NASA-CR-172149

NASA Contractor Report 172149

ICASE

FINITE-DIFFERENCE, SPECTRAL AND GALERKIN METHODS FOR TIME-DEPENDENT PROBLEMS

Eitan Tadmor

Contract No. NAS1-17070
June 1983

INSTITUTE FOR COMPUTER APPLICATIONS IN SCIENCE AND ENGINEERING
NASA Langley Research Center, Hampton, Virginia 23665

Operated by the Universities Space Research Association

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23665

FINITE-DIFFERENCE, SPECTRAL AND GALERKIN METHODS FOR TIME-DEPENDENT PROBLEMS

Eitan Tadmor
Institute for Computer Applications in Science and Engineering

ABSTRACT

We survey finite-difference, spectral and Galerkin methods for the approximate solution of time-dependent problems. A unified discussion on their accuracy, stability and convergence is given. In particular, the dilemma of high accuracy versus stability is studied in some detail.

1980 Mathematics Subject Classification. Primary 65M10, Secondary 35A12.

Research was supported by the National Aeronautics and Space Administration under NASA Contract No. NAS1-17070 while the author was in residence at ICASE, NASA Langley Research Center, Hampton, VA 23665.

0. Introduction

During the last decade, discrete methods -- other than the classical finite-differences -- have gained an increasing popularity while used for the approximate solution of time-dependent problems. Most noticable, are the (pseudo)spectral and Galerkin methods, e.g. [4-9], [13-14], [19], [21] and the references therein.

The purpose of these notes is to give a unified survey on these three classes of discrete methods -- finite differences, spectral and Galerkin, discussing some of the theoretical aspects with regard to their accuracy, stability and efficient implementation. As a model problem for our discussion, we consider the one-dimensional symmetric hyperbolic system

$$(0.1)\quad \partial_t u(x,t) = A(x,t)\partial_x u(x,t) + B(x,t)u(x,t), \qquad A(x,t) = A'(x,t).$$

Here and elswhere in the paper, w' denotes the transpose of a given vector, w^* its conjugate transpose, and $\|w\|=(w^*w)^{1/2}$ its Euclidean norm; similar notations are used for matrices. We will also briefly mention systems of parabolic type, which are more favored by the kind of arguments discussed below, due to the presence of dissipation. To avoid further complications that arise with time-discretization and handling boundary conditions, we restrict our attention to the periodic method of lines. This allows us to make a rather detailed study of various types of approximations to (0.1), where the spatial differentation is replaced by its discrete counterpart.

We begin, in Part I, discussing finite difference methods. Our approach -- slightly different than usual -- emphasizes the matrix representation of such methods. The reason is a two-fold one: first, the standard approach via Von Neumann analysis is by now classical and can be found in a variety of references, e.g. [15]; second, viewing these discretizations in the language of matrices allow us to move quite naturally to our discussion on spectral and Galerkin methods in the second and third parts. Indeed, the generality of the abstract discrete differentation operator dealt with in Part I will prevail for finite-difference as well as spectral and Galerkin methods. One of the main objectives of this review is, in fact, to show the intimate relation between the three: the spectral-Fourier method can be viewed as a special centered finite differencing based on an ever increasing number of gridpoints peridoically extended, and both result from an appropriate choice of basis functions used in the Galerkin method.

We focus our attention on the all important question of stability. It is shown that antisymmetry periodicity as well as a locality restriction are essential properties that a discrete differencing method should share with the differential problem, for the resulting discrete system to be stable. The locality restriction can be equivalently expressed by the boundedness of amplification blocks associated with the highest modes. The accuracy requirement, on the other hand, is determined by the exactness of differencing the lower modes. The combination of the two guarantee convergence, as the lower modes carrying most of the information are accurately represented, while the highest modes are not, yet stability assures us that they are not amplified and hence rapidly tend to zero, just as is the case with the differential problem.

Both properties of accuracy and stability are well accommodated in discrete methods having finite degree of accuracy; they contradict each other, however, with highly accurate methods. We are then led to a discussion on the skew-symmetric differencing and smoothing procedures. Both aim at dissolving this contradiction by bounding the highest modes' amplification blocks, yet leaving the lower, accurate modes, unharmed.

In Part II we amplify these points with regard to the spectral-Fourier method, from still a slightly different point of view. We base the whole stability analysis in this case on the aliasing formula, relating the Fourier coefficients of a given periodic function to those of its equidistant interpolant. This is the single most important formula, which dominates the question of accuracy versus stability in this case. It naturally arises with the Fourier method as the aliasing dilemma, and its usual remedy is again by either skew-symmetric differencing or via smoothing. From this point of view, finite difference methods having finite degree of accuracy can be viewed as special cases of the Fourier method with a built-in smoothing which guarantees their stability.

In Part III we discuss Galerkin-type methods. Again there is an emphasis on the close connection with finite-difference and spectral-Fourier methods. Stability follows in this case, due to lack of aliasing. Once the exact Fourier coefficients are discretized, we find ourselves dealing with exactly the same kind of arguments introduced before.

Finally, in order to make these notes self-contained, we have collected in the Appendix some basic properties of Toeplitz and circulant matrices; these play a vital role in the foregoing analysis.

Part I. Finite Difference Methods

1. Finite Difference Operators

Let $v(x)$ be a 2π-periodic m-dimensional vector function, whose values $v_\nu \equiv v(x_\nu)$ are assumed known at the gridpoints $x_\nu = \nu h$, $h = \frac{2\pi}{N}$, $\nu = 0,1,\cdots,N-1$. A second order accurate approximation to its derivative, $D_x v(x)$, is given by the centered divided difference

$$(1.1_2) \qquad D_2(h)[v(x)] = \frac{v(x+h) - v(x-h)}{2h} .$$

When augmented by the periodicity of v, these divided differences are well defined at all gridpoints $x = x_\nu$, $\nu = 0,1,\cdots,N-1$. The transformation which takes the vector of the assumed known gridvalues $\underset{\sim}{v} \equiv (v_0,\cdots,v_{N-1})^{\prime}$ into the vector of divided differences $\partial_{FD_2}[\underset{\sim}{v}] \equiv (D_2(h)[v_0],\cdots,D_2(h)[v_{N-1}])^{\prime}$ is linear, and hence has a matrix representation

$$(1.2_2) \qquad \partial_{FD_2}[\underset{\sim}{v}] = \underset{\sim}{D}_2\underset{\sim}{v};$$

here the matrix $\underset{\sim}{D}_2 \equiv \underset{\sim}{D}_2(h)$ is given by

$$(1.3_2) \qquad \underset{\sim}{D}_2 = \frac{1}{2h}\cdot\begin{bmatrix} 0 & I & 0 & \cdots & 0 & -I \\ -I & 0 & I & \ddots & & 0 \\ 0 & -I & 0 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ 0 & & & \ddots & 0 & I \\ I & \cdots & \cdots & 0 & -I & 0 \end{bmatrix},$$

its entries are being m-dimensional blocks. Similarly, fourth and sixth order accurate centered divided differences are given by

$$D_4(h)[v(x)] = \frac{4D_2(h) - D_2(2h)}{3}[v(x)]$$

(1.1$_4$)

$$= \frac{8[v(x+h)-v(x-h)] - [v(x+2h)-v(x-2h)]}{12h}$$

$$D_6(h)[v(x)] = \frac{15D_2(h) - 6D_2(2h) + D_2(3h)}{10}[v(x)]$$

(1.1$_6$)

$$= \frac{45[v(x+h)-v(x-h)] - 9[v(x+2h)-v(x-2h)] + [v(x+3h)-v(x-3h)]}{60h}$$

with their corresponding matrix representations

$$(1.3_4) \quad \underset{\sim}{D}_4 = \frac{1}{12h} \cdot \begin{bmatrix} 0 & 8I & -I & 0 & \cdot\;\cdot\;\cdot & 0 & I & -8I \\ -8I & 0 & 8I & \ddots & & & 0 & I \\ I & -8I & 0 & \ddots & \ddots & & & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & & \cdot \\ \cdot & & \ddots & \ddots & \ddots & \ddots & \ddots & \cdot \\ \cdot & & & \ddots & \ddots & \ddots & \ddots & \cdot \\ \cdot & & & & \ddots & \ddots & \ddots & 0 \\ 0 & & & & & \ddots & \ddots & -I \\ -I & 0 & & & & \ddots & 0 & 8I \\ 8I & -I & 0 & \cdot\;\cdot\;\cdot & 0 & I & -8I & 0 \end{bmatrix}$$

$$(1.3_6) \qquad \underset{\sim}{D}_6 = \frac{1}{60h} \cdot \begin{bmatrix} 0 & 45I & -9I & I & 0 & \cdots & 0 & -I & 9I & -45I \\ -45I & 0 & 45I & -9I & \ddots & & & 0 & I & 9I \\ 9I & -45I & 0 & 45I & \ddots & & & & 0 & -I \\ -I & 9I & -45I & 0 & \ddots & & & & & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & & & & & \vdots \\ \vdots & & & & & & & & & 0 \\ 0 & & & & & & & \ddots & & I \\ I & 0 & & & & & & \ddots & \ddots & -9I \\ -9I & I & 0 & & & & \ddots & \ddots & 0 & 45I \\ 45I & -9I & I & 0 & \cdots & 0 & -I & 9I & -45I & 0 \end{bmatrix}$$

Observe that the matrices $\underset{\sim}{D}_{2s}$, $s = 1,2,3$ are antisymmetric block circulant ones; by the latter we mean that their (j,k) block entry depends only on $(j-k)[\text{mod}N]$.

The above examples illustrate special cases of a general 2s-order accurate centered divided difference given by [9, Section 3]

$$(1.1_{2s}) \qquad D_{2s}(h) \equiv 2\sum_{k=1}^{s} \beta_k D_2(kh), \qquad \beta_k \equiv \beta_k(s) = \frac{(-1)^{k+1}(s!)^2}{(s+k)!(s-k)!};$$

likewise, it has an antisymmetric block circulant matrix representation $\underset{\sim}{D}_{2s} \equiv \underset{\sim}{D}_{2s}(h)$

$$(1.2_{2s}) \qquad \partial_{FD_{2s}}[\underset{\sim}{v}] = \underset{\sim}{D}_{2s}\underset{\sim}{v}.$$

As s increases, so does the amount of work required to perform the multiplication on the right-hand side of (1.2_{2s}). Traditionally, finite-difference methods are employed with small values of s, $s = 1,2,3$, requiring a total amount of work of $N \cdot s$ operations (1 operation = vector addition + vector multiplication by a scalar). For large values of s, $\underset{\sim}{D}_{2s}$ becomes a full matrix whose multiplication requires an increasing amount of work, up to N^2 operations. This number of operations can be reduced, however, by taking into acount that the matrix $\underset{\sim}{D}_{2s}$ is a circulant one, and as such, can be diagonalized by the block Fourier matrix **F**. Specifically, let the block Fourier matrix **F** be given by

(1.4a) $$[\mathbf{F}]_{jk} = \frac{1}{N} \cdot e^{-ij'kh} \cdot I_m, \qquad 0 \leqslant j,k \leqslant N-1$$

with the conventional notation

(1.4b) $$\ell' \equiv \ell - n, \qquad n \equiv \text{integral part of } N/2,$$

to be used throughout the paper. We then have (see the appendix for details)

($1.3a_{2s}$) $$\underset{\sim}{D}_{2s} = N\mathbf{F}^{*}\underset{\sim}{\Lambda}_{2s}\mathbf{F}$$

with the block diagonal matrix $\underset{\sim}{\Lambda}_{2s} \equiv \underset{\sim}{\Lambda}_{2s}(h)$ given by

($1.3b_{2s}$) $$[\underset{\sim}{\Lambda}_{2s}]_{jj} = \lambda_{2s}^{(j')} \cdot I_m \equiv \frac{2i}{h} \cdot \sum_{k=1}^{s} k^{-1} \beta_k \sin(j'kh) \cdot I_m, \qquad 0<j<N-1.$$

Multiplication of $\underset{\sim}{D}_{2s}$ in its <u>spectral representation</u> (1.3_{2s}) can be efficiently implemented by two FFT's and N scalar multiplications which amount to $8N\log N$ operations.

In general, we consider an abstract discrete differentation operator, whose matrix representation $\underset{\sim}{D} \equiv \underset{\sim}{D}(h)$ is only required for the two basic properties of being an antisymmetric block circulant one

$$(1.5) \qquad [\underset{\sim}{D}]_{jk} = d_{[k-j]}\cdot I_m = -d_{[j-k]}\cdot I_m, \qquad [\ell] \equiv \ell[\mathrm{mod}N].$$

(The antisymmetry requirement which corresponds to centered type differencing is in fact not as essential, but will suffice for our discussion below.) Such matrices admit the spectral representation

$$(1.6) \qquad \underset{\sim}{D} = NF^{*}\underset{\sim}{\Lambda}F$$

with a block diagonal matrix $\underset{\sim}{\Lambda}$, whose diagonal consists of the so called amplification blocks

$$(1.7a) \qquad [\underset{\sim}{\Lambda}]_{jj} = \lambda^{(j')}\cdot I_m \equiv \sum_{k=0}^{N-1} d_k e^{ij'kh}\cdot I_m, \qquad 0<j<N-1;$$

since $\underset{\sim}{D}$ is assumed to be antisymmetric, $d_k + d_{N-k} = 0$, and hence

$$(1.7b) \qquad \lambda^{(j')} = 2i\cdot \sum_{k=1}^{n} d_k \sin(j'kh), \qquad 0<j<N-1.$$

The discrete differentation as given by the spectral representation of $\underset{\sim}{D}$, $\underset{\sim}{D}=NF^{*}\underset{\sim}{\Lambda}F$ may be interpreted now as follows: from the gridvalues $v_{\nu|0<\nu<N-1}$, we have the discrete Fourier modes $\hat{v}_{\omega'}$ given by

$$(1.8) \qquad [\hat{v}]_{\omega'} \equiv [Fv]_{\omega'} = \frac{1}{N}\cdot \sum_{\nu=0}^{N-1} e^{-i\omega'\nu h} v_{\nu}, \qquad -n<\omega'<N-1-n;$$

then, each one of these modes is discretely differentiated as it multiplied by the ampliciation factor $\lambda^{(\omega')}$, and finally, the differentiated modes

$\lambda(\omega')\hat{v}_{\omega'}\big|_{-n\leq\omega'\leq N-1-n}$, are transformed back into the physical gridspace upon multiplication by $N\mathbf{F}^{*} \equiv \mathbf{F}^{-1}$.

2. Stability of Finite Difference Approximations

Replacing the spatial derivative in (0.1) by its discrete counterpart $D \equiv D(h)$, we end up with the finite-difference approximation[1]

$$(2.1a)\qquad \partial_t v_\nu(t) = A(x_\nu)D(h)\left[v_\nu(t)\right] + B(x_\nu)v_\nu(t);$$

introducing the block diagonal matrices $\underset{\sim}{A} = \mathrm{diag}[A(x_0),\cdots,A(x_{N-1})]$ $\underset{\sim}{B} = \mathrm{diag}[B(x_0),\cdots,B(x_{N-1})]$, it can be put in a matrix form

$$(2.1b)\qquad \partial_t \underset{\sim}{v}(t) = \underset{\sim}{A}\underset{\sim}{D}\underset{\sim}{v}(t) + \underset{\sim}{B}\underset{\sim}{v}(t).$$

The time-dependent difference equation (2.1) serves as an approximation to the differential problem (0.1), in the sense that <u>any</u> smooth solution, u, of (0.1), satisfies (2.1) modulo a small local truncation error $\underset{\sim}{\tau}(h) \equiv \underset{\sim}{\tau}(h;t)$

$$(2.2)\qquad \partial_t \underset{\sim}{u}(t) = \underset{\sim}{A}\underset{\sim}{D}\underset{\sim}{u}(t) + \underset{\sim}{B}\underset{\sim}{u}(t) + \underset{\sim}{\tau}(h;t).$$

The approximation is said to be <u>accurate of order</u> α if $\|\underset{\sim}{\tau}(h)\| = \mathcal{O}[h^\alpha]$. With $\underset{\sim}{D} = \underset{\sim}{D}_{2s}$ for example, one obtains a difference approximation which is accurate of order $2s$, $\|\underset{\sim}{\tau}_{2s}(h)\| = \mathcal{O}[h^{2s}]$. To link the local order of accuracy

[1] Also termed as the method of lines, to distinguish from the fully discretized problem in both space and time.

with the desired global convergence rate of the approximation, one has to verify its stability. That is, the approximation (2.1) is said to be <u>stable</u> if for all sufficiently small h we have

$$(2.3) \qquad \|\exp[\underset{\sim}{AD}t]\| < K \equiv K_T, \qquad 0<t<T.$$

Observe that the stability definition is independent of the lower order term, $\underset{\sim}{By}$, the reason being that stability in the above sense is, in fact, insensitive to such low-order perturbations. This is the content of the following classical perturbation lemma, whose proof is given here for completeness as it will play an essential role in our discussion below. (see e.g. [15, Section 3.9], [16] and under a much more general setup [20]).

<u>Perturbation Lemma</u> Let **A** be a given linear operator such that

$$\|\exp[\mathbf{A}t]\| < K_T, \qquad 0<t<T.$$

Then, altering **A** by adding a "low-order" bounded perturbation **B**, retains the exponent boundedness

$$\|\exp[(\mathbf{A}+\mathbf{B})t]\| < K(t), \qquad K(t) = K_T e^{K_T \cdot \|\mathbf{B}\| \cdot t}, \qquad 0<t<T.$$

<u>Proof</u> The solution of the inhomogeneous linear differential equation

$$(2.4a) \qquad \partial_t w(t) = \mathbf{L}w(t) + G(t), \qquad w(t=0)=w(0)$$

is given by

$$\text{(2.4b)} \qquad w(t) = e^{Lt}w(0) + \int_{\xi=0}^{t} e^{L(t-\xi)}G(\xi)d\xi.$$

Applying (2.4b) with $\mathbf{L} = \mathbf{A} + \mathbf{B}$ and $G \equiv 0$, we then get

$$w(t) = e^{[(\mathbf{A}+\mathbf{B})t]}w(0);$$

hence (2.4a) can be also written in this case as the <u>inhomogeneous</u> problem $\partial_t w(t) = \mathbf{A}w(t) + \mathbf{B}e^{[(\mathbf{A}+\mathbf{B})t]}w(0)$. Applying (2.4b) once more, this time with $\mathbf{L} = \mathbf{A}$ and $G = \mathbf{B}e^{[(\mathbf{A}+\mathbf{B})t]}w(0)$, we obtain

$$w(t) = e^{[\mathbf{A}t]}w(0) + \int_{\xi=0}^{t} e^{[\mathbf{A}(t-\xi)]}\mathbf{B}e^{[(\mathbf{A}+\mathbf{B})\xi]}w(0)d\xi.$$

Equating the last two representations of $w(t)$ which are valid for <u>arbitrary</u> initial data $w(0)$, we arrive at the well-known identity

$$\exp[(\mathbf{A}+\mathbf{B})t] \equiv \exp[\mathbf{A}t] + \int_{\xi=0}^{t} \exp[\mathbf{A}(t-\xi)]\cdot\mathbf{B}\cdot\exp[(\mathbf{A}+\mathbf{B})\xi]d\xi.$$

Taking norms on both sides we find

$$K(t) \leqslant K_T + K_T\cdot\|\mathbf{B}\|\cdot\int_{\xi=0}^{t} K(\xi)d\xi;$$

by Gronwall inequality, we conclude that $\int_{\xi=0}^{t} K(\xi)d\xi \leqslant \|\mathbf{B}\|^{-1}\cdot[e^{K_T\cdot\|\mathbf{B}\|\cdot t}-1]$, and hence

$$K(t) \leqslant K_T + K_T\cdot\|\mathbf{B}\|\cdot\|\mathbf{B}\|^{-1}\cdot[e^{K_T\cdot\|\mathbf{B}\|\cdot t}-1] = K_Te^{K_T\cdot\|\mathbf{B}\|\cdot t},$$

which completes the proof. (We remark that similar arguments apply for the analogous question which concerns the discrete framework discussed in [15, Section 3.9]).

From the preturbation lemma we see that stability is equivalent to the boundedness of $\|\exp[(\underset{\approx}{A}\underset{\sim}{D}+\underset{\sim}{B})t]\|$, or -- what amounts to the same thing -- to the continuous dependence of the discrete solution $\underset{\sim}{v}(t)$ on its initial data $\underset{\sim}{v}(0)$

$$\text{(2.5)} \qquad \|\underset{\sim}{v}(t)\| = \|e^{[(\underset{\approx}{A}\underset{\sim}{D}+\underset{\sim}{B})t]}\underset{\sim}{v}(0)\| \leqslant K(t)\cdot\|\underset{\sim}{v}(0)\|;$$

indeed if stability holds in the sense that $\|\exp[\underset{\approx}{A}\underset{\sim}{D}t]\|$ is bounded, see (2.3), then by the perturbation lemma with $A = \underset{\approx}{A}\underset{\sim}{D}$ and $B = \underset{\sim}{B}$, so is $\|\exp[(\underset{\approx}{A}\underset{\sim}{D}+\underset{\sim}{B})t]\|$. On the other hand, if $\|\exp[(\underset{\approx}{A}\underset{\sim}{D}+\underset{\sim}{B})t]\|$ is bounded then by the perturbation lemma with $A = \underset{\approx}{A}\underset{\sim}{D}+\underset{\sim}{B}$ and $B = -\underset{\sim}{B}$, so is $\|\exp[\underset{\approx}{A}\underset{\sim}{D}t]\|$.

Granted stability, we can now estimate the <u>global error</u> $\underset{\sim}{E}(t) \equiv \underset{\sim}{u}(t) - \underset{\sim}{v}(t)$: subtracting (2.1) from (2.2), we find that it is governed by

$$\partial_t\underset{\sim}{E}(t) = (\underset{\approx}{A}\underset{\sim}{D}+\underset{\sim}{B})\underset{\sim}{E}(t) + \underset{\sim}{\tau}(h;t)$$

and hence is of the form

$$\underset{\sim}{E}(t) = e^{[(\underset{\approx}{A}\underset{\sim}{D}+\underset{\sim}{B})t]}\underset{\sim}{E}(t=0) + \int_{\xi=0}^{t} e^{[(\underset{\approx}{A}\underset{\sim}{D}+\underset{\sim}{B})(t-\xi)]}\underset{\sim}{\tau}(h;\xi)d\xi;$$

using the perturbation lemma, we end up with the error estimate

$$\|\underset{\sim}{E}(t)\| \leqslant K(t)\cdot\|\underset{\sim}{E}(t=0)\| + \sup_{0\leqslant\xi\leqslant t}\|\tau(h;\xi)\|\cdot\int_{\xi=0}^{t} K(\xi)d\xi.$$

Thus, if an α-order accurate approximation is initialized with α-order accurate data, $\|\underset{\sim}{E}(t=0)\| = \mathcal{O}[h^{\alpha}]$, its stability will retain α-order of convergence later on, $\|\underset{\sim}{E}(t)\| = \mathcal{O}[h^{\alpha}]$.

Verifying stability is our main objective in the rest of this section. We start by rewriting

$$\underset{\sim}{A}\underset{\sim}{D}t = {}^1\!/\!_2\,(\underset{\sim}{A}\underset{\sim}{D}+\underset{\sim}{D}\underset{\sim}{A})t + {}^1\!/\!_2\,(\underset{\sim}{A}\underset{\sim}{D}-\underset{\sim}{D}\underset{\sim}{A})t$$

where by the symmetry of $\underset{\sim}{A}$ and antisymmetry of $\underset{\sim}{D}$, the first term on the right is antisymmetric and, therefore, has a bounded exponent

$$\|\exp[\,{}^1\!/\!_2\,(\underset{\sim}{A}\underset{\sim}{D}+\underset{\sim}{D}\underset{\sim}{A})t\,]\| = 1;$$

hence, by the perturbation lemma, with the second term on the right viewed as a low-order perturbation of the first, the exponent of the sum of the two terms is bounded provided the second is

$$\|\exp[\underset{\sim}{A}\underset{\sim}{D}t]\| \leq K_T e^{\|{}^1\!/\!_2\,(\underset{\sim}{A}\underset{\sim}{D}-\underset{\sim}{D}\underset{\sim}{A})\|\cdot T}, \qquad 0\leq t\leq T. \tag{2.6}$$

Thus we are left with finding a bound for the symmetric part of $\underset{\sim}{A}\underset{\sim}{D}$, $\mathrm{Re}(\underset{\sim}{A}\underset{\sim}{D}) \equiv {}^1\!/\!_2\,(\underset{\sim}{A}\underset{\sim}{D}-\underset{\sim}{D}\underset{\sim}{A})$, whose (p,q) block entry is given by

$$[\,{}^1\!/\!_2\,(\underset{\sim}{A}\underset{\sim}{D}-\underset{\sim}{D}\underset{\sim}{A})\,]_{pq} = {}^1\!/\!_2\, d_{[p-q]}\cdot[A(x_p)-A(x_q)], \qquad 0\leq p,q\leq N-1;$$

since $A(x)$ is assumed symmetric 2π-periodic, $\|A(x_p)-A(x_q)\| \leq h\cdot \underset{0\leq x\leq 2\pi}{\mathrm{Max}} \|A'(x)\|\cdot \mathrm{Min}[\,|p-q|,N-|p-q|\,]$, and hence ${}^1\!/\!_2\,(\underset{\sim}{A}\underset{\sim}{D}-\underset{\sim}{D}\underset{\sim}{A})$ is bounded entrywise and therefore in norm, by the matrix whose (p,q) block entry is given by

$$\frac{h}{2}\cdot\mathrm{Max}\|A'(x)\|\cdot|d_{[q-p]}|\cdot\mathrm{Min}[\,|p-q|,N-|p-q|\,]\cdot I_m.$$

This last matrix is a circulant one. In the appendix we show, see Corollary (A.8), that the norm of such matrix does not exceed (and in fact equals, in our case) the absolute value sum of its elements along its first (p = 0) row, $\frac{h}{2}\cdot \mathrm{Max}\|A'(x)\|\cdot \sum_{q=0}^{N-1} \mathrm{Min}[q,N-q]\cdot|d_q|$; recalling that $\underset{\sim}{D}$ is antisymmetric, $d_k + d_{N-k} = 0$, we finally end up with the desired bound

$$(2.7) \qquad \| {}^1\!/_2\,(\underset{\approx}{A}\underset{\approx}{D}-\underset{\approx}{D}\underset{\approx}{A})\| \leq h\cdot \sum_{k=1}^{n} k|d_k|\cdot \max_{0\leq x\leq 2\pi} \|A'(x)\|.$$

Insterting the last bound into (2.6) we find

$$(2.8) \qquad \|\exp[\underset{\approx}{A}\underset{\sim}{D}t]\| \leq e^{\left[h\cdot \sum_{k=1}^{n} k|d_k|\cdot \mathrm{Max}\|A'(x)\|\cdot T\right]}, \qquad 0\leq t\leq T.$$

The above estimate serves as a discrete analogue to the standard energy estimate one has in the differential case, whose abstract version amounts to

$$(2.9) \|\exp[A(x)D_x t]\| \leq e^{\| {}^1\!/_2\,(A(x)D_x - D_x A(x))\|\cdot T} \leq e^{{}^1\!/_2\,\cdot \mathrm{Max}\|A'(x)\|\cdot T}, \qquad 0\leq t\leq T.$$

In the case of a non-constant A, the two estimates, (2.8) and (2.9), differ, however, in the term $h\cdot \sum_{k=1}^{n} k|d_k|$ appearing in the first; to guarantee stability we assume this term to be bounded

$$(\mathrm{L}) \qquad h\cdot \sum_{k=1}^{n} k|d_k| \leq \mathrm{Const}.$$

Condition (L) assures us that the differencing operator $\underset{\sim}{D}$ is in a sense <u>local</u>, thus reflecting the local nature of differentiation D_x.

The antisymmetry, periodicity (= circulant form) and a locality characterization are essential properties shared by the differential operator, which the discrete differencing operator $\underset{\sim}{D}$ should retain as well, in order for an energy estimate (= stability) to be still valid under the discrete framework.

Regarding the locality requirement, we consider for example the centered divided differences D_{2s} for fixed values of s: these operators are clearly local as they employ information extracted from a fixed number of neighboring gridvalues; this is also reflected in their matrix representation $\underset{\sim}{D}_{2s}$ which has a finite width, w, defined as (see (1.5))

$$w(\underset{\sim}{D}) = \underset{1\leq k\leq n}{\text{Max}} \{k|d_k\neq 0\}.$$

We have $w(\underset{\sim}{D}=\underset{\sim}{D}_{2s}) = s$. Indeed, for such finite-width operators, $|d_k| \leq \text{Const.}h^{-1}$, and hence the locality condition $(\mathbb{L})$ is satisfied

$$h\cdot \sum_{k=1}^{n} k|d_k| \leq \text{Const.}w^2(\underset{\sim}{D}),$$

yielding stable approximations. As s increases, however, $\underset{\sim}{D}_{2s}$ becomes a full matrix which fails to satisfy the locality condition $(\mathbb{L})$. That is not to say that the approximation becomes unstable, since the locality condition we have obtained is sufficient yet unnecessary for stability. Sufficient and necessary locality conditions which guarantee stability are, as much as we are aware, not known; we expect, however, that a locality condition requiring $h\cdot|d_k|$ to decay faster than $1/k$, $k=1,\cdots,n$, is optimal for L_2-stability of the general variable-coefficients problem. (We remark that in the constant coefficient case [and in general, with a definite coefficient $A(x)$, where multiplication first by $A^{-1}(x)$ will bring us back to the former case], we have $\text{Max}\|A'(x)\| = 0$ and hence stability follows independently of a locality restriction, $\|\exp[\underset{\sim}{A}\underset{\sim}{D}t]\| < 1$, see (2.8).) To overcome the above difficulty, arising with "nonlocal" methods, two standard types of remedy can be employed — skew-symmetric differencing and introducing dissipation via smoothing; we discuss them next.

3. Skew-Symmetric Differencing and Smoothing

The spatial part of the differential system (0.1) is — apart from low order terms — a skew-selfajoint one

$$A(x)D_x + B(x) \equiv 1/2\,[A(x)D_x + D_x A(x)] + [B(x) - 1/2\,A'(x)];$$

skew-symmetric differencing is based on exploiting this formalism. Rewriting (0.1) in the form

$$\partial_t u(x,t) = \{1/2\,[A(x)\partial_x u(x,t) + \partial_x(A(x)u(x,t))] + [B(x) - 1/2\,A'(x)]\}u(x,t);$$

and replacing the spatial derivative by its discrete counterpart, we end up with the approximation

$$\partial_t \underset{\sim}{v}(t) = \{1/2\,[\underset{\sim}{A}\underset{\sim}{D} + \underset{\sim}{D}\underset{\sim}{A}] + [\underset{\sim}{B} - 1/2\,\underset{\sim}{A}']\}\underset{\sim}{v}(t).$$

The stability of the approximation in its skew-symmetric form is immediate: the first term inside the curly brackets is antisymmetric and hence its exponent is bounded by 1; by the perturbation lemma, therefore, the exponent of the sum of the two terms inside the curly brackets is bounded by the exponent of the norm of the second

$$\|\exp[\{\tfrac{1}{2}[\underset{\sim}{A}\underset{\sim}{D}+\underset{\sim}{D}\underset{\sim}{A}] + [\underset{\sim}{B}-\tfrac{1}{2}\underset{\sim}{A}']\}t]\| < e^{\|\underset{\sim}{B}-\frac{1}{2}\underset{\sim}{A}'\|T}, \qquad 0<t<T;$$

this is the exact energy estimate one has in the differential case. Thus skew-symmetric differencing, which is also available for a wide class of nonlinear problems, [17], maintains stability by retaining essential properties of the <u>whole</u> spatial operator $A(x)D_x + B(x)$ rather than differentiation itself; this is done, however, at the expense of doubling the total amount of work required.

An alternative less expensive procedure to maintain stability is smoothing, a topic which the rest of this section is devoted to. We start by going back to estimate (2.6) where we were left with bounding the symmetric part of $\underset{\sim}{A}\underset{\sim}{D}$, $\mathrm{Re}(\underset{\sim}{A}\underset{\sim}{D}) \equiv \tfrac{1}{2}(\underset{\sim}{A}\underset{\sim}{D}-\underset{\sim}{D}\underset{\sim}{A})$.

Employing the spectral representation of $\underset{\sim}{D}$, which we write as $\underset{\sim}{D} = (N^{1/2}\mathbf{F})^*\underset{\sim}{\Lambda}(N^{1/2}\mathbf{F})$, see (1.6), we obtain

$$(3.1) \qquad \tfrac{1}{2}(\underset{\sim}{A}\underset{\sim}{D}-\underset{\sim}{D}\underset{\sim}{A}) = \tfrac{1}{2}[\underset{\sim}{A}(N^{1/2}\mathbf{F})^*\underset{\sim}{\Lambda}(N^{1/2}\mathbf{F}) - (N^{1/2}\mathbf{F})^*\underset{\sim}{\Lambda}(N^{1/2}\mathbf{F})\underset{\sim}{A}];$$

multiplying by $N^{1/2}\mathbf{F}$ on the left, by $(N^{1/2}\mathbf{F})^*$ on the right, and observing that $N^{1/2}\mathbf{F}$ is unitary (e.g. (A.6) below), we find that the matrix above is unitarily similar and therefore equal in norm to

$$(3.2) \qquad \|\tfrac{1}{2}(\underset{\sim}{A}\underset{\sim}{D}-\underset{\sim}{D}\underset{\sim}{A})\| = \tfrac{1}{2}\|\{(N^{1/2}\mathbf{F})\underset{\sim}{A}(N^{1/2}\mathbf{F})^*\}\underset{\sim}{\Lambda} - \underset{\sim}{\Lambda}\{(N^{1/2}\mathbf{F})\underset{\sim}{A}(N^{1/2}\mathbf{F})^*\}\|.$$

Next, we turn to examine the matrix inside the curly brackets, whose (p,q) block entry is given by

$$(3.3) \qquad \{(N^{1/2}\mathbf{F})\underset{\sim}{A}(N^{1/2}\mathbf{F})^*\}_{p,q} = \frac{1}{N}\cdot\sum_{\nu=0}^{N-1} A(x_\nu)e^{i(q-p)\nu h};$$

using the Fourier expansion

$$A(x) = \sum_{\omega=-\infty}^{\infty} \hat{A}(\omega)e^{i\omega x}, \qquad \hat{A}(\omega) = \frac{1}{2\pi}\int_{\xi=0}^{2\pi} e^{-i\omega\xi}A(\xi)d\xi,$$

it can also be expressed as

$$\frac{1}{N}\cdot\sum_{\nu=0}^{N-1} A(x_\nu)e^{i(q-p)\nu h} = \frac{1}{N}\cdot\sum_{\nu=0}^{N-1}\Big(\sum_{\omega=-\infty}^{\infty}\hat{A}(\omega)e^{i\omega x_\nu}\Big)e^{i(q-p)\nu h}$$

$$(3.4) \qquad = \sum_{\omega=-\infty}^{\infty}\hat{A}(\omega)\cdot\frac{1}{N}\cdot\sum_{\nu=0}^{N-1}e^{i(q-p+\omega)\nu h} = \sum_{j=-\infty}^{\infty}\hat{A}(p-q+jN).$$

Having the representation of $\{(N^{1/2}\mathbf{F})\underset{\sim}{A}(N^{1/2}\mathbf{F})^*\}$ in (3.4) and recalling the diagonal structure of $\underset{\sim}{\Lambda}$ in (1.7), we conclude on account of (3.2) that the matrix $1/2\,(\underset{\sim}{A}\underset{\sim}{D}-\underset{\sim}{D}\underset{\sim}{A})$ is equal in norm to the matrix whose (p,q) block entry is given by

$$(3.5a) \qquad \big(\lambda^{(q')}-\lambda^{(p')}\big)\cdot\sum_{j=-\infty}^{\infty}\hat{A}(p-q+jN) \qquad 0\leqslant p,q\leqslant N-1.$$

We note that the locality condition $(\mathbb{L})$ can be deduced again at this stage, if we are to proceed as follows: from (1.7b) we find

$$(3.5b) \qquad \lambda^{(q')} - \lambda^{(p')} = 2i\cdot\sum_{k=1}^{n} d_k\cdot\big(\sin(q'kh)-\sin(p'kh)\big).$$

Since

$$|\sin(q'kh)-\sin(p'kh)| \leqslant k\cdot h\cdot \mathrm{Min}\big[|p-q|,N-|p-q|\big],$$

the matrix in (3.5) is bounded entrywise and therefore in norm, by the matrix whose (p,q) block entry is given by

$$2h\cdot\sum_{k=1}^{n} k|d_k|\cdot\{\mathrm{Min}[|p-q|,N-|p-q|]\cdot\sum_{j=-\infty}^{\infty}\|\hat{A}(p-q+jN)\|\}\cdot I_m.$$

The matrix in the above curly brackets is a circulant one. As before, its norm does not exceed the absolute value sum of its elements along the first row $(p=0)$ -- see Corollary (A.8) below; this sum in turn can be estimated in terms of the derivatives norm of $A(x)$. Thus, assuming the locality condition -- $h\cdot\sum_{k=1}^{n} k|d_k| < \mathrm{Const.}$ -- we conclude that $\frac{1}{2}(\underset{\sim}{A}\underset{\sim}{D}-\underset{\sim}{D}\underset{\sim}{A})$ and hence $\exp[\underset{\sim}{A}\underset{\sim}{D}t]$, $0<t<T$, have bounded norms, i.e., stability.

The merit of the representation (3.5) lies, however, in the possibility of expressing a locality condition in terms of the amplification blocks associated with $\underset{\sim}{D}$, $\lambda^{(k')}\cdot I_m$, rather than its entries $d_k\cdot I_m$. To this end we proceed as follows:

The matrix in (3.5) is written as the sum of two -- the first takes the zero j-index which we rewrite as

$$-\left(\frac{\lambda^{(q')}-\lambda^{(p')}}{q'-p'}\right)\cdot(p-q)\cdot\hat{A}(p-q); \tag{3.6a}$$

the second takes the rest of the j-indices

$$\left(\lambda^{(q')}-\lambda^{(p')}\right)\cdot\sum_{j\neq 0}\hat{A}(p-q+jN). \tag{3.6b}$$

It is the property of the finite difference methods that the first matrix in (3.6a) is bounded. For, $\lambda^{(j')} = 2i\cdot\sum d_k\sin(j'kh)$ represents the discrete differentation of the j' mode and as such, the order of magnitude of the difference $|\lambda^{(q')}-\lambda^{(p')}|$ should not exceed $\mathrm{Const.}|q'-p'|$. Hence the matrix in (3.6a) is bounded entrywise and therefore in norm, by the matrix whose (p,q) element is given by

$$\text{Const.}|p-q|\cdot\|\hat{A}(p-q)\|;$$

the norm of such a Toeplitz matrix -- see Corollary (A.11) -- does not exceed $\text{Const.}\cdot\sum_{\omega=0}^{N-1}|\omega|\|\hat{A}(\omega)\|$, which in turn, can be bounded by the norm of the derivatives of $A(x)$. Regarding the boundedness of the second matrix in (3.6b), we note that for $p-q$ bounded away from $jN, j \neq 0$, say $|p-q| < \theta N, \theta < 1$, we have $\|\sum_{j\neq 0}\hat{A}(p-q+jN)\| < C_{\gamma,\theta}N^{-\gamma}$ and hence for these nonextreme indices, the entries in (3.6b) are a'priori bounded -- in fact,they are negligibly small. For the rest of the indices, when $|p-q| \sim N$, i.e., when $\pm p' \sim \mp q' \sim n$, we must require the boundedness of $\lambda^{(q')},\lambda^{(p')}$. Thus the locality condition amounts to <u>the boundedness of the amplification blocks</u>, $\lambda^{(j')}\cdot I_m$, <u>associated with the high frequencies</u> $|j'| \sim n$. If this is the case, the matrix $\frac{1}{2}(\underset{\sim}{A}\underset{\sim}{D}-\underset{\sim}{D}\underset{\sim}{A})$ in its unitarily similar representation (3.5) is bounded and stability follows from (2.6).

The above situation is typical for all discrete methods, whose accuracy is determined by the <u>exactness</u> of differentiating the low modes, $\lambda^{(j')} \sim ij'$, while for their stability we need the <u>boundedness</u> of $|\lambda^{(j')}|$ associated with the highest modes, $|j'| \sim n$.[2] The combination of the two guarantee convergence, as the low modes carrying most of the information are accurately represented, while the highest modes are inaccurately represented, yet stability assures us that they are not amplified and hence rapidly tend to zero, just as is the case in the differential problem.

The two requirements -- accuracy and stability -- are well accommodated in difference methods having finite degree of accuracy; consider for example

[2] It should be emphasized that this stability restriction is, of course, only sufficient. Its necessity is still an open question.

the second accurate method where we have $\lambda_2^{(j')} = ih^{-1}\sin(j'h)$, see (1.3b$_2$), and hence $\lambda_2^{(j')} \sim ij'$ for $|j'| \sim 0$, i.e., accuracy, yet $|\lambda_2^{(j')}| = |h^{-1}\sin(j'h)| \leq$ Const. for $|j'| \sim n$, i.e., stability. The situation is less favorable, however, for highly accurate mehtods (of order N or more): the accuracy requirement $\lambda^{(j')} \sim ij'$ for the highest modes contradicts the stability restriction $|\lambda^{(j')}| \leq$ Const. as originated from the locality conditon. Observe that this latter contradiction still leads to a bound proportional at most to N, which corresponds in the differential case to the familiar situation of "losing one derivative."[3]

The smoothing procedure aiming at dissolving this contradiction by bounding the amplification factors associated with the high frequencies (or more generally -- the modes which these amplification factors multiply), yet leaving the lower accurate modes unharmed. For example, consider the so called Shuman filtering where

$$v_\nu \longrightarrow 1/4\left(v_{[\nu+1]}+v_{[\nu-1]}+2v_\nu\right)$$

is applied to the right hand side of (2.1a). In the Fourier space, it amounts to the further multipliction of the j' mode, $\hat{v}_{j'}$, by $1/2\cdot\left(1+\cos(j'h)\right)$; that is

$$\hat{v}_{j'} \longrightarrow 1/2\left(1+\cos(j'h)\right)\cdot\hat{v}_{j'}.$$

In other words, our smoothed discrete differentation operator $\underset{\sim}{D}_{\text{Shuman}}$ takes the form

[3] In fact, as we shall see later on, we have a loss of "one-half" derivative.

$$\underset{\sim}{D}_{\text{Shuman}} = NF^{*}\underset{\sim\sim}{\Lambda\Omega}_{\text{Shuman}}F$$

with

$$\underset{\sim}{\Omega}_{\text{Shuman}} = \operatorname{diag}[\tfrac{1}{2}(1+\cos(-nh))\cdot I_m,\cdots,\tfrac{1}{2}(1+\cos(((N-1-n)h))\cdot I_m],$$

which merely says that the amplification factors $\lambda^{(j')}$ were replaced by $\lambda^{(j')}\cdot\tfrac{1}{2}(1+\cos(j'h))$. For the highest modes we now have the desired boundedness -- in fact $|\lambda^{(j')}\cdot\tfrac{1}{2}(1+\cos(j'h))| \sim 0$ for $|j'| \sim n$. This is done, however, at the expense lowering the overall accuracy to a second one -- $\lambda^{(j')}\cdot\tfrac{1}{2}(1+\cos(j'h)) \approx ij' + \mathcal{O}[h^2]$ for $|j'| \sim 0$. In general, a linearly smoothed discrete differentation operator $\underset{\sim}{D}_{*}$ may take the form

$$(3.7a) \qquad \underset{\sim}{D}_{*} = NF^{*}\underset{\sim}{\Lambda}_{*}F, \qquad \underset{\sim}{\Lambda}_{*} \equiv \underset{\sim\sim}{\Lambda\Omega}$$

with

$$(3.7b) \qquad \underset{\sim}{\Omega} = \operatorname{diag}[\sigma^{(-n)}\cdot I_m,\cdots,\sigma^{(N-1-n)}\cdot I_m].$$

The requirement of both accuracy and stability can be now put in the concise form

$$(3.8) \qquad \sigma^{(j')} = \begin{cases} \approx 1 & \text{for } |j'| \text{ bounded away from } n \text{ (= accuracy)} \\ \downarrow 0 & \text{for } |j'| \uparrow n \qquad\qquad\qquad \text{(= stability)} \end{cases}$$

In [12], Majda et. al. advocated the use of exponential cut-off σ- smoothing when dealing with the <u>propagation of singularities</u> in linear problems. In [11], Kreiss and Oliger suggested a nonlinear smoothing, whose linearized version amounts to a polynomial cut-off of degree ≥ 2. In fact, a polynomial cut-off of degree <u>one</u> or more will suffice to compensate for the loss of <u>one</u> derivative we have observed earlier. To work out this last case in some detail, fix $\theta < 1$ and let

$$(3.9) \qquad \sigma^{(j')} = \begin{cases} 1 & |j'| < \theta n \\ \text{Const.}(|j'|-\theta n)^{-1} & \theta n < |j'| < n. \end{cases}$$

The adjusted amplification factors are now given by $\lambda^{(j')} \longrightarrow \lambda^{(j')}\sigma^{(j')}$. A fixed portion of the N frequencies is left unchanged maintaining the original order of accuracy. Regarding stability, we refer back to the real symmetric part $\underset{\sim}{AD}$ in its unitarily equivalent form (3.5), which is written as the sum of two, see (3.6): the first

$$-\left(\frac{\lambda^{(q')}\sigma^{(q')}-\lambda^{(p')}\sigma^{(p')}}{q'-p'}\right)\cdot(p-q)\cdot\hat{A}(p-q)$$

is bounded by the norm of the derivatives of $A(x)$ as we argued before; the second matrix

$$\left(\lambda^{(q')}\sigma^{(q')}-\lambda^{(p')}\sigma^{(p')}\right)\cdot \sum_{j\neq 0}\hat{A}(p-q+jN)$$

is likewise bounded. Indeed, for $|p-q| < \frac{1+\theta}{2}\cdot N$, its entries are negligibly small — they are bounded by $N\cdot \sum_{j\neq 0}\|\hat{A}(p-q+jN)\| < C_{\gamma,\theta}N^{-\gamma+1}$. For $|p-q| > \frac{1+\theta}{2}\cdot N$ we either have $p > (1+\theta)n$, $q < (1-\theta)n$, i.e. $p' > \theta n$, $q' < -\theta n$, or, the roles of p and q are reversed. In either case $|p'| > \theta n$, $|q'| > \theta n$ and therefore the latter matrix is essentially bounded entrywise and therefore in norm, by the matrix whose (p,q) entry is given by

$$(3.10) \qquad \left|\left(\frac{q'}{|q'|-\theta n} - \frac{p'}{|p'|-\theta n}\right)\right|\cdot \sum_{j\neq 0}\|\hat{A}(p-q+jN)\|\cdot I_m \qquad |p'|,|q'|>\theta n.$$

A direct calculation shows that this matrix is indeed bounded in terms of the norm of the derivatives of $A(x)$.

Part II. The Fourier Method

4. The Fourier Differencing Operator

As before, we let $v(x)$ be a 2π-periodic m-dimensional vector-function, whose values $v_\nu \equiv v(x_\nu)$ are assumed known at the gridpoints $x_\nu = \nu h$, $h = \frac{2\pi}{N}$; to simplify the notation we consider first the case of odd number of gridpoints, $N = 2n+1$, $\nu = 0,1,\cdots,2n$. By Fourier differentiation we merely mean differentiation of the trigonometric interpolant of these gridvalues. That is, one construct the trigonometric interpolant

$$\tilde{v}(x) = \sum_{\omega=-n}^{n} \hat{v}_\omega e^{i\omega x} \tag{4.1a}$$

where the discrete Fourier coefficients $\hat{v}_\omega$ are given by, compare (1.8),

$$\hat{v}_\omega = \frac{1}{N} \cdot \sum_{\nu=0}^{2N} v_\nu e^{-i\omega\nu h}, \qquad -n<\omega<n. \tag{4.1b}$$

The Fourier differentiation then takes the form

$$\frac{\partial\tilde{v}}{\partial x}(x_\nu) = \sum_{\omega=-n}^{n} i\omega\tilde{v}_\omega e^{i\omega x_\nu}. \tag{4.2}$$

The above procedure consists of the following three basic steps. First, transforming from the discrete space $\underset{\sim}{v} \equiv (v_0,\cdots,v_{2n})'$ into the Fourier space of amplitudes $\hat{\underset{\sim}{v}} \equiv (\hat{v}_{-n},\cdots,\hat{v}_n)'$:

$$\hat{\underset{\sim}{v}} = F\underset{\sim}{v}; \tag{4.3}$$

next, differentiation in the Fourier space takes place:

$$\hat{\underset{\sim}{v}} \longleftarrow \underset{\sim}{\Lambda}_F \hat{\underset{\sim}{v}}$$

with $\underset{\sim}{\Lambda}_F$ denoting the block diagonal matrix

(4.4a) $$\underset{\sim}{\Lambda}_F = \mathrm{diag}\left[-in\cdot I_m, -i(n-1)\cdot I_m, \cdots, i(n-1)\cdot I_m, in\cdot I_m\right];$$

finally, the differentiated amplitudes $\underset{\sim}{\Lambda}_F \hat{\underset{\sim}{v}}$ are transformed back into the discrete physical space:

$$\partial_F[\underset{\sim}{v}] = \mathbf{F}^{-1}[\underset{\sim}{\Lambda}_F \hat{\underset{\sim}{v}}], \quad \mathbf{F}^{-1} = N\mathbf{F}^*.$$

Added altogether, the Fourier differencing operator $\underset{\sim}{F}$ amounts to multiplication by

(4.4b) $$\underset{\sim}{F} = N\mathbf{F}^* \underset{\sim}{\Lambda}_F \mathbf{F},$$

which can be efficiently implemented by two FFT's and N scalar multiplications requiring $8N\log N$ operations.

An explicit representation of the Fourier differencing matrix, $\underset{\sim}{F}$, can be obtained by differentiating the interpolant formula, cf., [22, Chapter X]

$$\tilde{v}(x) = \frac{2}{2n+1} \cdot \sum_{\nu=0}^{2n} v_\nu \mathbf{K}(x - x_\nu), \qquad \mathbf{K}(\xi) = \frac{\sin[(n+{}^1\!/_2)\xi]}{2\sin({}^1\!/_2\,\xi)},$$

giving

(4.5) $$[\underset{\sim}{F}]_{jk} = -\frac{(-1)^{k-j}}{2\sin[(k-j)\pi/(2n+1)]} \cdot I_m, \qquad 0 < j,k < 2n.$$

Thus, it falls into the category of antisymmetric block circulant matrices discussed above, see (1.5),

(4.6) $$[\underset{\sim}{F}]_{jk} = d^{(F)}_{[k-j]} \cdot I_m \qquad d^{(F)}_{\ell} = \frac{(-1)^{\ell+1}}{2\sin[\ell\pi/(2n+1)]}$$

with a spectral representation given by (4.4). Indeed, a straightforward calcualtion, cf., Forenberg [3], shows

$$\underset{\sim}{\Lambda}_F = \lim_{s\to\infty}\underset{\sim}{\Lambda}_{2s};$$

that is, *the Fourier differencing can be viewed as a special centered finite differencing, based on an ever increasing number of gridpoints extended in a periodic way*, $\underset{\sim}{F} = \lim_{s\to\infty}\underset{\sim}{D}_{2s}$. Continuing with this point of view, we conclude that while the Fourier differencing enjoys an "infinite order of accuracy" -- a statement to be made precise below -- it is a nonlocal one. We would like to examine the role these properties play in the Fourier method, based on replacing spatial derivatives by Fourier differencing. We start by discussing the all important aliasing phenomenon.

5. Aliasing

Let $w(x)$ be a smooth 2π-periodic m-dimensional vector-function, with a formal Fourier expansion

$$w(x) = \sum_{\omega=-\infty}^{\infty} \hat{w}(\omega)e^{i\omega x} \tag{5.1a}$$

where the Fourier coefficients $\hat{w}(\omega)$ are given by

$$\hat{w}(\omega) = \frac{1}{2\pi}\int_{\xi=0}^{2\pi} w(\xi)e^{-i\omega\xi}d\xi \tag{5.1b}$$

Its interpolant $\tilde{w}(x)$ based on the sampled gridvalues $w(x_\nu), \nu = 0,1,\cdots,2n$, is given by

$$\tilde{w}(x) = \sum_{\omega=-n}^{n} \hat{w}_\omega e^{i\omega x} \tag{5.2a}$$

with the discrete Fourier coefficients

$$(5.2b) \qquad \hat{w}_\omega = \frac{1}{N}\cdot \sum_{\nu=0}^{2n} w(x_\nu)e^{-i\omega\nu h}, \qquad |\omega|\le n.$$

The relation between the Fourier coefficients $\hat{w}(\omega)$ of $w(x)$ and the coefficients $\hat{w}_\omega$ of its interpolant $\tilde{w}(x)$, is contained in the following

<u>Aliasing Lemma</u> For $w(x)$ as above we have

$$(5.3) \qquad \hat{w}_\omega = \sum_{k=-\infty}^{\infty} \hat{w}(\omega+kN).$$

<u>Proof</u> Inserting (5.1a) into (5.2b) we obtain

$$\hat{w}_\nu = \frac{1}{N}\cdot \sum_{\nu=0}^{2n} w(x_\nu)e^{-i\omega\nu h} = \frac{1}{N}\cdot \sum_{\nu=0}^{2n} \Big[\sum_{\mu=-\infty}^{\infty} \hat{w}(\mu)e^{i\mu x_\nu}\Big]e^{-i\omega\nu h}.$$

By the assumed smoothness of $w(x)$, summation can be interchanged, yielding

$$\hat{w} = \sum_{\mu=-\infty}^{\infty} \hat{w}(\mu)\cdot\frac{1}{N}\cdot\sum_{\nu=0}^{2n} e^{i\nu[\mu-\omega]h} = \sum_{k=-\infty}^{\infty} \hat{w}(\omega+kN),$$

as the second sum in the middle term is nonvanishing only for those indices μ such that $[\mu-\omega] = 0$, i.e., $\mu = \omega + kN$. This completes the proof.

Next, we consider the error between the gridfunction $w(x)$ and its equidistant interpolant $\tilde{w}(x)$. Rewriting $w(x) = \Big[\sum_{|\omega|\le n} + \sum_{|\omega|>n}\Big]\hat{w}(\omega)e^{i\omega x}$, and, with the help of the aliasing lemma,

$$\tilde{w}(x) = \sum_{|\omega|\le n} \hat{w}(\omega)e^{i\omega x} + \sum_{|\omega|\le n} \Big(\sum_{k\neq 0} \hat{w}(\omega+kN)\Big)e^{i\omega x},$$

we see that the difference $w(x) - \tilde{w}(x)$ is given as the sum of two basic contributions: the first, the truncation error, consisting of the higher truncated modes for $|\omega| > n$

$$(5.4a) \qquad \text{Truncation}\,[w(x)] = \sum_{|\omega|>n} \hat{w}(\omega)e^{i\omega x},$$

and the second, the aliasing error consisting of the higher aliased modes which were folded back on the lower ones, $|\omega| < n$, because of the finite resolution of our grid

$$(5.4b) \qquad \text{Aliasing}\,[w(x)] = -\sum_{|\omega|<n} \{ \sum_{k\neq 0} \hat{w}[\omega+k(2n+1)]\}e^{i\omega x}.$$

Observe that while the truncation error invloves modes higher than n, the aliasing error involves modes less or equal to n; hence the two are orthogonal with respect to each other, and the size of the difference $w(x) - \tilde{w}(x)$ is given by

$$(5.5a) \qquad \|w(x)-\tilde{w}(x)\|^2 = \|\text{Truncation}(w)\|^2 + \|\text{Aliasing}(w)\|^2.$$

By Parseval's relation, the two squared terms on the right are given respectively by

$$(5.5b) \qquad \|\text{Truncation}(w)\|^2 = \sum_{|\omega|>n} |\hat{w}(\omega)|^2$$

$$(5.5c) \qquad \|\text{Aliasing}(w)\|^2 = \sum_{|\omega|<n} | \sum_{k\neq 0} \hat{w}[\omega+k(2n+1)]|^2.$$

In both terms only the high amplitudes -- those associated with modes higher than n - are being summed. Since these high amplitudes tend rapidly to zero, i.e., for smooth $w(x)$ we have $|\hat{w}(\omega)| < C_\gamma(1+|\omega|)^{-\gamma}$ for any $\gamma > 0$, it follows that the two terms have the same error contribution of order $C_\gamma \cdot N^{(-\gamma+1)}$. Likewise we find that the derivative of $w(x)$, $\partial_x w(x)$, differs from the differentiated interpolant, $\partial_x \tilde{w}(x)$ by

$$\|\partial_x w(x)-\partial_x \tilde{w}(x)\|^2 = \sum_{|\omega|>n} |\omega|^2 \cdot |\hat{w}(\omega)|^2 + \sum_{|\omega|<n} |\omega|^2 \cdot | \sum_{k\neq 0} \hat{w}[\omega+k(2n+1)]|^2$$

which is of order $C_\gamma N^{(-\gamma+2)}$. As pointed out above, the Fourier differencing of $w(x)$ is in fact the exact differentiation of the interpolant $\tilde{w}(x)$. We therefore conclude that the error we commit by differentiating $\tilde{w}(x)$ rather than $w(x)$ is of the negligibly small order $C_\delta h^\delta$ for any $\delta > 0$. It is in this sense that we say the Fourier differencing has "infinite order accuracy."

Finally, we use the aliasing lemma to show the isometry between the discrete and continuous space functions. Precisely, consider the discrete space of gridfunctions $\underset{\sim}{y} = (y_0, \cdots y_{2n})', \underset{\sim}{z} = (z_0, \cdots, z_{2n})'$ equipped with the discrete inner product $(\cdot,\cdot)$

$$(5.6a) \qquad (\underset{\sim}{y},\underset{\sim}{z}) \equiv h \cdot \sum_{\nu=0}^{2n} z_\nu^* y_\nu,$$

as the discrete analogue of the space of 2π-periodic vector functions, $y(x), z(x)$ with

$$(5.6b) \qquad \left(y(x), z(x)\right) = \int_0^{2\pi} z^*(\xi) y(\xi) d\xi.$$

The above mentioned isometry now takes the concise form

$$(5.7) \qquad \left(\tilde{y}(x), \tilde{z}(x)\right) = (\underset{\sim}{y},\underset{\sim}{z}).$$

Indeed, consider the scalar 2π-periodic function $w(x) = 2\pi\cdot\tilde{z}^*(x)\tilde{y}(x)$. While the left-hand side is, by definition, $\hat{w}(\omega=0)$, the right hand side is, by definition, $\hat{w}_{\omega=0}$. According to the aliasing lemma, the two differ by the sum of amplitudes associated with aliased modes higher than $2n$, $\sum_{k\neq 0}\hat{w}[k(2n+1)]$. This sum is vanishing, however, since $w(x)$ being a trigonometric polynomial of degree $2n$ at most, contains no modes higher than $2n$.

6. Stability of the Fourier Method

In this section we study the stability of the Fourier method where spatial differentiation in (0.1) is carried out by Fourier differencing. According to the perturbation lemma we can safely neglect the low-order term assuming $B = 0$, and hence our approximation takes the form

$$\partial_t v_\nu(t) = L\tilde{v}|_{x=x_\nu} \tag{6.1a}$$

with the operator L given by

$$L = A(x)D_x. \tag{6.1b}$$

Indeed, the stability question as discussed above is relevant here, i.e., the unboundedness of the amplification blocks, see (4.4a), $\lambda_F^{(j')} = ij'\cdot I_m$ requires smoothing of the highest modes, in agreement with the nonlocality of the method, see (4.6), $h\cdot\sum_{k=1}^{n} k|d_k^{(F)}| = \mathcal{O}(1/h)$. The representation given below is from a somewhat different point of view, and in fact, it is the one that motivated our discussion in Section 3 above.

Multiplying (6.1a) by hv_ν^* and summing over all gridpoints we obtain

$$h\cdot\sum_{\nu=0}^{2n-1} v_\nu^* \partial_t v_\nu(t) = h\cdot\sum_{\nu=0}^{2n-1} v_\nu^* L\tilde{v}_{|x=x_\nu} = (\underset{\sim}{L\tilde{v}},\underset{\sim}{v}).$$

Taking real parts on both sides and making use of the isometry concluded in Section 5, we find

$$(6.2)\qquad \frac{d}{dt}\|\tilde{v}\|^2 = 2\mathrm{Re}\Big[h\sum_{\nu=0}^{2n-1} v_\nu^* \partial_t v_\nu(t)\Big] = 2\mathrm{Re}\big[(\underset{\sim}{L\tilde{v}},\underset{\sim}{v})\big] = 2\mathrm{Re}\big[(\widetilde{L\tilde{v}},\tilde{v})\big].$$

The crucial step now, involves splitting the right hand side into the sum of two terms: the first consists of the <u>exact</u> differentiation

$$(6.3a)\qquad 2\mathrm{Re}\big[(L\tilde{v},\tilde{v})\big] = \big([L+L^*]\tilde{v},\tilde{v}\big),$$

the second consists of the deviation from the exact differentiation

$$(6.3b)\qquad 2\mathrm{Re}\big[(\widetilde{L\tilde{v}}-L\tilde{v},\tilde{v})\big];$$

that is we have

$$(6.4)\qquad 2\mathrm{Re}\big[(\widetilde{L\tilde{v}},\tilde{v})\big] = 2\mathrm{Re}\big[(L\tilde{v},\tilde{v})\big] + 2\mathrm{Re}\big[(\widetilde{L\tilde{v}}-L\tilde{v},\tilde{v})\big]$$

in complete analogy to the splitting of the matrix in (3.5) into (3.6a) and (3.6b) as we introduced before.

That the first term in (6.3a) is bounded by $\mathrm{Const.}\|\tilde{v}\|^2$ is a property solely of the <u>differential</u> operator L, called semi-boundedness, which can be easily verified in our case by integration by parts,

$$(6.5)\qquad \big|\big([L+L^*]\tilde{v},\tilde{v}\big)\big| < \mathrm{Const.}\cdot\|\tilde{v}\|^2, \qquad \mathrm{Const} = {}^1\!/_2\cdot\mathrm{Max}\|A'(x)\|$$

in complete agreement with the exponential behavior indicated in (2.9). Thus we are left with estimating the second term in (6.3b). It is exactly this term which measures by how much we deviate from the differential energy estimate whose abstract version quoted in (2.9).

To this end we recall that the difference between $w = L\tilde{v}$ and its interpolant $\tilde{w} = \widetilde{L\tilde{v}}$ consists of two basic contributions -- the truncation error (5.4a) and the aliasing error (5.4b). The point to note here is that the truncation error being the sum of modes *higher* than n, is orthogonal to the n-degree interpolant $\tilde{v}$, and hence its contribution to the deviation term (6.3b) is completely suppressed. In other words, *it is solely the aliasing error in the representation of the differential operator* L -- *or what amounts to the same thing, of the coefficient matrix* $A(x)$ -- *which determines the stability of the discrete approximation* (6.1). To see how it comes about one compute the amplitudes of $L\tilde{v}$ as the convolution sum

$$\widehat{(L\tilde{v})}(\omega) = \sum_{q=-n}^{n} iq\cdot\hat{A}(\omega-q)\hat{v}_q \qquad -\infty<\omega<\infty;$$

hence the aliasing error is given by, see (5.4b)

$$\text{Aliasing}[L\tilde{v}] = -\sum_{|\omega|\le n}\Big\{\sum_{|q|\le n}\sum_{k\neq 0} iq\cdot\hat{A}[\omega-q+k(2n+1)]\hat{v}_q\Big\}e^{i\omega x}.$$

Multiplying by $\tilde{v}$ and making use of Parseval relation we find

$$\big(\widetilde{L\tilde{v}}-L\tilde{v},\tilde{v}\big) \equiv \big(\text{Aliasing}[L\tilde{v}],\tilde{v}\big) = -i\cdot\sum_{|p|,|q|\le n}\hat{v}^{*}_p\cdot q\cdot\sum_{k\neq 0}\hat{A}[p-q+k(2n+1)]\hat{v}_q;$$

taking the symmetric part we finally conclude that

$$(6.6) \qquad 2\mathrm{Re}\left(\widetilde{L\tilde{v}}-L\tilde{v},\tilde{v}\right) = i\cdot\sum_{|p|,|q|\leq n} \hat{v}_p^*\{(q-p)\cdot\sum_{k\neq 0}\hat{A}[p-q+k(2n+1)]\}\hat{v}_q$$

Our aim is trying to estimate the right-hand side in terms of $\|\tilde{v}\|^2$ -- by so doing, then together with (6.5) we will end up with an energy estimate

$$(6.7a) \qquad \frac{d}{dt}\|\tilde{v}\|^2 \leq \text{Const.}\cdot\|\tilde{v}\|^2,$$

whose integration assures us the continuous dependence of the solution on its initial data, i.e., stability, see (2.5)

$$(6.7b) \qquad \|\underset{\sim}{v}(t)\|^2 \equiv \|\tilde{v}(t)\|^2 \leq K(t)\cdot\|\tilde{v}(0)\|^2 \equiv K(t)\cdot\|\underset{\sim}{v}(0)\|^2.$$

To assert that the right-hand side of (6.6) does not exceed $\text{Const.}\|\tilde{v}\|^2 \equiv \text{Const.}\cdot\sum_{|\omega|\leq n}|\hat{v}_\omega|^2$ for <u>all</u> possible amplitudes $\hat{v}_\omega$, is, by definition, equivalent to assert the boundedness of the matrix whose (p,q) entry is given in the above curly brackets

$$(6.8) \qquad [\mathscr{A}]_{pq} = (q-p)\cdot\sum_{k\neq 0}\hat{A}[p-q+k(2n+1)], \qquad -n\leq p,q\leq n,$$

compare with (3.6b). The above terms represent the pure effect of aliasing on the coefficient matrix $A(x)$ -- in the constant coefficients case, for example, no aliasing occur, $\hat{A}(\omega) = 0$, $\omega \neq 0$, so the terms in (6.8) and hence that in (6.3b) are vanishing which agrees with the earlier deduced stability in this constant coefficients case. Returning to the general variable coefficients case we first note -- regarding the (p,q) entry in (6.8) -- that for $|p-q|$ bounded away from $2n$, $|p-q| \leq \theta\cdot 2n$, $\theta < 1$, these entries are negligibly small, since by the smoothness of $A(x)$ we have

$$\| \sum_{k\neq 0} \hat{A}[p-q+k(2n+1)] \| \le C_{\gamma,\theta} N^{-\gamma};$$

however, when $|p-q|$ approaches $2n$, that is, when $p\uparrow n$ and $q\downarrow -n$ or vice versa, $\sum_{k\neq 0} \hat{A}[p-q+k(2n+1)]$ contains the lower modes of $A(x)$ whose amplitudes are of size $\mathcal{O}(1)$, and hence these entries are of size $\mathcal{O}(N=2n+1)$ -- *the matrix whose* (p,q) *entry is given in (6.8) is, therefore, unbounded, no matter how smooth* $A(x)$ *is*. Consider for example the case where $A(x)$ consists of only one mode -- the only nonzero entries in (6.8) are then the $(p,q) = (\pm n,\mp n)$ ones, given respectively by $\mp 2n\hat{A}(\omega=\mp 1)$, which amount to the unboundedness of $\mathcal{A}$. (Putting it in a different way, we see that in constrast to local finite-difference methods, compare (2.7), $Re(\underset{\sim}{A}\underset{\sim}{F}) = \frac{1}{2}(\underset{\sim}{A}\underset{\sim}{F}-\underset{\sim}{F}\underset{\sim}{A})$ is unbounded, no matter how smooth $A(x)$ is; indeed up to unitary similarity -- the latter differ from $\mathcal{A}$ by the bounded term in (6.3a)).

Nevertheless, the above unboundedness does not necessarily imply instability, as much as it indicates the shortcomings of the above method of proving it. We observe that the difficulty arises when trying to estimate the (p,q) entries with $p\uparrow n$ and $q\downarrow -n$ or vice versa, in either case, when $|p|,|q| \sim n$. These aliased entries interact with amplitudes associated with high modes $\hat{v}^*_p, \hat{v}_q$, see (6.6), which are expected -- *if* the method is stable -- to be of a negligible small size. That is, despite the unboundedness of $\mathcal{A}$ in (6.8), we can still have the boundedness of the aliased term in (6.6) provided *a′ priori* information on the decay rate of $|\hat{v}_\omega|$, $|\omega| \sim n$ is at our disposal; it is well known, however, that the L_2-norm $\|\tilde{v}\|^2$ is too weak to derive such an a′ priori information.

With this an mind, smoothing may be viewed as a procedure aiming at giving us such a′ priori information about the size of the highest amplitudes. Consider for example the case where $A(x)$ consists of *fixed* number, say r modes; then smoothing by cutting a fixed number of modes -- the last r ones, $\hat{v}_\omega = 0$, $|\omega|>n-r$ -- will guarantee stability, as the term in (6.6) will vanish in this case. (In particualr, when $r = 1$, one only needs to estimate the last amplitude $\hat{v}_n$. In the case of even number of gridpoints, such an estimate exists, since $\underset{\sim}{F}$ being an even order antisymmetric matrix has a double zero eigenvalue; this leads to the H^1-stability in this case derived in [7].[4]) In general, a milder smoothing as the linear cut-off introduced in (3.9) will suffice for stability.

In closing this section, we remark that by rewriting (6.6) in the form

$$2\mathrm{Re}\left(\widetilde{L\tilde{v}-L\tilde{v}},\tilde{v}\right) = i\cdot\sum_{|p|,|q|\leq n} \sqrt{1+|p|}\cdot\hat{v}_p^*\cdot\left\{\frac{(q-p)}{\sqrt{1+|q|}\sqrt{1+|p|}}\cdot\sum_{k\neq 0}\hat{A}\left[p-q+k(2n+1)\right]\right\}\cdot\sqrt{1+|q|}\cdot\hat{v}_q,$$

where the matrix in the last curly brackets is bounded, then together with (6.5) we are lead to the estimate

$$(6.9) \qquad \frac{d}{dt}\|\tilde{v}\|^2 \leq \text{Const.}\|\tilde{v}\|^2_{H^{1/2}}, \qquad \|\tilde{v}\|^2_{H^{1/2}} \equiv \sum_{|\omega|\leq n}(1+|\omega|^2)^{1/2}|\hat{v}_\omega|^2.$$

That is, there is a loss of "one-half" derivative. If some dissipation is present in the system, e.g., with $L = A(x)D_x + D_x^2$, the gain of one derivative from the second spatial differentiation dominates, and we end up with the desired stability, e.g. [11].

[4] See the appendix for details.

Part III. Galerkin Methods

7. The Galerkin Procedure

In this section we start discussing the Galerkin procedure, whose basic idea is to reduce our infinite dimensional differential problem by projecting it on a finite-dimensional subspace. Let the latter be spanned by a system of linearly independent 2π- periodic functions $\phi_p(x)$, $-n \leqslant p \leqslant n$.

To project (0.1), we seek for approximation of the form

$$v(x,t) = \sum_{q=-n}^{n} \hat{v}(q,t)\phi_q(x) \tag{7.1}$$

satisfying

$$\left(\frac{\partial v}{\partial t} - Lv, \phi_p\right) = 0, \quad p=-n,\cdots,n. \tag{7.2$_p$}$$

Inserting (7.1) into (7.2), we obtain for the vector of generalized Fourier coefficients, $\hat{\underset{\sim}{v}}(t) = \left(\hat{v}(-n,t),\cdots,\hat{v}(n,t)\right)'$, the following system of ordinary differential equations

$$M\partial_t \hat{\underset{\sim}{v}}(t) = G\hat{\underset{\sim}{v}}(t); \tag{7.3a}$$

here, M and G are (2n+1)- dimensional block matrices whose (p,q) entry is given respectively by

$$[M]_{pq} = \left(\phi_q, \phi_p\right)\cdot I_m, \quad [G]_{pq} = \left(L\phi_q, \phi_p\right) \tag{7.3b}$$

The stability of the resulting system is a direct consequence of the semi-boundedness of the differential operator L, cf. (6.5) -- integration by parts yields

$$\operatorname{Re}(Lw,w) \equiv \tfrac{1}{2}\left(\left[L+L^*\right]w,w\right) \leqslant \text{Const}\cdot\|w\|^2,$$

for any 2π-periodic vector-function $w(x)$. Indeed, multiplying $\hat{v}(p,t)$ by

(7.2_p), adding and taking real parts we find

$$(7.4) \qquad \tfrac{1}{2}\frac{d}{dt}\|v(t)\|^2 \equiv \mathrm{Re}\left(\frac{\partial v}{\partial t},v\right) = \mathrm{Re}(Lv,v) \leq \mathrm{Const.}\|v(t)\|^2;$$

integration of the last inequality gives us the usual stability estimate -- compare (2.9).

Unless chosen with care, the basis functions $\phi_k(x)$ may lead to an ill-conditioned mass matrix, M, whose inversion required in (7.3a) can be still found numerically disastrous. The most extensively studied choices of basis functions which avoid such situations are essentially two. The first uses <u>local</u> base functions inducing sparse, well-behaved mass matrices, leading to finite-difference/finite-element like methods; the second uses <u>global, orthonormal</u> base functions like

$$\phi_p(x) = \frac{1}{\sqrt{2\pi}}e^{ipx}, \qquad -n \leq p \leq n,$$

where the mass matrix reduces to the identity, $M = I$. We continue by discussing the latter case.

The expansion we seek in (7.1), amounts now to the truncated Fourier expansion, whose Fourier coefficients $\hat{\underset{\sim}{v}}(t)$ are determined by the Galerkin procedure

$$(7.4a) \qquad \partial_t\hat{\underset{\sim}{v}}(t) = G\hat{\underset{\sim}{v}}(t)$$

where G is given by

$$(7.4b) \qquad [G]_{pq} = iq\cdot\frac{1}{2\pi}\int_0^{2\pi} A(x)e^{-i(p-q)x}dx \equiv iq\cdot\hat{A}(p-q), \qquad -n \leq p,q \leq n.$$

(As before, see (6.1b), we have neglected the lower order term assuming $L =$

$A(x)D_x$). We remark that implementation of the Galerkin procedure can be carried out fast, i.e., using $\mathcal{O}$ (NlogN) operations, provided the exact Fourier coefficients $\hat{A}(\omega), |\omega|<n$, are given. For, the procedure consists of two basic steps: first, differentation which is translated here to multiplication by the diagonal matrix $\underset{\sim}{\Lambda}_F, [\underset{\sim}{\Lambda}_F]_{qq} = iq \cdot I_m$ is taking place, requiring $N = 2n+1$ operations; and next, multiplication by $A(x)$ reflected as a convolution sum in the Fourier space is in order, which requires multiplication by the Toeplitz matrix $\hat{A}(p-q)$ -- indeed, multiplication by a general Toeplitz matrix can be carried out fast when first imbedded into a circulant one, see the appendix for details.

To obtain the Fourier coefficients

$$\hat{A}(p-q) \equiv \frac{1}{2\pi} \int_0^{2\pi} A(x) e^{-i(p-q)x} dx \tag{7.5}$$

one may use different quadrature rules approximating the integral on the right-hand side. This in turn leads to a whole variety of discrete Galerkin methods which include the Fourier method as a special case.

8. Discretization

The Fourier-Galerkin procedure in a component-wise form reads

$$\partial_t \hat{v}(p,t) = \sum_{q=-n}^{n} \hat{A}(p-q) \cdot iq \cdot \hat{v}(q,t) \tag{8.1a}$$

where $\hat{A}(\omega)$ is the Fourier coefficient

$$\hat{A}(\omega) = \frac{1}{2\pi} \int_{x=0}^{2\pi} e^{-i\omega x} A(x) dx. \tag{8.1b}$$

To approximate the integral in the right-hand side, we use the trapezoidal rule, based on the $N = 2n+1$ equidistant points $x_\nu = \nu h, h = \frac{2\pi}{N}$,

$$(8.2) \qquad \hat{A}(\omega) \sim \frac{1}{N}\cdot \sum_{\nu=0}^{N-1} A(x_\nu)e^{i\omega x_\nu};$$

since $A(x)$ is assummed periodic, the trapezoidal rule serves our purpose as any other high-order quadrature rule — in fact, it is "infinitely order accurate" in the precise sense discussed in Section 5 above, cf. [2, Section 2.9].

Introducing the approximation (8.2) into (8.1a), we find that the term $\hat{A}(p-q)$ is replaced by, see (3.3)

$$(8.3) \qquad \hat{A}(p-q) \sim \frac{1}{N}\cdot \sum_{\nu=0}^{N-1} A(x_\nu)e^{i(q-p)\nu h} \equiv [N\mathbf{F}\underset{\sim}{A}\mathbf{F}^*]_{pq};$$

thus, the above discretization result in a system of ordinary differential equation for the vector of unknown amplitudes, still denoted here by $\hat{\underset{\sim}{v}}$,

$$(8.4) \qquad \partial_t\hat{\underset{\sim}{v}}(t) = N\mathbf{F}\underset{\sim}{A}\mathbf{F}^*\underset{\sim}{\Lambda}_F\hat{\underset{\sim}{v}}(t).$$

This is exactly the Fourier method for the <u>discrete</u> Fourier amplitudes $\hat{\underset{\sim}{v}}(t) \equiv \mathbf{F}\underset{\sim}{v}(t) = (\hat{v}_{-n}(t),\cdots,\hat{v}_n(t))'$ -- multiplication by $\mathbf{F}^{-1}$ on the left brings it back into its familiar form in the physical space, see (4.4b)

$$(8.5) \qquad \partial_t\underset{\sim}{v}(t) = \underset{\sim}{A}(N\mathbf{F}^*\underset{\sim}{\Lambda}_F\mathbf{F})\underset{\sim}{v}(t) = \underset{\sim}{A}\underset{\sim}{F}\underset{\sim}{v}(t).$$

That is, the exact spatial differentiation is carried out on the interpolant $\tilde{v}$, compare (6.1a).

To summarize, we have seen that equidistant approximation of (8.3) based on N gridpoints reduces the Fourier-Galerkin procedure into the Fourier method; the difference between the two lies exactly in the aliasing term $\sum_{j\neq 0} \hat{A}(p-q+jN)$ -- according to (3.4), this is the exact difference between the right and left-hand sides of (8.3). Since the Fourier-Galerkin procedure was shown to be stable, we thus shed a different light on the conclusion that stability of the Fourier method is solely determined by aliasing errors. To suppress the latter, one may either smooth or, alternatively, discretize the integral on (8.1b) using more than N gridpoints. We turn now to discuss the details of the latter case.

Let $M = (1+\varepsilon)N$ be the number of gridpoints $x_\nu = \nu h, h = \frac{2\pi}{M}$, $\nu=0,1,\cdots M-1$, and use to trapezoidal rule to approximate

$$(8.6) \qquad \hat{A}(p-q) \sim \frac{1}{M}\cdot \sum_{\nu=0}^{M-1} A(x_\nu)e^{i(q-p)\nu h};$$

when inserted into (8.1a), the resulting system is given by

$$(8.7a) \qquad \partial_t \hat{v}_p(t) = \sum_{q=-n}^{n} [\frac{1}{M}\cdot \sum_{\nu-0}^{M-1} A(x_\nu)e^{i(q-p)\nu h}]\cdot iq\cdot \hat{v}_q(t).$$

Here, we adopt the notation of the discrete Fourier coefficients for the computed amplitudes, $\hat{v}_\omega(t)$, in the spirit of earlier agruments. Observe that the matrix whose (p,q) entry, $-n\leq p,q\leq n$, is given in the last curly brackets, is not a circulant anymore as in the Fourier case where $M = N$, yet its multiplication as a Toeplitz one can be carried out fast. To verify stability, we rewrite the (p,q) entry in the last curly brackets with the help of (3.4)

$$(8.7b) \qquad \partial_t \hat{v}_p(t) = \sum_{q=-n}^{n} \Big[\sum_{j=-\infty}^{\infty} \hat{A}(p-q+jM) \Big] \cdot iq \cdot \hat{v}_q(t).$$

As usual, we break the second summation into two parts

$$\sum_{j=-\infty}^{\infty} \hat{A}(p-q+jM) = \hat{A}(p-q) + \sum_{j\neq 0} \hat{A}(p-q+jM);$$

the first corresponds to the semi-bounded differential operator and can be estimated as before, while the second represent the pure effect of aliasing which in this case is completely controlled since by the smoothness of $A(x)$ we have

$$\sum_{j\neq 0} \|\hat{A}(p-q+jM)\| \leq C_\gamma(\varepsilon N)^{-\gamma}, \qquad \gamma>0, \ -n\leq p,q\leq n.$$

Indeed, a second look in (8.7b) reveals that the approximation (8.7) can be viewed as the standard Fourier method based on M modes, the last $(1+\varepsilon)^{-1}M$ of which were cut off (in the notation of (3.8), we have $\sigma^{(j)} = 0$ for $(1+\varepsilon)^{-1}M < |j| \leq M$) -- such smoothing guarantees stability.

Appendix

A. On Toeplitz and Circulant Matrices

In this section we record some well-known information about Toeplitz and circulant matrices which proves useful within the discussion above

A block Toeplitz matrix $\mathscr{T}$ consists of m-dimensional block entries, the (j,k) of which depends only on its distance from the main diagonal, $[\mathscr{T}]_{jk} = t_{k-j}$,

(A.1)

$$\mathscr{T} \equiv \mathscr{T}(t_{1-N},\cdots,t_0,\cdots,t_{N-1}) = \begin{bmatrix} t_0 & t_1 & t_2 \cdot\ \cdot\ \cdot & t_{N-2} & t_{N-1} \\ t_{-1} & & & & t_{N-2} \\ t_{-2} & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & t_2 \\ \cdot & & & & \\ t_{2-N} & & & & t_1 \\ t_{1-N} & t_{2-N} & \cdot\ \cdot\ \cdot\ t_{-2} & t_{-1} & t_0 \end{bmatrix}$$

Thus, a Toeplitz matrix is completely determined by a (2N-1)-dimensional vector $\underline{t} \equiv (t_{1-N},\cdots,t_0,\cdots t_{N-1})$, its entries, t_ℓ, $-(N-1)<\ell<N-1$, are being m-dimensional blocks.

If further, the vector $\underline{t}$ is defined on its negative indices as the periodic extension of the positive ones, $t_{-\ell} = t_{N-\ell}$, $0<\ell<N-1$, i.e., $\mathscr{T}_{jk}$ only depends on (k-j)[modN], then the matrix is a block circulant one, $\mathscr{C}$, $[\mathscr{C}]_{jk} = c_{(k-j)[\text{modN}]}$

$$\text{(A.2)}\qquad \mathscr{C} \equiv \mathscr{C}(c_0,\cdots,c_{N-1}) = \begin{bmatrix} c_0 & c_1 & c_2 & \cdots & c_{N-2} & c_{N-1} \\ c_{N-1} & & & & & c_{N-2} \\ c_{N-2} & & & & & \vdots \\ \vdots & & & & & c_2 \\ c_2 & & & & & c_1 \\ c_1 & c_2 & \cdots & c_{N-2} & c_{N-1} & c_0 \end{bmatrix}$$

Thus, a circulant matrix is completely determined by a N- dimensional vector $\underline{c} \equiv (c_0,\cdots,c_{N-1})$, its entries, c_ℓ, $0 \leqslant \ell \leqslant N-1$, being m-dimensional blocks.

The essential ingredient in studying circulant matrices, is that they admit the spectral representation

$$\text{(A.3)}\qquad \mathscr{C}(\underline{c}) = (N^{1/2}\mathbf{F})^* \underset{\sim}{\Lambda}_c (N^{1/2}\mathbf{F}).$$

Here, $\mathbf{F}$ denotes the block Fourier matrix, compare (1.4)

$$\text{(A.4a)}\qquad [\mathbf{F}]_{jk} = \frac{1}{N}\cdot e^{-ij'kh}\cdot I_m, \qquad 0 \leqslant j,k \leqslant N-1$$

with the conventional notation

$$\text{(A.4b)}\qquad \ell' = \ell - n, \qquad n \equiv \text{integral part of } N/2,$$

and $\underset{\sim}{\Lambda}_c$ is a block diagonal matrix given by

$$\text{(A.5)}\qquad [\underset{\sim}{\Lambda}_c]_{jj} = \sum_{\ell=0}^{N-1} e^{ij'\ell h}\cdot c_\ell$$

Verification of (A.3) is a straightforward one - the (j,k) entry of the right-hand side of (A.3) amounts to

$$\{(N^{1/2}\mathbf{F})^*\underset{\sim}{\Lambda}_c(N^{1/2}\mathbf{F})\}_{jk} = N\cdot\sum_{p=0}^{N-1}[F^*]_{jp}[\underset{\sim}{\Lambda}_c]_{pp}[F]_{pk}$$

$$= N\cdot\sum_{p=0}^{N-1}\frac{1}{N}\cdot e^{ip'jh}\cdot\Big[\sum_{\ell=0}^{N-1} e^{ip'\ell h}\cdot c_\ell\Big]\cdot\frac{1}{N}\cdot e^{-ip'kh}$$

$$= \frac{1}{N}\cdot\sum_{\ell=0}^{N-1} c_\ell\sum_{p=0}^{N-1} e^{ip'(\ell+j-k)h};$$

the second summation on the right is vanishing unless $\ell+j-k = 0[\mathrm{modN}]$, i.e., unless $\ell = (k-j)[\mathrm{modN}]$ where

$$\frac{1}{N}\cdot\sum_{\ell=0}^{N-1} c_\ell\sum_{p=0}^{N-1} e^{ip'(\ell+j-k)h}\Big|_{\ell=(k-j)[\mathrm{modN}]} = c_{(k-j)[\mathrm{modN}]} \equiv [\mathscr{C}]_{jk}.$$

Consideration of the block identity matrix I_N as a circulant one, with $\underline{c} = (I_m,0_m,\cdots 0_m)$, gives us from (A.3) and (A.5) that

$$I_N = (N^{1/2}\mathbf{F})^*(N^{1/2}\mathbf{F}); \tag{A.6}$$

that is, the matrix $N^{1/2}\mathbf{F}$ is a unitary one. Since the spectrum and the L_2-norm of a matrix are invariant under such unitary transformations, it follows from (A.3) that for general circulant matrices, $\mathscr{C}$, both are identical with those of block diagonal $\underset{\sim}{\Lambda}_c$. In particular we have

<u>Lemma (A.7)</u> For a block circulant matrix $\mathscr{C}(\underline{c})$ we have

$$\|\mathscr{C}(\underline{c})\| = \max_{0\leq j\leq N-1}\Big\|\sum_{\ell=0}^{N-1} e^{ij\ell\frac{2\pi}{N}}\cdot c_\ell\Big\|. \tag{A.7}$$

Proof The norm of a block diagonal matrix is given by the largest norm of its diagonal entries. Cosmetic reindexing of these diagonal entries in (A.5) gives us (A.7).

As an immediate corollary we have

Corollary (A.8) The norm of a scalar circulant matrix does not exceed the absolute value sum of its elements along its first row.

Proof In fact, from (A.7) we have the more general

$$\|\mathscr{C}(\underline{c})\| \leq \sum_{\ell=0}^{N-1} \|c_\ell\|. \tag{A.8}$$

The corollary is just a restatement of that last inequality for the scalar case, where $c_\ell = c_\ell \cdot I_m$.

Next, we employ the information just obtained for circulant matrices, for Toeplitz ones, with the help of the basic

Lemma (A.9) Any N-dimensional block Toeplitz matrix can be imbedded into a 2N-dimensional block circulant one.

Proof consider the block Toeplitz mattix $\mathscr{T} = \mathscr{T}(\underline{t})$ with $\underline{t} = (t_{1-N},\cdots,t_0,\cdots,t_{N-1})$. Denote $\underline{t}^- = (t_{1-N},\cdots,t_{-1})$, $\underline{t}^+ = (t_1,\cdots t_{N-1})$ and define the associated Toeplitz matrix $\mathscr{S} \equiv \mathscr{S}_s = \mathscr{T}(\underline{t}^+,s,\underline{t}^-)$ where s can be any fixed block. It is readily verified that

(A.9a)
$$\mathscr{C} = \begin{pmatrix} \mathscr{T} & \mathscr{S} \\ \mathscr{S} & \mathscr{T} \end{pmatrix}$$

is a 2N-dimensional block circulant

(A.9b)
$$\mathscr{C} = \mathscr{C}(\underline{c}), \qquad \underline{c} = (t_0, \underline{t}^+, s, \underline{t}^-);$$

in entrywise form we have

$$\mathscr{C} = \begin{bmatrix}
t_0 & t_1 & \cdots & t_{N-2} & t_{N-1} & s & t_{1-N} & \cdots & t_{-2} & t_{-1} \\
t_{-1} & & & & t_{N-2} & t_{N-1} & & & & t_{-2} \\
\vdots & & & & \vdots & \vdots & & & & \vdots \\
t_{2-N} & & & & t_1 & & & & & t_{1-N} \\
t_{1-N} & t_{2-N} & \cdots & t_{-1} & t_0 & t_1 & \cdots & & t_{N-1} & s \\
s & t_{1-N} & \cdots & t_{-2} & t_{-1} & t_0 & t_1 & \cdots & t_{N-2} & t_{N-1} \\
t_{N-1} & & & & t_{-2} & t_{-1} & & & & t_{N-2} \\
\vdots & & & & \vdots & \vdots & & & & \vdots \\
t_2 & & & & t_{1-N} & t_{2-N} & & & & t_1 \\
t_1 & t_2 & \cdots & t_{N-1} & s & t_{1-N} & t_{2-N} & \cdots & t_{-1} & t_0
\end{bmatrix}$$

Remark Rewriting $\mathscr{C}$ in (A.9c) as $\mathscr{T}(\underline{t})$, $\underline{t} = (\underline{t}^+, s, \underline{t}, s, \underline{t}^-)$ clarifies that the imbedding was made psooible by the process of periodic doubling.

Making use of Lemma (A.9), we have

Corollary (A.10) Multiplication of an N-dimensional block Toeplitz matrix can be implemented 'fast', i.e., using $\mathscr{O}(N\log N)$ block operations.

<u>Proof</u> We want to compute $\underset{\sim}{z} = \mathscr{T}\underset{\sim}{w}$, where $\mathscr{T}$ is an N-dimensional Toeplitz matrix and $\underset{\sim}{w}$ a given N-dimensional vector. For that purpose, imbed $\mathscr{T}$ into $\mathscr{C} = \begin{pmatrix} \mathscr{T} & \mathscr{S} \\ \mathscr{S} & \mathscr{T} \end{pmatrix}$ and compute $\underset{\sim}{z}_* = \mathscr{C}\underset{\sim}{w}_*$, $\underset{\sim}{w}_* = (\underset{\sim}{w}, \underset{\sim}{0}_N)'$ - as $\mathscr{C}$ being a circulant, this last multiplication can be implemented fast using its spectral representation (A.3) with two FFT's requiring $\mathscr{O}(N\log N)$ operations. The first N components of $\underset{\sim}{z}_*$ are then the desired vector $\underset{\sim}{z}$.

<u>Corollary (A.11)</u> For a block Toeplitz matrix $\mathscr{T}(\underline{t})$ we have

$$\text{(A.11)} \qquad \|\mathscr{T}(\underline{t})\| \leqslant \underset{0\leqslant j\leqslant 2N-1}{\text{Max}} \Big\| \sum_{\ell=-(N-1)}^{N-1} e^{ij\ell\frac{\pi}{N}} \cdot t_\ell \Big\|$$

<u>Proof</u> Imbed $\mathscr{T}(\underline{t})$ into $\mathscr{C}(\underline{c})$ with $\underline{c} = (t_0, \underline{t}^+, 0, \underline{t}^-)$ we then have from Lemma (A.7)

$$\|\mathscr{T}(\underline{t})\| \leqslant \|\mathscr{C}(\underline{c})\| = \underset{0\leqslant j\leqslant 2N-1}{\text{Max}} \Big\| \sum_{\ell=0}^{2N-1} e^{ij\ell\frac{2\pi}{2N}} \cdot c_\ell \Big\| ;$$

Insertion of the specific values of the blocks c_ℓ in this case, shows that the upper-bound on the right-hand side equals

$$\underset{0\leqslant j\leqslant 2N-1}{\text{Max}} \Big\| \sum_{\ell=0}^{N-1} e^{ij\ell\frac{\pi}{N}} \cdot t_\ell + (-1)^\ell \cdot 0 + \sum_{\ell=N+1}^{2N-1} e^{ij\ell\frac{\pi}{N}} \cdot t_{\ell-2N} \Big\| \equiv \underset{0\leqslant j\leqslant 2N-1}{\text{Max}} \Big\| \sum_{\ell=-(N-1)}^{N-1} e^{ij\ell\frac{\pi}{N}} \cdot t_\ell \Big\| .$$

<u>Remark</u> Making use of the freedom in choosing the block s along the main diagonal of the associated Toeplitz $\mathscr{S}_s$ (which was taken to be zero above), we similarily get

$$\|\mathscr{T}(\underline{t})\| \leqslant \inf_s \Big[\underset{0\leqslant j\leqslant 2N-1}{\text{Max}} \Big\| (-1)^j \cdot s + \sum_{\ell=-(N-1)}^{N-1} e^{ij\ell\frac{\pi}{N}} \cdot t_\ell \Big\| \Big] .$$

Corresponding to Corollary (A.8) we have

Corollary (A.12) The norm of a scalar Toeplitz matrix does not exceed the absolute value sum of its elements along its first and last rows.

B. The Fourier Method - The Case of Even Number of Gridpoints

The Fourier method is usually implemented with an even number of gridpoints, N = 2n; to be exact, with N being an integral power of 2, in which case the Cooley-Tukey variants of FFT are optimal. Here we record the slightly different formulas governing this case.

Assume v_ν are known gridvalues at $x_\nu = \nu h$, $h = \frac{2\pi}{N} \equiv \frac{\pi}{n}$, $\nu=0,1,\cdots 2n-1$. Their Fourier differencing amouonts to differentiation of their trigonometric interpolant

$$\text{(B.1a)} \qquad \tilde{v}(x) = \sum_{w=-n}^{n\;\prime\prime} \hat{v}_\omega e^{i\omega x}.$$

Here, the double prime denotes, as usual, halving the first and last terms, and the discrete Fourier coefficients $\hat{v}_\omega$ are given by

$$\text{(B.1b)} \qquad \hat{v}_\omega = \frac{1}{N}\cdot \sum_{\nu=0}^{2n-1} v_\nu e^{-i\omega\nu h}.$$

An explicit representation of the Fourier differencing matrix $\underset{\sim}{F}$ transforming $\underset{\sim}{v} \equiv (v_0,\cdots v_{2n-1})'$ into $\partial_F[\underset{\sim}{v}] \equiv (D_x\tilde{v}|_{x_0},\cdots,D_x\tilde{v}|_{x_{2n-1}})'$, can be obtained by differentiating the interpolant formula, e.g. [22, Chapter X]

$$\text{(B.2)} \qquad \tilde{v}(x) = \frac{1}{N}\cdot \sum_{\nu=0}^{2n-1} v_\nu K(x-x_\nu) \qquad K(\xi) = \frac{\sin(n\xi)}{2tg(\,^1\!/_2\,\xi)}$$

giving

(B.3) $$[\underset{\sim}{F}]_{jk} = -(-1)^{k-j}\cdot\cot\big((k-j)\pi/2n\big)\cdot I_m, \qquad 0<j,k<2n-1.$$

As a block circulant matrix, it admits the spectral representation

(B.4a) $$\underset{\sim}{F} = NF^{*}\underset{\sim}{\Lambda}_{F}F$$

with

(B.4b) $$\underset{\sim}{\Lambda}_{F} = \mathrm{diag}\big[0\cdot I_m, -i(n-1)\cdot I_m, \cdots, 0\cdot I_m, \cdots, i(n-1)\cdot I_m\big].$$

Observe that zero is a double eigenvalue in this case - this is necessairly so as $\underset{\sim}{F}$ being an antisymmetric <u>even</u> dimensional matrix, having the other complex eigenvalues in pairs. The left eigenvectors corresponding to the double zero eigenvalue are

(B.5a) $$(\underset{\sim}{z}^{(0)})'\underset{\sim}{F} = 0, \qquad (\underset{\sim}{z}^{(0)}) = \big(I_m, I_m, \cdots, I_m\big)'$$

(B.5b) $$(\underset{\sim}{z}^{(n)})'\underset{\sim}{F} = 0, \qquad (\underset{\sim}{z}^{(n)}) = \big(I_m, -I_m, \cdots, I_m, -I_m\big)'$$

asserting the exactness of the differentiation (B.3) for $\tilde{v}(x) = \text{Const.}$ and $\tilde{v}(x) = \cos(nx)$, respectively (compare [7, Lemma 1.1]).

The Fourier method for (0.1) with $L = A(x)D_x$, is of the form

(B.6) $$\partial_t\underset{\sim}{v}(t) = \underset{\sim}{A}\underset{\sim}{F}\underset{\sim}{v}(t).$$

Stability analysis in this case is similar to that introduced in Section 7 for the case of odd number of gridpoints. That is, to estimate the real symmetric part of $\big(\widetilde{L\tilde{v}},\tilde{v}\big)$, see (6.2), we use the aliasing formula which still reads, see (5.3)

$$\hat{w}_{\omega} = \sum_{k=-\infty}^{\infty} \hat{w}(\omega+kN), \qquad N=2n$$

leading us to an examination of the aliasing term, see (6.6)

$$\text{(B.7)} \qquad 2\mathrm{Re}\left(\widetilde{L\tilde{v}},\tilde{v}\right) = i\cdot\sum_{|p|,|q|<n} \hat{v}_p^*\left[(q-p)\cdot \sum_{k\neq 0} \hat{A}(p-q+2kn)\right]\hat{v}_q .$$

In this case, however, we have a priori information about the last discrete Fourier coefficient $\hat{v}_n$. To see how it comes about, multiply (B.6) by $\underset{\sim}{F}$ on the left, and rename the new variable $\underset{\sim}{w} = \underset{\approx}{Fv}$ for which we find

$$\partial_t \underset{\sim}{w}(t) = \underset{\approx}{FA}\underset{\sim}{w}(t);$$

next, multiplication by $(\underset{\sim}{z}^{(n)})^*$ on the left and using (B.5b) we conclude that $(\underset{\sim}{z}^{(n)})^*\underset{\sim}{w}(t)$ - which, by definition, coincides with $\hat{w}_{\pm n}(t) \equiv \sum_{\nu=0}^{2n-1} w_\nu \cos(nx_\nu)$ - remains constant in time, $\hat{w}_{\pm n}(t) = \hat{w}_{\pm n}(t=0) = 0$. Thus, returning to the aliasing term in (B.7), it is enough to sum only the first (n-1) modes

$$2\mathrm{Re}\left(\widetilde{L\tilde{w}}-L\tilde{w},\tilde{w}\right) = i\cdot\sum_{|p|,|q|<n-1} \hat{w}_p^*\left[(q-p) \sum_{k\neq 0} \hat{A}(p-q+2kn)\right]\hat{w}_q .$$

In particular, if A(x) contains only one mode, the vanishing right-hand side results in the desired energy estimate for $\underset{\sim}{w} = \underset{\approx}{Fv}$ which in turn amounts to an H^1- stability in this case -- see [7].

References

[1] J. W. Cooley and J. W. Tukey, An algorithm for the machine computation of complex Fourier series, Math. Comp., 19, (1965), pp. 297-301.

[2] P. J. Davis and P. Rabinowitz, Methods of Numerical Integration, Academic Press, New York, 1975.

[3] B. Fornberg, On a Fourier method for the integration of hyperbolic equations, SIAM J. Numer. Anal., 12, (1975), pp. 509-528.

[4] B. Fornberg and G. B. Whitham, A numerical and theoretical study of certain nonlinear wave phenomena, Philos. Trans. Roy. Soc. London, 289, (1978), pp. 373-404.

[5] D. Gottlieb and S. A. Orszag, Numerical Analysis of Spectral Methods: Theory and Applications, CBMS Regional Conference Series in Applied Mathematics 26, SIAM, Philadelphia, 1977.

[6] D. Gottlieb, L. Lustman and S. A. Orszag, Spectral calculations of one-dimensional inviscid compressible flow, SIAM J. Sci. Statist. Comput., 2, (1980), pp. 296-310.

[7] D. Gottlieb, S. A. Orszag and E. Turkel, Stability of pseudospectral and finite-difference methods for variable coefficient problems, Math. Comp., 37, (1981), pp. 293-305.

[8] O. H. Hald, Convergence of Fourier methods for Navier-Stokes equations, J. Comput. Phys., 40, (1981), pp. 305-317.

[9] H.-O. Kreiss and J. Oliger, Comparison of accurate methods for the integration of hyperbolic equations, Tellus, 27, (1972), pp. 199-215.

[10] H.-O. Kreiss and J. Oliger, Methods for the Approximate Solution of Time Dependent Problems, GARP Publications Series, No. 10, World Meteorological Organization, Geneva, 1973.

[11] H.-O. Kreiss and J. Oliger, Stability of the Fourier method, SIAM J. Numer. Anal., 16, (1979), pp. 421-433.

[12] A. Majda, J. McDonough and S. Osher, The Fourier method for nonsmooth initial data, Math. Comp., 32, (1978), pp. 1041-1081.

[13] S. A. Orszag, Numerical simulation of incompressible flows within simple boundaries: I. Galerkin (spectral) representations, Stud. Appl. Math., 50, (1971), pp. 293-327.

[14] S. A. Orszag, Spectral methods for problems in complex geometries, Adv. in Computer Meth. for Partial Differential Equations III., R. Vichnevetshy and R. S. Stepelman ed., IMCAS, 1979, pp. 149-157.

[15] R. D. Richmyer and K. W. Morton, Difference Methods for Initial Value Problems, Interscience, New York, 1967.

[16] G. Strang, On strong hyperbolicity, J. Math. Kyoto Univ., 6, (1967), pp. 397-417.

[17] E. Tadmor, Skew-Selfadjoint form for systems of conservation laws, J. Math. Anal. Appl., to appear.

[18] E. Turkel, Numerical methods for large-scale time-dependent partial differential equations, Computational Fluid Dynamics, 2, Hemisphere Publishing Corp., Von Karman Inst., 1980, pp. 127-262.

[19] R. Voigt, ed., Proceedings for the Symposium on Spectral Methods for Partial Differential Equations, CBMS Regional Conference Series in Applied Mathematics, SIAM, to appear.

[20] K. Yosida, Functional Analysis, 2nd ed., Springer Verlag, 1968.

[21] T. A. Zang and M. Y. Hussaini, Mixed spectral-finite difference approximations for slightly viscous flows, Proc. of the 7th Intl. Conf. in Numerical Methods in Fluid Dynamics, 1981, pp. 461-466.

[22] A. Zygmund, Trigonometrical Series, Volumes I & II, Cambridge University Press, Cambridge, 1968.

1. Report No. NASA CR-172149	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle Finite-Difference, Spectral and Galerkin Methods for Time-Dependent Problems	5. Report Date June 1983 6. Performing Organization Code
7. Author(s) Eitan Tadmor	8. Performing Organization Report No. 83-22
9. Performing Organization Name and Address Institute for Computer Applications in Science and Engineering Mail Stop 132C, NASA Langley Research Center Hampton, VA 23665	10. Work Unit No. 11. Contract or Grant No. NAS1-17070
12. Sponsoring Agency Name and Address National Aeronautics and Space Administration Washington, D.C. 20546	13. Type of Report and Period Covered Contractor Report 14. Sponsoring Agency Code

15. Supplementary Notes

Langley Technical Monitor: Robert H. Tolson
Final Report

16. Abstract

We survey finite-difference, spectral and Galerkin methods for the approximate solution of time-dependent problems. A unified discussion on their accuracy, stability and convergence is given. In particular, the dilemma of high accuracy versus stability is studied in some detail.

17. Key Words (Suggested by Author(s)) discrete methods accuracy stability smoothing	18. Distribution Statement 64 Numerical Analysis Unclassified-Unlimited

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 57	22. Price A04

For sale by the National Technical Information Service, Springfield, Virginia 22161

End of Document