SEISMIC MIGRATION BY CHEBYCHEV TRANSFORM: A NOVEL APPROACH

by

DIMITRIOS MICHAEL MITSAKIS

B.Sc. (Physics), University of Athens, 1983
Graduate Diploma (Applied Geophysics), Mc Gill University, 1984

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

Department of Geophysics and Astronomy

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

April 1987

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Geophysics and Astronomy

The University of British Columbia
1956 Main Mall
Vancouver, Canada
V6T 1Y3

Date 29/04/87

ABSTRACT

Chebychev semi-discretizations for both ordinary and partial differential equations are explored. The Helmholtz, heat, Schrödinger and 15° migration equations are investigated. The Galerkin, pseudospectral and tau projection operators are employed, while the Crank-Nicolson scheme is used for the integration of the time (depth) dependence. The performance of the Chebychev scheme is contrasted with the performance of the finite difference scheme for Dirichlet and Neumann boundary conditions. Comparisons between all finite difference, Fourier and Chebychev migration algorithms are drawn as well.

Chebychev expansions suffer from neither the artificial dispersion dispersion of finite difference approximations nor the demand for a periodic boundary structure of Fourier expansions. Thus, it is shown that finite difference schemes require at least one order of magnitude more points in order to match the accuracy level of the Chebychev schemes. In addition, the Chebychev migration algorithm is shown to be free of the wraparound problem, inherent in migration procedures based on Fourier transform.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

My supervisor Matt Yedlin has been an indispensable source of guidance, support and enthusiasm; his collaboration has been a revealing experience for me. Garry Clarke reviewed this thesis; his acute advice on the editing of the manuscript is most appreciated. Special thanks to Colin Walker for the many invaluable discussions on numerous aspects of this research. I would also like to thank Tim Scheuer and Dave Lumley for a few long stimulating conversations regarding the evanescent wavefield of chapter VI. Sonya Dehler and Barry Zelt provided great assistance in TEX-associated problems and misfortunes. My fellow graduate students in the Department of Geophysics and Astronomy have comprised a creative research environment. My girlfriend Anita has been a source of enduring patience, encouragement and companionship.

Last but certainly not least I want to thank deeply my father Michali, my mother Fitsa, my brother Kosta, my sister Evella and my cherished grandmother Evdokia for their love, devotion, support and unwavering belief in me. This thesis is dedicated to them.

Financial support for this research came from the Natural Sciences and Engineering Research Council of Canada (NSERC) and Esso Canada.

DEDICATION

ΣΤΗΝ ΟΙΚΟΓΕΝΕΙΑ ΜΟΥ

ΜΕ ΑΠΕΡΑΝΤΗ ΑΓΑΠΗ

I would perpetuate these nympths.

So clear,

Their skin's light bloom, it eddies in the air

Heavy with tufts of sleep.

Did I love a dream?

A Faun's Afternoon — Stéphane Mallarmé

CHAPTER I

SEISMIC MIGRATION

As a lily sways in newly stilled air, so my being moved
in its elements, in my ravishing dreams of her.
Hyperion or *The Hermit in Greece* — Friedrich Hölderlin

1.1 Migration Fundamentals

Simplifying assumptions regarding the subsurface's structure are involved in seismic data processing. Consequently, the final output does not correspond to the actual distribution of reflectors or diffractors in the earth's crust. The procedure that is responsible for repositioning the seismic data, so that associated reflectors or diffractors are properly reconstructed, is known as *migration* (Sheriff and Geldart, 1984).

Most migration techniques involve a lot of simplifying assumptions (either in early or later stages of their formulation) in order to achieve versatile and practical algorithms. The main problem — present due to the inverse problem nature of migration — is the lack of knowledge of the subsurface velocity distribution. Most techniques consider all the energy present in the seismogram as being primary; migration is mostly

done in two dimensions under the assumption that the cross-dip is zero. Moreover, lateral velocity variations cannot be accommodated both effectively and efficiently.

The reported simplifications usually result in undermigrated or overmigrated sections and in an improper handling of noise and multiples, which exhibit themselves as migration noise and improper reconstruction of the structure underneath. Nevertheless, migration usually improves the quality of the output section to be interpreted by enhancing its lateral resolution (constrained by the presence of spatial aliasing).

1.2 Modern Migration Algorithms

While the diffraction-stack (or wavefront-sum) method (Hagedoorn, 1954) dominated seismic migration up to the 70's, more recent migration techniques involve a two-step process. A wave-field extrapolation via a solution of the wave equation and an imaging principle, in order to define the end point of the downward continuation of the wavefield.

There exist three main methods for the implementation of the approximation of the particular equation to be solved, namely, the *Kirchhoff*, the *finite difference* and the *spectral* methods. The most common migration technique is the finite difference method in which finite difference approximations are employed in all the coordinates. Various parabolic approximation equations to the scalar wave equation are commonly solved. These are one-way equations and this property makes them easier to solve, more stable and more appropriate for application of the imaging principle (Claerbout, 1976). Their derivation is most readily understood on the basis of the *paraxial* approximation (see 6.1). Two of these equations have widely been applied in the migration

of seismic sections, namely, the 15° and 45° parabolic wave equations. (Loewental et al, 1976; Hood, 1978).

The other major class of migration algorithms is based on Fourier transforms and belongs to the spectral category (Bolondi et al, 1978). Its two basic variants are the $F - K$ (frequency-wavenumber) or Stolt migration (Stolt, 1978) and the phase-shift migration (Gazdag, 1978). An improved implementation of the latter (phase-shift plus interpolation) is presented in Gazdag and Sguazzero (1983). The major characteristic of these methods is accurate space derivative estimates due to the global nature of the Fourier spectral basis functions. However, this global property prohibits velocity variations in the coordinate to be transformed and it allows errors (that would be localized in finite difference formulations) to contaminate the whole spectrum and consequently, to propagate all over the domain of the solution (Claerbout, 1985). The Kirchhoff sum method (in principle an advanced and rigorous version of the older diffraction stack method) is based on the Kirchhoff's integral solution of the wave equation; it makes use of the Kirchoff-Huygens diffraction hypothesis, that is, every element of a reflector acts as a separate scatterer (French, 1975). Integral transform methods have been used for its solution (Schneider, 1978) and it has also been approached as a spatial deconvolution problem (Berkhout and van Wulfften Palthe, 1979).

None of above methods could claim to be the best. Their performance depends on the data set and on the particular implementation used. In general though, finite-difference methods are limited to either 15° or 45° dips, they can accommodate lateral velocity variations and they produce less migration noise. Frequency methods can migrate dips of any degree, but they produce more migration noise and they become very

cumbersome when lateral velocity changes need to be handled (Sheriff and Geldart, 1984).

Conventionally, migration is done after CMP stacking, so that the cost is reduced since the half-offset coordinate has been eliminated. The quality of a migrated section can be enhanced if migration is performed before stack (Dubrulle, 1983; Reshef and Kosloff, 1986). This is known as pre-stack migration and it is very costly. A partial pre-stack migration procedure has also been proposed (Yilmaz and Claerbout, 1980). Another interesting pre-stack kind of migration involves the migration of common-midpoint slant stacks (Ottolini and Claerbout, 1984), whereas even two-way wave equations have recently been used in migration (Kosloff and Baysall, 1983, Baysall et al, 1984).

Discussion of the many fine points involved in migration and an analysis of the numerous elaborate techniques that have been proposed goes beyond the scope of this brief summary of the migration fundamentals and this quick look at the most basic algorithms. An extensive and in-depth discussion of the subject can be found in Claerbout (1985).

1.3 Reflecting and Absorbing Boundary Conditions

In the numerical solution of initial boundary value problems, a finite grid is used to perform the computations. A seismic model cannot be represented by a finite medium and therefore, by retreating to numerical methods, we introduce artificial boundaries in our model. These boundaries create unphysical reflections which contaminate the solution, interfere with the true reflections and, possibly, cause instabilities. However,

the earth does not feature those boundaries and, consequently, the waves (obeying the physical free-space boundary conditions) will continue through them to die out at infinity.

We see the need to invent special boundary conditions which would simulate *transparent* boundaries. Two major directions in establishing them exist. The first involves introduction of numerical viscosity at an area close to the boundary, so that the wave's amplitude is reduced as they approach it; subsequently, we either get weak or no reflections at all (Lysmerand and Kuhlemeyer, 1969). This method performs well for compressional waves but diminishing of the reflected shear waves is not satisfactory. Recently, Cerjan et al (1985) proposed a modification of this technique to account for its shortcomings. In general, this first absorbing boundary condition technique has the advantage that is flexible enough to be employed regardless of which kind of discretization of the problem is used (Kosloff and Kosloff, 1986).

The second main approach is directed towards factoring the wave equations into incoming and outgoing waves. Modeling outgoing wave components at the boundaries, we succeed in absorbing energy over a wide range of angles of incidence and, thereafter, reduce artificial reflections. The *transparent* (the terms *absorbing*, *radiating* and *flowing* also appear in the literature) boundary conditions for the full elastic and scalar wave equation are essentially paraxial approximations of them. This technique has been developed for finite difference approximations and details can be found in Lindman (1975), Reynolds, (1978), Clayton and Engquist, (1977), Israeli and Orszag, (1981); an analysis of the well-posedness and stability of the resulting schemes has also been done (Engquist and Majda, 1977 and 1979). Absorbing boundary conditions for the parabolic wave equations used in migration have also been developed (Clayton and

Engquist, 1980). Problems associated with this technique are its inability to account for an adequate elimination of reflected energy in multi-dimensional problems when the angle of incidence is shallow and that its application to global discretizations (Fourier methods) is not clear, since although all the grid points are coupled, the boundary conditions are local. Total cancellation of the artificial reflections may be achieved by alternating between Dirichlet and Neumann boundary conditions; the reflection coeffients for these conditions are $+1$ and -1 respectively and therefore they cancel each other out (Smith, 1974). Nonetheless, the process requires excessive computations. Recently, a slightly different decomposition (Keys, 1985) helps the design of absorbing boundary operators by incorporating the direction of propagation. Coordinate transformations have also been investigated (Grosch and Orszag, 1977); however, significant errors might be introduced since waves in the vicinity of infinity cannot be resolved if the true solution is oscillatory there. Bayliss and Turkel (1982) developed a similar class of boundary conditions and showed that the behavior of the solution in the far field strongly affects the kind of absorbing boundary conditions to chosen.

1.4 Fourier Transform Migration and Associated Problems

Spectral Fourier methods impose periodic boundary conditions on the differential equation under consideration. Such an assumption for a seismic section is most likely invalid. However, if periodic boundary constraints are not imposed, the *Gibbs phenomenon* (see 2.2.4.1) will either slow down the convergence of the method or possibly lead to its divergence.

In addition to demanding *periodic boundary conditions*, the computational grid of a discrete Fourier transform cannot distinguish between a wavevector K and its aliases $K \pm N$ where N is the period of the transform. That could create very serious problems when evaluating convolutional sums in the wavenumber domain due to improper interaction among the various modes. This phenomenon is actually a manifestation of reflecting boundaries in the Fourier method (discretizing the continuous Fourier transform over a finite domain, forces the both the function and the transform to be periodic) and it is better known as *wrap-around* (periodic reflections or mirror images). In practice, this is avoided by doubling the period of the D.F.T and appending the extra space with zeros (Stolt, 1978); consequently, the various modes interact properly and we avoid having objects migrating across the boundaries of the computational domain and producing incorrect results near the boundaries (Gazdag, 1978).

1.5 Chebychev Transform as an Alternative

The induced periodicity and the introduction of extra computational work, due to need for padding with zeros, are undesirable features of Fourier methods. Achieving a relaxation of those requirements, while maintaining accurate derivative evaluation, implies that we employ spectral transforms that are non-periodic.

The Chebychev polynomials (see 2.2.4.2) emerge as a viable alternative for coping with the problem, since they enjoy the advantage of accurate derivative approximation without the need of any specific boundary structure. As far as the problem of the reflecting boundaries is concerned, a relevant variant of existing techniques should be employed.

Before studying the performance of Chebychev spectral methods in equations used for migration procedures (which are partial differential equations with two spatial dimensions and time derivatives), we should investigate the particularities and the details of the technique in simpler equations. In this thesis, four equations of increasing difficulty are solved with both finite difference and Chebychev schemes and a discussion of the relative performances is presented. These equations are:

1. The one-dimensional *Helmholtz* equation with the coefficient of the function term being positive; this is an eigenvalue-eigenfunction problem (chapter III).

2. The one-dimensional homogeneous *heat* equation, which is the natural extension of the previous ordinary differential equation to the equivalent partial differential equation; a first order time derivative has been introduced, so that the complications of time differencing may be studied (chapter IV).

3. The one-dimensional *Schrödinger* equation (the complex counterpart of the heat equation), since it is virtually identical to the diffraction term of the 15° migration equation (chapter V).

4. The 15° *migration* equation, for a 2-D earth model (chapter VI).

A presentation of the fundamentals of the various techniques and their particulars is given in chapter II; an updated summary of their performance for various problems has also been included.

CHAPTER II

SPECTRAL METHODS

But, on the other hand, in a universe suddenly divested of illusions and lights, man feels an alien, a stranger.

The myth of Sisyphus — Albert Camus

2.1 The Method of Weighted Residuals

All numerical methods seek the solution in a discrete finite point set and therefore, both the input and the output vector is given and sought, respectively, on those points. It is evident that an interpolation problem is encountered here, and it is essentially the particular form of the interpolation scheme used (either polynomial or trigonometric in general) that characterizes and distinguishes the various methods. In order to attempt a simultaneous description and comparison of the finite difference and the spectral methods, we may view them as special cases of discretization schemes known under the general name of the *Method of Weighted Residuals* (*M.W.R*) (Hussaini et al, 1983); finite elements may also be understood under this general frame. Detailed

presentations of the *M.W.R* method may be found in Strang and Fix (1973), Finlayson (1972) and Vichnevetsky (1981). A fundamental description is provided below.

The solution is sought in the form of a truncated series expansion $\sum_n^N a_n g_n$ in terms of the basis functions, i.e the g_n's. The expansion is then substituted in the differential equation to be solved and a residual is produced. In the next step another set of functions, usually referred to as weight functions, are used to manipulate this residual in a certain way, so that some norm of it would be either minimized or maximized.

In most practical cases, the manipulations follow the rules of the common inner product in the Hilbert space, resulting in the usual projection of the residual. In this way its L_2 norm — which is linked to its energy in a very straightforward manner — is minimized. This provides us with a set of simultaneous algebraic equations to be solved, so that the coefficients of the assumed expansion (the a_n's) will be obtained.

2.2 The Choice of Basis Functions

2.2.1 Finite difference methods

The *finite difference* methods involve a *polynomial interpolation*, whose explicit form can be viewed either as the Lagrange interpolating polynomial or the Newton finite difference polynomial. According to this scheme, the basis functions are polynomials of a certain degree (depending on the accuracy desired) having the property that every one of them is *local* on the computational grid.

2.2.2 Spectral methods

The *spectral* methods involve a *trigonometric interpolation*, where the basis functions belong to infinitely differentiable sets of linearly independent (and most commonly orthogonal) functions, which are usually chosen among linear combinations of the eigenfunctions of various (either singular or non-singular) Sturm-Liouville problems. Those contrary to the basis functions of the finite difference methods, are *global* on the computational grid.

2.2.3 Finite difference versus spectral methods

To show the close relationship of the polynomial interpolation, with Taylor-type expansions, we will use the Newton-Gregory formula for the approximation of a function $f(x)$ between the points α and $\alpha + (n-1)\epsilon$ using n equidistant interpolating points $\alpha, \alpha+\epsilon, \ldots, \alpha+(n-1)\epsilon$ (Lanczos, 1938):

$$\begin{aligned}\overline{f(\alpha+x)} = f(\alpha) + \left(\frac{\Delta f}{\Delta x}\right)_{x=\alpha}[x] + \frac{1}{2!}\left(\frac{\Delta^2 f}{\Delta x^2}\right)_{x=\alpha}[x]^2 \\ + \ldots + \frac{1}{(n-1)1}\left(\frac{\Delta^{n-1} f}{\Delta x^{n-1}}\right)[x]^{n-1}\end{aligned} \tag{2.1}$$

with $\Delta x = \epsilon$ and $[x]^\kappa = x(x-\epsilon)(x-2\epsilon)\ldots(x-(\kappa-1)\epsilon)$. When $f(x)$ is analytic at $x = \alpha$ with ϵ tending to 0, (2.1) becomes

$$f(\alpha+x) = f(\alpha) + f'(\alpha)x + \ldots + \frac{f^{n-1}(\alpha)}{(n-1)!}x^{n-1} \tag{2.2}$$

which is a Taylor expansion truncated to n terms. It is obvious then, that polynomial interpolation is based on a Taylor-type of expansion.

Taylor expansions show a local character. As a consequence, their extrapolating property limits us to estimating values only exceedingly close to the centre of the expansion. The accuracy of the approximation deteriorates rapidly, as we move farther away from its close vicinity. Therefore, due to the Taylor expansion characteristics, the distribution of the error is much less uniform when using finite differences. Even in the case of a convergent approximation significant departures from the true solution might arise due to three other important sources of discrepancies. These are the aliasing, the truncation error and the boundary conditions (Vichnevetsky, 1981).

Aliasing is due to high frequency characteristics of the function being sought, which might prohibit its adequate representation by a certain degree polynomial. Higher frequencies fold back and as they become indistinguishable from lower ones, degrade the approximation even more. The only way to reduce the effect of this type of error is to retain more terms in the truncated expansion or equivalently to increase the degree of the approximating polynomial. It is well understood that any finite function cannot be band-limited; therefore, by increasing the number of samples we cannot eliminate aliasing. However, we may be able to limit it to an acceptable level (Brigham, 1974).

Trigonometric interpolation cannot reduce the effects of aliasing either, but it can help us to eliminate completely the other source of serious problems, the so-called *truncation* error (Vichnevetsky, 1981) or *phase* error (Orszag, 1971b). The local character of the finite difference method and therefore, the small number of grid points involved in evaluating a certain order derivative, creates an error that becomes more

profound (Orszag, 1971e) for the short wavelengths (high wavenumbers or frequencies), as the grid cannot handle their spatial variational rates adequately. This error unfortunately propagates and results in an alteration of the characteristics of those ill-handled wavelengths introducing instability to the scheme, which inevitably degrades or even destroys completely the accuracy of the approximation. This phenomenon has long been recognized as a fundamental potential flaw in finite difference implementations and it has been termed *grid* (the terms *numerical* and *artificial* are used, as well) dispersion (Alford et al, 1974). In other words, the phase speed becomes a function of the discretization interval. Therefore, it alters accordingly the characteristics of propagation (Kelly et al, 1976).

The most common ways of overcoming this problem are oversampling and numerical viscosity. Since we are only interested in band-limited signals, we can choose the sampling rate such that numerical dispersion lies outside the band of interest. This approach is described as oversampling since we need to take 8-10 points per minimum wavelength of interest, as opposed to 2 according to the Nyquist aliasing criterion. This imposes a practical limitation when high-frequency resolution is sought (Kosloff and Baysal, 1982). Numerical viscosity amounts to an implicit filtering of high wavenumbers that are susceptible to artificial dispersion. Artificial viscosity is usually added to nondissipative schemes and it has the advantage, that the user controls the magnitude of the dissipation induced through the viscosity coefficient. Inherently dissipative schemes also exist (upstream differences, Lax-Wendroff). Employment of such schemes may be either advantageous or disadvantageous, depending on the particular characteristics of the problem being tackled. For hyperbolic problems dispersion accumulates with time. One manifestation of cumulative dispersion is dissipation. Attention must

be paid when elaborate schemes are used as parasitic waves might be introduced. It is also of interest to note that shear waves appear to undergo more severe numerical dispersion than longitudinal waves (Chin et al, 1984). Additionally, mixed finite difference schemes might conceal subtleties with respect to their stability properties. For example, the leap-frog Dufort-Frankel scheme for the advection-diffusion equation requires some stringent stability requirements in two dimensions, while it is unconditionally stable in one dimension. This is of significant interest in geophysical modeling, where the computational grid exhibits a small horizontal size and large velocity in the vertical direction (Coushman-Roisín, 1984).

Implementation of a higher order finite difference scheme reduces the truncation errors and consequently allows for coarser sampling. As an example of artificial dispersion consider the simple one-dimensional, homogeneous and non-dispersive hyperbolic equation $\partial^2 P/\partial t^2 = c_0^2 \partial^2 P/\partial x^2$. Time discretization results to an anomalous dispersion, whereas space discretization leads to normal dispersion. As a result when a scheme is controlled by temporal error, the numerical dispersion leads the signal, whereas it follows the signal when the spatial error is dominant (Dablain, 1986). Furthermore, artificial dispersion may give rise to *anisotropy* when more spatial coordinates are considered (Alford et al, 1974).

The analysis of the numerical dispersion is greatly facilitated through the calculation of the group velocity. The group velocity controls the energy propagation in dispersive partial differential equations. Therefore, a new kind of analysis — termed *group velocity* or *normal mode* analysis — has been proposed (Trefethen, 1982) for the investigation of the artificial dispersion associated with discrete representations of partial differential equations and its consequences.

Although this analysis is directly applicable to nondissipative finite difference schemes only, it can be extended to most dissipative ones, since dispersion usually dominates dissipation at low frequencies. A lot of insight and a quantitative comprehension of various differencing errors in wave propagation problems can be gained through the group velocity analysis. In addition, this analysis has been shown to provide us with a clear understanding of the Gustafsson, Kreiss and Sundstrom stability theory for hyperbolic initial boundary value problems (Gustaffson et al, 1972) — difficult in its original algebraic formulation (Trefethen, 1983).

The boundary conditions comprise the third source of discrepancy between the true and the discrete solution. The spectral methods handle specific boundary constraints in a very straightforward manner, which closely resembles the analytic way of handling them. This is not true for the finite-difference methods, which require the introduction of ficticious points across the boundaries in order to be able to represent certain type of boundary conditions. This results in expressions that are neither accurate nor do they remind us at all of the original expression. The reason, why boundary constraints formulations exhibit such significant differences depending on the method used, has to be found in either the local or the global character of the basis functions used.

Finite difference formulation of boundary conditions involves only the local boundary basis function, whereas a spectral formulation involves all the basis functions retained in the expansion. It would then be anticipated that boundary conditions would be much more accurately represented by spectral methods, although we may have to pay for an increased complexity in the calculations (details about the manipulation of the boundary conditions equations will be given in the discussion of the various

types of spectral methods). The late revival of the normal mode analysis (Gustafsson, 1982) has facilitated a better understanding of the effect of various inflow-outflow boundary conditions on the accuracy and stability properties of different finite difference schemes (Lax-Wendroff, folded Lax-Wendroff, leap-frog, Mac Cormack) applied to multi-dimensional initial boundary value hyperbolic equations (Abarnabel and Murman, 1982; Beam et al, 1982; Blotner, 1982; Coughran Jr.,1984).

All spectral methods are very sensitive to the correct implementation of the boundary equations; an inappropriate implementation could cause a hopeless divergence of the method (Orszag, 1971a).

2.2.4 Choice of spectral basis functions

Of equal importance is the choice of the basis functions to be used in a spectral method. The basis functions are frequently chosen among linear combinations of eigenfunctions of Sturm-Liouville problems. These eigenfunctions are linearly independent and they can be orthonormalized. A very detailed analysis of convergence rates can be found in the excellent monograph "Spectral Methods in Numerical Analysis" (Gottlieb and Orszag, 1977); here, we will restrict ourselves to a very comprehensive presentation of the most fundamental properties they possess.

2.2.4.1 The Fourier transform

The most popular set of orthogonal basis functions is the complex *Fourier* series, which is, in fact, the one that has extensively been used in geophysics. Several factors contribute to its great popularity. The transform (spectral) components are equidistantly placed, which makes both the concept of frequency and strength of various

frequency components very clear. It is very *clean* and *fast*, in the sense that the manipulations that involve differentiations and some integrations are very easy to handle efficiently. The discovery of the Wiener-Khintchine theorem, in the context of Fourier transform theory, allows for an interesting and very useful exploration of the spectral properties of the function under consideration and permits us to evaluate convolutional sums through equivalent simple vector multiplications in the spectral domain.

A very serious problem though, that prohibited for a long time the wide use of the Fourier methods in practical applications, was the great difficulty involved in performing the integration that defines the transform itself. There are also many cases for which it is impossible to evaluate the transform analytically. Consequently, numerical evaluation ended up being the only feasible alternative; however, it was slow and impractical. A major breakthrough took place in 1965 with the introduction of the *F.F.T* algorithm (Cooley and Tukey, 1965), which improved greatly the speed in evaluating the Fourier coefficients. The *F.F.T* algorithm is, in fact, a fast way of calculating trapezoidal sums involved in approximating the Fourier integrals and it, therefore, requires input at equidistant points; the spectrum is also given in a regular manner. However, Fourier coefficients may be computed otherwise and general methods — to account for irregularly spaced abscissas and for points outside the fundamental interval — using offset trapezoidal rules, have been developed (Lynes, 1984).

For a piecewise continuous functions $f(x)$, which has bounded total variation, we define its Fourier series for $0 \leq x \leq 2\pi$ as the periodic function (Gottlieb and Orszag,

1977)

$$g(x) = \sum_{k=-\infty}^{+\infty} a_k e^{ikx} \tag{2.3}$$

where

$$a_k = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx}\, dx \tag{2.4}.$$

The complex Fourier series $g(x)$ has a period of 2π, and it exhibits the following convergence properties:

$$g(x) = \frac{1}{2}\left[f(x+) + f(x-)\right] \qquad \text{for} \qquad x \in [0, 2\pi] \tag{2.5}$$

and in particular

$$g(0) = g(2\pi) = \frac{1}{2}\left[f(0+) + f(2\pi-)\right] \tag{2.6}$$

If $f(x)$ is smooth (infinitely differentiable) and periodic, the Riemann-Lebesgue lemma implies that a Fourier series truncated to L terms, converges to $f(x)$ faster than any finite power of $1/L$ as L tends to infinity, for every x. Problems arise however, when either of the two previous assumptions is not met.

Violation of the first assumption, implies a discontinuity of a certain order. Violation of the periodicity assumption ($f(0+) \neq f(2\pi-)$) is in fact equivalent to the introduction of discontinuities at $x = 0$ and $x = 2\pi$.

The existence of a discontinuity results in a non-uniform convergence of the truncated Fourier series in the neighborhood of the discontinuity, which is well known as the *Gibbs' phenomenon.* Alternatively, the Gibbs' phenomenon may be explained by imagining an attempt at reproducing a function at a discontinuity by using linear

combinations of continuous functions (Yedlin, 1985) or through the side-lobes of a box-car filter when truncation of the data does not occur at the multiple of its fundamental period (Brigham, 1974). Tapering is usually applied to minimize the jump at the discontinuity (Kanasewich, 1981). Early applications of FFT to the solution of differential equations are given in Orszag (1971a) and Le Bail (1971).

Under periodic boundary conditions Fourier methods are the natural choice and they can provide excellent results, but there exist a variery of diverse factors that can cause a Fourier method to converge slowly or even diverge, due to the presence of the Gibbs' phenomenon. Among those we may easily identify:

a) Non-periodic boundary conditions;

b) Discontinuity in a higher derivative due to the particular form of the differential equation when assuming a periodic solution;

c) Improper imposition of periodic boundary conditions to a problem that is non-periodic by structure;

d) Conflict in the nature of the initial condition with the method applied.

2.2.4.2 The Chebychev transform

In contrast to Fourier polynomials, which are linear combinations of eigenfunctions of non-singular Sturm-Liouville problems, eigenfunctions of singular Sturm-Liouville problems have a convergence rate that is controlled only by the infinite differentiability of the function. In other words, we may drop the periodicity requirement or equivalently the series will not exhibit the Gibbs' phenomenon at the endpoints. Some of these eigenfunctions could be found among the Chebychev, the Legendre, the Laguerre, the Hermite and other polynomials.

Among those, the Chebychev and the Legendre polynomials appear to have been applied most often. Chebychev polynomials have the advantage that the transform can be evaluated fast and efficiently via an algorithm (Gentleman, 1972) that incorporates the FFT, whereas there is no fast Legendre transform known yet; a relatively fast Legendre transform has been written, nevertheless (Orszag and Kells, 1980). Even though they share almost similar properties, the distribution of the error is more uniform in the Chebychev case, whereas Legendre succeeds in showing smaller errors at the endpoints (Lanczos, 1973).

The *Chebychev polynomials of the first kind* are defined as

$$T_n(x) = \cos n(\cos^{-1} x), \qquad \text{for} \qquad -1 \leq x \leq 1 \tag{2.7}$$

or (under the transformation $x = \cos\theta$)

$$T_n(\cos\theta) = \cos n\theta, \qquad \text{for} \qquad 0 \leq \theta \leq \pi \tag{2.8}$$

where n is the order of the polynomial.

They satisfy the orthogonality relation

$$\int_{-1}^{+1} T_n(x)T_m(x)(1-x^2)^{-1/2}\,dx = \frac{\pi}{2}c_n\delta_{nm} \tag{2.9}$$

with $c_0 = 2$ and $c_n = 1$ for $n > 0$. In the classical least squares sense the Chebychev coefficients of $f(x)$ in the expansion $\sum_{n=0}^{\infty} a_n T_n$ are given from the formula

$$a_n = \frac{2}{\pi c_n} \int_{-1}^{+1} f(x) T_n(x) (1 - x^2)^{-1/2} \, dx \tag{2.10}$$

The convergence rate of the expansion depend only on the smoothness of $f(x)$ in the interior of the domain of definition. Consequently, Chebychev expansions do not exhibit the Gibbs' phenomenon at the endpoints, while they do at any internal discontinuity, i.e if the n-th derivative of the function $u(x,t)$ is discontinuous somewhere, then, we do not obtain infinite order accuracy but accuracy of the order of $O(\Delta x^n)$ in the neighborhood of the discontinuity, instead (Orszag and Jayne, 1974). A variety of properties and recursive relations regarding the Chebychev polynomials can be found in many relevant books (Gottlieb and Orszag, 1977, Fox and Parker, 1968, Rivlin, 1974).

It is important to point out here that the Chebychev points are not equidistant in x but clustered near the endpoints instead. Therefore, they would provide a much better representation of functions that change quickly in narrow boundary layers, than polynomials using an equidistant distribution of points. However, this high resolution advantage, may cause problems when it comes to time differencing (Hussaini et al, 1983, Orszag, 1980). Chebychev expansions need at least π polynomials retained per wavelength in order to converge fast. In other words, if a function oscillates over a distance λ, we expect to retain $2\pi/\lambda$ polynomials for rapid convergence. That number would be of course smaller if rapid changes of the functions take place only close to the boundaries (Gottlieb and Orszag, 1977).

The exponential convergence (in the absence of discontinuities, of any kind) of Chebychev (or any other "spectral" type polynomials), as the number of retained polynomials increases, may be alternatively explained as follows. Spectral methods provide a much better estimate of the larger true eigenvalues of the differential operators involved than finite difference methods do (Zang et al, 1982).

Another set of polynomials $\{U_n\}$, known as the *Chebychev polynomials of the second kind*, are orthogonal in the interval $[-1,+1]$ with respect to the weighting factor $(1-x^2)^{1/2}$ and they are defined as

$$U_n(x) = \frac{\sin(n+1)\theta}{\sin\theta} \tag{2.11}$$

and the corresponding coefficients of an expansion $f(x) = \sum a_n U_n(x)$ are given as

$$a_n = \frac{2}{\pi}\int_{-1}^{+1} U_n(x)(1-x^2)^{1/2} f(x) \tag{2.12}$$

Some of the early applications of Chebychev polynomials in economization problems are discussed in Lanczos (1938) and Miller (1946), while the books by Lanczos (1957) and Fox (1962) include detailed presentations of the method and examples. Norton (1964) and Scraton (1964) studied Chebychev methods for the solution of linear o.d.e's and Clenshaw and Norton (1963) and Wright (1964) investigated non-linear o.d.e's; integral Fredholm equations were studied as well (Elliot, 1963).

2.3 The Choice of Weight Functions

Three different choices of weight functions are available; the corresponding spectral methods are known as the *Galerkin*, the *collocation* and the *tau* methods. In the Galerkin or spectral method the weight functions are exactly the basis functions used in the series expansion. The collocation or pseudospectral method employs weight functions which are shifted Dirac delta functions defined over a set of collocation points, on which the numerical and the exact solution are identical. The tau method could be viewed as a simplified version of the Galerkin method, because the weight functions are again the same as the basis functions. However, its particular treatment of the boundary constraints differentiates it from the latter.

The previous short description of our three options in choosing the weight functions does not reveal some of the essential characteristics of each one of them that are encountered in practice. All the methods work under the assumption of homogeneous boundary constraints. If this is not so, the differential equation needs to be modified so that the solution satisfies equivalent homogeneous boundary conditions.

In both the Galerkin and the collocation methods, all the basis functions have to satisfy the given boundary conditions on an individual basis. This is not a requirement for implementing the tau method. The above requirement makes the algebra more cumbersome in the first two methods (since we need to find a suitable but almost always non-orthogonal set of basis functions), and it therefore necessitates the introduction of relatively sophisticated tricks for the optimization of the resulting set of simultaneous algebraic equations that is developed in subsequent stages of the process (Orszag, 1971c). This superiority of the tau method should be credited to its special treatment of the boundary constraints (Lanczos, 1957). The higher modes of the expansion

are not considered at all; instead, the boundary constraints are substituted for them (Orszag, 1971d).

The tau method then emerges as an optimum choice for constant coefficient problems, although it is not expected to perform equally well for more complicated problems, such as variable coefficient or nonlinear ones. Special direct (full-diagonalization) or iterative ($A.D.I$) solvers may be devised to handle 2 or 3-D Chebychev tau formulations (Haidvogel and Zang, 1979; Haldenwang et al, 1984). Collocation appears to be an obvious choice here, since it simplifies the complicated algebra without losing the high mode contribution, the latter becoming significant in complicated problems. Although the pseudospectral method exhibits such an advantage, it suffers from introducing aliasing terms at full strength, a feature that might cause an inherent instability (Orszag, 1972; Fox and Orszag, 1973). On the other hand, Galerkin methods do not include aliasing terms but require complicated manipulations and more computational work (Orszag, 1969, 1970; Patterson and Orszag, 1971; Price and Varga, 1974). The Galerkin and the tau methods could also be viewed (at least in simpler problems) as equivalent to collocation methods with a different distribution of the collocation point set.

2.4 Problems, Advancements and Current Trends in Fourier and Chebychev Methods

Although spectral methods are particularly suited for simple geometries, efficient transformations might be employed to adjust them for distorted non-uniform domains (Orszag, 1980; Mc Crory and Orszag, 1980); an on-going research in this direction seems quite promising.

Pseudospectral schemes have been attractive since the early developments of spectral methods (Merilees, 1973; Schammel and Elsasser, 1976; Haidvogel, 1977), and they have also been applied to seismic migration (Pann et al, 1979). Subsequently, they have been progressively gaining in numericists' preference; this is especially pronounced for pseudospectral Chebychev approximations.

The main reason for such a dominance might be found in the easy handling of operators involving nonconstant coefficients, nonlinearities or non-smooth geometries, in the pseudospectral framework; complicated problems in flow and air dynamics have been handled satisfactorily (see 2.5 and 2.6). Pseudospectral Chebychev techniques have also performed very well for difficult domains as laminar heat transfer in pipelines (Hatziavramidis and Ku, 1983), the extended Graetz problem (Ku and Hatziavramidis, 1984), or axisymmetric flow in a heated, rotating spherical shell (Macaraeg, 1986). Galerkin or tau methods require excessive amounts of computations on similar complicated problems (Deville and Mund, 1985).

Pseudospectral Fourier semi-discretizations have already been applied in seismic migration; the transform provides an accurate evaluation of the spatial derivatives appearing in the equations. Gazdag (1981) employed a third-order Runge-Kutta scheme, whereas Kosloff and Baysal (1983) used a leap-frog (second-order accurate) scheme for the time differencing.

Zebib (1984) proposed an interesting variant of Galerkin's method: the higher derivative of the differential equation under consideration is expanded in Chebychev polynomials and the lower derivatives and the function itself are evaluated via successive integration of recursion formulae. The aim is to obtain higher accuracy than the tau method for the same amount of computations and satisfy non-homogeneous

conditions on the function and its derivatives directly; the manipulations involved, though, tend to become very complicated and obscure.

Finally, we should mention that among other applications, integral equations with weak singularities have been solved with Chebychev methods (Frenkel, 1983a) and the set of U_n's has also been used for both p.d.e's (Shoucri and Knopr, 1974) and integral equations (Frenkel, 1983b). In addition to these, the *shifted* Chebychev polynomials $T_n(x^2)$ have been successfully applied for the solution of Sturmian eigenalue equations (Delic and Rawitscher, 1985), while Boyd (1987) has studied the implementation of rational Chebychev basis functions in expansions on an infinite interval.

A brief but very enlightening review of the theory and the potential applications of the spectral methods could be found in Orszag and Israeli (1974), whereas an update with particular emphasis on pseudospectral techniques can be found in Gottlieb et al (1984b).

2.5 Time Integration in Spectral Methods

Although the above discussion provides a comprehensive analysis of the spectral methods, it is not complete in the sense that it does not cover mixed initial boundary value problems.

The approach mostly used, combines spectral representations of the spatial coordinates while maintaining some kind of finite difference scheme in time (finite elements may provide an equivalent or superior level of accuracy (Ku and Hatziavramidis, 1984)). Nevertheless, time-spectral expansions have been presented but they have not found wide applications (Morchoisne, 1984; Tal Ezer, 1984). The particular choice of

this scheme is not straightforward; many interrelated factors have to be considered so that an optimum choice for the specific p.d.e will be chosen.

2.5.1 Explicit schemes

Explicit time integration schemes enjoy easy formulation and one-step storage requirements; they do not require matrix inversions and when using them we can take advantage of the spectral representation and evaluate the derivatives either via recurrence formulae (tau) or via the FFT algorithm (Galerkin and pseudospectral). Despite these advantages, explicit schemes suffer either from unconditional instability (e.g second-order Runge-Kutta for Fourier methods, leap-frog for Chebychev problems) or from restrictive conditional stability which imposes severe limitations on the size of the time step Δt (Gottlieb and Orszag, 1977). Consequently, long-scale calculations become inefficient or even prohibited. The source of this problem can be linked to one of spectral methods' main advantages, namely, their high resolution. This is especially pronounced in the Chebychev case due to the *clustering* of the Chebychev nodal points near the boundaries; for the usual set of points $x_j = \cos(\pi j/N), j = 0, \ldots, N$, we see that the points x_1 and x_{N-1} are within about $\pi^2/2N^2$ of the boundary points x_0 and x_N respectively and therefore, the resolution there is $\Delta x = O(1/N^2)$. This severe limitation is especially annoying when we are dealing with problems that do not exhibit strong boundary-layer structure (Orszag, 1980).

2.5.2 Implicit schemes

The next obvious alternative is *implicit* schemes. Spectral implicit schemes allow much larger time steps to be taken but their employment for time-integration is hindered by some important drawbacks. The representations of the spectral operators are full, ill-conditioned matrices which cannot be easily inverted; the formulation becomes obscure and major problems arise in applying fast spectral algorithms for the evaluation of the derivatives. Furthermore, as the dimensionality of the problem is increased, formulations become complicated and the storage requirements and computational work using direct methods (e.g Gaussian elimination) become prohibitively high. One significant exception exists for implicit (usually Crank-Nicolson) tau schemes which exhibit a quasi-tridiagonal structure and as a result, they are amenable to a direct and fast inversion through a special modification of the classic LU decomposition (Golub and van Loan, 1983) tailored to their structure. This algorithm is readily expandable in higher dimensions, but unfortunately this conveniently exploitable structure is lost even for a linear but non-constant coefficient problem (Hussaini et al, 1984). Neither fully explicit nor fully implicit methods seem to provide a satisfactory answer to the time-differencing issue.

2.5.3 Semi-implicit and hybrid explicit-implicit schemes

Combinations of explicit and implicit schemes are viable alternatives. The motivation for such *semi-implicit* schemes is the implicit treatment of the boundary regions (where the high resolution occurs), while treating the interior of the computational domain explicitly (Gottlieb and Orszag, 1977). Mixed equations might result in very stiff systems; *hybrid explicit-implicit* schemes might be then applied so that various

terms are treated differently (Drummond et al, 1985). Another approach involves a finite difference predictor (implicit) combined with a spectral corrector (explicit). The idea here is to employ finite-differences in order to decrease the density of the matrix allowing in that way an easier inversion; the implicit nature of the scheme helps stabilizing the algorithm and permitting the use of larger time steps. That allows us (using a modified-Euler corrector scheme, for example), to increase the size of the time step to $O(1/N)$ as compared to $O(1/N^4)$ when applying Chebychev methods to equations involving second order derivatives (Taylor, 1984).

Such schemes have also been the answer to the numerical difficulties associated with employing explicit Chebychev integration for the Navier-Stokes equations; natural boundary conditions for the pressure are missing for incompressible flows and the classic ways of specifying them (proven to be successful in a finite differences environment) fail. The Adams-Bashworth-Crank-Nicolson ($ABCN$) scheme (Moin and Kim, 1980) overcomes the difficulties by resorting to a well-posed problem, i.e treatment of the problem as an initial boundary value one, as opposed to a simply initial value. Direct extension of this scheme to higher dimensions is cumbersome and time-splitting techniques have been augmented in the numerical process (Orszag and Kells, 1980); however, the new scheme exhibits a low order accuracy in time, i.e $O(\Delta t)$. Recently, an influence matrix technique for the boundary conditions specification has been incorporated into the scheme (Dennis and Quartelle, 1983; Le Quere and De Roquefort, 1985).

2.5.4 Unconditionally stable explicit schemes

Another alternative (in hyperbolic problems) is the use of unconditionally stable explicit schemes, where the size of Δt to be used, depends on accuracy requirements only; Δt still needs to remain fairly small (however, much bigger than the original stringent limit) in order to maintain good accuracy levels. The idea is to employ a purely explicit scheme (e.g Runge-Kutta, leap-frog, modified Euler) but to couple the temporal and spatial representations, instead. This is done by the application of an appropriate filter (whose choice depends both on the scheme used and the stiffness of the problem), so that phase-errors remain under control (Gottlieb and Turkel, 1980).

This approach has proven to work quite satisfactorily for the Fourier case but its success for the difficult Chebychev case remains controversial (Gottlieb et al, 1984b). It has been argued that the Gottlieb-Turkel filter does not result into absolutely stable schemes. Numerical investigation shows that limited filtering does not alter the absolute stability properties, whereas extended filtering causes the scheme to be absolutely unstable for any Δt. In addition, artificial dispersion, similar to the one introduced by finite differences, has also been observed. In more complicated problems, it is expected that although filtering would stabilize the fast modes, it would simultaneously distort the slow modes to a significant level, causing deterioration of the accuracy of the numerical method (Fulton and Taylor, 1984).

A powerful alternative, mainly for parabolic differential equations, has been found in modified versions of the Dufort-Frankel explicit scheme. Again here a controllable parameter γ is incorporated in the scheme so that phase-errors are damped during successive time steps. Consequently, time step restrictions are (at least in theory) relaxed

and only accuracy considerations are taken into account (Gottlieb and Gustafsson, 1976; Gottlieb and Lustman, 1983a).

2.5.5 Iterative techniques

The pronounced difficulties encountered due to the serious Δt limitations described in the foregoing paragraphs, are most important for those classes of equations in which the temporal behavior of the solution takes place on a much smaller scale than the spatial one. Similar considerations concern cases with solutions exhibiting temporal and spatial behavior occurring on the same scale.

Another class of problems includes equations for which we are simply interested in steady-state solutions. Then a technique which would converge fast in the steady-state solution (although being maybe inaccurate initially) would be quite adequate. Direct solvers are definitely impractical for the matrix inversions emerging in either implicit formulations or in steady-state problems. *Iterative* schemes have been proven to be the answer to this problem. The fundamental concepts involved in the iterative solution of those systems are presented briefly below (Gottlieb et el, 1984b).

Let $L(u) = f$ where L is the spectral matrix, u the approximate solution and f a known vector. Iterative solution of this system proceeds according to

$$u^{n+1} = u^n - H^{-1}[L(u^n) - f] \tag{2.13}$$

where H is a sparse, efficiently invertible matrix that approximates the Jacobian J_L of L.

Iterative schemes based on this formula are classified as stationary; various choices of H (called the preconditioning operator) define particular schemes exhibiting different convergence properties. In Jacobi's method $H = \mathrm{diag}(\mathrm{J_L})$, whereas in Gauss-Seidel's method H is the lower triangular part of J_L.

Nonstationary iterative schemes (e.g preconditioned Richardson's method) are based on

$$u^{n+1} = u^n - \alpha_n H^{-1}[L(u^n) - f] \tag{2.14}$$

where appropriate α_n's improve the convergence rate.

Nonstationary second-degree schemes based on polynomial acceleration provide further improvements. The iterations are done according to

$$u^{n+1} = \omega_n u^n - \alpha_n \omega_n H^{-1}[L(u^n) - f] + (1 - \omega_n)u^{n-1} \tag{2.15}$$

Different definitions of ω_n distinguish schemes based on the above formula (e.g Chebychev acceleration and conjugate gradient methods). All these iterative methods are heavily affected from the choice of the preconditioner H.

Routinely, a low-order finite difference approximation H_{FD} to L is chosen (Orszag, 1980); special remedies might be required for cases where the norm $||H_{\mathrm{FD}}^{-1}L||$ is unbounded (Hussaini and Zang, 1984). An increase in the dimensionality of the problem, though, imposes serious computational considerations in the evaluation of H_{FD}^{-1} and alternative preconditioners might have to be employed. Existing algorithms perform approximate factorizations or incomplete LU decompositions of H_{FD} to obtain a suitable operator fast (Zang et al, 1982, 1984). Minimal residual (MR) techniques can

be combined with relaxation schemes (e.g Richardson, Chebychev) to give a robust parameter-free scheme whose convergence is guaranteed subject to the condition that all the eigenvalues of $H_{\mathrm{FD}}^{-1}L$ lie in the right half of the complex plane (Wong et al, 1983).

Finite element matrices have been proposed for preconditioning spectral iterative schemes, as well. They exhibit properties similar to the finite difference preconditioners, being moreover symmetric (Canuto and Quarteroni, 1985). Finite elements provide for geometrical ease in handling irregular boundaries; the use of Lagrangian bilinear elements result in very sparse matrices, lowering the inversion cost with respect to the use of higher degree finite elements (Deville and Mund, 1985).

2.5.6 Spectral multigrid methods

Preconditioning is extremely crucial and much of the current research is devoted to preconditioning in spectral multigrid (SMG) techniques that have recently been introduced. *Multigrid* methods (MG) can achieve a dramatic acceleration of iterative techniques (Zang et al, 1982, 1984), as they take advantage of a common property of most relaxation schemes, namely, the potential efficient reduction of the high-frequency error components but inevitable slow reduction of the low-frequency components. The MG method differs from conventional solvers, in the sense that the discretization and solution processes are intermixed and they benefit from each other. A sequence of grids with varying mesh spacing is used. Relaxation is performed in each grid and various interpolation techniques are applied for transfer of data either in a coarse to fine or a fine to coarse grid fashion (Grinstein et al, 1983).

Pseudospectral applications to both subsonic (Euler equations) and transonic (full potential equations) inviscid flows have been investigated in association with MG methods. The double character of the transonic case (subsonic-supersonic) needs highly sophisticated time-differencing schemes as high levels of stiffness are present hindering straightforward formulations (Drummond et al, 1985; Street et al, 1985; Hussaini and Zang, 1984). The Douglas-Gunn alternating direction implicit ($A.D.I$) iterative method seems to perform quite satisfactorily for the foregoing problems in two and three dimensions.

Pseudospectral Fourier methods based on the FFT algorithm can be quite subtle. A fine point is associated with the evaluation of the derivatives in transform space. The highest (Nyquist) mode should be disregarded in the process to avoid contamination and divergence of the iterative MG scheme (two equivalent explanations of that could be found in Zang et al, 1982 and Brandt et al, 1985). This is known as the *two point oscillation phenomenon.* However, elimination of the Nyquist component might degrade the accuracy if a lot of information lies there; as an alternative Brandt et al (1985) performed the calculation of the derivative at the mid-points of the original collocation set. Thus, accuracy is preserved, at the expense of a small increase in the computational work needed.

2.6 Shock Handling in Spectral Methods

The transonic case is challenging since it involves a mixed type (hyperbolic- elliptic) equation, which further complicates an already difficult situation (inviscid compressible flows give rise to *shock* waves, that is to say, to *discontinuous* solutions). Shocks

require special attention: if a shock occurs in the interior of the computational domain, then the spectral methods' accuracy and convergence rate tends to deteriorate (shock capturing); for shocks occurring at the boundaries we do not need to worry (shock-fitting). A variety of filtering procedures can be applied to avoid the Gibbs phenomenon associated with the shock's formation (Gottlieb et al, 1984a, Hussaini et al, 1983). Evidently, in similar cases involving discontinuities, smoothing should play an integral role in the formulation of spectral methods so that they maintain stability and high levels of accuracy (Osher, 1984).

CHAPTER III

THE HELMHOLTZ EQUATION

The stars grew dim, the sky grew light and against this luminous background appeared, as if delicately traced in ink, the mountains, trees and gulls.

Dawn was breaking.

Zorba the Greek — Nikos Kazantzakis

3.1 Physical Aspects of the Helmholtz Equation

In many applications involving evolutionary problems in time, we solve the time independent equations directly; we usually seek a time-harmonic solution. The *Helmholtz* equation is obtained under the assumption of an $\exp(-i\omega t)$ harmonic time-dependence of the solution and in three dimensions it reads

$$\nabla^2\Phi + k^2\Phi = 0 \tag{3.1}$$

Depending on whether the Helmholtz equation is considered in the exterior or the interior of a body, it describes the scattering of the waves by the body or the propagation of the waves in it.

A variety of boundary conditions are applied to (3.1); their form depends on the dimensionality, the geometry and special physical aspects of the particular problem. In general, though, proper boundary conditions on the physical surfaces and an appropriate (depending on the nature and the geometry of the problem) radiation condition at infinity are essential (Bayliss and Turkel, 1982).

3.2 Numerical Aspects of the Helmholtz Equation

Wave propagation considerations involve at least three characteristic quantities associated with the wavelength (Bayliss et al, 1985).

1) The quantity $(k\Delta x)^{-1}$ measures the number of grid-points per wavelength and it is widely used as a measure of accuracy.

2) The quantity $k\ell$ (ℓ is the diameter of the computational region) gives the number of wavelengths in the computational domain. Although usually not considered, this quantity has an impact on the discretization error. The latter increases linearly with k, despite the fact, that the number of points per wavelength remains constant (Bayliss et al, 1985).

3) The quantity ka describes the number of wavelengths in the region in which k is spatially varying (a is the characteristic length of the inhomogeneous region).

Any kind of discretization leads to a large linear system of equations; this system becomes bigger as k increases because the solution becomes more oscillatory and,

therefore, finer sampling rates (or equivalently higher-order interpolation polynomials) are required. LU decomposition takes excessive amounts of work to solve those systems and consequently, we either have to limit ourselves to moderate values of k or to resort to iterative schemes.

The Helmholtz equation is difficult to solve using standard iterative methods (e.g SOR, Jacobi, conjugate gradient), because it allows both positive and negative eigenvalues for the difference operator (Nicolaides, 1979). That means that the Helmohltz operator is often indefinite. This happens when k^2 is larger than the smallest eigenvalue of the discrete approximation to the operator $-\nabla^2$ (Bayliss et al, 1983). That causes the low-frequency components of the error to be amplified leading to a divergence of the iterative scheme (Grinstein et al, 1983).

A variety of optimizations of these schemes has been proposed that work quite well for valus of k^2 that lead to slight indefiniteness (Nicolaides, 1979); preconditioning has been shown to accelerate the iterative procedures even for highly indefinite cases (Bayliss et al, 1983). Optimization has to circumvent additional problems introduced when radiation boundary conditions are used. These boundary conditions involve complex constants which destroy the self-adjointness of the Helmholtz matrix — a property that is a fundamental requirement in most iterative methods — (Bayliss et al, 1983).

3.3 The One-Dimensional Helmholtz Equation

The equation is

$$\frac{d^2}{dx^2}y(x) + k^2 y(x) = 0 \tag{3.2}$$

and two boundary conditions are applied to it. These may be *Dirichlet* (b.c's on the function itself), or *Neumann* (b.c's on the derivative of the function), or even *Robbins* (b.c's on linear combinations of both the function and its derivative). They will be either homogeneous or not. If we are applying the tau method we need not worry about this (see 3.3.1). For both the Galerkin and the collocation methods, though, we have to transform equation (3.2) so that it satisfies homogeneous boundary conditions. This is done by subtracting a suitable polynomial from $y(x)$ so that the updated function $g(x)$ obeys the desired homogeneous boundary conditions; equation (3.2) will be transformed as well. It is understood that after the updated function $g(x)$ has been computed, we add the subtracted polynomial to get the solution $y(x)$. For Dirichlet boundary conditions the polynomial is linear, whereas for Neumann is a quadratic function. Keeping the foregoing in mind, we continue assuming that the equation satisfies homogeneous boundary conditions.

3.3.1 The tau method

The expansion has to be done in terms of a set of orthogonal polynomials independently of the boundary constraints to be satisfied. The natural choice is, of course, the orthogonal set of the $T_m(x)$'s. We write $y(x) = \sum_{m=0}^{N} a_m T_m(x)$ retaining $N+1$ components of the infinite expansion. Substituting that in (3.2) and taking the inner product with each one of these T_n's, we project the equation onto the function space spanned by the $T_n(x)$'s . The resulting system reads

$$\sum_{n=0}^{N} \langle T_n, T_m \rangle + k^2 \sum_{n=0}^{N} a_n \langle T_n, T_m'' \rangle = 0 \quad \text{for } n = 0, \ldots, N \tag{3.3}$$

This involves the calculation of the inner products $\langle T_n, T_m \rangle$ and $\langle T_n, T_m'' \rangle$ (T_m'' denotes the second derivative of the Chebychev polynomial T_m), but before dealing with their calculation, we need to take care of the overdeterminacy that characterizes our system.

This system is, indeed, overdetermined because it has two more equations coming from the boundary conditions. General Robbins boundary conditions may be written as

$$\alpha_{\pm} u + \beta_{\pm} u_x = \gamma_{\pm} \quad \text{for } x = \pm 1 \tag{3.4}$$

Substituting the Chebychev series for $y(x)$, performing the differentiation and using the result $T_n'(\pm 1) = (\pm 1)^{n-1}\, n^2$, we obtain the two boundary equations

$$\sum_{n=0}^{N} (\pm 1)^n \left[\alpha_{\pm} \pm \beta_{\pm} n^2\right] a_n^{(0)} = \gamma_{\pm} \tag{3.5}$$

Deleting the equations for $n = N-1, N$ we end up having $N+1$ equations ($N-1$ from the projection operation plus two from the associated boundary conditions) to solve for the $N+1$ unknown coefficients of the original expansion for $y(x)$.

There are three ways of implementing the tau method; these are the direct, the differentiated and the integrated tau methods and we now proceed to present them in detail.

3.3.1.1 The direct system

Two inner products appear in (3.3). The first is simply the orthogonality relation between the Chebychev polynomials and it reads $(\pi/2) c_m \delta_{mn}$. The second is not easy to calculate directly and alternative procedures have been proposed. The differentiated

system (see 3.3.1.2) suffers from further truncation errors and might be susceptible to instabilities. The integrated system (see 3.3.1.3) may be cumbersome to apply, if analytic integration is prohibitive due to complicated coefficients present in the equation of interest.

As a consequence, a semi-analytic expression for the value of the integral has been developed (see Appendix A.1) and the appropriate computer algorithm written; although quite fast, this algorithm is doomed to be superseded by the other methods.

3.3.1.2 The differentiated system

This second implementation of the tau method tries to bypass the difficulties of evaluating the inner product $\langle T_n, T_m'' \rangle$ by expressing the second order derivative of the function as another Chebychev expansion $\sum_{n=0}^{N-2} a_m^{(-2)} T_m(x)$; the coefficients $a_m^{(-2)}$ can be found analytically as linear combinations of the original coefficients (equation A.40). This approach has been almost exclusively used in the literature. Although fast, convenient and well established, it results in matrix representations of differential operators that are not very stable. There might be cases where the results would be doubtful because of unacceptably high accumulation of round-off error. Equation (3.1) constitutes such a case. The reason may be that this is a *zero eigenvalue problem*; as the value of k^2 increases, it suppresses the diagonal of $\left(d^2/dx^2 + k^2\right)$ leading to a singular matrix. It is obvious that we must be careful in setting up an d^2/dx^2 matrix, which is as stable as possible.

Following this approach, we substitute the corresponding Chebychev expansions

for $y(x)$, $y''(x)$ in (3.1); this results in

$$\sum_{n=0}^{N-2} a_m^{(-2)} T_m(x) + k^2 \sum_{n=0}^{N} a_m^{(0)} T_m(x) = 0 \tag{3.6}$$

Completing the *tau* projection procedure we obtain

$$a_n^{(-2)} + k^2 a_n^{(0)} = 0 \qquad \text{for} \qquad 0 \leq n \leq N-2 \tag{3.7}$$

or

$$\sum_{\substack{p=n+2 \\ p+n:even}}^{N} p\,(p^2 - n^2)\, a_p^{(0)} + k^2\, a_n^{(0)} = 0 \quad \text{for } 0 \leq n \leq N-2 \tag{3.8}$$

and the boundary conditions

$$\sum_{n=0}^{N} (\pm 1)^n \left[\alpha_\pm \pm \beta_\pm n^2\right] a_n^{(0)} = \gamma_\pm \tag{3.9}$$

3.3.1.3 The integrated system

Problems susceptible to round-off (when the previous method is applied) ask for an advanced treatment. In these cases special stabilization procedures must be employed (Gottlieb and Orszag, 1977), or equivalently the integrated version of the tau method should be used (Fox and Parker, 1968).

The relevant procedure commences by integrating the equation twice, to obtain

$$y(x) + k^2 \int dx \int y(x)\, dx + Ax + B = 0 \tag{3.10}$$

where A, B are integration constants.

Substituting the Chebychev expansions for $y(x)$ and its second anti-derivative (where the coefficients $a_m^{(+2)}$, $m = 0, N+2$ are given as linear combinations of the $a_m^{(0)}$'s according to equation (A.49)) yields

$$\sum_{m=0}^{N} a_m^{(0)} T_m + k^2 \sum_{m=0}^{N+2} a_m^{(+2)} T_m + A\, T_1 + B\, T_0 = 0 \tag{3.11}$$

since $T_1 = x$ and $T_0 = 1$.

Taking then the inner product of the above expression with each one of the T_n's for $n = 0, \ldots, N+2$, gives a system of $N+3$ equations (from the projection, only).

The equations for $n = 0, 1$ contain the constants A, B and the coefficients $a_0^{(+2)}, a_1^{(+2)}$ which require knowledge of $\int y\, dx\, |_{x=-1}$ and $\int dx \int y\, dx\, |_{x=-1}$. These equations are, therefore, dropped. Proceeding, we examine the last two equations (for $n = N+1, N+2$); they seem to imply that $a_{N+1}^{(+2)} = a_{N+2}^{(+2)} \equiv 0$.

Employing expression (A.49), we find that

$$a_{N+1}^{(+2)} = \frac{c_{N-1} a_{N-1}^{(0)}}{4N(N+1)} \quad \text{and} \quad a_{N+2}^{(+2)} = \frac{c_N a_N^{(0)}}{4(N+1)(N+2)} \tag{3.12}$$

which, subsequently, require $a_{N-1}^{(0)} = a_N^{(0)} = 0$ for consistency.

Therefore, the tau projection is accomplished by disregarding the equations for $n = N+1$, $N+2$ and eliminating, simultaneously, $a_{N-1}^{(0)}$, $a_N^{(0)}$ in the projection equations. This is followed by substitution of these equations with the ones expressing the boundary conditions; the system is now fully determined.

That amounts to a completion of the tau projection and, thereafter, the final system reads

$$\frac{k^2 c_{n-2}}{4n(n-1)} a_{n-2}^{(0)} + \left(1 - \frac{k^2 e_{n+2}}{2(n^2-1)}\right) a_n^{(0)} + \frac{k^2 e_{n+4}}{4n(n+1)} a_{n+2}^{(0)} = 0, \qquad 2 \leq n \leq N \quad (3.13)$$

with $c_0 = 2$ and $c_n = 1$ for $n > 0$ and $e_n = 1$ for $n \leq N$, $e_n = 0$ for $n > N$ and the same boundary conditions.The same system has been obtained (Gottlieb and Orszag, 1977), by transforming results of the differentiated system as well.

The integrated system claims a stable structure (due to strong diagonal dominance), but it enjoys an additional advantage, as well. The system is quasi-tridiagonal, that is to say, the matrix is tridiagonal except for two rows; these come from the boundary conditions and are usually full. Although, this latter property of the integrated system is not of crucial importance, for the one-dimensional and time-independent Helmholtz equation under current consideration, it is extremely helpful in problems of higher dimensionality and in problems where time-dependence is present (see 4.3.2).

3.3.2 The Galerkin method

Considering now the Galerkin method, we are faced with the requirement of homogeneous boundary constraints; for non-homogeneous problems transformations are needed. If we are dealing with a non-homogeneous Dirichlet problem, that is to say,

$$\frac{d^2}{dx^2} y(x) + k^2 y(x) = 0 \quad \text{and } y(+1) = \alpha, y(-1) = \beta \qquad (3.14)$$

we define $g(x) = y(x) - (a + bx)$, where $a = (\alpha + \beta x)/2$ and $b = (\alpha - \beta)/2$. This results in a homogeneous Dirichlet problem

$$\frac{d^2}{dx^2}g(x) + k^2 g(x) = -k^2\ (a + bx) \quad \text{and } g(\pm 1) = 0 \tag{3.15}$$

This is solved and finally $y(x)$ is given as $g(x) + (a + bx)$.

If the problem is of non-homogeneous Neumann structure, namely,

$$\frac{d^2}{dx^2}y(x) + k^2 y(x) = 0 \quad \text{and } y_x(+1) = \gamma, y_x(-1) = \delta \tag{3.16}$$

we define $h(x) = y(x) - (dx + ex^2)$, where $d = (\gamma + \delta)/2$ and $e = (\gamma - \delta)/4$. That results in a homogeneous Neumann problem

$$\frac{d^2}{dx^2}g(x) + k^2 g(x) = -[2e + (k^2\delta)x + (k^2 e)x^2] \quad \text{and } g(\pm 1) = 0 \tag{3.17}$$

Adding $dx + ex^2$ back to $g(x)$, after solving the updated equation, completes the process.

Next comes the problem of choosing the appropriate basis functions (as a linear combination of Chebychev polynomials), so that the homogeneous boundary conditions (whatever the kind) are satisfied by each one of the elements of the set. For the Dirichlet case, a suitable set, consisted of the polynomials $q_n(x)$, with $n \geq 2$, is defined as

$$q_n(x) = \begin{cases} T_n(x) - T_0(x), & \text{for } n \text{ even;} \\ T_n(x) - T_1(x), & \text{for } n \text{ odd} \end{cases} \tag{3.18}$$

since $q_n(\pm 1) = 0$, always.

For the Neumann case, a suitable set is

$$q_{2n}(x) = T_{2n}(x) - n^2 T_2(x) + (n^2 - 1)T_0(x) \tag{3.19}$$

for $n \geq 2$, since $(q_n)_x(\pm 1) = 0$; incidentally these latter q_n's also satisfy $q_n(\pm 1) = 0$ (Orszag, 1971d).

For mixed boundary conditions (Robbins), i.e

$$\alpha u + \beta u_x = 0 \quad \text{at } x = -1 \quad \text{and} \quad \gamma u + \delta u_x = 0 \quad \text{at } x = +1 \tag{3.20}$$

the above defined set is still applicable; another suitable set $\{p_n\}$ $n = 4, \ldots, N$ may be defined as

$$p_n = \begin{cases} T_{2n+1} - 2T_{2n-1} + T_{2n-3} - T_3 + T_1, & \text{if } n \text{ is odd and for } n \geq 2; \\ T_{2n} - 2T_{2n-2} + T_{2n-4} - 2T_2 + 2T_0, & \text{if } n \text{ is even and } n \geq 2. \end{cases}$$

since $p_{2n}(\pm 1) = p_{2n+1}(\pm 1) = (p_{2n})_x(\pm 1) = (p_{2n+1})_x(\pm 1) = 0$ (Hatziavramidis and Ku, 1983).

Let us now assume that non-homogeneous Dirichlet boundary conditions are given. The problem is transformed and

$$g(x) = \sum_{m=2}^{N} \bar{a}_m^{(0)} q_m(x) \tag{3.22}$$

is substituted into it. This formulation leads to the system

$$\sum_{m=2}^{N} \bar{a}_m^{(0)} \left[\langle q_n, q_m'' \rangle + k^2 \langle q_n, q_m \rangle \right] = -k^2 \left[a \langle q_n, T_0 \rangle + b \langle q_n, T_1 \rangle \right], \qquad n = 2, \ldots, N \quad (3.23)$$

3.3.2.1 The direct system

The above set of equations includes $N-1$ equations for the $N-1$ coefficients $\bar{a}_2^{(0)}, \ldots, \bar{a}_N^{(0)}$ of the expansion.

The inner products appearing in the Galerkin formulation are decomposed into inner products involving the original Chebychev polynomials and their second derivatives. The open form solution is then applied and the resulting linear system is solved to obtain the coefficients $\bar{a}_2^{(0)}, \ldots, \bar{a}_N^{(0)}$.

Unfortunately, these coefficients cannot be used directly as input for the (inverse) fast Chebychev transform to evaluate the sum $g(x) = \sum_{m=2}^{N} \bar{a}_m^{(0)} q_m(x)$. The following transformation alleviates the problem: the function $g(x)$ is written as

$$g(x) = \sum_{m=0}^{N} a_m^{(0)} T_m(x) \quad (3.24)$$

and equivalencing that with the expansion

$$g(x) = \sum_{m=2}^{N} \bar{a}_m^{(0)} q_m(x) \quad (3.25)$$

we deduce that

$$a_0^{(0)} = -\sum_{n=1}^{N/2} \bar{a}_{2n}^{(0)} \quad \text{and} \quad a_1^{(0)} = -\sum_{n=1}^{N/2-1} \bar{a}_{2n+1}^{(0)} \tag{3.26}$$

where the rest of the these new coefficients are identical to the old ones, namely, $a_m^{(0)} = \bar{a}_m^{(0)}$ for $m = 2, \ldots, N$. The vector $\mathbf{a}^{(0)}$ can be efficiently mapped back onto x space to give $g(x)$. The addition of $a + bx$ terminates the procedure, giving $y(x)$.

3.3.2.2 The indirect (differentiated) system

This method bypasses the direct evaluation of the inner products, by employing the transformation discussed above in the early stages of the solution process. The differentiated system method's fundamental character constitutes an *orthogonalization* of the basis function set $\{q_n\}$. The members of this set enjoy the desired feature of satisfying the boundary constraints on an individual basis. However, this advantage is counterbalanced from the fact that $\{q_n\}$ is a *non-orthogonal* set and therefore, a projection would involve coupling among the various modes q_n. This non-orthogonality property of this set is rather repulsive, as it tends to complicate the formulation as seen previously. Orthogonalization of the set is definetely desirable and this is the essence of the procedures presented below. Of course, the original orthogonal set $\{T_n\}$ is the ultimate choice for the new basis function set.

The orthogonalization transformation might be visualized as being applied either at the pre- or the post-projection stage of the process. The post-projection version is of theoretical importance only, as it does not avoid evaluating the inner products $\langle q_n, q_m'' \rangle$.

Although an impractical procedure, it is presented, as it is easier to comprehend and it provides a rigorous justification of the pre-projection version given later on.

The *post-projection* version is nearly identical to the *direct* method. A subtle difference does exist and this is exactly the detail that justifies the orthogonalization. After the projection has been completed the coefficients $\bar{a}_m^{(0)}$ are coupled due to both the non-orthogonal nature of the set $\{q_n\}$ and the presence of the second derivative. The presence of the terms like

$$-\sum_{n=1}^{N/2} \bar{a}_{2n}^{(0)} \quad \text{and} \quad -\sum_{n=1}^{N/2-1} \bar{a}_{2n+1}^{(0)},$$

is due to coupling of non-orthogonal components or, equivalently, due to modification of the $\{T_n\}$ set (with the simultaneous elimination of two degrees of freedom), so that each element of the new set would satisfy the boundary conditions. These alterations involved the elimination of the polynomials $T_0(x)$ and $T_1(x)$ as individual components; concurrently, these polynomials are intermixed with the higher order Chebychev polynomials, so that the $\{q_n\}$ set can be constructed. The idea is, then, to return into an expansion of the orthogonal form $g(x) = \sum_{m=0}^{N} a_m^{(0)} T_m(x)$; the coupling sums would, subsequently, identify themselves as the coefficients $a_0^{(0)}$, $a_1^{(0)}$ of the autonomous components $T_0(x)$ and $T_1(x)$, respectively. The recovery of the two degrees of freedom calls for an augmentation of the system with two additional equations for the new coefficients. Finally, the constraints imposed on the coeffient set (being manifested in the above sums) need to be appropriately incorporated in the new system, so that their fullfillment is guaranteed.

Let us now discuss the *pre-projection* version, which is to be used in practice. According to this approach the expansion is sought in the form of the expansion $g(x) = \sum_{m=0}^{N} a_m^{(0)} T_m(x)$ and $g''(x)$ is approximated as $\sum_{m=0}^{N-2} a_m^{(-2)} T_m(x)$ with the $a_m^{(-2)}$'s obtained from the $a_m^{(0)}$'s, as usual. These expressions are then substituted into the transformed (homogeneous boundary conditions) problem. Projection with the original orthogonal polynomials T_n's follows and a system of $(N+1)$ equations in $(N+1)$ unknowns is formed. In order to fullfill the boundary conditions

$$\sum_{n=0}^{N} a_n^{(0)} = 0 \quad \text{and} \quad \sum_{n=0}^{N} (-1)^n a_n^{(0)} = 0 \tag{3.27}$$

we solve them for $a_0^{(0)}$ and $a_1^{(0)}$ to derive the equivalent pair

$$a_0^{(0)} = -\sum_{n=1}^{N/2} a_{2n}^{(0)} \quad \text{and} \quad a_1^{(0)} = -\sum_{n=1}^{N/2-1} a_{2n+1}^{(0)} \tag{3.28}$$

Building these constraints into the solution vector via appropriate changes in the coefficient matrix, we introduce an updated system — maintaining its $(N+1)\times(N+1)$ dimensional structure — which ensures satisfaction of the homogeneous boundary conditions through that additional coefficient coupling. During the last step, we see the boundary conditions spreading out in all the equations; this is another manifestation of the global property of spectral basis functions. Finally, the system of equations is solved for $\mathbf{a}^{(0)}$ which, subsequently, is inverted to obtain $g(x)$.

The differentiated system is absolutely essential in time dependent problems. Details, on the mathematical manipulations involved in the orthogonalization of the basis

function set, are presented during the discussion of the Galerkin formulation for the one-dimensional heat equation; a simplified version of those applies here. Indirect *integrated* systems are not employed, since they tend to lead to obscure formulations and they do not feature easily invertible matrices but rather full ones instead.

3.3.3 The collocation method

The formulation of the solution process of the Helmholtz equation, in a pseudospectral context, follows the Galerkin formulation of the problem in a nearly parallel fashion. The sources responsible for the reported resemblance between these two methods may be identified first, in the requirement that the equation must be associated with homogeneous boundary conditions and second and most important, in the requirement that each one of the basis functions used in the approximation, satisfies the homogeneous boundary conditions.

Let us now assume that we need to solve the Helmholtz equation with non-homogeneous Dirichlet boundary conditions (the same as in Galerkin's method). First, transformation of the problem to a new one with homogeneous conditions associated with it, is performed. Subsequently, the solution of the updated problem is pursued and both the direct and indirect variants of the method are investigated in detail.

3.3.3.1 The direct system

According to the *direct* approach, the use of the $\{T_n\}$ set is prohibited; a search for a suitable alternative leads us to the familiar set $\{q_n\}$, where the definition of the functions $q_n(x)$, for $n = 2, \ldots, N$ is identical with the one given in the Galerkin analysis of the problem. The expansion reads $g(x) = \sum_{m=2}^{N} \bar{a}_m^{(0)} q_m(x)$ and substitution in the equation follows. A pseudospectral projection is now expected to decompose

the differential equation into a system of $(N-1)$ algebraic equations for the $(N-1)$ unknowns $\bar{a}_m^{(0)}$'s.

The pseudospectral projection consists of taking the inner product of the equation with a set of δ functions $\{\delta_n\}$ where $\delta_n = \delta(x - x_n)$ for $n = 1, \ldots, N-1$. The points x_n constitute a $(N-1)$-long collocation-point set on which the expansion $g(x) = \sum_{m=2}^{N} \bar{a}_m^{(0)} q_m(x)$ satisfies the differential equation and the boundary conditions *exactly*.

A variety of collocation-point sets have appeared in the literature and a brief description of most promising ones is given below.

Chebychev's choice (Lanczos, 1957), consists of the zeros of the $(N-1)$-th order Chebychev polynomial $T_{N-1}(x)$, given as:

$$x_n = \cos\left(\frac{n+\frac{1}{2}}{N-1}\right)\pi \quad \text{for} \quad n = 0, \ldots, N-2 \tag{3.29}$$

Filippi's choice consists of the extrema of the N-th order Chebychev polynomial $T_N(x)$, which are located at the zeros of the derivative $T_N'(x)$ and may be obtained through the formula:

$$x_n = \cos\frac{\pi n}{N} \quad \text{for} \quad n = 0, \ldots, N \tag{3.30}$$

The latter set contains $(N+1)$ points, since it includes the endpoints -1 and $+1$, as well (corresponding to $n = 0$ and $n = N$ respectively). These points have to be omitted, as the choice of the $\{q_n\}$ set accounts for satisfaction of the imposed boundary conditions (Lanczos, 1957, Gottlieb and Orszag, 1977).

Other collocation sets include the extrema of the $(N-2)$-th order Chebychev polynomial $T_{N-2}(x)$ (*Clenshaw's* choice), the zeros of the the $(N-1)$-th order Legendre

polynomial $P_n(x)$ (*Legendre's* choice) and the extrema of the N-th order "stretched" Chebychev polynomial $T_N^{**}(x)$ (Kizner, 1964), called the *extremal* choice.

The pseudospectral projection reads

$$\sum_{m=2}^{N} \bar{a}_m^{(0)} \left\langle \frac{d^2}{dx^2} q_m(x), \delta(x - x_n) \right\rangle + k^2 \sum_{m=2}^{N} \bar{a}_m^{(0)} \left\langle q_m(x), \delta(x - x_n) \right\rangle$$
$$= -k^2 \left\langle (a + bx), \delta(x - x_n) \right\rangle \quad \text{for} \quad n = 2, \ldots, N-1 \tag{3.31}$$

or

$$\sum_{m=2}^{N} \bar{a}_m^{(0)} \frac{d^2}{dx^2} q_m(x_n) + k^2 \sum_{m=2}^{N} \bar{a}_m^{(0)} q_m(x_n) = -k^2 (a + bx_n) \quad \text{for} \quad n = 2, \ldots, N-1 \tag{3.32}$$

We can see then that the evaluation of the complicated inner products $\langle q_n, q_m'' \rangle$ or $\langle T_n, T_m'' \rangle$, present in the direct Galerkin and tau methods, respectively, has been substituted here, with the (much simpler) evaluation of the $(d^2/dx^2) q_m(x)$ at the $(N-1)$ collocation-point set $\{x_n\}$ (Appendix A.2)

The full matrix $(d^2/dx^2 + k^2) q_m(x_n)$ is then inverted to obtain $\bar{a}^{(0)}$. Direct application, of the inverse fast Chebychev transform (IFCT) routine with the vector $\bar{a}^{(0)}$ as input, is hindered by the fact that the elements of this vector correspond to an expansion in terms of the q_n's and not the T_n's. Employing the same transformation as in Galerkin's method (see 3.3.2.1), the problem is overcome and finally adding the solution of the original equation is obtained as $y(x) = g(x) + (a + bx)$.

3.3.3.2 The indirect (differentiated) system

The principal idea beyond this *indirect* variant of the pseudospectral method is the *orthogonalization* of the basis function set $\{q_n\}$. Since the arguments are, basically, the same to the ones presented in the discussion of the Galerkin system, we will focus on pointing out the differences between the two analyses only.

Starting with the *post-projection* version, we identify the coefficient coupling due to the use of the $\{q_n\}$ set; extension to the $\{T_n\}$ set, which also has two degrees of freedom more, is equivalent to allowing the boundary points -1 and $+1$ to enter the collocation set $\{x_n\}$.

This concept is properly implemented in the *pre-projection* version, where we start off with the familiar expansions $g(x) = \sum_{m=0}^{N} a_m^{(0)} T_m(x)$ and $g''(x) = \sum_{m=0}^{N-2} a_m^{(-2)} T_m(x)$. We then proceed with the pseudospectral projection, where the collocation-point set employed includes the boundary points as well. The next step involves imposing the appropriate constraints on the vector $\mathbf{a}^{(0)}$, so that compliance with the boundary conditions is guaranteed. After the manipulations demanded by the last step have been completed, we solve the system to obtain $\mathbf{a}^{(0)}$. This is inverted to give us $g(x)$; the addition of $a + bx$ to $g(x)$ completes the procedure, providing the solution $y(x)$ to the initial problem. Indirect *integrated* systems have not been in favour, due to the need for pre-processing (which might be burdensome) or because of the relatively more complicated form of the resulting problem.

3.3.4 Finite differences method

Finite differences may also be visualized as collocation with the requirement that the equation and its boundary conditions are to be satisfied exactly on a collocation point set consisting of the points $0, \epsilon, 2\epsilon, \ldots, n\epsilon$ as ϵ tends to 0. This set lies in the neighborhood of the origin and its choice corresponds to the Taylor expansion of $y(x)$ truncated to $N+1$ terms (Lanczos, 1957).

Let us start by defininining an $(N+1)$-long point set x_j (for $j = -N/2, \ldots, +N/2$) on the interval $[-1, +1]$. The corresponding discretization interval is $\Delta x = 2/N$ and the computational grid is given as

$$x_j = j\Delta x \quad \text{for} \quad j = -N/2, \ldots, +N/2 \tag{3.33}$$

The second derivative is approximated through the classic (second-order in accuracy) difference scheme

$$\frac{d^2}{dx^2} y(x) \simeq \frac{y(x+\Delta x) - 2y(x) + y(x-\Delta x)}{(\Delta x)^2} \tag{3.34}$$

and assuming arbitrary non-homogeneous Dirichlet boundary conditions ($y(-1) = \beta$ and $y(+1) = \alpha$), the Helmholtz equation (3.1) is transformed into the *tridiagonal* system

$$\begin{pmatrix} d & 1 & & & \\ 1 & d & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & d & 1 \\ & & & 1 & d \end{pmatrix} \begin{pmatrix} y_{-(N/2-1)} \\ \\ \vdots \\ \\ y_{+(N/2-1)} \end{pmatrix} = \begin{pmatrix} -\beta \\ 0 \\ \vdots \\ 0 \\ -\alpha \end{pmatrix} \tag{3.35}$$

where $d = (k\Delta x)^2 - 2$.

If the differential equation satisfies general non-homogeneous Neumann boundary conditions ($y'(-1) = \delta$ and $y'(+1) = \gamma$), a difference scheme for approximating the normal derivatives is needed. One way is to use *central differences*

$$\frac{d}{dx}y(x) \simeq \frac{y(x+\Delta x) - y(x-\Delta x)}{2\Delta x} \tag{3.36}$$

at both boundaries, whereas another implementation involves the use of the *one-sided* difference schemes

$$\frac{d}{dx}y(x) \simeq \frac{y(x+\Delta x) - y(x)}{\Delta x} \qquad \text{at} \quad x = -1 \tag{3.37}$$

and

$$\frac{d}{dx}y(x) \simeq \frac{y(x) - y(x-\Delta x)}{\Delta x} \qquad \text{at} \quad x = +1 \tag{3.38}$$

These last two schemes are not centered and they are, therefore, characterized by the *two-point oscillation* of $y(x)$ phenomenon (see 2.5.6 and 3.4.2).

Central differences yield the tridiagonal system

$$\begin{pmatrix} d & 2 & & & \\ 1 & d & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & d & 1 \\ & & & 2 & d \end{pmatrix} \begin{pmatrix} y_{-(N/2)} \\ \\ \vdots \\ \\ y_{+(N/2)} \end{pmatrix} = \begin{pmatrix} 2\delta\Delta x \\ 0 \\ \vdots \\ 0 \\ -2\gamma\Delta x \end{pmatrix} \tag{3.39}$$

Another tridiagonal system is obtained from the second technique; its explicit form is

$$\begin{pmatrix} d+1 & 1 & & & \\ 1 & d & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & d & 1 \\ & & & 1 & d+1 \end{pmatrix} \begin{pmatrix} y_{-(N/2-1)} \\ \\ \vdots \\ \\ y_{+(N/2-1)} \end{pmatrix} = \begin{pmatrix} \delta\Delta x \\ 0 \\ \vdots \\ 0 \\ -\gamma\Delta x \end{pmatrix} \tag{3.40}$$

Robbins boundary conditions may be constructed by combining the previous analyses; the obtained systems are still tridiagonal.

The inversion of these algebraic systems provides $y(x)$ at the points of the computational grid. These inversions can be performed very efficiently due to the tridiagonal structure of the matrices (see Appendix B.3). This feature is extremely important for time-dependent problems.

3.4 Discussion of Results

The performance of the numerical schemes discussed before, has been studied extensively by comparing the numerical results with the exact solution

$$y(x) = A\cos kx + B\sin kx \tag{3.41}$$

where A and B are integration constants and depend on the boundary conditions. The level of accuracy is evaluated in both the L_2 and the L_∞ norms by defining appropriate relative errors.

Let us say that $\bar{y}_i$ is the approximate solution vector calculated on the discrete point set x_i, $i = 1, \ldots, N$ and y_i contains the values of the exact solution at these points. We then define a relative L_2 error as

$$\bar{L}_2 = \frac{\sum_{i=1}^{N} (y_i - \bar{y}_i)^2}{\sum_{i=1}^{N} \bar{y}_i^2}$$

and a relative L_∞ error as

$$\bar{L}_\infty = \frac{\max_{1 \leq i \leq N} |y_i - \bar{y}_i|}{\max_{1 \leq i \leq N} |\bar{y}_i|}$$

and we may imagine contour maps of both these errors by tabulating their values for varying values of the parameters N and k^2. This procedure is repeated for various schemes and different kinds of boundary conditions.

3.4.1 Inhomogeneous Dirichlet boundary conditions

The specific choice of the boundary constraints $\alpha = y(x = +1)$ and $\beta = y(x = -1)$, does not affect the convergence properties of either the finite difference or the Chebychev schemes. Fourier series would exhibit a reduced convergence rate depending on the size of the discontinuity jump; therefore, the awkward pair of values $(\alpha, \beta) = (-2, +5)$ has been chosen in order to demonstrate the capability of Chebychev methods to handle efficiently such highly non-periodic constraints.

3.4.1.1 Finite differences

$\bar{L}_2$ and $\bar{L}_\infty$ values for the finite difference solution of the Dirichlet Helmholtz problem are displayed in tables (3.1-2) respectively, but before proceeding with a detailed discussion and interpretation of the results, we should pause to emphasize some fundamental aspects of both the continuous and the discrete problem.

$N \setminus k^2$	2	7	25	100	250	1000
5	4.75 (-2)	7.05 (-1)	5.78 (-1)	7.68 (-1)	9.84 (-1)	3.10 (-1)
9	1.96 (-3)	9.53 (-3)	2.42 (-1)	7.46 (-1)	1.00 (0)	9.72 (-1)
17	1.13 (-4)	4.73 (-4)	3.44 (-2)	1.14 (0)	1.00 (0)	1.00 (0)
33	8.00 (-6)	2.90 (-5)	2.98 (-3)	4.64 (-2)	7.66 (-1)	1.17 (0)
65	3.00 (-6)	2.00 (-6)	2.08 (-4)	3.92 (-3)	3.79 (-1)	1.86 (0)
129	1.40 (-5)	1.00 (-6)	1.60 (-5)	2.71 (-4)	7.94 (-2)	3.73 (-1)
257	2.22 (-4)	7.00 (-6)	7.00 (-5)	2.10 (-5)	8.29 (-3)	7.80 (-2)
513	3.75 (-3)	1.03 (-4)	6.10 (-5)	8.00 (-6)	1.18 (-3)	8.16 (-3)

Table 3.1 $\bar{L}_2$ values for the finite difference solution of the Dirichlet Helmholtz problem.

$N \setminus k^2$	2	7	25	100	250	1000
5	2.48 (-1)	9.40 (-1)	8.10 (-1)	1.01 (0)	1.00 (0)	6.53 (-1)
9	4.67 (-2)	1.13 (-1)	6.82 (-1)	9.81 (-1)	1.00 (0)	1.00 (0)
17	1.11 (-2)	2.51 (-2)	2.13 (-1)	1.58 (0)	1.17 (0)	1.00 (0)
33	2.99 (-3)	6.13 (-3)	6.31 (-2)	2.70 (-1)	8.94 (-1)	1.47 (0)
65	1.69 (-3)	1.63 (-3)	1.64 (-2)	7.72 (-2)	6.22 (-1)	1.87 (0)
129	3.96 (-3)	1.04 (-3)	4.50 (-3)	1.99 (-2)	2.82 (-1)	6.58 (-1)
257	1.56 (-2)	3.28 (-3)	2.79 (-3)	5.36 (-3)	9.15 (-2)	2.89 (-1)
513	6.34 (-2)	1.23 (-2)	7.91 (-3)	3.18 (-3)	3.43 (-2)	9.30 (-2)

Table 3.2 $\bar{L}_\infty$ values for the finite difference solution of the Dirichlet Helmholtz problem.

Dirichlet or Neuman boundary conditions correspond to standing waves and an increased k^2 value gives rise to more rapid oscillations. The eigenvalues of the continuous problem with eigenfunctions $y_n(x) = c_n \sin k(x+1)$ are $k_n^2 = -(n\pi/2)^2$, where c_n is an arbitrary constant. The second order finite difference representation of d^2/dx^2 reads $(1, -2, 1)/(\Delta x)^2$ and it is negative definite (i.e all eigenvalues are negative); its condition number's growth rate may be estimated from the ratio of its maximum over its minimum eigenvalue, that is to say $\lambda_{max}/\lambda_{min}$. The minimum eigenvalue may be estimated adequately from the continuous spectrum for $n = 1$, as $\lambda_{min} \sim (\pi/N)^2$ ($\Delta x = 1$ for normalization purposes). The eigenvalue spectrum has an upper bound (in absolute value) of approximately 4 as seen by application of Gerschgorin's theorem — the Gerschgorin's disks are $(-2, +2)$ (Δx is normalized to unity again). Combining this information for the minimum and the maximum eigenvalues, we can estimate a condition number which grows roughly as N^2. Furthermore, we should point out that the particular finite difference scheme is second order in accuracy and therefore its truncation error decays as N^2.

The Dirichlet Helmholtz equation gives rise to a symmetric eigenvalue problem when formulated in a finite difference environment. The matrix form of the problem is

$$\mathbf{A}\mathbf{y} + k^2\mathbf{I}\mathbf{y} = 0 \tag{3.44}$$

where $\mathbf{A}$ is the tridiagonal matrix given in (3.35). Applying a similarity transformation to $\mathbf{A}$, i.e $\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \mathbf{\Lambda}$, where $\mathbf{\Lambda}$ is a diagonal matrix with its elements being the eigenvalues of the matrix $\mathbf{A}$ and $\mathbf{S}$ is chosen such that its columns contain the

eigenvectors of $\mathbf{A}$, we obtain that

$$\mathbf{S^{-1}ASy} + k^2\mathbf{S^{-1}ISy} = 0 \tag{3.45}$$

or

$$(\mathbf{\Lambda} + k^2\mathbf{I})\mathbf{y} = \mathbf{0} \tag{3.46}$$

We now see that whenever $(k\Delta x)^2 < |\lambda_{min}|$, the Helmholtz matrix retains the negative definiteness of the second order difference operator $(+1, -2, +1)$. As $(k\Delta x)^2$ reaches $|\lambda_{min}|$ a zero-eigenvalue problem is encountered (the Helmholtz matrix is now singular), as the first of the resonant frequencies (eigen-frequencies) of the system is being reached. The moment $(k\Delta x)^2$ exceeds $|\lambda_{min}|$ the Helmholtz matrix becomes indefinite, for one of its eigenvalue turns positive. As $(k\Delta x)^2$ continues to increase, the indefiniteness persists (as more eigenvalues change sign) and singularities are identified, reflecting the presence of further resonances. This situation is interrupted as $|\lambda_{max}|$ is reached and subsequently, exceeded. No more singularities are present and the matrix becomes positive definite as all of its eigenvalues turn positive.

It is interesting to note that passing from indefiniteness to a positive definite structure occurs nearly simultaneously with the establishment of the diagonally dominant Helmholtz matrix, i.e for $(k\Delta x)^2 > 4$. The off-diagonal dominant case is obtained for the range $0 < (k\Delta x)^2 < 4$ which results in suppressing the diagonal of the matrix. In particular for the value of $(k\Delta x)^2 = 2$ a total collapse of the diagonal is experienced; that yields a singular matrix as it corresponds to the eigenvalue $\lambda_i = -2$ of the tridiagonal matrix $(+1, -2, +1)$.

Let us now proceed to discuss the results shown in the tables (3.1-2) on the basis of the previous analysis; the corresponding values of $k\Delta x$ are given in the table 3.3.

$N \setminus k^2$	2	7	25	100	250	1000
5	7.07 (-1)	1.32 (0)	2.50 (0)	5.00 (0)	7.91 (0)	1.58 (+1)
9	3.54 (-1)	6.61 (-1)	1.25 (0)	2.50 (0)	3.95 (0)	7.91 (0)
17	1.77 (-1)	3.31 (-1)	6.25 (-1)	1.25 (0)	1.98 (0)	3.95 (0)
33	8.84 (-2)	1.65 (-1)	3.13 (-1)	6.25 (-1)	9.88 (-1)	1.98 (0)
65	4.42 (-2)	8.27 (-2)	1.56 (-1)	3.13 (-1)	4.94 (-1)	9.88 (-1)
129	2.21 (-2)	4.13 (-2)	7.81 (-2)	1.56 (-1)	2.47 (-1)	4.94 (-1)
257	1.10 (-2)	2.07 (-2)	3.91 (-2)	7.81 (-2)	1.24 (-1)	2.47 (-1)
513	5.52 (-3)	1.03 (-2)	1.95 (-2)	3.91 (-2)	6.18 (-2)	1.24 (-1)
1025	2.76 (-3)	5.17 (-3)	9.77 (-3)	1.95 (-2)	3.09 (-2)	6.18 (-2)
2049	1.38 (-3)	2.58 (-3)	4.88 (-3)	9.77 (-3)	1.54 (-2)	3.09 (-2)

Table 3.3 Values of the quantity $k\Delta x$ for various values of the parameters k^2 and N.

Based on the behavior of both the $\bar{L}_2$ and the $\bar{L}_\infty$ computed estimates, we may identify three characteristics regions on the $(k\Delta x)$ contour map (Figure 3.1).

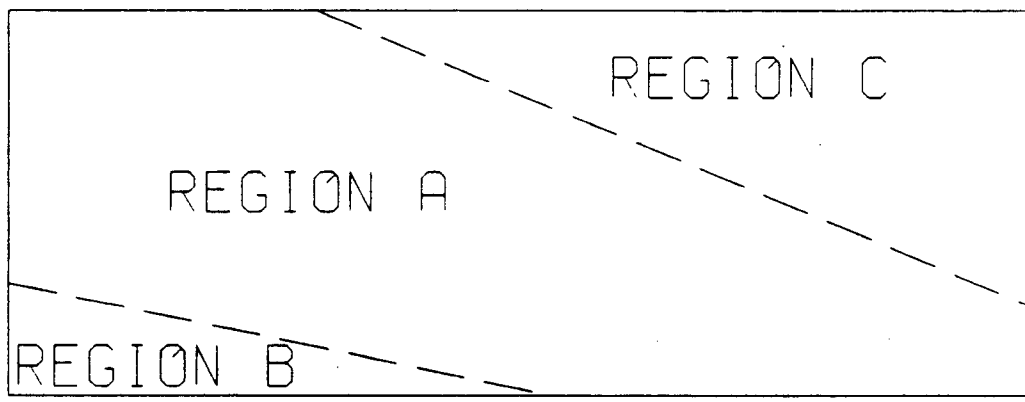

Figure 3.1 Characteristic regions on the $k\Delta x$ contours.

For a constant value of k^2 we would normally expect an improvement in accuracy as N increases, since the truncation error is reduced. Similarly, it is anticipated that, for constant N, the accuracy of the approximation diminishes as k^2 increases, due to an enhanced truncation error involved in the inadequate handling of the higher frequencies introduced in the system. Indeed, this kind of error behavior is observed in Region A.

Region B manifests the dominance of the round-off error (a direct consequence of a poor conditioning of the numerical system) over the truncation error. The latter is still being reduced (although as N^2 only) but the improvement is drowned in the excessive accumulation of round-off, which takes over and destroys the approximation. Furthermore, the values of $(k\Delta x)^2$ in Region B lie either very close or even within the machine accuracy and any attempt for higher resolution is futile.

Region C is quite interesting in the sense that it presents us with a surprising situation, at least at first glance. The matrix of the system is diagonally dominant and positive definite, while the crossing to Region A is associated with a loss of both these advantageous properties (the only exception is to be found for $k^2 = 2$ which yields a negative definite system always). Although, the matrix is better structured from a numerical viewpoint in Region C, the errors are much larger and they do not behave in any kind of consistent fashion at all.

The source of that anomalous error behavior may be traced in the high values of $k\Delta x$ in Region C; this $k\Delta x$ range results in a total deformation of the negative nature of the continuous spectrum, since the discrete scheme has only positive eigenvalues. Equivalently, the consequences of the high values of $k\Delta x$, may be understood as severe truncation error augmented with a profound aliasing, since the low sampling density

(denoted by the low values of the parameter $(k\Delta x)^{-1}$ which gives the number of points per wavelength) results in the difference Helmholtz operator failing ultimately to approximate the original differential operator. In other words, the underlying Taylor expansion is diverging and, thereafter, the numerical results are — to a great extent — meaningless both in an absolute and a relative quantitative sense. This is due to the inevitable inability of the numerical scheme to handle the solution properly. An additional deterioration is experienced, as the grossly inadequate sampling allows tremendous levels of aliasing to contaminate the solution. The approximation becomes so untrustworthy that the error estimates do not mean much, as they depend on a set of values given at a specific point set, whose low density is absolutely inadequate. Evidently, the error oscillates as the solution oscillates and it experiences a diminishing capability of resolving differences satisfactorily, when significant departures from the true solution are present; this problem is more profound in the $\bar{L}_2$ norm due to the squaring it involves.

Finally, the behavior of the errors for the same values of $(k\Delta x)^{-1}$, but for different values of k, reveals an interesting pattern. The error is progressively increasing with k, confirming the dependence of the numerical accuracy on the total number of wavelengths present in the computational grid; the latter is expressed by the quantity $k\ell$, i.e $2k$ in our particular 1-D simulation, since the length of the computational region is $\ell = 2$ and remains constant throughout the various computations.

3.4.1.2 Tau Chebychev methods

Chebychev spectral formulations of the Helmholtz problem give rise to an unsymmetric eigenvalue problem, involving a spectrum which quite often incorporates pairs

of complex conjugate eigenvalues. We commence our discussion of Chebychev methods with the tau method. Its direct implementation yields the results given in tables (3.4-50; the algorithm becomes inefficient for $N > 65$ and this has prohibited further computations. The combination $N = 129, k^2 = 250$ has been investigated in order to provide an indication of the system's performance for larger sizes of the problem and to help us enrich our comprehension of the discrepancies in the performance of the direct tau and the direct-indirect pseudospectral approaches on a Chebychev collocation point-set (see 3.4.1.3).

$N \setminus k^2$	2	7	25	100	250	1000
5	1.16 (-4)	9.67 (-2)	2.78 (+2)	2.79 (0)	1.18 (0)	2.29 (0)
9	$\star$	$\star$	1.31 (-3)	4.72 (-1)	1.21 (0)	1.71 (0)
17	$\star$	$\star$	$\star$	7.00 (-6)	1.29 (0)	7.20 (-1)
33	$\star$	$\star$	$\star$	$\star$	$\star$	1.92 (0)
65	$\star$	$\star$	$\star$	$\star$	$\star$	$\star$

Table 3.4 $\bar{L}_2$ values for the direct Chebychev tau solutions of the Dirichlet Helmholtz problem. Stars correspond to relative errors $\leq 10^{-7}$. Cost condiderations have prohibited computations for values of $N > 65$.

$N \setminus k^2$	2	7	25	100	250	1000
5	1.24 (-2)	3.50 (-1)	2.05 (+1)	1.69 (0)	1.15 (0)	5.03 (-1)
9	2.00 (-6)	2.28 (-4)	6.11 (-2)	9.50 (-1)	1.09 (0)	1.59 (0)
17	2.00 (-6)	3.00 (-6)	3.00 (-6)	3.59 (-3)	1.13 (0)	1.34 (0)
33	3.00 (-6)	3.00 (-6)	3.00 (-6)	8.00 (-6)	1.80 (-5)	1.40 (0)
65	3.00 (-6)	3.00 (-6)	3.00 (-6)	9.00 (-6)	1.80 (-5)	4.60 (-5)
129	—	—	—	—	2.90 (-5)	—

Table 3.5 $\bar{L}_\infty$ values for the direct Chebychev tau solutions of the Dirichlet Helmholtz problem. Cost considerations have not permitted computations for $N > 65$; however, the $N = 129$ system has been solved for $k^2 = 250$ to provide an idea of the performance of the method at this level.

Both the (indirect) differentiated and the integrated versions error estimates that are almost identical within round-off (inversion of all the linear systems of simultaneous equations arising in this chapter has been performed via the LU decomposition, augmented with iterative improvement of the solution vector whenever possible). Although the integrated d^2/dx^2 matrix enjoys a superior conditioning over the differentiated system's matrix, their Helmholtz counterparts do not exhibit dramatic conditioning differences. Tables (3.6-7) contain error estimates that reflect the performance of both methods.

The indirect methods' error estimates are virtually identical to the direct method's. Small differences do exist and they would tend to indicate that the direct system is slightly less accurate; these discrepancies are basically negligible and based on the presently computed error values and conditioning estimates, we can infer that present error behavior should evolve in a similar fashion for both the direct and the indirect implementations.

$N \setminus k^2$	2	7	25	100	250	1000
5	1.16 (-4)	9.67 (-2)	2.78 (+2)	2.79 (0)	1.18 (0)	2.30 (0)
9	$\star$	$\star$	1.31 (-3)	4.72 (-1)	1.21 (0)	1.71 (0)
17	$\star$	$\star$	$\star$	7.00 (-6)	1.29 (0)	7.20 (-1)
33	$\star$	$\star$	$\star$	$\star$	$\star$	1.92 (0)
65	$\star$	$\star$	$\star$	$\star$	$\star$	$\star$

Table 3.6 $\bar{L}_2$ values for the indirect Chebychev tau solutions of the Dirichlet Helmholtz problem. Stars correspond to relative errors $\leq 10^{-7}$ and they persist until N=513, at least.

$N \setminus k^2$	2	7	25	100	250	1000
5	1.24 (-2)	3.50 (-1)	2.05 (+1)	1.69 (0)	1.17 (0)	5.03 (-1)
9	1.00 (-6)	2.29 (-4)	6.11 (-2)	9.47 (-1)	1.09 (0)	1.59 (0)
17	1.00 (-6)	2.00 (-6)	3.00 (-6)	3.59 (-3)	1.13 (0)	1.34 (0)
33	2.00 (-6)	2.00 (-6)	4.00 (-6)	6.00 (-6)	1.20 (-5)	1.40 (0)
65	2.00 (-6)	3.00 (-6)	4.00 (-6)	6.00 (-6)	1.20 (-5)	3.00 (-5)
129	9.00 (-5)	1.10 (-5)	1.10 (-5)	1.90 (-5)	2.10 (-5)	5.60 (-5)
257	1.20 (-5)	1.30 (-5)	1.50 (-5)	7.00 (-6)	2.10 (-5)	3.40 (-5)
513	2.00 (-5)	2.60 (-5)	2.40 (-5)	1.40 (-5)	2.70 (-5)	3.90 (-5)

Table 3.7 $\bar{L}_\infty$ values for the indirect Chebychev tau solutions of the Dirichlet Helmholtz problem.

Continuing with the analysis of the relative performance of the indirect methods, we should point out that our reservations regarding the structure of the differentiated system are not unjustified, although in general its performance has been comparable integrated system's. The break-down of the former may be witnessed in the vicinity of a zero-eigenvalue area. As an example, for $N = 5$ and $k^2 = 24$ both systems experience an eigenvalue collapse, i.e $|\lambda_{min}| \sim 10^{-6} - 10^{-7}$) and thereby they appear to be nearly singular with the integrated system even more so, as it possesses the smallest eigenvalue ($\sim 10^{-7}$) and a determinant of $O(-5)$, that is, 3 orders of magnitude smaller than the determinant of the differentiated system. Despite that, the differentiated system shows an $\bar{L}_2$ of $O(14)$ and an $\bar{L}_\infty$ of $O(7)$, whereas the errors of the integrated system do not exceed $O(1)$. The explanation of such spectacular error discrepancies may be traced in the ill-conditioning of the differentiated system; its condition number is of $O(8)$, as contrasted to $O(5)$ of the integrated system, an excellent manifestation of the fact that the value of the determinant of a matrix is a very poor measure of its conditioning.

Let us now proceed to discuss the tau Chebychev performance and draw comparisons with the performance of the finite difference scheme. The same three regions of the finite difference tables may be identified again. Region C exhibits the same inconsistent and oscillatory error behavior; the explanation is basically the same in the finite differences case. Although truncation errors in the approximation of the differential operator are not present, the nature of the error originates in aliasing; the exponential convergence of the method eliminates truncation errors but it can do absolutely nothing to alleviate the contamination of the solution due to the severe levels of aliasing.

Exact comparisons under such circumstances are meaningless, especially provided that the computational grid is now quite different. The peaking of the error at $N = 5$ and $k^2 = 25$ might be either an arbitrary high attributed to the foregoing factors or it could be traced into a relative ill-conditioning of both the tau systems since it lies at the close vicinity of a singularity. Although either one or both the above possibilities might be partially responsible for that error high, a closer look at the true solution and a deeper understanding of the tau-projection, show that such a case suffers from overweighted boundary conditions at the expense of two lost highest modes. Furthermore, the collocation of points is rather unfortunate, since $k\Delta x_{max}$ is large exactly where a denser sampling would have been required.

Region B of the finite difference table can not be identified in the $\bar{L}_2$ map, as the nature of the Euclidian norm inhibits its resolution ability; recovery of this property would take place when the round-off would be dominating the solution completely and it would thus be meaningless. On the contrary the $\bar{L}_\infty$ map reveals the presence of such a zone which does not possess the expansion rate of the finite differences' one

though. These differences between the finite difference and the tau Chebychev scheme have their origins in the exponential convergence and the conditioning of the latter; the integrated system should be anticipated to exhibit the deterioration associated with Region B later than the differentiated system would. A comparison of the results (in both norms) in Region A shows that the tau Chebychev method enjoys an definite overall superiority over the finite differences.

For the same number of points per wavelength $(k\Delta x)^{-1}$ the tau Chebychev method is readily seen to have a smaller error; alternatively, the finite differences require more points to achieve the tau method's level of accuracy; furthermore the tau's superb maximum accuracy (basically $O(-6)$) is never reached by the finite differences as the round-off surpasses the low order algebraic convergence rate, overwhelming the accuracy improvements; this is best exhibited in the $\bar{L}_\infty$ tables.

Furthermore, the accuracy of the tau approximation is much less vulnerable to an increase of k (i.e increases in the total number of wavelengths $k\ell$ present), when compared to finite differences for the same value of $(k\Delta x)^{-1}$.

For small values of k ($k^2 \leq 25$), finite differences may achieve an accuracy comparable to the tau method's without requiring more than half an order of magnitude more points. As k increases, the demand for more points is more pronounced and we see that for $k^2 = 100$, about 1.5 orders of magnitude are needed, while for a k^2 of 250 and 1000, several orders of magnitude would be required to do so. Unfortunately, the computer code uses a Fast Chebychev Transform (which is based on a Fast Fourier Transform routine that only accepts input arrays dimensioned an integer power of 2) that does not permit experiments on grids of intermediate size; this prohibition resulted to the rather spiky structure of the tau map since the tau Chebychev operator

seems to exhibit an incredibly fast convergence rate as soon as the aliasing present in Region B is overcome.

Briefly, the overall comparison of the results shows that when aliasing and round-off factors are not dominating the numerical schemes, the tau-Chebychev method is both capable of exhibiting a superb level of accuracy provided that it is accompanied with an adequate number of points. Even more important, it is capable of providing approximations of moderate accuracy with many fewer points than the finite difference method.

3.4.1.3 Galerkin and pseudospectral Chebychev methods

Galerkin and pseudospectral approaches involve additional complications in their formulation; their direct versions were originally devised to alleviate these problems.

Tables (3.8-9) contain the errors associated with the Galerkin method, tables (3.10-11) with the pseudospectral using a Filippi collocation point set and tables (3.12-13) with the pseudospectral on a Chebychev collocation point set.

$N \setminus k^2$	2	7	25	100	250	1000
5	2.00 (-6)	1.81 (-2)	2.56 (0)	5.42 (-1)	9.95 (-1)	7.10 (-1)
9	$\star$	$\star$	5.40 (-5)	7.40 (-1)	7.35 (-1)	9.08 (-1)
17	$\star$	$\star$	$\star$	$\star$	1.67 (0)	9.19 (-1)
33	$\star$	$\star$	$\star$	$\star$	$\star$	3.69 (0)
65	$\star$	$\star$	$\star$	$\star$	$\star$	$\star$

Table 3.8 $\bar{L}_2$ values for the direct Chebychev Galerkin solution of the Dirichlet Helmholtz problem (stars correspond to relative errors $\leq 10^{-7}$). Computations for $N > 65$ have not been carried out because of cost considerations.

$N \setminus k^2$	2	7	25	100	250	1000
5	2.14 (-3)	1.50 (-1)	2.07 (0)	8.13 (-1)	1.00 (0)	9.52 (-1)
9	1.00 (-6)	4.10 (-5)	1.25 (-2)	1.05 (0)	9.92 (-1)	9.19 (-1)
17	1.00 (-6)	2.00 (-6)	6.00 (-6)	6.32 (-4)	1.27 (0)	9.80 (-1)
33	1.00 (-6)	2.00 (-6)	4.00 (-6)	9.00 (-6)	4.00 (-5)	1.89 (0)
65	2.00 (-6)	2.00 (-6)	3.00 (-6)	1.00 (-5)	1.84 (-4)	6.20 (-5)

Table 3.9 $\bar{L}_\infty$ values for the direct Chebychev Galerkin solution of the Dirichlet Helmholtz problem (see caption of Table 3.8).

We notice that Region C remains susceptible to aliasing and the irrelevant error oscillations are still being observed. Although the approximations remain still really bad, we do see the Galerkin and the Filippi-collocation methods exhibiting relatively smaller errors, while the tau and the Chebychev-collocation techniques are worse.

$N \setminus k^2$	2	7	25	100	250	1000
5	1.19 (-4)	1.82 (-2)	5.12 (-1)	5.37 (-1)	9.95 (-1)	7.32 (-1)
9	$\star$	$\star$	1.87 (-4)	7.45 (-1)	7.44 (-1)	9.04 (-1)
17	$\star$	$\star$	$\star$	$\star$	1.67 (0)	9.17 (-1)
33	$\star$	$\star$	$\star$	$\star$	$\star$	3.60 (0)
65	$\star$	$\star$	$\star$	$\star$	1.40 (-5)	2.78 (-3)
129	1.88 (-2)	6.33 (-3)	1.02 (-1)	7.44 (-1)	9.46 (-1)	1.42 (0)

Table 3.10 $\bar{L}_2$ values for the direct Chebychev pseudospectral solution of the Dirichlet Helmholtz problem on a Filippi collocation point-set (stars correspond to relative errors $\leq 10^{-7}$). Computations for $N > 129$ have been considered fruitless due to the greatly increased cost and the extremely high level of ill-conditioning present.

Proceeding to Region A, we notice that Galerkin is slightly superior to Filippi-collocation followed by tau and Chebychev-collocation in that order. This reaffirms that Galerkin (for simple linear constant-coefficient problems) is virtually equivalent to the collocation approach on the Filippi set, whereas the tau is identical to the

$N \setminus k^2$	2	7	25	100	250	1000
5	1.36 (-2)	1.54 (-1)	9.72 (-1)	9.11 (-1)	1.00 (0)	9.94 (-1)
9	$\star$	4.10 (-5)	2.13 (-2)	9.16 (-1)	9.83 (-1)	9.48 (-1)
17	2.00 (-6)	2.00 (-6)	8.00 (-6)	5.32 (-4)	1.28 (0)	1.00 (0)
33	1.00 (-6)	2.00 (-6)	1.10 (-5)	2.80 (-5)	4.77 (-4)	1.91 (0)
65	3.00 (-6)	1.80 (-5)	7.40(-5)	1.84 (-4)	5.46 (-3)	1.13 (-1)
129	1.45 (-1)	9.26 (-2)	3.98 (-1)	1.41 (0)	1.03 (0)	2.05 (0)

Table 3.11 $\bar{L}_\infty$ values for the direct Chebychev pseudospectral solution of the Dirichlet Helmholtz problem on a Filippi collocation point-set (see caption of Table 3.10).

collocation technique on the Chebychev set. The tau method solves exactly a perturbed problem on the zeros of a Chebychev polynomial of the appropriate order; in fact, Lanczos originally devised the method developing the general collocation projection principle into a technique which is both fast and very powerful for such kind of problems.

$N \setminus k^2$	2	7	25	100	250	1000
5	1.96 (-3)	9.72 (-2)	2.79 (+2)	6.24 (-2)	1.17 (0)	2.55 (-1)
9	$\star$	$\star$	7.03 (-3)	4.35 (-1)	1.21 (0)	1.38 (0)
17	$\star$	$\star$	$\star$	7.00 (-6)	1.29 (0)	6.54 (-1)
33	$\star$	$\star$	$\star$	$\star$	$\star$	1.91 (0)
65	$\star$	$\star$	$\star$	$\star$	$\star$	1.06 (-2)
129	6.71 (-1)	2.18 (-2)	2.08 (-2)	3.86 (-1)	9.99 (-1)	1.05 (-2)

Table 3.12 $\bar{L}_2$ values for the direct Chebychev pseudospectral solution for the Dirichlet Helmholtz problem on a Chebychev collocation point set (stars correspond to relative errors $\leq 10^{-7}$). Computations for $N > 129$ have been considered fruitless due to both the greatly increased cost and the very high level of ill-conditioning present.

$N \setminus k^2$	2	7	25	100	250	1000
5	5.26 (-2)	3.40 (-1)	2.08 (+1)	3.38 (-1)	1.10 (0)	5.51 (-1)
9	1.00 (-6)	2.23 (-4)	1.21 (-1)	7.57 (-1)	1.09 (0)	1.58 (0)
17	2.00 (-6)	2.00 (-6)	4.00 (-6)	3.84 (-3)	1.10 (0)	1.19 (0)
33	2.00 (-5)	2.00 (-5)	7.00 (-6)	3.60 (-5)	1.56 (-4)	1.40 (0)
65	3.00 (-6)	3.00 (-6)	3.3 (-5)	4.00 (-4)	1.64 (-3)	1.20 (-1)
129	8.71 (-1)	1.71 (-1)	2.98 (-1)	1.01 (0)	1.05 (0)	1.38 (0)

Table 3.13 L_∞ values for the direct Chebychev pseudospectral solution of the Dirichlet Helmholtz problem on a Chebychev collocation point-set (see caption of Table 3.12).

Region B is also present but as in tau's case, Galerkin shows a slow deterioration pace. However, this is not true for the pseudospectral methods and we see that accuracy starts rapidly being lost as the dimension of the problem increases; that usually occurs after maximum accuracy has been achieved and as anticipated, it is more pronounced for higher k^2 values. Such a behavior is very disturbing and it becomes shocking for high values of both parameters N and k^2; there a very fast and overwhelming loss of accuracy is experienced. It is worth pointing out that the Chebychev collocation point-set (which does not yield the outstanding resolution of the Filippi set otherwise) exhibits a less dramatic deterioration than the latter, as the size of the problem increases. The computational burden of the direct Galerkin approach hindered calculations for higher values of N but the presently computed error values (augmented with rough conditioning tests) may be considered as an indication that a similar breakdown is not likely to occur in that short range and with that rapid rate.

Excessive accumulation of round-off errors associated with the complicated computations involved in the construction of the pseudospectral matrices is not believed to carry the responsibility for the reported deterioration because of the following contradictory arguments. Firstly, these computations are performed in double precision

arithmetic (exactly, in order to avoid misfortunes of that type) and secondly the direct Galerkin technique does not exhibit these, athough for the same problem size, it involves more demanding and more vulnerable to round-off computations.

The cause of the problem is finally identified in the enormous level of ill-conditioning of the direct-collocation matrices; this prohibits even iterative improvement of the solution which is often seen diverging. The ill-conditioning of these systems (and probably the Galerkin's in a later stage) is then suspected to be associated to a destabilization of the procedure due to an improper numerical handling of the boundary conditions. The problem originates in the non-orthogonal nature of the $\{q_n\}$ basis function set as the resulting system is expected to lose some of the excellent conditioning properties associated with orthogonal sets; adequate conditioning may be maintained as long as these components retain a high-level of linear independence. Apparently, the direct algorithm does not succeed in fulfilling such expectations and then as the dimensionality and thereby, the "noise" level is increased, these weakly constructed high-resolution systems collapse; the Filippi point-set is obviously affected more than the having less resolving power Chebychev point-set. The rigid foundations of the Galerkin projection make the method more robust as far as ill-conditioning is concerned.

Confirmation of these suspicions demands a comparison with the corresponding results obtained from the indirect method; although the concepts behind the construction of the direct and the indirect systems have already been presented, we should point out that their main difference hinges on the boundary condition treatment.

$N \setminus k^2$	2	7	25	100	250	1000
5	2.00 (-6)	1.81 (-2)	2.56 (0)	5.42 (-1)	9.95 (-1)	7.10 (-1)
9	$\star$	$\star$	5.40 (-5)	7.40 (-1)	7.35 (-1)	9.08 (-1)
17	$\star$	$\star$	$\star$	$\star$	1.67 (0)	9.19 (-1)
33	$\star$	$\star$	$\star$	$\star$	$\star$	3.69 (0)
65	$\star$	$\star$	$\star$	$\star$	$\star$	$\star$

Table 3.14 $\bar{L}_2$ values for the indirect Chebychev Galerkin solution of the Dirichlet Helmholtz problem. Stars correspond to relative errors $\leq 10^{-7}$ and they persist until N=513, at least.

$N \setminus k^2$	2	7	25	100	250	1000
5	2.14 (-3)	1.50 (-1)	2.07 (0)	8.13 (-1)	1.00 (0)	9.52 (-1)
9	1.00 (-6)	4.00 (-5)	1.25 (-2)	1.05 (0)	9.92 (-1)	9.19 (-1)
17	1.00 (-6)	3.00 (-6)	8.00 (-6)	6.24 (-4)	1.27 (0)	9.80 (-1)
33	2.00 (-6)	5.00 (-6)	6.00 (-5)	2.80 (-5)	1.67 (-4)	1.89 (0)
65	9.00 (-6)	5.00 (-6)	9.00 (-6)	1.80 (-5)	9.80 (-5)	3.43 (-4)
129	8.00 (-6)	1.20 (-5)	1.80 (-5)	3.70 (-5)	1.69 (-4)	3.65 (-4)
257	1.10 (-5)	1.80 (-5)	1.90 (-5)	5.40 (-5)	4.20 (-5)	5.50 (-5)

Table 3.15 $\bar{L}_\infty$ values for the indirect Chebychev Galerkin solution of the Dirichlet Helmholtz problem (the smooth trend of the errors and the conditioning of the system have made the relatively expensive computations for N=513 unnecesary).

Results from the indirect systems are given in the tables (3.14-15) and (3.16-17) for the Galerkin and the Filippi-pseudospectral respectively.

Although, the results appear to be identical to the direct method's (within round-off precision), it is of crucial significance to recognize the stabilization induced in the system by the indirect method's algorithm, which yields systems of satisfactory conditioning that do not suffer from the reported instabilities and, thereafter, maintain their high accuracy in the same fashion as the tau method does.

The relative conditioning of these unsymmetric matrices arising from both the direct and the indirect techniques was studied through singular value decomposition

$N \setminus k^2$	2	7	25	100	250	1000
5	1.19 (-4)	1.82 (-2)	5.12 (-1)	5.37 (-1)	9.99 (-1)	7.32 (-1)
9	$\star$	$\star$	1.87 (-4)	7.45 (-1)	7.44 (-1)	9.04 (-1)
17	$\star$	$\star$	$\star$	$\star$	1.67 (0)	9.17 (-1)
33	$\star$	$\star$	$\star$	$\star$	$\star$	3.69 (0)
65	$\star$	$\star$	$\star$	$\star$	$\star$	$\star$

Table 3.16 $\bar{L}_2$ values for the indirect Chebychev pseudospectral solution of the Dirichlet Helmholtz problem on a Filippi collocation point-set. Stars correspond to relative errors $\leq 10^{-7}$ and they persist until N=513, at least.

$N \setminus k^2$	2	7	25	100	250	1000
5	1.36 (-2)	1.54 (-1)	9.72 (-1)	9.11 (-1)	1.00 (0)	9.94 (-1)
9	1.00 (-6)	4.30 (-5)	2.13 (-2)	9.16 (-1)	9.83 (-1)	9.45 (-1)
17	1.00 (-6)	3.00 (-6)	5.00 (-6)	5.19 (-4)	1.28 (0)	1.00 (0)
33	2.00 (-6)	4.00 (-6)	7.00 (-6)	1.00 (-5)	1.20 (-5)	1.93 (0)
65	6.00 (-6)	6.00 (-6)	1.00 (-5)	8.00 (-6)	1.40 (-5)	3.50 (-5)
129	1.00 (-5)	1.10 (-5)	2.10 (-5)	1.80 (-5)	2.40 (-5)	6.30 (-5)
257	1.50 (-5)	1.80 (-5)	2.90 (-5)	2.60 (-5)	2.20 (-5)	5.10 (-5)

Table 3.17 $\bar{L}_\infty$ for the indirect Chebychev pseudospectral solution of the Dirichlet Helmholtz problem on a Filippi collocation point-set. The smooth trend of the error and the conditioning of the system have made the relatively expensive computations for N=513 unnecessary.

(SVD); the condition numbers were estimated as $\sigma_{max}/\sigma_{min}$, where σ stands for singular value. This is necessary because of the complex contributions present in the eigenvalue spectrum and it is equivalent to calculating the eigenvalues of $\mathbf{A}\mathbf{A}^{\mathbf{T}}$ which are the squares of the singular values of the matrix $\mathbf{A}$. A comparison of the condition numbers confirmed our suspicions, since the direct methods suffer from a conditioning several orders of magnitude poorer. The poor conditioning causes rapid deterioration of any inherent stability of the algebraic systems. Indirect Chebychev-collocation tests have not been performed, for it may be readily seen that such method would not

suffer from ill-conditioning and it would, therefore, exhibit a behavior similar to tau method's. It is important to note, here, that the direct tau algorithm does not suffer from ill-conditioning and this shows that the tau projection is more numerically stable in its straightforward implementation than in its equivalent collocation form. That confirms the previous reasoning, since the direct tau approach does involve more complicated computations, whereas it exhibits a different boundary condition treatment. The consequence of this is that an exact equivalence relation between the direct and the indirect versions of any projection operator occurs only for the tau method.

Finally, completing the discussion of the relative performance of the finite difference and the Chebychev tau, Galerkin and pseudospectral schemes, we should emphasize the fact that the Chebychev methods are greatly superior to the finite difference scheme. Galerkin and the Filippi-pseudospectral techniques need fewer points to achieve an accuracy of a certain order than tau and Chebychev-pseudospectral techniques, whereas the finite difference require more points. The exact order of this demand for denser sampling varies both with the method and the value of k^2 and it ranges from half to several orders of magnitude.

3.4.2 Inhomogeneous von Neumann boundary conditions

The Chebychev superiority established for the Dirichlet problem that has just been discussed, is anticipated to carry over to Neumann problems as well. Choosing the Neumann boundary conditions to be $y_x(x=-1)=\delta=+5$ and $y_x(x=+1)=\gamma=-2$ as before, the Helmholtz equation (standing waves, still) was solved with both the finite difference and the tau Chebychev method. The tau method was chosen among

the Chebychev spectral projection techniques, because it is the easiest to formulate and it provides a lower bound in the resolution capacity of these methods.

Tables (3.18-19) contain the errors associated with the tau solution; their inspection reveals a behavior identical to the Dirichlet case.

$N \setminus k^2$	2	7	25	100	250	1000
5	3.32 (-2)	5.83 (-2)	1.07 (0)	1.40 (0)	1.08 (0)	2.06 (0)
9	$\star$	$\star$	1.35 (-2)	2.07 (0)	7.99 (-1)	9.55 (-1)
17	$\star$	$\star$	$\star$	2.00 (-6)	1.50 (0)	1.01 (0)
33	$\star$	$\star$	$\star$	$\star$	$\star$	3.01 (0)
65	$\star$	$\star$	$\star$	$\star$	$\star$	$\star$

Table 3.18 $\bar{L}_2$ values for the Chebychev tau solutions of the Neumann Helmholtz problem. Stars correspond to relative errors $\leq 10^{-7}$ and they persist until N= 513, at least.

$N \setminus k^2$	2	7	25	100	250	1000
5	1.71 (-1)	2.87 (-1)	1.66 (0)	1.24 (0)	1.13 (0)	1.64 (0)
9	5.00 (-6)	8.20 (-5)	1.28 (-1)	1.98 (0)	1.23 (0)	1.14 (0)
17	2.00 (-6)	3.00 (-6)	3.00 (-6)	2.32 (-3)	1.30 (0)	1.08 (0)
33	2.00 (-6)	3.00 (-6)	4.00 (-6)	5.00 (-6)	8.00 (-6)	1.96 (0)
65	5.00 (-6)	3.00 (-6)	4.00 (-6)	6.00 (-6)	1.10 (-5)	3.80 (-5)
129	2.50 (-5)	5.00 (-6)	2.30 (-5)	1.00 (-6)	2.10 (-5)	5.60 (-5)
257	2.80 (-5)	1.10 (-5)	4.60 (-5)	7.00 (-6)	2.90 (-5)	8.30 (-5)
513	5.90 (-5)	1.80 (-5)	7.10 (-5)	1.20 (-5)	6.60 (-5)	8.80 (-5)

Table 3.19 $\bar{L}_\infty$ values for the Chebychev tau solutions of the Neumann Helmholtz problem.

This is not astonishing at all, since the approximation of the first derivatives at the boundaries converges exponentially; therefore, there is no generation of truncation error which could contaminate the solution. However, this high quality performance level would be lost if wrong implementation of the boundary conditions was to take

place as the boundary errors would rapidly be transmitted all over the computational domain due to the global characteristics of spectral methods.

$N \setminus k^2$	2	7	25	100	250	1000
5	1.08 (-1)	2.70 (0)	1.46 (0)	7.76 (-1)	9.65 (-1)	9.80 (-1)
9	4.26 (-3)	3.30 (-2)	3.19 (-1)	7.85 (-1)	9.52 (-1)	9.50 (-1)
17	2.41 (-4)	1.56 (-3)	2.70 (-2)	1.89 (0)	1.82 (0)	9.60 (-1)
33	1.70 (-5)	9.20 (-5)	2.16 (-3)	3.40 (-2)	7.32 (-1)	1.07 (+1)
65	4.00 (-6)	6.00 (-6)	1.47 (-4)	2.66 (-3)	3.60 (-1)	2.16 (0)
129	1.90 (-5)	1.00 (-6)	1.30 (-5)	1.82 (-4)	7.54 (-2)	3.55 (-1)
257	2.79 (-4)	7.00 (-6)	1.00 (-5)	1.70 (-5)	7.87 (-3)	7.42 (-2)
513	4.15 (-3)	8.60 (-5)	1.01 (-4)	1.60 (-5)	1.07 (-3)	7.72 (-3)

Table 3.20 $\bar{L}_2$ values for the finite difference solution of the Neumann Helmholtz problem with cental differences at the boundaries.

On the contrary, finite differences are affected by this change of the boundary conditions applied to the equation and they do show a relative deterioration with respect to their accuracy level in the Dirichlet case, although not as dramatic as it would have been for a spectal method, had the latter been implemented inappropriately.

$N \setminus k^2$	2	7	25	100	250	1000
5	3.30 (-1)	1.64 (0)	1.45 (0)	1.00 (0)	1.00 (0)	9.96 (-1)
9	6.54 (-2)	1.85 (-1)	8.74 (-1)	1.01 (0)	1.00 (0)	1.00 (0)
17	1.54 (-2)	4.00 (-2)	2.32 (-1)	2.06 (0)	1.90 (0)	1.00 (0)
33	4.04 (-3)	9.65 (-3)	6.12 (-2)	2.91 (-1)	9.38 (-1)	4.02 (+1)
65	1.93 (-3)	2.48 (-3)	1.59 (-2)	7.49 (-2)	6.15 (-1)	2.15 (0)
129	4.28 (-3)	1.11 (-3)	4.66 (-3)	1.94 (-2)	2.79 (-1)	6.56 (-1)
257	1.66 (-2)	2.54 (-3)	3.89 (-3)	5.92 (-3)	8.98 (-2)	2.87 (-1)
513	6.40 (-2)	9.20 (-3)	1.18 (-2)	5.42 (-3)	3.31 (-2)	9.14 (-2)

Table 3.21 $\bar{L}_\infty$ values for the finite difference solution of the Neumann Helmholtz problem with central differences at the boundaries.

The explanation of the reported deterioration lies, naturally, in the lower convergence rate, since additional truncation errors involved in the finite difference approximation of the boundary conditions are present.

The magnitude of these errors depend on the particular finite difference employed at the boundaries. Tables (3.20-21) give the results associated with the classic second order finite difference scheme with central differences for the boundary conditions, whereas tables (3.22-23) correspond to one-sided differences at the boundaries.

Central differences are symmetric (by employing fictitious points adjacent to each boundary) and they definitely exhibit a higher quality of results with respect to one-sided differences.

$N \setminus k^2$	2	7	25	100	250	1000
5	4.00 (-1)	1.74 (-1)	2.32 (0)	9.83 (0)	2.58 (0)	8.63 (0)
9	2.33 (-1)	9.88 (-2)	2.71 (+2)	2.44 (0)	1.38 (0)	3.45 (0)
17	1.08 (-1)	4.27 (-2)	1.25 (+1)	2.33 (0)	1.06 (0)	1.49 (0)
33	4.03 (-2)	1.49 (-2)	5.55 (-1)	9.94 (-2)	4.99 (-1)	1.53 (+1)
65	1.26 (-2)	4.46 (-3)	7.82 (-2)	4.89 (-2)	1.93 (+1)	7.70 (-1)
129	3.25 (-3)	1.20 (-3)	1.52 (-2)	1.36 (-2)	1.07 (+1)	1.01 (-1)
257	3.14 (-4)	2.50 (-4)	3.12 (-3)	3.36 (-3)	9.76 (-1)	6.88 (-2)
513	1.95 (-3)	1.20 (-5)	3.38 (-4)	7.03 (-4)	1.27 (-1)	5.98 (-2)

Table 3.22 $\bar{L}_2$ values for the finite difference solution of the Neumann Helmholtz problem with one-sided differences at the boundaries.

The low level of accuracy exhibited by the latter approach should be attributed to the *two point oscillation*, the latter being intrinsic to these unsymmetric difference approximations. Assuming that $u(x) = \exp(ikx)$, its true first derivative is

$$\left(\frac{du}{dx}\right)_{TR} = ik\exp(ikx) \tag{3.47}$$

whereas central differences give

$$\left(\frac{du}{dx}\right)_{CD} = ik\exp(ikx)\frac{\sin k\Delta x}{k\Delta x} \tag{3.48}$$

and one-sided differences yield

$$\left(\frac{du}{dx}\right)_{OS} = ik\exp(ikx)\exp(\pm ik\Delta x/2)\frac{\sin(k\Delta x/2)}{(k\Delta x/2)} \tag{3.49}$$

Both the difference approximations converge to the true solution (for $k\Delta x \to 0$), but it is evident that the presence of the complex exponential $\exp(\pm ik\Delta x/2)$ complicates and decelerates the convergence of the *sinc* function.

$N \setminus k^2$	2	7	25	100	250	1000
5	6.34 (-1)	4.25 (-1)	2.29 (0)	4.41 (0)	2.14 (0)	5.56 (0)
9	4.82 (-1)	3.25 (-1)	1.68 (+1)	2.24 (0)	1.53 (0)	3.23 (0)
17	3.27 (-1)	2.09 (-1)	3.50 (0)	8.19 (-1)	1.15 (0)	2.03 (0)
33	2.01 (-1)	1.23 (-1)	7.37 (-1)	3.94 (-1)	7.98 (-1)	4.91 (0)
65	1.13 (-1)	6.74 (-2)	2.76 (-1)	2.34 (-1)	4.37 (0)	1.19 (0)
129	5.72 (-2)	3.50 (-2)	1.22 (-1)	1.20 (-1)	4.12 (0)	3.90 (-1)
257	1.79 (-2)	1.61 (-2)	5.54 (-2)	5.88 (-2)	9.88 (-1)	2.77 (-1)
513	4.36 (-2)	4.09 (-3)	1.96 (-2)	2.74 (-2)	3.57 (-1)	2.46 (-1)

Table 3.23 $\tilde{L}_\infty$ values for the finite difference solution of the Neumann Helmholtz problem with one-sided differences at the boundaries

This interference becomes most profound for $k\Delta x = \pm\pi$ (the corresponding wavelength is $2\Delta x$); this is the so-called $2\Delta x$ *wave* and the latter scheme is unable to treat it properly.

3.4.3 Robbins and radiation boundary conditions

The Helmholtz problem has been solved with mixed (Robbins) boundary conditions applied to it, as well. As anticipated, the error behavior is similar to the von Neumann case for both the finite difference and the tau Chebychev schemes.

Complex Robbins boundary conditions of the absorbing type, i.e $u(x) \pm iku(x) = 0$, are easy to accommodate. It would be expected that the performances' levels should be roughly equivalent to the ones exhibited for the case of real Robbins boundary conditions. However, these constraints have not been tested numerically, since the imposition of such a radiation boundary condition on one boundary, while maintaining a Dirichlet or Neumann condition at the other (simulating either a soft or a hard scatterer, respectively), defaults to trivial solutions in the fundamental interval, as no energy is allowed there.

CHAPTER IV

THE HEAT EQUATION

My soul, my soulf itself, is this flame:
insatiable for new horizons
its silent glowing passion blazes upward.
Dithyrambs of Dionysus — Friedrich Nietzsche

4.1 The Homogeneous One-Dimensional Heat Equation

The Helmholtz equation investigated in the last chapter is an ordinary differential equation of the elliptic type. We proceed with a simple parabolic partial differential equation, namely, the *heat* equation

$$\frac{\partial}{\partial t}u(x,t) = \sigma(x)\frac{\partial^2}{\partial x^2}u(x,t) \tag{4.1}$$

The presence of the time derivative term complicates the situation and demands advanced considerations for the numerical formulations of the problem. The heat equa-

tion describes the diffusion of the function $u(x,t)$ in the course of time; the parameter $\sigma(x)$ is termed the *diffusion* coefficient.

4.2 Finite Differences

Let us start with the classic finite-difference implementation using again a second order difference scheme (a summary of various alternative higher-order schemes can be found in Panov (1963)). Approximation of the second order spatial derivative is done according to (3.34) and it is, thereafter, centered at (x,t). Problems, however, arise in the approximation of the first order time derivative, since by employing (3.37) we center it at $(x, t + \Delta t/2)$.

The resulting scheme

$$u_x^{t+\Delta t} = u_x^t + \frac{\sigma \Delta t}{(\Delta x)^2}(u_{x+\Delta x}^t - 2u_x^t + u_{x-\Delta x}^t) \tag{4.2}$$

is $O(\Delta t)+O((\Delta x)^2)$ accurate, *explicit* and extremely easy to solve but the *non-uniform* centering of the temporal and the spatial derivatives does not allow for an appropriate handling (attenuation) of the short wavelengths and therefore, the numerical solution diverges as Δx becomes smaller. This is a case of conditional instability; a classic Fourier stability analysis shows that stability can be guaranteed under the severe limitation that $\left(\sigma \Delta t/(\Delta x)^2\right) \leq 0.5$ (Botha and Pinder, 1983). This restricton has an interesting physical interpretation; $\Delta t_{\max}$ is half the diffusion time $\tau = (\Delta x)^2/\sigma$ and it may be understood as the time needed for information to travel over a distance Δx (Vemuri and Karplus, 1981).

An obvious alternative is the application of the leap-frog scheme for the time differencing, that is to say

$$\frac{\partial u(x,t)}{\partial t} = \frac{u(x,t+\Delta t) - u(x,t-\Delta t)}{2\Delta t} \tag{4.3}$$

so that time derivative is centered at (x,t), i.e

$$u_x^{t+\Delta t} = u_x^{t-\Delta t} + \frac{2\sigma\Delta t}{(\Delta x)^2}(u_{x+\Delta x}^t - 2u_x^t + u_{x-\Delta x}^t) \tag{4.4}$$

Although this scheme is $O(\Delta t)^2 + O((\Delta x)^2)$ accurate, it leads to an unconditional instability, which is caused by the time difference being taken over two steps since that makes the difference equation second order in time. Although one of its solutions is the solution being sought, the parasitic solution (see 4.4.2) is an oscillating increasing exponential and it, inevitably, contaminates the results (Claerbout, 1976).

The backwards Euler *implicit* scheme offers a first answer to the problem of instability by considering the second derivative difference operator centered at $(x, t+\Delta t)$. The scheme exhibits an accuracy of $O(\Delta t) + O((\Delta x)^2)$, is unconditionally stable but it requires a matrix inversion for each time step. The relevant equations are

$$-bu_{x+\Delta x}^{t+\Delta t} + (1+2b)u_x^{t+\Delta t} - bu_{x-\Delta x}^{t+\Delta t} = u_x^t \tag{4.5}$$

where $b = (\sigma\Delta t/(\Delta x)^2)$. We can also see that the matrix to be inverted is triadiagonal; these systems can be solved very efficiently (see Appendix B.3). What is the underlying

physical mechanism? The backwards-Euler scheme drives high frequency features into equilibrium, that is, equations (4.5) converge to the steady-state equation

$$\frac{\partial^2}{\partial x^2} u(x,t) = 0 \tag{4.6}$$

for large time steps ($\Delta t \rightarrow \infty$). We are usually interested in the evolution of features with spatial scales $\lambda >> \Delta x$ and although, the backwards Euler scheme alleviates the instabilty problems, it is not considered to be satisfactorily accurate, because its temporal truncation error is $O(\Delta t)$ and subsequently, substantial step size restrictions might be demanded due to accuracy considerations (Press et al, 1985).

A significant improvement may be accomplished by taking the averages of the schemes (4.2) and (4.5); a rigorous justification of this scheme based on the underlying bilinear transformation is given in Chapter V. The resulting *implicit* scheme avoids all of the above misfortunes and it has long been presented by Crank and Nicolson (Claerbout, 1976). It maintains the centering of the time derivative at $(x, t + \Delta t/2)$ and it also succeeds in centering the second order spatial derivative there too; the improved truncation error, i.e $O((\Delta t)^2) + O((\Delta x)^2)$, allowes a much larger Δt to be taken (Peaceman, 1977). The fast convergence of the method in association with its unconditional instability and its efficiency (tridiagonal systems to be inverted) have contributed to its great popularity. Envisioning the Crank-Nicolson diffusion model from a physical perspective shows that the time evolution of the initial distribution makes less sense than in the case of the backwards-Euler difference approximation. The low frequency components evolve amidst a fluctuating, but bounded, background of the original high frequency distribution; the inaccuracies associated with these components

is what we have to trade for a both stable and fast algorithm. It might be advisable to conclude a Crank-Nicolson procedure by shifting to backwards-Euler at the end of our computation, driving, thereby, the mishandled small scale features into steady-state (Press et al, 1985).

The Crank-Nicolson scheme leads to a tridiagonal set of equations to be solved at each time step; for the Dirichlet conditions $u(x=-1,t)=\beta$ and $u(x=+1,t)=\alpha$, the system reads (with $a=\sigma\Delta t/2(\Delta x)^2$)

$$\begin{pmatrix} 1+2a & -a & & & \\ -a & 1+2a & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 1+2a & -a \\ & & & -a & 1+2a \end{pmatrix} \begin{pmatrix} u_{-(N/2-1)} \\ \\ \vdots \\ \\ u_{+(N/2-1)} \end{pmatrix}^{t=t_0+\Delta t} =$$

$$\begin{pmatrix} 1-2a & +a & & & \\ +a & 1-2a & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 1-2a & +a \\ & & & +a & 1-2a \end{pmatrix} \begin{pmatrix} u_{-(N/2-1)} \\ \\ \vdots \\ \\ u_{+(N/2-1)} \end{pmatrix}^{t=t_0} + \begin{pmatrix} 2a\beta \\ 0 \\ \vdots \\ 0 \\ 2a\alpha \end{pmatrix} \tag{4.7}$$

Of course, $u_{-(N/2)}=\beta$ and and $u_{+(N/2)}=\alpha$, always.

The Neumann conditions $u_x(x=-1,t)=\delta$ and $u_x(x=+1,t)=\gamma$ are handled through appropriate changes of the corner elements of the recursion matrices and the boundary vector. One-sided differences at the boundaries of the computational grid

result in the system

$$\begin{pmatrix} 1+a & -a & & & \\ -a & 1+2a & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 1+2a & -a \\ & & & -a & 1+a \end{pmatrix} \begin{pmatrix} u_{-(N/2-1)} \\ \\ \vdots \\ \\ u_{+(N/2-1)} \end{pmatrix}^{t=t_0+\Delta t} =$$

$$\begin{pmatrix} 1-a & +a & & & \\ +a & 1-2a & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 1-2a & +a \\ & & & +a & 1-a \end{pmatrix} \begin{pmatrix} u_{-(N/2-1)} \\ \\ \vdots \\ \\ u_{+(N/2-1)} \end{pmatrix}^{t=t_0} + \begin{pmatrix} -2a\delta\Delta x \\ 0 \\ \vdots \\ 0 \\ +2a\gamma\Delta x \end{pmatrix} \tag{4.8}$$

The values of $u(x)$ at the endpoints are calculated as $u(x = -1) = u_{-(N/2)} = u_{-(N/2-1)} - \delta\Delta x$ and $u(x = +1) = u_{+(N/2)} = u_{+(N/2-1)} + \gamma\Delta x$. Claerbout (1976) adopts this scheme, although he only considers zero slope boundary conditions ($\delta = \gamma = 0$).

The central difference approximation of the first derivatives at the boundaries (Keller, 1968) may be employed as well. Such an approach yields the recursion system

$$\begin{pmatrix} 1+2a & -2a & & & \\ -a & 1+2a & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 1+2a & -a \\ & & & -2a & 1+2a \end{pmatrix} \begin{pmatrix} u_{-(N/2)} \\ \\ \vdots \\ \\ u_{+(N/2)} \end{pmatrix}^{t=t_0+\Delta t} =$$

$$\begin{pmatrix} 1-2a & +2a & & & \\ +a & 1-2a & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 1-2a & +a \\ & & & +2a & 1-2a \end{pmatrix} \begin{pmatrix} u_{-(N/2)} \\ \\ \vdots \\ \\ u_{+(N/2)} \end{pmatrix}^{t=t_0} + \begin{pmatrix} -4a\delta\Delta x \\ 0 \\ \vdots \\ 0 \\ +4a\gamma\Delta x \end{pmatrix} \tag{4.9}$$

As the dimensionality of the equation increases though, implicit schemes become rather cumbersome. Alternatives include the Dufort-Frankel (explicit) scheme or symmetric semi-implicit schemes — especially if non-linearities are present (Livne and Glasner, 1985). A word of caution needs to be given for the unconditionally stable Dufort-Frankel scheme. Consistency and stability are both needed to ensure convergence of the numerical solution (Lax's equivalence theorem). The Dufort-Frankel scheme is consistent with the diffusion equation only if $\Delta t \rightarrow 0$ faster than $\Delta x \rightarrow 0$; otherwise, the scheme is still stable but consistent with the hyperbolic equation

$$u_t - u_{xx} + (\Delta t/\Delta x)^2 u_{tt} = 0 \tag{4.10}$$

instead and it therefore converges to the true solution of the latter equation as Δx and Δt tend to zero (Vemuri and Karplus, 1981).

4.3 Chebychev Methods

Spectral methods for the numerical solution of the heat equation present a viable alternative. Semi-discretizations involve a spectral representation of $\partial^2/\partial x^2$ combined with a finite difference approximation to $\partial/\partial t$. If the boundary conditions in the spatial coordinate are periodic, Fourier expansions are most appropriate; otherwise, a Chebychev expansion is the optimum alternative.

The homogeneous heat equation (4.1) is assumed to satisfy homogeneous Dirichlet boundary conditions and consequently, all the various Chebychev projection operators may be applied without any need for pre-processing. For the tau method the basis functions are the original Chebychev polynomials, while for the Galerkin and the

collocation methods, we need to make use of the appropriate basis functions $q_n(x)$ defined in (3.18). Although, both direct and indirect approaches may be pursued as before, complications associated with the the time derivative (see below) make the indirect systems easier to formulate; direct systems will not formulated.

4.3.1 The differentiated tau method

The expansions involve coefficients which depend on time, i.e $a_m^{(0)}(t)$ and $a_m^{(-2)}(t)$ for $u(x,t)$ and $u_{xx}(x,t)$, respectively.

Employing the usual procedure, we obtain

$$\frac{d}{dt}a_m^{(0)} = \sigma a_m^{(-2)} \quad \text{for} \quad m = 0, \ldots, N-2 \tag{4.11}$$

with the boundary conditions

$$\sum_{m=0}^{N} a_m^{(0)} = 0 \qquad \text{and} \qquad \sum_{m=0}^{N} (-1)^m a_m^{(0)} = 0 \tag{4.12}$$

Although we could use this mixed (o.d.e's and algebraic equations) system to apply the finite differences in time (since the o.d.e's will be transformed into algebraic equations), we would like to transform the system into one of a pure differential form, i.e

$$\frac{d}{dt}\mathbf{a}^{(0)} = \mathbf{A}\mathbf{a}^{(0)} \qquad \text{with} \quad \mathbf{a}^{(0)} = (a_0^{(0)}, \ldots, a_n^{(0)}) \tag{4.13}$$

that is to say, we need to augment the system with two o.d.e's for $(d/dt)a_{N-1}^{(0)}(t)$, $(d/dt)a_N^{(0)}(t)$ and build the boundary constraints into $\mathbf{A}$, so that compliance with

them is maintained. Differentiating the boundary conditions (4.12) with respect to t and combining the results with equations (4.11), yields a system for the unknowns $(d/dt)a_{N-1}^{(0)}$ and $(d/dt)a_N^{(0)}$, consisting of the equations

$$\sigma \sum_{m=0}^{N-2} a_m^{(-2)} = -\frac{d}{dt}a_{N-1}^{(0)} - \frac{d}{dt}a_N^{(0)} \tag{4.14}$$

$$\sigma \sum_{m=0}^{N-2} (-1)^m a_m^{(-2)} = -(-1)^{N-1}\frac{d}{dt}a_{N-1}^{(0)} - (-1)^N \frac{d}{dt}a_N^{(0)} \tag{4.15}$$

The solution may be expressed as

$$\frac{d}{dt}a_{N-1}^{(0)} = \sigma a_{N-1}^{(-2)} + \left(-\frac{1}{2}\right)\left[\sigma\frac{\partial^2 u}{\partial x^2}\right]_{x=+1} + (-1)^N \left[\sigma\frac{\partial^2 u}{\partial x^2}\right]_{x=-1} \tag{4.16}$$

$$\frac{d}{dt}a_N^{(0)} = \sigma a_N^{(-2)} + \left(-\frac{1}{2}\right)\left[\sigma\frac{\partial^2 u}{\partial x^2}\right]_{x=+1} + (-1)^{N+1} \left[\sigma\frac{\partial^2 u}{\partial x^2}\right]_{x=-1} \tag{4.17}$$

with $\left[\sigma\frac{\partial^2 u}{\partial x^2}\right]_{x=+1} = \sigma\sum_{m=0}^{N} a_m^{(-2)}$ and $\left[\sigma\frac{\partial^2 u}{\partial x^2}\right]_{x=-1} = \sigma\sum_{m=0}^{N}(-1)^m a_m^{(-2)}$. Augmenting equations (4.11) with (4.16) and (4.17), we achieve our goal; the new system may be written uniformly, as

$$\frac{d}{dt}a_m^{(0)} = \frac{\sigma}{c_m}\sum_{\substack{p=m+2 \\ p+m:\ even}}^{N} p(p^2-m^2)a_p^{(0)} + \delta_{m,N-1}b_1(t) + \delta_{m,N}b_2(t) \qquad \text{for } m = 0,\ldots,N \tag{4.18}$$

where

$$b_1(t) = -\frac{1}{2}\left[\sigma\frac{\partial^2 u}{\partial x^2}\right]_{x=+1} + \frac{1}{2}(-1)^N\left[\sigma\frac{\partial^2 u}{\partial x^2}\right]_{x=-1} \tag{4.19}$$

and

$$b_2(t) = -\frac{1}{2}\left[\sigma\frac{\partial^2 u}{\partial x^2}\right]_{x=+1} + \frac{1}{2}(-1)^{N+1}\left[\sigma\frac{\partial^2 u}{\partial x^2}\right]_{x=-1} \tag{4.20}$$

4.3.2 The integrated tau method

The relevant equations are developed following a procedure identical to the one presented for the Helmholtz equation.

A double indefinite integration of the equation leads to

$$\int dx \int \frac{\partial u}{\partial t}\, dx = \sigma[u + xA(t) + B(t)] \tag{4.21}$$

where A and B are arbitrary functions of t. Assuming the familiar expansions for $u(x,t)$ and $\int dx \int u(x,t)\, dx$, the projection with the T_m's (for $m = 0, \ldots, N+2$) is performed; the equations for $m = 0$ and $m = 1$ are omitted, since they contain the undetermined constants of integration A and B. The equations for $m = N+1$ and $m = N+2$, i.e

$$\frac{d}{dt}a_{N+1}^{(+2)} = 0 \qquad \text{and} \qquad \frac{d}{dt}a_{N+2}^{(+2)} = 0 \tag{4.22}$$

imply that (see 3.3.1.3)

$$\frac{d}{dt}a_{N-1}^{(0)} = 0 \qquad \text{and} \qquad \frac{d}{dt}a_{N}^{(0)} = 0 \tag{4.23}$$

in the projection equations for $m = N-4, \ldots, N$ and $m = N-1, N$, respectively.

The resulting system reads

$$\frac{d}{dt} a_m^{(+2)} = \sigma a_m^{(0)} \qquad \text{for } m = 0, \ldots, N \tag{4.24}$$

and substituting $a_m^{(+2)}$ according to (A.49) into (4.24), the final equations can be expressed, in terms of the $a_m^{(0)}$'s only, as

$$\frac{c_{m-2}}{4m(m-1)} \frac{d}{dt} a_{m-2}^{(0)} - \frac{e_{m+2}}{2(m^2-1)} \frac{d}{dt} a_m^{(0)} + \frac{e_{m+4}}{4m(m+1)} \frac{d}{dt} a_{m+2}^{(0)} = \sigma a_m^{(0)} \; ; \; m = 2, \ldots, N \tag{4.25}$$

augmented with the boundary conditions (4.12).

4.3.3 The Galerkin method

The $\{q_m\}$ set needs to be used in the expansion for the approximation

$$u(x,t) = \sum_{m=2}^{N} \bar{a}_m^{(0)} q_m(x) \tag{4.26}$$

The post-projection analysis shows that expressing $u_{xx}(x,t)$ in terms of the $q_m(x)$'s too, causes a coupling among the $\bar{a}_m^{(0)}$'s, while the use of the non-orthogonal $q_m(x)$'s introduces an additional coupling among the $(d/dt)\bar{a}_m^{(0)}$'s (in the Helmholtz equation this latter coupling involved the $a_m^{(0)}$'s, instead).

Consequently, the resulting system to be solved has the cumbersome form

$$\mathbf{B} \frac{d}{dt} \bar{\mathbf{a}}^{(0)} = \mathbf{C} \bar{\mathbf{a}}^{(0)} \tag{4.27}$$

(with the matrix **B** being non-diagonal); thereafter straightforward inversion is very inefficient. Diagonalization of **B** corresponds to orthogonalization of the $\{q_m\}$ set, that is to say, employing the $\{T_m\}$ set for the expansion of $u(x,t)$, i.e $u(x,t) = \sum_{m=0}^{N} a_m^{(0)} T_m(x)$. Subsequently, trivial normalizations transform **B** to the identity matrix **I**, yielding the desired output form

$$\frac{d}{dt}\mathbf{a}^{(0)} = \mathbf{A}\mathbf{a}^{(0)} \tag{4.28}$$

The manipulations involved may be briefly summarized as follows: summing up properly modified versions of equations (4.27), we identify the coefficients $a_0^{(0)}$ and $a_1^{(0)}$ (which are missing in the q_n expansion but are present in the desired T_n expansion) as linear combinations of the $\bar{a}_2^{(0)}, \bar{a}_3^{(0)}, \ldots, \bar{a}_N^{(0)}$, so that the boundary conditions are still satisfied. Two new equations for the time-derivatives of the augmented coefficients are constructed; these are then used to modify the rest of the equations, so that both consistency with the new expansion's characteristics and satisfaction of the boundary constraints are maintained at all times.

The equivalent pre-projection formulation (assuming a $\mathbf{a}_0^{(0)}$ solution vector associated with the $\{T_n\}$ set), imposes the constraints on $\mathbf{a}_0^{(0)}$ in a later stage; the system of equations reads (Gottlieb and Orszag, 1977)

$$\frac{d}{dt}a_m^{(0)} = \frac{\sigma}{c_m} \sum_{\substack{p=m+2 \\ p+m:\ even}}^{N} p(p^2 - m^2)a_p^{(0)} + \frac{1}{c_m}b_1(t) + \frac{(-1)^m}{c_m}b_2(t), \quad \text{for } m = 0, \ldots, N \tag{4.29}$$

with

$$b_1(t) = \left(-\frac{N+\frac{1}{2}}{N^2+N}\right)\left[\sigma\frac{\partial^2 u}{\partial x^2}\right]_{x=+1} + \frac{(-1)^N}{2(N^2+N)}\left[\sigma\frac{\partial^2 u}{\partial x^2}\right]_{x=-1} \tag{4.30}$$

$$b_2(t) = \frac{(-1)^N}{2(N^2+N)}\left[\sigma\frac{\partial^2 u}{\partial x^2}\right]_{x=+1} + \left(-\frac{N+\frac{1}{2}}{N^2+N}\right)\left[\sigma\frac{\partial^2 u}{\partial x^2}\right]_{x=-1} \tag{4.31}$$

4.3.4 The collocation method

The pseudospectral system exhibits disadvantages nearly identical to Galerkin's. The use of the $\{q_m\}$ set is, once more, responsible for the inefficient system that results. The post-projection analysis shows that the problems (due to further coupling among the $(d/dt)\bar{a}_m^{(0)}$'s) are more pronounced here and the decoupling transformation needs to overcome the dependence of the projection equations on the collocation points.

Diagonalization of the system advances as follows: augmentation of the collocation grid with the two boundary points is performed, and subsequently the collocation equations for these points are constructed. Special discrete orthogonality properties of the Chebychev polynomials, i.e

$$\sum_{j=0}^{N} \frac{1}{\bar{c}_j} T_l(x_j) T_n(x_j) = \frac{N}{2}\bar{c}_l \delta_{ln} \quad \text{for } l, n \in [0, N] \tag{4.32}$$

on the complete Filippi set i.e, $x_j = \cos(\pi j/N)$ for $j = 0, \ldots, N$, or

$$\sum_{j=0}^{N} T_l(x_j) T_n(x_j) = \frac{N+1}{2} c_l \delta_{ln} \quad \text{for } l, n \in [0, N] \tag{4.33}$$

on the zeros of T_{N+1}, i.e $x_j = \cos[(j+1/2)/(N+1)])$ for $j = 0, \ldots, N$, are then used to support a second stage of decomposing the collocation sums, eliminating, thereby, their dependence on the collocation set of points. A last stage involves the final diagonalization process, where the coefficients $a_0^{(0)}$ and $a_1^{(0)}$ are identified as certain sums (see 3.28) of the coupled coefficients $\bar{a}_m^{(0)}$. Manipulating the resulting system, in association with the discrete orthogonality relationships previously mentioned, allows construction of equations for $(d/dt)a_0^{(0)}$ and $(d/dt)a_1^{(0)}$. Incorporating those into the rest of the equations (to guarantee compliance with the boundary conditions in the course of time) completes the procedure.

The equivalent pre-projection approach might be used to derive the same final system of equations

$$\frac{d}{dt}a_m^{(0)} = \frac{\sigma}{c_m} \sum_{\substack{p=m+2 \\ p+m:\ even}}^{N} p(p^2-m^2)a_p^{(0)} + \frac{1}{\bar{c}_m}b_1(t) + \frac{(-1)^m}{\bar{c}_m}b_2(t) \quad \text{for } m = 0, \ldots, N \tag{4.34}$$

with

$$b_1(t) = -\frac{1}{N}\left[\sigma\frac{\partial^2 u}{\partial x^2}\right]_{x=+1} \qquad \text{and} \qquad b_2(t) = -\frac{1}{N}\left[\sigma\frac{\partial^2 u}{\partial x^2}\right]_{x=-1} \tag{4.35}$$

Collocation-Chebychev methods for the heat equation have been shown to be stable (Gottlieb, 1981, and Gottlieb and Lustman, 1983b).

4.3.5 Time differencing in Chebychev semi-discretizations

The final system of equations for all but the integrated tau methods may be written under the general form

$$\frac{da_n}{dt} = \frac{1}{c_n} \sum_{\substack{p=n+2 \\ p+n:\ even}}^{N} p(p^2 - n^2)a_p + b_1(t)B_{1n} + b_2(t)B_{2n} \qquad (4.36)$$

where the terms B_{in}, and

$$b_i(t) = c_{i+}\left[\sigma \frac{\partial^2 u}{\partial x^2}\right]_{x=+1} + c_{i-}\left[\sigma \frac{\partial^2 u}{\partial x^2}\right]_{x=-1} \qquad (4.37)$$

(which ensure compliance with the boundary constraints $\sum_{n=0}^{N} a_n = \sum_{n=0}^{N}(-1)^n a_n = 0$) vary in the different methods (compare equations 4.18-20, 4.29-31 and 4.34-35); the notation used here is after Gottlieb and Orszag, (1977).

Semi-discretizations usually proceed by employing a finite difference scheme for the remaining variable, i.e for t (finite elements in semi-discretizations may be used too). Time integration was discussed in (2.5). Thus, we limit ourselves in a brief presentation of the problems and the possible ways around them.

Explicit time differencing schemes enjoy remarkably easy formulation and they do not demand matrix inversions; these advantages are virtually counterbalanced by the overwhelming restrictions, which need to be imposed on the size of the time step Δt, to ensure stability of the semi-discrete scheme.

The source of this problem is found at the clustering of the Chebychev grid points near the endpoints -1 and $+1$. Such a feature supports high resolution in the vicinity of

the boundaries, but it, concurrently, imposes the severe limitation that Δt be smaller than $(1/\sigma M^4)$, where M is the number of Chebychev polynomials retained in the expansion in x. This is so, because the dense boundary regions have a resolution $\Delta x \sim O(1/M^2)$; alternatively, this might be explained on the basis of the fact that the largest eigenvalue of the Chebychev matrix representation to $\partial^2/\partial x^2$ grows as $1/M^4$, as it might be seen from application of Gershgorin's theorem (Zang et al, 1982). Despite those problems, explicit schemes are used quite often; the fourth-order Runge-Kutta scheme enjoys a great popularity (Hussaini et al, 1983).

Alternatives have been found in either explicit schemes involving some kind of filtering (see 2.5.2), i.e Dufort-Frankel (Gottlieb and Gustafsson, 1976; Gottlieb and Lustman, 1983a), or in schemes involving implicit treatment (see 2.5.2 and 2.5.3).

In order the express the system (4.36-37) in matrix notation, we define the $(N+1)$-long column vectors $\mathbf{B_1}$, $\mathbf{B_2}$, $\mathbf{d}$ and $\mathbf{e}$ as $(B_1)_n = B_{1n}$, $(B_2)_n = B_{2n}$,

$$d_n = (+1)^{n+2}\frac{n^2(n^2-1)}{3} \quad \text{and} \quad e_n = (-1)^{n+2}\frac{n^2(n^2-1)}{3} \tag{4.38}$$

The definitions of the last two vectors result from the evaluation of the second derivative at $x = \pm 1$, respectively; this is accomplished via the expression (Gottlieb and Orszag, 1977)

$$\frac{d^p}{dx^p}T_n(\pm 1) = (\pm 1)^{n+p}\prod_{k=0}^{p-1}\frac{n^2-k^2}{2k+1} \tag{4.39}$$

Denoting the Chebychev representation of $\partial^2/\partial x^2$ by the singular matrix $\bar{\bar{\mathbf{C}}}$ (the two last columns are zero reflecting the loss of two degrees of freedom), we may write our

system as (double bars as used to distinguish matrices from vectors also expressed with capital letters):

$$\frac{d}{dt}\mathbf{a}^{(0)} = \sigma[\bar{\bar{\mathbf{C}}}\mathbf{a}^{(0)} + c_{1+}\mathbf{B_1}(\mathbf{d}^T\mathbf{a}^{(0)}) + c_{1-}\mathbf{B_1}(\mathbf{e}^T\mathbf{a}^{(0)}) + c_{2+}\mathbf{B_2}(\mathbf{d}^T\mathbf{a}^{(0)}) + c_{2-}\mathbf{B_2}(\mathbf{e}^T\mathbf{a}^{(0)})] \tag{4.40}$$

or

$$\frac{d}{dt}\mathbf{a}^{(0)} = \sigma\bar{\bar{\mathbf{A}}}\mathbf{a}^{(0)} \tag{4.41}$$

where $\bar{\bar{\mathbf{A}}} = \bar{\bar{\mathbf{C}}} + \bar{\bar{\mathbf{B}}}$ and

$$\bar{\bar{\mathbf{B}}} = \{[c_{1+}(\mathbf{B_1}\mathbf{d^T})] + [c_{1-}(\mathbf{B_1}\mathbf{e^T})] + [c_{2+}(\mathbf{B_2}\mathbf{d^T})] + [c_{2-}(\mathbf{B_2}\mathbf{e^T})]\} \tag{4.42}$$

The Crank-Nicolson and the backwards Euler implicit schemes are unconditionally stable for Chebychev semi-discretizations (Gottlieb and Orszag, 1977). Formulating the Crank-Nicolson scheme for our system

$$u_{t+\Delta t} - u_t = \sigma\Delta t\,\bar{\bar{\mathbf{A}}}\left(\frac{u_{t+\Delta t} + u_t}{2}\right) \tag{4.43}$$

leads to the system

$$\left(\bar{\bar{\mathbf{I}}} - \frac{\sigma\Delta t}{2}\bar{\bar{\mathbf{A}}}\right)[\mathbf{a}^{(0)}]_{t+\Delta t} = \left(\bar{\bar{\mathbf{I}}} + \frac{\sigma\Delta t}{2}\bar{\bar{\mathbf{A}}}\right)[\mathbf{a}^{(0)}]_t \tag{4.44}$$

The backwards Euler scheme

$$u_{t+\Delta t} - u_t = \sigma \Delta t \, \bar{\bar{\mathbf{A}}} u_{t+\Delta t} \tag{4.45}$$

results into the system

$$\left(\bar{\bar{\mathbf{I}}} - \sigma \Delta t \bar{\bar{\mathbf{A}}}\right) [\mathbf{a}^{(0)}]_{t+\Delta t} = [\mathbf{a}^{(0)}]_t \tag{4.46}$$

The integrated scheme is consisted of both ordinary differential and algebraic equations; this has not been modified in order for the quasi-tridiagonal form to be maintained. Employing the Crank-Nicolson scheme leads into the system

$$\begin{aligned} &\alpha(n)[a^{(0)}_{n-2}]_{t+\Delta t} - \beta(n)[a^{(0)}_n]_{t+\Delta t} + \gamma(n)[a^{(0)}_{n+2}]_{t+\Delta t} = \\ &\alpha(n)[a^{(0)}_{n-2}]_t + \delta(n)[a^{(0)}_n]_t + \gamma(n)[a^{(0)}_{n+2}]_t \quad \text{for } n = 2, \ldots, N \end{aligned} \tag{4.47}$$

where $\alpha(n) = c_{n-2}/4n(n-1)$, $\beta(n) = \sigma\Delta t/2 + e_{n+2}/2(n^2-1)$, $\gamma(n) = e_{n+4}/4n(n+1)$ and $\delta(n) = \sigma\Delta t/2 - e_{n+2}/2(n^2-1)$.

The backwards Euler formulation of the integrated system reads

$$\begin{aligned} &\alpha(n)[a^{(0)}_{n-2}]_{t+\Delta t} - \epsilon(n)[a^{(0)}_n]_{t+\Delta t} + \gamma(n)[a^{(0)}_{n+2}]_{t+\Delta t} = \\ &\alpha(n)[a^{(0)}_{n-2}]_t - \varsigma(n)[a^{(0)}_n]_t + \gamma(n)[a^{(0)}_{n+2}]_t \quad \text{for } n = 2, \ldots, N \end{aligned} \tag{4.48}$$

with $\epsilon(n) = \sigma\Delta t + e_{n+2}/2(n^2-1)$ and $\varsigma(n) = e_{n+2}/2(n^2-1)$.

Systems (4.47) and (4.48) need two more equations that are given from the boundary conditions: These should be imposed on the vector at the new time step, i.e $t+\Delta t$ (backwards Euler) or they should be centered (Crank-Nicolson) at $t+(\Delta t/2)$. The corresponding equations are

$$\sum_{n=0}^{N}[a_n^{(0)}]_{t+\Delta t}=0 \qquad \text{and} \qquad \sum_{n=0}^{N}(-1)^n[a_n^{(0)}]_{t+\Delta t}=0 \tag{4.49}$$

or

$$\sum_{n=0}^{N}[a_n^{(0)}]_{t+\Delta t}=-\sum_{n=0}^{N}[a_n^{(0)}]_t=0 \text{ and } \sum_{n=0}^{N}(-1)^n[a_n^{(0)}]_{t+\Delta t}=\sum_{n=0}^{N}(-1)^{n+1}[a_n^{(0)}]_t=0 \tag{4.50}$$

4.4 Discussion of Results

The numerical algorithms have been tested against the exact solution to the heat equation (with $\sigma=1$ and $u(x=\pm 1,t)=0$) for the initial heat distribution $u(x,t=0)=\sin\pi x$ (after Hussaini et al (1983)), which reads

$$u(x,t)=e^{-\pi^2 t}\sin\pi x \tag{4.51}$$

Although the Crank-Nicolson scheme enjoys a great popularity and a high reputation for problems associated with heat conduction and diffusion, it involves a number of subtleties; those should be well understood before the technique is blindly applied.

4.4.1 Crank-Nicolson stability analysis

Assuming an $(N+1)$-long discretization of the fundamental interval $[-1,+1]$, let us write down the Crank-Nicolson system as

$$\mathbf{A}u' = \mathbf{B}u + k \tag{4.52}$$

where u' denotes the solution at the next time step. Stability is associated with the eigenvalues of the matrix $\mathbf{A^{-1}B}$, since

$$u' = \mathbf{A^{-1}B}u + \mathbf{A^{-1}}k \tag{4.53}$$

The eigenvalues μ_j of $A^{-1}B$ may be calculated as $\mu_j = (1 - a\lambda_j)^{-1}(1+a\lambda_j)$, where λ_j are the eigenvalues of the tridiagonal $(\Delta^2/\Delta x^2)$ operator (symmetric for the Dirichlet but unsymmetric for the Neumann problem).

The μ_j spectrum for Dirichlet conditions reads

$$\mu_j = \frac{1 - 4a\sin^2(j\pi/2N)}{1 + 4a\sin^2(j\pi/2N)} \quad \text{for} \quad j = 1,\ldots,N-1 \tag{4.54}$$

(two less than in the general case) and it immediately obvious that $|\mu_j| < 1$ for all positive values of a; an unconditional stability is always guaranteed.

The Neumann spectrum (equation 4.8) reads

$$\mu_j = \frac{1 - 4a\cos^2(j\pi/2N)}{1 + 4a\cos^2(j\pi/2N)} \quad \text{for} \quad j = 0,\ldots,N \tag{4.55}$$

and a trivial inspection reveals that the eigenvalues μ_j satisfy $|\mu_j| \leq 1$ since $\mu_N = 1$. The method is still stable but the presence of the unit eigenvalue is responsible for a persistent error in the numerical approximation (Mitchell and Griffiths, 1980). This error is coupled with the size of the parameter a; if a is not chosen appropriately then the numerical solution suffers from a high-frequency oscillation known as "Crank-Nicolson noise" (Wood and Lewis, 1975). This problem is also identified in problems of the mixed (Robbins) type and it is briefly discussed in Mitchell and Griffiths (1980); a detailed analysis is given in Keast and Mitchell (1967). Alleviation of the problem might require a significant decrease of the time step; alternative procedures involve noise elimination techniques, while trying to maintain an adequately fast time stepping pace (Wood and Lewis, 1975).

The Dirichlet case has been classified as unconditionally stable and free of persistent errors, since all the eigenvalues μ_j are clearly smaller than unity. A deeper understanding of the physics of the problem provides additional insight. Following Ames (1977), we compare the exact solution in terms of the infinite sine series

$$u(x,t) = \sum_{n=1}^{\infty} \hat{u}_n \exp(-n^2\pi^2 t) \sin(n\pi x) \tag{4.56}$$

with the finite difference approximation

$$v(i\Delta x, j\Delta t) = \sum_{n=1}^{N-1} \hat{v}_n \mu_n^j \sin(in\pi\Delta x) \tag{4.57}$$

The eigenvalues μ_n^j correspond to the numerical approximations to $\exp(-n^2\pi^2 t)$ and we can see that as Δx, $\Delta t \to 0$ (with a fixed), the eigenvalues $\mu_n^j \to \exp(-n^2\pi^2 t)$

ensuring that the scheme is consistent; Lax's equivalence theorem then guarantees that the scheme is convergent.

However, the μ_j spectrum depends on the size of the parameter $a = \Delta t/2(\Delta x)^2$ and it is of fundamental importance to realize that changes in the value of a affect, in a nonlinear fashion, the relative decay rates of the eigenmodes of the discrete model. Subsequently, it is crucial to make sure that the time evolution of the discrete spectrum follows the details of the evolution of the continuous one, both on an individual and a global basis. For the problem under current consideration, we need a spectrum that is dominated by the factor $\exp(-n^2\pi^2 t)$ or equivalently the largest eigenvalue μ_1 should be clearly controlling the spectrum. Additionally, we must take caution that μ_1 stays positive in order to avoid contamination from extraneous oscillations.

These constraints may be expressed as (Ames, 1977)

$$\mu_1 > 0 \qquad \text{and} \qquad \frac{|\mu_n|}{\mu_1} < 1 \quad \text{for all } n > 1 \tag{4.58}$$

and they impose a restriction on the size of Δt that can be used, which is

$$\Delta t < \frac{4}{N^2 \sin(2\pi/N)} \tag{4.59}$$

for the scale of our particular problem. Although an actual limitation to be taken into serious consideration, the above bound is not considered significant, since for a given spatial discretization, accuracy considerations would usually demand a step size of at least that level.

Furthermore, Flatt (1961) has shown that uniform stability (for Crank-Nicolson diffusion systems) depends on the length of the computational domain imposing, thereafter, further restrictions on the permissible value of the parameter a.

4.4.2 Absolute versus relative stability

Another important point should be discussed before the analysis of the numerical algorithms' performances. This is associated with the choice of the norm in which the accuracy of each method is to be measured. In the last chapter both the relative L_2 and L_∞ error norms were employed and it was decided that the $\bar{L}_\infty$ norm comprised a better indicator of the performance of a method. The infinity norm is a very pessimistic one and it may be misleading if extraneous solution points are present; in addition, its normalized version should be used with caution, for it would tend to magnify the error estimates when the solution itself becomes negligible. Despite this, we believe that an error analysis and an algorithm evaluation based on the $\bar{L}_\infty$ norm is absolutely adequate for our problem, at least as the solution remains both smooth (free of input noise) and far away from zero.

Both the Crank-Nicolson and the backwards-Euler solutions of either the finite difference or the Chebychev system, involve a repetitive application of a certain (linear) operator in order to advance the given initial condition successively in time. What is the nature of the errors arising in the recurrence? First, we have to consider that the initial condition fed into the discrete system is not exact due to round-off errors; second, we need to be aware of the fact that at every step of the repetitive application operator errors are both committed and transferred to the next step. The generation of errors is due to the fact that the operator itself is only an approximation of certain

order to the true differential one and therefore, a truncation error is always present. Additionally, even this inaccurate calculation is not done perfectly since the computations are susceptible to round-off problems. Furthermore, all the errors present in the solution vector at a certain time step are themselves being propagated to the next step through the computation.

Let us now return to the issue of stability and its exact meaning in a recurrence environment. The notion of stability has been presented in an absolute sense only. In other words, the schemes are stable in the sense that errors do not get magnified through the recurrence; consequently, absolute errors remain bounded. This kind of stability is termed *absolute stability.* It is obvious that such a stability definition is consistent with the L_∞ norm and it is doubtful whether an $\bar{L}_\infty$ error analysis is compatible with that. However, our persistence in examining the $\bar{L}_\infty$ errors instead of the L_∞ is further justified by the fact that the diffusion equation damps the initial distribution in the course of time and therefore, we should be investigating the decay of the approximate solution with respect to the decay of the true solution in a relative than in an absolute fashion.Thus a reliable estimate of the accuracy of the computed solution at some time step is obtained.

It would then be only natural to analyze the stability of the scheme in the normalized norm, thereby, investigating the *relative stability* of the given scheme, that is, the behavior of the relative (percentage) error as the number of iterations tends to infinity. Confusion prevails since numerical analysts themselves do not appear to follow a uniform path with respect to the definition of absolute and relative stability.

Pizer and Wallace (1983) define relative stability for both single- and multi-step recurrence methods as a description of the behavior of the relative error ϵ_i/y_i (where

i is the recurrence level) as $i \to \infty$. Naturally, absolute and relative stability do not coincide, unless the absolute error magnitude goes to zero at the same rate as the exact solution does. Otherwise, an absolutely stable scheme is unstable in a relative sense if $\epsilon_i \to 0$ more slowly than $y_i \to 0$ and an absolutely unstable scheme is relatively stable if $\epsilon_i \to \infty$ more slowly than y_i does (Pizer and Wallace, 1983).

The majority of numerical analysts (Ralston and Rabiniwitz, 1978, Gear, 1971) employ a different approach towards relative stability. In order to identify and explain this alternative and rather dominating second approach, let us consider the simple first order differential equation

$$y' = -Ky \qquad \text{with} \qquad y(x_0) = y_0 \tag{4.60}$$

with the true solution being

$$Y = y_0 \exp(-K(x - x_0)) \tag{4.61}$$

The general integration formula may be written as

$$y_{n+1}(1 + hKb_{-1}) = \sum_{i=0}^{p} (a_i - hKb_i) y_{n-i} \tag{4.62}$$

where p is the order of the integration scheme, h is the step size, and a_i and b_i ($b_{-1} \neq 0$ for implicit schemes) are integration constants depending on the particular

scheme employed (Ralston and Rabinowitz, 1978). The solution of the linear equation (4.60) may be expressed as

$$y_n = \sum_{i=0}^{p} c_i r_i^n \tag{4.63}$$

where the c_i's are constants and the r_i's are obtained by solving the equation

$$(1 + hKb_{-1})r^{p+1} = \sum_{i=0}^{p} (a_i - hKb_i)r^{p-i} \tag{4.64}$$

It may be shown that $c_0 \to y_0$ and $r_0^k \to \exp(-K(x_k - x_0))$ as $h \to 0$. Thus, the first term in (4.63) approaches the true solution as $h \to 0$. The rest of the r_i, $i = 1, \ldots, p$ are classified as *parasitic roots* (being present because the order of the difference equation is $p+1$ as opposed to 1 for the differential equation) and they should satisfy $|r_i| \leq |r_0|$ for all $i > 1$, so that they would not interfere destructively in the construction of the approximate solution. An analogous analysis may be carried out for the error components of the solution.

Now, if the solution is an increasing exponential, we cannot hope to keep the error bounded since the exact solution itself is not bounded. However, we need to ensure that the error stays small relative to the true solution, which means that the parasitic solutions should remain small with respect to the *non-parasitic* or *principal* solution and therefore errors will be magnified at a lower level than the true solution; we may then classify the method as being stable.

All the above point to the following definition of stability (Ralston and Rabiniwitz, 1978). A method is said to be absolutely stable on an interval $[\gamma, \delta]$ if for all hK in this interval $|r_i| < 1$ $i = 0, \ldots, p$, whereas this method is classified as relatively stable on

(in general) another interval $[\alpha, \beta]$ if for all hK in that interval $|r_i/r_0| \leq 1$ $i = 1, \ldots, p$ and if when $|r_i| = |r_0|$, r_i is a simple root.

If the solution is ever-decreasing in magnitude, absolute stability is meaningful only if all the parasitic solutions decrease in magnitude as well. Gear (1971) emphasizes that for the majority of the schemes and the problems being tackled, accuracy limitations (accurate approximation of $\exp(hK)$) overwhelm relative stability restrictions (parasitic roots smaller than the principal root). Furthermore, weakly stable or nearly weakly stable schemes (one or more parasitic roots lying at the vicinity of the unit circle for $hK = 0$) are likely to suffer from relative stability problems, while strongly stable schemes (all roots clearly well inside the unit circle) do not tend to exhibit such kind of problems. Although very enlightening, this second approach to the relative stability question is paradoxical, in the sense that it does not consider relative stability for the *single step* recursions present in our case, where there is only the principal root r_0 contributing in the numerical approximation.

The latter stability definition was presented for the case of a single scalar equation. For a system of equations though, r_0 is not a scalar anymore, but an eigenfunction matrix instead. A comparison of this definition with the restrictions (4.58), reveals a high degree of resemblance and we may conclude that, since we have to cope with an ever-decreasing solution, whose dependence on the first eigenfrequency increases dramatically with time, we may extend the notion of relative stability in the same fashion to single step recursions as well. In other words, we could view the higher eigenfrequencies as being of the parasitic kind and impose the requirement that they stay smaller from the principal mode at all times (compare with 4.58). Alternatively, it may be that the second approach does not consider the issue as one of a relative

stability nature, either because knowledge of the exact solution is not available a priori in general, or because the problem is classified as regarding accuracy, instead.

Before concluding this discussion on accuracy and stability, we should point out that the diffusion equation involves a natural damping (smoothing out irregularities) and therefore, not only the solution bur the errors themselves are susceptible to that. The effect of the errors may also be visualized from another perspective. They are virtually perturbing the original equation, augmenting it with an artificial heat source term, that is to say, our diffusion equation becomes inhomogeneous as "heat" is fed into the system superficially. These errors would decay as time progresses for an absolutely stable approximation, but they should have a faster relative decay rate with respect to the solution, if we are to retain our original relative accuracy in the course of time.

4.4.3 Additional numerical considerations

Numerical tests are made for $t = 1$ following the details of the test example of Hussaini et al (1983). Although computations for the Helmholtz equation (chapter III) are done in single-precision arithmetic, an inspection of the true solution at $t = 1$ reveals the inadequacy of such an accuracy level for our diffusion model computations. A major hurdle is the apparent magnification of the error at the zero crossings in the numerical solution. Stepping in time is accompanied by two negative effects insofar as accuracy is concerned. The magnitude of the solution itself is reduced and concurrently, cumulative round-off causes a drifting of the exact zero crossing away from zero. Consequently, after the time of interest for our problem has elapsed, numerical results become meaningless as numerical resolution of fundamental aspects of the solution has been lost.

This zero crossing phenomenon has more severe implications in the case of the Chebychev expansions. The two following factors are responsible for such a phenomenon. The first is associated with the fact that the endpoints are "frozen" in the finite difference system, whereas they have been included in the Chebychev system (so that the system is more efficient and better conditioned — see discussion in 3.3.2.2 and 3.3.3.2). Second, the boundary points are not amenable to direct manipulation, since their significance has been transferred to the spectrum through the transformation in the Chebychev space. The transform routine tends to operate as a "noise" generator, in the sense that it diminishes the resolution of the physical space and, therefore, it weakens the power of whatever protective precautions are taken in the physical space. Furthermore, re-imposition of the boundary conditions requires an inverse transform; continuation of the procedure demands a forward transform to recover the Chebychev spectrum at that time instant and thereby, efficiency is reduced and further (but rather minor) round-off's are experienced.

Experiments show a dramatic improvement when imposition of the boundary conditions is applied at every time step of the Chebychev system recursion. The presence of the interior zero crossing of the solution is rather intractable though. Either system (finite differences, Chebychev) exhibited a significant improvement when the interior zero crossing was kept satisfactorily close to zero, but, of course, such an approach is useless because it requires a priori information regarding details of the time evolution of the exact solution. A more efficient (but rather elementary) filtering based on a "minimum threshold" principle (either in the physical or the transform space) fails to improve the accuracy levels, indicating the need for a deeper understanding of

the problem and the ultimate dependence of such an approach on the specific initial condition being considered.

Obviously, shifting to double-precision arithmetic is a viable alternative. Nonetheless, the algorithms are doomed to suffer from exactly the same problems at larger times (employing double-precision merely procrastinates the occurr ence of the resolution loss); therefore, more sophisticated techniques would need to be devised in the future to account for a more efficient and productive handling of such problems; a detailed study of filtering in Chebychev space should be carried out but this beyond the scope of this thesis. Confirmation of the previous speculative analysis on the accuracy deterioration, associated with the pollution of the numerical algorithms due to such ill-conditioned intrinsic characteristics of the exact solution, has been obtained by comparing the results of our problem with results obtained for the initial condition $\sin\big((x+1)\pi/2\big)$, which does not have a zero crossing at $x = 0$; $\sin\big((x+1)\pi/2\big)$ has only half its period in $[-1,+1]$, while $\sin(\pi x)$ unfolds its full period in that same interval.

4.4.4 The finite difference Crank-Nicolson scheme

Let us now present some error estimates from our numerical experiments with the Crank-Nicolson finite-difference scheme. Tables (4.1-2) are concerned with the relative and the absolute L_∞ estimates for $u(x,0) = \sin \pi x$ at $t = 1$, respectively. Two different step sizes are considered, involving both constant and variable ($a = f(N)$) values for the parameter a.

The absolute stability of the scheme allows a fast time integration pace to be employed. Although fast integration is associated with an enhanced local temporal error, it also means that fewer time steps are needed to arrive at the desired time instant and

$N+1 \setminus \Delta t$	$1/N$ $(a = N/8)$	$1/N^2$ $(a = 1/8)$
5	0.10 (-1)	0.45 (+1)
9	0.58 (0)	0.62 (0)
17	0.18 (0)	0.13 (0)
33	0.46 (-1)	0.32 (-1)
65	0.12 (-1)	0.80 (-2)
129	0.29 (-2)	0.20 (-2)

Table 4.1 $\bar{L}_\infty$ values for the Crank-Nicolson solution of the heat equation with $u(x,0) = \sin \pi x$, at $t = 1$.

$N+1 \setminus \Delta t$	$1/N$ $(a = N/8)$	$1/N^2$ $(a = 1/8)$
5	0.52 (-4)	0.23 (-3)
9	0.30 (-4)	0.32 (-4)
17	0.91 (-5)	0.69 (-5)
33	0.24 (-5)	0.17 (-5)
65	0.60 (-6)	0.41 (-6)
129	0.15 (-6)	0.10 (-6)

Table 4.2 L_∞ values for the Crank-Nicolson solution of the heat equation with $u(x,0) = \sin \pi x$, at $t = 1$.

consequently, arithmetic error is reduced. This is clearly depicted in tables (4.3-4), which contain a description of these errors as time increases from $t = 0$ to $t = 1$; the given estimates are for $N = 16$, $\Delta t = 1/16$ $(a = 2)$ and $\Delta t = 1/256$ $(a = 1/8)$.

time steps	L_∞	$\bar{L}_\infty$
1	0.65 (-2)	0.12 (-1)
5	0.27 (-2)	0.58 (-1)
10	0.24 (-3)	0.11 (0)
17	0.91 (-5)	0.18 (0)

Table 4.3 L_∞ and $\bar{L}_\infty$ values for the Crank-Nicolson solution of the heat equation with $u(x,0) = \sin \pi x$ for N=16 and from $t = 0$ to $t = 1$; time advancing is performed in steps of $\Delta t = 1/16$.

time steps	L_∞	$\bar{L}_\infty$
1	0.47 (-3)	0.49 (-3)
50	0.36 (-2)	0.25 (-1)
150	0.23 (-3)	0.76 (-1)
256	0.69 (-5)	0.13 (0)

Table 4.4 L_∞ and $\bar{L}_\infty$ values for the Crank-Nicolson solution of the heat equation with $u(x,0) = \sin \pi x$ for N=16 and from $t = 0$ to $t = 1$; time advancing is performed in steps of $\Delta t = 1/256$.

Despite the fact that the recursion enjoys absolutely stability for both choices of Δt, either of those leads to relative instability (or relative inaccuracy), since the percentage error between the computed and the exact solution increases (linearly) with the number of time steps.

A similar behavior of the errors can be identified from the results given in table (4.5). These refer to the Crank-Nicolson solution of the heat equation with $u(x,0) = \sin x$ in $[0, \pi]$ and with exact solution $u(x,t) = \exp(-t) \sin x$. The number of x samples $N = 20$ and the parameter $a = 1/2\sqrt{20}$ were chosen as in Mitchell and Griffiths (1980).

time steps	$\bar{L}_\infty$	L_∞
1	0.11 (-4)	0.11 (-4)
2	0.22 (-4)	0.23 (-4)
4	0.44 (-4)	0.45 (-4)
8	0.87 (-4)	0.91 (-4)
16	0.17 (-3)	0.18 (-3)
80	0.58 (-3)	0.91 (-3)
160	0.75 (-3)	0.18 (-2)
320	0.62 (-3)	0.36 (-2)
640	0.21 (-3)	0.73 (-2)
800	0.11 (-3)	0.91 (-2)
1500	0.44 (-5)	0.17 (-1)
2000	0.37 (-6)	0.23 (-1)
3000	0.22 (-8)	0.35 (-1)

Table 4.5 L_∞ and $\bar{L}_\infty$ for the Crank-Nicolson solution of the heat equation with $u(x,0) = \sin x$ for different numbers of time steps ($N = 20$).

These results — virtually identical to theirs — demonstrate similar trends in the attitude of the errors. In the beginning both errors start to increase with time, but at a later stage the absolute error commences to descend as anticipated. A similar situation is not observed in the case of the percentage error, which appears to increase linearly and without having a finite bound.

4.4.5 The Chebychev Crank-Nicolson scheme

We now proceed with the performance of the Chebychev solutions to the diffusion equation. The Crank-Nicolson scheme has been chosen, over the backwards-Euler one, to simulate the time advancing of the solution, due to its reduced temporal error. The absolute stability having been ensured, we need only worry about the accuracy. A stepsize $\Delta t = 1/N^2$ has also been chosen; it is considerably larger that the explicit schemes' barrier and it approximately corresponds to the $(\Delta x)_{\min}$ of the Chebychev sampling.

$N+1$	*GAL*	*PSD*	*TAU*	*TIN*	*TIC*
5	0.94 (0)	0.94 (0)	0.10 (+1)	0.10 (+1)	0.10 (+1)
9	0.20 (-1)	0.20 (-1)	0.21 (-1)	0.21 (-1)	0.21 (-1)
17	0.12 (-2)	0.12 (-2)	0.12 (-2)	0.12 (-2)	0.12 (-2)
33	0.76 (-4)	0.76 (-4)	0.76 (-4)	0.76 (-4)	0.76 (-4)

Table 4.6 $\bar{L}_\infty$ for the Chebychev Crank-Nicolson solution of the heat equation with $u(x,0) = \sin \pi x$ from $t = 0$ to $t = 1$. The results listed refer to the Galerkin, pseudospectral, and differentiated, non-centred integrated and centred integrated tau systems, respectively. Extrapolation has provided error estimates for the systems with $N = 64, 128$; these are 0.50 (-5) and 0.30 (-6), respectively.

All three projections have been tested; tables (4.6-7) present $\bar{L}_\infty$ and L_∞ estimates for the initial condition $u(x, t = 0) = \sin \pi x$ and at $t = 1$. The linear systems to be inverted at each time step are not tridiagonal, as in the case of the finite differences and therefore, general inversion routines have to be employed. The results given have been obtained by applying an LU decomposition with partial pivoting. Improvement of the solutions using iterative improvement have not produced answers of better quality.

4.4.5.1 Conditioning and inversion of the propagation matrix

The conditioning of the Chebychev Crank-Nicolson matrices depends in a non-trivial fashion on both the values of σ and Δt used in a particular application.

$N+1$	*GAL*	*PSD*	*TAU*	*TIN*	*TIC*
5	0.39 (-4)	0.39 (-4)	0.41 (-4)	0.41 (-4)	0.41 (-4)
9	0.96 (-6)	0.96 (-4)	0.10 (-5)	0.10 (-5)	0.10 (-5)
17	0.62 (-7)	0.62 (-7)	0.62 (-7)	0.62 (-7)	0.62 (-7)
33	0.39 (-8)	0.39 (-8)	0.39 (-8)	0.39 (-8)	0.39 (-8)

Table 4.7 L_∞ for the Chebychev Crank-Nicolson solution of the heat equation with $u(x, 0) = \sin \pi x$ from $t = 0$ to $t = 1$. The results listed refer to the Galerkin, pseudospectral, and differentiated, non-centred integrated and centred integrated tau systems, respectively.

A condition number investigation (for the parameters of our current problem) nullified our suspicions with regard to the quality of their conditioning (see table 4.10); the tau systems are worse conditioned than the Galerkin and the pseudospectral matrices and between them, the integrated system exhibits definetely a looser structure. Partial pivoting, at least in theory, might fail even for well-conditioned matrices and it is important to understand that even a moderate failure could possibly lead to increased round-off over a large number of repetitions. The loss of relative accuracy, witnessed in the finite difference's evolution of the numerical solution and evident from a simple

comparison of the tables (4.8-9) for the Chebychev solutions, may be associated with this problem.

$N+1$	*GAL*	*PSD*	*TAU*	*TIN*	*TIC*
5	0.16 (0)	0.16 (0)	0.74 (0)	0.74 (0)	0.74 (0)
9	0.33 (-3)	0.33 (-3)	0.86 (-3)	0.86 (-3)	0.86 (-3)
17	0.48 (-5)	0.48 (-5)	0.48 (-5)	0.48 (-5)	0.48 (-5)
33	0.75 (-7)	0.75 (-7)	0.75 (-7)	0.75 (-7)	0.75 (-7)

Table 4.8 $\bar{L}_\infty$ for the Chebychev Crank-Nicolson solution of the heat equation with $u(x,0) = \sin \pi x$ after one time step. The results listed refer to the Galerkin, pseudospectral, and differentiated, non-centred integrated and centred integrated tau systems, respectively. Extrapolation has provided estimates for $N = 64, 128$; these are 0.12 (-8) and 0.18 (-10), respectively.

Eradication of our qualms demands the employment of the complete (total) pivoting strategy, but since the numerical tests do not indicate the presence of such an exceptional situation, the issue has been approached via less computationally intensive techniques. Preprocessing of the system has been considered. The *Crout LU* algorithm is a candidate due to its *implicit pivoting* strategy: the pivot element is chosen as if the entries of the matrix had been scaled to a maximum of unity in each row (Press et al, 1985). Instead, *row balancing* has been attempted; this amounts to a simplified version of the more general *row-column equilibrium*. The matrix $D = \mathrm{diag}(\beta^{r_1}, \cdots, \beta^{r_n})$ is computed appropiately, in order to produce a matrix $D^{-1}A$, which has roughly the same L_∞ norm for each of its rows (Golub and van Loan, 1983). The quantity β is the float-point arithmetic radix base of the machine used and thereafter, by choosing the scale factors only among the integer powers of β, no round-off errors are committed during the balancing process. Significant improvement of the conditioning may be

$N+1$	*GAL*	*PSD*	*TAU*	*TIN*	*TIC*
5	0.68 (-1)	0.68 (-1)	0.32 (0)	0.32 (0)	0.32 (0)
9	0.27 (-3)	0.27 (-3)	0.69 (-3)	0.69 (-3)	0.69 (-3)
17	0.45 (-5)	0.45 (-5)	0.45 (-5)	0.45 (-5)	0.45 (-5)
33	0.74 (-7)	0.74 (-7)	0.74 (-7)	0.74 (-7)	0.74 (-7)

Table 4.9 L_∞ for the Chebychev Crank-Nicolson solution of the heat equation with $u(x,0) = \sin \pi x$ after one time step. The results listed refer to the Galerkin, pseudospectral, and differentiated, non-centred integrated and centred integrated tau systems, respectively. The 65 and 129-long systems exhibit errors of 0.12 (-8) and 0.18 (-10) in value, respectively.

$N+1$	*GAL*	*PSD*	*TAU*	*TIN'S*
5	0.23 (+1)	0.18 (+1)	0.45 (+1)	0.45 (+1)
9	0.50 (+1)	0.39 (+1)	0.17 (+2)	0.46 (+3)
17	0.16 (+2)	0.14 (+2)	0.79 (+2)	0.28 (+4)
33	0.63 (+2)	0.59 (+2)	0.45 (+3)	0.19 (+5)
65	0.28 (+3)	0.27 (+3)	0.28 (+4)	0.13 (+6)
129	0.13 (+4)	0.13 (+4)	0.18 (+5)	0.85 (+6)

Table 4.10 Condition numbers for the Chebychev Crank-Nicolson systems with σ=1 and $\Delta t = 1/N^2$; the condition numbers given above refer to the Galerkin, pseudospectral, and differentiated, and integrated tau systems.

achieved and even though, this refers to a specific norm, comparable decrease of the condition number in any other norm should be anticipated (Press et al, 1985).

As a result, the probability of accuracy diminution, due to the addition of numbers that vary widely in magnitude during the elimination process, is greatly reduced. In addition to row balancing, the QR algorithm (with implicit shifts) has also been used to invert the propagation matrices.

Neither of the techniques mentioned above has succeeded in improving the accuracy, but the likelihood of witnessing such a stabilizing effect would tend to increase with the dimension of the spatial discretization. Furthermore, for such occurrences, it might be advisable to employ an (SVD) decomposition augmented with a threshold cut-off for the "noisy" singular values.

4.4.5.2 Analysis of results and comparison with finite differences

Let us now compare the performance of the Chebychev tecniques with finite differences; a variety of interesting points should be highlighted. The relative simplicity of the analytical problem and the well-posedness of its numerical counterpart under current consideration, defaults to a virtual equivalence between the Galerkin and the pseudospectral methods; the lack of ill-conditioning makes all the variants of the tau projection performing at the same level as well. Furthermore, as sampling becomes denser, the tau results improve rapidly, converging to the originally superior Galerkin (pseudospectral) approximations.

The integrated tau system's implementation demands a closer look. Although the boundary conditions should be imposed in a "centred" fashion according to equations (4.50), other investigators prefer a "non-centred" version, that is, equations (4.49) (Kim and Moin, 1980; Orszag and Kells, 1980). This latter choice is not consistent with the rest of the Crank-Nicolson system, but paradoxically, it amounts to a mere computational trick for minimizing round-off in the computation of the right-hand side vector. Again, no difference in the results is observed but this will not be so, if conditioning problems and incomplete boundary simulation are present; accumulation of the inner products (in updating the right-hand side vector) in an extended precision is another reason. The latter detail may be of fundamental importance, as a repetitive inaccurate accumulation of these inner products could weaken the imposition of the boundary conditions (primarily) to a significant extent and lead to woefully incorrect answers.

Some useful indications on the performance of all the different techniques, in the presence of unavoidable pitfalls, may be drawn by looking back at the troublesome

single precision solutions of the given problem. There, the Galerkin method enjoys an early relative superiority over all the other techniques, but as the dimension of the problem increases, the breakdown of the highly sensitive Galerkin and pseudospectral systems occurs faster than for their tau counterparts. Among the latter ones, the resistance to deterioration is more enhanced in the "non-centred" systems. The "centred" integrated and differentiated systems follow in this order. A final comment, here, concerns the superiority of the (pure differential) differentiated system (4.18-20) over its mixed (4.11-12) predecessor; the former may endure slight boundary imposition problems with relative success, whereas the latter rather dissolves.

$N+1$	*GAL*	*PSD*	*TAU'S*
9	0.24 (-3)	0.46 (-3)	0.16 (-2)
17	0.11 (-10)	0.21 (-10)	0.84 (-10)

Table 4.11 $\bar{L}_\infty$ values for the Chebychev Runge-Kutta solution of our model problem (after Hussaini et al, 1983).

$N+1$	*GAL*	*PSD*	*TAU'S*
9	0.26 (-1)	0.26 (-1)	0.31 (-0)
17	0.19 (-7)	0.19 (-7)	0.19 (-7)

Table 4.12 $\bar{L}_\infty$ values for the Chebychev Crank-Nicolson solution with $\Delta t = 1/N^4$ for our model problem; extrapolation from the errors after just one time step has again been employed.

Overall though, the Chebychev systems enjoy a great superiority over their finite difference rivals. They are able of attaining a moderate accuracy level by employing orders of magnitude fewer points than finite differences do; alternatively, they may be pushed into achieving extrere accuracy by dense sampling rates.

If the length of the calculations is not long, explicit schemes with high-order truncation error and, most importantly, provision for absolute and relative error control (accounting to an implicit control of the truncation error to predetermined tolerance levels) is the best choice, i.e fourth-order Runge-Kutta or high-order predictor-corrector methods. This may be manifested by considering the results given in table 4.11 (obtained via a fourth-order Runge-Kutta integration of the Chebychev systems) and comparing them with the error estimates displayed in table 4.12, which correspond to a Crank-Nicolson time integration with a time increment $\Delta t = 1/N^4$. This specific choice reflects the upper bound on Δt for explicit integrations (Runge-Kutta) and although it is absolutely meaningless in an "implicit" environment, it allows a comparison of the two temporal integrators to be carried out under a uniform time increment assumption. As anticipated, these Crank-Nicolson results are superior to the ones reported earlier, for $\Delta t = 1/N^2$. Despite this significant improvement, the Crank-Nicolson performance is definitely inferior to Runge-Kutta's. Furthermore, the special characteristics of the various Chebychev systems appear to be resolved better in a Runge-Kutta integration; the overlap of the Crank-Nicolson results is not witnessed there. What are the sources of these phenomena?

First, we ought to realize that the reported Runge-Kutta scheme has a temporal truncation error two orders of magnitude higher than the Crank-Nicolson's. Second, the former technique is "privileged" in the sense that it incorporates an adaptive time step size control, in order to maintain a predetermined absolute-relative accuracy level; on the contrary, a direct march in time was performed for the Crank-Nicolson systems. Therefore, the Runge-Kutta driving mechanism does not allow the accumulation of the

repetition of truncation errors with time. Finally, we should not forget that the Crank-Nicolson scheme amounts to a low-frequency approximation of the time propagator and therefore, a proper handling of the high-frequencies is neglected. The Runge-Kutta technique, on the other hand, propagates these higher frequencies more accurately, provided that the restrictions on the size of Δt are met; indeed, it is exactly this high-frequency part of the spectrum that would devastate the numerical approximation if the restrictions on Δt fail to be met. This last point refers to the characteristic feature of the implicit methods: they aim at producing the correct equilibrium solution and not at providing answers of high resolution.

The issue of resolution among the different Chebychev systems themselves is now to be discussed. The nature of the Runge-Kutta integrator allows the high-frequency superiority of the Galerkin and the pseudospectral methods to be reflected at their performance, compared to the tau method; incidentally, the Galerkin results are marginally better. However, as N increases and the spatial information gets almost saturated at $N = 16$ (see 4.4.5.3), we observe that the tau results merge into the level of the rest, even in the Runge-Kutta case.

Nevertheless, ambitions for answers of very high accuracy may be prohibited, due to time differencing obstacles when large time scale computations are considered. Implicit schemes manage to shrink these limitations to some extent at least, as long as the dimension of the problems remains moderate. In agreement with the previous observations, Hussaini et al (1983) propose the use of the Chebychev integrated Crank-Nicolson system for large time scale computations, when a relative accuracy of 10^{-3} is considered satisfactory (Hussaini et al, 1983).

Our results too point in the understanding that the most important feature of the Chebychev solutions is their ability of achieving moderate accuracy with much fewer spatial points. The Chebychev time step is much smaller than in finite differences (even with implicit schemes); nevertheless,the trade-off, subjective as it may be, is in favor of the Chebychev systems, especially as the spatial dimensionality of the problems increases.

The previous analysis reflects the case of $\sigma = 1$. How would it be affected by an increased σ? It is important to understand that either stability or accuracy considerations couple σ and Δt; as a result, it is the product of these, namely the quantity $\sigma\Delta t$, that controls their performance. Failure to satisfy the requirement of Δt not being larger than $O(1/\sigma N^4)$ would default to a destabilization of explicit integration schemes. On the contrary, implicit schemes will not suffer from instabilities, but inaccurate results would be obtained if accuracy considerations of the physical problem were not to be obeyed. Integration with either scheme (provided that the problem is treated correctly) should be pronouncing the superiority of the Galerkin and the pseudospectral methods over the tau systems; evidently, the higher resolution of the former techniques should be depicted better in the Runge-Kutta than in the Crank-Nicolson integration. In addition, properly implemented Chebychev systems would be exhibiting an enhanced superiority over the finite differences, as the magnitude of the diffusitivity parameter σ of the problem increases.

4.4.5.3 Analysis of the Chebychev spectrum of $\sin(\pi x)$.

We would now like to focus our attention back to the results given in table 4.6 and study them in association with the Chebychev spectrum of the initial condition $u(x, t = 0) = \sin(\pi x)$.

Concurrently, we would also like to identify how faithful the discrete Chebychev spectrum is to its true counterpart. This comprises an exceedingly important aspect of the effect of the discretization on the proper representation of the original information. Fortunately, the Chebychev transform of $\sin(\pi x)$ can be evaluated analytically and thereafter to machine-accuracy precision level. The derivation has been carried out in Appendix A.5; the magnitude of each spectral component according to (A.55) is given as

$$a_n = 2 \sum_{k=0}^{\infty} (-1)^k J_{2k+1}(\pi) \delta_{n,2k+1} \tag{4.65}$$

A mere inspection of this expression reflects the odd-character of the spectrum (the $\sin(\pi x)$ having odd parity in $[-1, +1]$), since the δ-function *ignites* for odd values of n, only. The absolute value of the a_n's (4.65) is plotted in figure (4.1a); the specrtum peaks for a Chebychev pseudo-wavenumber in the neighborhood of π. The latter marks the onset of a dramatic convergence, obvious in the corresponding logarithmic plot (figure 4.1b).

Table 4.13 displays information regarding the convergence of the Chebychev series; successive truncation of the true spectrum has been applied and the corresponding errors have been recorded.

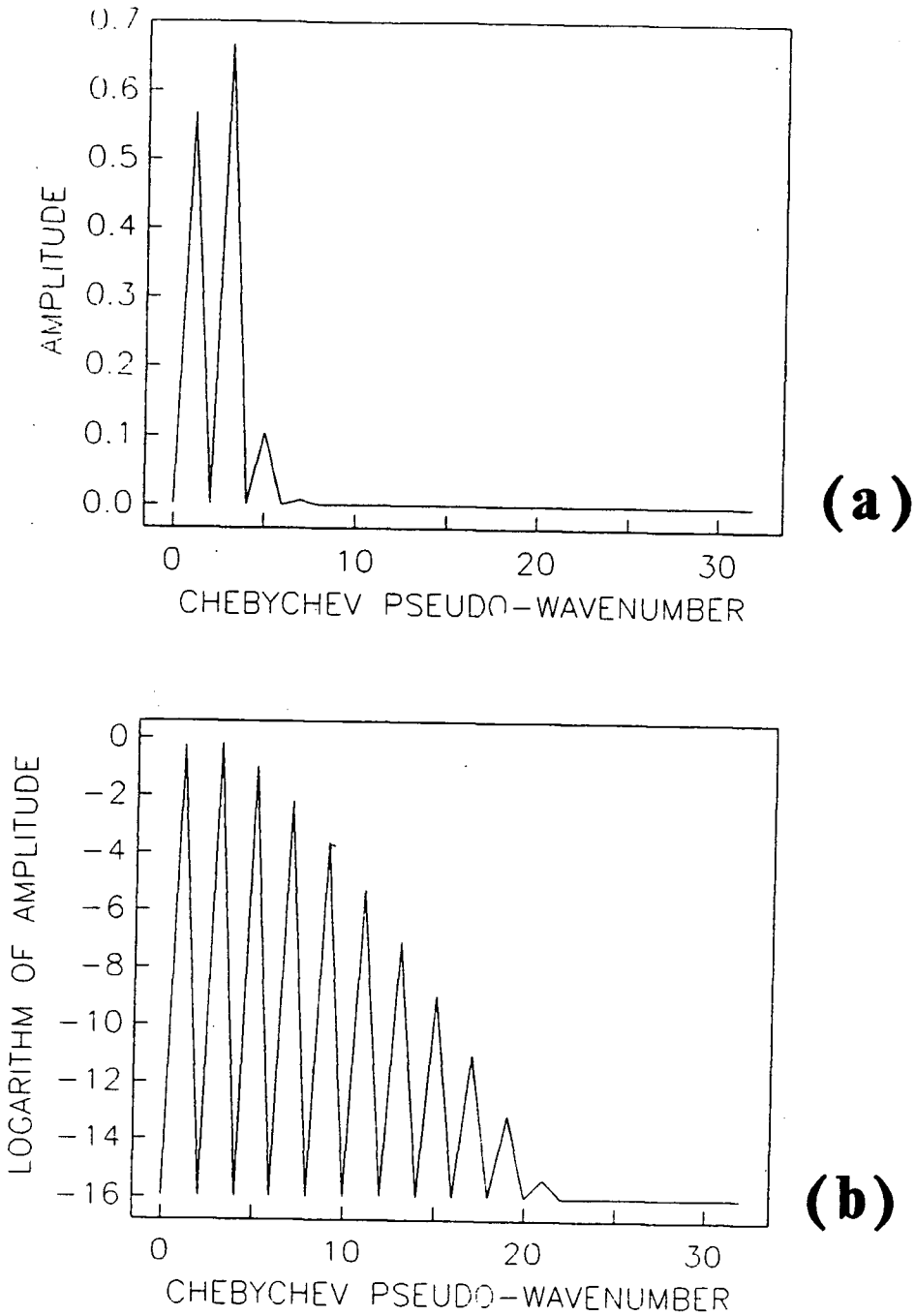

Figure 4.1 Amplitude spectum of the true Chebychev spectrum of $\sin(\pi x)$ in a (a) linear and (b) logaritmic scale.

Reconstruction of the sine is readily obtained as

$$\sin(\pi x) = 2 \sum_{n=0}^{\infty} \frac{1}{c_n} J_n(\pi) \sin\left(\frac{n\pi}{2}\right) T_n(x) \tag{4.66}$$

and table 4.14 contains error estimates concerning the quality of the reconstruction of the sine, employing only a certain portion of the true spectrum.

$N+1$	$\bar{L}_2$	$\bar{L}_\infty$
2	0.58 (0)	0.10 (+1)
4	0.14 (-1)	0.16 (0)
8	0.80 (-7)	0.25 (-3)
16	0.14 (-21)	0.16 (-10)
20	0.27 (-30)	0.69 (-15)
22	0.66 (-35)	0.34 (-17)

Table 4.13 $\bar{L}_2$ and $\bar{L}_\infty$ values for the true Chebychev spectrum of $\sin \pi x$, truncated to various cut-off levels.

$N+1$	$\bar{L}_2$	$\bar{L}_\infty$
2	0.61 (+1)	0.73 (+1)
4	0.15 (-1)	0.11 (0)
8	0.85 (-7)	0.25 (-3)
16	0.15 (-21)	0.11 (-10)
20	0.10 (-29)	0.14 (-14)
22	0.71 (-30)	0.10 (-14)

Table 4.14 $\bar{L}_2$ and $\bar{L}_\infty$ values for the true Chebychev series reconstruction of $\sin \pi x$, truncated to various cut-off levels.

The $\bar{L}_2$ behavior of table 4.14 verifies the *heuristic* criterion for the resolution requirements of Chebychev expansions, i.e "good" convergence demands π polynomials per wavelength, at the very least. Indeed, significant reduction of the reported error commences at $N_{\text{cut-off}} = 4$, which amounts to retaining 4 (2 non-zero) modes ($\sin \pi x$ has just one complete wavelength in the fundamental interval $[-1, +1]$). This,

alternatively, may be seen in table 4.13, where the relative energy discrepancies start exhibiting a dramatic drop-off at the same cut-off level.

Important as it is, the previous heuristic cannot provide a rigorous, quantitative interpretation of the rather vague classification *good*. The criterion has been established on the basis of the energy distribution in the spectrum and it might be misleading, when augmented with an analysis concerning pre-specified resolution (accuracy) ambitions. That this may be the case is evident from the $\bar{L}_{\infty}$ estimates in both tables. Athough substantial reduction of the inaccuracy level begins at $N_{\text{cut-off}} = 4$, it is the $N_{\text{cut-off}} = 8$, which witnesses an unquestionable diminution of the incomplete reconstruction's errors; still, it is obvious that high-resolution information is stretched up to the 21-st coefficient (for double precision computations). Everything, beyond this spectral boundary, floats in a noise swamp and provides no more information.

The previous discussion on the effect of the truncation of the true spectrum will facilitate the investigation of the quality of the discrete spectrum; we start by reminding ourselves of the most familiar discrete Fourier spectrum limitations. The only class of functions, whose discrete Fourier transform qualifies as an excellent simulator of the analytic integral, must satisfy all the following conditions:

1) The function is periodic;

2) The function is band-limited;

3) The sampling interval does not violate the Nyquist sampling criterion, i.e two points per cycle;

4) The truncation of the infinite time (space) function is performed exactly at a multiple of its period (Brigham, 1974).

Nevertheless, the results are "perfect" to the order of the accuracy of the numerical integration and limited to the round-off level.

The computation of the discrete Chebychev transform demands a discrete point-set as well. Although a quadrature on an equidistant point-set is usually preferred over other more elaborate and fancier non-equidistant quadrature schemes (for example the *DFT* algorithm is based on the *extended trapezoidal rule*), the Chebychev transform has been traditionally evaluated on specific non-equidistant point-sets for good reasons. These are associated with a dramatic acceleration in the speed of the relevant computations; the point-set $x_i = \cos\theta_i$ for θ_i in $[0, \pi]$ reduces the transform to a *cosine* one and it allows the use of the *FFT* algorithm, thereby introducing an overwhelming efficiency improvement (see Appendix B). Sampling in an non-equidistant fashion makes the aliasing analysis rather obscure, but on the other hand, a direct equidistant quadrature in x-space for the calculation of the integrals

$$a_n = \frac{2}{\pi c_n} \int_0^{\pi} \frac{f(x)T_n(x)}{\sqrt{1-x^2}}\, dx \tag{4.67}$$

is considered disadvantageous and therefore the issue is dropped.

The assumption of non-equidistant sampling may possibly lead into undesirable circumstances, in the sense that, depending on the particular function to be sampled, it may constitute a *bad* choice, gleaning information of inferior importance. On the other hand, practical applications usually involve initial conditions drawn from a sequence of points that most likely do not obey a nice, analytic formulation; thus, it is immediately understood that any further analysis in this direction, in a general framework, is fruitless, and it may be of some help only in a problem to problem basis.

Given the fact that the special *fast* point-set is employed, a cosine transform

$$a_n = \frac{2}{\pi c_n} \int_0^{\pi} f(\theta) \cos(n\theta)\, d\theta \tag{4.68}$$

is needed; traditionally, the extended trapezoidal rule has been employed on equidistant θ_i nodes in $[0, \pi]$. This is by no means mandatory, but its simplicity and its relative robustness have made its use almost exclusive; furthermore, it is amenable to the fast implementation of the FFT algorithm. Furthermore, fancier higher-order methods would tend to guarantee an improved accuracy of the integration, only for adequately smooth integrands; in addition, a desired accuracy may be readily controlled by monitoring of the relative change of the results between successive halving of the sampling rate.

The disadvantages of the trapezoidal quadrature are well-known: it is only second-order accurate and the improvement of its truncation error—being accomplished by finer sampling—is counterbalanced by an enhanced negative influence of round-off errors, the latter being inversely proportional to the mesh-size, i.e $O(1/\Delta x)$ (Mc Calla, 1967). Additionally, the trapezoidal approximation (expressed in terms of the *Euler-Maclaurin summation formula*) is only an asymptotic expansion—as opposed to a convergent one—with an error, that is always less than twice the amplitude of the first ommitted term in a certain truncation.

The cosine transform trapezoidal quadrature brings back the aliasing analysis of periodic functions that are equally sampled. *Aliasing* is the direct manifestation of the discretization procedure; the quality of the integration may be easily investigated via

the *aliasing summation formulas* (Lynes, 1984),

$$\bar{C}_r = C_r + \sum_{l=1}^{\infty} C_{lm+r} + C_{lm-r} \qquad \text{for} \quad |r| < m \tag{4.69}$$

where subscripts denote the order of the cosine coefficient, $m+1$ is the period of the discrete transform and $\bar{C}_r$ (C_r) is the discrete (exact) coefficient.

Assuming a convergent expansion, we may see that the error in $\bar{C}_r$ is $O(C_{lm-r})$ for $0 < r < m/2$, whereas higher frequencies are erroneously calculated, not being resolvable in that grid. Consequently, the aliasing contaminations are expected to increase with r and accordingly, the truncation error becomes more profound as the order of the coefficient r to be computed increases.

$N+1$	$\bar{L}_2$	$\bar{L}_\infty$
4	0.14 (-1)	0.16 (0)
8	0.80 (-7)	0.37 (-3)
16	0.14 (-21)	0.16 (-10)
32	0.26 (-29)	0.17 (-14)

Table 4.15 $\bar{L}_2$ and $\bar{L}_\infty$ values between the approximate (discrete) and the true (analytic) Chebychev spectral coefficients for various sampling densities.

The rather remarkable similarity between the errors estimates given in table 4.15 and the errors presented in tables (4.13-14) is difficult to neglect; failure to understand though, the fundamental differences in the nature of those quantities, may lead to false interpretations.

We do see, once more, that $N = 4$ gives us a somewhat reasonable approximation, while $N = 8$ appears to provide a fairly satisfactory approximation to those true spectral coefficients. Doubling the number of samples is accompanied with another

significant accuracy enhancement, which persists in up to another halving of the sampling interval, i.e $N = 32$. Finer sampling fails to improve the accuracy of the incorrect non-trivial higher-order coefficients; the observed accuracy improvement terminates at $N = 64$ and it is actually transformed to a deterioration as $N \leq 128$, affecting all the previously correctly computed, lower frequency coefficients as well.

Aliasing is definitely responsible for the incorrect results, obtained from the coarser sampling rates, since as the non-linear mapping $x = \cos\theta$ creates slight problems by forcing a non sinusoidal function to be decomposed in terms of periodic functions, i.e cosines and thereby, extending its apparent frequency content. The non-resolvable frequencies are folded back into the "visible" part of the spectrum; the contamination obeys the classic pattern recognized for periodic functions. The latter has been confirmed through a painstaking, high-resolution, numerical inspection of the various approximate Chebychev spectra; removal of aliasing leaves the coefficients accurate to $O(\Delta\theta^2)$. Aliasing is seen to have already been completely eliminated, at the givenprecision, at $N = 32$; nevertheless, further reduction of the truncation error, anticipated through a successive refinement of the sampling interval, is forced to a stall by *round-off*, since the maximum attainable resolution, offered by the machine at the given double precision level, has already been reached at the reported sampling density.

The last argument is true in an $\bar{L}_\infty$-based interpretation and it is consistent with the spirit of all the error analysis given before. The $\bar{L}_\infty$ norm comprises a pointwise error measure, normalized in a maximum global fashion and therefore tends to neglect local discrepancies that are considered unimportant in the overall comparison; it is this selective character of the $\bar{L}_\infty$ norm that originally weighed heavily in its choice.

Accurate measurement of the truncation error that affects severely the approximation of the higher order Chebychev coefficients, which are several orders of magnitude smaller than the low frequency part of the spectrum, necessitates the introduction of a different norm. This norm may be chosen to normalize the maximium pointwise error at the local level, i.e

$$\bar{L}_{\infty}^{\star} = \frac{\max_{0\leq i\leq N} |y_i - \bar{y}_i|}{|y_i|} \tag{4.70}$$

or alternatively, it could be defined as

$$\bar{L}_{\infty}^{\dagger} = \max_{0\leq i\leq N} \left| \frac{y_i - \bar{y}_i}{y_i} \right| \tag{4.71}$$

Neither one of these norms can provide an adequate description of the effect of the truncation error on the calculation of the high wavenumber Chebychev coefficients; nevertheless, the vast inaccuracies of the spectral section are readily seen by a comparison of the true spectrum (figure 4.1b) and the corresponding *instantaneous* percentage errors (figure 4.2).

Introducing a finer sampling, i.e $N = 64$, does not enhance the quality of the integration, because the truncation error's reduction is being counterbalanced by a respective deterioration of the round-off errors. Finally, the output of the integration quadrature with $N = 128$, reflects the onset of the highly undesirable situation, where the accumulation of round-off error overwhelms any improvement of the truncation error and we are, thereby, confronted with a round-off driving the computational mechanism, accumulating rapidly and soon, defaulting into an ultimate devastation of the algorithm's products.

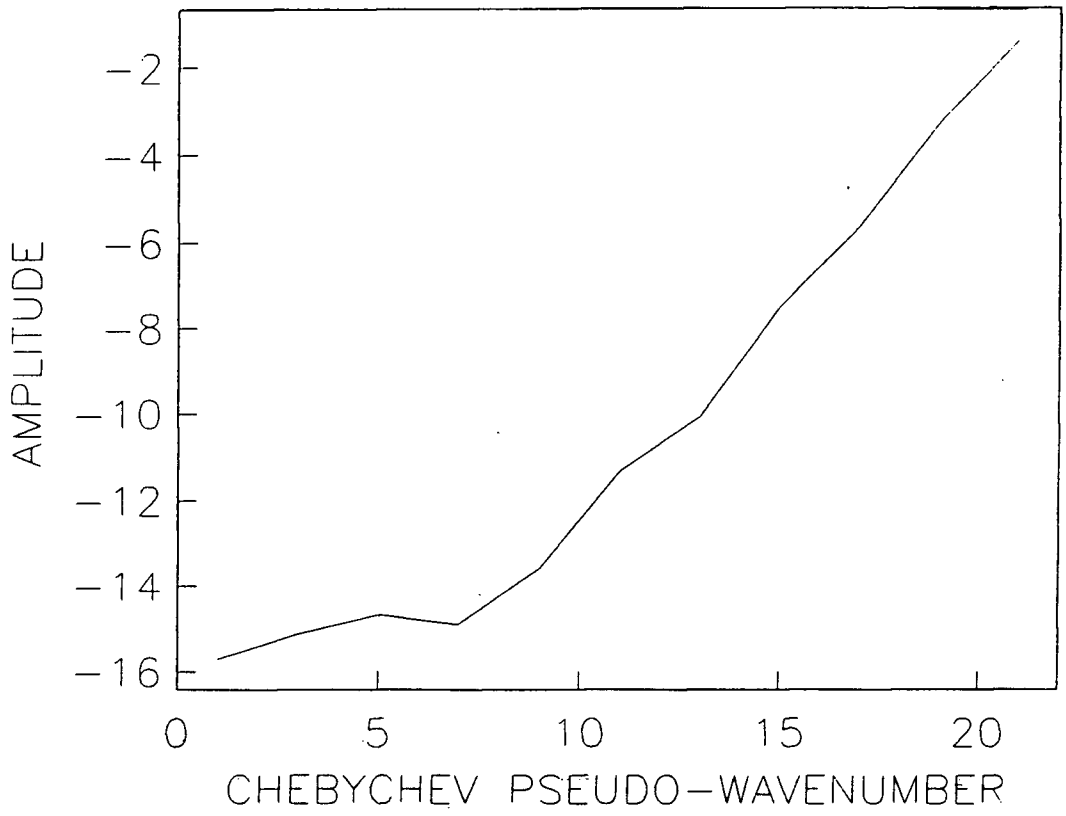

Figure 4.2 Local relative errors in the approximation of the Chebychev spectrum of $\sin(\pi x)$; 33 points has been employed for the numerical approximation. Notice that higher coefficients exhibit enhanced inaccuracies.

The absence of further information in the spectrum beyond a certain Chebychev mode has some interesting implications. Propagation of a spectrum with an extended insignificant part is both fruitless and meaningless; this part of the spectrum simulates an overdeterminancy composed of pure noise. This noise contribution, not adding anything to the solution itself, cultivates the rise of "numerical noise", which will gradually start interfering with the important part of the spectrum; subsequently, corruption of the latter should be anticipated over an adequately large number of time steps. Additionally, the increase in the spectrum order corresponds to a diminution

in the size of Δt to be used in the time integration, so that a desired accuracy level is maintained.

Despite all that, we still observe a continuous improvement of the accuracy levels, as the spatial discretization increases beyond $N = 32$. The answer lies in the size of Δt, which decreases as $1/N^2$ and in the fortunate fact, that the conditioning of the systems remains satisfactory for the present precision level. Consequently, the reduced temporal error due to the smaller Δt, overwhelms the "noise" interference due the trivial spectral extension and the relative degradation of the conditioning of the propagation matrix.

The previous analysis points to another major advantage shared and enjoyed by spectral techniques, namely *filtering*, which is expressed best in Fourier decompositions (the association of Fourier coefficients with temporal and spatial frequencies is most readily grasped from a physical viewpoint). Nevertheless, filtering is common to all spectral methods and it is evident that accuracy depends only on the temporal error, after the spatial decomposition has reached its maximum resolution for a given numerical precision level.

Incidentally, the $\bar{L}_\infty$ values for $N = 16$, after only one time step and for various sizes of Δt, are given in table 4.16; we do observe the positive effect of the improved temporal error on the relative accuracy.

The value $\Delta t = 1/32^2$ corresponds to $\Delta t = 1/N^2$ for $N = 32$; propagation of the 17-long spectrum at a rate of $1/32^2$ scores the same accuracy as the 33-long one does. The slight aliasing and the lower accuracy characterizing the former are, apparently, counterbalanced by the worse conditioning and the noise, present in the latter; this noise interference is not significant and even an 17-long truncation of the 33-long

Δt	$t \rightarrow \Delta t$	$t \rightarrow 1$
$1/N$	0.21 (-1)	0.28 (0)
$1/N^2$	0.48 (-5)	0.12 (-2)
$1/4N^2$	0.75 (-7)	0.76 (-4)
$1/N^4$	0.29 (-12)	0.19 (-7)

Table 4.16 L_∞ values obtained for the 17-long Chebychev Galerkin Crank-Nicolson diffusion system and for different Δt values. Both the errors after one time step and at $t = 1$ are given; the error value for $\Delta t = 1/N^4$ at $t = 1$ has been computed by extrapolating the error value obtained for the first time step.

spectrum gives identical results, as the cumulation of round-off nullifies its marginal superiority.

Nevertheless, the above a *posteriori* analysis of the numerical Chebychev spectrum of the initial condition clearly suggests the computation of a non-aliased spectrum, incorporating a proper balance of truncation error and round-off and its subsequent truncation to the desired order, the order being chosen on the basis of accuracy and efficiency considerations.

4.4.6 Finite difference and Chebychev backwards-Euler schemes

We have concentrated on the Crank-Nicolson formulation of the time integration and presented an analysis of the results obtained by using both finite differences and Chebychev methods in conjuction with it. Let us now show the effect of the smaller temporal truncation error enjoyed by the Crank-Nicolson compared to the backwards-Euler scheme. Tables (4.17-18) present relative and absolute infinity errors corresponding to the solution of the specific diffusion problem for the backwards-Euler scheme;

the four Chebychev systems use $\Delta t = 1/N^2$, whereas the finite difference results have been computed for both the $\Delta t = 1/N$ and $\Delta t = 1/N^2$ choices respectively.

$N+1$	FD $(1/N)$	FD $(1/N^2)$	GAL-PSD	TAU-TIN
5	0.24 (+3)	0.28 (+2)	0.15 (+1)	0.99 (0)
9	0.38 (+2)	0.21 (+1)	0.99 (+0)	0.99 (0)
17	0.86 (+1)	0.36 (0)	0.20 (0)	0.20 (0)
33	0.26 (+1)	0.82 (-1)	0.48 (-1)	0.48 (-1)
65	0.10 (+1)	0.20 (-1)	0.12 (-1)	0.12 (-1)
129	0.31 (0)	0.50 (-2)	0.29 (-2)	0.29 (-2)

Table 4.17 L_∞ values for the finite difference and the Chebychev backwards-Euler systems; the former are implemented with either $\Delta t = 1/N$ or $\Delta t = 1/N^2$, while the latter with $\Delta t = 1/N^2$. The Chebychev estimates for $N > 32$ have been computed by extrapolation.

The inferiority of this method may be identified immediately. Nevertheless, we can confirm its numerical absolute stability, the loss of relative accuracy with time and the superiority of the Chebychev systems over the finite differences.

$N+1$	FD $(1/N)$	FD $(1/N^2)$	GAL-PSD	TAU-TIN
5	0.12 (-1)	0.15 (-2)	0.62 (-4)	0.41 (-4)
9	0.20 (-2)	0.11 (-3)	0.48 (-4)	0.48 (-4)
17	0.44 (-3)	0.19 (-4)	0.10 (-4)	0.10 (-4)
33	0.14 (-3)	0.42 (-5)	0.25 (-5)	0.25 (-5)
65	0.52 (-4)	0.10 (-5)	—	—
129	0.23 (-4)	0.26 (-6)	—	—

Table 4.18 L_∞ values for the finite difference and the Chebychev backwards-Euler systems; the former are implemented with either $\Delta t = 1/N$ or $\Delta t = 1/N^2$, while the latter with $\Delta t = 1/N^2$. Results for the Chebychev systems have not been computed for $N > 32$ due to the excessive computational cost.

In addition, we observe identical results for the Galerkin and the pseudospectral methods; the tau methods produce same error estimates, which converge fast to the results of the former methods. The peculiarity of the tau projection is once more demonstrated, as it displays smaller errors than the Galerkin or the pseudospectral methods for $N = 4$.

The low temporal accuracy of the backwards-Euler propagation model has a strong influence on the size of Δt to be used in the time integration of the finite difference

matrix. The backwards-Euler scheme with $\Delta t = 1/N^2$ produces answers of a definite superiority over the results corresponding to the choice $\Delta t = 1/N$, in contrast to the Crank-Nicolson scheme.

4.4.7 Fast algorithms for the inversion of the integrated system

Despite the fact that implicit Chebychev methods lessen substantially the limitations associated with the size of the time integration step, it remains that full matrices need to be inverted at each new time step. Iterative methods are a promising alternative (see 2.5.5); remaining, however, in the field of direct methods, we concentrate on the integrated tau system, since its quasi-tridiagonal structural characteristics allow special inversion procedures to be devised. A detailed analysis of these algorithms is presented in Appendix B.2 from a general viewpoint. We now discuss particular details concerning their form in the case of the heat flow equation.

The importance of these techniques, when dealing with implicit time dependent problems, cannot be over-emphasized. The need for a repetitive inversion of the propagation matrix, makes their presence crucial. A trivial inspection of the quasi-tridiagonal even-odd component subsystems, reveals the dependence of the conditioning of these systems on the value of the product $(\sigma \Delta t)$; each one of these subsystems appears to be conditioned somewhat worse than the system as a whole. A rigorous analysis of the exact form of the reported dependence is not of an immediate interest to us and therefore, a deeper investigation is postponed until the next chapter, since it is the Schrödinger equation (chapter V) that is associated implicitly with the migration equation. Furthermore, the Schrödinger version should exhibit a different

conditioning dependence, due to the exclusive contribution of the factor $(\sigma\Delta t)$ in the imaginary component of the middle diagonal.

For the purposes of an introductory analysis, we constrain ourselves to the familiar choice of $\sigma = 1$ and $\Delta t = 1/N^2$. The conditioning of either subsystem is adequate, but this alone cannot guarantee success for the fast solvers. The latter employ a direct LU decomposition without pivoting and therefore they may easily fail even for perfectly conditioned matrices. Procedure **SLU1** is applied with the boundary row of 1's at the bottom of each subsystem; an upper tridiagonal matrix is obtained after forward elimination. The success of this extremely efficient algorithm, whose reliability is greatly desired, is directly associated with the *diagonal dominance* characteristics of the matrices involved. Diagonal dominance can never be achieved, in a strict sense, due to the presence of 1's in the boundary row. Despite that, we can still talk about diagonal dominance, in a looser sense, considering the rest of the matrix. Unfortunately, a strong off-diagonal dominance is present. The main diagonal being only $O(1/4n^2)$, the off-diagonal elements exhibit a magnitude superiority of $O(1/2n^2 - 1/2N^2)$; this amounts to a strong off-diagonal dominance since $n = 2, \ldots, N$.

Experiments for our given problem (the odd coefficients need not be propagated) show that **SLU1** is able to acquire solutions of a quality identical to the answers obtained via a general Gauss-elimination procedure with partial pivoting up to N=128. The algorithm has failed ultimately for $N = 256$, as a near-zero pivot demolishes its accuracy; the initial impact destabilizes the satisfaction of the boundary conditions and subsequently, a few more time steps destroy the approximation as a whole too. A row-balanced **SLU1** has not not show signs of improvement, either. Doubling the size of the problem results — as anticipated — in an absolute devastation, as the forward

elimination process is hindered, due to the encounter of a zero pivot. The reported off-diagonal dominance is $O(1/2n^2 - \Delta t/2)$, in general. Consequently, enlarged Δt values yield a more promising structure; Δt should, nevertheless, remain within a reasonable range for accuracy considerations.

The **SLU2** algorithm is implemented with the row boundary at the top; pivoting is not applied and a full upper triangular matrix is obtained after the completion of the forward elimination process. Inspection of the structure of the underlying matrix reveals only a weak off-diagonal dominance, since the off-diagonal elements' magnitude is $O(1/2n^2)$ compared to the magnitude of the diagonal element, i.e $O(1/2n^2 - 1/2N^2)$. The system should be more resistant to instabilities due to lack of pivoting; indeed, the **SLU2** procedure has performed exuberantly for our problem. We should point out, however, that an increase in the value of Δt results in a weaker main diagonal, the latter being $O(1/2n^2 - \Delta t/2)$ in general. The real value of this procedure is evident in problems with $\sigma = \sigma(t)$, which demand a repetition of the forward-elimination process at new time step; if σ is constant its efficiency contribution is negligible. Finally, the **SLU3** procedure, which is identical to the **SLU2** but for the partial pivoting that it incorporates, is to be used when no alternative exists and it can hardly be classified as fast.

Haidvogel and Zang (1978) claim that they " found pivoting to be unnecessary for this process", commenting on the inversion of the tau-integrated quasi-tridiagonal systems. There, they deal with the two-dimensional Poisson's equation and their ADI-SOR algorithm involves a direct integrated-tau solver, where the diagonals are $O(\omega_\nu/4n^2)$, $O(1-[\omega_\nu/2n^2])$ and $O(\omega_\nu/4n^2)$ respectively, with ω_ν being the relaxation parameter of the SOR process. In their case, the choice of ω_ν and the moderate size

of the systems may be responsible for the absolutely satisfactory performance of the fast solvers. The same applies for the experiments of Haldenwang et al (1984); they also do not seem to attempt the inversion for $N > 64$. Thus, $SLU1$ does not collapse and the inversion counts $O(8N)$ operations only.

CHAPTER V

THE SCHRÖDINGER EQUATION

On the beach in the foreground the painter had arranged that the eye should discover no fixed boundary, no absolute time of demarcation between earth and ocean

Remembrance of things past — Marcel Proust

5.1 The One-Dimensional Linear Schrödinger Equation

In quantum mechanics the state of a system is characterised by a wavefunction $\Psi(\mathbf{r},\mathbf{t})$. The time evolution is fixed by the *Schrödinger* equation

$$i\hbar\frac{\partial\Psi}{\partial\mathbf{t}} = \hat{H}\Psi \tag{5.1}$$

where $\hat{H}$ is the Hamiltonian operator of the system.

Solution of the Schrödinger equation and, therefore, determination of the wavefunction $\Psi(\mathbf{r},\mathbf{t})$, allows us to obtain all the dynamical information of the physical

system under consideration (Kosloff and Kosloff, 1983a, b). The general form of the Hamiltonian operator is (in the absence of a magnetic field) $\hat{H} = \hat{T} + \hat{V}$ (Bisseling and Kosloff, 1985). That is, the sum of the kinetic and potential energy operators, respectively.

The kinetic energy operator is equal to $\hat{p}^2/2m$ where $\hat{p}$ is the momentum operator and m is the mass of the system. The momentum operator $\hat{p}$ is $-i\hbar\nabla$ and since $\hat{V}$ operates on Ψ just as a simple multiplier, the above expression for the Hamiltonian becomes

$$\hat{H} = \frac{\hat{p}^2}{2m} + \hat{V} = -\frac{\hbar^2}{2m}\nabla^2 + \hat{V} \tag{5.2}$$

where ∇^2 is the well-known Laplacian operator. Before we proceed to discuss some numerical approaches and the problems associated with them, we will make a detour to present a brief review of the fundamentals concerning wavepackets in quantum mechanics.

5.1.1 Free propagation of a wavepacket

Let us consider the one-dimensional case with $\Psi = \Psi(x,t)$. At some instant $t = t_0$, we can write the plane-wave decomposition of the wavefunction $\Psi(x, t = t_0)$ as

$$\Psi(x,t_0) = \int_{-\infty}^{+\infty} f(k)\exp(ikx)\,dk \tag{5.3}$$

with k being the wavenumber. For a wavepacket to be satisfactorily defined, the amplitude spectrum $|\Psi(x)|$ should be sharp, being maximal for $k = k_0$ and having most of the energy only in a band Δk about k_0 (Hamilton et al, 1972). In addition,

the phase should be a slowly varying function of k in the interval Δk (Diu, 1980) and $k_0 \gg \Delta k$ for the group velocity, i.e $c_g = k_0/m$, to be well-defined. Such a wavepacket would exhibit similar features in position space; $|\Psi(x)|$ would be concentrated in an interval Δx around some $x = x_0$ value.

As time passes, the Schrödinger equation, which is basically a parabolic equation for the evolution of a complex quantity (Press et al, 1985), will cause a progressive diffusion of the position probability density function $\Psi\Psi^*$ (* denotes complex conjugate). Simultaneously the wavepacket will advance in x and it will spread out (as a direct consequence of the highly dispersive nature of the Schrödinger equation), so that the total probability is conserved, remaining equal to its normalized value (Lawden, 1967). It is obvious, of course, that according to the Heisenberg's principle of uncertainty, the sharper the distribution of $\Psi\Psi^*$ (enhanced resolution in position space), the broader the peak of ff^* (reduced resolution in momentum space).

Let us consider a free wavepacket ($\hat{V} = 0$) with initial wavefunction $\Psi(x, t = 0) = B(x)$. Decomposing $B(x)$ gives us

$$B(x) = \int_{-\infty}^{+\infty} B(k) e^{ikx}\, dk \tag{5.4}$$

with the Fourier spectrum given as

$$B(k) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} B(x) e^{-ikx}\, dx \tag{5.5}$$

In order to obtain the solution after time t, we propagate each of its Fourier components and we then synthesize the total field at the time of interest. For some specific $k = K$,

the component $B(K)e^{iKx}$ after time t will be equal to $B(K)e^{iKx}e^{-i\omega t}$ where

$$\omega(K) = \frac{E(K)}{\hbar} = \frac{p^2(K)}{2m\hbar} = \frac{\hbar K^2}{2m} \tag{5.6}$$

Thus $B(K)e^{iKx}$ becomes

$$B(K)e^{iKx}e^{-i\hbar K^2 t/2m} = B(K)e^{i(Kx-\hbar K^2 t/2m)} \tag{5.7}$$

that is, a plane-wave propagating in the direction of the positive x-axis with phase speed $c = \omega/K = \hbar K/2m$.

Superimposing all modes after time t, gives us the solution

$$\Psi(x,t) = \int_{-\infty}^{+\infty} B(K)e^{iK(x-c(K)t)}\,dK \tag{5.8}$$

Even hough each mode travels with speed c, the center of the wavepacket (and therefore the energy or the probability density) travels with the group velocity

$$c_g = \frac{d\omega}{dK} = \frac{\hbar K}{m} = 2c \tag{5.9}$$

which can be easily shown by the stationary phase method (Diu, 1980).

Either one of expressions (5.6) and (5.9) reveals the dispersive character of the Schrödinger equation. This dispersion is characterized as *anomalous* and it is quite profound. Consequently, short wavelengths accelerate with respect to long wavelengths; this amounts to the familiar situation, where an observer, who is travelling at c_g with

a finite wavetrain, sees wavecrests emerging at the leading edge. These are then incorporated into the wavetrain and finally, vanish at its trailing edge. (Tolstoy, 1973).

5.1.2 Free propagation of a Gaussian wavepacket

Most numerical methods for the integration of the Schrödinger equation are tested against the available analytic solution of a wavepacket that is amplitude modulated by a Gaussian distribution. This model avoids mathematical discontinuities, while it still decays quite rapidly exhibiting a satisfactory sharp peak. Let us proceed by defining a Gaussian wavepacket and calculate the exact analytic solution to be used for comparison with the numerical results.

Consider a one-dimensional Gaussian wavepacket at $t = 0$ with wavefunction

$$\Psi(x, t = 0) = Ae^{-(x^2/2\Delta^2)}e^{i\ k\ x} \tag{5.10}$$

(Tomonaga, 1966). We can see that the modulator, $\exp(-x^2/2\Delta^2)$, has non-negligible values only in the region of length Δ about the origin. With that choice, the total position probability density

$$\int_{-\infty}^{+\infty} \Psi\Psi^* \, dx \tag{5.11}$$

is normalized to $A^2/\sqrt{\pi\Delta}$ instead of unity (Bramhall and Casper, 1970), but this is rather a matter of definition. Its Fourier transform gives us the momentum (or equivalently wavenumber since we have adopted atomic units with $\hbar = 1$) space probability density

$$B(k') = \frac{A}{2\pi} \int_{-\infty}^{+\infty} \exp\left(-\frac{x^2}{2\Delta^2}\right) \exp(ikx) \exp(-ik'x) \, dx \tag{5.12}$$

or

$$B(k') = \frac{A}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\Delta^2(k-k')^2\right] \tag{5.13}$$

We observe that this function has a maximum at $k' = k$ and it takes an appreciable value only in a region $1/\Delta$ on both sides of this point.

We now see the effect of the modulation on the wavenumber spectrum of the original wave; a multitude of new wavenumbers, occupyg the two symmetrically disposed bands around the *carrier* wavenumber k, have been created.

It is apparent that the wavepacket can be obtained from the superposition of the plane waves with wavenumbers in the vicinity of k. The final result reads

$$\Psi(x,t) = \frac{A}{(1+it/m\Delta^2)^{1/2}} \exp\left\{-\frac{1}{2\Delta^2(1+it/m\Delta^2)}\left[x^2 - 2i\Delta^2 k\left(x - \frac{kt}{2m}\right)\right]\right\} \tag{5.14}$$

(Tomonaga, 1966) and $\Psi\Psi^*$ gives us the position probability density at time t. To simplify the notation, we extend our use of atomic units to include $m = 1/2$; we also assign A to unity.

5.2 Numerical Solution of the Schrödinger Equation

A numerical simulation of the wavepacket's propagation is not trivial in the sense that it involves many pitfalls. Let us then proceed and write down the Schrödinger equation in atomic units ($\hbar = 1$, $m = 1/2$) .

$$i\frac{\partial \Psi}{\partial t} = \hat{H}\Psi \tag{5.15}$$

assuming that the particles are free ($\hat{V}$=0). This leaves the Hamiltonian $\hat{H} = -\nabla^2$ (negative Laplacian), which in one dimension reduces further to $\hat{H} = -\partial^2/\partial x^2$. We can then rewrite the equation in the much simplified form

$$\frac{\partial \Psi}{\partial t} = i\frac{\partial^2 \Psi}{\partial x^2} \tag{5.16}$$

Let us begin by representing by Ψ_j^n the value of the discrete approximation to $\Psi(x,t)$ at the space-grid point $j\Delta x$ and the time-grid point $n\Delta t$.

5.2.1 Finite difference methods

We shall first discuss the existing finite-difference solutions to the Schrödinger equation. Traditionally, the second-order central difference operator has been employed for the representation of the $\partial^2/\partial x^2$; other approaches will be mentioned later.

Having decided upon the discretization of the spatial derivative, we only need to choose an appropriate scheme for the time derivative. The formal analytical solution reads

$$\Psi_j^{n+1} = e^{-i\,dt\hat{H}}\Psi_j^n \tag{5.17}$$

and it is obvious that we need an approximation to $\exp(-i\,dt\hat{H})$ for the time-advancing. Writing down the Taylor expansion

$$\exp(-i\,dt\hat{H}) = \sum_{k=0}^{\infty} \frac{(-i\,dt\hat{H})^k}{k!}\hat{H}^k \tag{5.18}$$

and truncating (5.18) after the first two terms, gives us $\exp(-i\,dt\hat{H}) \simeq 1 - i\,dt\hat{H}$ (Goldberg et al, 1967). Substituting this into the analytical solution we obtain the classic forward-Euler scheme, that is,

$$\Psi_j^{n+1} = (1 - i\,dt\hat{H})\Psi_j^n = \Psi_j^n - i\,dt\hat{H}\Psi_j^n \tag{5.19}$$

This is an explicit scheme that is very easy to manipulate but unfortunately it may be easily shown (Askar and Cakmak, 1978) that it is unconditionally unstable. We may think that by retaining another term in the Taylor expansion of the exponential evolution operator, we might be able to go around this problem. This does not happen, though, and

$$\exp(-i\,dt\hat{H}) \simeq 1 - i\,dt\hat{H} - (dt\hat{H})^2/2 \tag{5.20}$$

leads to another unstable scheme (Mc Cullough and Wyatts, 1971a, b).

Leaving explicit schemes aside for the moment, we proceed by constructing an implicit first-order accurate scheme as follows: Discretizing the analytical solution we obtain

$$\Psi_j^{n-1} = \exp(i\,dt\hat{H})\Psi_j^n \tag{5.21}$$

and the first order approximation to the exponential yields

$$\Psi_j^{n-1} = (1 + i\,dt\hat{H})\Psi_j^n \tag{5.22}$$

or

$$\Psi_j^n = (1 + i\,dt\hat{H})^{-1}\Psi_j^{n+1} \tag{5.23}$$

As we easily see this scheme is unconditionally stable (Press et al, 1985). Although we have ensured stability, we are not satisfied yet because it is obvious that the above scheme is dissipative, where the Schrödinger solution conserves the total probability density (energy) through the unitary nature of the evolution operator $\exp(-i\,dt\hat{H})$. Therefore, such a choice would exhibit a deteriorating accuracy performance in the course of time. The solution is found in Caley's form of approximating the exponential as

$$\exp(-i\,dt\hat{H}) \simeq \frac{1 - i\,dt\hat{H}/2}{1 + i\,dt\hat{H}/2} \tag{5.24}$$

(Goldberg et al, 1967) which clearly conserves energy with an amplification factor of one. The above difference form leads to the very familiar Crank-Nicolson method, where the uniform centering of both the time and the space derivatives at $(n + 1/2, j)$ ensures maintainance of the time reversal symmetry of the Schrödinger equation. In addition,this scheme is second-order accurate in time.

Although the Crank-Nicolson scheme enjoys unconditional stability and performs quite accurately when appropriately chosen $\Delta x, \Delta t$ are used (a detailed analysis of the procedure concerning the choice of these parameters is presented later), it still remains an implicit scheme and it, subsequently, requires the inversion of a system of simultaneous equations for every time step. This is a rather trivial task in one dimension, due to the tridiagonal structure of these matrices, but it can pose formidable problems as the increase in the dimensionality of the problem might force us to face a prohibitively large amount of computer work (Claerbout, 1985).

Recently an explicit scheme for the time integration of the Schrödinger equation has been presented. This is based on another uniform and symmetric centering of the derivatives appearing in the equation.

We write $\Psi_j^{n+1} = e^{-i\,dt\hat{H}}\Psi_j^n$ and $\Psi_j^{n-1} = e^{+i\,dt\hat{H}}\Psi_j^n$ as before; we then subtract by parts to obtain

$$\Psi_j^{n+1} - \Psi_j^{n-1} = (e^{-i\,dt\hat{H}} - e^{+i\,dt\hat{H}})\Psi_j^n \tag{5.25}$$

Taylor expansions, similar to the ones used previously, yield

$$\Psi_j^{n+1} - \Psi_j^{n-1} = -2i\,dt\hat{H}\Psi_j^n \tag{5.26}$$

which can be easily identified as the classic two step leap-frog scheme. This explicit scheme is conditionally stable and as accurate as the Crank-Nicolson one, i.e $O(\Delta x^2) + O(\Delta t^2)$. It is not strictly unitary, but the error involved is some orders of magnitude less that the error of the scheme itself and the stability of the scheme guarantees that it will not cumulate as time progresses (Askar and Cakmak, 1978). The Schrödinger equation is highly dispersive, the dispersion relation being $\omega = k^2$. Finite differences introduce phase-errors due to inadequate derivative approximations, which exhibit themselves as artificial dispersion altering the true dispersion relation.

Other approaches over the years, include a combination of a five point rather than the classic three point difference formula to approximate the Hamiltonian and a fifth order predictor-corrector scheme for the time derivative (Kulander, 1978), while another interesting approach involves a finite element formulation of the equation (Askar, 1981). Both the above techniques (although exhibiting relative advantages

over existing finite difference codes), suffer from artificial dispersions due to their non-spectral character.

5.2.2 Spectral methods

5.2.2.1 Spectral semi-discretizations

Lately pseudospectral Fourier and Hankel methods have been used to approximate the spatial derivative in the time dependent Schrödinger equation. For rectangular periodic geometries, Fourier spectral methods (Kosloff and Kosloff, 1983a, b) are most appropriate as the derivatives are evaluated exactly and boundary conditions are handled in a very natural way; for radial geometries Hankel spectral methods are most suitable (Bisseling and Kosloff, 1985). The spectral approach has some very important consequences. It maps the continuous physical Hilbert space to a discrete one, since Hermitian operators maintain their fundamental property under this mapping. Consequently, they still possess real eigenvalues and they satisfy the commutation relations of quantum mechanics to the degree of the machine accuracy. As anticipated, finite difference representations fail to do so, because they approximate a global quantity, such as the kinetic energy, using local functions (Tal Ezer and Kosloff, 1984).

Artificial spatial dispersion can be avoided altogether if aliasing is absent. However, a finite difference operator for the time derivative will still result in the introduction of artificial temporal dispersion. Unconditionally stable schemes would then exhibit a serious accuracy deterioration, if a sufficiently small Δt has not been chosen. On the other hand, a conditional stability of some scheme could be lost too, if an inappropriately large time step is chosen, causing the introduction of instabilities, which could render the numerical simulation meaningless. Fortunately, a proper decrease in

the size of the time step (which corresponds to a more accurate extrapolated value in the truncated Taylor series) to a fraction of the stability limit, can virtually provide us with a dispersion relation which matches the true one to a high degree of accuracy. That is not the case with the previous finite difference approaches since we do need an oversampling in the spatial dimension in order to eliminate spatial artificial dispersion (in addition to the same decrease in the size of the time step)(Kosloff and Kosloff, 1983a, b).

5.2.2.2 Full-spectral techniques

A spectral time propagation scheme has been recently proposed in order to introduce infinite accuracy in the approximation of the time derivative (Tal Ezer and Kosloff, 1984). This is based on a truncated expansion of the evolution operator in terms of the complex Chebychev polynomials in an appropriate coordinate system (Tal Ezer, 1984). There, it is the order of the complex time-interpolating polynomial that controls accuracy and not the actual size of Δt; consequently, this scheme can be used as an one step propagator if intermediate results are of no interest (Kosloff and Kosloff, 1986). Recently (Reshef and Kosloff, 1986) this technique was employed for migrating common-shot gathers in exploration geophysics. A somewhat different "spectral" technique solves the time dependent Schrödinger equation with a split operator FFT method followed by correlating the wavefunction with its initial state — a kind of numerical spectroscopy (Feit et al, 1982).

5.3 The Sommerfeld Radiation Condition

The continuous Schrödinger equation is associated with the radiation boundary conditions at infinity, that is, $\lim_{x \to \pm\infty} \Psi(x,t) = 0$. Our computational grid is finite though, and that is going to cause trouble, as time progresses, since with the wavepacket arriving at a rigid boundary wall, the physics of the problem will be violated. This artificially posed boundary will cause reflection of the wavepacket and it will, therefore, degrade the accuracy of the representation. The degradation is, originally, witnessed near the reflecting boundary only, but it soon pollutes the complete computational grid. It should be emphasized that finite differences tend to exhibit a slower contamination rate, when compared to spectral techniques, of the spatial features of the solution, which lie away from the reflecting boundary. This is due to the *local* nature of the former, as opposed to the *global* nature of the latter. In Fourier methods, in particular, the periodic character of the basis functions causes the familiar wraparound effect, which could be visualized as periodic reflections, as well.

A naive way of coping with the finite mesh proper boundary condition simulation involves an *adequate* increase of the size of the computational grid. Quantification of the classification *adequate* emerges from shifting the undesired, troublesome reflecting boundary outside the realm of interest, so that an unrealistic, but nevertheless, satisfactory numerical simulation of the radiation boundary condition is achieved. Although this trick practically solves the problem, it suffers from the disadvantage of introducing a lot of unneccesary memory requirements and a larger amount of computations. The familiar "padding with zeros" of Fourier methods corresponds to the described expansion of the grid size shifting the wraparound interference outside the realm of interest in the computational grid.

More sophisticated methods should involve some kind of absorbing boundaries. The highly dispersive Schrödinger equation poses severe difficulties for methods that impose an one way equation at the boundary allowing energy to propagate in the outward direction only (Israeli and Orszag, 1981); the situation becomes much more obscure in higher dimensions (Galbraith et al, 1984). In particular it is not at all clear how to formulate this approach for the Fourier scheme (due to the periodic structure implicitly imposed by the latter). Very recently (Kosloff and Kosloff, 1986), absorbing boundaries involving a complex potential, which attenuates the wave amplitude at the grid boundary regions, have been developed.

5.4 A New Implicit Chebychev Technique

We aim to maintain the high accuracy of the Fourier method, while relaxing the imposed periodic boundary structure. The answer lies in a spectral device of non-periodic character, since such a transform environment eliminates the wraparound reflections, inherent in a Fourier environment. This technique, nevertheless, needs to be augmented with an effective absorbing boundary condition mechanism. We thereby propose a new semi-discretization of the problem; the Hamiltonian operator is built in a Chebychev environment, while the Crank-Nicolson scheme is employed for the time integration.

All three different projection operators are implemented. The mathematical formulation follows closely the formulation for the heat equation, the only difference being the imaginary character of the diffusion-like coefficient of the Hamiltonian. Complex LU solvers (Bowdler et al, 1966) are used to invert the system of the linear simultaneous equations and evidently the fast LU solutions for the integrated-tau systems are

appropriately modified to match the needs of complex arithmetic. The performance of the method is thoroughly investigated and is compared with the performance of the classic second order accurate Crank-Nicolson finite difference scheme.

We commence the presentation with a brief quantitative summary of the fundamentals issues governing the important aspects of the numerical simulation of the wavepacket's propagation, in both schemes to be compared. The following presentation comprises an integration of the knowledge and the experience gained through the previous numerical experiments. Furthermore, it incorporates some indispensable insight into the problem given by Goldberg et al (1967) and Claerbout (1976, 1985).

5.4.1 Spatial aliasing

The first important issue is spatial aliasing. We calculate the maximum present wavenumber (momentum) in the wavepacket's spectrum, i.e $k_{\max}$, at a given machine precision level. Clearly, this value has to be smaller than the Nyquist angular wavenumber, i.e $k_{\mathrm{NYQ}} = \pi/\Delta x$, which identifies the highest wavenumber resolvable for a given mesh size. The described analysis is absolutely sufficient for the classic scheme's aliasing considerations but it is obviously inadequate to cover the needs of the Chebychev scheme, which, equivalently, demands a similar procedure but in a Chebychev fashion instead, i.e finding out the highest significant Chebychev coefficient.

5.4.2 Spatial artificial dispersion

A finite difference simulation of the Hamiltonian operator gives rise to spatial artificial dispersion. Diminution of the negative effects of the latter dictates an oversampling in the x-coordinate. The sampling density, for a particular wavenumber k,

is reflected in the quantity $(k\Delta x)^{-1}$; the latter measures the number of grid points employed in the reconstruction of the wavelength λ. Since the eigenfunctions of our model problem are $\sin(kx)$, we obtain

$$\frac{\Delta^2}{\Delta x^2}\sin(kx) = -\frac{2(1-\cos k\Delta x)}{\Delta x^2}\sin(kx) \tag{5.27}$$

and introducing the Taylor expansion of $\cos(k\Delta x)$, we see that

$$-\bar{k}^2_{\max} = -k^2_{\max}\left[1 - \frac{(k_{\max}\Delta x)^2}{12}\right] + O\Big((k_{\max}\Delta x)^4\Big) \tag{5.28}$$

Expression (5.28) shows clearly that the maximum (in absolute value) eigenvalue of the finite difference representation of the Hamiltonian, i.e $-\bar{k}^2_{\max}$ converges fast to the corresponding eigenvalue, i.e $-k^2_{\max}$, of the true Hamiltonian, if

$$\frac{(k_{\max}\Delta x)^2}{12} \ll 1 \tag{5.29}$$

Alternatively, the quality of the approximation may be measured from the equivalent relation

$$\bar{k}_{\max} = \frac{2}{\Delta x}\sin\frac{k_{\max}\Delta x}{2} \tag{5.30}$$

5.4.3 Temporal artificial dispersion

The bilinear approximation to the time propagation operator introduces artificial temporal dispersion into both schemes. Rewriting expression (5.24) as

$$\exp(-i\Delta t\hat{H}) \simeq \exp\left(-2i\arctan(\hat{H}\Delta t/2)\right) \tag{5.31}$$

allows us to identify the approximation of the true eigenvalues $-i\hat{H}$, of the $\partial/\partial t$ operator with the quantities

$$\overline{-i\hat{H}} = \frac{2}{\Delta t}\frac{1-\exp(i\hat{H}\Delta t)}{1+\exp(i\hat{H}\Delta t)} \tag{5.32}$$

The eigenvalues of $\hat{H}$ are k^2 (assuming no errors in its numerical simulation) and combining that with the dispersion relation $\omega = k^2$, we can easily derive a very useful expression relating ω and its bilinear approximates $\bar{\omega}$. This approximation reads

$$\bar{\omega} = \frac{2}{\Delta t}\tan\left(\frac{\omega\Delta t}{2}\right) \tag{5.33}$$

and this quality is linked to the quantity $(\omega\Delta t)^{-1}$, which measures the number of samples in a period of a given frequency ω. Consequently, phase errors are introduced and they accumulate over a given number of time steps; computing the error per time step for a particular frequency, multiplying it with the number of time steps, i.e let us say M, we may compute a relative (among the various modes) phase error. Diminution of this error, i.e

$$(M\Delta t^3/12)[k_{\max}^6 - k_0^6] \ll 1 \tag{5.34}$$

in the case of a wavepacket with average wavenumber k_0, is a definite prerequisite for an appropriate numerical simulation of the wavepacket's propagation.

5.4.4 The initial condition and the propagation parameters

5.4.4.1 A Gaussian wavepacket

A Gaussian amplitude modulator is chosen and it is initialized in such a way that it decays quite rapidly; it therefore possesses a well-defined group velocity and it lies far away from the boundaries of our computational grid (which correspond to mathematical infinities). We then normalize our spatial computational grid from -1 to $+1$ and center our wavepacket at $x = 0$. The wavepacket is defined as

$$u(x, t = 0) = \exp(ik_0 x) \exp(-x^2/2\sigma_0^2) \tag{5.35}$$

and by choosing $\sigma_0 = 0.05 = 2/20$ (where 2 is the total length of our grid), we succceed in obtaining a rapidly vanishing Gaussian around its rather sharp peak (so that it is virtually negligible in a distance more than σ_0 from $x = 0$). Furthermore, its spectrum is not too spead out about k_0 in the wavenumber space and it also satisfies homogeneous boundary conditions.

5.4.4.2 Boundary reflections and the choice of the average wavenumber

The alleviation of the artificial reflections from the boundary $x = +1$ is the last issue to be considered.

We adopt, for the purposes of this thesis, a variant of the simple trick described previously. Thus we predetermine the distance to be travelled by the wavepacket,

assuring that it does not impinge on the rigid boundary at $x = +1$. In that way, the wavefunction remains practically zero on and in the neighborhood of the artificial boundary and, thereafter, an important physical property of the true solution is satisfied.

The materialization of this approach, though, imposes certain restrictions on the parameters of the problem. The choice of k_0 defaults to a certain value of group velocity for the position probability density (energy). However, the obvious interdependence, between the average wavenumber of the wavepacket and the duration of the propagation, can lead to boundary reflections if incorrect choices are made. The group velocity of the centre of the wavepacket is (recall expression 5.9) $c_g = p/m = \hbar k_o/m = 2k_0$. Let us assume $L = N \cdot \Delta x$ is the total length of the grid where N is the number of spatial grid points. Furthermore, let us write $T = M \cdot \Delta t$ with T being the duration of propagation and M the total number of time steps corresponding to a time step of Δt.

We proceed to propagate the wavepacket, originally centered at $x = 0$, a certain distance X, so that it does not come too close to the $x = +1$ boundary (since it moves to the right). Then

$$X = c_g T = 2k_0 T \Rightarrow k_0 = X/2T \tag{5.36}$$

This indirect way of specifying the average wavenumber should guarantee us the desired simulation, but unfortunately that is not so, since there is an additional major source of problems, namely the highly dispersive nature of the Schrödinger equation, which causes the solutions as time progresses. As a result, the wavepacket spreads out and a k_0 defined through expression (5.36) might not be adequate to prevent artificial

reflections, as an excessive spatial spreading might cause the right flank of the packet to impinge on the boundary much faster than the analysis shows.

Therefore, care needs to be taken, in order to sustain the original spread allowing only negligible and controllable amount of additional spreading to occur as time advances. The spread σ after time T can be calculated as a function of both the initial spread σ_0 and the time T as

$$\sigma^2 = \left(\sigma_0^4 + 4T^2\right)^{1/2} \tag{5.37}$$

provides an additional constraint that will now guarantee an indirect satisfaction of the radiation boundary conditions of the continuous problem. A concurrent limitation of the maximum duration of the propagation is also imposed.

5.7 Choice of Error Norms

A certain norm needs to be chosen for the quantification of the quality of the approximate solution $\bar{y}_i$. The issue is rather unclear and it depends upon the special characteristics of the problem under consideration and the specific features of the solution, we are particularly interested in. The matter touches upon the general question of comparing two discrete vectors of complex elements.

Various choices are available but, for our problem, we feel that a normalized L_2 energy norm suffices. This norm is defined as

$$\bar{L}_2^{(0)} = \frac{\sum_{i=0}^{N} |y_i - \bar{y}_i|^2}{\sum_{i=0}^{N} |y_i|^2} \tag{5.38}$$

and in its "infinity" form, as

$$\bar{L}_{\infty}^{(0)} = \frac{\max_{0\leq i\leq N} |y_i - \bar{y}_i|}{\max_{0\leq i\leq N} |y_i|} \tag{5.39}$$

Justification of this *classic* choice may be acquired by envisioning the behavior of (5.38) in the complex plane. The norm follows the "beak" of the updated resultants of the complex vectors; it, therefore, amounts to computing global energy discrepancies normalized to the energy of the reference vector, i.e the exact solution vector. We could, then, term this norm as the "small" norm; the superscript $^{(0)}$ refers to the order of the derivative of the vectors, whose energy differences are being measured.

The analysis of the performances of the various schemes to follow relies on information gathered using the norms (5.38) and (5.39); the latter reveals pointwise departures, which could probably be inconsistent with the global estimates. Although, erroneous isolated local deviations are not anticipated in our problem, the $\bar{L}_{\infty}^{(0)}$ has occasionally been proven indispensable in revealing boundary reflections, whose contamination effects have become invisible in the global estimate. However, we ought to note, a viable alternative or counterpart to the $\bar{L}^{(0)}$ norms. These are the "flat" $\bar{L}^{(1)}$ norms, which tend to measure the energy discrepancies between the first derivatives of the vectors to be compared, either in a global sense, i.e $\bar{L}_{2}^{(1)}$, or in a maximum pointwise fashion, i.e $\bar{L}_{\infty}^{(1)}$.

The nature of these "flat" norms consists in tracing the trails of the vectors point by point, as contrasted to marking their successive leading edges only. That defaults in an annihilation of the DC shifts from the origin, inherent in the vector sections

themselves, amounting, thereafter, to an evaluation of the derivatives' differences. The "flat" norms are expressed as

$$\bar{L}_2^{(1)} = \frac{\sum_{i=0}^{N-1} \left|(y_{i+1} - y_i) - (\bar{y}_{i+1} - \bar{y}_i)\right|^2}{\sum_{i=0}^{N-1} \left|y_{i+1} - y_i\right|^2} \tag{5.40}$$

and

$$\bar{L}_\infty^{(1)} = \frac{\max_{0 \le i \le N-1} \left|(y_{i+1} - y_i) - (\bar{y}_{i+1} - \bar{y}_i)\right|}{\max_{0 \le i \le N-1} \left|y_{i+1} - y_i\right|} \tag{5.41}$$

A fundamental characteristic common in both the "small" and the "flat" norms is that they receive contributions due to both the magnitude and phase differences indiscriminately. Distinguishing between the partial discrepancies is not feasible, as the existing ambivalence is not resolvable.

Ambitions of further resolution between magnitude and phase errors, demand the construction of additional norms, especially designed to measure the "closeness" of the two vectors with respect to one particular parameter only. The interpretation is thus aided by augmenting the coupled norms' estimates with the auxiliary norms.

The issue is not trivial but it appears that magnitude norms are easier to visualize and to construct, either in a "small" or a "flat" environment. Nevertheless, these norms suffer from an improper balancing, with respect to the coupled estimates, which hinders an accurate quantification of their differences.

Phase norms' structures are much more difficult to decide upon. The fundamental issue of *phase-unwrapping* is inherent to the problem and furthermore, computations of the phase angle, through the inverse tangent in the complex plane, are seen to be very

vulnerable to numerical noise, frequently providing meaningless results. Phase computations, involving the difference vector or, alternatively, the individual vectors separately, have their share of advantages and drawbacks. The former approach maintains an absolute phase difference between 0 and 2π, at the expense of an approximation being introduced.

Finally, the logarithmic approach to the problem, which tends to operate as a compressor of magnitude differences, amplifying, therefore, the phase discrepancies, is not considered.

5.6 Numerical Experiments and Analysis of Performance

5.6.1 Parameter initialization

The analysis presented in (5.4.4.2) has led us to choose $X = L/8 = +0.25$ and $T = 0.005$, which default to an average wavenumber $k_0 = 25$ with a standard deviation $\Delta k = 1/\sigma_0 = 10$. The choices $\sigma_0 = 0.1$ and $T = 0.005$ allow a final spreading $\sigma \simeq 0.19$, only; the latter indicates an 19% increase of the initial σ_0 value.

Furthermore, we maintain $\Delta t = 1/N^2$ for an $(N + 1)$-long spatial discretization; compliance of this time-stepping pace with a uniform length of propagation T is achieved by introducing a rational number of time steps.

5.6.2 Analysis of the initial condition

5.6.2.1 The unmodulated signal

A comparison of the Fourier and the Chebychev amplitude spectrum of the unmodulated wave $\exp(ik_0x)$ may be made by inspection of figures (5.1a) and (5.2a), respectively.

In the Fourier spectrum, we identify the anticipated "spike" at $k = k_0$, in the wavenumber space, i.e

$$a(k) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \exp(ik_0x)\exp(-ikx)dx = \delta(k-k_0) \tag{5.42}$$

The Chebychev spectrum of $\exp(ik_0x) = \cos(k_0x) + i\sin(k_0x)$ may be computed exactly; the procedure is an extended version of the computational process given for the cosine. The transform is, obviously, complex and it reads

$$a_n = 2\sum_{l=0}^{\infty}\frac{(-1)^l}{c_l}J_{2l}(k)\delta_{n,2l} \;+\; 2i\sum_{l=0}^{\infty}(-1)^l J_{2l+1}(k)\delta_{n,2l+1} \tag{5.43}$$

The amplitude and the phase of the a_n's are plotted in figure (5.2). The amplitude distribution is quite stimulating, as its peak is identified at the close vicinity of $n = 25$, while significant amounts of energy are present at the lower modes, only. We observe, once more, the characteristic convergence pattern of Chebychev expansions. The function $\exp(25ix)$ has approximately 8 complete wavelengths inside $[-1,+1]$; its Chebychev spectrum reaches a global maximum at about 25 and and it, then, exhibits higher coefficients characterized by negligible amplitudes, demonstrating that 8π polynomials suffice for a good convergence. Nevertheless, maximum resolution for single-precision extends up to $n = 40 - 45$.

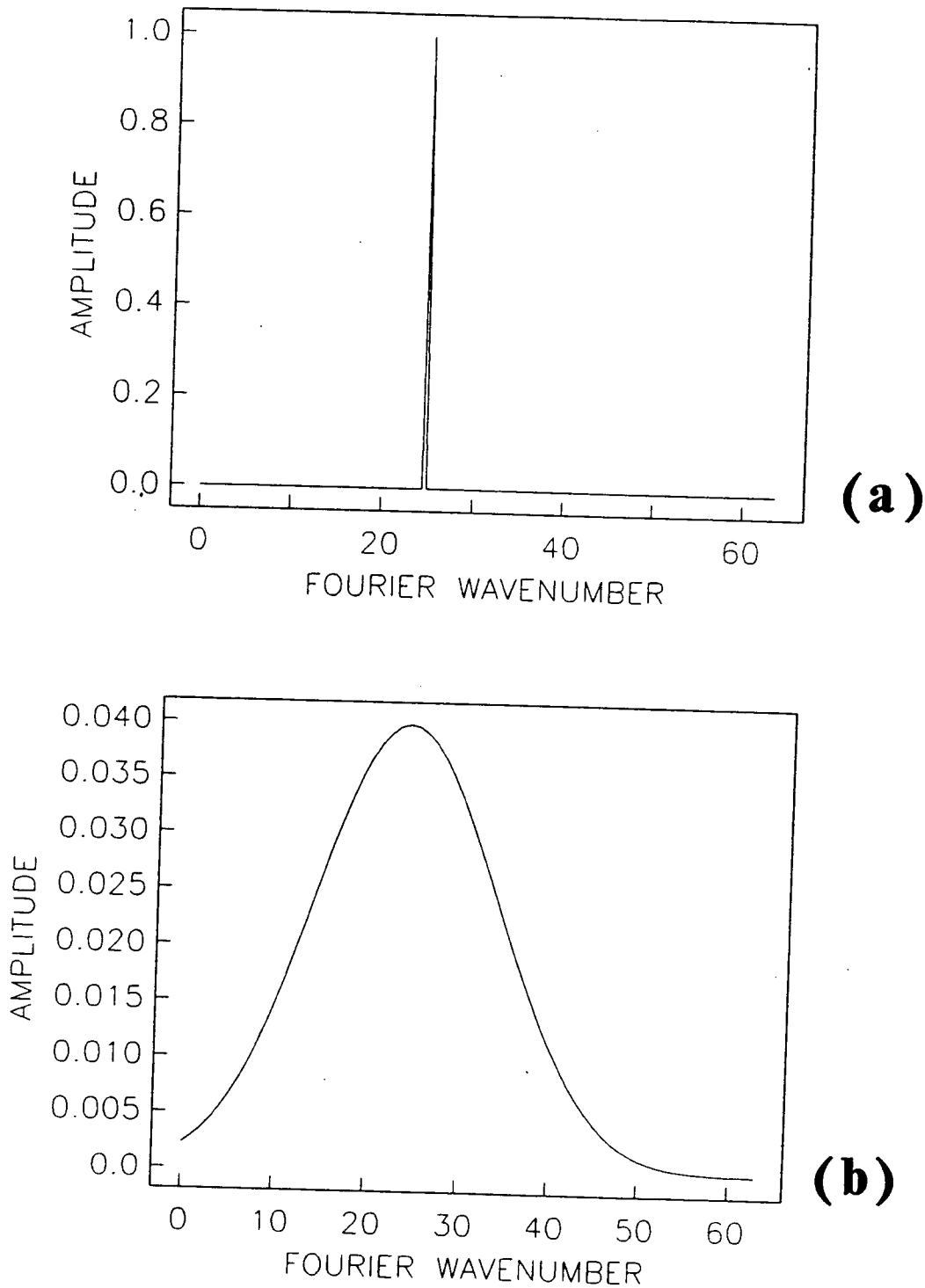

Figure 5.1 Amplitude spectrum of the true Fourier transform of (a) the unmodulated and (b) the modulated signal.

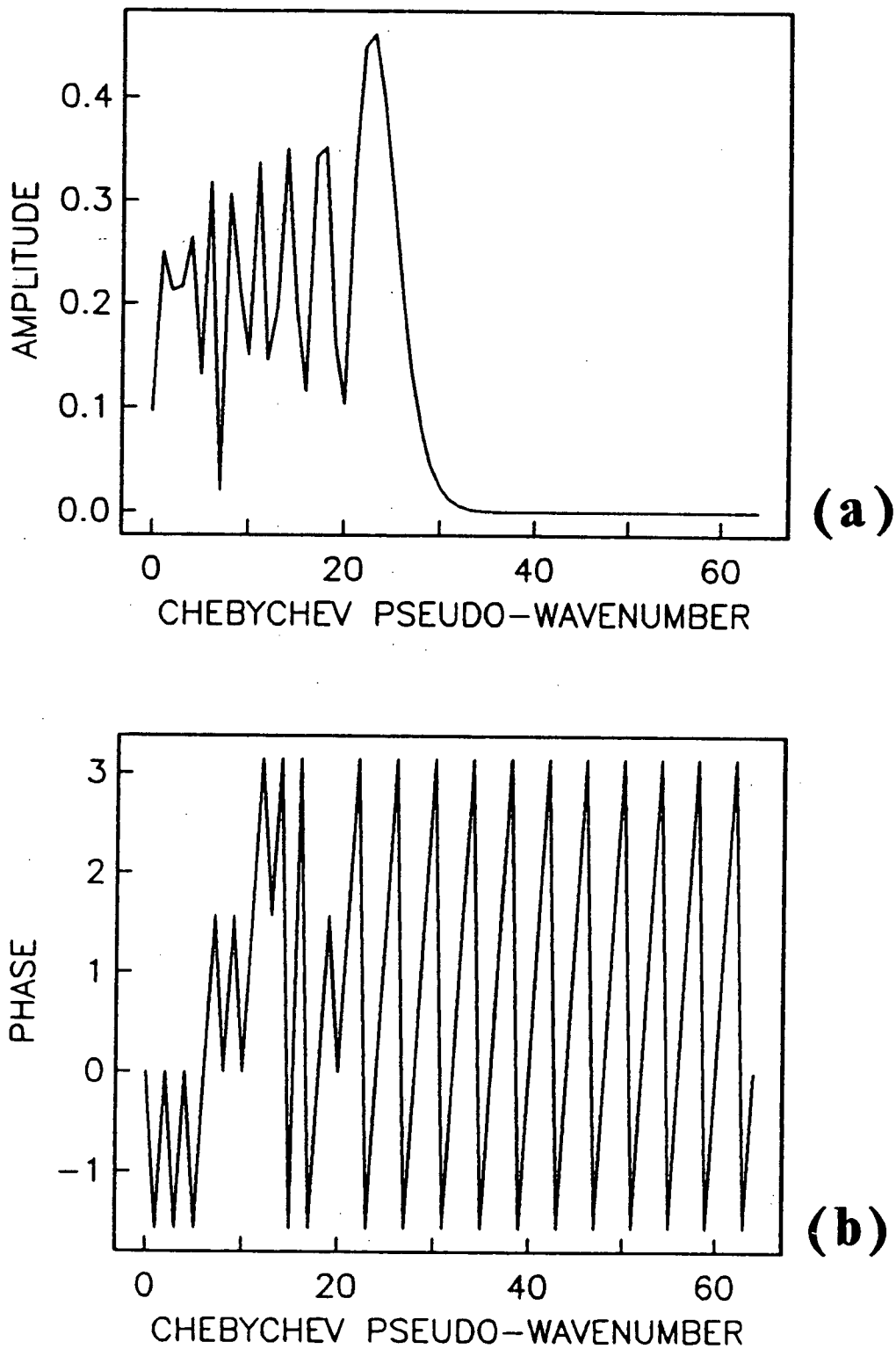

Figure 5.2 *(a)* Amplitude and (b) phase spectrum of the true Chebychev transform of the unmodulated signal.

5.6.2.2 The modulated signal

The amplitude modulation of $\exp(ik_0x)$ with the Gaussian $\exp(-x^2/2\sigma_0^2)$ affects both the Fourier and the Chebychev spectra in a very definite manner.

The Fourier coefficients, i.e

$$a(k) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \exp\left(-\frac{x^2}{2\sigma_0^2}\right)\exp(ik_0x)\exp(ikx)dx \tag{5.44}$$

can be computed analytically and they read

$$a(k) = \frac{\sigma_0}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\sigma_0^2(k-k_0)^2\right] \tag{5.45}$$

(Tomonaga, 1966). The modulation in the x-domain results in a roughly equivalent modulation in the k-domain, where the amplitude peaks at $k = k_0$, while decaying fast around it; the amplitude at $k = k_0 \pm (1/\sigma_0)$ is 60% of its maximum value (figure 5.1b), only.

The Chebychev spectrum of the modulated wave does not lend itself to an obvious analytical evaluation and thus we have to resort to the familiar numerical quadrature expressed by the $F.C.T$ algorithm. To avoid a biasing of our comparisons, we compute the numerical counterpart of the unmodulated spectrum (figure 5.3); an error analysis between the numerical and the analytic spectra indicates the sampling density needed for a satisfactory quality level of the numerical approximate.

An $\bar{L}_2$ of 0.36 (-09), and a corresponding $\bar{L}_\infty$ of 0.18 (-4) signal that the resolution border has been virtually reached at $N = 64$; the excellent quality of this numerical approximate is clearly seen by comparing figures (5.2-3). Furthermore, figure (5.4)

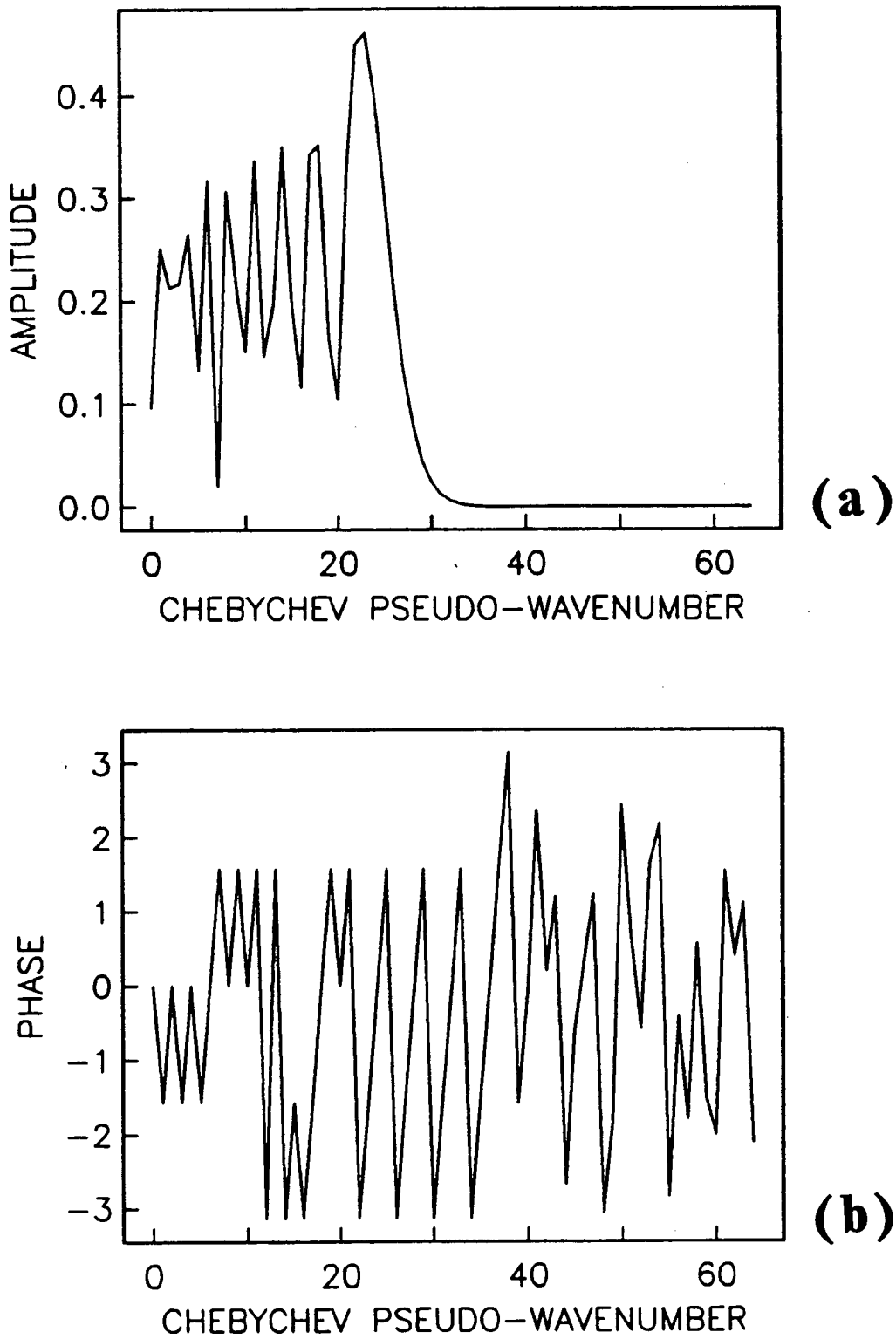

Figure 5.3 *(a)* Amplitude and *(b)* phase of the discrete Chebychev transform of the unmodulated signal; 65 samples have been employed.

displays the instantaneous relative errors; thus it unveils the gross inaccuracies of the high-order and less significant coefficients due to truncation error. However, the wide spread of the coefficients, combined with the reality of the round-off, would nullify any improvement of truncation error to be obtained in finer grids.

The complications arising in measuring and comparing the phases may be exemplified in this ideal case. The phase of the difference vector is plotted in figure in (5.5a), while the differences of the individual phases are depicted in figure (5.5b). The latter reveals instantaneous phase discrepancies, that are either 0 or 2π in the significant part of the spectrum. This behavior is a rather innocuous consequence of the fact, that the purely real, even, true coefficients are approximated by elements incorporating minor imaginary contributions with a negative sign; approaching the branch cut from the third quadrant, results in a phase of $-\pi$, instead of the correct π. The phase of the trivial part of the spectrum is totally meaningless; this insignificance carries on, to characterize figure (5.5a) in a global scale. The very small size of the vector elements and the"noisy" distribution of signs, results in a failure to provide interpretable information.

As we have established a satisfactory amount of faith in the the numerical spectrum of the unmodulated signal, we proceed to compute the spectrum after the signal's modulation (figure 5.6). This computation reveals a spectacular, somewhat similar to its Fourier analog, modulation of the Chebychev spectrum, around $k_0 = 25$, as well. Minor departures, from the typical Gaussian shape, may be noticed at the low order modes, whereas an appreciable amount of energy has been transferred to higher modes (figure 5.6a). The spectral coefficients displayed have been obtained from an $N = 128$ discretization. Noise prevails after $n \simeq 75 - 80$ (an identical observation may

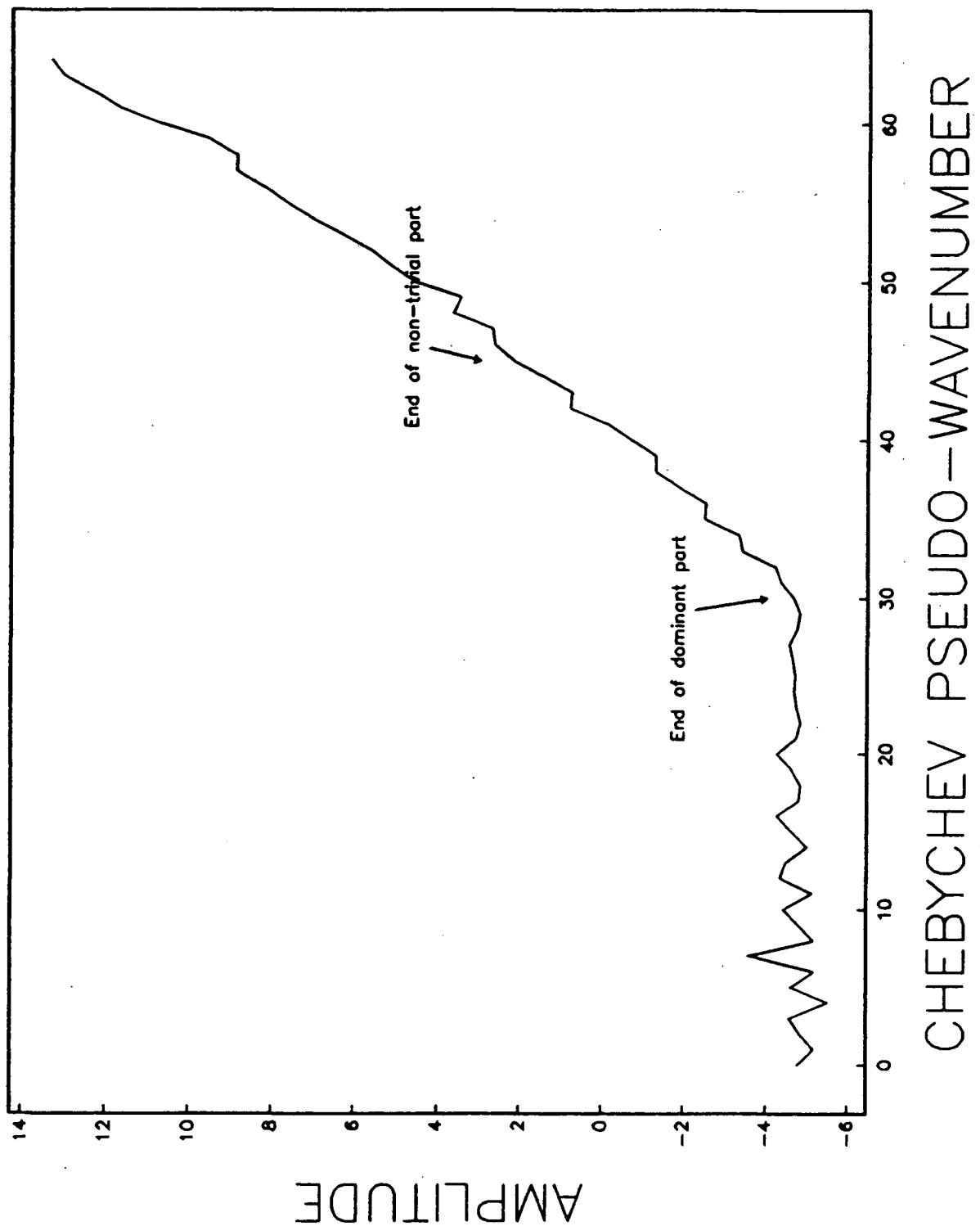

Figure 5.4 Amplitude spectrum of the error vector, between the analytic and the discrete Chebychev transform of the unmodulated signal, normalized with respect to the local magnitude of the true spectrum; 65 coefficients have been considered.

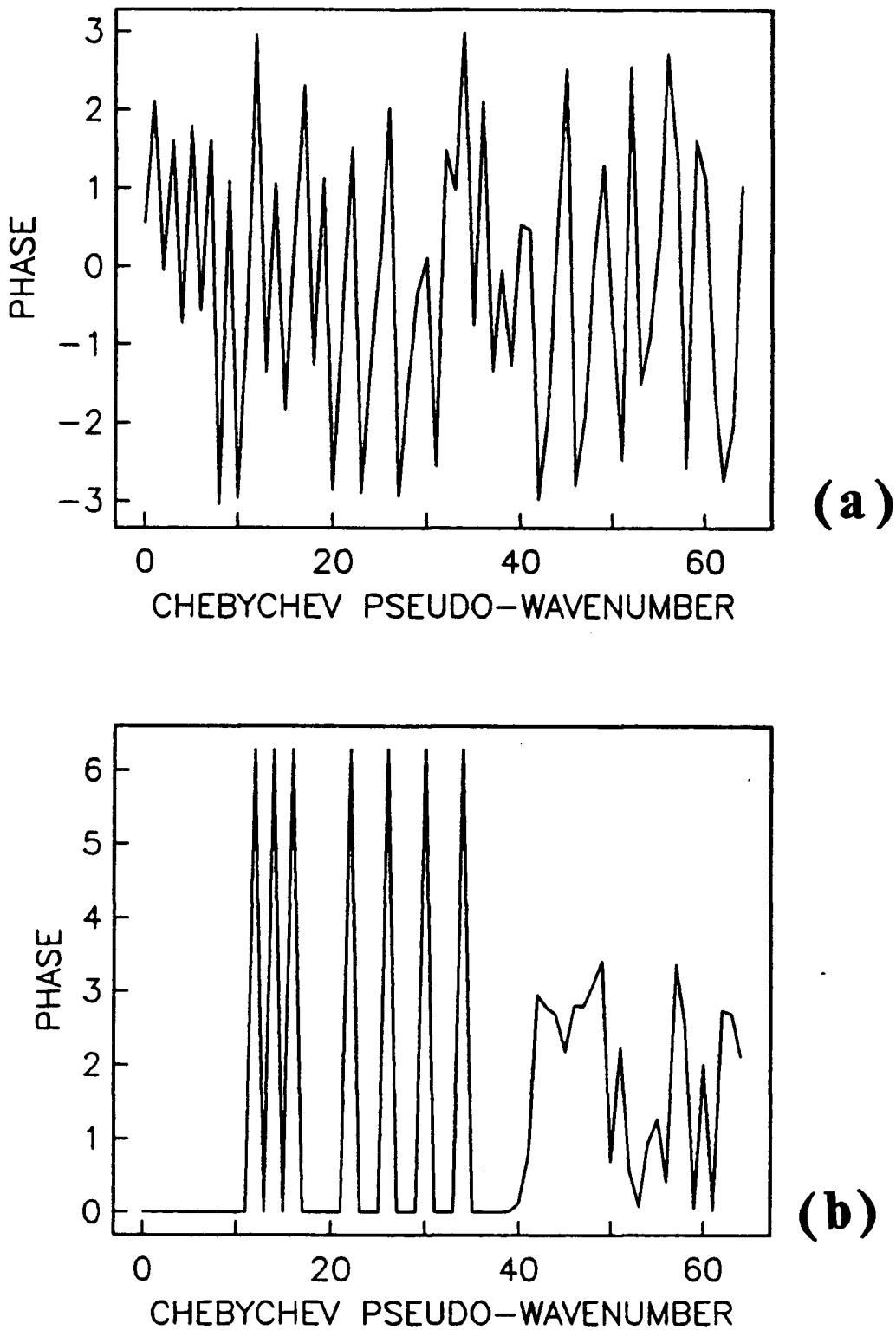

Figure 5.5 *(a)* Phase spectrum of the error vector between the analytic and the discrete Chebychev transform of the unmodulated signal and *(b)* error vector between the analytic and the discrete Chebychev phase spectrum of the unmodulated signal; both graphs involve 65-long vectors.

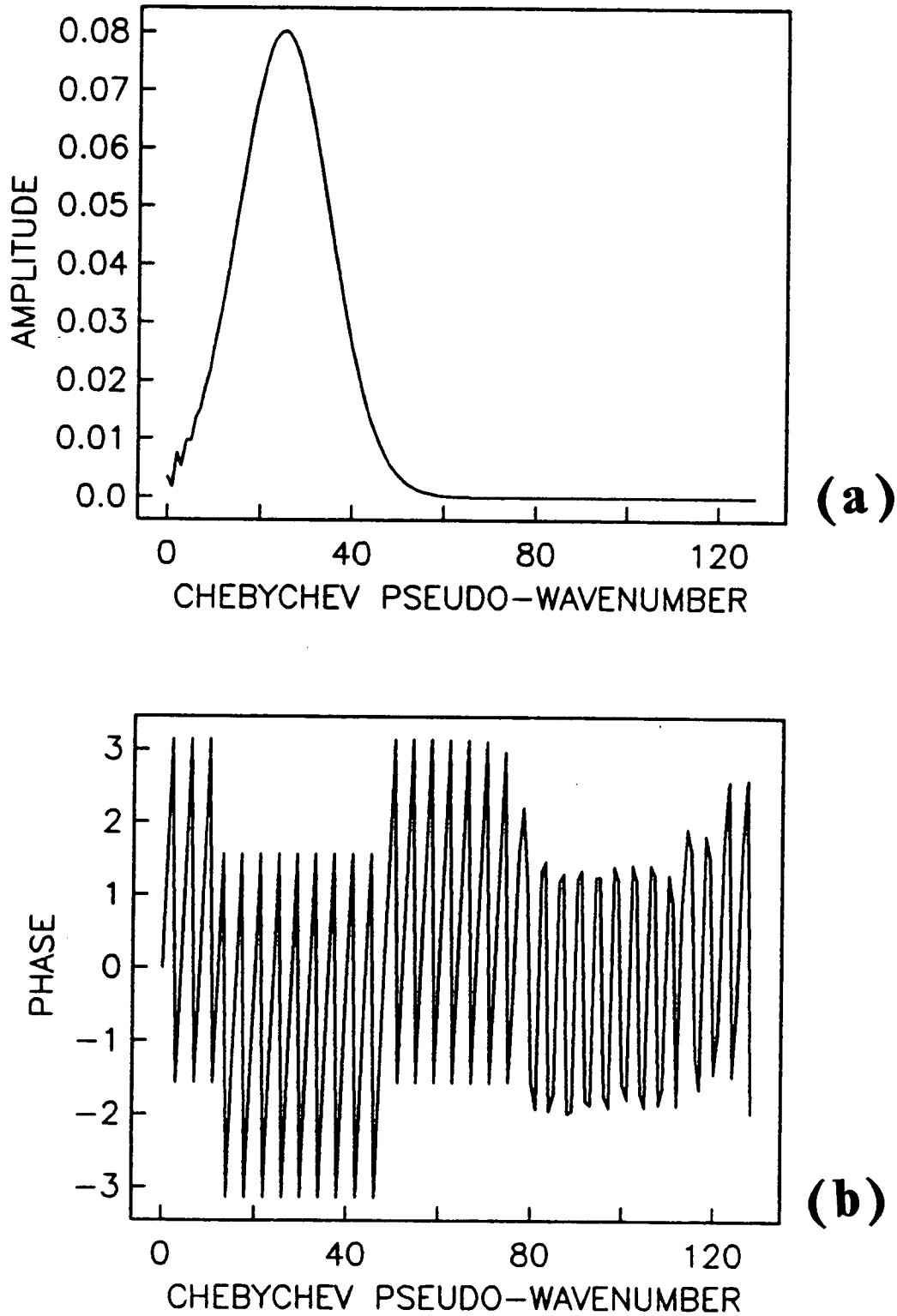

Figure 5.6 *(a)* Amplitude and *(b)* phase discrete Chebychev spectrum of the modulated signal.

be made for the Fourier spectrum), while the magnitude of the last $10-30$ resolvable coefficients is significantly smaller than the magnitudes of the lower ones.

Calculation of the post-modulation spectrum with $N = 64$ verifies an anticipated small but definite amount of aliasing, due to the inadequate representation of the non-trivial higher modes. The aliasing is especially pronounced in the less dominant tail of the spectrum; doubling the density of the x-sampling, we eliminate aliasing and improve the truncation error. The trailing coefficients are expected to still suffer from inaccuracies due to truncation errors, but the minor contribution of these higher-order modes to the representation of the function, combined with the possibility of a pronounced round-off accumulation, in both the transform calculation and the subsequent numerical solution procedures, in denser grids, renders further resolution attempts hopeless.

Another spectacular result, with respect to the phase characteristics of the modulated spectrum, is obtained; the instantaneous phase is plotted in figure (5.6b). The phase of the dominant Gaussian part of the spectrum oscillates between $-\pi$ and $\pi/2$, whereas the phase of the less significant Gaussian flanks oscillates between $-\pi/2$ and π; this phase modulation-like behavior persists, to a remarkable extent, in the trivial part of the spectrum, as well.

5.6.3 Discussion of results

Tables 5.1 and 5.2 display the $\bar{L}_2$ and the $\bar{L}_\infty$ computed estimates for all the Chebychev and the finite difference methods presented previously. The Chebychev values of these tables refer to a spectrum computed directly from an equal available number of x-samples; no appreciable differences between the Chebychev variants exist

and the inferiority of the classic finite difference scheme is readily identified. Table 5.3 is concerned with the errors that correspond to spectra obtained via successive truncations of a spectrum computed originally from $N = 128$ x-samples; only the Galerkin results are given.

$N+1$	*GAL*	*PSD*	*TAU*	*TIN*	*FD*
5	0.37 (+2)	0.37 (+2)	0.36 (+2)	0.37 (+2)	0.19 (+2)
9	0.41 (+1)	0.43 (+1)	0.35 (+1)	0.45 (+1)	0.25 (+1)
17	0.20 (+1)	0.20 (+1)	0.20 (+1)	0.21 (+1)	0.20 (+1)
33	0.31 (0)	0.33 (0)	0.35 (0)	0.35 (0)	0.90 (0)
65	0.89 (-3)	0.89 (-3)	0.89 (-3)	0.89 (-3)	0.11 (0)
129	—	—	—	—	0.77 (-2)
257	—	—	—	—	0.48 (-3)

Table 5.1 $\bar{L}_2$ values for the various numerical solutions of the Schrödinger equation; results for the Galerkin, pseudospectral, tau differentiated and tau integrated Chebychev schemes are displayed versus results for the classic finite difference scheme; $\Delta t = 1/N^2$, while Chebychev estimates for $N > 64$ have not been computed due to the excessive computational cost involved.

$N+1$	*GAL*	*PSD*	*TAU*	*TIN*	*FD*
5	0.61 (+1)	0.61 (+1)	0.59 (+1)	0.61 (+1)	0.61 (+1)
9	0.19 (+1)	0.20 (+1)	0.17 (+1)	0.20 (+1)	0.12 (+1)
17	0.12 (+1)	0120 (+1)	0.11 (+1)	0.12 (+1)	0.10 (+1)
33	0.29 (0)	0.35 (0)	0.31 (0)	0.31 (0)	0.87 (0)
65	0.29 (-1)	0.29 (-1)	0.29 (-1)	0.29 (-1)	0.35 (0)
129	—	—	—	—	0.88 (-1)
257	—	—	—	—	0.22 (-1)

Table 5.2 $\bar{L}_\infty$ values for the various numerical solutions of the Schrödinger equation; results for the Galerkin, pseudospectral, tau differentiated and tau integrated Chebychev schemes are displayed versus results for the classic finite difference scheme; $\Delta t = 1/N^2$, while Chebychev estimates for $N > 64$ have not been computed due to the excessive computational cost involved.

$N+1$	$\bar{L}_2$	$\bar{L}_\infty$
5	0.11 (+1)	0.10 (+1)
9	0.10 (+1)	0.10 (+1)
17	0.93 (0)	0.83 (0)
33	0.20 (0)	0.26 (0)
65	0.89 (-3)	0.29 (-1)

Table 5.3 $\bar{L}_2$ and $\bar{L}_\infty$ values for the Chebychev Galerkin solution of the Schrödinger equation; $N+1$ corresponds to the spectral cut-off level of an original $N = 128$ spectrum, fed into the system.

Finite differences are severely aliased up to N=16, while an N=32 discretization covers virtually all the spatial wavenumbers of primary significance (figure 5.1b). The Chebychev transform is heavily aliased up to N=32; nonetheless, a sampling density corresponding to N=64 allows most of the dominant spatial information to be transferred into the Chebychev domain. An investigation of the errors given in tables (5.1-2), in conjuction with a detailed study of corresponding graphs of the amplitude of the approximate solution plotted versus the exact wavefunction, reveals some important characteristics of the numerical schemes.

The limited local character of the finite differences is associated with the fact that the classic scheme (irrespective of being aliased or not) preserves the Gaussian shape of the true solution; nevertheless, it is the very same intrinsic feature of the finite difference scheme that obstructs its convergence; artificial spatial dispersion is injected into the system, the group velocity is underestimated and the numerical solution is superseded by the analytic solution, the later being centerred at $x = +0.25$ (figure 5.7). An aliased Chebychev spectrum (for N=32) yields some unexpected results, as the biased spectrum (figure 5.8a—solid line) nullifies the *spreading-control* analysis (recall discussion in 5.4.4.2). A revealing demonstration of this (with $\Delta t = 1/32^2$) is presented in figures 5.8b, 5.9a and 5.9b (dashed lines), where the Galerkin, the

pseudospectral and the tau solutions are displayed versus the analytic solution (solid line).

The time-advanced numerical solution exhibits translucent evidence of appreciable boundary reflections; it may be contemplated that it is the altered dispersion relation (due to time differencing), which gives rise to this artificial spreading, but numerical experiments, employing much reduced time steps, exhibited similar reflections.

This result points to the real source of problems, namely, the propagation of an aliased spectrum, which definitely cannot account properly for the Gaussian shape of the initial condition. The spectrum incorporates barely 60% of the spatial information with its higher part being rich in aliasing contamination; its great departure from the correct Gaussian-like shape is clearly noticeble in figure 5.8a. This misrepresentation manifests itself with a spreading factor much larger than anticipated; finite difference stencils apply in the x-space instead and therefore their implicit spectral aliasing does not transmit this misrepresentation in a global manner. The central bulk of the Chebychev spread-out Gaussian may, nevertheless, claim a smaller shift from the exact solution than in finite differences.

A deeper understanding of the reported anomaly may be gained by repeating the experiments with an initial spectrum consisting of a truncated portion of the non-aliased, more-accurate and complete spectrum corresponding to $N = 128$ (figure 5.8a—dotted line). These filtered versions of the previous experiments display a rather ambivalent character. We do see (dotted lines in figures 5.8b and 5.9) that elimination of aliasing (decreased truncation error is of secondary importance) smooths out the reflections at the left boundary significantly, but it does not succeed in accomplishing a similar operation at the right boundary. The answer lies in the inability of the

filtered spectrum to satisfy the boundary conditions; the latter are not homogeneous any more and their effect may be envisioned as either an inhomogeneity factor not being accounted for or equivalently a Gibbs' phenomenon at both boundaries due to the non-uniform convergence of the expansion at these points. The generated errors are, thereafter, gradually moving inwards affecting the rest of the solution; at the left boundary, the major improvement due to the non-aliased character of the new spectrum overwhelms the deterioration introduced by the incomplete satisfaction of the boundary condition at $x = -1$. The contamination is, nevertheless, visible and it is especially pronounced at the right boundary, the boundary $x = +1$ being much closer to the Gaussian; the Gibbs' oscillation interferes much faster with the Gaussian's right flank and the final outcome can not claim an ameliorated simulation of the physical process. Major reductions on the size of Δt failed to improve this result, again. However, a comment on the improvement observed with respect to the alignment of main bulk of the numerical solution with the true Gaussian should be made. Furthermore, we should note the relatively reduced Gibbs' phenomenon in the tau variant (dotted line in figure 5.9b) compared to the size of the phenomenon in the Galerkin and the pseudospectral simulations (dotted lines in figures 5.8b and 5.9a, respectively). The stronger emphasis of the boundary conditions that characterizes the former seems to decelerate the ill-conceived Gibbs' radiation.

Introducing 65 coefficients in the Chebychev spectrum has dramatic consequences. Most of the spatial information having been fed in, the algorithm produces a considerably improved output (figure 5.10a). Identical results for its filtered version reflect the counterbalancing effect of the limited aliasing of the direct spectrum with the weak Gibbs' phenomenon accompanying the use of the truncated spectrum since

only a minor amount of energy lies beyong the current cut-off. Additionally, the improved truncation error of the latter cannot be resolved because of the accumulation of round-off during the process. The reported improvement in the quality of the numerical simulation is also partly due to the decreased time step (Δt goes as $1/N^2$) and it is obvious that further significant decrease of the error is coupled with a decrease in the size of Δt. In theory, maximum accuracy demands retaining 75-80 coefficients; the practical advantages are doubtful though, especially if we recall that the accuracy of that additional higher portion of the numerical spectrum is severely damaged by truncation error.

The Fourier spectrum of the initial condition extends up to $k = 75 - 80$, but it is readily seen that the energy beyond $k = 50$ is of secondary importance for all practical purposes, although (maximum attainable resolution would require the whole non-trivial spectrum). The finite difference scheme appears to converge much more slowly than its Chebychev rival, despite it being slightly aliased at $N = 32$ and aliasing-free at $N = 64$.

Accuracy loss is also coupled with time-differencing, but comparison between the Chebychev and the classic scheme for the same Δt size demonstrates clearly the inferior quality of the classic scheme. This trend is hardly astonishing because of formidable amounts of artificial spatial dispersion still present at these discretization levels, since $(k_{\max}\Delta x)^{-1} = N/100$ for our problem. The anticipated improvement of the classic scheme's performance with an oversampling in the x-direction seems to match the Chebychev accuracy at $N = 256$. However, this is misleading since the latter has a time step 16 times larger than the former does; repeating the Chebychev experiment with $N = 64$ and the time step of the finite difference scheme of $N = 256$ (figure 5.10b),

yields an $\bar{L}_2$ of 0.60 (-5) and an $\bar{L}_\infty$ of 0.18 (-2), indicating that finite difference would need a whole order of magnitude more points in order to attain a comparable accuracy level, i.e an oversampling of $N = 512$ would be needed.

Finite differences do not appear capable of achieving their reported accuracy levels with time steps significantly smaller than the current $\Delta t = 1/N^2$ pattern. Considering $N = 256$, so that only limited amounts of artificial spatial dispersion contaminate the classic scheme, we obtain error estimates that vary considerably for different Δt values. As an example, let us consider $\Delta t = 1/64^2$; repeating the experiment, we compute an $\bar{L}_2$ of 0.11 (0) and an $\bar{L}_\infty$ of 0.35 (0), which constitute a pair of inferior accuracy estimates.

Either scheme is coupled with a low-order finite difference approximation of the time derivative and thus an implicit temporal oversampling is needed to reduce time truncation errors; temporal dispersion is acceptable at $\Delta t = 1/128^2$, i.e $(\omega_{\max}\Delta t)^{-1} \simeq 6$ and it is definetely minimal at $\Delta t = 1/256^2$, i.e $(\omega_{\max}\Delta t)^{-1} \simeq 26$, with a relative phase error (expression 5.34) of 0.1 (-3).

This analysis tends to indicate that the obtained error estimates are greater than it would be anticipated. Part of the explanation lies in the incomplete handling of the short wavelengths, an intrinsic feature of the Crank-Nicolson formulation. A relative instability is coupled with the latter and we thus witness a relative accuracy loss in the course of time. A quantitative idea of the magnitude of the reported loss is given below. The Chebychev Galerkin scheme with $N = 64$ and $\Delta t = 1/256^2$ exhibits $\bar{L}_2$ and $\bar{L}_\infty$ of 0.33 (-9), 0.10 (-4) and 0.58 (-6), 0.56 (-3), for $M = 1$ and $M = 100$ iterations respectively; the error levels at $t = 0.005$ reported above involve an M of 328 iterations, approximately.

Concluding the analysis, we should briefly touch upon two interesting points. The spatial character of our specific initial condition, i.e high concentration in a narrow middle portion of the computational grid, makes an all-grid sampling quite inefficient. The latter is more distressing in the Chebychev case, due to the contradiction between the absence of information near the boundary regions and the high density of the Chebychev nodes there. Memory is thereby wasted, while high boundary resolution is definitely not required in our problem. On the contrary, the boundaries are artificial and there is no desire to approach them; we actually want to make sure that we stay away from them. This enhanced boundary sensitivity has undoubtedly its own share of responsibility for the troublesome reflections characterizing the "incomplete" Chebychev experiments with $N = 32$. The boundary clustered points aid in the the incorrect spreading of the solution; boundary reflections and boundary-generated Gibbs oscillations are accelerated and carried faster in the interior of the computational domain.

A note of the increased vulnerability of the boundary condition imposition in the Chebychev scheme, due to the intervening transform, should be made again.

Furthermore, the dissipation of the solution is bound to be more troublesome in the Chebychev case than in the finite difference scheme, because of the increased number of trivial points at the boundaries' neighborhoods, which would tend to accelerate the onset of resolution loss.

A final comment on the behavior of the absolute errors is worth mentioning. A gradual deterioration of absolute accuracy is observed in the course of time. This deterioration is of a mild character; this is due to the fact that the amplification factor of the Crank-Nicolson formulation is exactly unity. Consequently, the round-off

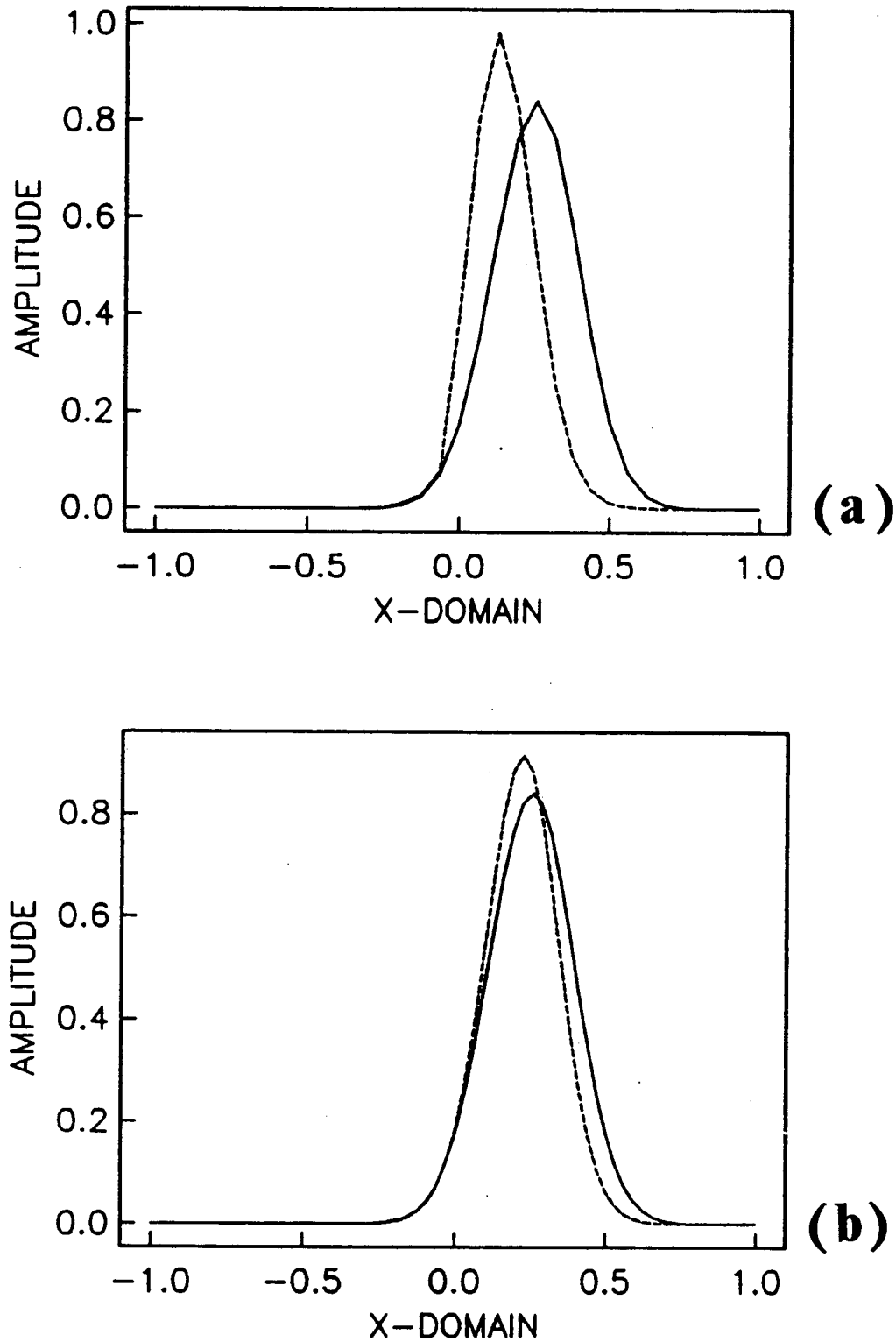

Figure 5.7 The finite difference (dashed line) versus the analytic (solid line) solution; (a) N=32 and (b) N=64. The time step is $1/N^2$.

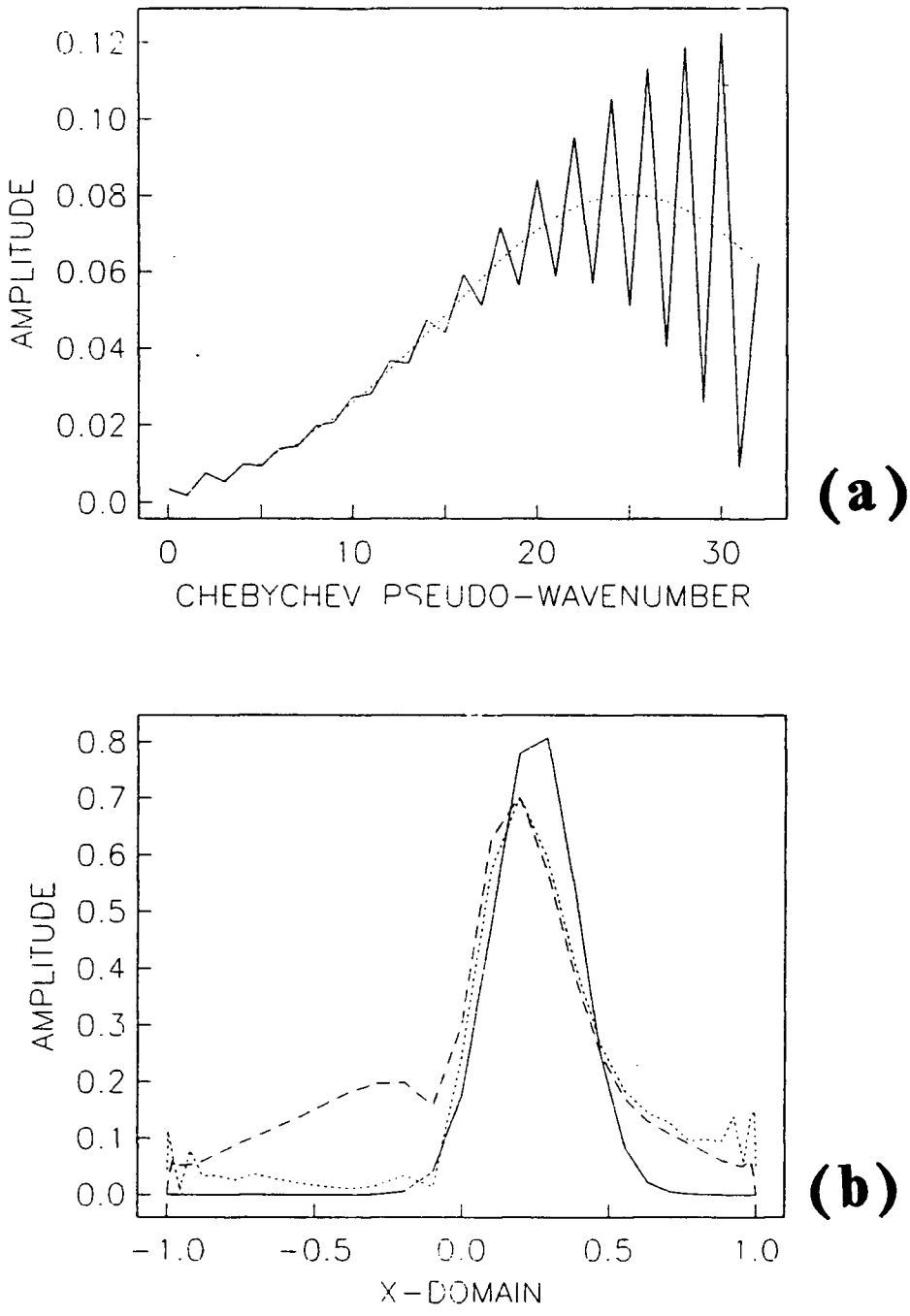

Figure 5.8 (a) Chebychev spectrum for $N = 32$; both the aliased (dashed line) and the truncated (dotted line) spectra are given. (b) The aliased (dashed line) and the truncated (dotted line) Chebychev Galerkin solution with $N = 32$ and $\Delta t = 1/32^2$ versus the analytic (solid line) solution.

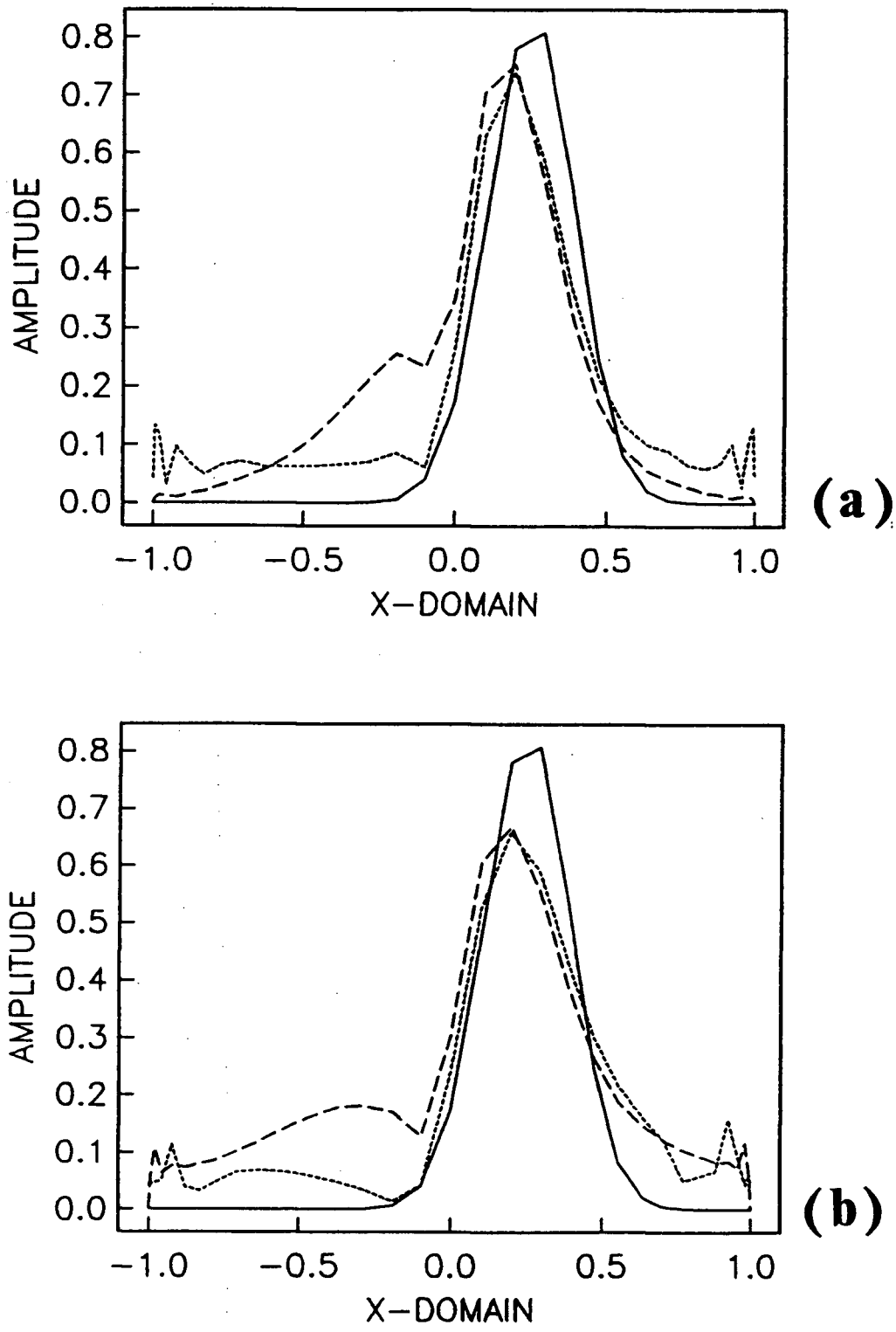

Figure 5.9 (a) The aliased (dashed line) and the truncated (dotted line) Chebychev pseudospectral solution versus the analytic (solid line) solution. (b) The aliased (dashed line) and the truncated (dotted line) tau solution versus the analytic (solid line) solution; Δt and N are the same as in figure (5.8b).

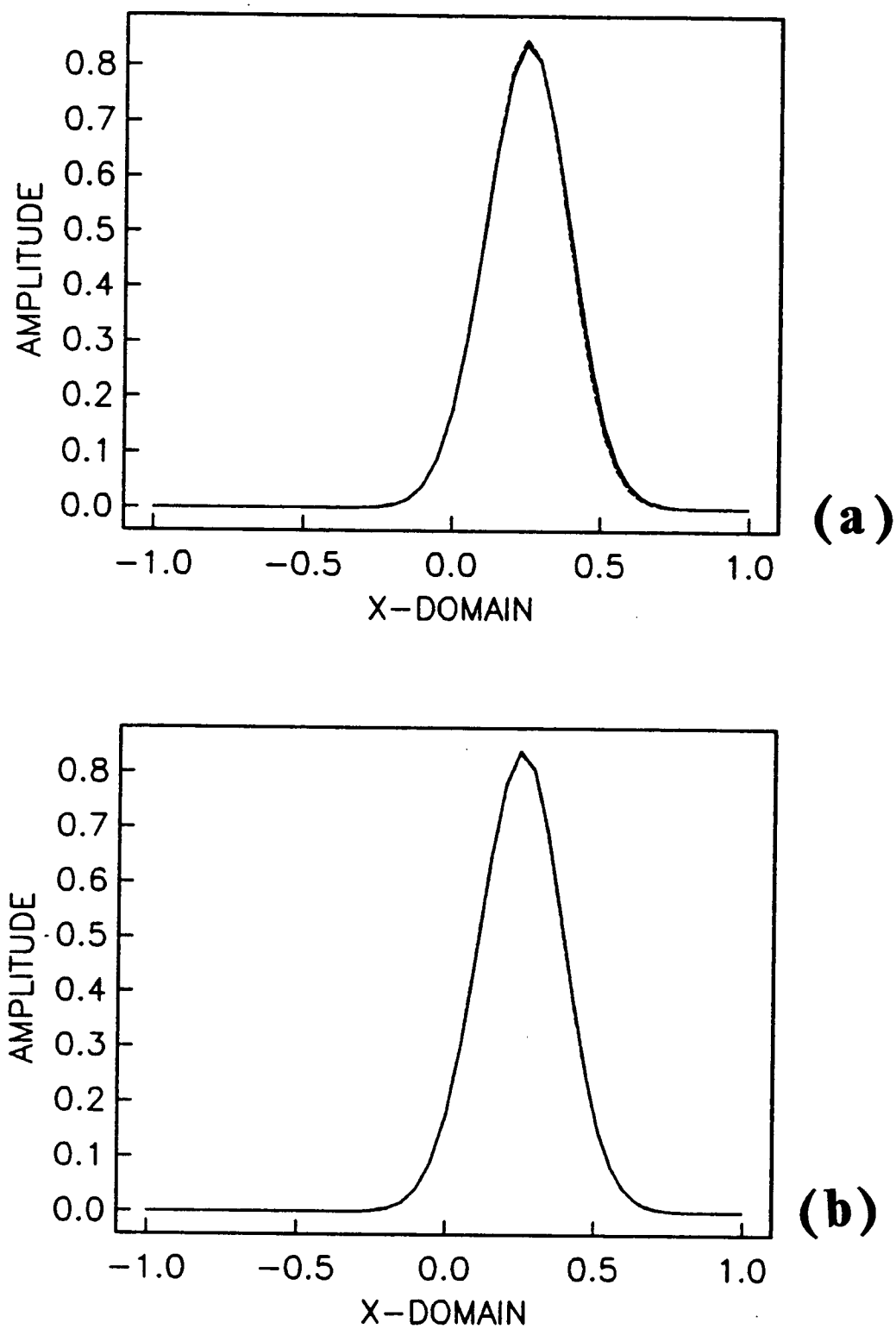

Figure 5.10 The Chebychev Galerkin (dashed line) versus the analytic (solid line) solution; $N = 64$ with (a) $\Delta t = 1/64^2$ and (b) $\Delta t = 1/256^2$.

generated at each iteration (and all other errors that are present) are transferred over to the next step, being neither amplified, i.e absolute stability, nor getting damped. This is a significant departure from the heat equation; in the latter, the amplification factor is smaller than unity and therefore, there exists a time after which the error gets successfully damped.

5.7 The Fast Complex Integrated Tau Solver.

Procedures **SLU1-3** (see Appendix B.2) can be readily extended to complex arithmetic, in order to deal with the complex diffusitivity parameter of the Schrödinger problem.

It is immediately seen that the Schrödinger system exhibits an essential new structural characteristic, namely, a complex middle diagonal. The latter marks a significant departure from the heat equation and necessitates a careful analysis of the system's structure. In the numerical environment simulated by the "super-fast" **SLU1** algorithm, we recognize a high degree of off-diagonal dominance. The main diagonal is $O(1/4n^2)$, while the off-diagonal contribution is $O(|\imath\Delta t/2 + 1/2n^2| + 1/4n^2)$. The reported off-diagonal dominance is more pronounced here than in the heat equation (see 4.4.7). In addition, it is obvious that increases in Δt worsen the situation; a decreased Δt improves the situation, but unfortunately off-diagonal dominance persists even in the limit of $\Delta t \rightarrow 0$.

Our intuition has been confirmed by a number of numerical experiments. Under the familiar choice $\Delta t = 1/N^2$, experiments have shown, that, for $N \leq 8$, the **SLU1**-based solution is identical to the solution obtained via a straightforward Gaussian elimination

with partial pivoting. Mild deterioration characterizes $N = 16$, i.e 4-5 significant figures are now accurate, to be continued as N reaches 32 or 64; the latter choices yield results accurate to 4 and 3-4 digits, respectively. The accuracy is severely damaged at $N = 128$ and inevitably, near-zero pivots annihilate the approximation at $N = 256$. The choice $\Delta t = 1/N$, for systems with $N > 16$, has yielded meaningless answers, due to division by near-zero pivots. On the other hand, the choice $\Delta t = 1/N^3$ has resulted in much improved answers, restoring the accuracy to 4-5 figures for systems up to $N = 128$.

What about procedures **SLU2** and **SLU3**? Significant efficiency gains in the solution of any quasi-tridiagonal system are achieved only when the **SLU1** algorithm is applied. The **SLU3** algorithm is equivalent to a complete LU decomposition with partial pivoting and it is to be used in the lack of any alternatives, only. Finally, the **SLU2** procedure exhibits an efficiency gain which is important only on theoretical grounds. The analysis of the Schrödinger tau-integrated systems also serves as a predecessor for the incresed demands of the migration problem of the next chapter. However, the **SLU2** algorithm cannot achieve but a minor reduction of the computational cost in the repetitive inversions involved in the migration downward continuation process. Thus, we have restricted ourselves to a theoretical investigation with respect to the anticipated performance level of the technique; numerical experiments have not been conducted. The diagonal shift, inherent in the **SLU2** original system, gives rise to a diagonally-dominant system (the structure is also superior to the equivalent system of the heat equation). The diagonal vigor of the system is easily identified; the main diagonal features elements of $O(|\imath\Delta t/2 + 1/2n^2|)$, whereas the off-diagonals' strength

is just $O(1/2n^2)$. This quarantees a performance of high quality (see 4.4.7). Furthermore, increasing Δt enhances the diagonal dominance of the system; the limit $\Delta t \to 0$ corresponds to neutrally-dominant systems.

CHAPTER VI

THE PARABOLIC EQUATION

Then Myskin stretched out his trembling hand to him
and softly touched his head, his hair, stroking them
and stroking his cheeks ... he could do nothing else!
The Idiot — Fyodor Dostoyevsky

6.1 The Paraxial Approximation in Exploration Geophysics

The application of the paraxial approximation (see below), in order to derive one-way equations, results in a class of equations known as the *parabolic approximation.* This equation is often employed when the desired propagation is at a "small" angle to a preferred direction. The main advantage is that instead of having to solve the elliptic Helmholtz equation (boundary value problem), we only need to solve a parabolic equation (initial value problem) which is easier to handle numerically. Applications of the parabolic equation abound and they cover a wide range of diverse fields including

underwater wave propagation, laser beam propagation, quantum mechanics, electromagnetic diffraction and propagation, plasma physics, optical waves and, of course, seismic waves (Mary and Lee, 1985).

The parabolic wave equation was originally proposed in geophysics by Claerbout (1976) and it was introduced in an effort to expand the existing vertical-incidence theory to include some angular bandwidth around vertical-incidence. A detailed derivation of the equation may be found in Claerbout (1976, 1985); it basically involves various approximations to the square root $\sqrt{k^2 - k_x^2}$, the latter corresponding to the vertical wavenumber k_z, as may readily be seen by inspection of the dispersion relation of the full wave equation, i.e

$$k_z^2 + k_x^2 = \frac{\omega^2}{v^2} \tag{6.1}$$

The 15° equation corresponds to a linear approximation, i.e retaining the first two terms of its Taylor expansion, while a bilinear approximation yields the 45° equation. Muir's recursion formula, i.e

$$R_{n+1} = 1 - \frac{X^2}{1 + R_n} \tag{6.2}$$

can be used to obtain higher order approximations (Claerbout, 1985); the phase shift method (Gazdag, 1978) implements the full square root operator and it is capable of migrating dips up to 90°.

6.2 The 15° Migration Equation

6.2.1 Inherent limitations

The 15° equation has extensively been used in the migration of seismic stacked sections, since, due to the assumptions made during its derivation, it enjoys characteristic computational advantages. These include an easy handling of the numerical aspects of the equation and an accurate migration of dips up to 15° for a velocity function varying slowly with depth, i.e the vertical component of the velocity gradient $\partial U/\partial z$ is assumed to be small.

6.2.2 Formulation for a CMP gather

For the purposes of this study, we only consider the migration of *zero offset CMP gathers*; the *exploding reflector model (ERM)* is employed for the modeling process. The equations may be set up in the regular stationary coordinate frame (x, z, t) with the resulting migration scheme known as the (x, z, t) migration. Alternatively, *time retardation* is customarily invoked; the relevant equations then involve a *diffraction* term

$$\frac{\partial^2 U}{\partial z \partial t} = -\frac{v}{2}\frac{\partial^2 U}{\partial x^2} \tag{6.3}$$

and a *thin lens* term

$$\frac{\partial U}{\partial z} = \frac{1}{v}\frac{\partial U}{\partial t} \tag{6.4}$$

where x, z and t denote half offset, midpoint and 2-way travel time respectively. The velocity $v = v(x, z)$ is half the true medium velocity to account for the ERM hypothesis and U is the upgoing wavefield since we are interested in migration than

in diffraction, instead. The *Fresnel* diffraction term accounts for the lateral diffusion of the waves, while the thin-lens term performs the depth extrapolation. The time-retarded migration equations are traditionally solved in the frequency domain; Fourier transformation in time, yields

$$-i\omega\frac{\partial U}{\partial z} = -\frac{v}{2}\frac{\partial^2 U}{\partial x^2} \tag{6.5}$$

or

$$\frac{\partial U}{\partial z} = \frac{v}{2i\omega}\frac{\partial^2 U}{\partial x^2} \tag{6.6}$$

for the diffraction term (6.3), while the thin lens term reads

$$\frac{\partial U}{\partial z} = -i\frac{\omega}{v}U \tag{6.7}$$

The full equation may be then written as

$$\frac{\partial U}{\partial z} = -i\frac{\omega}{v}U + \frac{v}{2i\omega}\frac{\partial^2 U}{\partial x^2} \tag{6.8}$$

6.2.3 Variants of the 15° ω-migration

The time dependence having been transformed, we are only now being confronted with a number of monochromatic wave equations. The procedure involves a separate depth extrapolation of each of the available temporal frequencies, to be followed by the appropriate superposition according to the imaging principle.

6.2.3.1 The finite difference $(\omega - x)$-migration

An *splitting* method is customarily employed for the solution of the full extrapolation equation (6.8). That has the advantage of incorporating an exact analytic solution for the thin lens component, while the familiar second order accurate Crank-Nicolson finite difference configuration is used for the solution of the diffraction component. No stability problems arise since each component is stable by itself. The splitting procedure introduces an additional error in the numerical solution; this error is zero if the two operators commute and in such a case a full separation is possible (Brown, 1983). The thin lens and the diffraction operators do commute for horizontally stratified media, i.e v=v(z) only, but lateral velocity variations introduce a noncommutativity which increases with the lateral inhomogeneity. Nevertheless, the finite difference formulation allows a straightforward accommodation of horizontal velocity changes.

6.2.3.2 The frequency-wavenumber $(\omega - k_x)$ migration

In the presence of lateral inhomogeneities this approach is not applicable; otherwise, Fourier transformation of the x-coordinate into the corresponding k_x axis, allows the complete extrapolation process to be performed via a single *phase shift*, manifesting the separability of the components of the extrapolation in such a case. This is easily deduced from the transformed version of (6.8), i.e

$$\frac{\partial U}{\partial z} = -i\left(\frac{\omega}{v} - \frac{vk_x^2}{2\omega}\right)U \tag{6.9}$$

6.3 Analysis of Migration Parameters

6.3.1 The depth extrapolation step Δz

6.3.1.1. Evanescent aliasing

Finite difference migration proceeds in the x-domain and we, therefore, do not have direct access to the (ω, k_x) plane. Considering each monochromatic wavefield separately, the multiplication of the k_x-transformed wavefields with the shift operator, may be equivalently expressed as the convolution of the wavefield with a certain operator, known as the *spatial wavelet*, the latter and the shift operator constituting a Fourier transform pair, i.e

$$W(x, \omega, \Delta z) \longleftrightarrow \exp(-i\Delta z\sqrt{k^2 - k_x^2}) \tag{6.10}$$

(Berkhout, 1981). There is, thereby, an explicit dependence of the downward continuation process on the extrapolation step. The character and the magnitude of this dependence may be best understood in the (ω, k_x) plane and they are directly associated with the concept of propagation and evanescence. The shift operator, for a given temporal frequency ω, operates as a depropagator of a certain range of horizontal wavenumbers. The vertical wavenumber k_z, i.e

$$k_z = -\sqrt{k^2 - k_x^2} = -(1/v)\sqrt{\omega^2 - (vk_x)^2} \tag{6.11}$$

is considered as *evanescent* if $|vk_x| > |\omega|$ and it attenuates rapidly in the direction of propagation. Introduction of the evanescent modes into the extrapolation process

involving a small Δz step is likely to lead to instability of the scheme. A small Δz might not account for a proper attenuation of these "non-physical" solutions and as the error cumulates with a large number of extrapolation steps, the migrated section becomes absolutely useless. The situation is characteristic of a numerical "blow-up" of the solution. The migration involves the depropagation of upcoming waves and evanescent modes result into growing exponentials; the stability criterion is violated and a blow-up is shortly witnessed. However, the issue involves another factor, namely, the validity of the wave equation-based extrapolation, when a *near-field* case is considered, i.e $k_z \Delta z \ll 1$. The wave equation is basically a *far-field* approximation, i.e $k_z \Delta z \gg 1$, while small depth steps tend to give rise a near-field situation. Conciliation with the described inconsistency imposes an even more stringent lower bound on Δz (Berkhout, 1981). It is common practice to mute out the evanescent area before commencing the migration procedure, nullifing in that way, stability limitations with regard to the lower bound of the extrapolation step.

Migration in the x-domain, e.g finite difference, finite element, spatial deconvolution, does not enjoy a direct access of the (ω, k_x) plane. If evanescent energy is present, the spatial wavelet will experience a severe contamination (evanescent aliasing) and consequently the migration algorithm will be unstable (Nautiyal, 1986). An increase in ω defaults to a bigger value of $k^\star$, that is, the transition point from propagation to evanescence on the k_x-axis. As the evanescent region shrinks, the evanescent contribution for a given Δz diminishes as well. At $\omega^\star = v(k_x)_{\mathrm{NYQ}}$, the transition wavenumber is the Nyquist value itself and therefore, for $\omega > \omega^\star$, the evanescent zone lies beyond the $(k_x)_{\mathrm{NYQ}}$ value.

6.3.1.2 Depth aliasing

The extrapolation step Δz must also satisfy upper bound constraints. The reason is that Δz has to be at least equal to half the minimum wavelength involved in the downward extrapolation. In practice, temporal frequencies are limited by ω_{NYQ} and this imposes an upper bound on the maximum vertical wavenumber allowed to propagate, i.e

$$(k_z)_{\text{NYQ}} \leq \frac{\omega_{\text{NYQ}}}{v} \tag{6.12}$$

Consequently, the restriction becomes

$$\Delta z \leq \frac{\pi}{(k_z)_{\text{NYQ}}} = \frac{\pi v}{\omega_{\text{NYQ}}} \tag{6.13}$$

assuming that all information up to the Nyquist frequency is being used in the extrapolation.

Although, the previous analysis is rigorous for a constant velocity function only, it may be approximately applied even in the case of mild x, z perturbations. If $\partial U/\partial z$ is not adequately small, a more stringent upper bound on Δz must be imposed. This is required in order to compensate for the loss of justification in dropping the $\partial^2 U/\partial x^2$ term when deriving the paraxial equation.

6.3.2 The 15° dispersion relation

The 15° equation is incapable of properly migrating dips that exceed its maximum capacity, namely a dip of 15°. This limitation is clearly identified in the altered dispersion relation obeyed by the equation, i.e

$$k_z = \frac{\omega}{v}\left(1 - \frac{(vk_x)^2}{2\omega^2}\right) \tag{6.14}$$

Expression (6.14) discloses two major characteristics of the 15° equation. First, the equation does not exhibit a frequency dispersive character, maintaining in that respect a very important feature of the full wave equation. Second, the dispersion relation (6.14) divulges a strong anisotropic character, departing from the isotropic semicircle; equation (6.14) is a parabola and its departure from the true semicircle increases with the dip angle vk_x/ω. The modified dispersion relation of the 15° equation gives rise to "dispersion effects" when migrating dips greater than 15°. These effects manifest themselves in a potential separation of the low from the high frequencies, during the downward extrapolation of the wavefield. The errors increase with the dip; in addition, incorrect repositioning, for a certain dip, becomes more pronounced as the frequency increases as well (Hatton et al, 1986). These migration errors degrade the quality of the migrated steeply dipping reflectors; "ghosts" often accompany the true reflectors (Gazdag and Sguazzero, 1984).

The described phenomenon is known as "dispersion"; it owes its existence to the limited dip capability of the 15° equation. Even unaliased data suffer from dispersion effects and the addressed problem is absolutely independent of spatial aliasing and dip reversal, to be examined shortly. Furthermore, another kind of dispersion

consideration arises, namely artificially induced dispersion effects, arising in numerical simulations operating in finite difference environments.

6.3.3 Artificial dispersions

In two-dimensional migration algorithms, two spatial coordinates are considered. Migrating in the ω-domain assumes a harmonic time dependence and subsequently, no temporal artificial temporal dispersion problems are encountered. Both the x and the k_x algorithms employ an exact solution for the thin-lens term. Consequently, no truncation errors are committed and the differential operator is represented exactly. However, the situation becomes more involved for the Fresnel diffraction term. The wavenumber technique involves a Fourier spectral interpolation in x and it is, therefore, free of artificial spatial dispersion. Furthermore, the extrapolation in z can be performed via another simple analytic solution; a numerical approximation for the depth derivative $\partial U/\partial z$ is then avoided. The finite difference approximation of the $\partial^2 U/\partial x^2$ operator creates artificial dispersion in this axis and the lack of an analytic extrapolation solution necessitates a finite difference approximation for the depth derivative, introducing similar problems there as well.

Stability considerations are relaxed by employing the Crank-Nicolson scheme but nevertheless, accuracy demands that Δx and Δz are small enough to account for a minimization of these artificial dispersion problems. In principle, the imaginary diffusitivity $v/2i\omega$ depends on ω; therefore, for a given v, a uniform Δz stepping rate should cover the minimum available temporal component of the complete wavefield. In practice this fine point seems never to be addressed. Increased velocities result in greater values of σ, demanding appropriate reduction in the size of Δz. It is interesting

to note that the stability criterion for the explicit solution of this Schrödinger-like equation corresponds to the Nyquist criterion regarding anti-aliasing considerations in the z-axis, i.e

$$2\Delta z \leq \frac{(\Delta x)^2}{\sigma} \tag{6.15}$$

where the right hand side of the inequality represents the diffusion depth that corresponds to a lateral wavelength of Δx. Elimination of the x-dispersion and simultaneous handling of a $v(x)$ case may be achieved through higher-order schemes in z. This approach allows the use of the Fourier transform for the accurate compuatation of the horizontal spatial derivatives; Gazdag (1980) and Kosloff and Baysal (1983) have proposed a third-order and a fourth-order Runge-Kutta scheme, respectively.

6.3.4 Aliasing in a seismic section

Spatial and temporal aliasing in a seismic section are now examined. A good understanding of the character of aliasing in two dimensions requires the identification of an essential difference in the nature of aliasing between one and two dimensions. The concept of the Nyquist frequency as a folding frequency may often be misleading. This interpretation is indeed valid in one dimension, but it only applies to real input functions, since the amplitude of the spectrum of a complex input function is not symmetric, in general. The introduction of a second dimension has a significant consequence. Despite the fact that aliasing overlap (folding) still occurs on each transform coordinate separately, aliases are decoupled in the (ω, k_x) plane (Hatton et al, 1986). Alternatively, the concept of two-dimensional aliasing as a reflection or a folding operation about the ω and the k_x Nyquist boundaries is incorrect; a wraparound due to an overlap of repeating spectra is a valid description of it (Clement, 1973).

6.3.4.1 Spatial aliasing and dip reversal

Let us assume that reflectors of all dips are present. It is trivial then to see that all ω's beyond ω^* are spatially aliased, that is, the propagating spectral regions not covered by $(k_x)_{\text{NYQ}}$ wrap around and reappear in the negative k_x region. This is well-known as ***spatial aliasing*** and it is directly associated with phenomenon of ***dip-reversal.***

Spatial aliasing varies with the reflector's dip, i.e θ, the latter being measured from the horizontal. As the plane-wave wavefront reaches the surface of the earth, the geophone array records an apparent wavelength λ_x, namely the horizontal component of the true wavelength λ. Spatial aliasing is avoided if

$$\lambda_x \geq 2\Delta x \tag{6.16}$$

according to the Nyquist criterion. Substituting $\lambda_x = \lambda / \sin\theta$ and $\lambda = 2\pi v/\omega$, expression (6.16) becomes

$$\omega_{\text{alias}} \leq \frac{v(k_x)_{\text{NYQ}}}{\sin\theta} \tag{6.17}$$

Expression (6.17) reveals that $\theta = 90°$ corresponds to the worst case presented previously, that is, horizontal rays (vertically dipping reflector). The best case is $\theta = 0°$; this amounts to a flat reflector (vertical arrivals at the geophone group) and then, $\omega_{\text{alias}} \rightarrow \infty$.

In the intermediate region, as θ is gradually reduced from 90° to 0°, the aliased k_x range for a given ω shrinks. The maximum non-aliased frequency ω_{alias} increases as θ is reduced from 90° to 0° or, equivalently, the aliased k_x range for a given ω shrinks, accompanied with an expansion of the available k_z range.

6.3.4.2 Temporal aliasing

In principle *temporal aliasing* can also occur. Nevertheless, *anti-aliasing* techniques may be employed to ensure that there are no temporal frequencies beyond the ω_{NYQ} boundary; otherwise, temporal aliasing occurs and aliased frequencies fold in the negative ω-quadrants. In practice, a $(k_x)_{\text{NYQ}}$ is uniquely specified by the particular geophone spacing; temporal aliasing is then avoided, if

$$\omega_{\text{NYQ}} \leq v(k_x)_{\text{NYQ}} \tag{6.18}$$

Finally, an event can be both spatially and temporally aliased if the above conditions are not met; spatial and temporal wavenumber wraparound then results.

6.3.4.3 Migration of aliased data

Temporal aliasing is usually avoided by means of anti-aliasing filtering, but unfortunately, spatial aliasing is a common phenomenon in seismic data. Spatially aliased data suffer from dip reversals; migration then, band-limiting the reflectors, proceeds repositioning the reversed dips, that is, the aliased frequencies, in an erroneous fashion. Additional dip limitations are introduced by the limited recording time and the limited line extent (Lynn and Deregowski, 1981). If only mild amounts of aliasing are present, migration is still recommended; severely aliased data are likely to result in a migrated section of inferior quality. The danger is enhanced in either high velocity, high noise situations or cases characterized by low velocity, steep dips and wide receiver separation. The migration of dip aliased data is a major topic; a variety of filtering and preprocessing remedies have been proposed in the past, the most famous being the dip

moveout correction (DMO) by Fourier transform (Hale, 1983). This technique exhibits the classic advantages of Fourier techniques over similar finite diffence algorithms, and it does not break down at large offsets and steep dips as the latter do.

6.4 Absorbing Boundary Conditions

Absorbing boundary conditions for the finite difference formulations of the 15° paraxial equation in geophysics, have long been presented (Clayton and Engquist, 1980). These conditions are linear, first order in k_x, stable and local, i.e they are confined to a few traces in the boundaries' neighborhood. Their derivation is based upon the concept of impedance and in differential form, they have as follows

$$U_x + (b/v)U_t = 0 \tag{6.19}$$

$$U_z - bU_x - (a/v)U_t = 0 \tag{6.20}$$

and

$$U_{zt} + cvU_{xz} - bU_{xt} - (a/v)U_{tt} = 0 \tag{6.21}$$

The constants a, b and c are determined by matching the boundary condition at the right or the left boundary, with the right or the left side of the dispersion relation of the equation in the interior of the computational domain. The reflection coefficient R, for a monochromatic plane wave, can be chosen such that reflections are completely suppressed. In general though, we are interested in general wavefields; R is designed to be minimum for the wavenumber band that carries most of the significant energy.

Lastly, the effectiveness of the boundary conditions (6.19-21) depends on the incident angle to the boundary, i.e $\sin^{-1}(vk_x/\omega)$.

6.5 Analysis of an Example of 15° Finite Difference Migration of CMP Data

We are given the surface data, e.g pressure field $P(x, z = 0, t)$ obtained after CMP stacking, where x is the half-offset, z is the midpoint, and t is the two-way travel-time coordinate, respectively on a $P(x_i, t_j)$ grid. The classic finite difference approach of migrating these data according to the the 15° equation can be briefly summarized as follows.

6.5.1 The algorithm

We first Fourier transform the data over time to obtain $P(x, z = 0, \omega)$, that is, to obtain the decomposition in terms of the frequency harmonics. Then, the surface (complex) wavefield for each frequency ω is considered and we deal independently with each one of them, since the problem has been analyzed into a sum of monochromatic problems. Each frequency is propagated downwards a depth Δz via a combined solution of both the diffraction and the thin lens terms. We then inverse Fourier transform the extrapolated monochromatic wavefields and we take the strip corresponding to $t = 0$ to be the migrated wavefield at that depth. This procedure is repeated for each new time step until we reach the desired one. Each time the calculated extrapolated monochromatic wavefields are being fed into the equations in order to advance the solution in depth.

The solution of the diffraction term involves the inversion of a tridiagonal system for each frequency and for every depth step. Additionally, the elements of this matrix

depend in a very straightforward manner on the values of the frequency, the velocity and the boundary conditions. From a computational point of view we do not need to perform an Inverse Fourier transform for all frequencies at each time step, since by simply superimposing all of them we can obtain the desired reflector surface for the specific z value.

6.5.2 Computational details

A program based on an algorithm given by Claerbout (1985) utilizing the above has been written; a number of important computational details are addressed and analyzed in the following lines.

6.5.2.1 The input model

The idealized surface data used are made of broadened half impulses in a triangle form. Broadening implies an absence of high frequencies and therefore, the minimization of artificial dispersions caused by the difference operators.

6.5.2.2 The ω_0 and the ω_{NYQ} frequencies

Neither the DC nor the Nyquist temporal mode is used in the depth extrapolation, each one for different reasons. The DC is excluded because it does not satisfy the equation; it actually satisfies the Laplace's equation. No complications arise since setting it to zero amounts to a mere mean-value removal in the time coordinate. A non-zero Nyquist component indicates aliasing; however, the absence of any energy at the Nyquist does not guarantee an alias-free spectrum. If temporal aliasing is present, a dip reversal in ω is expected with the ω_{NYQ} marking the discontinuity point. While removal of the Nyquist contribution is the least we can do, eliminating this component

has another computational merit. Using the complex conjugate property of the Fourier transform, we can simply superimpose the real parts of the positive frequencies up to but excluding the Nyquist, in order to apply the imaging condition.

6.5.2.3 The $(k_x)_{\mathrm{NYQ}}$ component

The significance of the Nyquist mode in the spatial coordinate is now discussed. The Nyquist waveumber is associated with the discontinuity involved in the dip reversal in k_x; when $k_x < (k_x)_{\mathrm{NYQ}}$ left dips are migrated, but for $k_x > (k_x)_{\mathrm{NYQ}}$ right dips are processed since the aliased wavenumber wraps back into the negative k_x-region. The wavenumber discontinuity causes spurious ringing in the x-coordinate. This is usually reduced by appropriate filtering; Claerbout (1985) gives such a filter, i.e

$$W(k_x\Delta x) = \frac{1+\cos(k_x\Delta x)}{1+0.85\cos(k_x\Delta x)} \tag{6.22}$$

6.5.2.4 Boundary conditions

We proceed to discuss the boundary condition issue. Claerbout obtains his results by imposing zero-slope b.c's (figure 6.1), which are maintained constant for all depths. Figure (6.2) gives the migrated output but for homogeneous boundary conditions.

6.5.2.5 Interpretation of the migrated section

The highly idealized input model helps to obtain a migrated section of high quality. Most of the troublesome factors addressed previously are either absent or virtually negligible. The trouble-free input is aided by a meticulous choice of parameters, so that all sources of problems are practically eliminated. The migrated section features, nevertheless, certain artifacts at large depths values (figure 6.1); these are believed

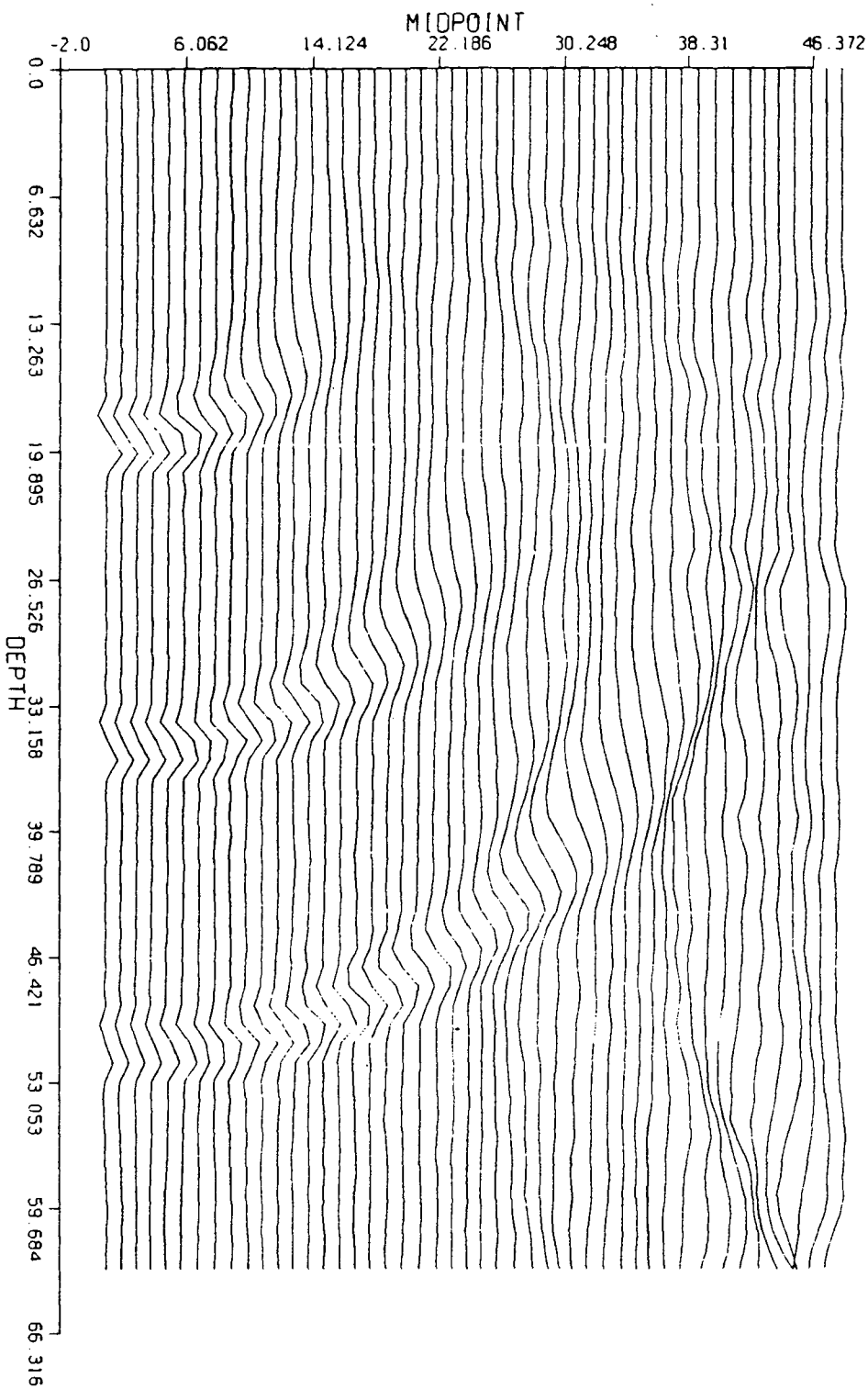

Figure 6.1 Migrated section following Claerbout's (1985) example with zero-slope boundary conditions.

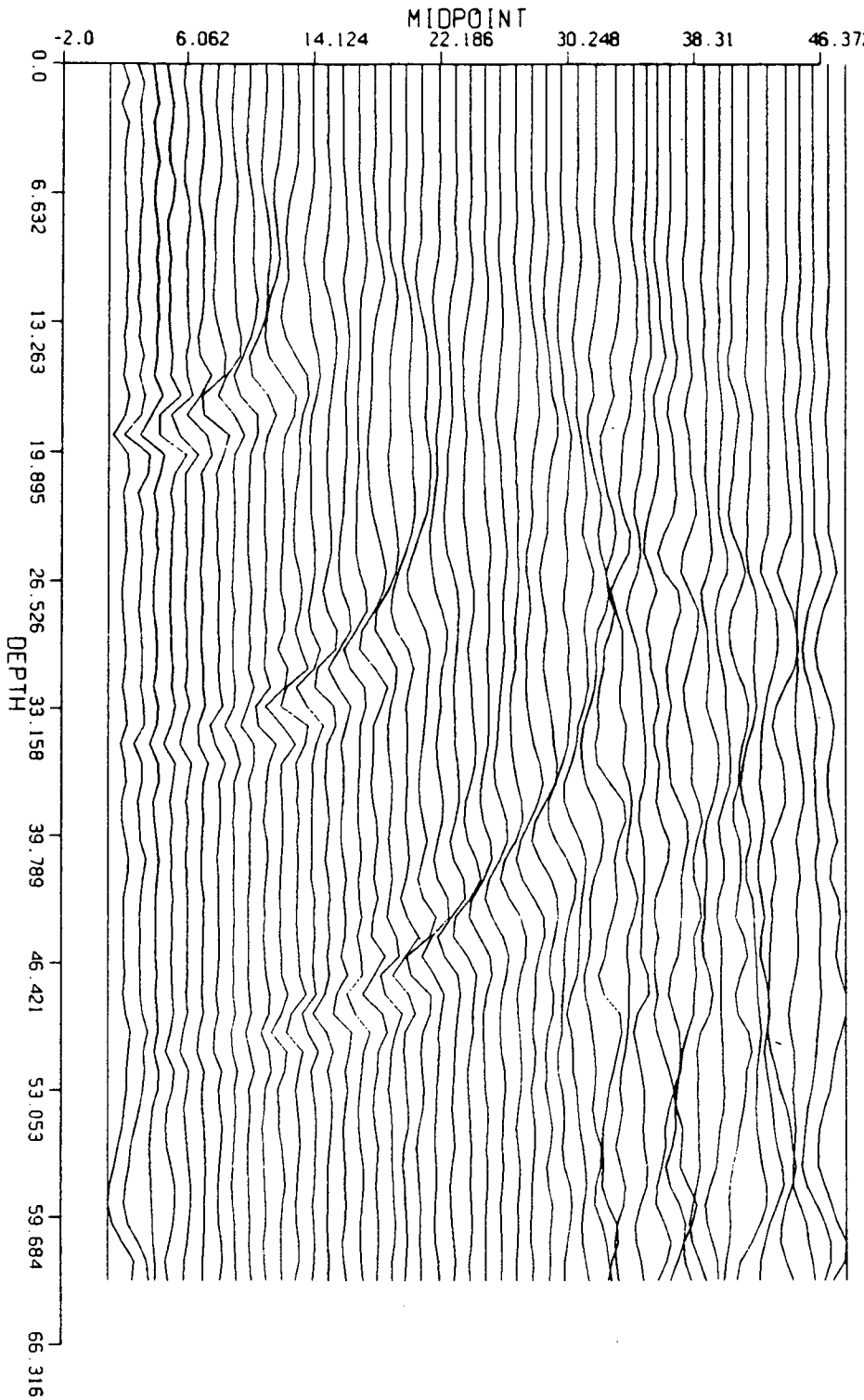

Figure 6.2 Migrated section following Claerbout's (1985) example with homogeneous boundary conditions, instead.

to be ω-domain wraparound effects (see also 6.9.3 for plotting details). The original smoothing of the input spikes is also seen in the broadening of the approximate semicircles. Finally, it is interesting to observe that the migrated waveforms are not symmetric functions of time, as the unmigrated impulses are. Migration (in two dimensions) yields an output that is neither symmetric nor antisymmetric, but a 45° phase-shifted pulse (Claerbout, 1985).

6.5.2.6 A theroretical inconsistency

An interesting point regarding Claerbout's example should be made. By imposing homogeneous Neumann b.c's, we are practically introducing discontinuities in our scheme since the initial conditions fail to obey them. The reason why such a fundamental theoretical flaw does not seem to affect at all the solution might be identified in a number of ways. The Schrödinger equation is naturally diffusive and it therefore tends to smooth any kind of discontinuities. Errors caused by the discontinuity do not propagate fast in the interior of our computational domain, due to the local nature of the finite difference basis functions. Spectral methods would be unforgiving in a similar situation, as the error would rapidly pollute the solution everywhere in the grid. In addition, this problem is eliminated after the first Δz extrapolation has been completed, since the advanced wavefields are now consistent with the imposed boundary constraints.

The lack of boundary reflection contamination is partially explained by the rather mild intensity of the described inconsistency. The original impulses are located at the close vicinity of the left boundary with their peaks on the boundary trace itself. Fourier transformation of the time coordinate involves an implicit normalization of the

Fourier coefficients with the number of points per trace, scaling the discontinuity jump accordingly. The region around the right boundaries is ideally zero and it is, therefore consistent with the imposed boundary condition.

However, the severe loss of quality due to an enlarged boundary discontinuity jump is clearly demonstrated in figure (6.2). The implementation of zero-endpoint boundary conditions amounts to a significant discontinuity jump at the left boundary; the deterioration of the output's quality is rather dramatic. The central sections of the semicircles are lost; a pronounced dispersion alters the upper sections of the reflectors and an enhanced wraparound interference is also present.

6.6 Migration in the Fourier—Chebychev Plane

6.6.1 The Schrödinger and the Fresnel diffraction equations

Let us take a closer look at the diffraction term

$$\frac{\partial U}{\partial z} = \frac{v}{2i\omega}\frac{\partial^2 U}{\partial x^2} \tag{6.23}$$

Assigning $v = 1$ and $\omega = 1/2$, reduces (6.23) to

$$\frac{\partial U}{\partial z} = -i\frac{\partial^2 U}{\partial x^2} \tag{6.24}$$

This is identical to the Schrödinger equation but for a minus sign; it is, in fact, the diffraction term of the forward modeling parabolic equation that is identical to the Schrödinger equation.

For notational simplicity, let us write the diffraction term as

$$\frac{\partial U}{\partial z} = \sigma \frac{\partial^2 U}{\partial x^2} \tag{6.25}$$

with $\sigma = v/2i\omega$ being the equivalent of the $i\hbar/2m$ in the Schrödinger equation; $\sigma(\omega, x, z)$ could be visualized as the complex diffusitivity function.

6.6.2 The fundamental algorithm

The Chebychev solution of the Schrödinger equation having extensively been investigated in the last chapter, there is no need for a detailed presentation of the scheme here. However, a brief discussion of the incorporation of the scheme in the migration algorithm as a whole is essential.

The Chebychev algorithm falls naturally in the spectral (wavenumber) class of algorithms. The spatial decomposition is identified in terms of Chebychev rather than Fourier coefficients. Despite the fact that the notion of the spatial wavenumber k_x is primarily lost in the new decomposition, it may be recovered by sampling at the familiar fast-node set and the subsequent application of the transformation $x = \cos\theta$. This process yields a Chebychev spectrum, where successive coefficients correspond to equidistantly located Chebychev wavenumbers, the difference being that the new wavenumber is understood as k_θ, as opposed to k_x.

The Chebychev representation of the $\partial^2 U/\partial x^2$ operator cannot be computed simply as $-k_x^2\ U$; therefore, we cannot obtain the extrapolated wavefield by the means of a simple phase shift operation, i.e an analytic solution, as in the Fourier migration. Consequently, the computational procedure comprises an intermediary between the

finite difference and the Fourier algorithms. The Chebychev spectrum of a certain ω and at a particular z level is continued downwards via the combined solution of the diffraction term and the thin-lens, as in the finite differences. The thin-lens term is solved analytically, whereas the solution of the diffraction term is done numerically according to the procedures presented in the last chapter.

6.6.3 The boundary conditions issue

Ideally, absorbing boundaries are desirable. Setting them up is an elaborate task and it lies beyond the scope of this first investigation of migration in the $(\omega - k_\theta)$ space. It would seem that a proper implementation of equations (6.19-21) should account correctly for the implicit global character of the boundary constraints in Chebychev techniques. The described equations are local to the boundaries; Chebychev environments are expected to be more sensitive to such absorbing boundaries. Absolute care must be taken to assure that the new boundary conditions are consistent with the global character of Chebychev simulations and that the resulting integration scheme is stable.

It also seems that absorbing boundary formulations of the "sponge" kind would be easier to incorporate in the Chebychev algorithm. An indication is given by the relatively successful application of these techniques not only for finite difference but for Fourier algorithms. A recent generalized approach, which allows even finite element discretizations to be accommodated, is given in Sochacki et al (1987). For the purposes of the current introductory study, we restrict ourselves to homogeneous boundary conditions. Migration input and parameters are then chosen, in such a way, that the

boundary conditions are not dominant and subsequently, the grid satisfactorily mimics an infinite domain.

6.7 Optimization Procedures in the Chebychev Migration Algorithm

The basic Chebychev algorithm presented above is quite primitive; a careful analysis reveals a number of optimization possibilities which, when correctly applied, can reduce the cost essentially. The applicability of these short cuts is highly coupled, primarily, with the boundary conditions used and the level of inhomogeneity. A presentation follows.

6.7.1 Homogeneous boundary conditions

Our first attempt to migrate in Chebychev space involves homogeneous Dirichlet boundary conditions in the x-boundaries of the computational domain. These boundary constraints are maintained throughout the extrapolation process, that is, at all depths (times). This boundary character may be exploited to optimize the fundamental algorithm discussed previously.

6.7.1.1 The thin lens term in Chebychev space

Solution of the Fresnel diffraction term gives the extrapolated wavefield in k_θ. The next step augments this solution with its thin lens counterpart. This operation is a multiplication in the x-domain and it corresponds to a convolution in the θ space. However, the thin lens contribution is merely a constant and therefore the convolution reduces to a simple multiplication. This is an essential point, because a variable thin lens contribution would demand an inverse Chebychev transform to map the

extrapolation back in x, so that the direct application of the thin lens solution is possible.

6.7.1.2. Imaging in k_θ space

The significance of the previous discussion is better understood, when optimization of the imaging is considered. Imaging is, naturally, performed in x-space superimposing twice the real parts of the positive frequencies. This means that an inverse Chebychev transform is required at every (ω, z) level to obtain the wavefield in x there. Fortunately, we can reduce the number of inverse transforms to the number of the desirable depth levels. This is achieved by the superposition of twice the real parts of the extrapolated Fourier-Chebychev spectra, at a given z level; this is a valid procedure since the real character of the Chebychev transform leaves the conjugate symmetry of the ω-axis unaltered. This procedure amounts to imaging in k_θ space, that is to say, computing the inverse Fourier transform at $t = 0$. Having completed this first phase of imaging, we can invert the imaged Chebychev spectra to produce the migrated section in the physical (x, z) space. It is obvious, that the described procedure demands the validity of the analysis given in (6.7.1.1), so that the complete solution is readily attainable in the (ω, k_θ) spectral plane.

6.7.1.3 Constant velocity function

The construction of the Chebychev matrices depends on ω, v and Δz. Assuming a uniform extrapolation step Δz and a constant velocity function v, we need to construct the Chebychev Crank-Nicolson systems only once for each frequency used. Although the loops on ω and z are interchangeable in general, nesting the z-loop inside the ω-loop is optimal.

6.7.1.4 The final algorithm

A brief summary of the complete algorithm for the solution of the 15° migration equation in the (ω, k_θ) plane is presented below. The algorith proceeds as follows.

1.) Initialize the surface pseudodata for the input model in the unmigrated (real) section $P(x_k, z = 0, t_i)$, where x_k $(k = 0, \ldots, N)$ is the half-offset, z is the midpoint and t_i $(i = 0, \ldots, M-1)$ is the two-way travel-time coordinate. M and N must be integer powers of 2 due to FOUR2 input requirements.

2.) Fourier transform the time dependence (column-wise) to obtain its frequency representation ω_i with $-M/2 + 1 \leq i \leq M/2$. Information contained in the negative frequencies is the complex conjugate of the positive frequencies' content and therefore, Chebychev transformation of the x dependence is applied for the positive frequencies ω_i, $0 < i < M/2$, only. The DC $(i = 0)$ and the Nyquist component $(i = M/2)$ are not considered. The input is complex and transformation of both the real and the imaginary parts of the (ω, x) complex wavefields has to be done. The described double Fourier-Chebychev transform may be optionally performed in reverse order. The alternative equivalent process consists of a (real) Chebychev transform in x (all rows), to be followed by a Fourier transform (all columns). The same amount of computational work is involved in both procedures and the output is understood as $P(\omega_i, Ch_k)$, where Ch_k refers to the k-th order Chebychev mode, i.e $k = 0, \cdots, N$.

For all ω_i ; $0 < i < NYQ$

Perform a monochromatic wavefield extrapolation at all the desired depth levels as described in the following loop.

3.) Calculate the diffusivity-like coefficient, which corresponds to the current value of ω_i and the velocity function v.

4.) Compute the amount of phase shift from the analytic solution of the thin lens term for the current ω_i, the velocity v and the particular choice of the extrapolation step Δz. This constant quantity (since $v \neq v(x)$) is denoted as s.

5.) Construct the appropriate Chebychev Crank-Nicolson L (left-hand side) and R (right-hand side) matrices, according to the projection operator chosen (Galerkin, pseudospectral, tau-differentiated or tau-integrated) and the appropriate homogeneous boundary conditions.

6.) Isolate the Chebychev spectrum of each monochromatic ω_i wavefield at the z_{j-1} depth level, i.e $r_i^{j-1}(Ch_k) = P(Ch_k, z_{j-1}, \omega_i)$, where the depth index j runs from 1 to a predetermined total number of extrapolation steps NZ.

For all $[\Delta z]_j$ $j = 1, \cdots, NZ$

7.) Update the right hand side of the simultaneous equations by multiplying the complete solution at the previous level with the R matrix, i.e $Rr_i^{j-1}(Ch_k) = h_i^{j-1}(Ch_k)$.

8.) Solve the system $Lr_i^j(Ch_k) = h_i^{j-1}(Ch_k)$.

9.) Incorporate the thin lens solution by the multiplication, i.e $r_i^j(Ch_k) = s * r_i^j(Ch_k)$

10.) Image in Ch_k space (compute the inverse Fourier transform at $t = 0$) through the superposition $\hat{P}(Ch_k, z_j) = \hat{P}(Ch_k, z_j) + 2 * Real[r_i^j(Ch_k)]$, where $\hat{P}$ contains the reflectors' surface in Chebychev space.

continue with the z loop

continue with the ω loop

11. Conclude the migration process by inverting the imaged Chebychev spectra of the complete wavefields at all the desired depth levels, i.e $P(Ch_k, z_j) \longleftrightarrow P(x_k, z_j)$.

6.7.2 Other kind of boundary conditions

A more demanding future treatment of the problem would require a more elaborate boundary structure. Although it is premature to refer to optimization techinques and details for algorithms to be developed in the future, a brief investigation of a number of problems is worthwhile.

Theoretically, the level of cccomplication for boundary conditions corresponds to non-homogeneous Dirichlet conditions, i.e imposing the requirement of maintaining the original x-boundary values during the depth extrapolation. This minor complication, at first glance, hinders a straightforward implementation of the presented optimized algorithm. The need for subtracting the (complex) linear trends prior to the solution of the diffraction term, the intervention of the thin lens solution (a boundary free complex constant) and the anticipated addition of the trend back to the computed solution to complete one solution cycle, affect the efficiency of the algorithm. Nevertheless, numerical experiments done with a test algorithm based on the described data manipulation prove the validity of the described procedure and demonstrate its ability to cope with the boundary condition alteration caused by the thin lens solution.

The need for introducing additional transforms at each extrapolation phase, for going forth and back in the transform space to accommodate the cumbersome solution process, is uninviting. Improvement gains may be achieved by a skillful exploitation (when applicable) of the full separation of the solution components; the incorporation

of simple analytic solutions for the imaging of the troublesome trends is another issue be explored. A simplification of this complicated situation may be also achieved by an a priori global removal of the existing trends. Then, a proper augmentation of the final solution of the intermediate well-known homogeneous problem with the trend map, would conclude the migration; nevertheless, a meticulous analysis is required to prevent unexpected failures. The issue is of theoretical importance only, since such a boundary structure is no better than its homogeneous counterpart, as far as artificial reflections are concerned. Nonetheless, it gives an idea of possible complications.

Another kind of boundary conditions is the homogeneous or the non-homogeneous Neumann type. The homogeneous problem is solved as its Dirichlet counterpart; the boundary structure is incorporated in the propagating matrices themselves. The intervention of the thin lens solution remains an issue of consideration, while the non-homogeneous version involves manipulation involving a complex quadratic trend (its Laplacian is not zero, as in the case of the linear trend). Ultimately, an absorbing boundary structure should be implemented; these conditions can be formulated as an extended version of Robbins boundary constraints, merging the analysis for their Dirichlet and von Neumann components given previously. We should once more emphasize that the imposed conditions must lead to stable problems; extreme care must be taken, due to global character of spectral discretizations. Sponge-type boundary treatment introduces an implicit degree of locality to the boundary conditions and it is probably easier to apply. The danger of introducing instabilities is still high and a painstaking investigation is a prerequisite.

6.7.3 Variable velocity functions

A horizontally stratified medium does not affect the versatility of the main algorithm but it slows it down. The propagation systems have to be reconstructed at each new velocity level; furthermore, the thin lens solution needs to be recomputed there.

Lateral velocity variations are, in principle, difficult to handle with spectral techniques. Galerkin and tau implementations have to be solved in the transform space; thereafter we are confronted with the evaluation of inner products of the form

$$\int_{-1}^{+1} \sigma(x) \frac{\partial^2 u_N}{\partial x^2} \frac{\phi_n}{\sqrt{1-x^2}}\, dx \tag{6.26}$$

or

$$\int_{-1}^{+1} \sigma(x) \frac{\partial^2 u_N}{\partial x^2} \frac{T_n}{\sqrt{1-x^2}}\, dx \tag{6.27}$$

for the Galerkin and the tau method respectively.

The major difficulties encountered in the evaluation of (6.26) and (6.27) are avoided by the application of the pseudospectral technique; the latter permits the accurate derivative computation in the transform space, while solving the system in the physical space, i.e

$$\frac{\partial}{\partial z} u_N(x_j) = \sigma(x_j) \frac{\partial^2}{\partial x^2} u_N(x_j) \qquad ; \qquad j = 1, \ldots, N-1 \tag{6.28}$$

with

$$u_N(x_0) = u_N(x_N) = 0 \tag{6.29}$$

The pseudospectral technique is capable of handling velocity variations easily; the function and the derivative can be computed efficiently in $O(2N \log N)$ operations.

6.8 The $(\omega - k_\theta)$ Transform of a Seismic Section

The proposed migration procedure may be identified as a downward continuation of the Fourier-Chebychev transformed original (t, x) seismic section. An advanced version of the described algorithm should incorporate a rigorous analysis of the (ω, k_θ) plane. The investigation should be capable of uniquely determining the undesirable evanescent regions and the occurence of dip aliasing. Identification of the these most important aspects in the mixed-tranformed seismic section necessitates the derivation of the appropriate dispersion relation in a consistent fashion.

A rigorous global approach to the problem is definitely non-trivial, since the particular combination of Chebychev polynomials, used as basis functions, has to be manipulated. Fourier implementations of absorbing (non periodic) boundary constraints overcome the problem at the expense of strict rigor. The boundary conditions are designed to have an indirect boundary locality and thus an approximation of the significant interior domain of the computational grid with the same (periodic) basis functions can be partly justified. The approach is reminiscent of local mode analysis (simulation of inaccessible boundaries); the dispersion relation is assumed to be unaltered. The difficulties of the (ω, k_θ) transform may be attacked by a similar approximation, that is, employment of the original $T_n(x)$ polynomials in the calculation of the dispersion relation. This promising approach should employ the fast x-nodes, in order to reduce the $T_n(x)$ functions of the x-coordinate to *cosines*, i.e $\cos n\theta$, in the θ-coordinate axis.

The Chebychev transform becomes a (real) cosine transform; on the contrary, the Fourier transform is complex. Consequently, it is not obvious, how spatial aliasing (dip reversal) is manifested. Is a radial line in (t, x) transformed into a radial line in the

(ω, k_θ) plane? The Chebychev spectrum "sees" positive frequencies only and it might be that the apparent lack of the *sine* component (an additional degree of freedom in a complex space) defaults to a folding of the dip, i.e contamination of the lower Chebychev wavenumbers, as opposed to the dip reversal characterizing the (ω, k_x) plane of spatially aliased data, the latter letting the "visible" lower wavenumbers free of aliasing contaminations. On the other hand, the Chebychev transform of a complex function does not involve a coupling between the transforms of its real and the imaginary components. This feature is not shared by Fourier transformations and it should have a very definite effect on the behavior of the aliased modes in two dimensions; some kind of aliasing decoupling should be anticipated. The issue is, nonetheless, not trivial and we do not pursue it further. The necessary mathematical analysis and the appropriate numerical experiments are needed to provide the answers and a full understanding of the (ω, k_θ) transform plane; mapping of the evanescent areas and a better grasp of aliasing and dip filtering in these double mixed-transform coordinates are also anticipated.

Alternatively, the well-known Fourier analysis may be used to aid in the identification of all these important parameters. Various filtering operations may be performed in (ω, k_x) to remove the undesirable spectral regions. Subsequently, the filtered version of the (ω, k_x) spectrum may be inverted back, to be transformed again in Chebychev space for the extrapolation of the retained portion of the spectrum itself. The procedure involves interpolation of complex quantities; therefore a careful analysis should be made to ensure that unwanted surprises do not occur. This alternative might be worthwhile, saving trouble and unnecessary fustration.

6.9 Chebychev Migration of a Model Problem

6.9.1 Parameter initialization

The input model is chosen to follow the rules of a two-dimensional Gaussian distribution in the original (t, x) space of the CMP gather. This choice comprises a versatile input, since a direct control of its attenuation characteristics is readily managed. The input "spike" is defined as

$$P(t,x) = \exp[(-\sigma_t t^2 - \sigma_x x^2)/2] \tag{6.29}$$

The x-line is taken to be the fundamental Chebychev interval, i.e $[-1,+1]$. This choice merely serves convenience; a general interval $[a,b]$ can be mapped onto the fundamental interval by means of the simple linear transformation

$$y = \frac{1}{2}(b-a)x + \frac{1}{2}(b+a) \tag{6.30}$$

The novel character of the Chebychev algorithm demands a model migration, so that the important aspects of the new technique are illustrated properly. A first oversimplification is found in the very choice of an appropriately smoothed "spike" model; the model is often employed as a first order evaluation of various implementations. In theory, an ideal spike has infinite temporal and spatial bandwidths and the Gaussian parameters σ_t and σ_x can be picked such that a smoothed Gaussian "hill" is obtained. We aim at a clear, sharp migration semicircle; too much smoothing results in blurred outputs, which, naturally, suffer from resolution losses. Additionally, we do not want

boundary reflection contaminations due to the homogeneous boundary conditions. Assuring the absence of evanescent aliasing and the virtual lack of migration "ghosts" are equally important; none of the algorithms uses the troublesome information at the ω_0 and the ω_{NYQ} components. Enlargement of the grid sizes imposes prohibitively large amounts of computational work, especially in the case of the Chebychev version. An acceptable compromise may be found by allowing some spatial and temporal aliasing; the simplified unique character of the input model permits the migration algorithms to proceed readily and obtain migrated sections of reasonably high quality.

In the following examples, we employ a 17-long x-discretization (17 traces) and a temporal set of 64 samples (64 points per trace). Consequently, the finite difference and the Fourier techniques use $\Delta x = 2/16$, while the usual fast-node sampling pattern is employed for the Chebychev procedure. Furthermore, the temporal sampling interval is chosen to be $\Delta t = \Delta x/4$ (implying $X = T$), a constant velocity function $v = 2$ is assigned and a $\Delta z = \Delta t$ is implemented in the extrapolation process (satisfying the depth anti-aliasing requirements). The input Gaussian is centered on the x-axis, i.e at $x = 0$ and it is located at one third the length of the travel time axis; the parameters σ_t and σ_x are given the uniform value of 4500. A minor disadvantage of the described model is its predisposal against the Chebychev algorithm, since the the fast-node sampling is expected to glean an inferior amount of information, due to its coarse character in the central part of the x-domain.

6.9.2 Comparison of migrated sections

The migration of our model with the phase shift method is featured in figure (6.3); this result is being included in order to demonstrate a migration output correct to 90°.

The familiar wraparound effect is not witnessed because the periodic interference lies outside the $[-\pi, +\pi)$ fundamental interval (which has been mapped onto the $[-1, +1)$ interval for consistency). Figure (6.4) displays the migrated section obtained when the 15° approximation is invoked. Only a minor amount of negative interference is discernible; the explanation lies in the fact that both the primary evanescent energy and the energy lying at dips higher than 15° have not been incorporated in the extrapolation process. The small amounts of "ringing" present in figures (6.4-5) are caused by the sharp (ω, k_x) filter edges.

The next two figures demonstrate the destructive interference of these two factors, despite the fact that care has been taken to keep this at a minimum level. Figure (6.5) portrays the situation where only the primary evanescent energy has been filtered, while figure (6.6) is concerned with the migration outcome when the whole (ω, k_x) spectrum has been employed. This last figure will be later compared with the Chebychev and the finite difference results. The finite difference algorithm's artificial spatial dispersion has catastrophic consequences; the problem is clearly depicted in figure (6.7) where the true reflector is highly overmasked by dispersion noise. The boundary traces are zero, but no boundary reflection contamination is experienced, as in the case of the Fourier technique. The observed dispersion effects, due to the low-order polynomial representation of $\partial^2/\partial x^2$, demand a much denser spatial sampling of at least an increase of one order of magnitude.

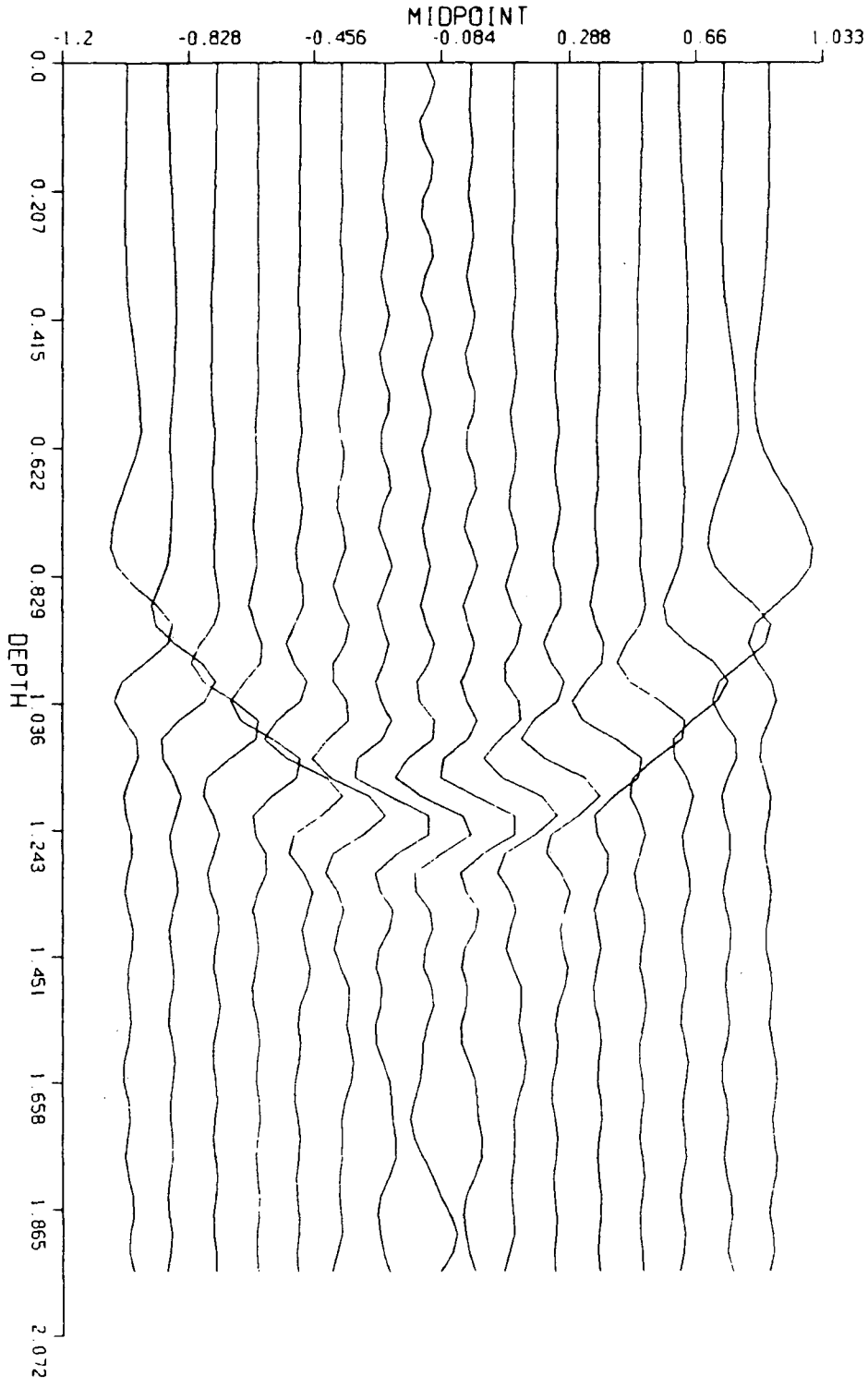

Figure 6.3 Migration with the phase shift method.

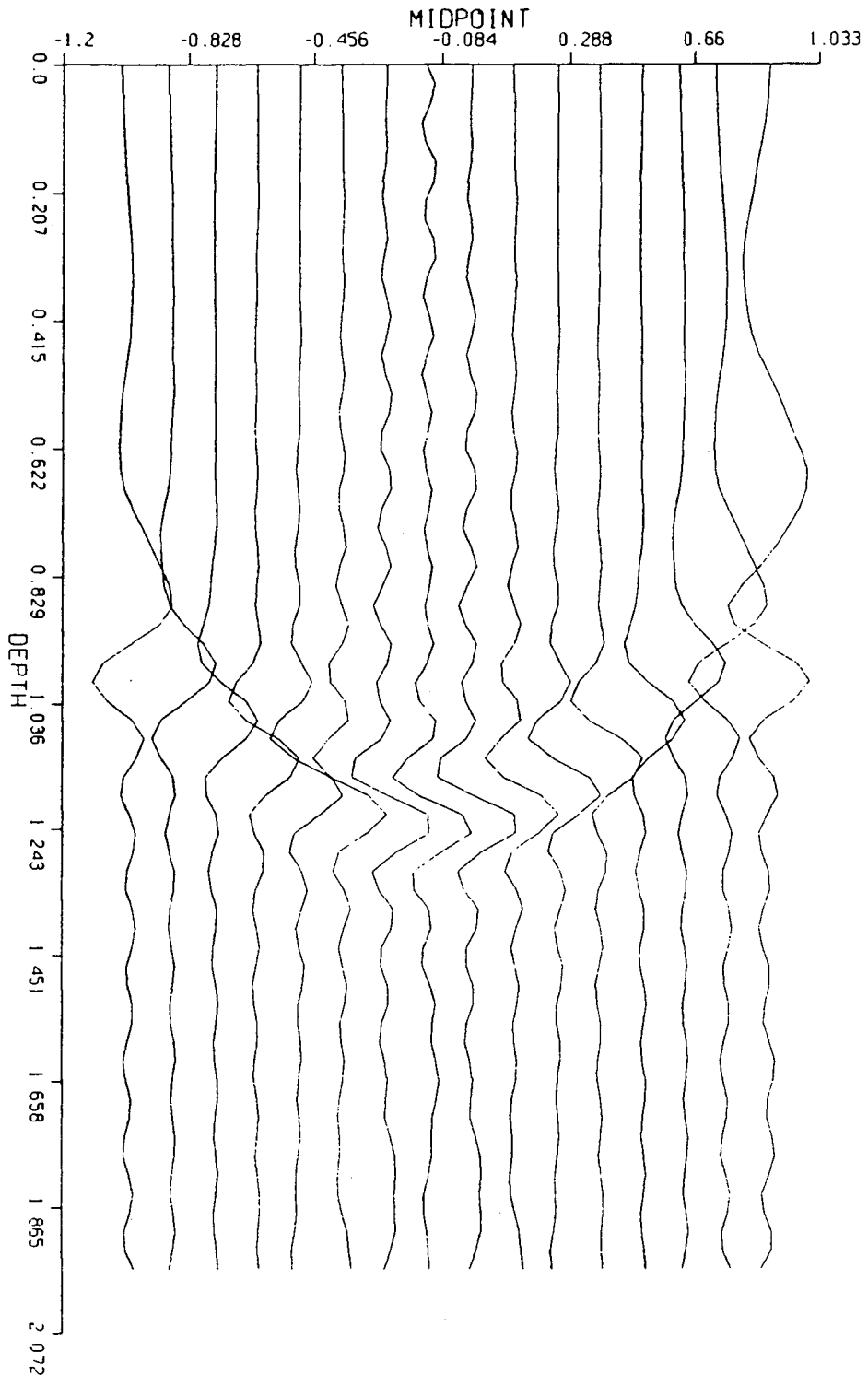

Figure 6.4 15° Fourier migration; evanescent and high dip energy have been filtered.

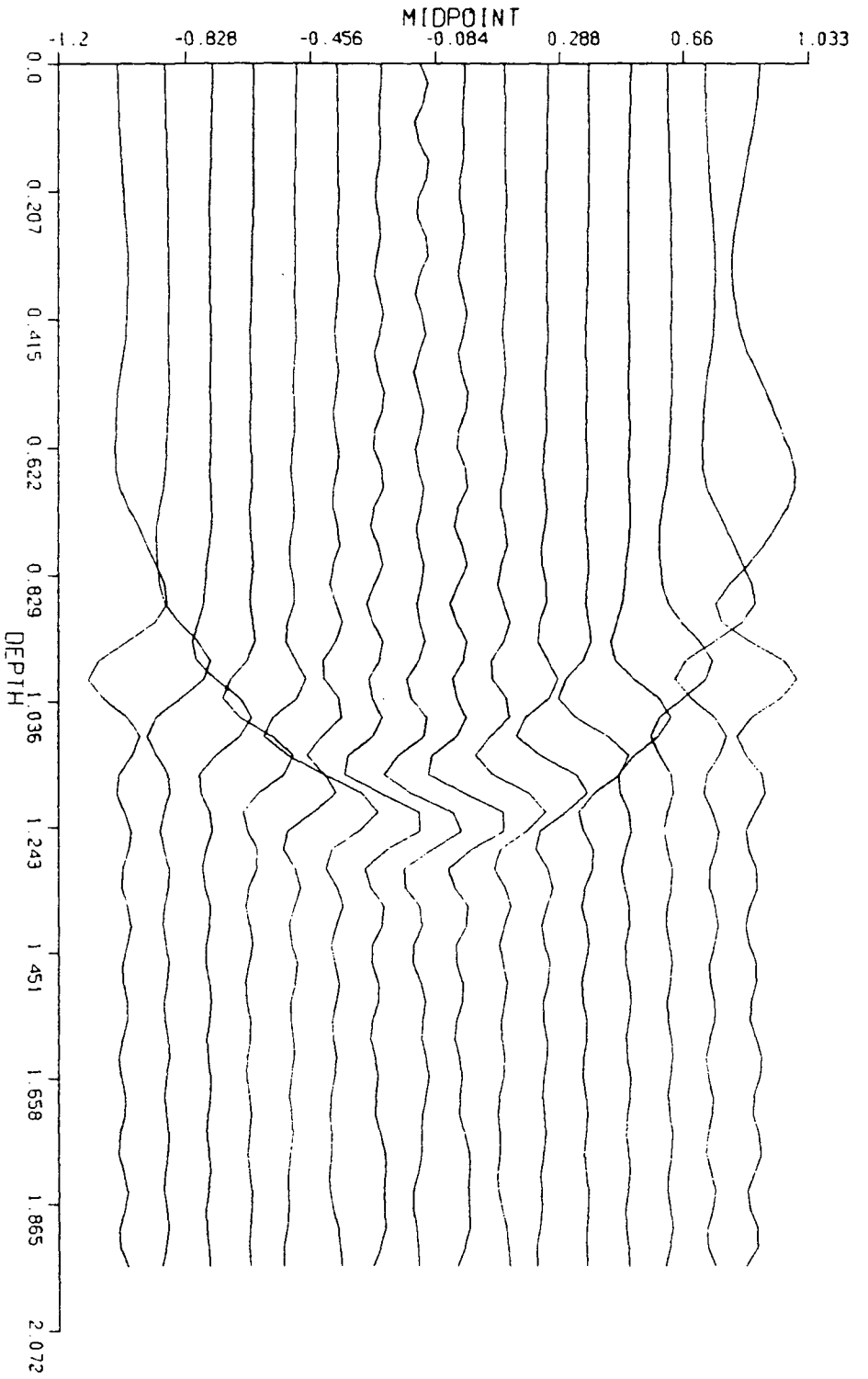

Figure 6.5 15° Fourier migration; only the evanescent energy has been filtered.

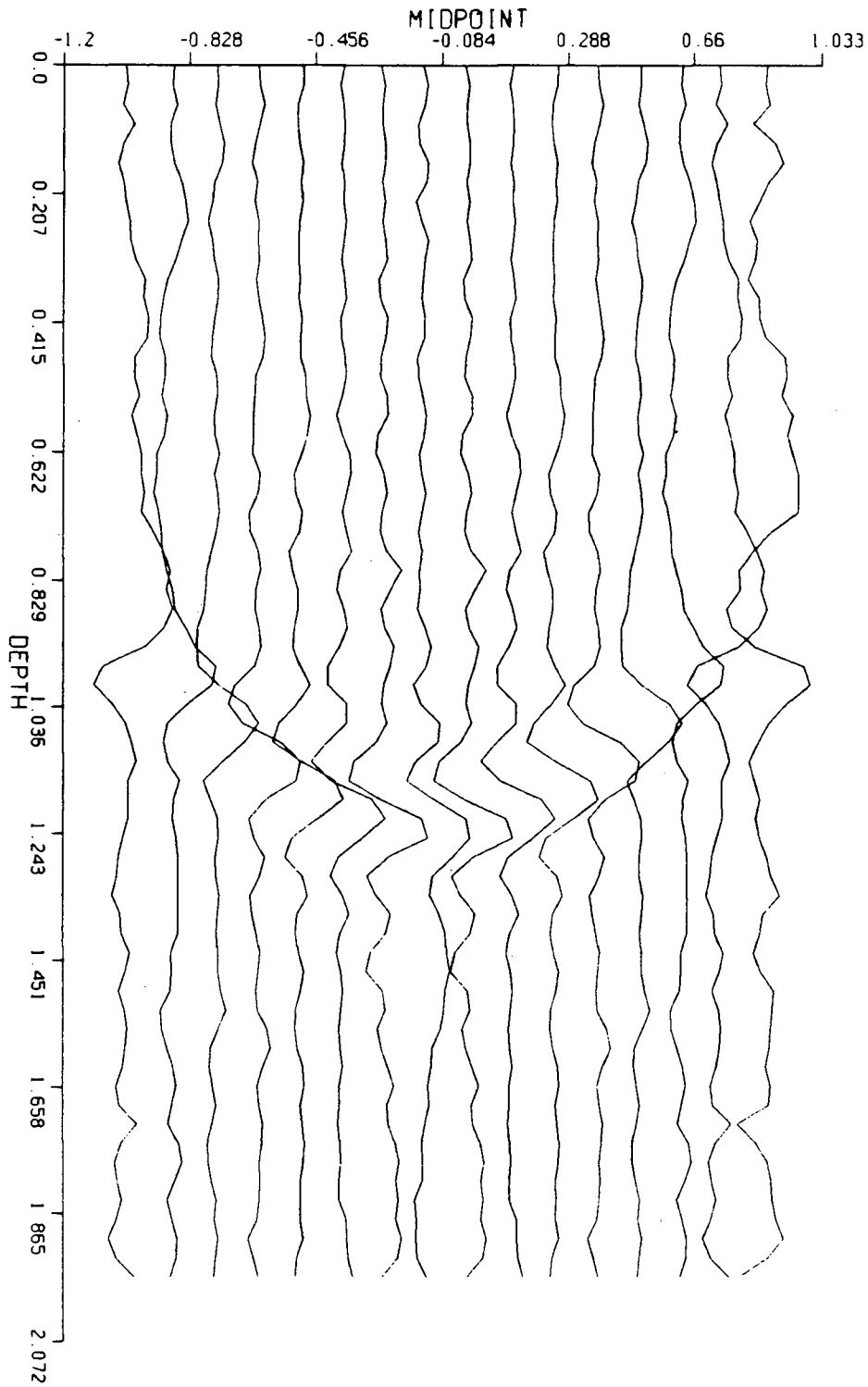

Figure 6.6 15° Fourier migration; no filtering has been applied.

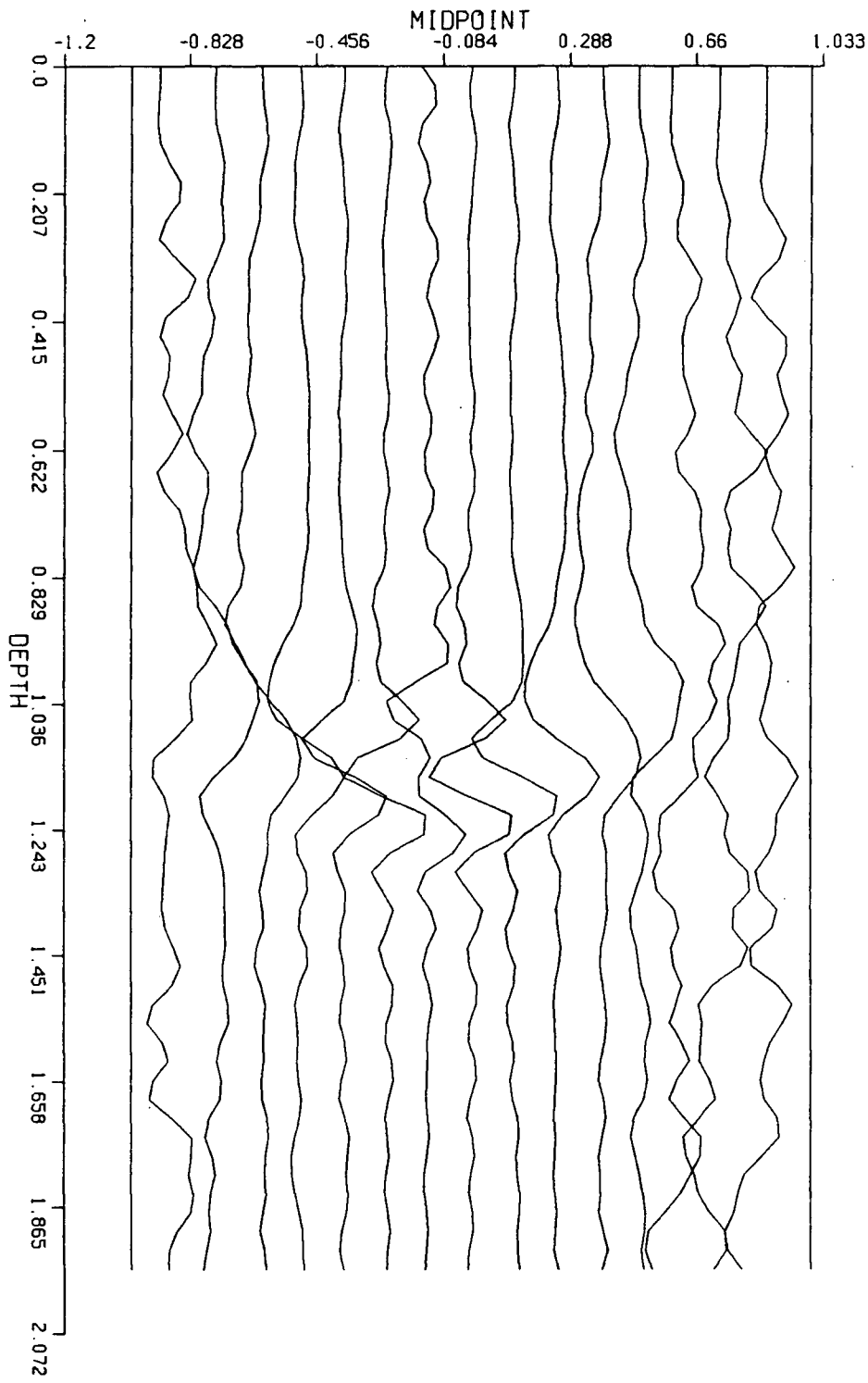

Figure 6.7 15° finite difference migration.

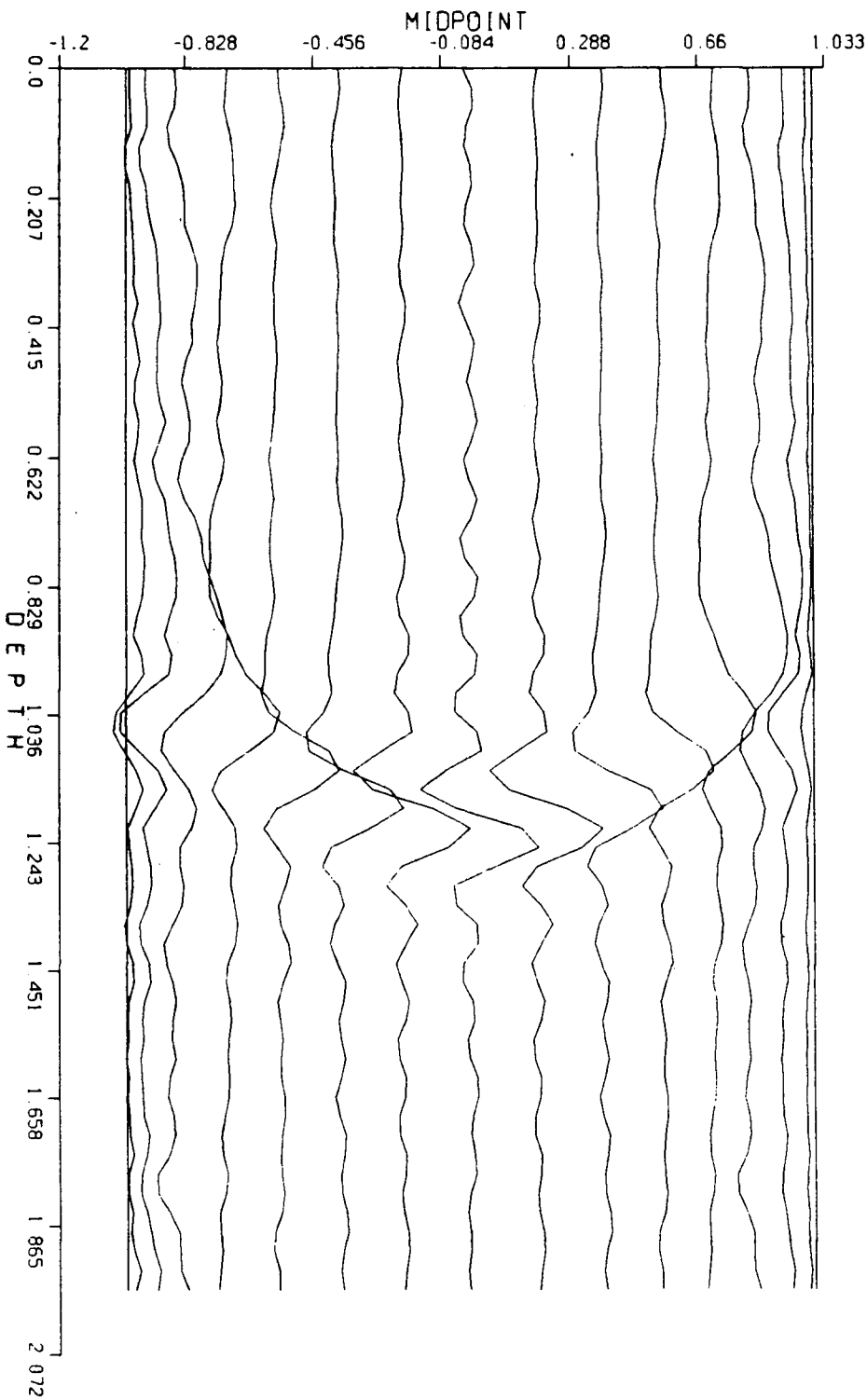

Figure 6.8 15° Chebychev-Galerkin migration.

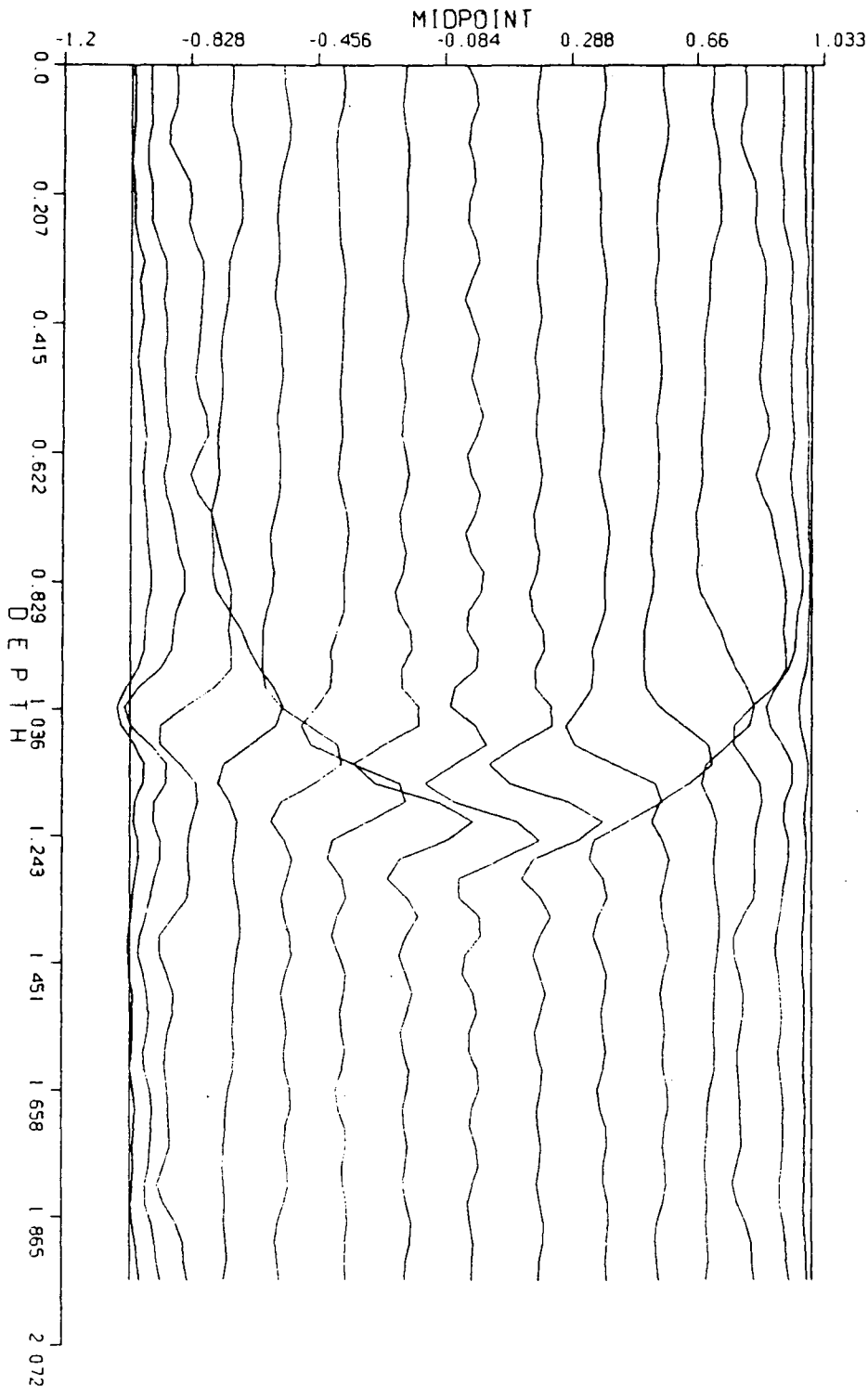

Figure 6.9 15° Chebychev-pseudospectral migration

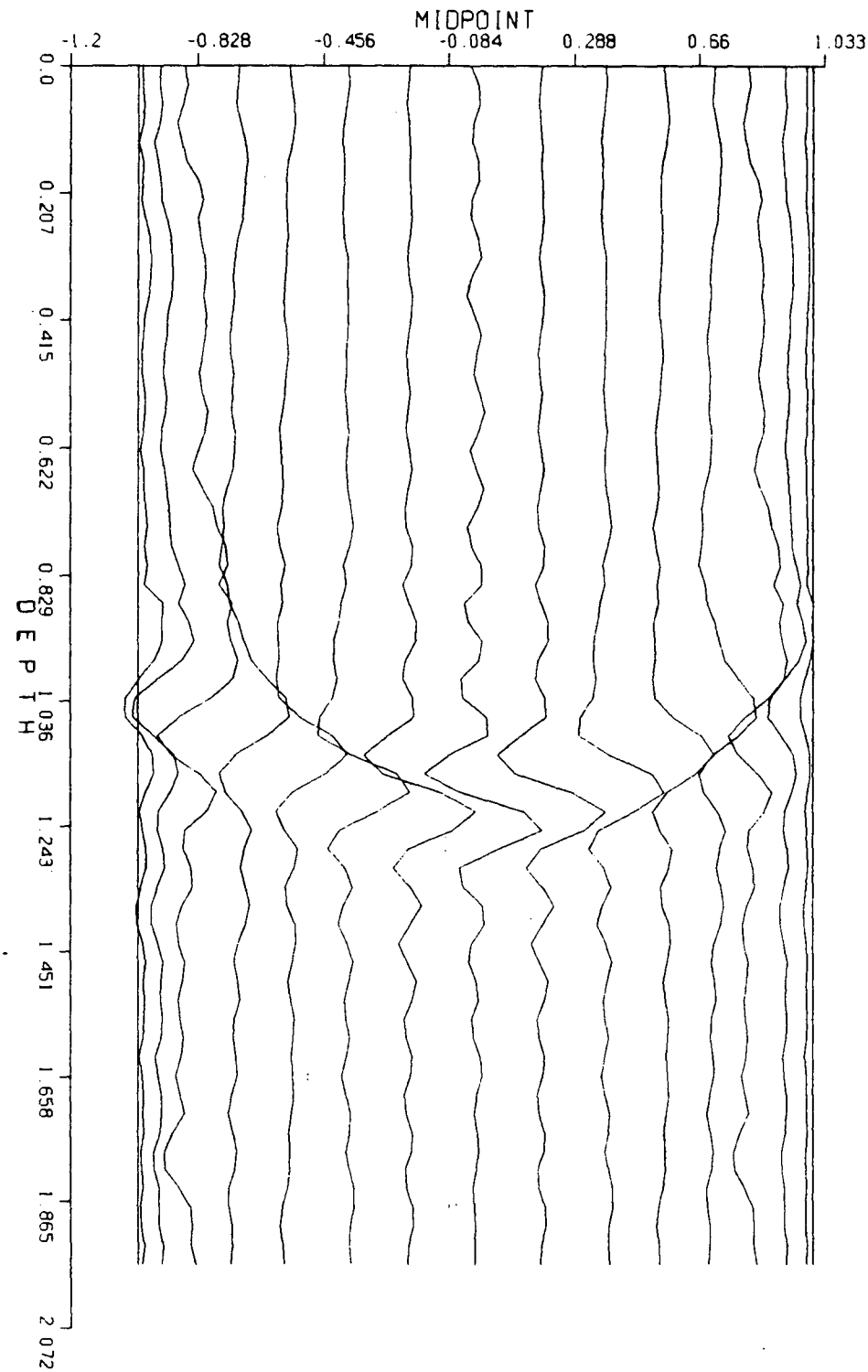

Figure 6.10 15° Chebychev tau-differentiated migration.

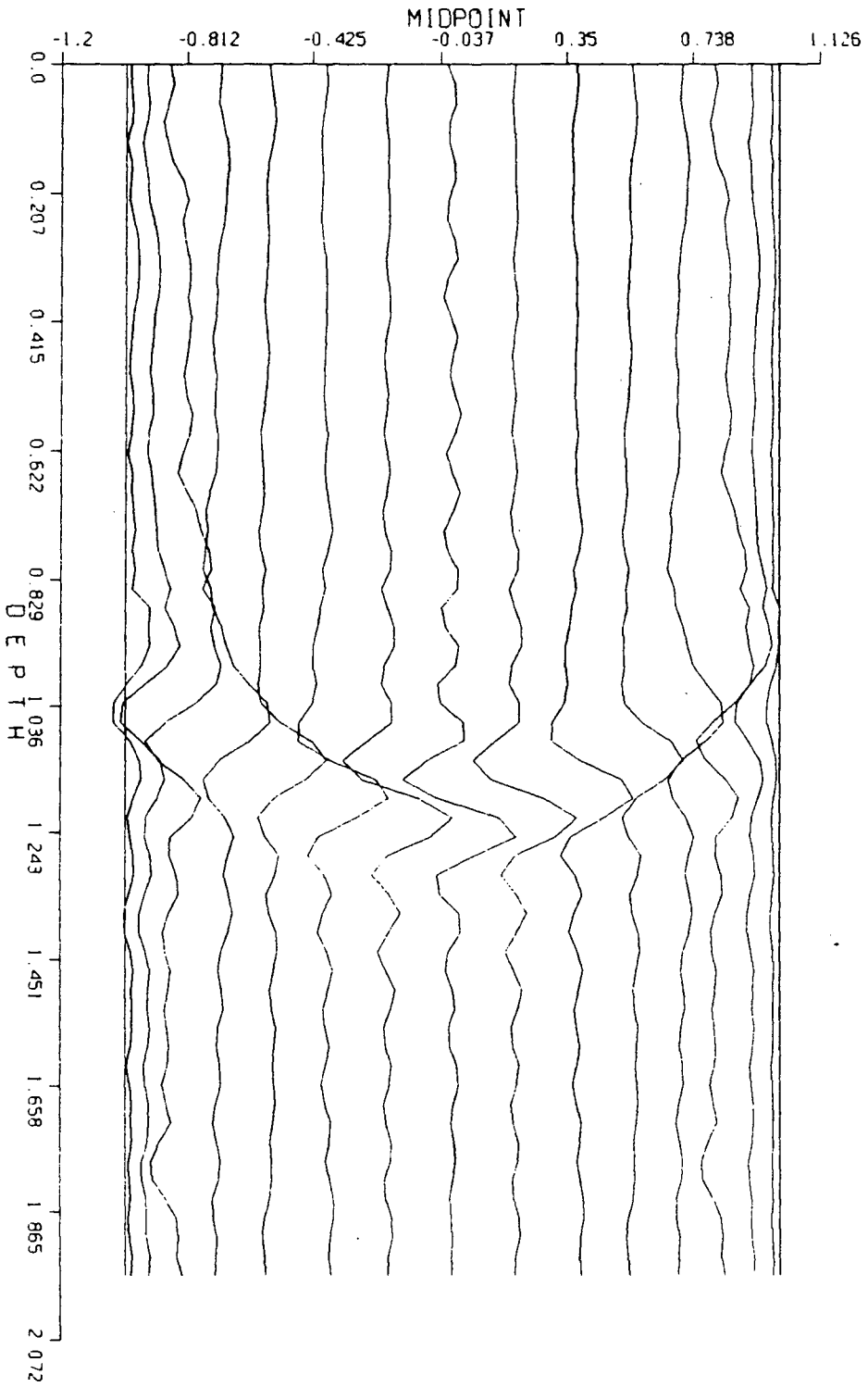

Figure 6.11 15° Chebychev tau-integrated migration.

The Chebychev algorithm's performance is unfolded in figures (6.8-11) corresponding to its Galerkin, pseudospectral, tau-differentiated and tau-integrated variants, respectively. The results appear to be quite similar demonstrating the approximate equivalence of the various projection choices for correctly implemented, trouble-free, constant-coefficient linear problems. Minor differences are observed in the close vicinity of the boundaries, between the tau variants and the other projections. The two tau techniques perform identically, while the same may be said for the Galerkin and the pseudospectral projections. The Chebychev migration compares favorably with its Fourier rival; artificial reflections are absent and no appreciable contamination due to the z-dispersion effects is experienced. Nevertheless, the Galerkin and the pseudospectral techniques would exhibit a marginally superior performance over the tau implementations, under certain conditions, as expected from our previous experience and as seen in a number of other migration experiments, not shown here.

6.9.3 Plotting resolution

The limited number of traces and the aliasing present hinder the production of an output of a high cosmetic value. The need for writing a simple plotting routine is a direct consequence of the lack of a readily accessible, highly sophisticated plotting program, which would allow an input of non-equidistant traces to be fed into it. As a result, another limitation in obtaining pictures satisfying high esthetic demands is imposed. Despite the fact that it is only the Chebychev sections that demand the use of this special routine, all results have been plotted in this way, to allow for a more objective comparison among them. Trace to trace normalization, i.e local as opposed to global, is applied and while a uniform normalization is incorporated

in the finite difference and the Fourier graphs, the Chebychev plots invoke an non-uniform normalization pattern, which involves the particular adjacent traces. Finally, an overlap parameter of 100% is used in all plots.

6.10 The Tau-Integrated Chebychev Algorithm

The short investigation of the properties of the fast inversion of the complex-valued quasi-tridiagonal systems of the Chebychev Crank-Nicolson tau-integrated formulation of the Schrödinger equation (see 5.7) has not yielded promising results. Nevertheless, the high quality of the Chebychev-tau migrated section (figure 6.11) points to the need for further research on the application of procedure **SLU1** in this migration algorithm.

6.10.1 Theoretical insights

The analysis presented in (5.7) concerns the special case $\sigma = \imath$. The more general problem of a non-unit valued diffisivity coefficient is encountered in the Fresnel diffraction term, i.e $\sigma = -\imath v/2\omega$ and therefore, the magnitude of the reported off-diagonal dominance becomes of $O(|1/2n^2 - \imath v\Delta z/4\omega|)$. Despite the fact that the situation remains basically the same as before, the amplitude of the ratio $v/2\omega$ has a definite effect on the analysis of the Schrödinger case. It only appears to be natural that for $|\sigma| > 1$ the off-diagonal dominance is pronounced and thus a reduced Δz is needed to counterbalance the large diffusivity coefficient. On the other hand if $|\sigma| < 1$, the situation is reversed and the restrictions on the size of Δz can be moderately relaxed.

6.10.2 Numerical experiments

The migration procedures for our previous model example use a depth step $\Delta z \sim O(1/N)$. According to the analysis for $|\sigma| = 1$, procedure **SLU1** is incapable of coping with systems with $N \geq 32$ when such a Δz is employed.

6.10.2.1 The low frequency instability of procedure SLU1

A similar behavior has been identified in the inversion of the Fresnel term. An excellent performance is seen for $N = 16$; the output is identical with the migrated section displayed in figure (6.11), while the cost is reduced by roughly 90%, just marginally greater than for the finite difference solution. For our problem the parameter $v/2\omega$ ranges from π^{-1} to $(32\pi)^{-1}$ and it thus defaults to an implicit reduction of the "apparent" step size, i.e $\sigma\Delta z$. However, an implementation of the fast inversion for $N = 32$ and a step size of $O(1/N)$ has failed ultimately.

The answer to this can be traced in the unsuccesful extrapolation of the low frequencies. That may be easily seen by recalling the heuristic discoveries of the Schrödinger analyisis (see 5.7). There it has been found that a step size of $O(1/N)$ results in an irredeemable instability of the procedure, whereas a step size of $O(1/N^2)$ yields results of high accuracy. The high frequencies have indeed been scaled, such that they are extrapolated with an "apparent" Δz of $O(1/N^2)$. However, the scaling of the low frequencies has left the $\Delta z \sim O(1/N^2)$ pattern basically unaltered. Consequently the extrapolation of the high frequencies is performed correctly, but the inversion algorithm collapses for the low frequency components of the solution and erroneous results are obtained.

6.10.2.2 A second migration input-model

We now examine a Chebychev-migrated section of a denser spatial sampling. For demonstration purposes, we take $N = 32$, $DT = DX/2$ and a depth step $DZ = DT/10$. The scaling factor 10, for the depth step, has been chosen such that the low frequency inversion is stabilized while the computational cost is still manageble. Figures (6.12-15) demonstrate the performance of the Fourier (without and with filtering), the finite difference and the **SLU1**-Chebychev algorithm for the the migration of the second model problem.

A comparison of figures (6.12) and (6.13) unveils the character of the evanescent energy in the context of the 15° propagation. Recall the Taylor expansion of the full square root $\sqrt{k^2 - k_x^2}$, i.e

$$k_z = \sqrt{1 - \left(\frac{vk_x}{\omega}^2\right)} = \frac{|\omega|}{v}\left(1 - \frac{vk_x^2}{2\omega^2} - \frac{vk_x^4}{8\omega^4} + \ldots\right) \tag{6.31}$$

Expansion (6.31) converges for $vk_x < \omega$ only, while for $vk_x > \omega$ diverges rapidly. Nonetheless, there is a major difference in the way that evanescent aliasing manifests itself in the full-angle and the 15° migration equations. In the former, evanescent modes give rise to growing exponentials; on the contrary, the latter is incapable of exhibiting the characteristic evanescent blow-up.

How does evanescent energy behave under expansion (6.31)? Let us first examine the 15° case of current interest. We do see that for $vk_x \in (\omega, \sqrt{2}\omega)$ the computations are incorrect but energy is sent in the correct direction. As $vk_x > \omega$ the complex exponential changes sign and incorrect high frequency energy is sent in the opposite

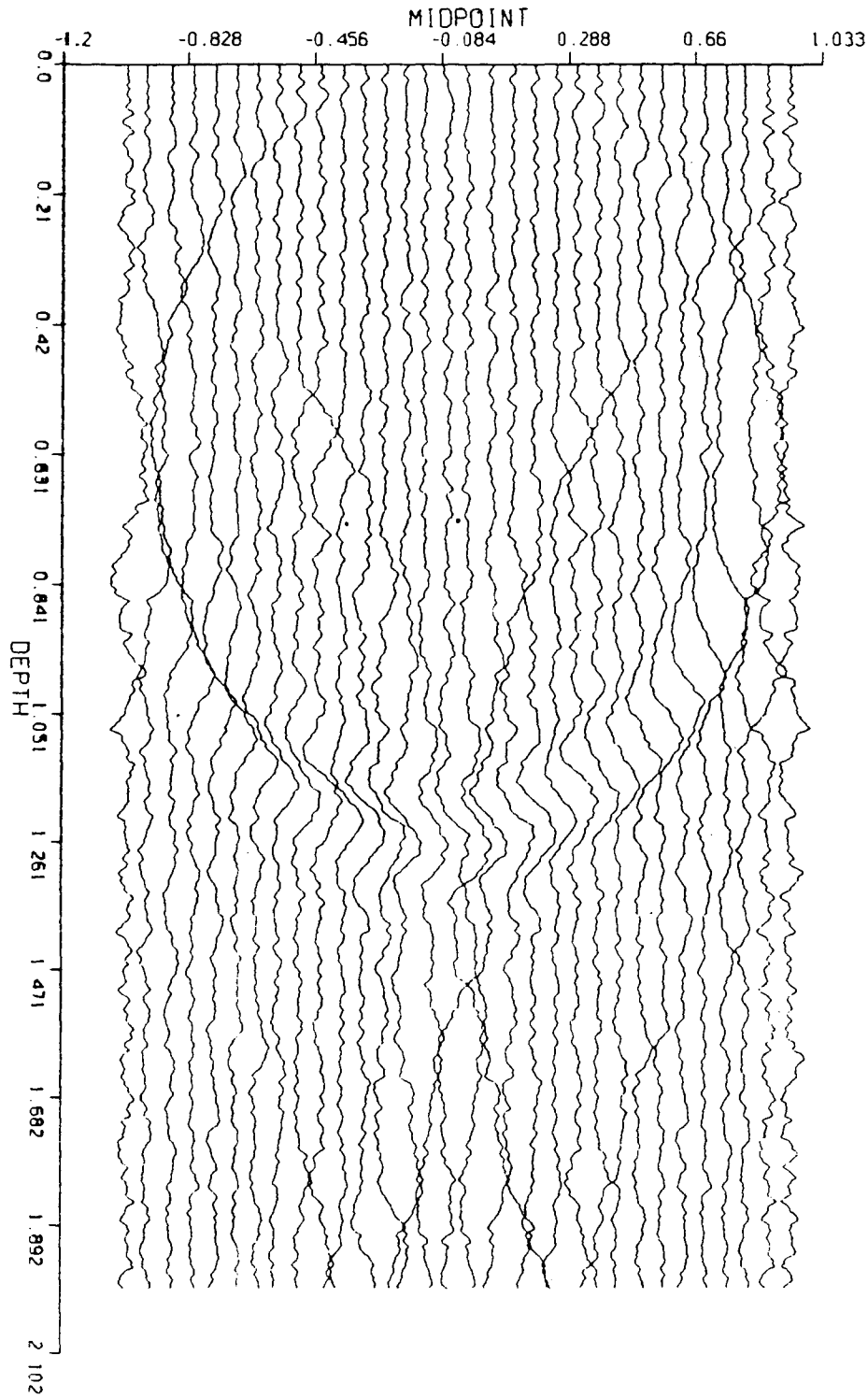

Figure 6.12 Fourier migration of the second model problem; evanescent filtering has not been incorporated.

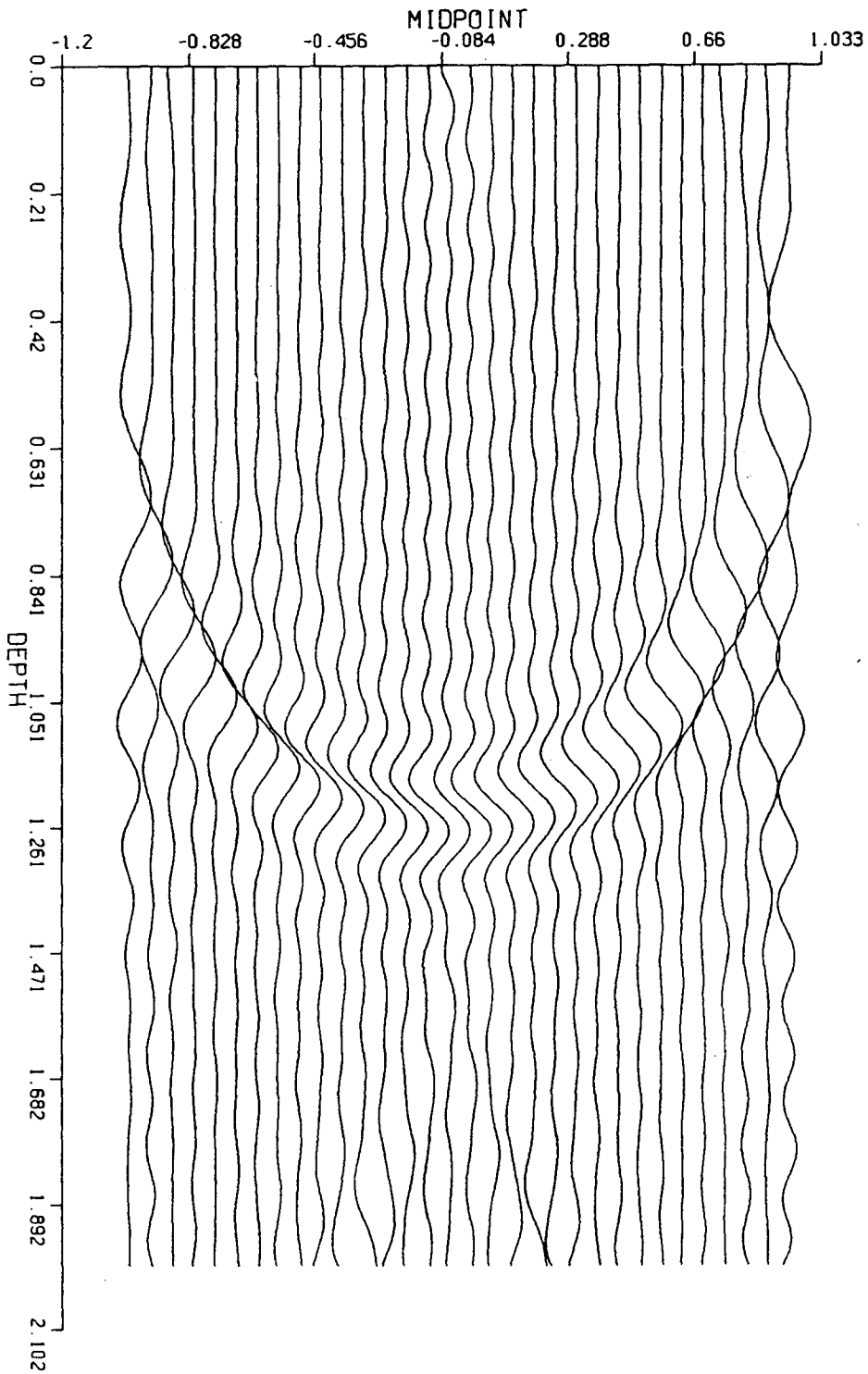

Figure 6.13 Fourier migration of the second model problem; evanescent filtering has been incorporated.

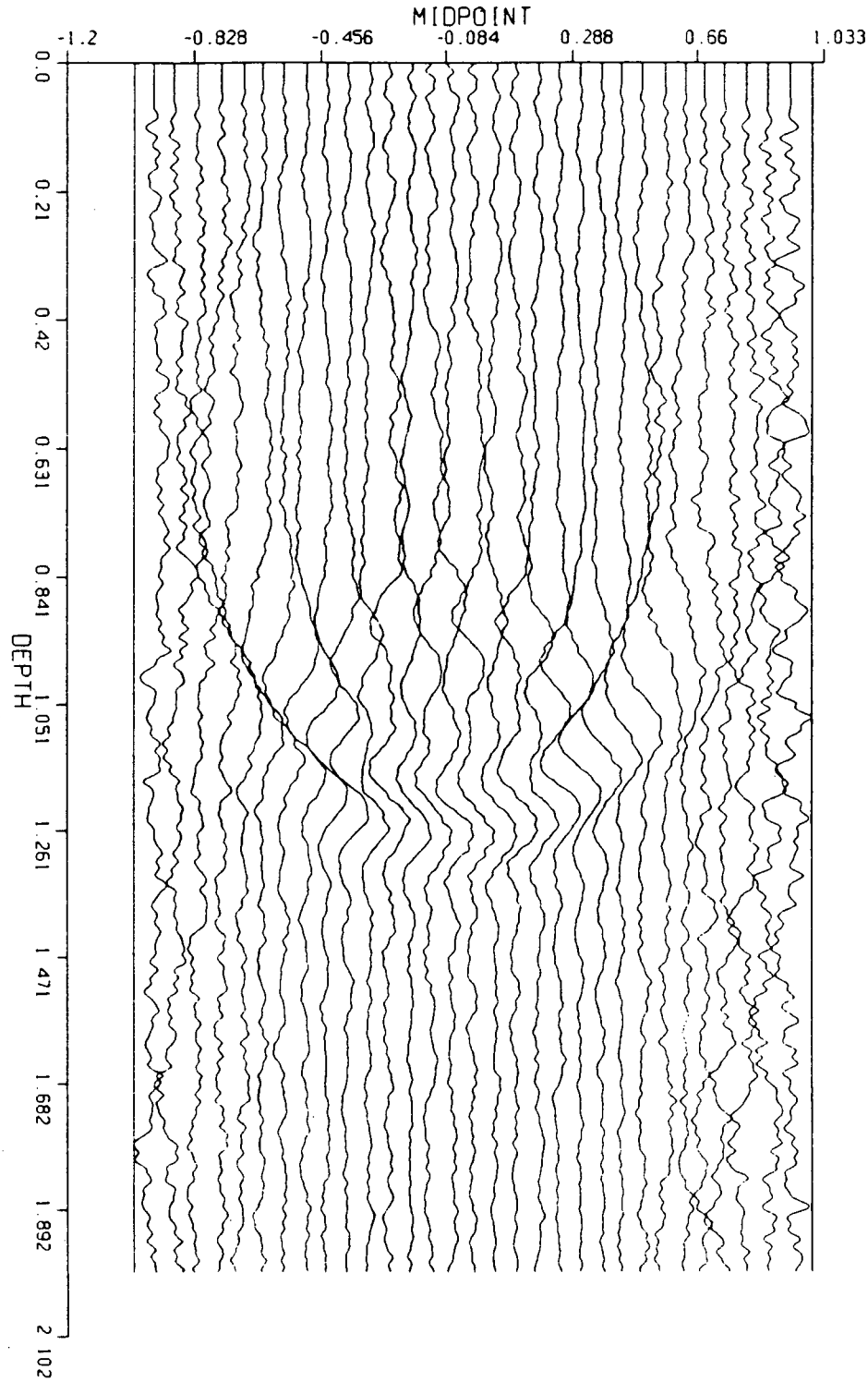

Figure 6.14 Finite difference migration of the second model problem.

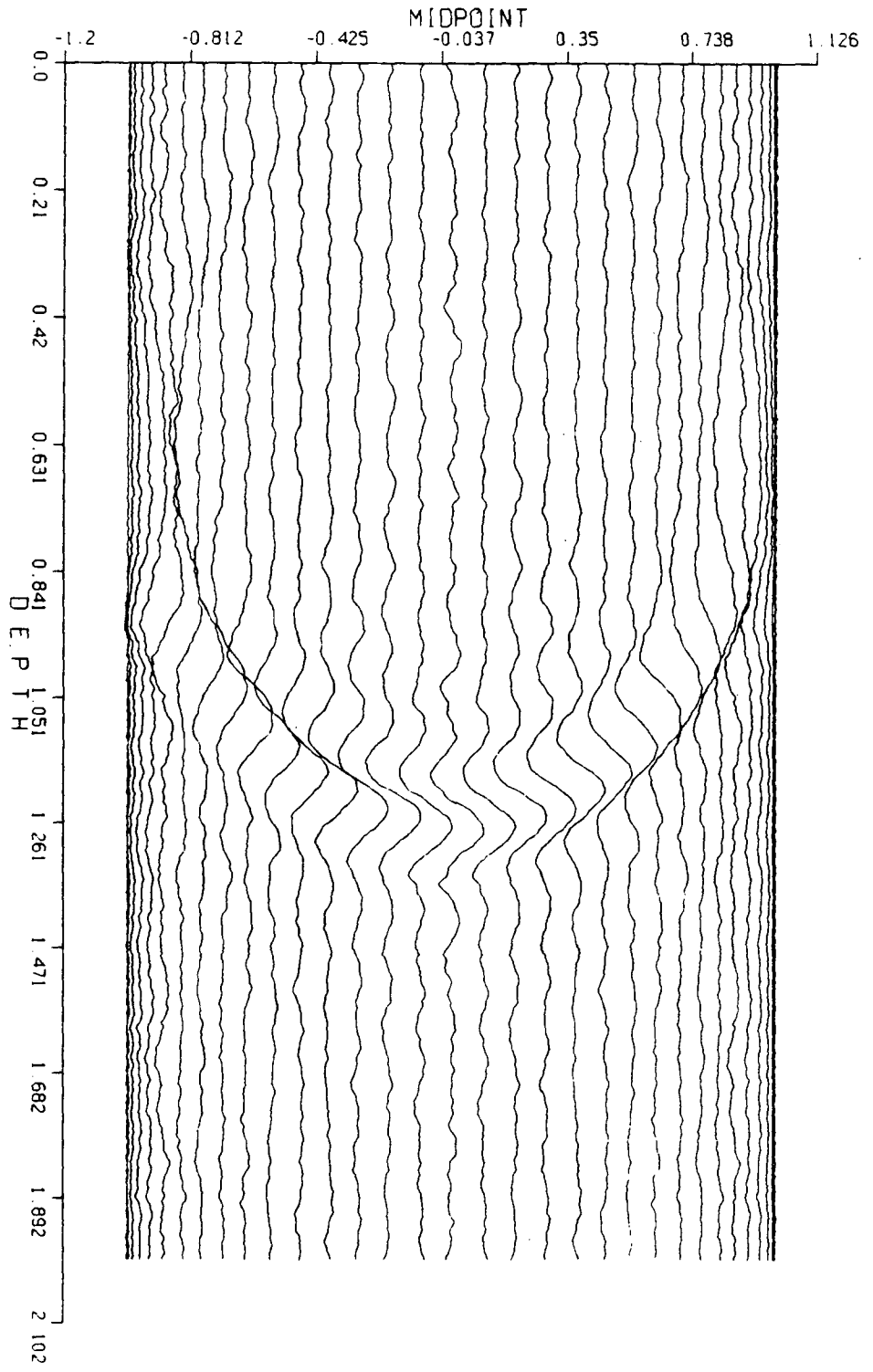

Figure 6.15 Chebychev migration of the second model problem.

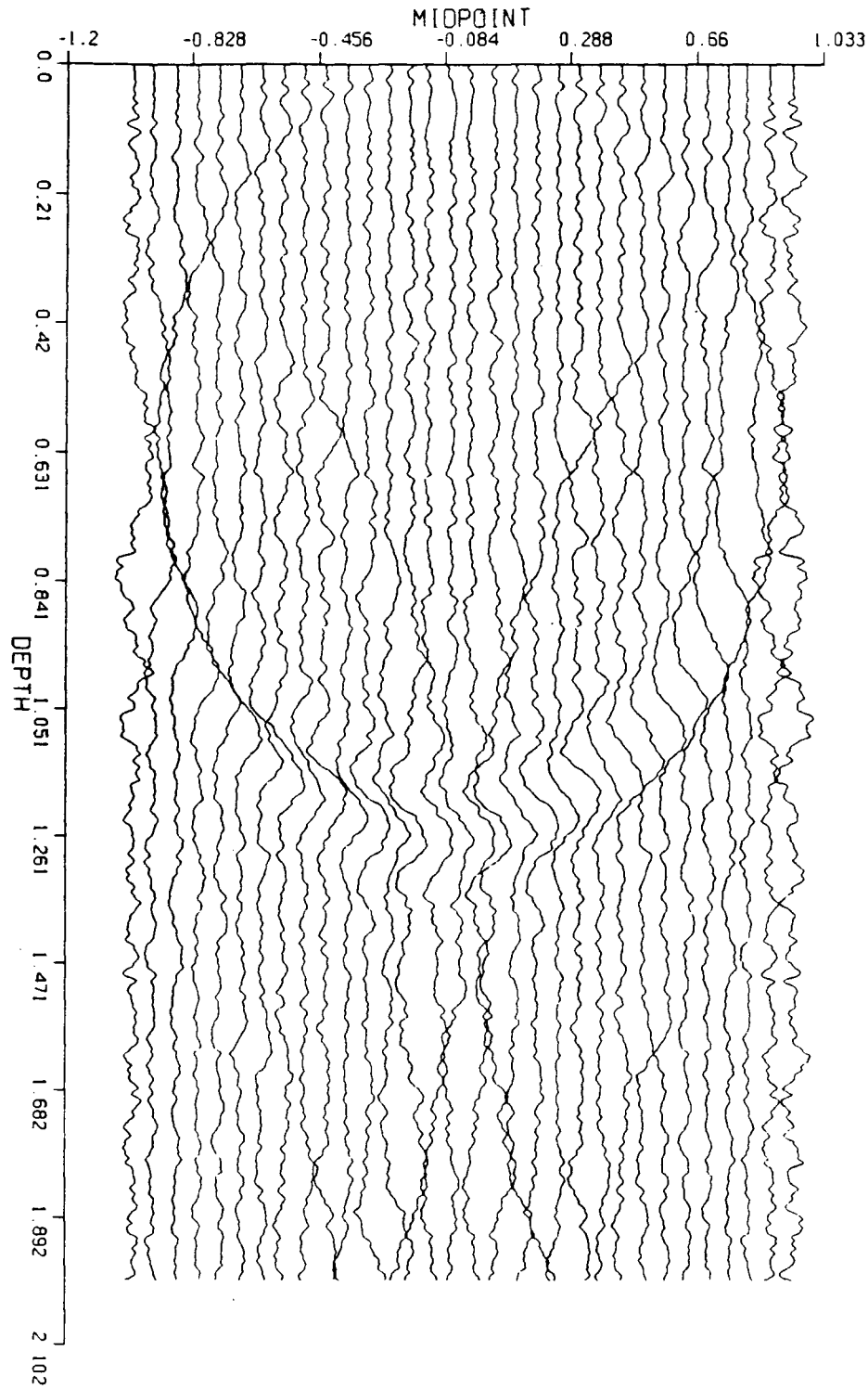

Figure 6.16 Fourier Crank-Nicolson migration of the second model problem.

(incorrect) direction. Similar considerations apply when more terms in (6.31) are retained. The validity of this analysis is demonstrated in figure (6.12) which concerns the 15° exact Fourier migration of the second input model where the evanescent region has not been filtered. High frequency noise has been introduced everywhere and the true semicircle has been augmented by another semicircle-artifact due to the improperly directed energy. Moreover, characteristic wraparound artifacts are clearly visible in figure (6.12). Filtering of the troublesome evanescent energy enhances the quality of the migrated section (figure 6.13) dramatically. Nonetheless, some "ringing" due to not tapering the filter edges and a mild ω-domain wraparound interference can be identified in figure (6.13).

The finite difference migration output shown in figure (6.14) is very poor as well. In addition to the intractable evanescent high frequency noise, severe dispersion contamination is also present. However, the improperly conceived semicircle of figure (6.12) is not as easy to identify and the noise is of relatively lower frequency than in the Fourier migration. The Chebychev section in figure (6.15) has been obtained via the **SLU1** procedure; the low frequencies have been properly handled and the computational burden of this Chebychev migration is comparable to the demands of the finite difference migration. The quality of figure (6.15) is striking indeed. The evanescent noise is basically of the same frequency as in the finite difference output; however, we should note that the boundary traces are noisier than the interior of the computational domain. The semicircular artifact is only slightly discernible again and the dispersion effects due to the finite difference approximation of the $\partial/\partial z$ operator are very mild too.

An interesting question arises. Why do figures (6.14) and (6.15) exhibit a lower frequency content than figure (6.15) does? The Fourier algorithm differs from the finite difference and the Chebychev procedures, in the sense that the former implements an exact solution for the Fresnel term, whereas the latter involve the Crank-Nicolson approximation which is well-known to be a primarily low frequency approximation to the true exponential solution. Thus, the evanescent modes are allowed only a limited exposure in the course of the extrapolation. A weak amplitude and a significant phase distortion characterize the semicircle-artifact, while an overall smoothing of the output section results as well. This smoothing is interrupted by spatial dispersion noise in the finite difference output (figure 6.12). The ongoing discussion points out the need for another numerical experiment (figure 6.16). There, the Fourier algorithm has been employed, but the available exact solution of the Fresnel term has been substituted with the corresponding Crank-Nicolson approximation. The Fourier eigenvalues of $\partial^2/\partial x^2$ are $-k_x^2$ and therefore, the Fourier Crank-Nicolson approximation reads

$$\exp\left(\frac{\imath v k_x^2}{2\omega}\right)\Delta z \simeq \frac{1+(\imath v k_x^2/4\omega)\Delta z}{1-(\imath v k_x^2/4\omega)\Delta z} \tag{6.32}$$

However, figure (6.16) features only a very minor smoothing with respect to the exact migration shown in figure (6.12), probably because of the persistent contaminating effect of the Fourier transform's periodicity. Incidentally, procedure **SLU1** cannot be held responsible for the lower frequency content of the Chebychev output of figure (6.15), since the high frequencies have actually been treated very accurately during the inversion associated with their depth extrapolation.

Finally, we should note that the stable Crank-Nicolson scheme is able to manifest the evanescent blow-up if the full square root operator is to be used. Then whenever $vk_x > \omega$, the Crank-Nicolson amplification factor, i.e

$$\xi = \frac{1 - \imath k_z a}{1 + \imath k_z a} \tag{6.33}$$

(with a real and $|a| \leq 1$) will lose its conjugate symmetry, since the wavenumber k_z turns imaginary. Thus, with $|\xi|$ exceeding unity, the stability criterion is not met and the evanescent modes will default to the correct growing exponentials. Obviously, for the 15° equation, k_z is always real and therefore the scheme is always stable; as a result, the evanescent aliasing results to the reported erroneous high frequency oscillations.

6.10.3 The balancing of the boundary condition row in SLU1

The **SLU1** fast inversion procedure has been seen to falter for large systems and large depth steps. The source of this instability has been identified in the off-diagonal dominance characterizing the underlying systems; this becomes more pronounced by the large multipliers involved in the pivotless forward elimination phase of the algorithm. At the very first elimination step the 1 of the last and the 0.25 of the first row yield a multiplier of 4. The situation gets worse as the elimination proceeds, since the last row is being successively magnified. This points to another question for future investigation. The last row of 1's corresponds to homogeneous boundary conditions and we may scale it arbitrarily, i.e $\varepsilon \sum a_n = 0$, without modifying the boundary structure of the system or altering its solution. The parameter ϵ should be chosen in such a way, that the magnitude of the last row is appropriately reduced and large multipliers are

avoided during the elimination. Simultaneously, care should be taken not to destabilize the conditioning of the system, since the matrix becomes singular in the limit of $\varepsilon = 0$ (one row of the matrix is zero).

6.11 A Synopsis of Results and Future Targets

Chebychev semi-discretizations have been investigated for both ordinary and partial differential equations. Comparisons with similar approaches involving finite difference and Fourier solutions have been performed and a great deal of insight into the details and the intricacies of the various implementations has been gained.

Chebychev schemes do not demand periodic boundary conditions, thus allowing the imposition of arbitrary boundary structures. Furthermore, they belong to the spectral category and thus they do not suffer from dispersion errors in the computation of derivatives. The tremendous impact of this fact on the accuracy and the efficiency of the solution has been first seen in the case of the Helmholtz equation. Finite differences require an extended oversampling to attain results comparable with the extemely accurate Chebychev solutions. Among the available projection choices, the Galerkin and the pseudospectral perform better than the simpler tau projections.

The time derivative appearing in the heat and the Schrödinger equations cannot be handled efficiently by explicit schemes, due to the very stiff character of the Chebychev systems; a viable alternative has been found in the absolutely stable Crank-Nicolson scheme. The superior performance of the Chebychev method has been clearly demonstrated for both problems, although severe computational considerations arise for large time calculations. A first answer to the problem has been given by the

tau-integrated implementation, which exhibits a quasi-tridiagonal character and is thereafore amenable to a very efficient inversion. In addition, a first investigation for one-dimensional Chebychev high-pass filtering has been undertaken, with promising results; a deeper analysis is required to extend these results. The Schrödinger problem also introduces the issue of artificial boundaries present in every discretization; Chebychev expansions are susceptible to such contamination as well.

Chapter VI explores the solution of the 15° migration equation. The approach involves a Fourier transform in time, a Chebychev in midpoint and a Crank-Nicolson finite difference scheme in depth. The familiar splitting technique is employed to account for the thin-lens and the diffraction contributions to the solution. The first results are of high quality and show promise for possible future work. Both the full dip one-way and the full two-way wave equations should be carefully investigated in the context of a Chebychev environment. A relevant dispersion relation needs to be discovered facilitating filtering in the (ω, k_θ) space; stretched coordinates may be employed to achieve a mode-decoupling so that a versatile expression will be obtained. Efficient reduction of the undesirable boundary reflections should be an indispensable component of future efforts; an associated stability analysis of the resulting equation is absolutely necessary. Direct solution of the large Chebychev systems is hopeless; the lack of alternative approaches tends to nullify any possibility of extrapolation in Chebychev space. Iterative techniques (perhaps the MG device) need to be considered to provide a satisfactory resolution of the problem. A final point requiring a closer look is the effect of the non-equidistant sampling, recommended in Chebychev simulations. Spatial aliasing, robustness and the undesirable boundary clustering of the Chebychev nodes need to be explored.

BIBLIOGRAPHY

Abarnabel S. S. and Murman M. E. (1982), Stability of two dimensional hyperbolic initial boundary value problems for explicit and implicit schemes, *J. Comput. Phys.* **48**, 160-167

Abramowitz M. and Stegun A. L. (1964), *Handbook of mathematical functions with formulas, graphs and mathematical tables*, Department of Commerce, U.S.A

Alford R. M., Kelly R. K. and Boore M. D. (1974), Accuracy of finite difference modeling of the acoustic wave equation, *Geophysics* **39**, 834-842

Ames F. W. (1977), *Numerical methods for partial differential equations*, Academic Press, NY

Askar A. (1981), A finite element method with local trigonometric basis for close coupling equations, *J. Chem. Phys.* **74** (11), 6133-6143

Askar A. and Cakmak S. A. (1978), Explicit integration method for the time-dependent Schrödinger equation for collision problems, *J. Chem. Phys.* **68** (6), 2794-2798

Bayliss A., Goldstein I. C. and Turkel E. (1983), An iterative method for the Helmholtz equation, *J. Comput. Phys.* **49**, 443-457

Bayliss A., Goldstein I. C. and Turkel E. (1985), On accuracy conditions for the numerical calculations of waves, *J. Comput. Phys.* **59**, 396-404

Bayliss A. and Turkel E. (1982), Far field boundary conditions for compressible flows, *J. Comput. Phys.* **48**, 182-199

Baysal E., Kosloff D. D., and Sherwood C. W. J. (1984), A two-way non-reflecting wave equation, *Geophysics* **49**, 132-141

Beam M. R., Warming F. R. and Yee C. H. (1982), Stability analysis of numerical boundary conditions and implicit difference approximations for hyperbolic equations, *J. Comput. Phys.* **48**, 200-222

Berkhout J. A. (1981), Wave-field extrapolation in seismic migration, a tutorial, *Geophysics* **46**, 1638-1656

Berkhout A. J. and van Wulfften Palthe D. W. (1979), Migration in terms of spatial deconvolution, *Geophys. Prospect.* **27**, 261-291

Bisseling R. and Kosloff R. (1985), The Fast Hankel Transform as a tool in the solution of the time-dependent Schrödinger equation, *J. Comput. Phys.* **59**, 136-151

Blottner G. F. (1982), Influence of boundary approximations and conditions on finite difference solutions, *J. Comput. Phys.* **48**, 246-269

Bolondi. J., Rocca F. and Savelli S. (1978), A frequency domain approach to two-dimensional migration, *Geophys. Prospect.* **26**, 750-772

Botha F. J. and Pinder F. G. (1983), *Fundamental concepts in the numerical solution of differential equations*, John Wiley and Sons, Inc

Bowdler J. H., Martin S. R., Peters G. and Wilkinson H. J. (1966), Solution of real and complex systems of linear equations, *Numerische Mathematik* **8**, 217-234

Boyd J. (1987), Spectral methods using rational basis functions on an infinite interval, *J. Comput. Phys.* **69**, 112-142

Bramhall H. M. and Casper M. B. (1970), Reflections on a wave packet approach to quantum mechanical barrier penetration, *Amer. J. Phys.* **38** (9), 1136-1148

Brandt A., Fulton R. S. and Taylor D.G. (1985), Improved spectral MG methods for periodic elliptic problems, *J. Comput. Phys.* **58**, 96-112

Brigham O. (1974), *The Fast Fourier Transform*, Prentice Hall, Inc, Englewoods Cliffs, New Jersey

Brown L. D. (1983), Applications of operator separation in reflection seismology, *Geophysics* **48**, 288-294

Canuto C. and Quarteroni A. (1985), Preconditioned minimal residual methods for Chebychev spectral calculations, *J. Comput. Phys.* **60**, 315-337

Cerjan C., Kosloff D., Kosloff R. and Reshef M. (1985), A nonreflecting boundary condition for the discrete acoustic and elastic equations, *Geophysics* **50**, 705-708

Chin Y. C. R., Hedstrom G. and Thigpen L. (1984), Numerical aspects in seismology, *J. Comput. Phys.* **54**, 18-56

Claerbout J. (1976), *Fundamentals of Geophysical Data Prospecting*, Mc Graw Hill Book Company, New York

Claerbout J. (1985), *Imaging the earth's interior*, Bracewell Scientific Publications, Oxford

Clayton R. W. and Enguist B. (1977), Absorbing boundary conditions for acoustic and elastic wave-equation: *Bull. Seism. Soc. Am.* **6**, 1529-1540

Clayton R. W. and Enguist B. (1980), Absorbing boundary conditions for wave-equation migration, *Geophysics* **45**, 895-904

Clement G. W. (1973), Basic principles of two-dimensional digital filtering, *Geophys. Prospect.* **21**, 125-141

Clenshaw W. C. and Norton J. H. (1963), The solution of ordinary differential equations in Chebychev series, *Comp. J.*, **6**, 88-92

Cooley J. W., Lewis P. A. and Welch P. D. (1970), The fast Fourier transform algorithm: programming considerations in the calculation of sine, cosine and Laplace transforms, *J. Sound Vib.* **12**, 315-337

Cooley J. W. and Tukey J. W. (1965), An algorithm for the machine computation of the complex Fourier series, *Math. Comp.* **19**, 297

Coughran M. W. Jr. (1984), On noncharacteristics boundary conditions for discrete hyperbolic initial boundary value problems, *J. Comput. Phys.* **60**, 135-154

Cushman-Roisin B. (1984), Analytical, linear stability criteria for the leap-frog, Dufort Frankel method, *J. Comput. Phys.* **53**, 227-239

Dablain A. M. (1986), The application of high order differencing to the scalar wave equation, *Geophysics* **51**, 54-66

Delic G. and Rawitscher H. G. (1985), Sturmian eigenvalue equations with a Chebychev polynomial basis, *J. Comput. Phys.* **57**, 188-209

Dennis R. C. S. and Quartapelle L. (1983), Direct solution of the vorticity-stream function ordinary differential equations by a Chebychev approximation, *J. Comput. Phys.* **52**, 448-463

Deville M. and Labrosse G. (1982), An algorithm for the evaluation of the multidimensional (direct and inverse) discrete Chebychev transforms, *J. Comput. Appl. Math.* **8** (4), 293-304

Deville M. and Mund E. (1985), Chebychev pseudospectral solution of second-order elliptic equations with finite-element preconditioning, *J. Comput. Phys.* **60**, 517-533

Diu B. (1980), Plane waves and wave packets in elementary quantum mechanical problems, *Eur. J. Phys.* **1**, 231-240

Drummond P. J., Hussaini Y. M. and Zang A. T. (1985), Spectral methods for modelling supersonic chemically reacting flow fields, *ICASE*, **Contract No. NAS1-17070**

Dubrulle A. A. (1983), Numerical methods for the migration of constant-offset sections in homogeneous and horizontally layered media, *Geophysics* **48**, 1195-1203

Elliot D. (1963), A Chebychev series method for the numerical solution of Fredholm integral equations, *Comp. J.* **6**, 102-110

Engquist B. and Majda A. (1977), Absorbing boundary conditions for the numerical simulation of waves, *Math. Comp.* **31**, 629-651

Engquist B. and Majda A. (1979), Radiation boundary conditions for acoustic and elastic wave calculations, *Comm. Pure and Appl. Math.* **32**, 313-357

Faddeev D. K. and Faddeeva V. N. (1963), *Computational methods of linear algebra*, W. H. Freeman and Co., San Francisco

Feit D. M., Fleck A. J. Jr. and Steiger A. (1982), Solution of the Schrödinger equation a spectral method, *J. Comput. Phys.* **47**, 412-433

Finlayson B. A (1972), *The method of weighted residuals and variational principles*, Academic Press, Inc, NY

Flatt H. P. (1961), Chain matrices and the Crank-Nicolson equation, in *Advances in Computers* **2** (F. L. Alt, ed.), Academic Press, NY

Fox L. (1962), *Numerical solution of ordinary and partial differential equations*, Pergamon Press, Oxford

Fox L. and Orszag A. S. (1973), Pseudospectral approximation to two-dimensional turbulence, *J. Comput. Phys.* **11**, 612-619

Fox L. and Parker B. I. (1968), *Chebychev polynomials in Numerical Analysis*, Oxford University Press, London

French W. S. (1975), Computer migration of oblique seismic reflection profiles, *Geophysics* **40**, 961-980

Frenkel A. (1983a), A Chebychev expansion of singular integral equations with a logarithmic kernel, *J. Comput. Phys.* **51**, 326-334

Frenkel A. (1983b), A Chebychev expansion of singular integrodifferential equations with a $\partial^2 \ln|s-t|/\partial s \partial t$ kernel, *J. Comput. Phys.* **51**, 335-342

Fulton R. S. and Taylor D. G. (1984), On the Gottlieb-Turkel time filter for Chebychev spectral methods, *J. Comput. Phys.* **55**, 302-312

Galbraith I., Ching S. Y. and Abraham E. (1984), Two-dimensional time-dependent quantum-mechanical scattering event, *Amer. J. Phys.* **52** (1), 60-68

Gazdag J. (1978), Wave equation migration with the Phase Shift method, *Geophysics* **43**, 1342-1351

Gazdag J. (1981), Modeling of the acoustic wave equation with transform methods, *Geophysics* **46**, 854-859

Gazdag J. and Sguazzero P. (1983), Migration of seismic data by phase shift plus interpolation, *Seismic Acoustic Laboratory* **11**, University of Houston

Gazdag J. and Sguazzero P. (1984), Migration of seismic data, *Proceedings of the IEEE* **72**, 1302-1315

Gear W. C. (1971), *Numerical initial value problems in ordinary differential problems*, Prentice Hall, Inc, Englewood Cliffs, NJ

Gentleman M. W. (1972), Implementing Clenshaw-Curtis quadrature, I. Methodology and Experience II. Computing the Cosine Transform, *Comm. ACM* **15** (5), 337-346

Goldberg A., Schey M. H. and Schwartz L. J. (1967), Computer-generated motion pictures of one-dimensional quantum-mechanical transmission and reflection phenomena, *Amer. J. Phys.* **35** (3), 177-186

Golub H. G. and van Loan F. C. (1983), *Matrix computations*, John Hopkins Univ. Press, Baltimore, Maryland

Gottlieb D. (1981), The stability of Pseudospectral Chebychev methods, *Math. Comp.* **36** (153), 107-118

Gottlieb D. and Gustaffson B. (1976), Generalized Dufort-Frankel methods for parabolic initial-boundary value problems, *SIAM. J. Numer. Anal.* **13** (1), 129-144

Gottlieb D. and Lustman L.(1983a), The Dufort-Frankel Chebychev method for parabolic initial-boundary value problems, *Computers and Fluids* **11**, 107-120

Gottlieb D. and Lustman L. (1983b), The spectrum of the Chebychev collocation operator for the heat equation, *SIAM J. Numer. Anal.* **20** (5), 909-921

Gottlieb D., Lustman L. and Street G. (1984a), Two dimensional shocks in *Proceedings of the Symposium on spectral methods for PDE* (R.Voigt, D.Gottlieb and M. Y. Hussaini, Eds), SIAM Monograph, Philadelphia

Gottlieb D.,Hussaini Y. M. and Orszag A. S. (1984b), Theory and application of spectral methods in *Proceedings of the Symposium on spectral methods for PDE* (R.Voigt, D.Gottlieb and M. Y. Hussaini, Eds), SIAM Monograph, Philadelphia

Gottlieb D. and Orszag A. S. (1977), *Numerical analysis of spectral methods: Theory and Applications*, CBMS-NSF,Regional Conference Series in Applied Mathematics, SIAM 1977

Gottlieb D. and Turkel E. (1980), On time-discretizations for spectral methods, *Stud. Appl. Math.* **63**, 67-86

Grinstein F. F., Rabitz H. and Askar A. (1983), The multigrid method for accelerated solution of the discretized Schrödinger equation, *J. Comput. Phys.* **51**, 423-443

Grosch E. C. and Orszag A. S. (1977), Numerical solution of problems in unbounded regions: Coordinate transforms, *J. Comput. Phys.* **25**, 273-296

Gustafsson B. (1982), The choice of numerical boundary conditions for hyperbolic systems, *J. Comput. Phys.* **48**, 270-283

Gustaffson B., Kreiss O. H. and Sundstrom A. (1972), Stability theory of difference approximations for mixed initial boundary value problems II, *Math. Comput.* **26**, 649

Hagedoorn G. (1954), A process of seismic reflection interpretation, *Geophys.Prospect.* **2**, 85-127

Haidvogel D. (1977), Quasigeostrophic regional and general circulation modelling: An efficient pseudospectral approximation technique, in *Computing methods for geophysical mechanics* (R.P.Shaw,Ed) **25**, American Society of Mechanical Engineers

Haidvogel B. D. and Zang T. (1979), The accurate solution of Poisson's equation by expansion in Chebychev polynomials, *J. Comput. Phys.* **30**, 167-180

Haldenwang P., Labrosse G., Abboudi S. and Deville M. (1984), Chebychev 3-D spectral and 2-D pseudospectral solvers for the Helmholtz equation, *J. Comput. Phys.* **55**, 115-128

Hale D. (1983), Dip-moveout by Fourier Transform, *Geophysics* **49**, 741-757

Hamilton C. J., Schwartz L. J. and Bowers A. W. (1972), Computer- generated films for solid-state physics, *Amer. J. Phys.* **40**, 1656-1661

Hatton L., Worthington H. M. and Makin J. (1986), *Seismic Data Processing*, Blackwell Scientific Publications, Oxford

Hatziavramidis D. and Ku H-C. (1983), Pseudospectral solutions of laminar heat transfer problems in pipelines, *J. Comput. Phys.* **52**, 414-424

Hood P. (1978), Finite difference and wavenumber migration, *Geophys. Prospect.* **26**, 773-789

Hussaini Y. M., Street L. G. and Zang A. T. (1983), Spectral methods for partial differential equations, *ICASE*, **Report No. 83-46**

Hussaini Y. M. and Zang A. T. (1984), Iterative spectral methods and spectral solutions to compressible flows in *Proceedings of the Symposium on spectral methods for PDE* (R.Voigt, D.Gottlieb and M. Y. Hussaini, Eds), SIAM Monograph, Philadelphia

Israeli M. and Orszag A. A. (1981), Approximation of radiation boundary conditions, *J. Comput. Phys.* **41**, 115-135

Kanasewich R. E. (1981), *Time sequence analysis in geophysics*, The University of Alberta press, Edmonton

Keast P. and Mitchell R. A. (1967), Finite difference solution of the third boundary problem in elliptic and parabolic equations, *Numer. Math.* **10**, 67-75

Keller B. H. (1968), *Numerical methods for two-point boundary-value problems*, Blaisdell Publishing Company

Kelly R. K., Ward W. R., Treitel S. and Alford M. R. (1976) Synthetic seismograms: A finite difference approach, *Geophysics* **41**, 2-27

Keys G. R. (1985), Absorbing boundary conditions for acoustic media, *Geophysics* **50**, p.892-902

Kizner W. (1964), Error curves for Lanczos' "selected points" method, *Comp. J.* **8**, 372

Kosloff D. and Kosloff R. (1983a), A Fourier method solution for the time-dependent Schrödinger equation as a tool in molecular dynamics, *J. Comput. Phys.* **52**, 35-53

Kosloff R. and Kosloff D. (1983b), A Fourier method solution for the time-dependent Schrödinger equation: A study of the reaction $H^+ + H_2, D^+ + HD$, and $D^+ + H_2$, *J. Chem. Phys.* **79** (4), 1823-1833

Kosloff R. and Kosloff D. (1986), Absorbing boundaries for wave propagation problems, *J. Comput. Phys.* **63**, 363-376

Kosloff D. D. and Baysal E. (1982), Forward modeling by a Fourier method, *Geophysics* **47**, 1402-1412

Kosloff D. D. and Baysal E. (1983), Migration with the full acoustic wave equation, *Geophysics* **48**, 677-687

Ku H-C. and Hatziavramidis D. (1984), Chebychev expansion methods for the solution of the extended Graetz problem, *J. Comput. Phys.* **56**, 495-512

Kulander C. K. (1978), Collision induced dissociation in collinear $H + H_2$: Quantum mechanical probabilities using the time-dependent wavepacket approach, *J. Chem. Phys.* **69** (11), 5064-5072

Lanczos C. (1938), Trigonometric interpolation of empirical and analytical functions, *J. Math. Phys.* **17**, 123-199

Lanczos C. (1957), *Applied Analysis*, Prentice Hall, Englewood Cliffs, NJ

Lanczos C. (1973), Legendre versus Chebychev polynomials, in *Proc. Roy. Irish Acad. Conf. Numer. Anal.* (John J. H. Miles, Ed.), Academic Press

Lawden F. D. (1967), *The mathematical principles of Quantum Mechanics*, Methuen and Co. Ltd

Le Bail C. R. (1971), Use of fast Fourier transforms for solving partial differential equations in physics, *J. Comput. Phys.* **9**, 440-465

Le Quere P. and De Roquefort A. T. (1985), Computation of natural convection in two dimensional cavities with Chebychev polynomials, *J. Comput. Phys.* **57**, 210-228

Lindman E. L. (1975), "Free-Space" boundary conditions for the time dependent wave equation, *J. Comput. Phys.* **18**, 66-78

Livne E. and Glasner A. (1985), A finite difference scheme for the heat conduction equation, *J. Comput. Phys.* **58**, 59-66

Lyness N. J. (1984), The calculation of trigonometric Fourier coefficients, *J. Comput. Phys.* **54**, 57-73

Lynn B. H. and Deregowski S. (1981), Dip limitations on migrated sections as a function of line length and recording time, *Geophysics* **46**, 1392-1397

Lysmerand J. and Kuhlemeyer R. L. (1969), Finite dynamic model for infinite media, *J. Eng. Mech. Div. Proc. Am. Soc. Civil Eng.* **95**, 859-877

Loewenthal D., Lu L., Robertson R. and Sherwood J. (1976), The wave equation applied to migration, *Geophys. Prospect.* **24**, 380-399

Macaraeg G. M. (1986), A mixed pseudospectral-finite difference method for the axisymmetric flow in a heated, rotating spherical shell, *J. Comput. Phys.* **62**, 297-320

Marchuk I. G. (1982), *Methods of numerical mathematics*, Spinger-Verlag, NY

Mary St. F. D. and Lee D. (1985), Analysis of an implicit finite-difference solution to an underwater wave propagation problem, *J. Comput. Phys.* **57**, 378-390

Mc Calla R. T. (1967), *Introduction to numerical methods and FORTRAN programming*, John Wiley & Sons, Inc.

McCrory L. R. and Orszag A. S. (1980), Spectral methods for the multi-dimensional diffusion problems, *J. Comput. Phys.* **37**, 93-112

Mc Cullough A. E.,Jr, and Wyatt E. R. (1971a), Dynamics of collinear $H+H_2$ reaction. I. Probability density and flux, *J. Chem. Phys.* **54** (8), 3578-3591

Mc Cullough A. E.,Jr, and Wyatt E. R. (1971b), Dynamics of collinear $H+H_2$ reaction. I. Probability density and flux, *J. Chem. Phys.* **54** (8), 3792-3600

Merilees E. P. (1973), The pseudospectral approximation applied to the shallow water equations on a sphere, *Atmosphere* **11** (1), 13-20

Mitchell R. A. and Griffiths F. D. (1980), *The finite difference method in partial differential equations* , John Wiley and Sons

Miller P. C. J. (1946), Two numerical applications of Chebychev polynomials,*Proc. Roy. Soc. Edinburgh* **62**, 204-210

Moin P. and Kim J. (1980), On the numerical solution of time dependent viscous incompressible flud flows involving solid boundaries, *J. Comput. Phys.* **35**, 381-392

Morchoisne Y. (1984), Inhomogeneous flow calculations by spectral methods: mono-domain and multi-domain techniques in *Proceedings of the Symposium on spectral methods for PDE* (R.Voigt, D.Gottlieb and M. Y. Hussaini, Eds), SIAM Monograph, Philadelphia

Nautiyal A. (1986), *Aspects of spatial wavelets and their application to modelling seismic reflection data*, M.Sc. thesis, University of British Columbia

Nicolaides R. (1979), On some theoretical and practical aspects of multigrid methods, *Math. Comp.* **33**, 933-952

Norton J. H. (1964), The iterative solution of non-linear ordinary differential equations in Chebychev series, *Comp. J.* **7**, 76-85

Orszag A. S. (1969), Numerical methods for the simulation of turbulence, *Phys. Fluids (Suppl. II)* **12**, 250-257

Orszag A. S. (1970), Transform methods for the calculation of vector-coupled sums: Application to the spectral form of the vorticity equation, *J. Atmos. Sci.* **27**, 890-895

Orszag A. S. (1971a), Numerical simulation of incompressible flows within simple boundaries. I. Galerkin (spectral) representations, *Stud. Appl. Math.* **L** (4), 293-327

Orszag A. S. (1971b), Numerical simulation of incompressible flows within simple boundaries: accuracy,*J. Fluid Mech.* **49**, 75-112

Orszag A. S. (1971c), Galerkin approximations to flows within slabs, spheres and cylinders, *Phys. Rev. Letters* **26** (18), 1100-1103

Orszag A. S. (1971d), Accurate solution of the Orr-Sommerfeld stability equation, *J. Fluid Mech.* **50**, 689-703

Orszag A. S. (1971e), On the resolution requirements of finite differences schemes, *Stud. Appl. Math.* **L** (4), 395-397

Orszag A. S. (1972), Comparison of pseudospectral and spectral approximation, *Stud. Appl. Math.* **LI** (3), 253-259

Orszag A. S. (1980), Spectral methods for problems in complex geometries, *J. Comput. Phys.* **37**, 70-92

Orszag A. S. and Israeli M. (1974), Numerical simulation of viscous incompressible flows, *Ann. Rev. Fluid Mech.* **5**, 281-318

Orszag A. S. and Jayne W. L. (1974), Local errors of difference approximations to hyperbolic equations, *J. Comput. Phys.* **14**, 93-103

Orszag A. S. and Kells C. L. (1980), Transition to turbulence in plane Poiseuille and plane Couette flow, *J. Fluid Mech.* **96**, 159-205

Osher S. (1984), Smoothing for spectral methods in *Proceedings of the Symposium on spectral methods for PDE* (R.Voigt, D.Gottlieb and M. Y. Hussaini, Eds), SIAM Monograph, Philadelphia

Ottolini R. and Claerbout F. J. (1984), The migration of common midpoint slant stacks, *Geophysics* **49**, 237-249

Pann K., Shin Y. and Eisner E. (1979), A collocation formulation of wave equation migration, *Geophysics* **44**, 712-721

Panov J. D. (1963), *Formulas for the numerical solution of partial differential equations by the method of differences*, Frederick Ungar Publishing Co, NY

Patterson S. G. and Orszag A. S. (1971), Spectral calculations of isotropic turbulence: efficient removal of aliasing interactions, *Phys. Fluids* **14**, 2538-2541

Peaceman W. D. (1977), *Fundamentals of numerical reservoir simulation*, Elsevier Scientific Publishing Company

Pizer M. S. and Wallace L. V. (1983), *To compute numerically; concepts and strategies*, Little, Brown and Company

Press H. W., Flanney P. B., Tenkolsky A. S. and Vetterling T. W. (1985), *Numerical Recipies — The art of scientific programming* Cambridge University Press

Price S.H. and Varga S. R. (1974), Errors bounds for semi-discrete Galerkin approximations of parabolic problems with applications to petroleum reservoir mechanics, in *Numerical Solution of Field Problems in Continuum Physics*, American Mathematical Society, Providence

Ralston A. and Rabinowitz P. (1978), *A first course in numerical analysis*, McGraw-Hill Company

Reshef M. and Kosloff D. (1986), Migration of common-shot gathers, *Geophysics* **51**, 324-331

Reynolds C. A. (1978), Boundary conditions for the numerical solution of wave propagation problems, *Geophysics* **43**, 1099-1110

Rivlin J. T. (1974), *The Chebychev polynomials*, Wiley, New York

Schammel H. and Elsasser K. (1976), The application of the spectral method to nonlinear wave propagation, *J. Comput. Phys.* **22**, 501-516

Schneider W. A. (1978), Integral formulation for migration in two and three dimensions, *Geophysics* **43**, 49-76

Scraton E. P. (1964), The solution of linear differential equations in Chebychev series, *Comp. J.* **8**, 57-61

Sheriff E. R. and Geldart P. L. (1984), *Exploration seismology*, Cambridge Univ. Press, Cambridge

Shoucri M. and Knopr G. (1974), Numerical integration of the Vlasov equation, *J. Comput. Phys.* **14**, 84-92

Smith W. D. (1974), A non-reflecting plane boundary for wave propagation problems, *J. Comput. Phys.* **15**, 492-503

Sochacki J., Kubichek R., George J., Fletcher R. W. and Smithson S. (1987), Absorbing boundary conditions and surface waves, *Geophysics* **52**, 60-71

Spiegel R. M. (1968), *Mathematical handbook of formulas and tables*, Schaum's outline series in mathematics, Mc Graw-Hill book company

Stolt H. R. (1978), Migration by Fourier transform, *Geophysics*, **43**, 23-48

Strang G. and Fix J. G. (1973), *An analysis of the finite element method*, Prentice Hall, Englewood Cliffs, NJ

Street L. G., Zang A. T. and Hussaini Y. M. (1985), Spectral MG methods with applications to transonic potential flow, *J. Comput. Phys.* **57**, 43-76

Tal-Ezer H. (1984), Spectral methods in time for hyperbolic equations, *ICASE*, **Report 172302**, NASA Langley Research Center

Tal-Ezer H. and Kosloff R. (1984), An accurate and efficient scheme for propagating the time-dependent Schrödinger equation, *J. Chem. Phys.* **81** (9), 3967-3971

Taylor D. T. (1984), Recent advances in pseudo-spectral methods, in *Proceedings of the Symposium on spectral methods for PDE* (R.Voigt, D.Gottlieb and M. Y. Hussaini, Eds), SIAM Monograph, Philadelphia

Tolstoy I. (1973), *Wave propagation*, Mc Graw Hill

Tomonaga S. I. (1966), *Quantum Mechanics*, North Holland Publishing Company

Trefethen N. L. (1982), Group velocity in finite difference schemes, *SIAM Review* **24**, 113-133

Trefethen N. L. (1983), Group velocity interpretation of the stability theory of Gustafsson, Kreiss and Sundstrom, *J. Comput. Phys.* **49**, 199-217

Vemuri V. and Karplus J. W. (1981), *Digital computer treatment of partial differential equations*, Prentice Hall, Inc, Englewood Cliffs, NJ

Vichnevetsky R. (1981), *Computer methods for partial differential equations, Volume 1: Elliptic equations and the finite element method*, Prentice Hall Whitham B. G. (1974), *Linear and Nonlinear Waves*, Wiley ans Sons, N.Y

Wright K. (1964), Chebychev collocation methods for ordinary differential equations, *Comp. J.* **6**, 358-365

Wong S. Y., Zang A. T. and Hussaini Y. M. (1983), Efficient iterative techniques for the solution of spectral methods, in *Symposium on spectral methods in partial differential equations*, ICASE, NASA Langley Research Center

Wood W. L. and Lewis W. R. (1975), A comparison of marching schemes for the transient heat conduction equation, *Int. J. Num. Meth. Engng.* **9**, 679-689

Yedlin M. (1985, 1986) personal communication.

Yilmaz O. and Claerbout F. J. (1980), Prestack partial migration, *Geophysics* **45**, 1753-1779

Zang A. T., Wong S. Y. and Hussaini Y. M. (1982), Spectral MG methods for elliptic equations, *J. Comput. Phys.* **48**, 485-501

Zang A. T., Wong S. Y. and Hussaini Y. M. (1984), Spectral MG methods for elliptic equations II, *J. Comput. Phys.* **57**, 485-501

Zebib A. (1984), A Chebychev method for the solution of boundary value problems, *J. Comput. Phys.* **53**, 443-455

APPENDIX A

A.1 Analytic Evaluation of $\langle T_n, T_m'' \rangle$.

The analytic evaluation of the inner product of a Chebychev polynomial $T_n(x)$ with the second derivative of another Chebychev polynomial $T_m(x)$, is given below. The inner product of interest, is $\langle T_n, T_m'' \rangle$ or (neglecting normalization constants)

$$\int_{-1}^{+1} T_n(x) \frac{d^2}{dx^2} T_m(x) (1 - x^2)^{-1/2} \, dx \qquad (A.1)$$

Substituting $x = \cos\theta$ and calculating the second derivative of T_m, yields

$$-m \int_0^{\pi} \left[\frac{m \cos n\theta \cos m\theta}{\sin^2 \theta} - \frac{\cos n\theta \sin m\theta \cos \theta}{\sin^3 \theta} \right] d\theta \qquad (A.2)$$

Both integrands are singular; the corresponding integrals are divergent and, therefore, direct integration is prohibited. Instead, we consider both integrands together, in order to achieve a mutual cancellation of the troublesome singularities. Their current form though, hinders this desirable elimination and it necessitates the application of a decomposition which would allow the cancellation to take place.

The first integrand contains a $\cos n\theta \cos m\theta$ term and a $\cos n\theta \sin m\theta$ term is included in the second.

The idea is to expand both of these products into power series of $\cos\theta \sin\theta$; to accomplish that we combine the results of applying both the De Moivre's and the binomial theorems for the expansion of $(\cos\theta + i\sin\theta)^n$, as described below.

According to De Moivre's theorem, we write

$$(\cos\theta + i\sin\theta)^n = \cos n\theta + i\sin n\theta \qquad (A.3)$$

Now, applying the binomial theorem, we get

$$(\cos\theta + i\sin\theta)^n = (\cos\theta)^n + \binom{n}{1}(\cos\theta)^{n-1}(i\sin\theta) + \binom{n}{2}(\cos\theta)^{n-2}(i\sin\theta)^2 + \cdots + (i\sin\theta)^n \qquad (A.4)$$

where

$$\binom{n}{k} = \binom{n}{n-k} = \frac{n!}{k!(n-k)!} \quad \text{for } n, k \in N \qquad (A.5)$$

Equating both the real and the imaginary parts of equations (A.3) and (A.4), we obtain the following expansions for $\cos\theta$ and $\sin\theta$ respectively,

$$\cos n\theta = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n}{2k}(-1)^k(\cos\theta)^{n-2k}(\sin\theta)^{2k}, \quad \text{for } n \geq 0 \quad \text{and even} \qquad (A.5)$$

and

$$\sin n\theta = \sum_{k=0}^{[n-1/2]} \binom{n}{2k+1} (-1)^k (\cos\theta)^{n-(2k+1)} (\sin\theta)^{2k+1}, \quad \text{for } n \geq 1 \quad \text{and odd} \tag{A.6}$$

where [ ] denotes integer part.

Let us now expand $\cos n\theta$, $\cos m\theta$ and $\sin m\theta$ according to (A.5) and (A.6), respectively.

$$\cos n\theta = \sum_{k=0}^{N} a_k (\cos\theta)^{n-2k} (\sin\theta)^{2k} \tag{A.7}$$

$$\cos m\theta = \sum_{l=0}^{L} c_l (\cos\theta)^{m-2l} (\sin\theta)^{2l} \tag{A.8}$$

and

$$\sin m\theta = \sum_{j=0}^{M} b_j (\cos\theta)^{m-(2j+1)} (\sin\theta)^{2j+1} \tag{A.9}$$

where the coefficients are given as

$$a_k = (-1)^k \binom{n}{2k}, \quad c_l = (-1)^l \binom{m}{2l}, \quad b_j = (-1)^j \binom{m}{2j+1} \tag{A.10}$$

with

$$N = \left[\frac{n}{2}\right], \quad L = \left[\frac{m}{2}\right], \quad \text{and } M = \left[\frac{m-1}{2}\right] \tag{A.11}$$

Multiplying now (A.7) with (A.8) gives

$$\cos n\theta \cos m\theta = \sum_{k=0}^{N} \sum_{l=0}^{L} a_k\, c_l\, (\cos\theta)^{n+m-(2k+2l)} (\sin\theta)^{2k+2l} \tag{A.12}$$

and multiplying (A.7) with (A.9) gives

$$\cos n\theta \sin m\theta = \sum_{k=0}^{N} \sum_{j=0}^{M} a_k b_j (\cos\theta)^{n+m-(2k+2j+1)} (\sin\theta)^{2k+2j+1} \qquad (A.13)$$

The next step involves forming the complete integrands. Therefore we multiply (A.12) with $1/\sin^2\theta$ and (A.13) with $\cos\theta/\sin^3\theta$, to obtain (A.14) and (A.15) respectively.

$$\frac{\cos n\theta \cos m\theta}{\sin^2\theta} = \sum_{k=0}^{N} \sum_{l=0}^{L} a_k c_l (\cos\theta)^{(n+m-2)-(2k+2l-2)} (\sin\theta)^{2k+2l-2} \qquad (A.14)$$

and

$$\frac{\cos n\theta \sin m\theta \cos\theta}{\sin^3\theta} = \sum_{k=0}^{N} \sum_{j=0}^{M} a_k b_j (\cos\theta)^{(n+m-2)-(2k+2j-2)} (\sin\theta)^{2k+2j-2} \qquad (A.15)$$

Substituting (A.14) and (A.15) into (A.2) and making some suitable rearrangements, yields

$$\begin{aligned} -m \sum_{k=0}^{N} a_k \Big[m \sum_{l=0}^{L} c_l \int_0^{\pi} (\cos\theta)^{(n+m-2)-(2k+2l-2)} (\sin\theta)^{2k+2l-2}\, d\theta \\ - \sum_{j=0}^{M} b_j \int_0^{\pi} (\cos\theta)^{(n+m-2)-(2k+2j-2)} (\sin\theta)^{2k+2j-2}\, d\theta \Big] \end{aligned} \qquad (A.16)$$

Some important points concerning the above formula should be identified. For $m = 0$ or $m = 1$ the integral vanishes identically. Singularities at both integrands

are still present and they correspond to the combinations $k = l = 0$ and $k = j = 0$ yielding a $\sin^{-2}\theta$ in the denominators. The advantage of this formula is obvious from the fact that it unfolds the removable singularity nature of the prohibited combination $k = l = j = 0$. Evidently, the singularities cancel out and the remaining combinations of indices form an integrand, which gives rise to a convergent integral.

Moving towards a matrix formulation of the problem, we define

$$d_{kl} = \int_0^{\pi} (\cos\theta)^{\alpha-\beta} (\sin\theta)^{\beta}\, d\theta \tag{A.17}$$

and

$$f_{kj} = \int_0^{\pi} (\cos\theta)^{\alpha-\delta} (\sin\theta)^{\delta}\, d\theta \tag{A.18}$$

with

$$\alpha = n + m - 2,\ \beta = 2k + 2l - 2,\ \delta = 2k + 2j - 2 \tag{A.19}$$

The expression to be evaluated is

$$-m\left[\sum_{k=0}^{N} a_k \left(m \sum_{l=0}^{L} c_l d_{kl} - \sum_{j=0}^{M} b_j f_{kj}\right)\right] \tag{A.20}$$

where $d_{00} = f_{00} \equiv 0$ to account for the removal of the singularities.

To complete this procedure, we only need to consider the integration of an integral of the form

$$I(\alpha - \beta, \beta) = \int_0^{\pi} (\cos\theta)^{\alpha-\beta} (\sin\theta)^{\beta}\, d\theta \tag{A.21}$$

This is evaluated recursively as (Spiegel, 1968)

$$I\Big(\alpha-\beta,\beta\Big)=\frac{\alpha-(\beta+1)}{\alpha}\,I\Big(\alpha-(\beta+2),\beta\Big) \qquad (A.22)$$

Although (A.22) looks quite innocent, extreme caution needs to be exercised. An in-depth analysis of complications emerging during the recursion, helped us to reveal some tricky points and allowed an efficient routine to be written. After this algorithm had been implemented successfully, an idea based on the use of beta functions for the evaluation of those last integrals was proposed (Yedlin, 1986). The efficiency of the previous algorithm can be greatly enhanced if these integrals are to be evaluated directly through built in beta function routines. The reason why this is possible is obvious from the following definition of the beta function

$$\mathrm{B}(z,w)=2\int_0^{\pi/2}(\cos\theta)^{2z-1}(\sin\theta)^{2w-1}\,d\theta \quad \text{for } \mathcal{R}\,(z),\mathcal{R}\,(w)>0 \qquad (A.23)$$

(Abramowitz and Stegun, 1964).

The beta function is also connected with the gamma function through the relation

$$\mathrm{B}(z,w)=\frac{\Gamma(z)\Gamma(w)}{\Gamma(z+w)}=\mathrm{B}(w,z) \qquad (A.24)$$

(Abramowitz and Stegun, 1964).

A.2 Analytic Evaluation of $T''_m(x_n)$

The direct evaluation of $T''_m(x)$ on a collocation-point set $\{x_n\}$, which includes the endpoints $x = +1$ and $x = -1$ or $\theta = 0$ and $\theta = \pi$, respectively, has to cope with the problems associated with the presence of poles of second order at these boundary points. The explicit form of $T''_m(x_n)$

$$\frac{d^2}{dx^2}T_m(x) = -m\left[\frac{m\cos m\theta}{\sin^2\theta} - \frac{\sin m\theta\cos\theta}{\sin^3\theta}\right] \tag{A.25}$$

shows that both ratios forming the above expression feature these singularities.

Following the previous procedure, we transform the above expression using the expansions (A.9) and (A.10); the coefficients and the limits of the expansions are given according to (A.11) and (A.12), respectively.

This procedure results in the expressions

$$\frac{\cos m\theta}{\sin^2\theta} = \sum_{l=0}^{L} c_l\,(\cos\theta)^{(m-2l)}(\sin\theta)^{(2l-2)} \tag{A.26}$$

$$\frac{\sin m\theta\cos\theta}{\sin^3\theta} = \sum_{j=0}^{M} b_j\,(\cos\theta)^{(m-2j)}(\sin\theta)^{(2j-2)} \tag{A.27}$$

The removable singularity nature of the poles $\theta = 0, \pi$ is now revealed; the troublesome poles correspond to $l = j = 0$ and it may readily be seen that they cancel each other out.

The final output is obtained through the formula

$$-m\Big[\sum_{l=0}^{L} c_l e_l - \sum_{j=0}^{M} b_j g_j\Big] \tag{A.28}$$

where

$$e_l = (\cos\theta)^{m-2l}(\sin\theta)^{2l-2} \quad \text{and} \quad g_j = (\cos\theta)^{m-2j}(\sin\theta)^{2j-2} \tag{A.29}$$

The imposition of $e_0 = g_0 \equiv 0$ accounts for the removal of the poles.

A.3 The Differentiated System's Coefficients

Let us write

$$u(x) = \sum_{n=0}^{N} a_n^{(0)} T_n(x) \tag{A.30}$$

Expanding both the first and the second derivative of $u(x)$ in Chebychev polynomials, yields

$$\frac{d}{dx}u(x) = \sum_{n=0}^{N-1} a_n^{(-1)} T_n(x) \tag{A.31}$$

and

$$\frac{d^2}{dx^2}u(x) = \sum_{n=0}^{N-2} a_n^{(-2)} T_n(x) \tag{A.32}$$

The fact that $a_N^{(-1)} \equiv 0$ and $a_{N-1}^{(-2)} = a_N^{(-2)} \equiv 0$, reflects the loss of one and two degrees of freedoms respectively.

We aim to derive an expression which gives the coefficients $a_n^{(-1)}$ as a function of the coefficients $a_n^{(0)}$'s; the derivation follows.

Evaluating the indefinite integral of $T_n(x)$ and differentiating both sides of the resulting expression, yields

$$T_n(x) = \frac{d}{dx}\left[\frac{1}{2}\left(\frac{c_n}{n+1}T_{n+1}(x) - \frac{d_{n-2}}{n-1}T_{n-1}(x)\right)\right] \tag{A.33}$$

where

$$c_n = \begin{cases} 0 & \text{for } n < 0; \\ 2 & \text{for } n = 0; \\ 1 & \text{for } n > 0. \end{cases} \qquad d_n = \begin{cases} 0 & \text{for } n < 0; \\ 1 & \text{for } n \geq 0. \end{cases} \tag{A.34}$$

We now write

$$\sum_{n=0}^{N-1} a_n^{(-1)} T_n(x) = \frac{d}{dx}u(x) = \frac{d}{dx}\sum_{n=0}^{N} a_n^{(0)} T_n(x) \tag{A.35}$$

Using (A.33), we obtain

$$\frac{d}{dx}\left[\frac{1}{2}\sum_{n=0}^{N-1} a_n^{(-1)}\left(\frac{c_n}{n+1}T_{n+1}(x) - \frac{d_{n+2}}{n-1}T_{n-1}(x)\right)\right] \tag{A.36}$$

Equating the coefficients of $T_n(x)$ in (A.35) and (A.36),

$$2na_n^{(0)} = c_{n-1}a_{n-1}^{(-1)} - d_{n-1}a_{n+1}^{(-1)} \quad \text{for } n \in [1, N] \tag{A.37}$$

For the purposes of the differentiated system, expression (A.37) is not adequate, since two of the coefficients being sought are given as a function of only one of the

known coefficients. To achieve the desired inverse relationship, we only need to apply an appropriate summing procedure in (A.37), in order to cancel the unwanted coefficients.

The final formula reads

$$c_n a_n^{(-1)} = 2 \sum_{\substack{p=n+1 \\ p+n:\ odd}}^{N} p a_p^{(0)} \quad \text{for } n \in [0, N] \tag{A.38}$$

where the coefficient $a_N^{(-1)}$ is clearly zero.

However, our ultimate target is finding a relationship that gives us the coefficients $a_n^{(-2)}$'s as functions of the $a_n^{(0)}$'s.

Applying (A.38) with regard to the first and second derivatives's coefficients, we get

$$c_p a_p^{(-2)} = 2 \sum_{\substack{q=p+1 \\ q+p:\ odd}}^{N} q a_q^{(-1)} \quad \text{for } p \in [0, N] \tag{A.39}$$

and substituting (A.38) in (A.39) yields (since $c_p \geq 1$ always)

$$\begin{aligned} c_n a_n^{(-2)} &= 4 \sum_{\substack{p=n+1 \\ p+n:\ odd}}^{N} p \sum_{\substack{q=p+1 \\ q+p:\ odd}}^{N} q\, a_q^{(0)} \\ &= 4 \sum_{\substack{q=n+2 \\ q+n:\ even}}^{N} q\, a_q^{(0)} \sum_{\substack{p=n+1 \\ p+n:\ odd}}^{q-1} p \\ &= 4 \sum_{\substack{q=n+2 \\ q+n:\ even}}^{N} q\, a_q^{(0)} \frac{(n+1)+(q-1)}{2} \left(\frac{q-n}{2}\right) \\ &= \sum_{\substack{q=n+2 \\ q+n:\ even}}^{N} q\,(q^2-n^2)\, a_q^{(0)} \end{aligned} \tag{A.40}$$

Clearly enough, both $a_{N-1}^{(-2)}$ and $a_N^{(-2)}$ are zero. Formulas (A.38) and especially (A.40), are susceptible to large truncation errors which could affect their convergence severely.

A.4 The Integrated System's Coefficients

In an attempt to overcome the difficulties associated with the round-off problems of the previous approach, the integrated system method seeks to express the coefficients of integrals of $u(x)$, as functions of the coefficients of the expansion of the function itself.

Therefore if

$$u(x) = \sum_{n=0}^{N} a_n^{(0)} T_n(x) \tag{A.41}$$

then

$$\int u(x)\,dx = \sum_{n=0}^{N+1} a_n^{(+1)} T_n(x) \tag{A.42}$$

and

$$\int dx \int u(x)\,dx = \sum_{n=0}^{N+2} a_n^{(+2)} T_n(x) \tag{A.43}$$

where the presence of $a_{N+1}^{(+1)}$ and $a_{N+1}^{(+2)}$, $a_{N+2}^{(+2)}$ corresponds to the gain of one and two degrees of freedom, respectively.

It is straightforward that

$$2na_n^{(+1)} = c_{n-1}a_{n-1}^{(0)} - a_{n+1}^{(0)} \quad \text{for } n \in [1, N+1] \tag{A.44}$$

The fact that the lower limit of the definition interval of n is 1 instead of being 0, is a consequence of neglecting the integration constant in (A.42).

We now proceed to accomplish the task of expressing $a_n^{(+2)}$'s as functions of the $a_n^{(0)}$'s; applying (A.44) for $a_n^{(+2)}$ and $a_n^{(+1)}$ gives

$$2na_n^{(+2)} = c_{n-1}a_{n-1}^{(+1)} - a_{n+1}^{(+1)} \quad \text{for } n \in [2, N+2] \tag{A.45}$$

where $n \neq 0$, 1 due to neglecting both a constant and a linear term in (A.43). Shifting the indices in (A.44), we obtain

$$2(n-1)a_{n-1}^{(+1)} = c_{n-2}a_{n-2}^{(0)} - a_n^{(0)} \quad \text{for } n \in [2, N+2] \tag{A.46}$$

and

$$2(n+1)a_{n+1}^{(+1)} = c_n a_n^{(0)} - a_{n+2}^{(0)} \quad \text{for } n \in [0, N] \tag{A.47}$$

and sybstituting (A.46) and (A.47) in (A.45), yields

$$a_n^{(+2)} = \frac{1}{2n}\left[c_{n-1}\frac{1}{2(n-1)}\left(c_{n-2}a_{n-2}^{(0)} - a_n^{(0)}\right) - \frac{1}{2(n+1)}\left(c_n a_n^{(0)} - a_{n+2}^{(0)}\right)\right] \tag{A.48}$$

for $n \in [2, N]$ since $[2, N+2] \cap [0, N] = [2, N]$.

Going through the algebra and eliminating c_{n-1}, c_n (always equal to one), we finally arrive at the desired expression, which reads

$$a_n^{(+2)} = \frac{c_{n-2}}{4(n-1)n}a_{n-2}^{(0)} - \frac{1}{2(n^2-1)}a_n^{(0)} + \frac{1}{4n(n+1)}a_{n+2}^{(0)} \quad \text{for } n = 2, \ldots, N \tag{A.49}$$

A.5 The Chebychev Transform of $\sin(\pi x)$.

The analytic derivation of the Chebychev transform of $u(x) = \sin \pi x$ proceeds as follows:

The magnitude of the n-th spectral component is given as

$$a_n = \frac{2}{\pi c_n}\int_{-1}^{+1} \frac{\sin(\pi x)\cos(n\cos^{-1}x)}{(1-x^2)^{-1/2}}\,dx \tag{A.50}$$

or under the familiar transformation $x = \cos\theta$

$$a_n = \frac{2}{\pi c_n}\int_0^{\pi} \sin(\pi\cos\theta)\cos(n\theta)\,d\theta \tag{A.51}$$

At this point, an appropriate manipulation of the integrand allows us to proceed with the integration directly. This involves an expansion of $\sin(\pi \cos\theta)$ into a particular infinite cosine series according to the formula

$$\sin(z\cos\theta) = 2\sum_{k=0}^{\infty}(-1)^k J_{2k+1}(z)\cos(2k+1)\theta \qquad (A.52)$$

(Abramowitz and Stegun, 1964), where $J_l(z)$ is a Bessel function of the first kind.

Substitution of $\sin(\pi\cos\theta)$ according to (A.52) results into

$$\frac{4}{\pi c_n}\sum_{k=0}^{\infty}(-1)^k J_{2k+1}(\pi)\int_0^{\pi}\cos(2k+1)\theta\cos n\theta\, d\theta \qquad (A.53)$$

and applying the well-known formula

$$\int_0^{\pi}\cos mx\cos lx\, dx = \frac{\pi}{2}c_m\delta_{ml} \qquad (A.54)$$

where m, l are integers, we obtain the desired expression, which reads

$$a_n = 2\sum_{k=0}^{\infty}(-1)^k J_{2k+1}(\pi)\ \delta_{n,2k+1} \qquad (A.55)$$

APPENDIX B

B.1 The Fast Chebychev Transform (F.C.T) Algorithm

The *direct Chebychev transform* of a function $f(x)$ amounts to the calculation of the coefficients of the decomposition

$$f(x) = \sum_{n=0}^{N} a_n^{(0)} T_n(x) \qquad (B.1)$$

via the integrals

$$a_n^{(0)} = \frac{2}{\pi c_n} \int_{-1}^{+1} \frac{T_n(x)}{\sqrt{1-x^2}} f(x)\, dx \qquad (B.2)$$

The *inverse Chebychev transform* performs the summation, after knowledge of the $a_n^{(0)}$'s has been acquired.

Two factors contribute to devising a fast algorithm for evaluating both the direct and the inverse Chebychev transforms.

First comes the realization that by substituting $x = \cos\theta$, the Chebychev transform reduces to a *cosine transform*, that is to say,

$$a_n^{(0)} = \frac{2}{\pi c_n} \int_0^{\pi} \cos n\theta\, f(\theta)\, d\theta \qquad (B.3)$$

and

$$f(\theta) = \sum_{n=0}^{N} a_n^{(0)} \cos n\theta \qquad (B.4)$$

where $f(\theta)$ needs to known (or it will be evaluated, respectively) at the nodes

$$x_i = \cos\theta_i, \qquad 0 \leq \theta_i \leq \pi \qquad (B.5)$$

The second factor is the possibility of constructing a *fast cosine transform* through the use of the FFT algorithm.

Before proceeding further, we would like to clarify a point, which appears to be a common source of confusion. This regards the *normalization* of the transform, where two different types appear to be applied. If consistency when applying both the direct and the inverse transforms is maintained, then both methods lead to identical formulations.

According to the first version, the inverse transform is defined as

$$f(\cos\frac{\pi i}{N}) = \sum_{n=0}^{N} \left(\frac{1}{\bar{c}_n}\right) a_n^{(0)} \cos\left(\frac{n\pi i}{N}\right) \qquad \text{for } 0 \leq i \leq N \qquad (B.6)$$

with the coefficients given as

$$a_n^{(0)} = \frac{2}{N} \sum_{i=0}^{N} \left(\frac{1}{\bar{c}_i}\right) f(\cos\frac{\pi i}{N}) \cos\left(\frac{n\pi i}{N}\right) \qquad \text{for } 0 \leq n \leq N \qquad (B.7)$$

where the vector $\bar{c}_j$ is given as

$$\bar{c_j} = \begin{cases} 2 & \text{if } j = 0; \\ 1 & \text{if } 1 \leq j \leq N-1; \\ 2 & \text{if } j = N. \end{cases} \qquad (B.8)$$

(Fox and Parker, 1968).

In the second version, the expansion reads defined as

$$f(\cos\frac{\pi i}{N}) = \sum_{n=0}^{N} a_n^{(0)} \cos\left(\frac{n\pi i}{N}\right) \qquad \text{for } 0 \leq i \leq N \qquad (B.9)$$

and the forward transform is defined as

$$a_n^{(0)} = \frac{2}{N}\left(\frac{1}{\bar{c}_n}\right) \sum_{i=0}^{N}\left(\frac{1}{\bar{c}_i}\right) f(\cos\frac{\pi i}{N}) \cos\left(\frac{n\pi i}{N}\right) \qquad \text{for } 0 \leq n \leq N \qquad (B.10)$$

(Haidvogel and Zang, 1979). In this thesis, the Chebychev transform pair follows the definition expressed by (B.9-10).

The development of the fast cosine transform is briefly outlined below (for further details and justification of the process, see Cooley et al, (1970), Gentleman, (1972) and Deville and Labrosse, (1982)).

Let us assume, that the real, $(N+1)$-long vector $\mathbf{z}$ for $j = 0, \ldots, N$ is given as input to be cosine transformed.

We start by defining the $2N$-long vector $\mathbf{y}$ as

$$y_j = z_j \quad \text{for } j = 0, \ldots, N \quad \text{and} \quad y_{2N-j} = y_j \quad \text{for } j = 1, \ldots, N-1 \qquad (B.11)$$

This vector is real and symmetric.

Now, we define a N-long vector $\mathbf{x}$ as

$$x_j = y_{2j} + i[y_{2j+1} - y_{2j-1}] \quad \text{for } j = 0, \ldots, N-1 \qquad (B.12)$$

The elements of this array are complex and conjugate symmetric.

In the next step, we construct another complex array $\mathbf{b}$, which is $(N/2)$-long, as follows

$$b_j = \left(x_j + x_{(N/2)+j}\right) + ie^{\pm 2\pi i n/N}\left(x_j - x_{(N/2)+j}\right) \quad \text{for } j = 0, \ldots, N/2-1 \quad (B.13)$$

The plus sign $(+)$ corresponds to direct transform, whereas the minus sign $(-)$ to the inverse transform. The Fourier transform of $\mathbf{b}$

$$B_j = \sum_{n=0}^{N/2-1} b_n e^{\pm 2\pi i/(N/2)} \quad \text{for } j = 0, \ldots, N/2-1 \qquad (B.14)$$

is then taken using a complex FFT subroutine on $N/2$ points; the array $\mathbf{B}$ is, in general, complex.

The recovery of the Fourier transform $\mathbf{X}$ of $\mathbf{x}$ is done according to

$$X_{2j} = \mathcal{R}(B_j) \quad \text{and} \quad X_{2j+1} = I(B_j) \quad \text{for } j = 0, \ldots, N/2-1 \qquad (B.15)$$

and, subsequently, the Fourier transform $\mathbf{Y}$ of $\mathbf{y}$ are obtained via the expression

$$Y_j = \frac{1}{2}\left(X_j + X_{N-j}\right) \pm \frac{1}{2\sin\left(\pi j/N\right)}\left(X_j - X_{N-j}\right) \quad \text{for } j = 1,\ldots,N-1 \quad (B.16)$$

The elements of the N-long array $\mathbf{Y}$ are real and symmetric ($Y_{2N-j} = Y_j \quad \text{for } j = 1,\ldots,N-1$).

Finally, we obtained the desired $(N+1)$-long real coefficient vector $\mathbf{Z}$ as

$$Z_j = 2Y_j \quad \text{for } j = 1,\ldots,N-1 \quad (B.17)$$

and

$$Z_0 = X_0 + \sum_{n=0}^{N-1} z_{2n+1} \quad , \quad Z_N = X_0 - \sum_{n=0}^{N-1} z_{2n+1} \quad (B.18)$$

Appropriate normalizations depending on the version and the direction of the transform complete the procedure.

When more than one dimension needs to be transformed, a straightforward application of the above algorithm in each coordinate may not be considered efficient enough, as it might lead to excessive memory requirements and computational burden. For multi-dimensional cases, then, an extensive use of symmetry considerations combined with the proper pre- and post-processing can provide both computational acceleration and memory economization gains (Deville and Labrosse, 1982).

B.2 Fast Inversion of Quasi-Tridiagonal Systems

Employing the Chebychev integrated tau method for the solution of either ordinary differential equations, or (in combination with a Crank-Nicolson time differencing scheme) partial differential equations, which involve second order differential operators, results in linear systems of quasi-tridiagonal structure.

Let us say, that the system to be solved is

$$\mathbf{F}_{nm}\mathbf{a}_{\mathbf{m}}^{(0)} = \mathbf{b}_n \qquad \text{for } n, m = 0, \ldots, N \tag{B.19}$$

The elements of the matrix $\mathbf{F}_{nm}$ differ slightly from equation to equation, but the structure remains the same and the system to be solved is

$$\begin{pmatrix}
f_{00} & & f_{02} & & f_{04} & & & & \\
& f_{11} & & f_{13} & & f_{15} & & & \\
& & \ddots & & \ddots & & \ddots & & \\
& & & \ddots & & \ddots & & \ddots & \\
& & & & f_{n,n} & & f_{n,n+2} & & f_{n,n+4} \\
& & & & & \ddots & & \ddots & \\
& & & & & & \ddots & & \ddots \\
+1 & -1 & & & & \ldots & & -1 & +1 \\
+1 & +1 & & & \ldots & & & & +1
\end{pmatrix}
\begin{pmatrix} a_0^{(0)} \\ \\ \\ \vdots \\ \\ \\ a_N^{(0)} \end{pmatrix}
=
\begin{pmatrix} b_0 \\ \\ \\ \vdots \\ \\ \\ b_N \end{pmatrix}$$

(B.20)

A closer look at the matrix of the expansion coefficients

$$
\begin{array}{c}
 \\
n=0 \\
 \\
 \\
 \\
\vdots \\
 \\
 \\
n=N-1 \\
n=N
\end{array}
\begin{array}{c}
\begin{array}{cccccccc} n=0 & & \rightarrow & & \cdots & & & n=N \end{array} \\
\left(\begin{array}{cccccccc}
f_{00} & & f_{02} & & f_{04} & & & \\
 & f_{11} & & f_{13} & & f_{15} & & \\
 & & \ddots & & \ddots & & \ddots & \\
 & & & \ddots & & \ddots & & \ddots \\
 & & & & f_{n,n} & & f_{n,n+2} & & f_{n,n+4} \\
 & & & & & \ddots & & \ddots & \\
 & & & & & & \ddots & & \ddots \\
+1 & -1 & & & & \cdots & & -1 & +1 \\
+1 & +1 & & & \cdots & & & & +1
\end{array}\right)
\end{array}
\qquad (B.21)
$$

shows that, it is the the last two rows (which correspond to homogeneous Dirichlet boundary conditions at $x = -1$ and $x = +1$, respectively) that cause a departure from a tridiagonal structure.

For the Helmholtz equation, the non-zero elements are defined as

$$
f_{n,n} = \frac{k^2 c_{n-2}}{4n(n-1)}, \quad f_{n,n+2} = 1 - \frac{k^2 e_{n+2}}{2(n^2-1)} \quad \text{and} \quad f_{n,n+4} = \frac{k^2 e_{n+4}}{4n(n+1)} \qquad (B.22)
$$

for $2 \leq n \leq N$ and

$$
c_n = \begin{cases} 2, & \text{if } n = 0; \\ 1, & \text{if } 1 \leq n \leq N \end{cases} \quad \text{and} \quad \begin{cases} 1, & \text{if } 1 \leq n \leq N; \\ 0, & \text{if } n > N \end{cases} \qquad (B.23)
$$

The elements of the vector **b** are zero for this case.

For the heat equation (with conductivity unity), these elements are

$$f_{n,n} = \frac{c_{n-2}}{4n(n-1)}, \quad f_{n,n+2} = \left(\frac{-\Delta t}{2}\right) + \frac{e_{n+2}}{2(n^2-1)} \quad \text{and } f_{n,n+4} = \frac{e_{n+4}}{4n(n+1)} \tag{B.24}$$

whereas for the Schrödinger equation (in atomic units), the presence of the imaginary unity i results in

$$f_{n,n} = \frac{c_{n-2}}{4n(n-1)}, \quad f_{n,n+2} = \left(\frac{-i\Delta t}{2}\right) + \frac{e_{n+2}}{2(n^2-1)} \quad \text{and } f_{n,n+4} = \frac{e_{n+4}}{4n(n+1)} \tag{B.25}$$

The elements b_n above are calculated from the multiplication of the right hand side Crank-Nicolson matrix with the solution vector from the previous time step and they are, in general, non-zero.

We observe that the even and the odd $a_n^{(0)}$'s could be decoupled, if it was not for the last two rows. To decouple them, we transform the boundary conditions

$$\sum_{n=0}^{N} a_n^{(0)} = 0 \quad \text{and} \quad \sum_{n=0}^{N} (-1)^n a_n^{(0)} = 0 \tag{B.26}$$

to

$$\sum_{n=0}^{N/2} a_{2n}^{(0)} = 0 \quad \text{and} \quad \sum_{n=1}^{N/2} a_{2n-1}^{(0)} = 0 \tag{B.27}$$

Now the $(N-1)$th row elements alternate between 0 and 1 (odd coefficients), whereas the elements of the (N)th row alternate between 1 and 0 (even coefficients).

Now the matrix looks like

$$
\begin{array}{c}
\\
n=0 \\
\\
\downarrow \\
\\
\vdots \\
\\
\\
n=N-1 \\
n=N
\end{array}
\begin{array}{c}
\begin{array}{ccccccccc}
n=0 & & \rightarrow & & \cdots & & & & n=N
\end{array} \\
\left(\begin{array}{ccccccccc}
f_{00} & & f_{02} & & f_{04} & & & & \\
& f_{11} & & f_{13} & & f_{15} & & & \\
& & \ddots & & \ddots & & \ddots & & \\
& & & \ddots & & \ddots & & \ddots & \\
& & & & f_{n,n} & & f_{n,n+2} & & f_{n,n+4} \\
& & & & & \ddots & & \ddots & \\
& & & & & & \ddots & & \ddots \\
& +1 & & +1 & & \cdots & & +1 & \\
+1 & & +1 & & \cdots & & & & +1
\end{array}\right)
\end{array}
\tag{B.28}
$$

and the transformed system

$$
\left(\begin{array}{ccccccccc}
f_{00} & & f_{02} & & f_{04} & & & & \\
& f_{11} & & f_{13} & & f_{15} & & & \\
& & \ddots & & \ddots & & \ddots & & \\
& & & \ddots & & \ddots & & \ddots & \\
& & & & f_{n,n} & & f_{n,n+2} & & f_{n,n+4} \\
& & & & & \ddots & & \ddots & \\
& & & & & & \ddots & & \ddots \\
& +1 & & +1 & & \cdots & & +1 & \\
+1 & & +1 & & \cdots & & & & +1
\end{array}\right)
\left(\begin{array}{c} a_0^{(0)} \\ \\ \\ \vdots \\ \\ \\ a_N^{(0)} \end{array}\right)
=
\left(\begin{array}{c} b_0 \\ \\ \\ \vdots \\ \\ \\ b_N \end{array}\right)
\tag{B.29}
$$

is equivalent to the pair

$$\begin{pmatrix} g_{00} & g_{01} & g_{02} & & \\ & g_{11} & g_{12} & g_{13} & \\ & & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots \\ +1 & & \cdots & & +1 \end{pmatrix} \begin{pmatrix} a_0^{(0)} \\ a_2^{(0)} \\ \vdots \\ \\ a_N^{(0)} \end{pmatrix} = \begin{pmatrix} b_0 \\ b_2 \\ \vdots \\ \\ b_N \end{pmatrix} \qquad (B.30)$$

for the even part of $\mathbf{a}^{(0)}$ and

$$\begin{pmatrix} h_{11} & h_{12} & h_{13} & & \\ & h_{22} & h_{23} & h_{24} & \\ & & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots \\ +1 & & \cdots & & +1 \end{pmatrix} \begin{pmatrix} a_1^{(0)} \\ a_3^{(0)} \\ \vdots \\ \\ a_{N-1}^{(0)} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_3 \\ \vdots \\ \\ b_{N-1} \end{pmatrix} \qquad (B.31)$$

for the odd part of it.

The elements of the matrices **G** and **H** are given directly from the elements of the decomposed matrix **F** as $g_{n,m} = f_{2n,2m}$ for $n = 0, \ldots, N/2 - 1$, $m = 0, \ldots, N/2$ and $g_{N/2,m} = f_{N,2m}$ for $m = 0, \ldots, N/2$. The elements of the matrix **H** are given, in a similar fashion, as $h_{n,m} = f_{2n-1,2m-1}$ for $n = 1, \ldots, N/2 - 1$, $m = 1, \ldots, N/2$ and $h_{N/2,m} = f_{N-1,2m-1}$ for $m = 1, \ldots, N/2$.

Let us proceed with the solution process for the first of these sub-systems (since they are identical).

A special forward elimination procedure is applied. Thus, the bottom row of **G** is eliminated, in such a way that the sparseness the matrix is fully exploited. At this point the question of whether *pivoting* should be applied (when appropriate) arises.

This is an essential point, since if the elimination is performed without pivoting, there will be no row interchanges and subsequently the resulting matrix $\bar{\mathbf{G}}$ will be upper triangular with bandwidth 2 (or *upper tridiagonal*). Therefore, the resulting system is

$$\bar{\mathbf{G}}\mathbf{a}_{\mathbf{2n}}^{(0)} = \mathbf{L}^{-1}\mathbf{b}_{\mathbf{2n}} = \bar{\mathbf{b}}_{\mathbf{2n}} \tag{B.32}$$

(where information on the multipliers are kept in $\mathbf{L}^{-1}$) or

$$\begin{pmatrix} \bar{g}_{00} & \bar{g}_{01} & \bar{g}_{02} & & & \\ & \bar{g}_{11} & \bar{g}_{12} & \bar{g}_{13} & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \\ & & & & & \bar{g}_{N/2,N/2} \end{pmatrix} \begin{pmatrix} a_0^{(0)} \\ a_2^{(0)} \\ \vdots \\ \\ a_N^{(0)} \end{pmatrix} = \begin{pmatrix} \bar{b}_0 \\ \bar{b}_2 \\ \vdots \\ \\ \bar{b}_N \end{pmatrix} \tag{B.33}$$

Such a system may then be solved very efficiently via back substitution procedures specialized on banded matrices (Golub and van Loan, 1983). It is interesting to observe, however, that pivoting is advisable from the very first elimination step, because the presence of 1 in the last row makes the multiplier of the first row larger than unity. Application of pivoting tends to expand the upper bandwidth of $\bar{\mathbf{G}}$ resulting in

$$\bar{\mathbf{G}}\mathbf{a}_{\mathbf{2n}}^{(0)} = \mathbf{M}^{-1}\mathbf{b}_{\mathbf{2n}} = \bar{\mathbf{b}}_{\mathbf{2n}} \tag{B.34}$$

(where information regarding both the multipliers and the permutations are kept in

$\mathbf{M^{-1}}$) or

$$\begin{pmatrix} \bar{g}_{00} & & \cdots & & \bar{g}_{0,N/2} \\ & \bar{g}_{11} & & \cdots & \\ & & \ddots & & \cdots \\ & & & \ddots & \cdots \\ & & & & \bar{g}_{N/2,N/2} \end{pmatrix} \begin{pmatrix} a_0^{(0)} \\ a_2^{(0)} \\ \vdots \\ \\ a_N^{(0)} \end{pmatrix} = \begin{pmatrix} \bar{b}_0 \\ \bar{b}_2 \\ \vdots \\ \\ \bar{b}_N \end{pmatrix} \qquad (B.34)$$

The matrix $\bar{\mathbf{G}}$ is upper triangular with full bandwidth and a straightforward back substitution is necessary.

If the former approach fails to work (due to a zero pivot during the elimination), the latter pivoting procedure will have to be implemented, degrading inevitably the efficiency of the algorithm. The prospect of confronting an upper triangular matrix with full bandwidth is not particularly inviting; it is, nevertheless, inevitable due to the transfer of the full row of one's at the top of the matrix during the initial pivoting. If implementation of the no-pivoting procedure is not possible, we can only consider an optimization of the alternative algorithm. Minimization of "fill-in" may not be possible but an improvement of the forward elimination should be explored. Pivoting results in a system of an awkward form and subsequently, the amount of searches can not be reduced; eliminations of trivial entries can not be bypassed either. Continuation of the elimination-pivoting process does not permit any shortcuts, leading to a general Gaussian elimination algorithm.

The following reformulation of the problem allows the derivation of an equivalent system. A simple transformation induces diagonal dominance in the system and it also gives rise to elegant pivoting algoritms. We start with a simple transformation of the original system, so that the rows of alternating 0's and 1's occupy its first two

rows. This introduces a slight modification in the structure of the matrix **F**

$$
\begin{array}{c}
\\
n=0 \\
n=1 \\
\\
\downarrow \\
\\
\vdots \\
\\
\\
n=N
\end{array}
\begin{array}{c}
\begin{array}{ccccccccc}
n=0 & & \rightarrow & & \cdots & & & & n=N
\end{array} \\
\left(\begin{array}{ccccccccc}
 & & +1 & & \cdots & & +1 & & \\
+1 & & & & \cdots & & & & +1 \\
f_{20} & & f_{22} & & f_{24} & & & & \\
 & f_{31} & & f_{33} & & f_{35} & & & \\
 & & \ddots & & \ddots & & \ddots & & \\
 & & & \ddots & & \ddots & & \ddots & \\
 & & & & f_{n,n-2} & & f_{n,n} & & f_{n,n+2} \\
 & & & & & \ddots & & \ddots & \\
 & & & & & & \ddots & & \ddots
\end{array}\right)
\end{array}
\tag{B.36}
$$

where the elements f_{nm} that lie in the diagonals are obtained by a mere shift of the row index n by a factor of 2.

Evidently, the components **G**

$$
\begin{array}{c}
\\
n=0 \\
\\
\downarrow \\
\vdots \\
n=N/2
\end{array}
\begin{array}{c}
\begin{array}{ccccc}
n=0 & & \rightarrow & \cdots & n=N/2
\end{array} \\
\left(\begin{array}{ccccc}
+1 & & \cdots & & +1 \\
g_{10} & g_{11} & g_{12} & & \\
 & g_{21} & g_{22} & g_{23} & \\
 & & \ddots & \ddots & \ddots \\
 & & & \ddots & \ddots
\end{array}\right)
\end{array}
\tag{B.37}
$$

(with $g_{n,m} = f_{2n,2m}$ for $n = 1,\ldots,N/2$, $m = 0,\ldots,N/2$ and $g_{0,m} = f_{0,2m}$ for $m = 0,\ldots,N/2$) and $\mathbf{H}$

$$
\begin{array}{c}
 \\
n=1 \\
 \\
\downarrow \\
\vdots \\
n=N/2
\end{array}
\begin{array}{c}
\begin{array}{cccccc} n=1 & & \rightarrow & \cdots & n=N/2 \end{array} \\
\left(\begin{array}{cccccc}
+1 & & \ldots & & +1 \\
h_{21} & h_{22} & h_{23} & & \\
 & h_{32} & h_{33} & h_{34} & \\
 & & \ddots & \ddots & \ddots \\
 & & & \ddots & \ddots
\end{array}\right)
\end{array}
\qquad (B.38)
$$

(with $h_{n,m} = f_{2n-1,2m-1}$ for $n = 2,\ldots,N/2$, $m = 0,\ldots,N/2$ and $h_{1,m} = f_{1,2m-1}$ for $m = 1,\ldots,N/2$) of the even—odd matrix decomposition of $\mathbf{F}$ exhibit corresponding alterations; the row of 1's has been transferred to the first row and the diagonals have undergone a lateral shift, introducing a new main diagonal. Two fundamental points are to be identified: the introduction of a more vigorous main diagonal (recall equations (B.22-25) and the *upper Hessenberg* structure of the decomposed systems. The reinforcement of the main diagonal tends to allow an accurate and efficient forward elimination without pivoting. Furthermore, the reported structure of the matrix may be exploited to provide a substantial decrease in the number of searche-comparisons associated with pivoting, if the latter is necesary.

Concluding the presentation of the special LU decompositions of the tau integrated systems analysed previously, we summarize them briefly as follows:

1) **SLU1**: The boundary row is at the bottom of the matrix originally. Its elimination proceeds without pivoting and, subsequently, backsubstitution is applied on an upper tridiagonal matrix.

2) **SLU2** The boundary row is originally at the top of the matrix. Elimination of lower off-diagonal is performed without pivoting. A full upper triangular matrix is obtained and a full bandwidth backsubstitution is applied.

3) **SLU3** Identical to the **SLU2** procedure, but for the pivoting incorporated in the **SLU3** algorithm. The pivoting process follows an "upper Hessenberg" pattern.

An analysis of the applicability and the performance of fast solvers for the heat, the Schrödinger and the migration problems is given in (4.4.7), (5.7) and (6.11) respectively.

B.3 Inversion of Tridiagonal Systems

Tridiagonal systems may be inverted very efficiently; the following algorithm is taken from Vichnevetsky (1981). Let us assume that we have to solve the system $\mathbf{Au} = \mathbf{d}$, i.e

$$\begin{pmatrix} a_1 & c_1 & & & \\ b_2 & a_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & a_{N-1} & c_{N-1} \\ & & & b_N & a_N \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_N \end{pmatrix} = \begin{pmatrix} d_1 \\ \vdots \\ d_N \end{pmatrix} \qquad (B.39)$$

The algorithm is based on a special LU decomposition and includes the following steps:

An *LU decomposition* of the problem is performed:

$\alpha_1 = a_1$

$\beta_1 = c_1/\alpha_1$

For $i = 2, \ldots, N$

$\alpha_i = a_i - b_i\beta_{i-1}$

$\beta_i = c_i/\alpha_i$

The system $\mathbf{Ly} = \mathbf{d}$, is then solved by *forward substitution*:

$y_1 = d_1/\alpha_1$

For $i = 2, \ldots, N$

$y_i = (d_i - b_i y_{i-1})/\alpha_i$

The procedure is completed by solving $\mathbf{Uu} = \mathbf{y}$ with *backsubstitution*:

$u_N = y_N$

For $i = N-1, \ldots, 1$

$u_i = y_i - \beta_i u_{i+1}$

Besides its dramatic increase in speed, the method succeeds in bypassing the error-growth due to the backsubstitution and in reducing the storage requirements significantly (Mitchell and Griffiths, 1980). Pivoting has not been introduced in this algorithm. Consequently, this routine fails if a zero pivot is encountered; such a case is possible even for a non-singular matrix. Despite this theoretical flaw, tridiagonal solvers very rarely fail in practice. Furthermore, if the tridiagonal matrix is diagonally dominant, i.e $|a_i| > |b_i| + |c_i|$ for $i = 1, \ldots, N$, it may be shown that division by zero never occurs (Press et al, 1985). Of course, if a zero pivot is encountered, we have to introduce pivoting in the algorithm.