator is at most countable with 0 as the only possible accumulation point.

Theorem 1.5.7

If H is a separable Hilbert space and $K : H \rightarrow H$ is a compact operator, the null space of $K-\lambda I$ is finite dimensional for λ in $\sigma_p(K)$.

This leads to the spectral decomposition theorem.

Theorem 1.5.8

Let H (,) be a separable Hilbert space and $K : H \rightarrow H$ be a compact, self-adjoint, linear operator. Let the point spectrum of K be denoted by the sequence $\{\lambda_n\}$ ($\{|\lambda_n|\}$ non-increasing). Let $\{\varphi_n\}$ be orthonormal eigenvectors, φ_n corresponding to λ_n.* Then K has the decomposition defined by $Kx = \sum_{n=1}^{\infty} \lambda_n(x, \varphi_n)\varphi_n$ (x arbitrary in H).

1.6 An Example

To illustrate the ideas of this chapter and how they tie in with a practical problem, a famous example (the backwards heat equation) "solved" by the technique generally referred to as "Simple Spectral Cut-off" will be considered. The point of view outlined in the previous sections will be the one adopted. "Spectral Cut-off" will be appraised as an extension method of the kind described.

The "forwards heat problem" to be inverted will be the simplest one imaginable. If we were required to solve

$$\begin{aligned} &u_t = u_{xx} && \text{for } t > 0 \quad x \in (0,\pi) \\ &u(0,t) = u(\pi,t) = 0 && \qquad\qquad (1.6.1) \\ &u(x,0) = f(x) && f \in L^2[0,\pi] \end{aligned}$$

----

*Possible multiplicity of eigenvalues is accounted for by allowing the same eigenvalue more than one index in that eventuality, i.e., $\lambda_n = \lambda_{n+1}$ is possible.

for some $t = T > 0$, we would find, by separation of variables,

$$u(x, T) = \sum_{n=1}^{\infty} e^{-n^2 T} (f, \sqrt{\tfrac{2}{\pi}} \sin nx) \sqrt{\tfrac{2}{\pi}} \sin nx . \qquad (1.6.2)$$

Here the inner product (,) used is defined by:

$$(f_1, f_2) = \int_0^{\pi} f_1(x)\, f_2(x)\, dx \quad \text{for } f_1, f_2 \quad L^2[0, \pi] . \qquad (1.6.3)$$

The inverse problem would be: Given $u(x, T) = g(x)$,

$$(1.6.4)$$

find $f(x)$ such that $g(x) = \sum_{n=1}^{\infty} e^{-n^2 T} (f, \sqrt{\tfrac{2}{\pi}} \sin nx) \sqrt{\tfrac{2}{\pi}} \sin nx$.

Identify: $B_1 = B_2 = L^2[0, \pi] = H$, a separable Hilbert space. K is the (forwards) heat operator.

Kf is defined by 1.6.2 $Kf(x) = \sum_{n=1}^{\infty} e^{-n^2 T} (f, \sqrt{\tfrac{2}{\pi}} \sin nx) \sqrt{\tfrac{2}{\pi}} \sin nx$. Its eigenvalues are $e^{-n^2 T}$ with corresponding orthonormal eigenvectors $\sqrt{\tfrac{2}{\pi}} \sin nx$. We recognize the spectral decomposition form of Theorem 1.5.8 and indeed, K is a compact, linear, self-adjoint operator.

If g_0 is exactly given, solving this problem is easy. We could simply recover the Fourier sine series for f_0 by:

$$\begin{aligned}(g_0, \sqrt{\tfrac{2}{\pi}} \sin mx) &= \sum_{n=1}^{\infty} e^{-n^2 T} (f_0, \sqrt{\tfrac{2}{\pi}} \sin nx)\, \delta_n^m \\ &= e^{-m^2 T} (f_0, \sqrt{\tfrac{2}{\pi}} \sin mx) \qquad (1.6.5)\end{aligned}$$

$$\Longrightarrow (f_0, \sqrt{\tfrac{2}{\pi}} \sin mx) = e^{m^2 T} (g_0, \sqrt{\tfrac{2}{\pi}} \sin mx) . \qquad (1.6.6)$$

whence $f_0(x) = \sum_{m=1}^{\infty} e^{m^2 T} (g_0, \sqrt{\tfrac{2}{\pi}} \sin mx) \sqrt{\tfrac{2}{\pi}} \sin mx$. (1.6.7)

If g_0 is perturbed by $\varepsilon \sqrt{\frac{2}{\pi}} \sin px$ (p integer; norm of perturbation ε), f_0 would be altered by $e^{p^2T} \varepsilon \sqrt{\frac{2}{\pi}} \sin px$ (norm $e^{p^2T}\varepsilon$). By choosing p large enough, $e^{p^2T}\varepsilon$ can be made arbitrarily large.

If g_0 is perturbed by $\sum_{n=1}^{\infty} \varepsilon \sqrt{\frac{6}{\pi^2}} \frac{1}{n} \sqrt{\frac{2}{\pi}} \sin nx$, (of norm ε), f_0 would be altered by $\sum_{n=1}^{\infty} \varepsilon\sqrt{\frac{6}{\pi^2}} \frac{e^{n^2T}}{n} \sqrt{\frac{2}{\pi}} \sin nx$ which does not represent an $L^2[0,\pi]$ function. ($\sum_{n=1}^{\infty} \varepsilon^2 \cdot \sqrt{\frac{6}{\pi^2}} \frac{e^{2n^2T}}{n^2}$ is a rapidly divergent series.) This is a manifestation of the data restriction mentioned in section 1.2. The range of K can not be all of $L^2[0,\pi]$ and 1.6.7 can only be applied sensibly to g in the range of K. The range of K will be dense in $L^2[0,\pi]$. For any g in $L^2[0,\pi]$, choosing N large enough enables us to approximate g arbitrarily accurately by $\sum_{m=1}^{N} (g, \sqrt{\frac{2}{\pi}} \sin mx) \sqrt{\frac{2}{\pi}} \sin mx$ which is the image under K of $\sum_{m=1}^{N} e^{m^2T} (g, \sqrt{\frac{2}{\pi}} \sin mx) \sqrt{\frac{2}{\pi}} \sin mx$.

Since the ill-posedness arose from high frequency perturbations, simple spectral cut-off rules these out by truncating 1.6.7 after a finite number of terms:

$$f_N(x) = \sum_{m=1}^{N} e^{m^2T} (g, \sqrt{\frac{2}{\pi}} \sin mx) \sqrt{\frac{2}{\pi}} \sin mx . \qquad (1.6.8)$$

The only f's this method considers as possible solutions are those lying in the N dimensional subspace of $L^2[0,\pi]$ spanned by $\sqrt{\frac{2}{\pi}} \sin mx \quad m = 1, \ldots\ldots, N$. The image under 1.6.8 of the unit ball about a data point g will be a hyperellipsoid within that subspace.

The closure of the hyperellipsoid will be compact and can be identified as the admissibility set A of section 1.4.

Assuming f_0 is in A and is mapped to some point g_0 within ε of our data g, applying 1.6.8 to g gives an estimate of f_0. Having made these assumptions, our estimate can be out by at most $\eta_A(g,\varepsilon) \equiv \sup\limits_{\substack{f_1, f_2 \,\varepsilon\, A \\ Kf_1 \; Kf_2 \epsilon N\varepsilon(g)}} \|f_1 - f_2\|$ which, by inspection, is $2\varepsilon e^{N^2 T}$. Simple spectral cut-off is thus a convergent extension.

To be sure, there are more sophisticated spectral cut-off methods than this; (hence the modifier "simple"). It may seem rather crass to make a restriction of this sort except as an approximating endeavour. However, this was only intended as an illustrative example of an extension method.

1.7 Additional Remarks

The point of view of well-posed extension methods the preceding sections may lead one to adopt is a trifle too narrow. It is important to realize the limitations of such analysis.

First of all, not all good methods are convergent. Suppose in solving a related problem to $Kf_0 = g_0$, we could demonstrate an error estimate of the form: $\eta + \rho(\varepsilon)$ (where ε is the error associated with our knowledge of g_0, $\rho(\varepsilon) \rightarrow 0$ and η is some small bias inherent in the method and independent of ε). Depending on the size of η and the speed with which $\rho(\varepsilon) \rightarrow 0$ with ε, we may be quite satisfied with such a method. After all, in practical problems the ε's encountered are non-zero, the whole point being that the data is not perfect. Having to live with the fact that $\eta + \rho(0) = \eta$ is non-zero need not be any great hardship. An example of such a method for the problem of section 1.6 would be to place a non-zero bound on the "tail" of f_0's Fourier sign series, i.e., $\sum_{n=N+1}^{\infty} (f_0, \sqrt{\frac{2}{\pi}} \sin nx) \sqrt{\frac{2}{\pi}} \sin nx$ has norm $\leqslant \eta$. Then our error estimate for the solution by spectral cut-off would be $\eta + 2\varepsilon\, e^{N^2 T}$. If a functional dependence of η on N were assumed, a minimization of $\eta(N) + 2\varepsilon\, e^{N^2 T}$ would indicate where we should cut off the series.

The previous sections will not, however, be irrelevant to the appraisal of these biased methods. The assumptions made about f_0 are not quite commensurate with our genuine knowledge of f_0 but do lead to a convergent extension to which all the foregoing applies. Having done so, we will be left with assessing the bias in producing our two-part error analysis.

Our knowledge of g_0 may be statistical in nature in which case our estimate made for f_0 should be statistical too. If all we know about g_0 is that our actual data g is a sample of a Gaussian random variable with mean g_0 and variance σ, it will not be possible to put a deterministic error bound on whatever estimate of f_0 is made. Different yardsticks will be required to measure the success of a statistical method from those described thus far.

The relevance of the value judgements to be made on the extension methods to be considered in later chapters will remain slightly subjective. That being understood, such assessment will be made without further ado.

CHAPTER 2

SOME OBSERVATIONS ON TIKHONOV'S METHOD OF REGULARIZATION

2.1 Overview of the Method

Consider the Fredholm integral equation of the first kind:

$$\int_0^b K(x, y)\, f_0\,(y)\, dy = g_0(x) \qquad (2.1.1)$$

where f_0 is in $B_1 \| \cdot \|_1$ (a Banach subspace of $L^2[a, b]$); g_0 is in $L^2[c, d]$; $K(x, y)$ is an L^2 kernel (is in $L^2\{[c, d] \times [a, b]\}$).

Equation 2.1.1 can also be written in operator notation:

$$Kf_0 = g_0 \;\; . \qquad (2.1.2)$$

K so defined is a compact mapping from B_1 to $L^2[c, d]$ (proof in Taylor [27] page 277) and will be self-adjoint if the kernel is symmetric ($K(x, y) = K(y, x)$ a. e.). We assume that $Kf_0 = 0$ has only the trivial solution in B_1.

In Tikhonov [28], a convergent extension method was introduced via the "regularizing assumption" that f_0 belongs to a class of functions $f \in B_1$ satisfying:

$$\Omega^2(f) \equiv \int_a^b \{p(x)\,[f'(x)]^2 + q(x) f^2(x)\}\, dx \leqslant \omega_1^2 \qquad (2.1.3)$$

for some real number ω_1. (Here $p(x)$ and $q(x)$ are positive and continuous on $[a, b]$.)

The very definition of B_1 is incomplete until the norm $\| \cdot \|_1$ is specified. A convergent extension will be achieved via this restriction whenever 2.1.3 defines a compact set in the *norm topology* of B_1 $\| \cdot \|_1$.

Definition 2.1.4

The assumption 2.1.3 will then be said to regularize K under the norm $\| \cdot \|_1$.

If we take B_1 to be the absolutely continuous functions with square integrable derivatives on $[a, b]$ and associate the norm

$$\|f\|_1 = \max_{x \in [a,b]} |f(x)| \ , \tag{2.1.5}$$

the family defined by 2.1.3 is equicontinuous; uniformly bounded and hence compact by Ascoli's theorem.

In Franklin [8], attention was paid to the effectiveness of the regularizing assumption on various operators K. The notion of a rate of convergence was introduced and calculated for a few examples. Following the notation of that paper, the norm $\| \cdot \|_1$ will be denoted $\mu(\cdot)$. The effect of the norm μ on convergence will be of paramount importance here.

2.2 Applying the Method

It is supposed that K(x, y) is known and a function g given satisfying $\|g-g_0\|^* \leqslant \varepsilon$ for some $\varepsilon > 0$. Then a number α is chosen related to ε by the inequalities:

$$C_1 \varepsilon^2 \leqslant \alpha \leqslant C_2 \varepsilon^2 \tag{2.2.1}$$

for two positive constants C_1 and C_2. The function f which minimizes

----

*Unless otherwise specified $\|\cdot\|$ will denote the L^2 norm $\|g\|^2 = \int_a^b g^2(x)\, dx$.

the quantity

$$\|Kf-g\|^2 + \alpha\,\Omega^2(f) \qquad (2.2.2)$$

is taken as the approximation to f_0. Using the fact that Ω and $\|\cdot\|$ define norms obeying the parallelogram law, Franklin demonstrated the uniqueness of the minimizing function f.

Since f minimizes 2.2.2, f_0 must satisfy:

$$\|Kf-g\|^2 + \alpha\Omega^2(f) \leqslant \|Kf_0-g\|^2 + \alpha\Omega^2(f_0) \ . \qquad (2.2.3)$$

Assuming $\Omega^2(f_0) \leqslant \omega^2$ and $Kf_0 = g_0$, this implies:

$$\|Kf-g\|^2 + \alpha\Omega^2(f) \leqslant \|g_0-g\|^2 + \alpha\omega^2 \leqslant \varepsilon^2 + \alpha\omega^2 \ .$$

From this we obtain:

$$\|Kf-g\|^2 \leqslant \varepsilon^2 + \alpha\omega^2 \leqslant \varepsilon^2\,(1 + C_2\omega^2) \ , \qquad (2.2.4)$$

and

$$\alpha\Omega^2(f) \leqslant \varepsilon^2 + \alpha\omega^2 \leqslant \alpha(\frac{1}{C_1} + \omega^2) \ . \qquad (2.2.5)$$

From 2.2.4,

$$\|Kf-g_0\| \leqslant \|Kf-g\| + \|g-g_0\| \leqslant \varepsilon(1+C_2\omega^2)^{\frac{1}{2}} + \varepsilon \ . \qquad (2.2.6)$$

And from 2.2.5, $\Omega^2(f) \leqslant \frac{1}{C_1} + \omega^2$ (independent of ε) . (2.2.7)

If 2.1.3 defines a compact set in the norm topology of B_1 $\mu(\cdot)$, 2.2.6 and 2.2.7 enable us to apply theorem 1.4.2 to conclude $\mu(f-f_0) \to 0$ as $\|g-g_0\| \to 0$.

In Chapter 1, rapidity of convergence was measured by

$$\eta_A(g,\varepsilon) = \sup_{\substack{f_1,f_2 \in A \\ Kf_1, Kf_2 \in N_\varepsilon(g)}} \mu(f_1 - f_2)$$

where A is the compact set in which f_0 is assumed to lie. Here $A = A(\omega_1) \equiv \{f \in B_1 \mid \Omega^2(f) \leqslant \frac{1}{C_1} + \omega^2\}$.

The f minimizing 2.2.2 lies in $A(\omega_1)$ as does f_0.

For regularization, Franklin [8] introduced the following measures of convergence.

Definition 2.2.8

The modulus of regularization $\rho_\mu(\varepsilon)$ is given by:

$$\rho_\mu(\varepsilon) \equiv \sup_{\substack{\|Kf\| \leqslant \varepsilon \\ \Omega^2(f) \leqslant 1}} \mu(f) \ .$$

Tikhonov's method provides a mapping $T_\alpha : L^2[c,d] \rightarrow B_1$.

Definition 2.2.9

The modulus of convergence $\sigma_\mu(\varepsilon, \alpha)$ is defined by

$$\sigma_\mu(\varepsilon, \alpha) \equiv \sup_{\substack{\|g - Kf_0\| \leqslant \varepsilon \\ \Omega^2(f_0) \leqslant 1}} \mu(T_\alpha g - f_0) \ .$$

If the bound on $\Omega^2(f_0)$ is ω^2 (not necessarily 1), we notice that

$$\sup_{\substack{\|g - Kf_0\| \leqslant \varepsilon \\ \Omega^2(f_0) \leqslant \omega^2}} \mu(T_\alpha g - f_0) = \omega \sup_{\substack{\|\frac{g}{\omega} - K\frac{f_0}{\omega}\| \leqslant \frac{\varepsilon}{\omega} \\ \Omega^2(\frac{f_0}{\omega}) \leqslant 1}} \mu(T_\alpha \frac{g}{\omega} - \frac{f_0}{\omega}) = \omega\, \sigma_\mu(\tfrac{\varepsilon}{\omega}, \alpha).$$

Definition 2.2.10

The rate of convergence is given by $\omega\sigma_{\mu}(\frac{\varepsilon}{\omega}, \alpha)$.

Franklin proved that the modulus of regularization and rate of convergence are related by the inequalities:

$$\omega\rho(\varepsilon/\omega) \leqslant \omega\sigma(\varepsilon/\omega, \alpha) \leqslant \omega'\rho(\varepsilon'/\omega') \tag{2.2.11}$$

where $\varepsilon' \equiv \left(1 + \sqrt{1 + \frac{\alpha\omega^2}{\varepsilon^2}}\right)\varepsilon \leqslant \left(1 + \sqrt{1 + C_2\omega^2}\right)\varepsilon$ (2.2.12)

and $\omega' \equiv \left(1 + \sqrt{1 + \frac{\varepsilon^2}{\alpha\omega^2}}\right)\omega \leqslant \left(1 + \sqrt{1 + \frac{1}{C_1\omega^2}}\right)\omega$. (2.2.13)

All we wish to know about the rate of convergence is known once we have found the modulus of regularization. In practice, asymptotic estimates of $\rho_{\mu}(\varepsilon)$ valid as $\varepsilon \to 0$ are what we try to obtain. We seek a function $h(\varepsilon)$ (with which we have some familiarity as $\varepsilon \to 0$) for which we can exhibit constants r_1 and r_2 satisfying

$$r_1 h(\varepsilon) \leqslant \rho_{\mu}(\varepsilon) \leqslant r_2\, h(\varepsilon) \tag{2.2.14}$$

for sufficiently small ε. The function $h(\varepsilon)$ typically proves to be a power (ε^x for some $x < 1$) or logarithmic : $(-\log\varepsilon)^{-y}$ for some $y > 0$.

Since $p(x)$ and $q(x)$ are positive and continuous on $[a,b]$, they take on maximum values (denoted by P and Q respectively) and minimum values (denoted p and q). Clearly then, $\min(p;q)\left\{\int_a^b [f'^2(x) + f^2(x)]\,dx\right\} \leqslant \Omega^2(f_0) \leqslant \max(P,Q)\left\{\int_a^b [f'^2(x)+f^2(x)]\,dx\right\}$. So choosing the functions $p(x)$ and $q(x)$ to be other than identically 1 gives a value for $\Omega^2(f_0)$ which can be related to that obtained with $p(x) \equiv q(x) \equiv 1$ by scale factors. The effect of changing ω by a modest multiplicative factor on the rate of convergence is not profound as is seen from a glance at 2.2.11. Generally, p(x) and q(x) are chosen to be identically 1 unless convenience suggests otherwise.

2.3 The Modulus of Regularization for the Maximum Norm

Although it is often easiest to bound the modulus of regularization (in the manner of 2.2.14) for the L_2 norm, $\mu(\cdot) = \|\cdot\|$, we are often more interested in the maximum norm defined by $\mu_\infty(f) \equiv \max\limits_{x \in [a,b]} |f(x)|$ (which exists since $f \in B_1$ are certainly continuous on $[a,b]$). Denote, for convenience, $\rho_\infty(\varepsilon) \equiv \sup\limits_{\substack{\Omega^2(f) \leqslant 1 \\ \|Kf\| \leqslant \varepsilon}} \mu_\infty(f)$ and

$$\rho_2(\varepsilon) \equiv \sup_{\substack{\Omega^2(f) \leqslant 1 \\ \|Kf\| \leqslant \varepsilon}} \|f\| .$$

Theorem 2.3.1

If $p \equiv \min\limits_{x \in [a,b]} p(x)$, $\qquad \rho_\infty(\varepsilon) \leqslant \frac{2}{p^{\frac{1}{4}}} [\rho_2(\varepsilon)]^{\frac{1}{2}}$ for sufficiently small ε.

Proof:

Let x and y be arbitrary points in $[a,b]$ and f satisfy $\Omega^2(f) \leqslant 1$. Then

$$|f^2(x) - f^2(y)| = \left| \int_x^y 2ff'(t)dt \right| \leqslant 2\left\{\int_x^y f^2(t)dt\right\}^{\frac{1}{2}} \left\{\int_x^y f'^2(t)dt\right\}^{\frac{1}{2}}$$

$$\leqslant 2 \|f\| \frac{1}{\sqrt{p}} \left\{\int_x^y p(x)\, f'^2(t)dt\right\}^{\frac{1}{2}} \leqslant 2 \frac{\|f\|}{\sqrt{p}} .$$

Let y be such that the minimum of $|f|$ on $[a,b]$ occurs at y. Then denoting this minimum by f_{min}, f(x) satisfies:

$$f^2(x) \leq f^2_{\min} + \frac{2}{\sqrt{p}} \|f\| \Rightarrow \mu_\infty(f) \leq \left\{\frac{\|f\|^2}{b-a} + \frac{2}{\sqrt{p}} \quad \|f\|\right\}^{\frac{1}{2}}$$
$$= \|f\|^{\frac{1}{2}} \left\{\frac{\|f\|}{b-a} + \frac{2}{\sqrt{p}}\right\}^{\frac{1}{2}} .$$

So $$\rho_\infty(\varepsilon) = \sup_{\substack{\Omega^2(f) \leq 1 \\ \|Kf\| \leq \varepsilon}} \mu_\infty(f) \leq \sup_{\substack{\Omega^2(f) \leq 1 \\ \|Kf\| \leq \varepsilon}} \|f\|^{\frac{1}{2}} \left\{\frac{\|f\|}{b-a} + \frac{2}{\sqrt{p}}\right\}^{\frac{1}{2}}$$

$$\leq \left(\sup_{\substack{\Omega^2(f) \leq 1 \\ \|Kf\| \leq \varepsilon}} \|f\|^{\frac{1}{2}}\right) \left(\sup_{\substack{\Omega^2(f) \leq 1 \\ \|Kf\| \leq \varepsilon}} \left[\frac{\|f\|}{b-a} + \frac{2}{\sqrt{p}}\right]^{\frac{1}{2}}\right)$$

$$= [\rho_2(\varepsilon)]^{\frac{1}{2}} \left\{\frac{\rho_2(\varepsilon)}{b-a} + \frac{2}{\sqrt{p}}\right\}^{\frac{1}{2}} .$$

For small enough ε, $\frac{\rho_2(\varepsilon)}{b-a} \leq \frac{2}{\sqrt{p}}$ and the result follows. □

Since $\|f\| \leq \mu_\infty(f)\,(b-a)^{\frac{1}{2}}$, we have a lower bound on $\rho_\infty(\varepsilon)$ of $\frac{\rho_2(\varepsilon)}{(b-a)^{\frac{1}{2}}}$.

2.4 An Example

The general notion of when the regularizing assumption results in a convergent extension remains somewhat elusive. For the following problem, related, (as will be shown), to a backwards problem of heat flow with variable coefficients, this issue can be approached in a more concrete fashion than our appealing to the topological property of compactness.

Consider the linear two point boundary value problem:

$$\frac{d}{dx}\,[p(x)\,g'(x)] - q(x)\,g(x) = f(x), \qquad \text{for } a < x < b$$

$$p(x) \text{ positive and in } C^1[a,b]^* \qquad (2.4.1)$$

$$q(x) \text{ positive and in } C^0[a,b]$$

subject to the boundary conditions $g^1(a) = g^1(b) = 0.$ (2.4.2)

If we are asked to find g(x) given f(x), (assuming the homogeneous problem with $f(x) \equiv 0$ has only the trivial solution), our problem would be well-posed. A Green's function representation

$$Kf(x) = \int_a^b G(x,y)\,f(y)\,dy = g(x) \qquad (2.4.3)$$

for the solution g(x) would exist. Since the problem is self-adjoint, G(x, y) will be symmetric. The spectral decomposition of the operator K defined by 2.4.3 is obtainable numerically. Indeed, the Sturm-Liouville system:

$$Lu(x) \equiv \frac{d}{dx}\,[p(x)\,u'(x)] - q(x)\,u(x) = -\frac{1}{\lambda}\,u(x)\,; \qquad (2.4.4)$$

$u'(a) = u'(b) = 0$ has a countably infinite set of eigenvalues $\frac{1}{\lambda_n}$ and an associated orthonormal set of eigenfunctions ψ_n. The ψ_n will be complete in $L^2[a,b]$. Expanding both sides of 2.4.1 in the ψ_n gives:

$$L\,(\sum_{n=1}^{\infty} (g,\psi_n)\psi_n) = \sum_{n=1}^{\infty} (f,\psi_n)\psi_n$$

----

*p(x) has a continuous first derivative on [a, b] ; q(x) is continuous on [a, b] .

$$\Rightarrow \sum_{n=1}^{\infty} (g,\psi_n)\frac{1}{\lambda_n}\psi_n = \sum_{n=1}^{\infty} (f,\psi_n)\psi_n$$

$$\Rightarrow (g,\psi_n)\frac{1}{\lambda_n} = (f,\psi_n) \quad \Longrightarrow (f,\psi_n)\lambda_n = (g,\psi_n)$$

whence $$g = \sum_{n=1}^{\infty} \lambda_n (f,\psi_n)\psi_n = Kf \tag{2.4.5}$$

and we recognize the form of Theorem 1.5.8 for compact self-adjoint operators K. In passing, we note that

$$g(x) = \sum_{n=1}^{\infty}\lambda_n \left\{\int_a^b \psi_n(y)f(y)\,dy\right\}\psi_n(x)$$

$$= \int_a^b \left[f(y)\sum_{n=1}^{\infty} \lambda_n\,\psi_n(x)\,\psi_n(y)\right] dy$$

and we identify $G(x,y) \equiv \sum_{n=1}^{\infty} \lambda_n\,\psi_n(x)\,\psi_n(y)$.

The ill-posed problem to be considered here is the inverse problem: given g find f. Of course, if the data are exact and we know how to compute the derivatives analytically, this is no problem at all. Our problem is to invert $Kf_0 = g_0$ given g such that $\|g-g_0\| < \varepsilon$, assuming a regularizing condition.

The nice feature of this problem is that if

$$\Omega^2(f_0) \equiv \int_a^b [p(x)\,f_0'^2(x) + q(x)\,f_0^2(x)]\;dx \leq \omega^2 \tag{2.4.6}$$

and we choose our functions p(x) and q(x) to be the same as in 2.4.1, $\Omega^2(f_0)$ has a convenient representation.

$\Omega^2(f_0)$ is derived from the inner product:

$$\langle f, g\rangle \equiv \int_a^b [\, p(x)f'(x)g'(x) + q(x)f(x)g(x)]\, dx \;; \qquad (2.4.7)$$

$\Omega^2(f_0) = \langle f_0, f_0\rangle$. The ψ_n are orthogonal in this inner product; in fact,

$$\langle \psi_n, \psi_m\rangle = \int_a^b [p(x)\psi'_n(x)\,\psi'_m(x) + q(x)\,\psi_n(x)\,\psi_m(x)]\;\; dx$$

$$= \Big[^{b}_{a}\; p(x)\,\psi_n(x)\,\psi'_m\;\;(x) - \int_a^b \psi_n(x)\left\{\frac{d}{dx}\,\psi(x)\,\psi'_m(x) - q(x)\psi_m(x)\right\} dx$$

$$= 0 \;-\; \int_a^b \psi_n(x)\left[\frac{-1}{\lambda_m}\;\psi_m(x)\right] dx \;=\; \frac{1}{\lambda_m}\,\delta^m_n \;.$$

$$\text{So } \Omega^2(f_0) = \sum_{n=1}^{\infty} \frac{1}{\lambda_n}\,(f_0, \psi_n)^2 \;. \qquad (2.4.8)$$

We will consider regularization with respect to norms of the form:

$$\mu^2(f) \equiv \sum_{n=1}^{\infty} \mu_n (f, \psi_n)^2 \qquad (2.4.9)$$

(The μ_n are to be real and positive.) Bounded sequences $\{\mu_n\}$ will give rise to norms μ satisfying $m\,\|f\|^2 \leqslant \mu^2(f) \leqslant M\,\|f\|^2$ and hence equivalent to the L^2 norm. More interesting are the cases where $\mu_n \to \infty$. For the sum 2.4.9 defining $\mu(f)$ to converge, the $(f, \psi_n)^2$ must tend to zero correspondingly quickly.

Now we ask, "Under what conditions on the μ_n does the regularizing assumption $\Omega^2(f_m) \leqslant 1$ ensure that $\mu^2(f_m) \to 0$ when $\|Kf_m\| \to 0$ for a sequence $\{f_m\} \subset B_1$?"

Theorem 2.4.10

A necessary and sufficient condition for $\Omega^2(f) \leqslant 1$ to regularize the operator K (defined in 2.4.3) under the norm μ (defined in 2.4.8) is that $\lim_{n \to \infty} \lambda_n \mu_n = 0.$

Proof:

First of all, assume $\lim_{n \to \infty} \lambda_n \mu_n = 0$ and define: $\xi_N \equiv \sup_{n > N} \mu_n \lambda_n$. Then $\lim_{N \to \infty} \xi_N = 0$ by hypothesis. Let $\{f_m\}$ be a sequence of functions satisfying $\Omega^2(f_m) \leqslant 1$; $\|Kf_m\| \to 0$. For all m,

$$\mu^2(f_m) = \sum_{n=1}^{\infty} \mu_n (f_m, \psi_n)^2 = \sum_{n=1}^{N} (f_m, \psi_n)^2 \lambda_n^2 \frac{\mu_n}{\lambda_n^2}$$

$$+ \sum_{n=N+1}^{\infty} (f_m, \psi_n)^2 \frac{1}{\lambda_n} u_n \lambda_n$$

$$\leqslant \sup_{n \leqslant N} \frac{\mu_n}{\lambda_n^2} \|Kf_m\|^2 + \xi_N \sum_{n=N+1}^{\infty} (f_m, \psi_n)^2 \frac{1}{\lambda_n} .$$

$$\leqslant \|Kf_m\|^2 \sup_{n \leqslant N} \frac{\mu_n}{\lambda_n^2} + \xi_N .$$

For any η, there exists N_η such that $\xi_N < \frac{\eta}{2}$ if $N \geqslant N_\eta$. $\sup_{n \leqslant N_\eta} \frac{\mu_n}{\lambda n^2}$ is some finite number; call it M_η. Let m be chosen large enough so that $\|Kf_m\|^2 \leqslant \frac{\eta}{2M_\eta}$. Then for sufficiently large m, $\mu^2(f_m) \leqslant \frac{\eta}{2} + \frac{\eta}{2} = \eta$.

Conversely, suppose $\lim_{n \to \infty} \lambda_n \mu_n > \xi > 0$. Then for any N, there exists $n > N$ such that $\lambda_n \mu_n > \xi$. So a subsequence of $\{\lambda_n \mu_n\}$

exists which is larger than ξ - call it $\{\lambda_{n_k}\mu_{n_k}\}$. Consider the sequence $f_k \equiv \sqrt{\lambda_{n_k}}\,\psi_{n_k}$. $\|Kf_k\| = \lambda_{n_k}{}^{3/2} \to 0$ as $k \to \infty$. $\Omega^2(f_k) = \frac{1}{\lambda_{n_k}} \cdot (\sqrt{\lambda_{n_k}})^2 = 1$ and $\mu^2(f_k) = \mu_k\lambda_{n_k} > \xi$ so $\mu^2(f_k)$ does not tend to zero as $k \to \infty$. □

The proof of 2.4.10 gives us the following estimate (upper bound for $\rho_\mu(\varepsilon)$).

$$\rho_\mu^2(\varepsilon) \leqslant \inf_N \left\{ \sup_{n \leqslant N} \varepsilon^2 \frac{\mu_n}{\lambda_n^2} + \sup_{n>N} \mu_n\lambda_n \right\} . \tag{2.4.11}$$

For the L^2 norm, all the μ_n are identically 1 and

$$\rho_2^2(\varepsilon) \leqslant \inf_N \left\{ \sup_{n\leqslant N} \frac{\varepsilon^2}{\lambda_n^2} + \sup_{n>N} \lambda_n \right\} . \tag{2.4.12}$$

The λ_n are decreasing so $\sup\limits_{n\leqslant N} \frac{\varepsilon^2}{\lambda_n^2} = \frac{\varepsilon^2}{\lambda_N^2}$ and $\sup\limits_{n>N} \lambda_n = \lambda_{N+1}$.

Thus 2.4.12 becomes:

$$\rho_2^2(\varepsilon) \leqslant \inf_N \left\{ \frac{\varepsilon^2}{\lambda_N^2} + \lambda_{N+1} \right\} . \tag{2.4.13}$$

Assuming ε is at least small enough that $\lambda_1 \geqslant (2\varepsilon^2)^{\frac{1}{3}}$, there exists n_0 such that $\lambda_{n_0+1} \leqslant (2\varepsilon^2)^{\frac{1}{3}} \leqslant \lambda_{n_0}$. So $\frac{\varepsilon^2}{\lambda_{n_0}^2} + \lambda_{n_0+1} \leqslant \frac{\varepsilon^2}{(2\varepsilon^2)^{\frac{2}{3}}} + (2\varepsilon^2)^{\frac{1}{3}} = \left(\frac{1}{2^{\frac{2}{3}}} + 2^{\frac{1}{3}}\right)\varepsilon^{\frac{2}{3}}$.

$$\Rightarrow \qquad \rho_2(\varepsilon) \leqslant \varepsilon^{\frac{1}{3}} \left(\frac{1}{2^{\frac{2}{3}}} + 2^{\frac{1}{3}}\right)^{\frac{1}{2}} . \qquad (2.4.14)$$

Now consider the sequence of functions $\{f_k\} \equiv \{ \frac{\varepsilon}{\lambda_k} \psi_k\}$.

$$\|Kf_k\| = \varepsilon \; ; \; \Omega^2(f_k) = \frac{\varepsilon^2}{\lambda_k^3} \; ; \; \|f_k\| = \frac{\varepsilon}{\lambda_k} .$$

For f_k to satisfy the constraints:

$$\Omega^2(f_k) \leqslant 1 \quad \text{and} \quad \|Kf_k\| \leqslant 1 \quad ,$$

it is required that $\varepsilon^2 \leqslant \lambda_k^3$. If, in fact, $\varepsilon^2 = \lambda_k^3$ for some value k_0 of k, then $\|f_{k_0}\| = \varepsilon^{\frac{1}{3}}$. So there is an infinite sequence of $\varepsilon \to 0$ for which $\rho_2(\varepsilon)$ is bounded below by $\varepsilon^{\frac{1}{3}}$. This is the same power law as demonstrated for the upper bound in 2.4.14. Although $\varepsilon^{\frac{1}{3}}$ may not be a lower bound for all values of ε, there is little value in carrying the analysis of $\rho_2(\varepsilon)$ further. In making error estimates, one proceeds as if $\rho_2(\varepsilon)$ obeys, (for all $\varepsilon \leqslant \frac{\lambda_1^{\frac{2}{3}}}{2}$),

$$\varepsilon^{\frac{1}{3}} \leqslant \rho_2(\varepsilon) \leqslant \varepsilon^{\frac{1}{3}} \left(\frac{1}{2^{\frac{2}{3}}} + 2^{\frac{1}{3}}\right)^{\frac{1}{2}} . \qquad (2.4.15)$$

We identify the form of 2.2.14 with $h(\varepsilon) = \varepsilon^{\frac{1}{3}}$.

2.5 The Backwards Heat Equation

The main reason for considering regularization of the foregoing problem is that it is related to the backwards problem of heat flow with variable coefficients. Consider finding u(x, T) satisfying

$$u_t = (p(x)u_x)_x - q(x)u \text{ for } a < x < b; \quad 0 < t \qquad (2.5.1)$$

(p(x) and q(x) as in 2.4) subject to the initial condition $u(x, 0) = f_0(x)$ and the "no-flux" boundary conditions $u_x(a, t) = u_x(b, t) = 0$.

This problem is well-posed. By separation of variables, its solution is readily found to be:

$$u(x, T) = \sum_{n=1}^{\infty} e^{-\zeta_n T} (f_0, \psi_n)\, \psi_n(x) \qquad (2.5.2)$$

where ζ_n and $\psi_n(x)$ are respectively the eigenvalues and orthonormal eigenfunctions associated with the Sturm-Liouville problem:

$$\frac{d}{dx}(p(x)\psi_n'(x)) - q(x)\, \psi_n(x) = -\zeta_n\, \psi_n(x) \qquad (2.5.3)$$

subject to $\psi_n(a) = \psi_n(b) = 0$.

This defines a compact operator $K_T : L^2[a, b] \to L^2[a, b]$.

$$K_T f_0 = \sum_{n=1}^{\infty} e^{-\zeta_n T} (f_0, \psi_n)\, \psi_n \ . \qquad (2.5.4)$$

K_T has the same eigenfunctions as the Sturm-Liouville problem 2.5.3 but the eigenvalues ζ_n of 2.5.3 become $e^{-\zeta_n T}$ for the operator K_T. K_T has the integral representation:

$$K_T f_0(x) = \int_a^b f(y) \left[\sum_{n=1}^{\infty} e^{-\zeta_n T} \psi_n(x)\, \psi_n(y) \right] dy = \int_a^b f(y) K(x, y)\, dy \qquad (2.5.5)$$

where the (symmetric) kernel $K(x, y) \equiv \sum_{n=1}^{\infty} e^{-\zeta_n T} \psi_n(x)\, \psi_n(y)$. (2.5.6)

The inverse problem would be to recover f_0 given $g \approx g_0$. To use Tikhonov's method, we would make the regularizing assumption:

$$\Omega^2(f_0) \equiv \int_a^b [p(x)\, f'^2(x) + q(x)\, f^2(x)]\, dx \leqslant \omega^2 . \qquad (2.5.7)$$

By the same method used in section 2.4, we find

$$\Omega^2(f_0) = \sum_{n=1}^{\infty} \zeta_n (f_0, \psi_n)^2 \qquad (2.5.8)$$

while

$$\| K_T f_0 \|^2 = \sum_{n=1}^{\infty} e^{-2\zeta_n T} (\psi_n, f_0)^2 \ . \qquad (2.5.9)$$

In considering regularization with respect to the norm $\mu(f)$ defined by 2.4.9, a necessary and sufficient condition is that $\lim_{n \to \infty} \frac{\mu_n}{\zeta_n} = 0.$

In Franklin [8], the L_2 modulus of regularization $\rho_2(\varepsilon)$ for the case $p(x) \equiv q(x) \equiv 1$; $[a, b] = [0, \pi]$ was found to go down like $(-\log \varepsilon)^{-\frac{1}{2}}$. Not surprisingly, this more general case can be shown to exhibit the same behaviour. This is extremely slow convergence. To use regularization as a numerical method for this problem would be a mistake. We generally expect our convergence estimates to be somewhat modest; $\varepsilon^{\frac{1}{3}}$ convergence means, roughly speaking, that our solution will have $\frac{1}{3}$ as many decimal places of accuracy as our data and this would be a reasonably successful application. However, requiring data accurate to order e^{-10} to get a solution accurate to order 10^{-1} is most unsatisfactory. To get a better return for our data accuracy, we will require stronger restrictive assumptions. That is the subject of Chapter 4.

CHAPTER 3

SOME OBSERVATIONS ON QUASIREVERSIBILITY METHODS

3.1 Quasireversibility Methods

Among the extension methods for solution of the so-called "backwards problems of evolution" (e.g., the backwards problems of heat flow) are the quasireversibility methods.

Lattès and Lions [17] developed such a method and were likely the first to use the term "quasireversibility." Since then, the term has become generic in describing a class of related approaches. In all quasireversibility methods, a differential operator is perturbed by the addition of an extra term modified by a small parameter. Appropriate additional boundary and/or initial conditions are added if the extra term has changed the order of the equation. In this chapter, two such methods will be applied to the backwards heat problem and analyzed from the point of view of Chapter 1.

It is possible to be fairly general in the description of the backwards heat problems to be tackled with no inconvenience. The following problem will thus be considered.

Let L_x be a linear differential operator of the form:

$$L_x u \equiv \frac{\partial}{\partial x}(p(x)u_x) - q(x)u \,, \qquad (3.1.1)$$

where $p(x)$ is positive and continuously differentiable on $[a,b]$; $q(x)$ is continuous on $[a,b]$. Let $B(u, c_j) = 0$ denote a linear boundary condition of the form

$$B(u;c_j) \equiv \alpha_j\, u_x(c_j) + \beta_j\, u(c_j) = 0 \quad \text{for} \quad j = 1,2\,. \tag{3.1.2}$$

The problem of interest is to find $u(x,0) \equiv f_0(x)$ such that

$$u_t = L_x u \quad \text{for} \quad a < x < b\,; \qquad 0 < t < T \tag{3.1.3}$$

subject to

$$B(u;a) = B(u;b) = 0$$

and $u(x,T) = g_0(x)$. (Our actual data will be $g(x) \approx g_0(x)$.) The problem of finding $g_0(x)$ given $f_0(x)$ would be the well-posed forward heat problem. 3.1.3 is, of course, ill-posed.

3.2 Outline of the Method of Lattès and Lions

The related problem to be solved is that of finding $u(x,t;\eta)$ satisfying

$$u_t = L_x\, u + \eta\, L_x^2\, u \quad \text{for} \quad a < x < b; \quad 0 < t < T$$

subject to

$$B(u;a) = B(u;b) = 0$$

$$B(L_x\, u;a) = B(L_x u;b) = 0\,, \tag{3.2.1}$$

and $\quad u(x,T) = g(x)\,.$

Here η is some small parameter. The QR approximation taken for $f_0(x)$ is $f(x) \equiv u(x,0;\eta)$. Lattès and Lions demonstrate that the problem 3.2.1 is well-posed.

3.3 The QR Method of Lattès and Lions as a Well-posed Extension

Consider the Sturm-Liouville system:

$$L_x w = -\xi\, w \qquad (3.3.1)$$

subject to $B(w;a) = B(w;b) = 0$.

It has a countably infinite set of eigenvalues ξ_n and an associated orthonormal system of eigenfunctions ψ_n. These eigenfunctions are complete in $L^2[a,b]$.

Through separation of variables in 3.1.3, we find that $f_0(x)$ and $g_0(x)$ are related by

$$g_0(x) \equiv u(x,T) = \sum_{n=1}^{\infty} (f_0, \psi_n)\, e^{-\xi_n T}\, \psi_n(x) \;; \qquad (3.3.2)$$

the forward heat operator (through time T), being defined by

$$g_0 = K_T\, f_0 = \sum_{n=1}^{\infty} e^{-\xi_n T} (f_0, \psi_n)\, \psi_n \;. \qquad (3.3.3)$$

Separation of variables in 3.2.1 leads to the system:

$$\eta\, L_x^2\, v + L_x\, v = -\lambda\, v$$

subject to

$$B(v,a) = B(v;b) = 0 \qquad (3.3.4)$$

and $\quad B(L_x\, v;a) = B(L_x\, v;b) = 0$.

The eigenfunctions of 3.3.1 are also eigenfunctions of 3.3.4. Certainly, $B(\psi_n;a) = B(\psi_n;b) = 0$ and $B(L_x\psi_n;a) = B(-\xi_n\psi_n;a) = 0$. ($B(L_x\psi_n;b) = 0$ similarly.) The associated eigenvalues λ_n for 3.3.4 are related to the ξ_n of 3.3.1 by $\lambda_n = \xi_n - \eta\xi_n^2$. So the solution to 3.2.1 is

$$f(x) \equiv u(x,0;\eta) = \sum_{n=1}^{\infty} e^{(\xi_n - \eta\xi_n^2)T} (g,\psi_n)\psi_n(x) \,. \qquad (3.3.5)$$

(It is interesting to note in passing that were we solving a forwards version of 3.2.1, it would be ill-posed!) Equation 3.3.5 enables us to define a QR operator $Q_{T;\eta}$ whose domain is all of $L^2[a,b]$ by

$$f = Q_{T;\eta}\, g \equiv \sum_{n=1}^{\infty} e^{(\xi_n - \eta\xi_n^2)T} (g,\psi_n)\,\psi_n \,. \qquad (3.3.6)$$

In estimating f_0 by f, there are two distinct sources of error. One stems from the inexactitude of our knowledge of g_0 - the fact that our data g is only approximate - $\|g-g_0\| \leqslant \varepsilon$. The fact that a different problem is being solved also gives a "bias" error. $Q_{T;\eta}\, g_0$ will not be precisely f_0. Specifically,

$$\begin{aligned} \|f-f_0\| &= \|Q_{T;\eta} g - f_0\| = \|Q_{T;\eta}(g-g_0) + Q_{T;\eta} g_0 - f_0\| \\ &\leqslant \|Q_{T;\eta}(g-g_0)\| + \|Q_{T;\eta} g_0 - f_0\| \\ &\leqslant \|Q_{T;\eta}\| \; \|g-g_0\| + \|(Q_{T;\eta} K_T - I)\, f_0\| \\ &\leqslant \|Q_{T;\eta}\|\varepsilon + \|(Q_{T;\eta} K_T - I)\, f_0\| \,. \end{aligned} \qquad (3.3.7)$$

The first term is our error from our data; the second is our bias. As $\eta \to 0$, we expect the bias to g_0 to zero and our $\|Q_{T;\eta}\|$ to become infinite. Indeed as $\eta \to 0$, the original problem 3.1.3 is recovered; hence no bias but complete instability. How quickly bias disappears and how quickly instability returns as $\eta \to 0$ are key questions in gauging the effectiveness of the method.

First, $\|Q_{T,\eta}\|$ will be computed.

$$\|Q_{T;\eta}\|^2 = \sup_{\|g\|=1} \|Q_{T,\eta}g\|^2 = \sup_{\|g\|^2=1} \sum_{n=1}^{\infty} e^{2(\xi_n-\eta\xi_n^2)T}(g,\psi_n)^2$$

$$\leqslant \sup_n e^{2(\xi_n-\eta\xi_n^2)T} \sup_{\|g\|^2=1} \sum_{n=1}^{\infty}(g,\psi_n)^2 = \sup_n e^{2(\xi_n-\eta\xi_n^2)T}$$

$$\leqslant \sup_{x>0} e^{2(x-\eta x^2)T} \quad \text{where } x \text{ is a continuous variable.}$$

By calculus, $\sup_x e^{2(x-\eta x^2)T} = e^{\frac{T}{2\eta}}$ and the estimate $\|Q_{T;\eta}\| \leqslant e^{\frac{T}{4\eta}}$ 3.3.8 results. Furthermore, a lower bound valid uniformly in η does not exist. The maximum of $e^{2(x-\eta x^2)T}$ occurs at $x = \frac{1}{2\eta}$. Since $\xi_n \to \infty$, for appropriate η, there will exist $\xi_N = \frac{1}{2\eta}$ in which case $\|Q_{T,\eta}\psi_N\|$ will achieve the bound 3.3.8.

Now, the bias will be considered. For f_0 in $L^2[a,b]$,

$$\|(Q_{T;\eta}K_T-I)\,f_0\|^2 = \sum_{n=1}^{\infty}[e^{(\xi_n-\eta\xi_n^2)T}\,e^{-\xi_nT}-1]^2\,(f_0,\psi_n)^2$$

$$= \sum_{n=1}^{\infty}[e^{-\eta\xi_n^2T}-1]^2\,(f_0,\psi_n)^2 . \qquad (3.3.9)$$

So for fixed f_0, the bias goes to zero as $\eta \to 0$. This occurs, however, in a highly nonuniform fashion in f_0. If the value of η is fixed and f_0 taken to be $f_0 = \|f_0\|\,\psi_m$, $\|Q_{T;\eta}K_T-I)\;\|f_0\|\,\psi_m\|^2 =$ $\|f_0\|^2\,(1-e^{-\eta\xi_m^2T})^2$. This quantity approaches 1 as $m \to \infty$ so for any $\eta > 0$, $\|Q_{T;\eta}K_T-I\| = 1$. The convergence of the bias to zero may be arbitrarily slow in η depending on the value of f_0.

The QR method has no built-in mechanism for restricting the f_0's to some admissibility class. It is thus not surprising that this nonuniform bias occurs. One might consider adjoining, in an ad hoc fashion, some restrictive assumption. This can assuredly be done. For example, bounding $\|f_0\|$ and adding a spectral cut-off assumption: $f_0 = \sum_{n=1}^{N} (f_0, \psi_n)\psi_n$ will control bias. But then one would have no particular reason not to use straight spectral cut-off. The beauty of the QR method lies in the ease of its implementation. Its extended problem is easily solved numerically through finite difference schemes. Additional restrictions will tend to get in the way; to compromise that desirable feature. If the restrictions are conducive to some other approach, the usefulness of the QR method will be largely obviated.

The other quasireversibility method to be examined improves upon the bias estimates but at some cost to stability.

3.4 The Method of Gajewski and Zacharias

This approach, described in Gajewski and Zacharias [10] extends the problem 3.1.3 as follows. Find $u(x,t;\eta)$ satisfying

$$u_t = L_x u + \eta\, L_x u_t \qquad \text{for} \qquad a < x < b \qquad 0 < t < T$$

$$\text{subject to} \qquad B(u;a) = B(u;b) = 0 \;; \tag{3.4.1}$$

$$u(x,T) = g(x) \,.$$

The approximation for $f_0(x)$ taken is $f(x) \equiv u(x,0;\eta)$. This will be denoted the G-Z method.

Separation of variables leads to the system:

$$-\lambda(w-\eta\ L_x w) = L_x w$$

$$B(w;a) = B(w;b) = 0 \quad . \tag{3.4.2}$$

Again, the eigenfunctions ψ_n of 3.3.1 are eigenfunctions of this system. The eigenvalues λ_n of 3.4.2 are related to the ξ_n of 3.3.1 by $-\lambda_n(1+\eta\xi_n) = -\xi_n$ or $\lambda_n = \frac{\xi_n}{1+\eta\xi_n}$. By separation of variables, the solution to 3.4.1 is

$$u(x,t;\eta) = \sum_{n=1}^{\infty} e^{\frac{\xi_n}{1+\eta\xi_n}(T-t)} (g,\psi_n)\psi_n(x) \ , \tag{3.4.3}$$

whence

$$f(x) \equiv u(x,0;\eta) = \sum_{n=1}^{\infty} e^{\frac{\xi_n T}{1+\eta\xi_n}} (g,\psi_n)\psi_n(x) \ . \tag{3.4.4}$$

This defines an operator $S_{T,\eta}$ by

$$S_{T,\eta}\, g = \sum_{n=1}^{\infty} e^{\frac{\xi_n T}{1+\eta\xi_n}} (g,\psi_n)\ \psi_n \ . \tag{3.4.5}$$

The bias for the G-Z method will be, (for a given f_0), $\|(S_{T,\eta}\, K_T - I)\ f_0\|$. This is given by

$$\|(S_{T,\eta} K_T - I)\ f_0\|^2 = \sum_{n=1}^{\infty} \left\{[e^{\frac{\xi_n T}{1+\eta\xi_n}} \ e^{-\xi_n T} - 1]\ (f_0,\psi_n)\right\}^2$$

$$= \sum_{n=1}^{\infty} \left\{[e^{\frac{-\eta\xi_n^2}{1+\eta\xi_n}} - 1]\ (f_0,\psi_n)\right\}^2 \quad . \tag{3.4.6}$$

A glance at 3.3.9 shows this bias estimate is always superior. However, there is still the problem of nonuniformity in the convergence of the bias to zero and, in fact, $\|S_{T,\eta}K_T - I\| = 1$ just as before.

The stability of the G-Z method depends upon the norm of $S_{T,\eta}$. From 3.4.5,

$$\|S_{T,\eta}\|^2 = \sup_{\|g\|=1} \sum_{n=1}^{\infty} e^{\frac{2\xi_n T}{1+\eta\xi_n}} (g,\psi_n)^2 \leqslant \sup_{x>0} e^{\frac{2xT}{1+\eta x}}$$

where x is a continuous variable. This sup is approached as $x \to \infty$ and is easily seen to be $e^{\frac{2T}{\eta}}$. No better estimate is possible. This bound is approached as $m \to \infty$ by $\|S_{T,\eta}\psi_m\|$. This is worse than the QR method by a power of four.

For the G-Z method, the counterpart to the overall error estimate 3.3.7 (for the QR method) is

$$\|f-f_0\| \leqslant \|S_{T,\eta}\| \varepsilon + \|(S_{T;\eta}K_T - I) f_0\|$$

$$= e^{\frac{T}{\eta}} \varepsilon + \|(S_{T;\eta} K_T - I) f_0\| \quad . \qquad (3.4.7)$$

3.5 Summary

In studying these quasireversibility methods, one is immediately impressed by the simple elegance of the underlying ideas and the ease with which they can be implemented numerically.

When the analysis is carried further, one discovers the difficulties mentioned with bias error. That argument, however, is by no means a compelling reason for abandoning the QR or G-Z method. Competitive methods, of the kind described in Chapter 1, make restrictive assumptions about the solution f_0. If a method's comparison is to be fair, then those restrictive assumptions should be invoked as side conditions in working out QR or G-Z bias. Why should one method be made to cope with outlandish possibilities which its competition disallows? It will be possible, with appropriate restrictions, to make our bias error behave linearly in the small parameter η^* as $\eta \to 0$. So at this level of discussion, quasireversibility might still compare very favourably.

Allowing the linear bias estimate, our QR full error estimate 3. 3. 7 would take the form $\varepsilon\, e^{\frac{T}{4\eta}} + r\eta$ where r is some constant. It is natural to choose $\eta(\varepsilon)$ to minimize this quantity. By calculus, one finds that this minimum occurs when η has the value η_0 satisfying $\varepsilon \frac{T}{4\eta^2} e^{\frac{T}{4\eta}} = r$, the value of the minimum being $\frac{4r\eta_0^2}{T} + r\,\eta_0$. The value of η_0 can be found approximately

$\eta_0 \approx \frac{-T}{4\log\varepsilon}$. The behaviour of our error estimate, granting these quasireversibility methods all possible concessions, becomes logarithmic in ε. We are unable to improve on Tikhonov's method by either QR or G-Z.

The difficulty is not easily shrugged off. At the outset it was known that stability would be lost as the small parameter η went to zero. The misfortune is that it is lost so very rapidly as η becomes small. The exponential growth in norms of the significant

----

*Bias which is $O(\eta)$ as $\eta \to 0$ is the best behaviour one can force through restrictions on f_0 of the kind considered.

operators $Q_{T,\eta}$ and $S_{T,\eta}$ is unacceptable and in any natural implementation, the effect on computed solutions will be quite unavoidable.

Quasireversibility methods might be found which will give better results. Such a method would have somehow to avoid the exponential growth of instability possessed by QR and G-Z. It is not at all obvious how this should be done.

CHAPTER 4

THE BACKWARDS HEAT EQUATION - A NUMERICAL SOLUTION

The methods discussed in Chapters 2 and 3 give logarithmic convergence when applied to backwards heat flow problems. To improve upon this, strong restrictive assumptions about our class of admissible solutions will be required.

In fact, such an improvement has been around since the mid 1950's when John and Pucci were active in ill-posed problem theory. The first part of this chapter (sections 4.1-4.3) will be, in a sense, going over old ground in discussing this extension; logarithmic convexity and spectral analysis, (now standard methods in the field), will be used to establish its effectiveness. The rest of the chapter will be concerned with numerical solution of the extended problem. Section 4.4 introduces a related problem whose solution is shown to approximate that of the extension. This related problem is amenable to numerical solution. Computations performed on it are presented in section 4.5.

4.1 The Problem and Its Extension

Define the linear differential operator $\mathcal{L}_x$ by

$$\mathcal{L}_x u \equiv qu - (pu_x)_x \tag{4.1.1}$$

where $p(x) > 0$, $q(x) > 0$ on $[a,b]$; $p \in C^1[a,b]$; $q \in C^0[a,b]$. As in Chapter 3, boundary conditions will be applied having the form

$$\alpha_1 u(a) + \beta_1 u_x(a) = 0^* \quad \text{and} \quad \alpha_2 u(b) + \beta_2 u_x(b) = 0$$

----

* We rule out the trivial cases $\alpha_1 = \beta_1 = 0$ and $\alpha_2 = \beta_2 = 0$.

and these will be denoted $B(u;a) = B(u;b) = 0$ for convenience.

Let $\rho(x)$ be positive and continuous on $[a,b]$ and consider the problem of finding $u_0(x,\tau)$ for some τ in $(0,T)$ given that

$$\rho u_t = -\mathcal{L}_x u_0 \qquad \text{for} \qquad 0 < t < T; \;\; a < x < b$$

subject to

$$B(u_0;a) = B(u_0;b) = 0 \tag{4.1.2}$$

and $\quad u_0(\,,T) = g_0{}^{*}.$

This problem is ill-posed: unstable with respect to perturbations in g_0. We shall, of course, assume that our knowledge of g_0 is approximate; that we, in fact, know $g \approx g_0$.

To treat this problem it will be convenient to define a different inner product on $L^2[a,b]$.

Definition 4.1.3

Let f and g be in $L^2[a,b]$. Define the inner product $\langle f, g\rangle$ by

$$\langle f, g\rangle \equiv \int_a^b f(x)\, g(x)\, \rho(x)\, dx$$

and denote its accompanying norm by $\|\cdot\|_2$. $(\|f\|_2^2 = \langle f, f\rangle)$.

In terms of our usual inner product $(\,,\,)$ (corresponding to the norm $\|\cdot\|$), we would have $\langle f, g\rangle = (f, \rho g)$ and $\|f\|_2 = \|\sqrt{\rho}\, f\|$.

----

*In this chapter, a function's dependence on a spatial variable will not be displayed when it is being considered as an element of a function space. The function $w(x,t)$ if being considered as an element of $L^2[a,b]$ would be denoted $w(\,,t)$.

Since $\rho(x)$ is continuous on $[a, b]$ it has upper and lower bounds: $C_1 > 0$ and $C_2 > 0$. Thus, the norms $\| \cdot \|_2$ and $\| \cdot \|$ are comparable: $\sqrt{C_2} \ \|f\| \leqslant \|f\|_2 \leqslant \sqrt{C_1} \ \|f\|$. Both kinds of inner product will be used.

If one seeks a solution to 4.1.2, he will be led to the Sturm-Liouville system:

$$(p\,w_x)_x + (\lambda\rho - q)\,w = 0 \tag{4.1.4}$$

subject to $B(w;a) = B(w;b) = 0$. This system has a set of eigenfunctions φ_n, complete and orthonormal with respect to the inner product $\langle \cdot\,, \cdot \rangle$ and norm $\| \cdot \|_2$.

Noting that any eigenfunction φ_n and its associated eigenvalue λ_n satisfy

$$\mathcal{L}_x\,\varphi_n = \lambda_n\,\rho\,\varphi_n\,, \tag{4.1.5}$$

one sees that any f in the domain of $\mathcal{L}_x$ satisfies

$$(f, \mathcal{L}_x f) = \sum_{n=1}^{\infty} \lambda_n \langle f, \varphi_n \rangle^2 .$$

So $(f, \mathcal{L}_x f)$ will be positive for all f in the domain of $\mathcal{L}_x$ if and only if all the eigenvalues of system 4.1.4 are positive. In such an instance, $\mathcal{L}_x$ will be said to be positive definite.

It is now possible to define the extension to 4.1.2. Find any $u_o(,\tau)$ satisfying

$$\rho\, u_{0_t} = -\mathcal{L}_x u_0 \qquad \text{for} \quad 0 < t < T\ ;\ a < x < b \tag{4.1.6}$$

subject to $B(u_0;a) = B(u_0;b) = 0$,

$$\|u_0(,T) - g\|_2 \leqslant \varepsilon ,$$

and $$\|u_0(,0)\|_2 \leqslant M.$$

It is assumed that a solution exists to 4.1.6. Physically, we would be asking for the temperature profile at time τ in a finite rod given the measured profile at time $T > \tau$, assuming that at time 0, the net heat contained in the rod was bounded by some known quantity. If $\mathcal{L}_x$ is positive definite, heat will be dissipated in the rod, any initial profile gradually decaying away as time progresses.

4.2 Logarithmic Convexity

Inasmuch as the solution to 4.1.6 is not unique, we need to know how far apart its solutions can be. The simplest and perhaps most elegant means of assessing this is by a logarithmic convexity argument. What one obtains is an upper bound for the error inherent in the extension. A more detailed appraisal by spectral analysis will show us that it is not just an upper bound but the best one* we can, in general, obtain.

We need a few interim results at this point.

Lemma 4.2.1 (Self-adjointness of $\mathcal{L}_x$)†

Let h_1 and h_2 be in $C^2[a,b]$ and satisfy

$B(h_j;a) = B(h_j;b) = 0 \quad$ for $j = 1,2$.

Then $(h_1, \mathcal{L}_x h_2) = (h_2, \mathcal{L}_x h_1)$.

----

*The meaning of that statement will be clarified in section 4.3.

† Proof is well-known. See, for example, Stakgold [24].

Now we can prove

Theorem 4.2.2

Let u(x, t) satisfy

$$\rho\, u_t = - \mathcal{L}_x u \qquad \text{for} \qquad 0 < t < T \ ; \quad a < x < b$$

subject to $B(u;a) = B(u;b) = 0$ and $u(\,,0) = f \in L^2[a,b]$. Then $\log \|u(\,,t)\|_2$ is a convex function of time.

Proof:

It is sufficient to show that $\frac{d^2}{dt^2} \log \langle u(\,,t), u(\,,t)\rangle \geqslant 0.$

$$\frac{d^2}{dt^2} \log \langle u(\,,t), u(\,,t)\rangle =$$

$$= \frac{\langle u(\,,t), u(\,,t)\rangle \frac{d^2}{dt^2} \langle u(\,,t), u(\,,t)\rangle - [\frac{d}{dt} \langle u(\,,t), u(\,,t)\rangle]^2}{\langle u(\,,t), u(\,,t)\rangle^2} \ .$$

$$\frac{d}{dt} \langle u(\,,t), u(\,,t)\rangle = 2\langle u(\,,t),\ u_t(\,,t)\rangle = 2\Big(u(\,,t),\ \rho u_t(\,,t)\Big)$$

$$= -2\Big(u(\,,t),\ \mathcal{L}_x u(\,,t)\Big) = -2\langle u(\,,t),\ \frac{\mathcal{L}_x u(\,,t)}{\rho}\rangle \ .$$

$$\frac{d^2}{dt^2} \langle u(\,,t), u(\,,t)\rangle = -2\Big(u(\,,t), \mathcal{L}_x u_t(\,,t)\Big) -2\Big(u_t(\,,t),\ \mathcal{L}_x u(\,,t)\Big).$$

By hypothesis,

$$B(u(,t);a) = B(u(,t);b) = 0 \ \Rightarrow\ B(u_t(\,,t);a) = B(u_t(\,,t);b) = 0;$$

so, (by Lemma 4.2.1),

$$\frac{d^2}{dt^2} \langle u(\,,t), u(\,,t)\rangle = -4\Big(u_t(\,,t),\ \mathcal{L}_x u(\,,t)\Big) = 4\Big(\frac{\mathcal{L}_x u(\,,t)}{\rho}\ , \mathcal{L}_x u(\,,t)\Big)$$

$$= 4\ \langle \frac{\mathcal{L}_x u(\,,t)}{\rho}\ ,\ \frac{\mathcal{L}_x u(\,,t)}{\rho}\rangle \ .$$

The denominator of the expression for $\frac{d^2}{dt^2} \log \langle u(,t), u(,t)\rangle$ is positive; the numerator is

$$4\langle u(,t), u(,t)\rangle \langle \frac{\mathcal{L}_x u(,t)}{\rho}, \frac{\mathcal{L}_x u(,t)}{\rho}\rangle - 4\langle u(,t), \frac{\mathcal{L}_x u(,t)}{\rho}\rangle^2 .$$

This is non-negative by Schwartz' inequality.* □

Now suppose $u_1(,\tau)$ and $u_2(,\tau)$ are both solutions to 4.1.6. Their difference $(u_1-u_2)(x,t)$ will satisfy the differential equation and boundary conditions. Furthermore,

$$\|u_1(,T)-u_2(,T)\|_2 \leq \|u_1(,T)-g\|_2 + \|u_2(,T)-g\|_2 \leq 2\varepsilon ,$$

and

$$\|u_1(,0)-u_2(,0)\|_2 \leq \|u_1(,0)\|_2 + \|u_2(,0)\|_2 \leq 2M .$$

By theorem 4.2.2, $\log \|(u_1-u_2)(,t)\|_2$ is a convex function of time so

$$\log\|(u_1-u_2)(,\tau)\|_2 \leq \frac{T-\tau}{T}\log\|(u_1-u_2)(,0)\|_2 + \frac{\tau}{T}\log\|(u_1-u_2)(,T)\|_2$$

$$\leq \frac{T-\tau}{T}\log 2M + \frac{\tau}{T}\log 2\varepsilon$$

$$= \log\left[(2M)^{\frac{T-\tau}{T}}(2\varepsilon)^{\frac{\tau}{T}}\right]$$

$$\Rightarrow \|(u_1-u_2)(,T)\|_2 \leq (2M)^{\frac{T-\tau}{T}}(2\varepsilon)^{\frac{\tau}{T}} = 2M^{\frac{T-\tau}{T}}\varepsilon^{\frac{\tau}{T}} .$$

So, two functions satisfying the conditions of the extended problem 4.1.6 can differ by, at most, $2\varepsilon^{\frac{\tau}{T}} M^{\frac{T-\tau}{T}}$ at time $t = \tau$.

----

*Applied to $\langle u(,t), \frac{\mathcal{L}_x u(,t)}{\rho}\rangle^2$.

4. 3 Estimate by Spectral Analysis

This extension can be examined from the viewpoint of Chapter 1. Section 4. 2 has given us an upper bound on the uncertainty associated with a solution of 4. 1. 6. However, is this the best we can claim ? The answer will prove to be yes, but it is not yet obvious that such is the case.

Define a heat operator K_t mapping (forward) through time t as follows. The solution to

$$\rho u_t = - \mathcal{L}_x u \qquad 0 < t < T \; ; \quad a < x < b \tag{4. 3. 1}$$

subject to $B(u;a) = B(u;b) = 0$

and $u(, 0) = f$, evaluated at $t = \tau$ will be $K_\tau f$. Separation of variables leads to the spectral representation

$$K_\tau f = \sum_{n=1}^{\infty} \langle f, \varphi_n \rangle \, e^{-\lambda_n \tau} \varphi_n$$

where the φ_n and λ_n refer to the Sturm-Liouville system 4. 1. 4.

Its inverse will be denoted by $K_{-\tau}$ (the backwards operator) but care must be taken in all manipulations that $K_{-\tau}$ is only being applied to elements in its domain $\subset L^2[a, b]$.

In this notation, 4. 1. 6 can be reformulated. Find $u_0(, \tau)$ satisfying

$$\| K_{T-\tau} u_0(, \tau) - g \|_2 \leqslant \varepsilon$$

such that $\| K_{-\tau} u_0(, \tau) \|_2$ exists and is bounded by M.

In this form, 4. 1. 6 can readily be cast into the language of Chapter 1. Specifically, identify the operator K of Chapter 1 as

$K_{T-\tau}$ and the set A to which our solutions will be restricted as the closure of the set

$$\{f_0 \in L^2[a,b] \mid f_0 \in \text{Dom}\,(K_{-\tau})\,; \quad \|K\text{-}\tau\ f_0\| \leqslant M\}\ .$$

We wish to compute $\eta_A(g;\varepsilon)$ defined by

$$\eta_A(g;\varepsilon) = \sup_{\substack{f_1;f_2 \in A \\ K_{T-\tau} f_1;\, K_{T-\tau} f_2 \in N_\varepsilon(g)}} \|f_1 - f_2\|_2$$

and, by the definition of A, this is

$$= \sup_{\substack{\|h_1\| \leqslant M;\ \|h_2\| \leqslant M \\ \|K_T h_1 - g\| \leqslant \varepsilon\ ;\ \|K_T h_2 - g\| \leqslant \varepsilon}} \|K_\tau\ (h_1 - h_2)\|_2 \qquad (h_1;h_2 \text{ respectively } K_{-\tau} f_1 \text{ and } K_{-\tau} f_2)\ ,$$

$$\leqslant \sup_{\substack{\|h_1 - h_2\| \leqslant 2M \\ \|K_T(h_1 - h_2)\| \leqslant 2\varepsilon}} \|K_\tau\ (h_1 - h_2\|_2 \qquad \text{and denoting } h_1 - h_2 \text{ by h, this is}$$

$$= \sup_{\substack{\|h\| \leqslant 2M \\ \|K_T\ h\| \leqslant 2\varepsilon}} \|K_\tau\ h\|_2\ .$$

Under the restrictions $\|h\| \leqslant 2M \quad \|K_T h\| \leqslant 2\varepsilon$,

$$\|K_\tau h\|_2^2 = \sum_{n=1}^{\infty} \langle h, \varphi_n\rangle^2 e^{-2\lambda_n \tau} = \sum_{n=1}^{N} \langle h, \varphi_n\rangle^2 e^{-2\lambda_n T} e^{2\lambda_n (T-\tau)}$$

$$+ \sum_{n=N+1}^{\infty} \langle h, \varphi_n\rangle^2 e^{-2\lambda_n \tau}$$

$$\leqslant 4\varepsilon^2 e^{2\lambda_N (T-\tau)} + 4M^2 e^{-2\lambda_{N+1}\tau} \quad . \tag{4.3.2}$$

This estimate is valid for all N; so we have that

$$\sup_{\substack{\|h\|_2 \leqslant 2M \\ \|K_T h\|_2 \leqslant 2\varepsilon}} \|K_\tau h\|_2 \leqslant 4 \inf_N \left[\varepsilon^2 e^{2\lambda_N (T-\tau)} + M^2 e^{-2\lambda_{N+1}\tau}\right].$$

Since the λ_n are monotonically increasing to ∞ as $n \to \infty$, there exists a value n_0 of N for which

$$\lambda_{n_0} \leqslant \frac{1}{2T} \log\left[\frac{M^2 \tau}{\varepsilon^2 (T-\tau)}\right] \leqslant \lambda_{n_0+1} \tag{4.3.3}$$

(provided that ε is small enough that $\lambda_1 \leqslant \frac{1}{2T} \log\left[\frac{M^2\tau}{\varepsilon^2(T-\tau)}\right]$.)

So bounding $\inf\limits_N$ using the above estimate with n_0,

$$\sup_{\substack{\|h\|_2 \leqslant 2M \\ \|K_T h\|_2 \leqslant 2\varepsilon}} \|K_\tau h\|_2 \leqslant 4(\varepsilon^2)^{\frac{\tau}{T}} (M^2)^{\frac{T-\tau}{T}} \left\{\frac{1}{(1-\frac{\tau}{T})^{1-\frac{\tau}{T}} (\frac{\tau}{T})^{\frac{\tau}{T}}}\right\} .$$

(after some algebraic simplification.) (The value $x_0 = \frac{1}{2T} \log\left[\frac{M^2 \tau}{\varepsilon^2 (T-\tau)}\right]$

is where $\varepsilon^2 e^{-2x(T-\tau)} + M^2 e^{-2x\tau}$ takes on its minimum value; hence the interest in λ_n near that point.)

So we get the estimate:

$$\eta_A(g;\varepsilon) \leqslant 2\varepsilon^{\frac{\tau}{T}} M^{\frac{T-\tau}{T}} \left\{ \frac{1}{(1-\frac{\tau}{T})^{1-\frac{\tau}{T}} (\frac{\tau}{T})^{\frac{\tau}{T}}} \right\}^{\frac{1}{2}} .$$

The factor in braces goes between 1 and $\sqrt{2}$ as τ goes from 0 to $\frac{T}{2}$ and is symmetric about $\tau = \frac{T}{2}$. This estimate is larger than that of section 4.2 because at least one of the inequalities in 4.3.3 is strict. The effect is giving up a factor of (up to) $\sqrt{2}$ in the estimate.

Now consider the sequence $\{h_k\} \equiv \{2M\varphi_k\}$. Certainly $\|h_k\| = 2M$ for all k and $\|K_T h_k\| = 2Me^{-\lambda_k T}$. To satisfy $\|K_T h_k\| \leqslant 2\varepsilon$, k must be such that $2Me^{-\lambda_k T} \leqslant 2\varepsilon$. This implies $-\lambda_k T \leqslant \log \frac{\varepsilon}{M}$. Let k_0 be the smallest such k. $e^{-\lambda_{k_0}\tau}$ will be approximately $(\frac{\varepsilon}{M})^{\frac{\tau}{T}}$ if $\lambda_{k_0} \approx -\frac{1}{T} \log \frac{\varepsilon}{M}$ (i.e., comes close to satisfying the constraint exactly). In this case $\|K_\tau h_{k_0}\| \approx 2M^{\frac{T-\tau}{T}} \varepsilon^{\frac{\tau}{T}}$. There will be a sequence of $\varepsilon \to 0$ for which $-\frac{1}{T}\log \frac{\varepsilon}{M}$ is precisely an eigenvalue. Then the estimate $2M^{\frac{T-\tau}{T}} \varepsilon^{\frac{\tau}{T}}$ will be attained. For a general small ε, the lowest bound we can guarantee for $\eta_A(g;\varepsilon)$ is $2M^{\frac{T-\tau}{T}} \varepsilon^{\frac{\tau}{T}} \leqslant \eta_A(g;\varepsilon)$. We can not improve on the bound obtained so easily and cleanly through logarithmic convexity.

4.4 A Related Problem for Numerical Solution

In its form 4.1.6, the extension does not lend itself to numerical computation. Many solutions exist and what we now seek is a good numerical algorithm giving a discretized approximation to any one of them.

Suppose we sought to minimize $\|K_{T-\tau} u-g\|_2$ subject to the constraint that $\|K_{-\tau} u\|_2 \leqslant M$. Then $u_0(,\tau)$ of 4.1.6 would satisfy $\|K_{T-\tau} u_0(,\tau)-g\|_2 \leqslant \varepsilon$. The minimum would do at least as well and we might hope that this minimum u and $u_0(,\tau)$ would be close together; that something analogous to logarithmic convexity might prevail. In Franklin [8], it is pointed out that Tikhonov's method of regularization has an interpretation as a constrained extremum problem and the parameter α may be thought of as a Lagrange multiplier. This motivates the following related problem.

Find $u(,\tau)$ minimizing the quadratic functional

$$\|K_{T-\tau} w-g\|_2^2 + \alpha \|K_{-\tau} w\|_2^2 .^* \qquad (4.4.1)$$

(It must be stressed that the foregoing argument was purely of motivational intent. A more thorough analysis will soon be given.)

The minimum of 4.4.1 is readily found through spectral analysis. Indeed, if w is in the domain of $K_{-\tau}$ and denoted

----

*For w not in the domain of $K_{-\tau}$, take $\|K_{-\tau} w\|_2$ as being infinite.

$$w \equiv \sum_{n=1}^{\infty} w_n \varphi_n; \quad g \equiv \sum_{n=1}^{\infty} g_n \varphi_n$$

$$(w_n \equiv \langle w, \varphi_n \rangle \ ; \ g_n \equiv \langle g, \varphi_n \rangle)$$

then

$$\| K_{T-\tau} w - g \|_2^2 + \alpha \| K_{-\tau} w \|_2^2 = \sum_{n=1}^{\infty} (e^{-\lambda_n (T-\tau)} w_n - g_n)^2 + \alpha e^{2\lambda_n \tau} w_n^2 \ .$$

This is

$$\sum_{n=1}^{\infty} \left\{ (e^{-2\lambda_n (T-\tau)} + \alpha e^{2\lambda_n \tau}) \left[w_n - \frac{e^{-\lambda_n (T-\tau)} g_n}{e^{-2\lambda_n (T-\tau)} + \alpha e^{2\lambda_n \tau}} \right]^2 + \right.$$

$$\left. + \frac{\alpha e^{2\lambda_n \tau} g_n^2}{e^{-2\lambda_n (T-\tau)} + \alpha e^{2\lambda_n \tau}} \right\}$$

(by a process of completing the square term by term in the w_n). Cancelling some $e^{\lambda_n \tau}$ factors and splitting into two parts, we get

$$\sum_{n=1}^{\infty} (e^{-2\lambda_n (T-\tau)} + \alpha e^{2\lambda_n \tau}) \left[w_n - \frac{e^{-\lambda_n (T+\tau)} g_n}{e^{-2\lambda_n T} + \alpha} \right]^2 +$$

$$+ \sum_{n=1}^{\infty} \frac{\alpha g_n^2}{e^{-2\lambda_n T} + \alpha} \ .$$

The second term is independent of w, the first will vanish if

$$w_n = \frac{e^{-\lambda_n (T+\tau)} g_n}{e^{-2\lambda_n T} + \alpha} \ ;$$

otherwise the first term is positive. w, so represented, is an L^2 function; is in the domain of $K_{-\tau}$ and solves the minimum problem. Thus

$$u(,\tau) = \sum_{n=1}^{\infty} \frac{\langle g, \varphi_n\rangle e^{-\lambda_n(T+\tau)}}{e^{-2\lambda_n T}+\alpha} \varphi_n \quad . \tag{4. 4. 2}$$

Then we observe that 4. 4. 2 is the spectral representation of the solution w to

$$(K_{2T} + \alpha I)w = K_{T+\tau}\, g \quad . \tag{4. 4. 3}$$

Solving 4. 4. 3 thus replaces the problem of minimizing 4. 4. 1. We may do this either by taking a few terms in the expansion 4. 4. 2, (taking care to account for error in truncating the series), or we may develop a finite difference method for direct solution of 4. 4.3.

In 4. 4. 2, a linear operator is defined. Call it $F_{\tau,\alpha}$

$$F_{\tau,\alpha} g = u(,\tau) = \sum_{n=1}^{\infty} \langle g, \varphi_n\rangle \frac{e^{-\lambda_n(T+\tau)}}{\alpha+e^{-2\lambda_n T}} \tag{4. 4. 4}$$

$F_{\tau,\alpha}$ is bounded and hence 4. 4. 3 is well-posed. In fact,

$$\|F_{\tau,\alpha}\| = \sup_n \frac{e^{-\lambda_n(T+\tau)}}{\alpha+e^{-2\lambda_n T}} \quad .$$

This can be bounded by

$$\sup_{x>0} \frac{e^{-x(T+\tau)}}{e^{-2xT}+\alpha}$$

which is

$$\alpha^{-(\frac{1}{2} - \frac{\tau}{2T})} \left(\tfrac{1}{2} + \frac{\tau}{2T}\right)^{(\frac{1}{2} + \frac{\tau}{2T})} \left(\tfrac{1}{2} - \frac{\tau}{2T}\right)^{(\frac{1}{2} - \frac{\tau}{2T})} . \quad * \qquad (4.4.5)$$

Now let $u_0(,\tau)$ be any solution to 4.1.6. The error made in solving 4.4.3 will consist of two parts: a bias associated with solving a different problem and an error associated with the uncertainty in our data g (the fact that ε is non-zero).

Specifically, if $g_0 = u_0(, T)$, then

$$\begin{aligned} \|u(,\tau)-u_0(,\tau)\|_2 &= \|F_{\tau,\alpha}\, g - K_{-(T-\tau)}\, g_0\|_2 \\ &= \|F_{\tau,\alpha}(g-g_0) + (F_{\tau,\alpha} - K_{-(T-\tau)})g_0\|_2 \\ &\leqslant \|F_{\tau,\alpha}\|\ \|g-g_0\|_2 + \|(F_{\tau,\alpha} - K_{-(T-\tau)})g_0\|_2 . \end{aligned} \qquad (4.4.6)$$

The first term is bounded by $\|F_{\tau,\alpha}\|\,\varepsilon$; the second is the bias. To bound the bias, use the fact that $g_0 = K_T u_0(, 0)$ with $\|u_0(, 0)\|_2 \leqslant M$. Then

$$\begin{aligned} \|(F_{\tau,\alpha} - K_{-(T-\tau)})g_0\| &= \|[F_{\tau,\alpha} - K_{-(T-\tau)}]\, K_T u_0(, 0)\|_2 \\ &\leqslant \|F_{\tau,\alpha} K_T - K_\tau\|\ M. \end{aligned}$$

The operator $F_{\tau,\alpha} K_T - K_\tau$ is bounded and has the spectral representation defined by

----

*This bound on $\|F_{\tau;\alpha}\|$ is valid for $0 < \tau \leqslant \frac{1-\alpha}{1+\alpha}\, T$. All our τ's will be in that range.

$$(F_{\tau,\alpha}K_T - K_\tau)h = \sum_{n=1}^{\infty}\left[\frac{e^{-\lambda_n(T+\tau)}}{e^{-2\lambda_n T}+\alpha}\, e^{-\lambda_n T} - e^{-\lambda_n \tau}\right]\langle h, \varphi_n\rangle\, \varphi_n$$

$$= \sum_{n=1}^{\infty} \frac{-\alpha\; e^{-\lambda_n \tau}}{e^{-2\lambda_n T}+\alpha}\, \langle h, \varphi_n\rangle\, \varphi_n \qquad (h \in L^2[a,b])$$

so

$$\|F_{\tau,\alpha}K_T - K_\tau\| = \sup_n \frac{\alpha\; e^{-\lambda_n \tau}}{e^{-2\lambda_n T}+\alpha} \leqslant \sup_{x>0} \frac{\alpha\; e^{-x\tau}}{\alpha + e^{-2xT}}$$

$$= \alpha^{\frac{\tau}{2T}} \left(\frac{\tau}{2T}\right)^{\frac{\tau}{2T}} \left(1 - \frac{\tau}{2T}\right)^{1-\frac{\tau}{2T}} . \qquad (4.4.7)$$

So finally,

$$\|u(,\tau) - u_0(,\tau)\|_2 \leqslant \varepsilon\, \alpha^{-(\frac{1}{2}-\frac{\tau}{2T})} \left(\tfrac{1}{2} + \frac{\tau}{2T}\right)^{\frac{1}{2}+\frac{\tau}{2T}} \left(\tfrac{1}{2} - \frac{\tau}{2T}\right)^{\frac{1}{2}-\frac{\tau}{2T}}$$

$$+ M\alpha^{\frac{\tau}{2T}} \left(\frac{\tau}{2T}\right)^{\frac{\tau}{2T}} \left(1 - \frac{\tau}{2T}\right)^{1-\frac{\tau}{2T}}$$

$$= \alpha^{\frac{\tau}{2T}} \left\{ \varepsilon\, \alpha^{-\frac{1}{2}} \left(\tfrac{1}{2} + \frac{\tau}{2T}\right)^{\frac{1}{2}+\frac{\tau}{2T}} \left(\tfrac{1}{2} - \frac{\tau}{2T}\right)^{\frac{1}{2}-\frac{\tau}{2T}} + M\left(\frac{\tau}{2T}\right)^{\frac{\tau}{2T}} \left(1 - \frac{\tau}{2T}\right)^{1-\frac{\tau}{2T}} \right\} .$$

$$(4.4.8)$$

So far, no relationship has been specified between α, ε and M. A look at 4.4.8 suggests taking $\varepsilon\alpha^{-\frac{1}{2}} = M$. Then 4.4.8 becomes

$$\|u(,\tau)-u_0(,\tau)\|_2 \leq \left(\frac{\varepsilon}{M}\right)^{\frac{\tau}{T}} \cdot M \left\{ \left(\tfrac{1}{2}+\frac{\tau}{2T}\right)^{\frac{1}{2}+\frac{\tau}{2T}} \left(\tfrac{1}{2}-\frac{\tau}{2T}\right)^{\frac{1}{2}-\frac{\tau}{2T}} + \left(\frac{\tau}{2T}\right)^{\frac{\tau}{2T}} \left(1-\frac{\tau}{2T}\right)^{1-\frac{\tau}{2T}} \right\}$$

$$= \varepsilon^{\frac{\tau}{T}} M^{\frac{T-\tau}{T}} r\left(\frac{\tau}{T}\right) . \qquad (4.4.9)$$

Here

$$r(\mu) \equiv \left(\tfrac{1}{2}+\frac{\mu}{2}\right)^{\frac{1}{2}+\frac{\mu}{2}} \left(\tfrac{1}{2}-\frac{\mu}{2}\right)^{\frac{1}{2}-\frac{\mu}{2}} + \left(\frac{\mu}{2}\right)^{\frac{\mu}{2}} \left(1-\frac{\mu}{2}\right)^{\frac{1-\mu}{2}}$$

for $0 < \mu < 1$. As μ goes from 0 to $\frac{1}{2}$, $r(\mu)$ goes from $\frac{3}{2}$ to $\frac{1}{2}\, 3^{\frac{3}{4}}$ and $r(\mu)$ is symmetric about $\mu = \frac{1}{2}$.

For all τ which will be of interest to us,

$$\|u(,\tau) - u_0(,\tau)\|_2 \leq \frac{3}{2}\, \varepsilon^{\frac{\tau}{T}} M^{\frac{T-\tau}{T}} .$$

At first glance, this seems too good to be true since even were $u(,\tau)$ an exact solution to 4.1.6, it could only be guaranteed that it is within $2\varepsilon^{\frac{\tau}{T}} M^{\frac{T-\tau}{T}}$ of $u_0(,\tau)$. The explanation is as follows. The solutions to 4.1.6 are contained within a region in $L^2[a,b]$ whose diameter is $2\varepsilon^{\frac{\tau}{T}} M^{\frac{T-\tau}{T}}$. This method simply chooses a point from somewhere in the middle of that region.

4.5 Numerical Solution

The numerical problem from 4.4 is to find $u(, \tau)$ solving

$$(K_{2T} + \alpha I)u(, \tau) = K_{T+\tau} \, g \quad . \tag{4.4.3}$$

The solution has the expansion

$$u(, \tau) = \sum_{n=1}^{\infty} \frac{e^{-\lambda_n (T+\tau)} \langle g, \varphi_n \rangle}{\alpha + \varepsilon^{-2\lambda_n T}} \varphi_n \tag{4.4.2}$$

where the φ_n and λ_n refer to the Sturm-Liouville system 4.1.4 and α is ε^2/M^2 (cf. 4.1.6).

For the simple examples to be tried in testing the method, the numerical "path of least resistance" seemed to be straightforward application of 4.4.2. It will be pointed out, however, that there are cases where one should not do this. (Had this been fully appreciated at the outset, more effort would have been devoted to direct solution of 4.4.3.)

Denote the m term approximation to an L^2 function f in the eigenfunctions φ_n of 4.1.4 by f^m:

$$f^m \equiv \sum_{n=1}^{m} \langle f, \varphi_n \rangle \, \varphi_n \quad .$$

If $u_0(, \tau)$ solves 4.1.7 and $u(, \tau)$ solves 4.4.3, we need to know $\| u^m(, \tau) - u_0(, \tau) \|_2$.

$$\| u^m(, \tau) - u_0(, \tau) \|_2 = \| u^m(, \tau) - u_0^m(, \tau) + u_0^m(, \tau) - u_0(, \tau) \|_2$$

$$\leqslant \| u(, \tau) - u_0(, \tau) \|_2 + \| u_o^m(, \tau) - u_0(, \tau) \|_2 \; .$$

The first term gave rise to the estimate 4.4.9 ; the second is the truncation error. Since $u_0(,\tau)$ solves 4.1.6,

$$\|u_0^m(,\tau)-u_0(,\tau)\|_2^2 = \sum_{n=m+1}^{\infty} \langle u_0(,\tau),\varphi_n\rangle^2 = \sum_{n=m+1}^{\infty} e^{-2\lambda_n\tau} \langle u_0(,0),\varphi_n\rangle^2$$

$$\leqslant e^{-2\lambda_{m+1}\tau} M^2 .$$

The truncation error after m terms satisfies

$$\|u_0^m(,\tau)-u_0(,\tau)\|_2 \leqslant e^{-\lambda_{m+1}\tau} M .$$

In calculations performed, T was taken to be 1 as was M. The truncation error after m terms will then be smaller than error from other sources if $e^{-\lambda_{m+1}\tau} \leqslant \varepsilon^\tau$.

$$e^{-\lambda_{m+1}\tau} \leqslant \varepsilon^\tau \quad \Rightarrow \quad \lambda_{m+1} \geqslant \log\frac{1}{\varepsilon} . \qquad (4.5.1)$$

The values of ε to be considered will be no smaller than e^{-16} so we'll take $\lambda_{m+1} \geqslant 16$ as our cut-off criterion. In fact, the drill will be to compute eigenvalues and eigenfunctions until an eigenvalue as large as 16 is encountered. Then we will be taking at least one term more than is needed for 4.5.1 to be maintained and the truncation error's relative importance will be very slight.

To get a rough idea of where our cut-off point will be, an à priori estimate of the n^{th} eigenvalue is needed. Sturm-Liouville theory provides such an estimate but it must be used with some caution.

The eigenvalues come from the system

$$(p\,w_x)_x + (\lambda\rho - q)\,w = 0 \qquad (4.1.4)$$

subject to

$$\alpha_1 w(a) + \beta_1 w_x(a) = 0$$

$$\alpha_2 w(b) + \beta_2 w_x(a) = 0 \quad .$$

Defining the constant β by $\beta \equiv \int_a^b \left[\frac{\rho(x)}{p(x)}\right]^{\frac{1}{2}} dx$, the λ_n go up at least as quickly as is suggested by the asymptotic formula (for large n).

$$\lambda_n \sim \left(\frac{(n-1)\pi}{\beta}\right)^2 + O(1)\,.^{*} \qquad (4.5.2).$$

The danger in using this for small n is that the constant in O(1) may be very large. One can readily construct an example in which it bypasses any specified number. (This typically happens when $|q|$ is much larger than p and ρ.[†]) Provided q is comparable to p and ρ and we do not expect too much from 4.5.2, it seems to work fairly well in providing a rough idea of the cut-off point. It proved quite satisfactory in all the examples tried.

So one expects to cut off at that value of n for which $\left(\frac{(n-1)\pi}{\beta}\right)^2 \geqslant 16$. Call this value n_c.

Actual computations were performed on the following equations:

----

*They may go up like $(\frac{n\pi}{\beta})^2 + O(1)$ or $\left(\frac{(n-\frac{1}{2})\pi}{\beta}\right)^2 + O(1)$.

[†] See Appendix B.

$$u_t = u_{xx} - u \qquad [a,b] = [0,\pi] \qquad (4.5.3)$$

$\Rightarrow \beta = \pi$ and $n_c = 5$.

$$\frac{1}{x} u_t = (xu_x)_x + \frac{1}{x} u \qquad [a,b] = [1,2] \qquad (4.5.4)$$

$\Rightarrow \beta = \log 2$ and $n_c = 2$.

$$\frac{1}{\sqrt{x}} u_t = (\sqrt{x}\, u_x)_x - x^2 u \quad [a,b] = [1,2] \qquad (4.5.5)$$

$\Rightarrow \beta = 2\sqrt{2} - 2$ and $n_c = 2$.

$$\sec x\, u_t = (\cos x\, u_x)_x - \log(\tfrac{1}{2} + x)u \qquad [a,b] = [0, \tfrac{\pi}{4}] \qquad (4.5.6)$$

$\Rightarrow \beta = \log(1 + \sqrt{2})$ and $n_c = 3$.

It will now be apparent that one should not use the expansion 4.4.2 when β is large for then n_c will also be large and too many eigenfunctions will have to be computed. This is an expensive numerical proposition if more than a few are required.

Numerical computation of eigenvalues and eigenfunctions was performed using the methods outlined in Chapter 5 of Keller [15]. The approach for finding φ_n and λ_n solving

$$(pw_x)_x + (\lambda\rho - q)w = 0 \qquad (4.5.7)$$

subject to $\alpha_1 w(a) + \beta_1\, p(a)w'(a) = 0$ and $\alpha_2 w(b) + \beta_2\, p(b)w'(b) = 0$ is by a shooting method. Solve the initial value problem:

$$(p\, w_x)_x + (\lambda\rho - q)w = 0 \qquad (4.5.8)$$

subject to $w(a) = \beta_1 \, p(a)$

$w'(a) = -\alpha_1$

for a given value of λ. Call the solution $w(x;\lambda)$. $w(x;\lambda)$ satisfies the differential equation and left-hand boundary condition of 4. 5. 7. If λ is an eigenvalue, $\phi(\lambda) \equiv \alpha_2 w(b;\lambda) + \beta_2 p(b) w_x(b;\lambda)$ will be zero. An ingenious approach is given in Keller [15] whereby Newton's method can be used to locate the zeros of $\phi(\lambda)$. $\phi(\lambda)$ and $\phi'(\lambda)$ are both obtained in the solution of a single initial value problem. (If $\lambda_n^{(\nu)}$ denotes the ν^{th} iterate by Newton's method, the $(\nu+1)^{st}$ iterate is given by

$$\lambda_n^{(\nu+1)} = \lambda_n^{(\nu)} - \frac{\phi(\lambda_n^{(\nu)}}{\phi'(\lambda_n^{(\nu)}} \,.)$$

So each iteration requires the solution of an initial value problem.

The first guess is quite critical. Newton's method gives quadratic convergence when we are near a zero. A special initial guess routine was employed using the following algorithm. Assume that the asymptotic expansion of the n^{th} eigenvalue goes like

$$\lambda_n \quad (\frac{n\pi}{\beta})^2 + c_1 + \frac{c_2}{n^2} + \ldots\ldots *$$

1) If $n = 1$, take $(\frac{\pi}{\beta})^2$ as the initial guess.

2) If $n > 1$, compute $\lambda_{n-1}^{(f)} - (\frac{(n-1)\pi}{\beta})^2$ $(\lambda_{n-1}^{(f)}$ being the final approximation computed for $\lambda_{n-1})$. This should give roughly $c_1 + \frac{c_2}{(n-1)^2} + \ldots.$

*The boundary conditions actually used were compatible with this assumption. If others had been used, we might have needed $[(n-\frac{1}{2})\,\frac{\pi}{\beta}\,]^2$ or $[(n-1)\frac{\pi}{\beta}]^2$ as the first term in the expansion.

(if $\lambda_{n-1}^{(f)}$ is an accurate approximation to λ_{n-1}). Then take

$$\lambda_n^{(0)} = (\frac{n\pi}{\beta})^2 + \lambda_{n-1}^{(f)} - (\frac{(n-1)\pi}{\beta})^2$$

as an initial guess. We expect $\lambda_n - \lambda_n^{(0)} = \frac{c_2}{n^2} - \frac{c_2}{(n-1)^2} + \ldots = O(\frac{1}{n^3})$ so $\lambda_n^{(0)}$ becomes a good initial guess when n attains a modest size. Usually 1 or 2 iterations of Newton's method were sufficient to give adequate approximate eigenvalues and eigenfunctions. Seven decimal places of accuracy were good enough for these computations.

Also needed was a forward heat-equation solver. Some very minor modifications of the scheme presented in Keller [16] to include a variable $\rho(x)$ were required; it proved tailor-made for this job. If h and k are respectively, the space and time step sizes, this "Box Scheme" is accurate to $O(h^2+k^2)$. Richardson extrapolation was used to increase the accuracy to $O(h^4+k^4)$. Sufficient accuracy was obtained with h and k chosen to be (b-a)/100 and T/60 (T = 1).

In each numerical example the procedure was:

1) Begin with an initial profile $f = u_0(, 0)$.

2) Use the box scheme to compute $u_0(, 0.1)$; $u_0(, 0.25)$; $u(, 0.5)$; $u_0(, .667)$, $u_0(, .75)$ and $u_0(, 1.0)$.

3) Calling $u_0(, 1.0) \equiv g_0$, perturb it by an amount ε and call the result g.

4) Map g backwards via the truncated expansion of 4.4.2 obtaining $u^m(, .75)$, $u^m(, .667)$, etc.

5) Compute $\|u^m(,\tau) - u_0(,\tau)\|_2$ and compare with predicted values.

Four different standard initial profiles $u_0(, 0)$ were used for each equation. They were: a parabola concave down, a sine wave, a triangular wave and a constant. All were scaled to ensure that $\|u_0(,0)\|_2 = 1$.

It will suffice to tabulate the results of one such computation. Problem 4.5.6 is as interesting as any* having the coefficients $p(x) = \cos x$; $\rho(x) = \sec x$; $q(x) = \log(\frac{1}{2} + x)$. The boundary conditions used were $u(a) = u(b) = 0$.† Table 4.5.9 lists the results obtained with $\varepsilon = 3.0 \times 10^{-6}$ and a constant initial profile. Predicted error is $r(\tau)\,\varepsilon^{\frac{\tau}{T}}$ (cf. 4.4.9).

TABLE 4.5.9

τ	COMPUTED ERROR	PREDICTED ERROR BOUND
0.00	1.2	1.5
0.10	1.3×10^{-1}	4.2×10^{-1}
0.25	2.0×10^{-2}	6.3×10^{-2}
0.333	7.0×10^{-3}	2.2×10^{-2}
0.50	8.6×10^{-4}	2.6×10^{-3}
0.667	1.0×10^{-4}	3.1×10^{-4}
0.75	3.7×10^{-5}	1.1×10^{-4}

The numerical work does not claim completeness in any sense. No such claim would be made until, at least, the matter of what to do when β is larger than π by a significant amount was

----

*Problems 4.5.3 and 4.5.4 were used primarily to test the various subprograms. Their Sturm-Liouville problems can be solved analytically.

†This corresponds to $\alpha_1 = \alpha_2 = 1$; $\beta_1 = \beta_2 = 0$.

satisfactorily resolved. A few ideas for direct inversion of 4.4.3 have been conceived but to implement them, much thought and some consultation with numerical specialists will be required. In their present form, they do not merit expounding here.

4.6 Remark About Time-Dependent Coefficients

Other workers have interested themselves in cases where p and q are allowed to be time dependent. Agmon and Nirenberg, by a more sophisticated convexity argument (see Friedman [9], page 182) developed the much less generous estimates applicable to such cases. In particular, with $\rho \equiv 1$, constants c and m exist such that if $\mu(\tau) = \frac{e^{cT}-1}{e^{c\tau}-1}$, then two solutions $u_1(,\tau)$ and $u_2(,\tau)$ to 4.1.6 with these coefficients will satisfy, (for $0 < \tau < T$),

$$\|u_1(,\tau) - u_2(,\tau)\|_2 \leqslant 2\, e^{m\tau}\; e^{\frac{-mT}{\mu(\tau)}}\, M^{\frac{\mu(\tau)-1}{\mu(\tau)}}\, \varepsilon^{\frac{1}{\mu(\tau)}} .$$

The significant factor is $\varepsilon^{\frac{1}{\mu(\tau)}}$ which tells us what kind of convergence rate is being realized. If c is small, then $[\mu(\tau)]^{-1} \approx \frac{\tau}{T}$ yielding the approximation $\varepsilon^{\frac{1}{\mu(\tau)}} \approx e^{\frac{\tau}{T}}$, the rate obtained when p and q were time independent. The point is that, depending on the nature of the time dependence, c may be large. Suppose we can approximate $e^{c\tau} - 1 \approx e^{c\tau}$ and $e^{c\tau} - 1 \approx e^{c\tau}$. Then $\frac{1}{\mu(\tau)} \approx e^{-c(T-\tau)}$. One still has convergence by a power law but the power may be rather small $(\varepsilon^{\frac{1}{\mu(\tau)}} \approx \varepsilon^{e^{-c(T-\tau)}})$.

The attitude one will adopt in this eventuality might be to seek a method obtaining this very modest rate of convergence. But he might also decide that 4.1.6 will not do for the general time-dependent case and seek a more effective extension. That is largely a matter of individual persuasion.

4.7 Some Related Work on the Problem

As was mentioned earlier, the extension 4.1.6 has been known for some time. Not surprisingly, it has attracted the attention of theorist and computational specialist alike. A few key references will be given here but no semblance of completeness to the list is claimed. A massive bibliography appears in the excellent survey paper on the ill-posed problem methods: — Payne [19]. Many of these are relevant to this chapter.

It will already be apparent that the theoretical results needed to establish the well-posedness of 4.1.6 are known. Agmon, Nirenberg and Payne have carried logarithmic convexity much further than needed for this modest application. Besides logarithmic convexity and the spectral theory, one can exploit functional analysis to demonstrate well-posedness. This was done in Franklin [7] and Saylor [22].

Computational work on the "simplest case" $u_t = u_{xx}$ began with John [13]. It has become more popular in recent years to look at the (spatially) variable coefficients case (considered here). Buzbee and Carasso [2] tackled 4.1.2 by introducing a related fourth order boundary value problem in space-time. Their method works on the time-dependent coefficients case as well. (However, see section 4.6). Douglas [5] develops a linear programming technique employed by Cannon [3] in actual computations. Good results were claimed by both the above parties.

There have been numerous contributions, (especially to the operator theory side of ill-posed problem methods), by Russians. It has already been mentioned that Tikhonov [28] motivated the establishment of the related problem 4.4.1 (→ 4.4.3) as a means of finding an approximating algorithm for 4.1.6. Bakusinskii [1] introduces a regularized approach for a wide varity of abstract ill-posed problems on Hilbert spaces. Applying it in a straightforward manner to 4.1.2, one can be led to the analogue of 4.4.3

$$(K_{2T-2\tau} + \alpha I)\, f = K_{T-\tau}\, g \,. \qquad (4.7.1)$$

More will be said about this in Chapter 5. For the moment, it suffices to say that 4.7.1 is a well-posed operator equation in $L^2[a,b]$ which degenerates appropriately when α is set to zero.

To the best of our present knowledge, solution of 4.1.6 via 4.4.3 has not been attempted by others. To be truthful, the error estimates of Cannon [3] and Buzbee and Carasso [2] were not subjected to careful scrutiny before the computations of section 4.5 were performed. Certainly for the examples tried, there would be no embarrassment in any comparison with other methods. Quite apart from the application to this problem per se, however, this chapter introduces a means for incorporating a variety of solution set restrictions into numerical computations in a very natural way. It will make the motivation of Chapter 5, an attempt to expand Tikhonov's method, much easier. Now we have familiarity with what will prove to be a special case of the theory to be presented there.

CHAPTER 5

T-METHODS

We saw in Chapter 2 how Tikhonov incorporated a solution set constraint $\Omega^2(f) \leqslant \omega^2$ * into a numerical algorithm. In Chapter 4, the same principle was exploited to impose the constraint

$$\langle f, K_{-2\tau} f \rangle \leqslant M^2 \quad \dagger$$

on solutions to the backwards heat equation. The success of the application in Chapter 4 encourages an attempt at some generalization. The host of regularizing algorithms to be constructed in this chapter will be given the generic name "T-methods," Tikhonov's method being the model.

Before beginning, however, a comment must be made about a mathematical pitfall into which there is a very real danger of falling. One can so easily become far too enamoured with what amounts to mathematical formalism. We know that a convergent extension to $Kf_0 \approx g$ results, $(K : B_1 \|\cdot\|_1 \rightarrow B_2 \|\cdot\|_2)$, whenever f_0 is constrained to lie within a compact set (see Chapter 1). That leaves us with an impressive array of "suitable" constraints if convergence is our sole concern. However, convergence had better not be our sole concern. If there is a tendency for us to snub the practical man who asks "Why the restriction chosen as opposed to any other?", then we have forgotten or failed to appreciate a most important feature

----

*See section 2.2 for the definition of $\Omega^2(f)$.

†See section 4.4 for the definition of $K_{-2\tau}$.

of ill-posed problems. The effect of K is to cause information necessary to the approximate description of f_0 to be lost in the uncertainty associated with g. The nature of the information about f_0 put back via the constraint is every bit as important as whatever convergence properties are realized in so doing.

What follows assumes a useful constraint of a particular form has been found and then constructs an approximation for the elements f_0 satisfying $Kf_0 \approx g$ subject to said restriction.

Quite apart from the construction of new algorithms, however, this theory has another application. A method whose motivation has a different philosophical origin may turn out to have an interpretation as a T-method. Sometimes, this alternative interpretation adds to one's comprehension of the method.

5.1 The General T-method

Let $B_1 \; \|\cdot\|_1$ be a Banach space and $H_2(\cdot\,,\cdot)$ be a Hilbert space. Let $\|\cdot\|_\chi$ be a norm defined on a subspace of B_1 which satisfies a parallelogram law on its domain of definition.

Let $K : B_1 \to H_2$ be a bounded, linear, one-to-one operator whose range is dense in H_2. Consider finding f_0 in B_1 satisfying

$$\|Kf_0 - g\|_2 \leqslant \varepsilon \qquad (5.1.1)$$

(ε and γ positive numbers)

$$\|f_0\|_\chi \leqslant \gamma$$

for some g in B_2. Assume such an f_0 exists.

If $\|f_0\|_\chi \leqslant \gamma$ has compact closure in the $\|\cdot\|_1$ topology on B_1, this represents a convergent extension as $\varepsilon \to 0$ of $Kf_0 \approx g$.

Introduce a parameter α related to ε^2 by

$$C_1 \varepsilon^2 \leqslant \alpha \leqslant C_2 \varepsilon^2 \qquad (5.1.2)$$

and look for u = f in B_1 minimizing the quadratic functional

$$\|Ku-g\|_2^2 + \alpha \|u\|_\chi^2 {}^{*} . \qquad (5.1.3)$$

The f minimizing this quantity is unique. The proof given in Franklin [8] for Tikhonov's method generalizes immediately. It will be given here for completeness.

Lemma 5.1.4

The minimum to 5.1.3 occurs for a unique f in B_1.

Proof:

Suppose f_1 and f_2 both in B_1 minimize 5.1.3. A norm $\|\cdot\|$ defined on a normed space B satisfies the parallelogram law if, for all pairs u_1 and u_2 in its domain of definition,

$$\|u_1+u_2\|^2 + \|u_1-u_2\|^2 = 2(\|u_1\|^2 + \|u_2\|^2) .$$

With $\|\cdot\| = \|\cdot\|_\chi$; $B = B_1$, pick

$u_1 = \frac{1}{2} f_1$ and $u_2 = \frac{1}{2} f_2$;

then

$$\left\|\frac{f_1+f_2}{2}\right\|_\chi^2 + \left\|\frac{f_1-f_2}{2}\right\|_\chi^2 = \tfrac{1}{2}\|f_1\|_\chi^2 + \tfrac{1}{2}\|f_2\|_\chi^2 . \qquad (\bigstar)$$

----

*For u not in the domain of $\|\cdot\|_\chi$, adopt the convention $\|u\|_\chi = \infty$. Such a u will not be considered a candidate for minimizing 5.1.3.

With $\|\cdot\| = \|\cdot\|_2$; $B = H_2$, pick

$$u_1 = \tfrac{1}{2}(Kf_1 - g) \text{ and } u_2 = \tfrac{1}{2}(Kf_2 - g);$$

so

$$\left\|K\left(\frac{f_1+f_2}{2}\right)-g\right\|_2^2 + \left\|K\left(\frac{f_1-f_2}{2}\right)\right\|_2^2 = \tfrac{1}{2}\left(\|Kf_1-g\|_2^2 + \|Kf_2-g\|_2^2\right). (\bigstar\bigstar)$$

Add ★★ to α times ★ and use the hypothesis that f_1 and f_2 both minimize 5.1.3. Then

$$\left\|K\left(\frac{f_1+f_2}{2}\right)-g\right\|_2^2 + \alpha\left\|\frac{f_1+f_2}{2}\right\|_\chi^2 + \left\|K\left(\frac{f_1-f_2}{2}\right)\right\|_2^2 + \alpha\left\|\frac{f_1-f_2}{2}\right\|_\chi^2 =$$

$$= \text{the minimum value of 5.1.3.}$$

So $\frac{f_1+f_2}{2}$ also minimizes 5.1.3 and $f_1 - f_2 = 0$. □

Because f minimizes 5.1.3, it must satisfy

$$\|Kf-g\|_2^2 + \alpha\|f\|_\chi^2 \leqslant \|Kf_0-g\|_2^2 + \alpha\|f_0\|_\chi^2 \qquad (5.1.5)$$

$$\leqslant \varepsilon^2 + \alpha\gamma^2 .$$

One is led to

$$\|Kf-g\|_2^2 \leqslant \varepsilon^2(1+C_2\gamma^2) \quad \text{and} \qquad (5.1.6)$$

$$\|f\|_\chi^2 \leqslant \frac{1}{C_1} + \gamma^2 . \qquad (5.1.7)$$

So if $\|f\|_\chi^2 \leqslant$ constant has compact closure in the $\|\cdot\|_1$-topology, the minimization of 5.1.3 is a convergent extension as $\varepsilon \to 0$ of $Kf_0 \approx g$.

The machinery developed in Franklin [8] for assessing the effectiveness of Tikhonov's method on various problems $Kf_0 \approx g$ generalizes to all T-methods. Let $T_\alpha : H_2 \to B_1$ denote the solution operator to the minimization problem 5.1.3. That is

$$f \equiv T_\alpha g$$

where f minimizes 5.1.3. Make the following definitions.

Definition 5.1.8

The modulus of regularization $\rho(\varepsilon)$ will be given by

$$\rho(\varepsilon) \equiv \sup_{\substack{\|Kf_0\|_2 \leqslant \varepsilon \\ \|f_0\|_X \leqslant 1}} \|f_0\|_1 \ .$$

Definition 5.1.9

The modulus of convergence $\sigma(\varepsilon, \alpha)$ will be given by

$$\sigma(\varepsilon, \alpha) \equiv \sup_{\substack{\|Kf_0 - g\|_2 \leqslant \varepsilon \\ \|f_0\|_X \leqslant 1}} \|T_\alpha g - f_0\|_1 \ .$$

Definition 5.1.10

The rate of convergence will be the name given

$$\sup_{\substack{\|Kf_0 - g\|_2 \leqslant \varepsilon \\ \|f_0\|_X \leqslant \gamma}} \|T_\alpha g - f_0\|_1 \qquad (\gamma \text{ replacing } 1 \text{ in } 5.1.9) \ .$$

The results quoted in Chapter 2 for Tikhonov's method relating $\rho(\varepsilon)$, $\sigma(\varepsilon, \alpha)$ and the rate of convergence hold here as well. The rate of convergence in 5.1.10 is $\gamma\sigma(\frac{\varepsilon}{\gamma}, \alpha)$ and

$$\gamma\,\rho(\frac{\varepsilon}{\gamma}) \leqslant \gamma\sigma(\frac{\varepsilon}{\gamma}, \alpha) \leqslant \gamma'\rho(\frac{\varepsilon'}{\gamma'})$$

where

$$\varepsilon' = \left(1 + \sqrt{1 + \frac{\alpha\gamma^2}{\varepsilon^2}}\right)\varepsilon \quad \text{and} \quad \gamma' = \left(1 + \sqrt{1 + \frac{\varepsilon^2}{\alpha\gamma^2}}\right)\gamma \;.$$

5.2 T_H-Methods

A useful subclass of T-methods is that in which $B_1 \|\cdot\|_1 = H_1(\cdot, \cdot)_1$ is a separable Hilbert space and $\|\cdot\|_\chi$ is derived from an operator $\chi : H_1 \to H_1$. χ is to be self-adjoint, positive definite and have 0 in its continuous spectrum. Define $\|u\|_\chi$ for u in the range of χ by

$$\|u\|_\chi^2 \equiv (u, \chi^{-1}u)_1 \;. \tag{5.2.1}$$

T-methods of this sort will be called "T_H-methods" because a Hilbert space H_1 is involved.

The quadratic functional 5.1.3 to be minimized by u = f has the form

$$(Ku-g, Ku-g)_2 + \alpha(u, \chi^{-1}u)_1 \;. \tag{5.2.2}$$

K will have an adjoint operator $K^* : H_2 \to H_1$ defined by

$$(K^* g, f)_1 = (g, Kf)_2 \qquad \text{for } f \in H_1;\ g \in H_2 . \tag{5.2.3}$$

Consider perturbing f minimizing 5.2.2. by δf in Ran χ. ($\|\delta f\|_1$ is not necessarily small.) For any such δf,

$$(K(f+\delta f)-g,\ K(f+\delta f)-g)_2 + \alpha(f+\delta f,\ \chi^{-1}[f+\delta f])_1$$

$$-\{(Kf-g, Kf-g)_2 + \alpha(f, \chi^{-1}f)_1\} \geqslant 0 \ .$$

Expanding out the inner products and simplifying,

$$\Leftrightarrow \quad 2(K\delta f, Kf-g)_2 + \alpha(\delta f, \chi^{-1}f)_1 + \alpha(f, \chi^{-1}\delta f)_1$$

$$+ (K\delta f, K\delta f)_2 + (\delta f, \chi^{-1}\delta f)_1 \geqslant 0$$

$$\Leftrightarrow \quad 2(\delta f, K^*(Kf-g) + \alpha\chi^{-1}f)_1 + (K\delta f, K\delta f)_2 + (\delta f, \chi^{-1}\delta f)_1 \geqslant 0 \ .$$

A necessary and sufficient condition for this to occur, (as is seen by taking $\delta f = \pm\eta^2\ \widetilde{\delta f}$: $\|\widetilde{\delta f}\|_1 = 1;\ \eta \to 0$) is for all δf in Ran χ,

$$(\delta f, K^*(Kf-g) + \alpha\chi^{-1}f)_1 = 0 \ .$$

Ran χ has a trivial orthogonal complement for, if h $\perp$ Ran χ, then

$$h \perp \chi h \ \Rightarrow\ (h, \chi h)_1 = 0 \qquad \Rightarrow\ h = 0$$

since χ is positive definite. So our minimum f satisfies

$$K*(Kf-g) + \alpha\,\chi^{-1}f = 0$$

$$\Rightarrow \quad (K^*K + \alpha\,\chi^{-1})f = K^*g \ . \tag{5.2.4}$$

This equation's solution is unique since $K^*K + \alpha\chi^{-1}$ has a trivial null space. If no extraneities are introduced by multiplication by χ, 5.2.4 becomes

$$(\chi K^*K + \alpha I)f = \chi K^*g \ . \tag{5.2.5}$$

If $-\alpha$ is not in the spectrum of χK^*K, 5.2.5 is a well-posed operator equation for f minimizing 5.2.2. Another approach to solving 5.2.4 is to set $f = \chi h$ solving

$$(K^*K\chi + \alpha I)h = K^*g \tag{5.2.6}$$

for h.

There would have been little value in introducing T- or T_H-methods if no good means for finding f existed. The utility of a T_H-method rests on the solubility of 5.2.5 and/or 5.2.6 (numerically or otherwise).

So the solution operator T_α for the minimizing problem 5.2.2 is

$$T_\alpha = \chi(K^*K\chi + \alpha I)^{-1} K^* = (\chi K^*K + \alpha I)^{-1}\chi K^*.$$

5.3 Examples

To illustrate T and T_H methods, let us look at some examples amongst the problems encountered in previous chapters.

Example 5.3.1 Tikhonov's Method

Certainly, the foregoing must apply to the original model T-method. Making the appropriate identifications is easy in this case; it will be seen that many of the terms in Chapter 2 are given the same labels as their counterparts in this chapter.

First, let us quickly review section 2.1 in which Tikhonov's method was introduced for the Fredholm equation of the first kind:

$$\int_a^b K(x,y)f_0(y)dy = g_0(x) \tag{2.1.1}$$

where f_0 is in $B_1\ \mu(\cdot)$ (a Banach subspace of $L^2[a,b]$), g_0 is in $L^2[c,d]$; $K(x,y)$ is an L^2 kernel.

In operator notation, 2.1.1 was written

$$Kf_0 = g_0 \tag{2.1.2}$$

defining a compact operator $K : B_1 \rightarrow L^2[c,d]$. It was assumed that K had only a trivial null space.

In solving $Kf_0 \approx g$, the regularizing assumption

$$\Omega^2(f_0) \equiv \int_a^b \{p(x)\,[f_0'(x)]^2 + q(x)f_0^2(x)\}\,dx \leq \omega_1^2 \tag{2.1.3}$$

was imposed: -

$p \in C^1[a,b]$; $q \in C^0[a,b]$; $p(x) > 0$; $q(x) > 0$ on $[a,b]$.

So now, identify*: (see section 5.1)

$K = K$ (the integral operator)

$B_1\ \|\cdot\|_1 = B_1\ \mu(\cdot)$ (A family of $\mu(\cdot)$ was discussed as well as $\mu(\cdot) = \|\cdot\|_\infty$ and $\mu(\cdot) = \|\cdot\|_2$ the standard L^2-norm.)

----

*Items pertaining to Chapter 5 appear on the left of '='.

$$H_2(\cdot,\cdot)_2 = L^2\lfloor c,d\rfloor\ (\cdot,\cdot) \qquad \text{(The standard inner product on } L^2\text{).}$$

$$\|\cdot\|_\chi^2 = \Omega^2(\cdot)$$

$$\gamma = \omega_1 .$$

In the special case $B_1\ \mu(\cdot) = L^2[a,b]\ \|\cdot\|$, Tikhonov's method becomes a T_H-method for then we have $H_1(\cdot,\cdot)_1 = L^2[a,b]\ (\cdot,\cdot)$; $K : H_1 \to H_2$ as required. We wish to identify the operator

$$\chi : L^2[a,b] \to L^2[a,b]$$

for which

$$(f_0,\ \chi^{-1} f_0)_1 = \Omega^2(f_0) \ .$$

Integrate by parts.

$$\Omega^2(f_0) = \int_a^b \{p(x)\,[f_0'(x)]^2 + q(x)\ f_0^2(x)\}\,dx$$

$$= \Big[_a^b \quad p(x)f_0(x)f_0'(x) + \int_a^b f_0(x)\{q(x)f_0(x) - \frac{d}{dx}[p(x)f_0'(x)]\}\,dx \ .$$

Can a χ be found for which the following identification is legal?

$$\chi^{-1} f_0 = qf_0 - (pf_0')'$$

$$f_0(a) = f_0(b) = 0 \quad \text{for} \quad f_0 \in \text{Ran}\ \chi.$$

Since χ^{-1} is a differential operator, it is natural to look for an integral operator χ. In fact,

$$h_0 = \chi^{-1} f_0 = qf_0 - (pf_0')' \qquad (\bigstar)$$

$$f_0(a) = f_0(b) = 0 \; ,$$

means that the Green's function for the boundary value problem ★ supplies the operator χ. Call this Green's function $\chi(x, y)$. So χ is given by

$$\chi h_0 = \int_a^b \chi(x, y) h_0(y) \, dy \; .$$

χ is self-adjoint; positive definite as it should be.

Applying 5.2.5 to find f minimizing 5.2.2,

$$(\chi K^* K + \alpha I) f = \chi K^* g \; , \qquad (\bigstar\bigstar)$$

will clearly give us a Fredholm equation of the second kind - $\chi K^* K$ is an integral operator and $\chi K^* g$ is a known right-hand side. A simple choice of p and q in the regularizing assumption (like $p \equiv q \equiv 1$) will enable us to find $\chi(x, y)$ analytically. The Fredholm equation ★★ is well-posed.

Example 5.3.2 The Backwards Heat Problem of Chapter 4

In Chapter 4, the heat operator $K\tau$ was defined by

$$K_\tau f_0 = u_0(, \tau) \qquad 0 < \tau < T$$

where u_0 satisfied

$$\rho u_{0_t} = qu - (pu_x)_x \qquad \text{for} \quad 0 < t < T$$

subject to

$$\alpha_1 u(a,t) + \beta_1 u_x(a,t) = 0 \quad ,$$

$$\alpha_2 u(b,t) + \beta_2 u_x(b,t) = 0 \quad ,$$

and

$$u(;0) = f_0 \; .$$

Here p, q and ρ are positive functions on $[a,b]$ and

$$p \in C^1[a,b] \; ; \quad q \in C^0[a,b] \, ; \quad \rho \in C^0[a,b] \; .$$

An inner product $\langle \cdot , \cdot \rangle$ was defined on $L^2[a,b]$ with ρ as the weight function:

$$\langle f_1, f_2 \rangle \equiv \int_a^b f_1(x) f_2(x) \rho(x) dx \quad \text{for } f_1 ; f_2 \in L^2[a,b] \; .$$

Inverse operators to the K_t were denoted by

$$K_{-t} : \text{Ran } K_t \to L^2[a,b] \; .$$

The operators K_t are self-adjoint and sub-additive. That is

1) $K_t{}^* = K_t$

2) $K_{t_1} K_{t_2} = K_{t_1 + t_2}$.

Desired was an approximation for $u_0(,\tau)$ $0 < \tau < T$ given $u_0(T) \approx g$ and $u_0(,0)$ bounded in norm, specifically,

$$\langle K_{T-\tau}\, u_0(,\tau)-g,\ K_{T-\tau}\, u_0(,\tau)-g\rangle \leqslant \varepsilon^2$$

$$\langle K_{-\tau}\ u_0(,\tau),\ K_{-\tau}\ u_0(,\tau)\rangle \leqslant M^2 \ .$$

This is soluble via a T_H-method. Identify*:

$$K = K_{T-\tau}\ ;$$

$$H_1(\cdot\,,\cdot)_1 = L^2[a,b]\ \langle\cdot\,,\cdot\rangle\ ;\ \ H_2\ (\cdot\,,\cdot)_2 = L^2[a,b]\ \langle\cdot\,,\cdot\rangle$$

$$\|u\|_\chi^2 = \langle K_{-\tau}\ u,\ K_{-\tau}\ u\rangle = \langle u, K_{-2\tau}\ u\rangle \ .$$

So $\chi^{-1} = K_{-2\tau} \Longrightarrow \chi = K_{2\tau}$.

Applying 5.2.5 gives

$$(\chi K^* K + \alpha I)f = \chi K^* g$$

$$\Longrightarrow \qquad (K_{2\tau}\ K^*_{T-\tau}\ \ K_{T-\tau} + \alpha I)\ f = K_{2\tau}\ K^*_{T-\tau}\ g$$

$$\Longrightarrow \qquad (K_{2T} + \alpha I)\ f = K_{T+\tau}\ g$$

and we recognize 4.4.3. Recall that in Chapter 4, it was discovered by spectral analysis.

Example 5.3.3 Bakusinskii's Method

This regularizing algorithm is introduced in Bakusinskii [1] in slightly more general terms than will be considered here. Let $H_1(\cdot\,,\cdot)$ be a separable Hilbert space and $H_2(\cdot\,,\cdot)$ be a separable Hilbert space having an orthonormal basis $\{\varphi_n\}$.

----

*Items pertaining to Chapter 5 on left-hand side of '='.

Let $K : H_1 \rightarrow H_2$ be a linear, one to one, bounded operator. Assume

$$Kf_0 = g_0 \qquad f_0 \in H_1, \qquad g_0 \in H_2$$

has a solution.

The idea is to find the projection of f_0 onto the span of the elements $K^*\varphi_1, \ldots\ldots, K^*\varphi_n$. That is, find

$$f_{0_N} = \sum_{i=1}^{N} c_i K^* \varphi_i$$

where the c_i are to be determined from the equations

$$\sum_{i=1}^{N} c_i \, (K^*\varphi_i, \; K^*\varphi_j)_1 = (g_0, \varphi_j)_2 \qquad j = 1, \ldots\ldots, N \, .$$

Since this system is ill-posed with respect to perturbations in the data g_0, replace it with the regularized system:

$$\sum_{i=1}^{N} c_i [\alpha \, \delta_{ij} + (K^*\varphi_i, \; K^*\varphi_j)_1] = (g, \varphi_j) \qquad (\bigstar)$$

where $\alpha > 0$. (Note that g_0 has been replaced by g.) Denote by f_N the resulting approximation to f_{0_N}.

Let us specialize the problem to K compact. Then K^* is compact and KK^* will be a compact mapping, $(KK^* : H_2 \rightarrow H_2)$, which is self-adjoint. Choose the $\{\varphi_n\}$ to be the orthonormal eigenelements of KK^* and let $\lambda_n > 0$ be the corresponding eigenvalues. The system $\bigstar$ becomes

$$\sum_{i=1}^{N} C_i [\alpha \, \delta_{ij} + \delta_{ij} \, \lambda_j] = (g, \varphi_j)_2$$

$$\Longrightarrow \qquad C_j (\alpha + \lambda_j) = (g, \varphi_j)_2 \, .$$

So
$$f_N = \sum_{j=1}^{N} \frac{(g, \varphi_j)_2 K^* \varphi_j}{\alpha + \lambda_j} .$$

Let N → ∞ to get the approximation f for f_0 suggested by this approach:

$$f = \sum_{j=1}^{\infty} \frac{(g, \varphi_j)_2}{\alpha + \lambda_j} K^* \varphi_j .$$

This says that f satisfies the equation

$$(KK^* + \alpha I)Kf = KK^* g$$

$$\Longrightarrow \quad (K^* K + \alpha I)f = K^* g . \qquad \bigstar\bigstar$$

Taking the problem of Chapter 4, $K = K_{T-\tau}$, gives

$$(K_{2T-2\tau} + \alpha I)f = K_{T-\tau}\, g$$

as quoted in section 4.7.

Compare ★★ with 5.2.5 and observe that $\chi = I$ makes those equations identical. But $\chi = I$ is not suitable to give rise to a convergent extension of $Kf_0 \approx g$.

Indeed, the set of points f_0 satisfying

$$(f_0, I^{-1} f_0)_1 \leqslant \gamma^2 \Longleftrightarrow \|f_0\|_1 \leqslant \gamma$$

will simply be a closed ball in H_1. If H_1 is not finite dimensional, this is certainly not a compact set.

This suggests that as the error ε in the data g tends to zero, this method pulls a point from a closed ball which may wander about within it never settling down about a fixed location.

That appraisal is not entirely fair, however. The fact that the sum was truncated after a finite number of terms suggests that spectral cut-off was meant to accompany the regularizing algorithm. No connection was suggested between ε and α so none should perhaps be attached. Instead, just take α as a small parameter. Note that as $\alpha \to 0$, $Kf \to g_0$. Furthermore, for a finite value of α, the expression for f is prevented from blowing up by α's damping effect if too many terms in the sum have been taken.

Certainly in that sense, this technique will be an improvement on simple spectral cut-off. Maybe it was never envisioned as being more than that.

5.4 A Few Convergence Results for T_H-Methods

A few results will be given here which will help to decide when a T_H-method will give good convergence properties and which will facilitate error estimation in applications.

Theorem 5.4.1

Let $H(\cdot, \cdot)$ be a separable Hilbert space and $\{\varphi_n\}$ be a complete orthonormal set of elements in H. Let $\gamma > 0$ be given and $\{\lambda_n\}$ be a sequence of positive numbers tending to zero. Then the set A defined by

$$A = \left\{ f \in H \mid \sum_{n=1}^{\infty} (f, \varphi_n)^2 \frac{1}{\lambda_n} \leqslant \gamma^2 \right\}^*$$

has compact closure.

----

*The sum does not converge for all $f \in H$.

Proof:

First of all, we observe that A is bounded. Let $f \in A$,

$$\|f\|^2 = \sum_{n=1}^{\infty} (f, \varphi_n)^2 = \sum_{n=1}^{\infty} \lambda_n \frac{1}{\lambda_n} (f, \varphi_n)^2 \leqslant \gamma^2 \sup_n \lambda_n$$

and the λ_n are bounded by hypothesis.

Let $\{f_\ell\}$ be a sequence of elements in A and show that there is a convergent subsequence. The f_ℓ are bounded and thus have a weakly convergent subsequence f_{ℓ_k} (see Taylor [27], page 209). Denote the (weak) limit by f. We must show

$$f_{\ell_k} \rightharpoonup f \; ; \; \{f_{\ell_k}\} \subset A \implies f_{\ell_k} \to f \text{ as } k \to \infty \, .$$

The (weak) limit f is itself in A. For any positive integer M

$$\sum_{n=1}^{M} (f, \varphi_n)^2 \frac{1}{\lambda_n} = \lim_{k \to \infty} \sum_{n=1}^{M} (f_{\ell_k}, \varphi_n)^2 \frac{1}{\lambda_n} \leqslant \gamma^2$$

by the weak convergence of f_{ℓ_k} to f. This is true for all M so

$$\sum_{n=1}^{\infty} (f, \varphi_n)^2 \frac{1}{\lambda_n} \leqslant \gamma \implies f \in A \, .$$

Now consider the difference $\|f-f_\ell\|^2$.

$$\|f-f_{\ell_k}\|^2 = \sum_{n=1}^{N} (f-f_{\ell_k}, \varphi_n)^2 + \sum_{n=N+1}^{\infty} (f-f_{\ell_k}, \varphi_n)^2$$

$$\leqslant \sum_{n=1}^{N} (f-f_{\ell_k}, \varphi_n)^2 + \sup_{n>N} \lambda_n \sum_{n=N+1}^{\infty} \frac{1}{\lambda_n} (f-f_{\ell_k}, \varphi_n)^2$$

$$\leqslant \sum_{n=1}^{N} (f-f_{\ell_k}, \varphi_n)^2 + 2\gamma^2 \sup_{n>N} \lambda_n \, .$$

The second term can be made arbitrarily small by choosing N large enough. Having chosen N, choose k large enough that the first term is made small by virtue of $(f-f_{\ell_k}, \varphi_n) \rightarrow 0$ for each $n \leqslant N$. So $\|f-f_{\ell_k}\| \rightarrow 0$ as $k \rightarrow \infty$. □*

An immediate consequence of 5.4.1 is that if the operator χ in 5.2.1 should have the form

$$\chi f = \sum_{n=1}^{\infty} (f, \varphi_n)_1 \lambda_n \varphi_n \qquad \text{for} \quad f \in H_1; \;\; \lambda_n \text{ positive} \rightarrow 0$$

$(\{\varphi_n\}$ some complete orthonormal system in $H_1)$, then the extension 5.1.1 is convergent for any operator K of the kind discussed. The T_H-method for obtaining an approximate solution will also be convergent. In particular, one can see that χ compact, self-adjoint and positive definite will always work.

In the language of Chapter 2, Theorem 5.4.1 provides a sufficient condition that the operator K be regularized under the norm $\|\cdot\|_1$ by the assumption $(f, \chi^{-1} f) \leqslant \gamma$. In that chapter, the family of norms of the form

$$\mu^2(f) = \sum_{n=1}^{\infty} \mu_n (f, \varphi_n)^2$$

was discussed. The result 5.4.1 has a simple generalization to the norm $\mu(\cdot)$. The set A of 5.4.1 will be compact in the $\mu(\cdot)$-topology if

$$\lim_{n \rightarrow \infty} \lambda_n \mu_n = 0 \; .$$

----

*This elegant proof was suggested by Professor Franklin.

The proof of 5. 4. 1 only needs to be altered where we show that weak convergence in A implies convergence in the norm $\mu(\cdot)$.

$$\mu^2(f-f_{\ell_k}) = \sum_{n=1}^{N} \mu_n (f-f_{\ell_k}, \varphi_n)^2 + \sum_{n=N+1}^{\infty} \frac{\mu_n}{\lambda_n} \lambda_n (f-f_{\ell_k}, \varphi_n)^2$$

$$\leqslant \sum_{n=1}^{N} \mu_n (f-f_{\ell_k}, \varphi_n)^2 + 2\gamma^2 \sup_{n>N} \mu_n \lambda_n$$

and the rest of the argument is much the same.

Write $\mu_n \equiv \frac{1}{\zeta_n}$ and define an operator Φ by

$$\Phi(f) = \sum_{n=1}^{\infty} (f, \varphi_n) \frac{1}{\mu_n} \varphi_n \equiv \sum_{n=1}^{\infty} (f, \varphi_n) \zeta_n \varphi_n .$$

The norm $\mu(\cdot)$ is then generated by

$$\mu^2(f) \equiv (f, \Phi^{-1} f) \quad . \tag{5. 4. 2}$$

Boundedness in the norm topology generated by χ yields compactness in the norm topology generated by Φ when the eigenvalues λ_n of χ go down faster than those ζ_n of Φ, (Φ and χ having the same eigenelements φ_n corresponding to ζ_n and λ_n respectively).

When the foregoing is put together with Theorem 1. 4. 2, the following convergence result for T_H-methods is obtained.

Theorem 5. 4. 3

Let H_1 $(\cdot, \cdot)_1$ and H_2 $(\cdot, \cdot)_2$ be separable Hilbert spaces and let $K : H_1 \rightarrow H_2$ be a bounded, one-to-one operator whose range is dense in H_2. Let $\{\varphi_n\}$ be an orthonormal basis for H_1 and norms $\|\cdot\|_\chi$ and $\mu(\cdot)$ be defined on subspaces of H_1 by

$$\|f\|_\chi^2 = \sum_{n=1}^{\infty} \frac{1}{\lambda_n} (f, \varphi_n)^2 \quad \text{and} \quad \mu^2(f) = \sum_{n=1}^{\infty} \frac{1}{\zeta_n} (f, \varphi_n)^2$$

where $\lambda_n;\zeta_n > 0$ for all n. A sufficient condition that the operator K be regularized under the norm μ by the condition $(f, \chi^{-1}f) \leqslant \gamma^2$ is that

$$\lim_{n\to\infty} \frac{\lambda_n}{\zeta_n} = 0 .$$

Having obtained an idea of when regularization is obtained and hence solution via a T_H-method is feasible, it would be useful to get a feeling for what sort of convergence rates will be obtained. Observe that, in 5.4.3, a detailed description of K is not needed. The details of K's structure within the broad limitations imposed by 5.4.3 will not affect whether or not convergence occurs but have a great deal to do with the rate of convergence. Knowing that the modulus of regularization (see section 5.2) is of great interest in this respect , a few cases will be considered wherein bounds upon it can be imposed. Only regularization under the norm $\|\cdot\|_1$ will be treated. At first glance, all situations will seem rather special. Reflection on the experiences of past chapters, however, shows them to occur rather frequently in practice.

Theorem 5.4.4

Let $H_1(\cdot,\cdot)_1$ and $H_2(\cdot,\cdot)_2$ be separable Hilbert spaces and let $K : H_1 \to H_2$ be one to one, bounded, linear and compact. Let the eigenvalues of K^*K be denoted $\{\lambda_n\}$ and its corresponding orthonormal eigenelements by $\{\varphi_n\}$. Let the norm $\|\cdot\|_\chi$ be defined by

$$\|f\|_\chi^2 = \sum_{n=1}^{\infty} \frac{1}{\eta_n} (f, \varphi_n)^2,^* \qquad \text{for} \quad f \in H_1$$

----

* $\|f\|_\chi = \infty$ if the sum does not converge.

where η_n is a positive sequence; $\eta_n \to 0$ as $n \to \infty$. Assume the λ_n are monotone non-increasing. Recall that $\rho(\varepsilon)$ is defined by

$$\rho(\varepsilon) \equiv \sup_{\substack{\|Kf\|_2 \leqslant \varepsilon \\ \|f\|_\chi \leqslant 1}} \|f\|_1 = \sup_{\substack{(f, K^*Kf) \leqslant \varepsilon^2 \\ \|f\|_\chi^2 \leqslant 1}} \|f\|_1 \quad .$$

a) If there exist positive constants C and p such that

$$\eta_n \leqslant C\lambda_n^p \quad \text{for} \quad n \geqslant n_0 ,$$

then

$$\rho(\varepsilon) \leqslant \varepsilon^{\frac{1}{p+1}} C^{\frac{1}{2(p+1)}} \left[p^{\frac{1}{p+1}} + p^{-\frac{p}{p+1}} \right]^{\frac{1}{2}}$$

$$\text{for } \varepsilon \leqslant Cp\,\lambda_{n_0}^{\frac{p+1}{2}} \quad .$$

b) If there exist positive constants C and p such that

$$\eta_n \geqslant C\lambda_n^p \quad \text{for} \quad n \geqslant n_0 ,$$

then there exists a sequence of ε values tending to zero for which $\rho(\varepsilon)$ has the lower bound:

$$\rho(\varepsilon) \geqslant C^{\frac{1}{2(p+1)}} \varepsilon^{\frac{p}{p+1}} .$$

c) If there exist positive constants C and p and R > 1 such that

$$\lambda_n \leqslant R^{-[\frac{C}{\eta_n}]^p} \qquad \text{for } n \geqslant n_0 ,$$

convergence is (at best) logarithmic.

d) If there exist positive constants C and p and R > 1 such that

$$\lambda_n \geqslant R^{-[\frac{C}{\eta_n}]^p} \quad \text{for } n \geqslant n_0 \;,$$

convergence is (at worst) logarithmic.

Proof of a) and c)*

a)
$$\|f\|_1^2 = \sum_{n=1}^{N} \frac{\lambda_n}{\lambda_n} (f, \varphi_n)_1^2 + \sum_{n=N+1}^{\infty} \frac{\eta_n}{\eta_n} (f, \varphi_n)_1^2$$

$$\leqslant \frac{1}{\lambda_N} \sum_{n=1}^{\infty} \lambda_n (f, \varphi_n)_1^2 + \sup_{n > N} \eta_n \sum_{n=1}^{\infty} \frac{1}{\eta_n} (f, \varphi_n)_1^2$$

If $\|Kf\|_2 \leqslant \varepsilon$ and $\|f\|_\chi^2 \leqslant 1$, then for all $N \geqslant n_0$,

$$\|f\|_1^2 \leqslant \frac{1}{\lambda_N} \varepsilon^2 + \sup_{n > N} \eta_n \leqslant \frac{1}{\lambda_N} \varepsilon^2 + C \lambda_{N+1}^p \qquad (\bigstar)$$

(by hypothesis). If $\varepsilon \leqslant C p \lambda_{n_0}^{\frac{p+1}{2}}$, there exists $N \geqslant n_0$ such that

$$\lambda_{N+1} \leqslant \left[\frac{\varepsilon^2}{Cp}\right]^{\frac{1}{p+1}} \leqslant \lambda_N \,.$$

Substitute into ★ and the result follows.

c) It will be shown that an infinite sequence of values of ε tending to zero have $\rho(\varepsilon)$ bounded below by a logarithmic function of ε.

$$\text{Let } f_m \equiv \frac{\varepsilon}{\sqrt{\lambda_m}} \;.$$

Then

$$\|Kf_m\|_2^2 = \varepsilon^2 \;; \quad \|f_m\|_\chi^2 = \frac{1}{\eta_m} \frac{\varepsilon^2}{\lambda_m} \;; \quad \|f_m\|_1 = \frac{\varepsilon}{\sqrt{\lambda_m}} \;.$$

----

*The same ideas are involved in proving 2) and 4) left out for brevity.

Let us maximize $\frac{\varepsilon}{\sqrt{\lambda_m}}$ subject to

$$\frac{1}{\eta_m}\frac{\varepsilon^2}{\lambda_m} \leqslant 1.$$

Fix attention on those ε for which $\varepsilon^2 = \lambda_m \eta_m$ for some m. There are infinitely many such ε and they do define a sequence tending to zero as $m \to \infty$. Then

$$\frac{\varepsilon}{\sqrt{\lambda_m}} = \sqrt{\eta_m} = \|f_m\|_1 \quad .$$

For such an ε, $\sqrt{\eta_m}$ will be a lower bound for $\rho(\varepsilon)$. η_m can be bounded above under our hypothesis about the λ_n.

$$\varepsilon^2 = \lambda_m \eta_m \leqslant \eta_m R^{-[\frac{C}{\eta_m}]^p}$$

For C, R, p as specified, the function

$$z(x) \equiv x R^{-(\frac{C}{x})^p}$$

rises monotonically from 0 to ∞ as x goes from 0 to ∞.

There is thus one root $x_0(\varepsilon)$ to

$$\varepsilon^2 = x R^{-(\frac{C}{x})^p}$$

and $x_0(\varepsilon)$ will be $\leqslant \eta_m$. An asymptotic formula for $x_0(\varepsilon)$ valid for small ε can be found.

$$2\log\varepsilon = \log x_0 - (\frac{C}{x_0})^p \log R \sim -(\frac{C}{x_0})^p \log R$$

$$\Rightarrow \quad x_0 = C\left(\frac{-\log R}{2\log\varepsilon}\right)^{\frac{1}{p}} + o\left(-\frac{1}{\log\varepsilon}\right)^{\frac{1}{p}} \quad .$$

So for sufficiently small ε which happens to be $\lambda_m \eta_m$ for some m,

$$\rho(\varepsilon) \geqslant\sim C\left(\frac{\log R}{-2\log\varepsilon}\right)^{\frac{1}{p}} \quad . \quad \square$$

Every rate of convergence discussed in Chapters 2 and 4 could have been speedily estimated using 5. 4. 4-a) and 5. 4. 4-b). The statement to the effect that Tikhonov's method gets logarithmic convergence on the backwards heat equation follows from 5. 4. 4-c) and 5. 4. 4-d).

A consequence of 5. 4. 4-a) is that if η_n goes down faster than any power of λ_n, then for small enough ε, $\rho(\varepsilon)$ tends towards linearity. It goes without saying, however, that ε may have to be unrealistically small (from the point of view of numerical computation) before anything like linear behaviour prevails.

One is not always able to apply 5. 4. 4 conveniently - or indeed at all. However, the following estimate is always available:

$$\rho^2(\varepsilon) \leqslant \inf_N \left[\varepsilon^2 \sup_{n \leqslant N} \frac{1}{\lambda_n} + \sup_{n > N} \eta_n\right] \quad . \qquad (5.4.5)$$

For ε's of sizes encountered in computations, direct use of 5. 4. 5 will often be, by far, the best way to find $\rho(\varepsilon)$. The λ_n and η_n need not be monotone although lack of monotonicity will make the estimation more difficult.

CHAPTER 6

SOME OBSERVATIONS ON A STATISTICAL METHOD

A few approaches for dealing with linear, ill-posed problems of the form $Kf_0 \approx g$ have now been examined. As we have seen, the K's of interest to us suppress crucial information about f_0 in an irrecoverable fashion; successful extension entails, among other things, a replacement of missing information.

The methods studied so far seek to accomplish this by placing very definite restrictions on f_0, forcing f_0 to lie within some admissibility class if it is to be considered an allowable solution. Let us call these "D-methods."*

Suppose, however, that our physical application is such that we have access to a wealth of statistical data about the solutions to be encountered and the errors associated with our knowledge of g. Then we might well seek to give a statistical estimate of f_0, replacing the information suppressed by K statistically. A classic case is that of mathematical weather prediction† where the accumulated records compiled over the years by meteorologists are available. Many problems involving interpretation of distorted signals also suggest this approach, planetary radar and photographic image enhancement being two suitable areas of application.

The idea of applying statistical methods to ill-posed problems goes back at least to Sudakov and Kalfin [26] (1957). Lavrentiev [18]

----

*"D" is for deterministic. There is no possibility admitted that f_0 might lie outside the admissibility class.

†See Courant and Hilbert [4], page 231.

mentions their work and suggests a statistical approach of his own. He points out how D-methods can be thought of as limiting cases wherein the probability that the solution lies in some set goes to 1. This set would be identified as the D-method's admissibility class. Strand and Westwater [25] worked on statistical extension of Fredholm integral equations of the first kind. There are several meteorological references in their bibliography.

Many of the earlier workers in the field discretized their problems by suitable quadrature techniques as a preliminary step to their analysis. That made the problem finite dimensional, the ill-posed operator K typically replaced by an ill-conditioned matrix.* A statistical approach which addresses itself directly to the original problem will require the notion of a random process over an infinite dimensional space. (See Gelfand [11] and Lavrentiev [18].)

The method of Franklin [6], developed for direct extension of problems on Hilbert spaces will be the subject of this chapter. As far as is possible, discussion of statistics will be avoided as the work of previous chapters lends itself to the appraisal of D-methods only. An interesting fact to be brought to light is that a T_H-method is readily constructible whose solution is the estimate suggested by Franklin's method. On its own, the T_H-method would be unmotivated mathematical extension but, in this context, it becomes somewhat more respectable. Its usefulness lies in the alternative viewpoint it affords of the statistical estimate. What we know about convergence

----

*See Conclusions for further discussion of this point.

in T_{11}-methods can be invoked to account for certain stability features in Franklin's method which are not obvious from the statistical standpoint. In section 6.4, the problem of harmonic continuation on which computations were performed by Franklin will be discussed in an illustrative capacity.

6.1 Franklin's Method

For the purposes of this discussion, a very sketchy outline of the method will suffice. (Anyone interested in pursuing the details is referred to Franklin [6].)

Let $H_1(\cdot\,,\cdot)_1$ and $H_2(\cdot\,,\cdot)_2$ be Hilbert spaces and $K : H_1 \rightarrow H_2$ be bounded, linear, and one to one. The equation $Kf_0 \approx g$ is replaced by

$$Kf_0 + n = g \tag{6.1.1}$$

$$f_0 \in H_1; \quad n \in H_2; \quad g \in H_2 \,.$$

Having carefully defined random processes over Hilbert spaces and their correlation operators, Franklin views f_0, n and g as samples from processes referred to respectively as the signal, noise and data processes.

Labeling these processes P_1, P_2 and P_3, one relates them by

$$Ku_1 + u_2 = u_3 \tag{6.1.2}$$

(u_3 = g would represent a particular sample from the data process P_3.) Sought is a best linear estimate (in a well-defined sense) of the signal u_1 in terms of the data u_3. It is assumed that the autocorrelation operators R_{11} and R_{22} (of the signal and noise processes respectively) are known as is their cross-correlation operator R_{12}.

The solution to this best linear estimate problem is found. For a sample $u_3 = g$ from the data process, u_1 should be estimated by

$$f = (R_{11}K^{*}+R_{12})\,(KR_{11}K^{*}+KR_{12}+R_{12}^{*}K^{*}+R_{22})^{-1}g. \qquad (6.1.3)$$

The correlation operators have the following properties:

a) $R_{ij} : H_i \rightarrow H_j$ are bounded. $i, j = 1, 2.$ (6.1.4)

b) $R_{ij} = R_{ji}^{*}$. $i, j = 1, 2.$

c) $(h_i, R_{ii}h_i)_i > 0$ for $0 \neq h_i \in H_i$. $i = 1, 2.$

In actual computations, further restrictions are usually imposed on the operators R_{ij}. In effect, assumptions about the signal and noise processes are made which are reasonable for most applications and which make the implementation easier. Specifically,

a) $R_{12} \equiv 0$ (signal and noise are uncorrelated) (6.1.5)

b) $R_{22} = \nu^2 I$ (where ν is a small parameter).

6.1.5-b) is referred to as the "white noise condition," ν being the "white noise amplitude."

Under these assumptions, 6.1.3 becomes

$$f = R_{11}K^{*}\,(KR_{11}K^{*} + \nu^2 I)^{-1}\,g. \qquad (6.1.6)$$

6.2 An Equivalent T_H-Method

At this point, the question confronting us is, "How can 6.1.6 be interpreted as an attempt to find an approximation f to f_0 satisfying

$$\|Kf_0 - g\|_2 \leqslant \varepsilon \tag{6.2.1}$$

$$(f_0, \chi^{-1} f_0)_1 \leqslant \gamma^2$$

via the minimization of

$$\|Kf-g\|_2^2 + \alpha(f, \chi^{-1} f)_1 \text{ ?"} \tag{6.2.2}$$

Here $\chi : H_1 \rightarrow H_1$ it to be a positive definite, bounded, self-adjoint linear operator. It is also required that α be related to ε by

$$C_1 \varepsilon^2 \leqslant \alpha \leqslant C_2 \varepsilon^2 \tag{6.2.3}$$

for positive constants C_1 and C_2.

Answering this question entails identifying χ, ε, α and γ, within the framework just outlined, in terms of R_{11} and ν in 6.1.6. We know from Chapter 5 that the solution to the minimization problem 6.2.2 will be given by

$$f = (\chi K^* K + \alpha I)^{-1} \chi K^* g \tag{6.2.4}$$

as a consequence of 5.2.5.

Let us therefore try to arrange that

$$R_{11}\,K^*(KR_{11}K^* + \nu^2 I)^{-1} = (\chi K^*K + \alpha I)^{-1}\,\chi K^*$$

$$\Longleftrightarrow \quad (\chi K^*K + \alpha I)\,R_{11}K^* = \chi K^*(KR_{11}K^* + \nu^2 I)$$

$$\Longleftrightarrow \quad \alpha R_{11}K^* = \nu^2 \chi K^* \ .$$

$$\Longleftrightarrow \quad \alpha\, R_{11} = \nu^2 \chi$$

So choose α and χ:

$$\alpha = \nu^2; \quad \chi = R_{11} \ . \tag{6.2.5}$$

The operators K and χ, (now fixed), determine the modulus of regularization $\rho(\varepsilon)$. When ε, γ and α have been specified, the difference between f(given by 6.2.4) and any f_0 satisfying the conditions of 6.2.1 is bounded by the rate of convergence $\gamma\,\sigma(\frac{\varepsilon}{\gamma}, \alpha)$. This is related to the modulus of regularization by the inequalities:

$$\gamma\,\rho(\tfrac{\varepsilon}{\gamma}) \leq \gamma\,\sigma(\tfrac{\varepsilon}{\gamma}, \alpha) \leq \gamma'\rho(\tfrac{\varepsilon'}{\gamma'}) \tag{6.2.6}$$

where (see section 5.1)

$$\varepsilon' = \left(1 + \sqrt{1 + \frac{\alpha\gamma^2}{\varepsilon^2}}\,\right)\varepsilon \quad \text{and} \quad \gamma' = \left(1 + \sqrt{1 + \frac{\varepsilon^2}{\alpha\gamma^2}}\,\right)\gamma \ . \tag{6.2.7}$$

There is still considerable leeway in how ε and γ must be specified. In fact, there are many problems 6.2.1 whose solutions by 6.2.4 with $\alpha = \nu^2$ would lead to the same approximation f. In identifying a companion for Franklin's method, choose one giving good error estimates in terms of ν. What interest error estimates are in understanding a statistical method has not yet been made clear - that being the subject of section 6.3 - for the time being, it will suffice to note that choosing

$$\gamma = \gamma_0 \; ; \qquad \varepsilon = \gamma_0 \nu \; ; \qquad \alpha = \nu^2 \tag{6.2.8}$$

is legitimate and causes 6.2.6 (in view of 6.2.7) to become

$$\gamma_0 \, \rho(\nu) \leqslant \gamma_0 \sigma(\nu, \nu^2) \leqslant \gamma_0 (1 + \sqrt{2}) \, \rho(\nu) \, . \tag{6.2.9}$$

At first glance, it appears, somewhat alarmingly, that γ_0 is a free parameter and that the error estimate $\gamma_0 \sigma(\nu, \nu^2)$ can be made arbitrarily small. This is not the case, however, since it was assumed in deriving the relationships 6.2.6 that f_0 existed satisfying 6.2.1. Taking γ_0 too small will cause the sets

a) $$\{ f_0 \in H_1 \mid \|Kf_0 - g\|_2 \leqslant \nu \gamma_0 \} \tag{6.2.10}$$

and

b) $$\{ f_0 \in H_1 \mid (f_0, \chi^{-1} f_0)_1 \leqslant \gamma_0^2 \}$$

to become disjoint. For the estimate 6.2.9 to be valid, γ_0 must be chosen larger than some value $\gamma_{\min}(\nu, g)$. If g is in the range of $K\chi$, a value of γ_0 can be found which is applicable for all values of ν. Indeed, if

$$g = K\chi h \qquad \text{for } h \in H_1$$

$$\gamma_0 = (h, \chi h) \; ; \quad f_0 = \chi h$$

satisfies

$$\|Kf_0 - g\|_2 = 0 \leqslant \nu \gamma_0 \qquad \text{for all } \nu$$

and

$$(f_0, \chi^{-1} f_0) = \gamma_0^2 .$$

So, if $g \in \operatorname{Ran} K\chi$, there is always at least one choice of f_0 satisfying the two constraints 6.2.10 whatever the value of ν.

If g is not in the range of $K\chi$, it is indeed possible for $\gamma_{\min}(\nu, g)$ to approach infinity as $\nu \rightarrow 0$. Fortunately, for practical purposes, this pathology can never make itself felt. Our ε's can never be smaller than the limitations imposed by the machine representation of g. The actual machine description of the data always refers equally well to any of a collection of points in the abstract Hilbert space some of which will be in the range of $K\chi$. This entitles us to say "Without loss of generality, assume g is in the range of $K\chi$."

In what is to follow, an intuitive understanding of certain qualitative phenomena is all that is desired; the estimates to be made will be very liberal. It will be assumed throughout that a modest value of γ_0 (order 1) exists for which Franklin's method can be viewed as equivalent to a T_H-method, the rate of convergence satisfying 6.2.9 for any ν of computational relevance.

6.3 Insensitivity of Estimate to Certain Statistical Assumptions

Franklin observed in his numerical computations that the solution f to 6.1.6 was rather insensitive to the statistical assumptions made. In particular, the white noise amplitude ν could be varied over several orders of magnitude without notably affecting f. We are now in a position to explain this.

The white noise amplitude has now been identified as a small convergence parameter. For each value of ν, an element f_ν is associated by way of 6.1.6. As $\nu \rightarrow 0$, f_ν approaches some limit. When ν is small enough, f_ν is quite near the limit and making it smaller will not have much effect.

Let $\nu_1 > \nu_2$ be distinct values of ν corresponding to f_{ν_1} and f_{ν_2}. Let f_0 satisfy

$$\|Kf_0 - g\| \leqslant \nu_2\gamma_0 < \nu_1\gamma_0 \tag{6.3.1}$$

$$(f_0, R_{11}^{-1} f_0) \leqslant \gamma_0^2$$

(γ_0 as discussed in section 6.2). Then by 6.2.9,

$$\|f_{\nu_1} - f_{\nu_2}\|_1 \leqslant \|f_{\nu_1} - f_0\|_1 + \|f_0 - f_{\nu_0}\|_1$$

$$\leqslant (1 + \sqrt{2})\,\gamma_0\,[\rho(\nu_1) + \rho(\nu_2)] \quad . \tag{6.3.2}$$

This is a very crude estimate, the distance between f_{ν_1} and f_{ν_2} being bounded by the sum of their distances to a third (pessimistically located) point. Nonetheless, 6.3.2 does tell us that stability of the kind mentioned will be observed in computations if the regularizing effect of R_{11} on K yields a good modulus of regularization.

At this point, it would be desirable to see what sort of convergence occurs with the R_{11}'s and K's typically used. In the next section, we shall delve into the computational example actually worked by Franklin and check this out.

6.4 The Numerical Example of Harmonic Continuation

The ordinary Dirichlet problem for the unit circle is to find $u(r,\theta)$ satisfying

$$\Delta u = 0 \qquad \text{for} \qquad 0 < r < 1 \; ; \quad 0 \leqslant \theta < 2\pi \tag{6.4.1}$$

subject to $u(1,\theta) = f_0(\theta)$.

The solution is given by the well-known Poisson formula:

$$u(r,\theta) = \frac{1}{2\pi} \int_0^{2\pi} \frac{(1-r^2)\, f_0(\varphi)}{1-2r\cos(\theta-\varphi)+r^2}\, d\varphi \; . \tag{6.4.2}$$

Suppose we were given

$$g_0(\theta) \equiv u(\rho,\theta)$$

on some interior circle of radius $\rho < 1$ and asked to recover f_0. That is, find $f_0(\varphi)$ satisfying

$$g_0(\theta) = \frac{1}{2\pi} \int_0^{2\pi} \frac{(1-\rho^2)\, f_0(\varphi)}{1-2\rho\cos(\theta-\varphi)+\rho^2}\, d\varphi \; . \tag{6.4.3}$$

This is a Fredholm integral equation of the first kind.

Let us write this in operator notation:

$$Kf_0 = g_0 \; . \tag{6.4.4}$$

K, regarded as a mapping from $L^2[0,2\pi]$ into $L^2[0,2\pi]$, is self-adjoint, compact; has the eigenvalues ρ^n, (n a non-negative integer), and the orthonormal eigenfunctions:

$\frac{1}{\sqrt{2\pi}}$ corresponding to the eigenvalue $\rho^0 = 1 \equiv \sqrt{\lambda_1}$ *

$\frac{\cos n\theta}{\sqrt{\pi}}$ corresponding to ρ^n for $n \geqslant 1 \equiv \sqrt{\lambda_{2n}}$

$\frac{\sin n\theta}{\sqrt{\pi}}$ corresponding to ρ^n for $n \geqslant 1 \equiv \sqrt{\lambda_{2n+1}}$

K will thus have the spectral decomposition:

$$Kf = (f, \frac{1}{\sqrt{2\pi}})\frac{1}{\sqrt{2\pi}} + \sum_{n=1}^{\infty}\left[\rho^n(f, \frac{\cos n\theta}{\sqrt{\pi}})\frac{\cos n\theta}{\sqrt{\pi}} + \rho^n(f, \frac{\sin n\theta}{\sqrt{\pi}})\frac{\sin n\theta}{\sqrt{\pi}}\right]. \tag{6.4.5}$$

(All these facts about K are easily discovered when separation of variables in r and θ is performed in solving 6.4.1).

In passing, let us see what convergence Tikhonov's method would achieve. Take the simplest regularizing functional:

$$\Omega^2(f) = \int_0^{2\pi} \{f^2(\theta) + [f'(\theta)]^2\}\, dx = (f, f) + (f', f')$$

If f has the Fourier series:

$$f \sim (f, \sqrt{\frac{1}{2\pi}})\sqrt{\frac{1}{2\pi}} + \sum_{n=1}^{\infty}\left[(f, \frac{\cos n\theta}{\sqrt{\pi}})\frac{\cos n\theta}{\sqrt{\pi}} + (f, \frac{\sin n\theta}{\sqrt{\pi}})\frac{\sin n\theta}{\sqrt{\pi}}\right]$$

then

$$\Omega^2(f) = (f, \sqrt{\frac{1}{2\pi}})^2 + \sum_{n=1}^{\infty}\left[(f, \frac{\cos n\theta}{\sqrt{\pi}})^2 (n^2+1) + (f, \frac{\sin n\theta}{\sqrt{\pi}}) (n^2+1)\right]$$

$$= (f, \chi^{-1} f)$$

----

*Recall that λ_n was the designation in Theorem 5.4.4 of the eigenvalues of $K^*K = K^2$ (in this case).

where χ has the spectral representation:

$$\chi f = (f, \frac{1}{\sqrt{2\pi}}) \frac{1}{\sqrt{2\pi}} + \sum_{n=1}^{\infty} \frac{1}{1+n^2} (f, \frac{\cos n\theta}{\sqrt{\pi}}) \frac{\cos n\theta}{\sqrt{\pi}} + \frac{1}{1+n^2} (f, \frac{\sin n\theta}{\sqrt{\pi}}) \frac{\sin n\theta}{\sqrt{\pi}} \qquad (6.4.6)$$

The eigenfunctions of χ are:

$\frac{1}{\sqrt{2\pi}}$ corresponding to the eigenvalue 1 $\quad (\equiv \eta_1)$ *

$\frac{\cos n\theta}{\sqrt{\pi}}$ corresponding to $\frac{1}{1+n^2} \quad \eta \geqslant 1 \quad (\equiv \eta_{2n})$

$\frac{\sin n\theta}{\sqrt{\pi}}$ corresponding to $\frac{1}{1+n^2} \quad \eta \geqslant 1 \quad (\equiv \eta_{2n+1})$

$K^*K = K^2$ has the same eigenfunctions as χ, the eigenvalues λ_k of K^*K are monotone non-increasing. Thus 5.4.4 is applicable. The eigenvalues to be compared are λ_k and η_k identified here by

$$\lambda_1 = 1;\ \lambda_{2n} = \rho^{2n};\ \lambda_{2n+1} = \rho^{2n} \qquad n \geqslant 1$$

$$\eta_1 = 1;\ \eta_{2n} = \frac{1}{1+n^2};\ \eta_{2n+1} = \frac{1}{1+n^2} \qquad n \geqslant 1$$

Denote $\rho \equiv \frac{1}{R}$; $R > 1$.

Let $C = 2$; $p = \frac{1}{2}$.

$$\lambda_{2n} = R^{-2n} = R^{-[2Cn^2]^p} \leqslant R^{-[C(n^2+1)]^p} = R^{-[\frac{C}{\eta_{2n}}]^p}.$$

Now let $C = 1$; $p = 1$.

$$\lambda_{2n} = R^{-2n} \geqslant R^{-(n^2+1)} = R^{-[\frac{1}{\eta_{2n}}]} = R^{-[\frac{C}{\eta_{2n}}]^p}.$$

----

*Recall that the η_k are the eigenvalues of χ defining $\|\cdot\|_\chi$ in 5.4.4.

The same calculations hold with λ_{2n} and η_{2n} replaced by λ_{2n+1} and η_{2n+1}. By Theorem 5.4.4 (parts c) and d)), convergence will be logarithmic if Tikhonov's method is used. This observation was made in Franklin [ ]. Note how slowly the η_k go down as compared with the λ_k. This is in marked contrast to what we shall observe when the $\chi = R_{11}$ of Franklin's computation is taken as the regularizer.

The operator R_{11} used was defined by

$$R_{11}f(\theta) = \int_0^{2\pi} \alpha \exp\{-\beta \sin^2 [\tfrac{1}{2}(\theta-\varphi)]\} f(\varphi)\, d\varphi \qquad (6.4.7)$$

which is the convolution h*f of f with the function h:

$$h(\varphi) \equiv \alpha \exp\{-\beta \sin^2 \tfrac{\varphi}{2}\} \quad . \qquad (6.4.8)$$

The parameters α and β were related to how large and how oscillatory the anticipated solutions would be on the average. The "size" and "roughness" associated with R_{11} were defined* and shown to be equal to $\sqrt{\alpha}$ and $\sqrt{\frac{\beta}{2}}$ respectively. Computations were performed with size = 1 and roughness = 1, 2, 5 and 7. So $\alpha = 1$ and β ranges in value from 2 to 98. ρ was chosen to be $\frac{1}{2}$.

To get the Fourier series for $R_{11}f$ from f and hence obtain the spectral decomposition of R_{11}, use the convolution theorem for Fourier series. That is, if f has the series

----

*Size and roughness will not be defined for this non-statistical discussion.

$$f \sim f_0 \frac{1}{\sqrt{2\pi}} + \sum_{n=1}^{\infty} \left[f_n \frac{\cos n\theta}{\sqrt{\pi}} + f_n{}' \frac{\sin n\theta}{\sqrt{\pi}} \right]$$

and

$$h \sim h_0 \frac{1}{\sqrt{2\pi}} + \sum_{n=1}^{\infty} \left[h_n \frac{\cos n\theta}{\sqrt{\pi}} + h_n{}' \frac{\sin n\theta}{\sqrt{\pi}} \right]$$

then h*f has the series:

$$h{*}f \sim \sqrt{2\pi}\, f_0 h_0 \frac{1}{\sqrt{2\pi}} + \sum_{n=1}^{\infty} \left[\sqrt{\pi}\,(f_n h_n - f_n{}' h_n{}') \frac{\cos n\theta}{\sqrt{\pi}} + \sqrt{\pi}\,(f_n' h_n + f_n h_n') \frac{\sin n\theta}{\sqrt{\pi}} \right].$$

So all that is needed to complete the spectral representation of R_{11} is the Fourier expansion of h. The h_n' are zero since $h(\varphi)$ is 2π periodic in φ and is even when extended to negative values. To get the h_n, evaluate

$$\int_0^{2\pi} \cos n\varphi\, \alpha \exp[\beta \sin^2 \tfrac{\varphi}{2}]\, d\varphi = \alpha \int_0^{2\pi} \cos n\varphi \exp[-\beta(\tfrac{1-\cos\varphi}{2})]\, d\varphi$$

$$= 2\alpha\, e^{-\frac{\beta}{2}} \pi \frac{1}{\pi} \int_0^{\pi} e^{\frac{\beta}{2}\cos\varphi} \cos n\varphi\, d\varphi = 2\pi\alpha\, e^{-\frac{\beta}{2}} I_n(\tfrac{\beta}{2}) .$$

(I_n is the n^{th} order modified Bessel function.)

So $h_0 = \sqrt{2\pi}\ \alpha\ e^{-\frac{\beta}{2}} I_0(\frac{\beta}{2}) \qquad h_n = 2\sqrt{\pi}\ \alpha\ e^{-\frac{\beta}{2}} I_n(\frac{\beta}{2})$.

So R_{11} has the form:

$$R_{11}f(\theta) = 2\pi\alpha e^{-\frac{\beta}{2}} I_0(\tfrac{\beta}{2}) (f, \tfrac{1}{\sqrt{2\pi}}) \frac{1}{\sqrt{2\pi}} + \sum_{n=1}^{\infty} 2\pi\alpha e^{-\frac{\beta}{2}} I_n(\tfrac{\beta}{2}) (f, \tfrac{\cos n\theta}{\sqrt{\pi}}) \frac{\cos n\theta}{\sqrt{\pi}} +$$

$$+ \sum_{n=1}^{\infty} 2\pi\alpha e^{-\frac{\beta}{2}} I_n(\tfrac{\beta}{2}) (f, \tfrac{\sin n\theta}{\sqrt{\pi}}) \frac{\sin n\theta}{\sqrt{\pi}} . \qquad (6.4.9)$$

R_{11} has the same orthonormal eigenfunctions as K^*K and we identify, for purposes of applying Theorem 5.4.4,

$$\eta_1 = 2\pi\alpha e^{-\frac{\beta}{2}} I_0(\tfrac{\beta}{2})$$

$$\eta_{2n} = 2\pi\alpha e^{-\frac{\beta}{2}} I_n(\tfrac{\beta}{2}) \qquad n \geqslant 1$$

$$\eta_{2n+1} = 2\pi\alpha e^{-\frac{\beta}{2}} I_n(\tfrac{\beta}{2}) \qquad n \geqslant 1 . \qquad (6.4.10)$$

The comparison will this time be between

$$\rho^{2n} \equiv (\tfrac{1}{R})^{2n} \quad \text{and} \quad 2\pi\alpha e^{-\frac{\beta}{2}} I_n(\tfrac{\beta}{2}) .$$

From the series expansion for $I_n(x)$, (an entire function),

$$I_n(x) = (\tfrac{1}{2}x)^n \sum_{k=0}^{\infty} \frac{(\frac{1}{2}x)^{2k}}{k!(k+n)!} \leqslant (\tfrac{1}{2}x)^n \frac{1}{n!} \sum_{k=0}^{\infty} \frac{(\frac{1}{2}x)^k}{k!\,k!}$$

$$= (\tfrac{1}{2}x)^n \frac{1}{n!} I_0(x) .$$

Comparison of the behaviour in n of

$$\frac{2\pi\alpha e^{-\frac{\beta}{2}}}{n!} (\tfrac{\beta}{4})^n I_0(\tfrac{\beta}{2}) \quad \text{and} \quad (\tfrac{1}{R})^{2n}$$

shows that the η_k go down much more rapidly than the λ_k. For small enough values of its argument, the modulus of regularization $\rho(\varepsilon)$ will exhibit very rapid (→ linear) convergence to zero with ε. But what of values to be encountered in computations such as $\varepsilon = 10^{-3}$ or $\varepsilon = 10^{-6}$? Asymptotic estimates are not helpful here and we are not in a position to usefully exploit Theorem 5.4.4. Instead one appeals to 5.4.5.

That is, use the estimate for $\rho(\varepsilon)$ given by

$$\rho^2(\varepsilon) \leqslant \inf_N \left\{ \varepsilon^2 \sup_{k \leqslant N} \frac{1}{\lambda_k} + \sup_{k > N} \eta_k \right\} . \tag{5.4.5}$$

Our λ_k and η_k are monotone. In fact, they are such that the above reduces to

$$\begin{aligned} \rho^2(\varepsilon) &\leqslant \inf_n \left\{ \varepsilon^2 \frac{1}{\lambda_{2n+1}} + \eta_{2n+2} \right\} \\ &= \inf_n \left\{ \varepsilon^2 R^{2n} + 2\pi\alpha e^{-\frac{\beta}{2}} I_{n+1}(\tfrac{\beta}{2}) \right\} . \end{aligned} \tag{6.4.11}$$

For specified values of ε, α, β and R, the above inf can be found numerically. Let us find it for the following values of the parameters:

$$\begin{aligned} \frac{\beta}{2} &= 1;4;25;49, \\ \varepsilon &= 10^{-3};\ 10^{-6}, \\ \alpha &= 1, \\ R &= 2 . \end{aligned} \tag{6.4.12}$$

The variable parameters are ε and β so denote the modulus of regularization $\rho(\varepsilon;\frac{\beta}{2})$.

$$\rho^2(\varepsilon;\tfrac{\beta}{2}) \leqslant \inf_n \left\{ \varepsilon^2 4^n + 2\pi e^{-\frac{\beta}{2}} I_{n+1}(\tfrac{\beta}{2}) \right\} .$$

The computed results appear in Table 6.4.13.

TABLE 6.4.13

$\frac{\beta}{2}$	ε	$\approx \rho(\varepsilon;\frac{\beta}{2})$	$\log \rho(\varepsilon;\frac{\beta}{2})/\log\varepsilon$
1.0	10^{-3}	2.97×10^{-2}	.509
	10^{-6}	2.80×10^{-4}	.592
4.0	10^{-3}	9.41×10^{-2}	.342
	10^{-6}	2.35×10^{-3}	.438
25.0	10^{-3}	1.36×10^{-1}	.136
	10^{-6}	6.55×10^{-2}	.197
49.0	10^{-3}	2.02×10^{-1}	.116
	10^{-6}	1.36×10^{-1}	.136

An interesting quantity is the ratio of the log of $\rho(\varepsilon;\frac{\beta}{2})$ to $\log\varepsilon$. It is a measure of the fraction of significant figures of accuracy in the data, a T_H-method retains in the solution. We expect it to go to 1 slowly in ε* since in the linear limit,

$$\rho \sim C\varepsilon \Longrightarrow \log\rho \sim \log C + \log\varepsilon$$

$$\Longrightarrow \frac{\log\rho}{\log\varepsilon} \sim \frac{\log C}{\log\varepsilon} + 1 = 1 + o(\varepsilon) \ .$$

For all the values of $\frac{\beta}{2}$ considered, this ratio did increase as ε was lowered from 10^{-3} to 10^{-6}.

----

*For this example, we do obtain linear limiting behaviour.

The tabulated values of $\log \rho(\varepsilon)/\log\varepsilon$ would tend to suggest that the smaller β is chosen to be, the better will be the convergence. This corresponds to making statistical assumptions favouring low frequency oscillations. Interestingly enough, this postdiction is not borne out by Franklin's computations. This T_H-analysis very definitely has its limitations.

Roughly speaking, "too many possibilities" are admitted by the combined assumptions:

a) $(f_0, \chi^{-1} f_0) \leqslant \gamma^2$

b) $\|Kf_0-g\|^2 \leqslant \varepsilon$

for a really good comparison with the statistical method. In the spectral decomposition of f_0, high frequency terms are damped by the assumption a) and low frequency ones are pinned down by the assumption b). For small enough values of ε, assumption b) locates arbitrarily many low frequency terms arbitrarily accurately. (That is why spectral cut-off works.) But exceedingly small ε's may be required before the combined effects of a) and b) really give a good rate of convergence. (One must beware of asymptotically pleasing estimates.)

The statistical viewpoint would have us focus attention on a range of frequencies with a lower as well as an upper bound. Perhaps replacing a) with

$$\gamma_1^2 \leqslant (f_0, \chi^{-1} f_0) \leqslant \gamma_2^2$$

would be more appropriate but then the analysis would become more difficult.

Rather than trying to push a point further, let us just accept the fact that all qualitative features of the statistical method will not be explained through the T_H-interpretation. We merely make the observation that the statistically constructed R_{11} is an excellent regularizer for this operator K. Whatever the more appropriate D-analogue happens to be, the regularizing effect of R_{11} can be expected to force rapid convergence.

CONCLUSIONS

This work has been concerned with ill-posed linear problems resulting from an attempt to invert compact mappings between Banach spaces. The mechanism for well-posed extension has been the imposition of an additional constraint that admissible solutions lie within given compact sets. Ultimate numerical solution of appropriate related problems has been in the background of all discussion.

When the mapping is between separable Hilbert spaces (as is so often the case), compact restriction involves suppressing the contributions from most of the terms in the expansion of allowed elements. The mathematical study in Hilbert function spaces such as $L^2[a,b]$ would necessarily begin and end with suppression of high frequency modes. It began with simple spectral cut-off and ended with T_H-methods.

Motivation of constraints is truly of paramount importance although the preponderance of effort has been upon the analysis of their effects and upon their incorporation via related (well-posed) problems. The condition $Kf_0 \approx g$ is very weak; satisfied by an enormous and diverse collection of f_0's. An astronomer asked to approximate the location of a star given only that it lay in a designated unbounded sector of the universe would be confronted with no more impossible a task. In reaching into that morass of f_0's and pulling one out, we must have a reason for focusing our attention on the one chosen. If we have no such reason, anyone else's choice is at least as good as ours.

Order of priority in the reduction of a problem to a form suitable for machine computation is a topic meriting some mention.* Numerical discretization must sooner or later be performed and a finite dimensional problem solved. At what stage in the analysis should this be considered done? Many are in favour of discretizing immediately, much happier in dealing with ill-conditioned matrices than with compact operators. This approach is not endorsed here. We know that incorporation of additional information about the solution will be necessary. That information, be it statistical or deterministic, will be in the nature of statements about elements in the Banach space in which the problem was initially cast. It is difficult to imagine a means whereby the information can be utilized without loss in authenticity. It is recommended instead that we hold off on the numerics until our interest has been established in a well-defined element in the Banach solution space, that element being the solution of a well-posed problem. Then finding a numerical approximation for that quantity becomes a worthy and plausible goal.

One who adopts the philosophy of this work with regard to ill-posed problems must find the subject discouraging until he learns to be content with rather meagre returns. He must learn to call a method good when in his answer he expects "only" to lose half the significant figures of accuracy supplied in his data.

When strong restrictions in a D-method promote very rapid convergence, it means that good à priori knowledge of the solution

----

*It was touched obliquely in the opening remarks of Chapter 6.

was available. Under no circumstances is something given to us for nothing. Very simply, there is a limited amount of information present in the statement $Kf_0 \approx g$. What little there is can be coupled with whatever else we know in our attempt to propose approximate solutions. The optimal campaign of action towards this end will so often yield results below the expectations of one whose past experience is with well-posed problems. We must not be disappointed just because our honestly established claims of accuracy prove to be rather unspectacular. It may be impossible to do significantly better.

APPENDIX A

A FEW DEFINITIONS* FROM TOPOLOGY AND FUNCTIONAL ANALYSIS

It is assumed the reader has some familiarity with linear (vector) spaces, elementary set concepts, convergence of sequences of real numbers.

A. 1 A norm $\|\cdot\|$ defined on a subset V of a linear space X is a rule assigning real numbers to elements $v_1;v_2 \in V$ which has the properties:

a) $\|v_1\| \geqslant 0$; $\|v_1\| = 0$ only if $v_1 = 0$. (positive definiteness)

b) $\|\alpha v_1\| = |\alpha| \, \|v_1\|$. (homogeneity)

c) $\|v_1+v_2\| \leqslant \|v_1\| + \|v_2\|$. (triangle inequality)

A. 2 An inner product $(\cdot\, ,\cdot)$ defined on a linear space X is a complex-valued function defined on XxX with the properties:

a) $(x_1+x_2, x_3) = (x_1, x_3) + (x_2, x_3)$. (bilinearity)

b) $(x_1, x_2) = \overline{(x_2, x_1)}$. (conjugate symmetry)

c) $(\alpha x_1, x_2) = \alpha(x_1, x_2)$.

d) $(x_1, x_1) \geqslant 0$; $(x_1, x_1) \neq 0$ if $x_1 \neq 0$ (positive definiteness)

$x_1, x_2, x_3 \in X$. (A norm $\|\cdot\|$ will be defined by $\|x_1\|^2 = (x_1, x_1)$. Inner product spaces are thus normed spaces.)

A. 3 A normed linear space X $\|\cdot\|$ is a linear space X on which a norm $\|\cdot\|$ is defined.

----

*It will actually be definition interspersed with remarks about provable facts. Only points which will be required to understand the main text are raised.

A. 4 A linear inner product space $X(\cdot,\cdot)$ is a linear space on which an inner product $(\cdot,\cdot)$ is defined.

A. 5 A Cauchy sequence in a normed space $X\|\cdot\|$ is a sequence of elements $\{x_n\} \subset X$ with the property that for each $\varepsilon > 0$, there exists $N(\varepsilon)$ such that

$$\|x_n - x_m\| < \varepsilon \quad \text{if } n;m \geqslant N(\varepsilon) .$$

A. 6 a) A subset V of a normed linear space $X\|\cdot\|$ is said to be closed if all Cauchy sequences $\{v_n\} \subset V$ converge to an element contained in V. That is, if $\{v_n\} \subset V \subset X$ is a Cauchy sequence, there exists $v \in V$ such that

$$\|v_n - v\| \to 0 \text{ as } n \to \infty.$$

b) The smallest closed set containing a given set V is called the closure of V and is denoted $\overline{V}$.

c) A normed space $X\|\cdot\|$ which is closed is said to be complete.

A. 7 A complete, normed, linear space is called a Banach space.

A. 8 A complete, normed, linear, inner product space is called a Hilbert space.

A. 9 A subset V of a normed space $X\ \|\cdot\|$ is said to be dense in X if each element x in X is the limit of a sequence in V. That is if for each $x \in X$, there exists $\{v_n\} \subset V$ such that

$$\|v_n - x\| \to 0 \quad \text{as} \quad n \to \infty.$$

A. 10 An operator K mapping from a subset V of a space X into a subset W of a space Y is a rule which associates a unique element $w \equiv K(v)$ or $Kv \in W$ for each $v \in V$. The largest subset $V \subset X$ on which K is defined is called the domain of K and is denoted DomK. The subset $W \subset Y$ defined by $W = \{y \in Y \mid y = Kv$

for some $v \in \text{Dom}K\}$ is called the range of K and is denoted Ran K. Write $K : \text{Dom } K \rightarrow Y$

$\{Kv \in Y \mid v \in V \subset \text{Dom } K\} \equiv KV$ (the image of V under K).

A. 11 a) An operator is onto (surjective) if Ran K = Y.

b) An operator is one to one (injective) if for $v_1, v_2 \in \text{Dom } K$.

$Kv_1 = Kv_2$ only if $v_1 = v_2$.

c) An operator K is linear if for all $v_1, v_2 \in \text{Dom } K$ and all scalars α and β,

$$K(\alpha v_1 + \beta v_2) = \alpha K v_1 + \beta K v_2.$$

A. 12 If $K : \text{Dom } K \subset X \rightarrow Y$ is one to one, then for each $w \in \text{Ran } K$, there is a unique element $v \in \text{Dom } K$; denote $v \equiv K^{-1}w$. The operator K^{-1} so defined is called the inverse of K.

A. 13 A topology on X is a class T of subsets of X satisfying:

a) The union of every class of sets in T is a set in T.

b) The intersection of every finite class of sets in T is a set in T.

A. 14 An open set O in a Banach space $B\|\cdot\|$ is a set with the property that for each $x \in O$, there exists a positive scalar ε such that

$$\{b \in B \mid \|b-x\| < \varepsilon\} \subset O.$$

The class T of open sets of $B\|\cdot\|$ defines the norm topology on $B\|\cdot\|$.

A. 15 A subset V of a Banach space $B \|\cdot\|$ is compact (in the norm $\|\cdot\|$ topology) if every class of open sets whose union contains V has a finite subclass whose union contains V.

A. 16 A linear operator $K : X_1 \rightarrow X_2$ where $X_1 \|\cdot\|,$ and $X_2 \|\cdot\|_2$ are normed linear spaces is said to be bounded if there exists a real number M (≥ 0) such that for all $x_1 \in X_1$,

$$\|K x_1\|_2 \leq M\|x_1\|_1 .$$

The smallest such M is denoted $\|K\|$ and is called the norm of the operator K.

A. 17 A linear functional F defined on a linear space X is a linear operator whose range is contained in the space of scalars and whose domain is X. The space of bounded linear functionals defined on a normed space $X \|\cdot\|$ is denoted X^*.

A. 18 A sequence $\{x_n\} \subset X$ is said to be weakly convergent to $x \in X$ if for all bounded linear functionals $F \in X^*$,

$$|F(x_n) - F(x)| \rightarrow 0 \text{ as } n \rightarrow \infty. \text{ Denote } x_n \rightharpoonup x.$$

A. 19 A linear operator $K : X_1 \rightarrow X_2$ where $X_1 \|\cdot\|_1$ and $X_2 \|\cdot\|_2$ are normed spaces is said to be compact if the image KV of every bounded set $V \subset X_1$ has compact closure in X_2. That is if $\overline{KV}$ is compact (in the $\| \ \|_2$-topology).

A. 20 Let $B_1 \|\cdot\|_1$ and $B_2 \|\cdot\|_2$ be Banach spaces and $T : B_1 \rightarrow B_2$ be a bounded linear operator. Define the adjoint operator $T^* : B_2^* \rightarrow B_1^*$ by

$$T^*F(b_1) = F(Tb_1)$$

for all $F \in B_2^*$; $b_1 \in B_1$. (It is easily shown that T^* is bounded, linear, and $T^* : B_2^* \rightarrow B_1^*$.)

Remark

In a Hilbert space $H(\cdot, \cdot)$, a general theorem due to Riesz says that any member of H^* can be identified with a member of H. Specifically, for $F \in H^*$, there exists $f \in H$ such that for all $h \in H$,

$$F(h) = (h, f) .$$

H and H^* are thus identified.

If $H_1(\cdot, \cdot)$, and $H_2(\cdot, \cdot)_2$ are Hilbert spaces and $T : H_1 \rightarrow H_2$ is a bounded linear operator, then $T^* : H_2 \rightarrow H_1$ is defined by

$$(T^* h_2, h_1)_1 = (h_2, Th_1)_2$$

for $h_1 \in H_1$; $h_2 \in H_2$.

A. 21 Let $H(\cdot, \cdot)$ be a Hilbert space and T be a bounded linear operator mapping from H into itself $(T : H \rightarrow H)$.

a) T is said to be normal if $TT^* = T^*T$.

b) T is said to be self-adjoint if $T = T^*$.

Remark

In all the foregoing, details which will not be needed are omitted. Convergence, unless otherwise specified, is in the norm; compactness will always be in the norm topology; open and closed sets have been defined for the norm topology. Ill-posed problem theory could be generalized to much more abstract topologies; convergence in nets considered - indeed, work of this sort has been done. Abstraction was carried as far as practical utility seemed to suggest.

APPENDIX B

Remark on the Estimate of Sturm-Liouville Eigenvalues

In Chapter 4, the following estimate was given for the nth eigenvalue λ_n of the Sturm-Liouville problem:

$$(pw_x)_x + (\lambda\rho - q)w = 0 \qquad (4.1.4)$$

subject to

$$\alpha_1 w(a) + \beta_1 w'(a) = 0$$

$$\alpha_2 w(b) + \beta_2 w'(b) = 0 \text{ —}$$

$$\lambda_n \sim \left(\frac{(n-1)\pi}{\beta}\right)^2 + O(1) \text{ ,}^* \qquad (4.5.2)$$

β being given by

$$\beta \equiv \int_a^b \left[\frac{\rho(x)}{p(x)}\right]^{\frac{1}{2}} dx \quad .$$

It was mentioned that this formula's application in deciding a truncation value in a series expansion (see section 4.5) should not be taken too seriously; that the constant in O(1) could be large. Here, a couple of instances are cited in which this is the case.

First take:

$$\alpha_1 = \alpha_2 = 0 \text{ ; } \quad \beta_1 = \beta_2 = 1 \text{ ; } \quad [a,b] = [0,\pi] \text{ ;}$$

$$p \equiv \rho \equiv 1 \text{ ; } \quad q \equiv Q^2 = \text{constant .}$$

----

*It was stated that for large n, eigenvalues increase <u>at least</u> this rapidly.

That gives the system:

$$w'' + (\lambda - Q^2)\, w = 0$$

$$w'(0) = w'(\pi) = 0 \,.$$

The eigenvalues λ_n and the eigenfunctions φ_n (normalized) are easily found to be:

$$\varphi_1 \equiv \frac{1}{\sqrt{\pi}} \quad \text{corresponding to } \lambda_1 = Q^2 \;;$$

$$\varphi_n \equiv \sqrt{\frac{2}{\pi}}\, \cos\left[(n-1)x\right] \quad \text{corresponding to } \lambda_n = (n-1)^2 + Q^2 \quad n > 1 \;;$$

$$\beta = \int_0^{\pi} \left[\frac{1}{1}\right]^{\frac{1}{2}} dx = \pi \,.$$

4.5.2 would have $\lambda_n = (n-1)^2 + O(1)$ and the constant in $O(1)$ is identified as Q^2. This may be chosen arbitrarily large at will.

Now take:

$$p(x) = x; \quad \rho(x) = \frac{1}{x} \;; \quad q(x) = \frac{Q^2}{x} \quad (Q^2 = \text{const.})$$

$$\alpha_1 = \alpha_2 = 0 \;; \quad \beta_1 = \beta_2 / 1 \;; \quad [a, b] = [1, 2] \,.$$

That yields the system:

$$\frac{d}{dx}\left(x \frac{d}{dx} w\right) + \left(\frac{\lambda}{x} - \frac{Q^2}{x}\right) w = 0$$

$$w'(1) = w'(2) = 0 \,.$$

The eigenvalues λ_n and eigenfunctions φ_n (normalized; weight function $\rho(x) = \frac{1}{x}$) are

$$\varphi_1 = \sqrt{\frac{1}{\ell n 2}} \quad \text{corresponding to} \quad \lambda_1 = Q^2$$

$$\varphi_n = \sqrt{\frac{2}{\ell n 2}} \cos\left[\frac{(n-1)\pi \, \ell n x}{\ell n\, 2}\right] \quad \text{corresponding to}$$

$$\lambda_n = \left[\frac{(n-1)\pi}{\ell n\, 2}\right]^2 + Q^2 \qquad \text{for } n > 1 \ .$$

$$\beta = \int_1^2 \left(\frac{1}{x^2}\right)^{\frac{1}{2}} dx = \ell n\, 2 \ .$$

4.5.2 yields $\lambda_n = \left[\frac{(n-1)\pi}{\ell n\, 2}\right]^2 + O(1)$ and the constant is again identified as Q^2; arbitrarily large. It is easy to see that, in general, increasing q by $Q^2 \rho(x)$ will increase the eigenfunctions by Q^2 in any system 4.1.4.

REFERENCES

[1] Bakusinskii, A. B., "A New Regularizing Algorithm for the Solution of a Linear Ill-posed Equation in a Hilbert Space." (Russian) Computing Methods and Programming XII, (Russian), pp. 53-55. Izdat. Moskov Univ. Moscow (1969). Math. Rev. V-44, 1972 (review by G. Vainikko).

[2] Buzbee, B. L. and Carasso, A., "On the Numerical Computation of Parabolic Problems for Preceding Times," Math. Comp. 27, 122 (1973), pp. 237-266.

[3] Cannon, J. R., "Some Numerical Results for the Solution of the Heat Equation Backwards in Time," Numerical Solutions of Nonlinear Differential Equations. (Proc. Adv. Sympos. Num. Sol. of Nonlinear D. E.'s, Madison, Wisc., 1966), Wiley, New York (1966), pp. 21-54.

[4] Courant, R. and Hilbert D., Methods of Mathematical Physics Volume II, Interscience, New York (1966).

[5] Douglas, J., "The Approximate Solution of an Unstable Physical Problem Subject to Constraints," Functional Analysis and Optimization, Academic Press, New York (1966), pp. 65-66.

[6] Franklin, J. N., "Well-Posed Stochastic Extensions of Ill-Posed Linear Problems," J. Math. Anal. Appl. 31 (1970), pp. 682-716.

[7] Franklin, J. N., Stability of Bounded Solutions of Linear Functional Equations," Math. Comp., 25, 115 (1971), pp. 413-424.

[8] Franklin, J. N., "On Tikhonov's Method for Ill-posed Problems," Math. of Comp., 28, 128 (1974).

[9] Friedman, A.. Partial Differential Equations, Holt, Rhinehart and Winston, New York (1969).

[10] Gajewski, H. and Zacharias, K., "Zur regularisierung einer Klasse nicht-korrekter Probleme bei Evolutionsgleichungen," J. Math. Anal. Appl., 38 (1972), pp. 784-789.

[11] Gelfand, I. M., "Generalized Random Processes," Dokl. Akad. Nauk, USSR 100, 853-56 (1955), Math. Rev. V-16, 1955 (part 2 of rev. by Doob).

REFERENCES (Cont'd)

[12] Hadamard, J., Lectures on Cauchy's Problem in Linear Partial Differential Equations, Yale Univ. Press, New Haven, Conn. (1923).

[13] John, F., "Numerical Solution of the Equation of Heat Conduction for Preceding Times," Ann. Mat. Pura. Appl., 40 (1955), pp. 129-142.

[14] John, F., "Continuous Dependence on Data for Solutions of Partial Differential Equations with a Prescribed Bound," Comm. Pure Appl. Math., 13 (1960), pp. 551-585.

[15] Keller, H. B., Numerical Methods for Two-Point Boundary Value Problems, Blaisdell, Waltham, Mass. (1968).

[16] Keller, H. B., "A New Difference Scheme for Parabolic Problems," Numerical Solution of Partial Differential Equations - II (Proc. Second Sympos. on the Num. Sol. of P.D.E.'s, Univ. of Md. 1970), Academic Press, New York (1971), pp. 327-350.

[17] Lattès, R. and Lions, J.-L., The Method of Quasireversibility Applications to Partial Differential Equations, American Elsevier, New York (1969).

[18] Lavrentiev, M. M., Some Improperly Posed Problems of Mathematical Physics, Springer Tracts in Natural Philosophy, Vol. 11, Springer-Verlag, Berlin (1967)

[19] Payne, L. E., "Some General Remarks on Improperly Posed Problems for Partial Differential Equations," Symposium on Non-Well-Posed Problems and Logarithmic Convexity, (Heriot-Watt Univ., Edinburgh, Scotland, 1972), Springer-Verlag, Berlin (1973).

[20] Payne, L. E., "On Some Nonwell Posed Problems for Partial Differential Equations," Numerical Solutions of Nonlinear Differential Equations (Proc. Adv. Sympos. Num. Sol. of Nonlinear D.E.'s, Madison, Wisc, 1966), Wiley, New York (1966), pp. 239-263.

[21] Riesz, F. and Sz.-Nagy, B., Functional Analysis, Frederick Ungar, New York (1955).

[22] Saylor, R., "Time Reversal in Abstract Cauchy Problems," SIAM J. Math. Anal., 2 (1971), pp. 454-457.

REFERENCES (Cont'd)

[23] Simmons, G. F., Introduction to Topology and Modern Analysis, McGraw-Hill, New York (1963).

[24] Stakgold, I., Boundary Value Problems of Mathematical Physics, Volume I, Macmillan, New York (1967).

[25] Strand, O. N. and Westwater, E. R., "Statistical Estimation of the Numerical Solution of a Fredholm Integral Equation of the First Kind," J. of the Assoc. for Computing Machinery, 15 (1968), pp. 100-114.

[26] Sudakov, V. N. and Kalfin, L. A., "A Statistical Approach to Improperly Posed Problems of Mathematical Physics" (Russian), Dokl. Akad. Nauk. SSSR, 5 (1957).

[27] Taylor, A. E., Introduction to Functional Analysis, Wiley, New York (1958).

[28] Tikhonov, A. N., "Solution of Incorrectly Formulated Problems and the Regularization Method," Dokl. Akad. Nauk. SSR 153 (1963), 49 = Soviet Math. Dokl. 4 (1963).