SIAM REVIEW
Vol. 29, No. 4, December 1987

001

STABILITY ANALYSIS OF FINITE-DIFFERENCE, PSEUDOSPECTRAL AND FOURIER-GALERKIN APPROXIMATIONS FOR TIME-DEPENDENT PROBLEMS*

EITAN TADMOR†

Abstract. We consider finite-difference, pseudospectral and Fourier–Galerkin methods for the approximate solution of time-dependent problems. The paper provides a *unified* framework for the stability analysis of all three discrete methods. In particular, the problem of stability for highly accurate stencils is studied in some detail.

Key words. accuracy, aliasing, circulant matrices, convergence, discrete Fourier transform, finite-difference methods, Fourier method, Fourier–Galerkin method, hyperbolic equations, stability, truncation error

AMS(MOS) subject classifications. Primary 65M10; secondary 35A12

Introduction. Finite-difference methods are today a classical tool for the approximate solution of time-dependent problems. These methods are equipped with a well-developed theory for analyzing their various properties [34], [21]. Here, the concepts of consistency, stability and convergence, familiar to all practitioners in the field, play an essential role.

During the past decade, other discrete methods for the approximate solution of such problems have gained popularity. Primary examples are the pseudospectral and Galerkin-type methods, e.g., [2], [6], [8], [10]–[12], [20], [28]–[30]. At the same time the stability and convergence analysis of these methods has also rapidly developed [2]–[4], [6], [8]–[11], [13]–[15], [17], [18], [20], [27], [36].

The purpose of this paper is to provide a *unified* framework for studying the properties of consistency, stability and convergence of all three discrete methods, namely the finite-difference, pseudospectral, and Galerkin schemes.

As a model problem we consider the symmetric hyperbolic system

$$\frac{\partial}{\partial t}u(x,t)=A(x)\frac{\partial}{\partial x}u(x,t)+B(x)u(x,t), \qquad t\geqq 0, \tag{0.1}$$

with initial conditions, $u(x, 0)$, given at $t = 0$. For this problem to be correctly posed, appropriate boundary conditions should also be specified. At present, however, a satisfactory analysis of such *numerical* boundary conditions is available only in the case of finite-difference methods [16], [19], [31]; consequently, we restrict our attention to a unified discussion for the *periodic* problem, which requires no special boundary treatment.

To approximate the problem in hand, one may proceed in two stages. First, at each time level the solution is projected into a finite-dimensional space, say of

* Received by the editors June 21, 1983; accepted for publication (in revised form) May 20, 1986. This work was supported in part by the National Aeronautics and Space Administration under contract NAS1-17070 while the author was in residence at the Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, Virginia 23665. Additional support was provided by U.S.–Israel Binational Science Foundation grant 85-00346, and by National Science Foundation grant DMS85-03294 and Army Research Office grant DAAG-85-K-0190. The author is a Bat-Sheva Foundation Fellow.

† School of Mathematical Sciences, Tel Aviv University, Ramat Aviv, 69 978 Tel Aviv, Israel.

dimension N. Thus, in the case of finite-difference and pseudospectral methods the N-projection consists of N equidistant grid values, while the (periodic) Fourier–Galerkin schemes are usually calculated in terms of their first N Fourier coefficients. In either case, it is the finite-dimensional projection which is actually calculated by the numerical scheme. The numerical scheme itself is then constructed in the second stage by replacing the spatial differentiation with its discrete counterpart. Being finite-dimensional, this discrete version of differentiation admits an $N \times N$ matrix representation, the so-called differentiation matrix.

At this point our numerical scheme is a semidiscrete algorithm, i.e., it amounts to a system of ordinary differential equations governing the calculated N-projection. We focus our attention on the important question of the stability of such a system. Once stability is established, the resulting stable system can be integrated by an appropriate ODE solver.

We begin in Part I by discussing finite-difference methods. Such methods are usually employed with a fixed, N-independent order of accuracy. Yet, by periodic extension of the differencing stencil, one may consider highly accurate finite-difference methods, say of order N or more, e.g., [20], [9]. Our discussion will apply equally to the low, fixed order accurate methods as well as to the highly accurate ones.

Studying the properties of finite-difference methods can be carried out by the familiar von Neumann analysis, where the Fourier symbols of the difference scheme are examined [34], [21, §9]. In our discussion, there is a particular emphasis on how the accuracy and stability properties of finite-difference methods are directly determined by the corresponding differentiation matrices, rather than by their corresponding Fourier symbols. The reason for taking this approach is two-fold. First, we do so in order to shed light on the classical von Neumann analysis from a point of view slightly different than the usual one. The second and main motivation in choosing this approach is its relevance to the other two methods discussed in later sections. In other words, the discussion on finite-difference methods in terms of their differentiation matrices, carried out in the first part, will lead into the second and third parts where the same *unified* stability discussion will apply to the pseudospectral as well as the Galerkin methods. We will thus achieve the main objective of this paper, that is, to point out the intimate relation within the stability analysis of all three methods.

Each of the three methods is identified with a different differentiation matrix. In order to maintain a unified discussion with regard to all three of them, we consider a rather general differentiation matrix. Such a matrix is assumed to share with the differentiation operator the properties of being an antisymmetric and periodic, i.e., circulant, matrix. To guarantee stability, we also require that the differentiation matrix meet a certain locality restriction. Specifically, this locality restriction requires the *boundedness* of the Fourier symbols associated with the high modes of the scheme. On the other hand, a reliable numerical scheme should satisfy an accuracy requirement as dictated by the *exactness* of differencing the lower modes. The combinaton of these two requirements guarantees the *convergence* of the calculated numerical N-projection to the exact solution. Indeed, the lower modes carrying most of the information will be accurately represented by the numerical model; while this need not be the case with the higher modes, stability will assure us that these high modes are not amplified and hence rapidly tend to zero in complete analogy with the differential set-up.

We distinguish between discrete methods having a fixed N-independent order of accuracy such as the classical fixed "low" order accurate difference methods, and highly accurate methods of order N or more. In the first case, the fixed order of accuracy concerns only the lower modes, and therefore does not interfere with the stability restriction placed on the higher ones. However, in the second case of high accuracy, the accuracy and stability requirements contradict each other. That is, the exactness in differencing the high modes (= high accuracy), results in the unboundedness of the corresponding Fourier symbols (= instability). Trying to resolve this contradiction, we are led to the discussion closing the first part of the paper, where we deal with skew-symmetric differencing and smoothing procedures. In both cases one wishes to bound the Fourier symbols associated with the high modes while retaining the accuracy of the lower modes.

We continue in Part II, studying the (periodic) pseudospectral Fourier method. This method can be viewed as the infinite limit of periodic center differencing [9]. Consequently, our previous discussion on the highly accurate finite-difference methods equally applies. In fact, it is particularly relevant for the infinite-order accurate Fourier method. Here, we also analyze the stability question from still a slightly different point of view. To this end, we first introduce the aliasing formula, relating the Fourier coefficients of a periodic function to those of its equidistant interpolant. We then proceed to develop the stability analysis of the Fourier method, based upon this aliasing formula. The aliasing errors are then shown to play a key role here, since they dominate the problem of stability versus high accuracy in this case. In order to resolve this problem and to guarantee stability, two commonly used solutions are suggested in the literature. These were already encountered in connection with the highly accurate finite-difference methods. Namely, one may use skew-symmetric Fourier differencing [20], [37] or employ an appropriate smoothing procedure [1], [10], [15], [22], [25]. As a final note to the second part, we close the circle here with the previously studied finite-difference methods: the latter can be viewed as special cases of the Fourier method which differ in their *built-in* smoothing recipe.

In Part III, we conclude with a third type of discrete methods—the Galerkin-type methods. In order to stay within the unified framework, we confine ourselves to the periodic Fourier–Galerkin method. This, in turn, enables us to elaborate on the close connection between the stability analysis of the Fourier–Galerkin method on the one hand, and both finite-difference and pseudospectral methods on the other hand.

The Fourier–Galerkin schemes are evaluated in terms of the first N *exact* Fourier coefficients of the solution. The error committed in this case consists solely of *truncation errors*. Consequently, the stability of these schemes is intimately related to the correctness of the differential model itself. Once the exact Fourier coefficients are discretized, additional aliasing errors are introduced, and our stability study carried out in the previous sections becomes relevant. In particular, finite-difference and pseudospectral methods, with or without smoothing, can be viewed as special cases of the Fourier–Galerkin method; they can be classified according to the specific quadrature rule they employ in order to discretize the exact Fourier–Galerkin coefficients.

Finally, in order to make the paper self contained, we have collected in the appendix some basic properties of Toeplitz and circulant matrices; these will play a vital role in the 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 $x_\nu = \nu h$, $h = 2\pi/N$, $\nu = 0, 1, \cdots, N-1$. 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[1] $\underset{\sim}{v} \equiv (v_0, \cdots, v_{N-1})'$ into the vector of divided differences $\partial_{FD_2}[\underset{\sim}{v}] \equiv (D_2(h)[v_0], \cdots, D_2(h)[v_{N-1}])'$ 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 $\underset{\sim}{D}_2 \equiv \underset{\sim}{D}_2(h)$ consists of m-dimensional block entries given by

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

Similarly, fourth- and sixth-order accurate centered divided differences are given respectively by

$$(1.1_4) \qquad \begin{aligned} D_4(h)[v(x)] &= \frac{4D_2(h) - D_2(2h)}{3}[v(x)] \\ &= \frac{8[v(x+h)-v(x-h)] - [v(x+2h)-v(x-2h)]}{12h}, \end{aligned}$$

$$(1.1_6)$$

$$\begin{aligned} &D_6(h)[v(x)] \\ &\quad = \frac{15D_2(h) - 6D_2(2h) + D_2(3h)}{10}[v(x)] \\ &\quad = \frac{45[v(x+h)-v(x-h)] - 9[v(x+2h)-v(x-2h)] + [v(x+3h)-v(x-3h)]}{60h} \end{aligned}$$

with the corresponding matrix representations

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

[1] We denote the transpose of a vector, w, by a prime, w', we use a star to denote its conjugate transpose w^*, and we let $\|w\| \equiv (w^* w)^{1/2}$ denote the usual Euclidean norm. Similar notation is used for matrices.

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

Observe that the matrices $\underset{\sim}{D}_{2s}$, $s=1,2,3$ are antisymmetric matrices which admit a block circulant form. By this we mean that the (j,k) block entry of such matrices depends only on $(j-k)[\text{mod } N]$.

The examples above illustrate special cases of a more general recipe for $2s$-order accurate centered differencing which is given by [20, §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, each one of these differencng stencils is connected with an antisymmetric block circulant differentiation matrix, $\underset{\sim}{D}_{2s}\equiv\underset{\sim}{D}_{2s}(h)$, such that

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

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, fixed (i.e., N independent) values of s, e.g., $s=1,2,3$, where a total amount of work of $N\cdot s$ operations is required (1 operation = vector addition + vector multiplication by a scalar). For large values of s, of order N or more, $\underset{\sim}{D}_{2s}$ becomes a full matrix whose multiplication requires an increasing amount of work up to N^2 operations. However, the number of operations can be substantially reduced, due to the circulant form of the matrices $\underset{\sim}{D}_{2s}$ which enables their efficient diagonalization by a block Fourier matrix $\mathbf{F}$. To be more specific, let us denote by n the integral part of $N/2$,

$$(1.4\text{a})\qquad n\equiv\text{integral part of }\frac{N}{2};$$

then, the above-mentioned block Fourier matrix, $\mathbf{F}$, consists of m-dimensional block entries given by

$$(1.4\text{b})\qquad [\mathbf{F}]_{jk}=\frac{1}{N}\cdot e^{-ijkh}\cdot I_m,\qquad -n\leqq j,k\leqq N-n-1.^{2}$$

In the Appendix, (A.3), we verify that circulant matrices such as $\underset{\sim}{D}_{2s}$ admit the following *spectral representation* in terms of a block *diagonal* matrix $\underset{\sim}{\Lambda}_{2s}\equiv\underset{\sim}{\Lambda}_{2s}(h)$,

$$(1.3_{2s})\qquad \begin{aligned} &\underset{\sim}{D}_{2s}=\mathbf{F}^{-1}\underset{\sim}{\Lambda}_{2s}\mathbf{F},\qquad \mathbf{F}^{-1}\equiv N\mathbf{F}^*,\\ &[\underset{\sim}{\Lambda}_{2s}]_{jj}=\lambda_{2s}^{(j)}\cdot I_m\equiv\frac{2i}{h}\cdot\sum_{k=1}^{s}k^{-1}\beta_k\sin(jkh)\cdot I_m,\qquad -n\leqq j\leqq N-n-1.\end{aligned}$$

[2] Alternatively, one may consider $[\mathbf{F}]_{jk}=1/N\cdot e^{-i(j-n)(k-n)h}\cdot I_m$, $0\leqq j$, $k\leqq N$. However, in order to simplify the notation later on, we prefer the (j,k) entries of $\mathbf{F}$ to lie within the range $-n\leqq j$, $k\leqq N-n-1$.

Multiplication by $\underset{\sim}{D}_{2s}$ in its spectral representation (1.3_{2s}) can now be efficiently implemented by two FFTs and N scalar multiplications which amount to $8N\log N$ operations. Thus, for high $2s$-order accurate differencing, we have regained the (almost) linear rather than quadratic dependence on N.

Next, we extend our discussion by considering rather general discrete differentiation operators. Motivated by the centered differencing examples above, we make one assumption regarding the corresponding matrix representation of such operators, $\underset{\sim}{D} \equiv \underset{\sim}{D}(h)$. Specifically, we assume that the differentiation matrices have the following antisymmetric block circulant form:

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

The antisymmetric form is related to centeral differencing, while the circulant form reflects the assumed periodicity. We require both in order to enable a unified framework for our later discussion on the pseudospectral and Galerkin methods. The circulant matrix $\underset{\sim}{D}$ in (1.5) admits the following spectral representation (see (A.3))

$$(1.6) \qquad \underset{\sim}{D} = \mathbf{F}^{-1}\underset{\sim}{\Lambda}\mathbf{F}, \qquad \mathbf{F}^{-1} \equiv N\mathbf{F}^*$$

with a block diagonal matrix, $\underset{\sim}{\Lambda}$, whose diagonal consists of the so-called *Fourier symbols*

$$(1.7\text{a}) \qquad [\underset{\sim}{\Lambda}]_{jj} = \lambda^{(j)} \cdot I_m \equiv \sum_{k=0}^{N-1} d_k e^{ijkh} \cdot I_m, \qquad -n \leqq j \leqq N-n-1.$$

Since $\underset{\sim}{D}$ is assumed to be antisymmetric, $d_k + d_{N-k} = 0$, and hence the Fourier symbols can be rewritten in the simpler form

$$(1.7\text{b}) \qquad \lambda^{(j)} = 2i \cdot \sum_{k=1}^{n} d_k \sin(jkh), \qquad -n \leqq j \leqq N-n-1.$$

The discrete differentiation described by the spectral representation of $\underset{\sim}{D}$, $\underset{\sim}{D} = \mathbf{F}^{-1}\underset{\sim}{\Lambda}\mathbf{F}$, can be interpreted now as follows. Using the gridvalues $v_{\nu|0 \leqq \nu \leqq N-1}$, one forms the discrete Fourier modes $\hat{v}_\omega$ given by

$$(1.8) \qquad [\hat{v}]_\omega \equiv [\mathbf{F}v]_\omega = \frac{1}{N} \cdot \sum_{\nu=0}^{N-1} e^{-i\omega\nu h} v_\nu, \qquad -n \leqq \omega \leqq N-n-1.$$

Then, each one of these modes is differentiated as it is being multiplied by the Fourier symbol $\lambda^{(\omega)}$. Finally, the differentiated modes

$$\lambda^{(\omega)}\hat{v}_{\omega|-n \leqq \omega \leqq N-n-1}$$

are transformed back into the physical gridspace upon multiplication by $\mathbf{F}^{-1}$.

2. Stability of finite difference approximations. Replacement of the spatial derivative in the symmetric hyperbolic system (0.1) by the differentiation matrix $D \equiv D(h)$, results in the semidiscrete approximation

$$(2.1\text{a}) \qquad \frac{\partial}{\partial t} v_\nu(t) = A(x_\nu)D(h)[v_\nu(t)] + B(x_\nu)v_\nu(t), \qquad \nu = 0, 1, \cdots, N-1.$$

Let us introduce the block diagonal matrices $\underset{\sim}{A} = \operatorname{diag}[A(x_0), \cdots, A(x_{N-1})]$ and $\underset{\sim}{B} = \operatorname{diag}[B(x_0), \cdots, B(x_{N-1})]$; then we can rewrite the approximation (2.1a) in a concise matrix formulation

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

We note that the symmetric hyperbolicity of the system (0.1) is reflected by the coefficient matrix $\underset{\sim}{A}$ being symmetric.

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

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

The approximation is said to be *accurate of order* α if $\|\underset{\sim}{\tau}(h)\| = O[h^{\alpha}]$. For example, with $\underset{\sim}{D} = \underset{\sim}{D}_{2s}$ one obtains a difference approximation which is accurate of order $2s$, $\|\underset{\sim}{\tau}_{2s}(h)\| = O[h^{2s}]$. In order to link the local *order* of accuracy with the desired global convergence *rate* of the approximation, one has to verify stability. We say that approximation (2.1) is *stable*, if for all sufficiently small h the following estimate holds

$$\| \exp [(\underset{\sim}{A}\underset{\sim}{D}+\underset{\sim}{B})t]\| \leqq K \equiv K_T, \qquad 0 \leqq t \leqq T. \tag{2.3}$$

Observe that the stability definition takes into account the lower order term, $\underset{\sim}{B}\underset{\sim}{v}$, appearing in (2.1b). We claim, however, 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. This perturbation lemma, whose proof is given here for completeness, will play an essential role in our discussion (see, e.g., [34, §3.9], [35], [40]).

Perturbation Lemma. *Let* $\mathbf{A}$ *be a given linear operator whose exponent is bounded*:

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

Then, after adding a "low-order" bounded perturbation, $\mathbf{B}$, *we still have a bounded exponent, that is,*

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

Proof. The solution of the inhomogeneous linear differential equation

$$\frac{\partial}{\partial t} w(t)=Lw(t)+G(t), \qquad w(t=0)=w(0) \tag{2.4a}$$

is given by

$$w(t)=e^{\mathbf{L}t}w(0)+\int_{\xi=0}^{t} e^{\mathbf{L}(t-\xi)}G(\xi)\, d\xi. \tag{2.4b}$$

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

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

Hence, (2.4a) can also be written as the *inhomogeneous* problem $w_t(t) = \mathbf{A}w(t) + \mathbf{B}e^{[(\mathbf{A}+\mathbf{B})t]}w(0)$. 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 *arbitrary* 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

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

Using the Gronwall inequality, we conclude that

$$\int_{\xi=0}^{t} K(\xi)\,d\xi \leqq \|\mathbf{B}\|^{-1}\cdot[e^{K_T\cdot\|\mathbf{B}\|\cdot t}-1],$$

and hence

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

as asserted. We remark that similar arguments apply for the analogous question concerning low-order perturbations of power-bounded operators, e.g., [34, §3.9].

Making use of the perturbation lemma, we can now show that stability is equivalent to the boundedness of $\|\exp[\utilde{A}\utilde{D}t]\|$. In other words, if we consider a discrete scheme with or, in particular, without its low-order terms, then the solution of such a stable scheme, $\utilde{v}(t)$, depends continuously on the initial data, $\utilde{v}(0)$,

$$\|\utilde{v}(t)\| = \|e^{[\utilde{A}\utilde{D}t]}\utilde{v}(0)\| \leqq K(t)\cdot\|\utilde{v}(0)\|. \tag{2.5}$$

Indeed, if stability holds in the sense that $\|\exp[(\utilde{A}\utilde{D}+\utilde{B})t]\|$ is bounded, then by the perturbation lemma with $\mathbf{A} = \utilde{A}\utilde{D}+\utilde{B}$ and $\mathbf{B} = -\utilde{B}$, $\|\exp[\utilde{A}\utilde{D}t]\|$ is also bounded. On the other hand, if the exponent $\exp[\utilde{A}\utilde{D}t]$ is bounded as in (2.5), then by the perturbation lemma with $\mathbf{A} = \utilde{A}\utilde{D}$ and $\mathbf{B} = \utilde{B}$, so is the exponent $\exp[(\utilde{A}\utilde{D}+\utilde{B})t]$.

Given stability, we can now estimate the *global error* $\utilde{E}(t) \equiv \utilde{u}(t) - \utilde{v}(t)$. To accomplish this, we subtract (2.1) from (2.2) to find that the error $\utilde{E}(t)$ is governed by the error equation

$$\frac{\partial}{\partial t}\utilde{E}(t) = (\utilde{A}\utilde{D}+\utilde{B})\utilde{E}(t) + \utilde{\tau}(h;t).$$

The solution of the error equation is given by

$$\utilde{E}(t) = \exp[(\utilde{A}\utilde{D}+\utilde{B})t]\utilde{E}(t=0) + \int_{\xi=0}^{t}\exp[(\utilde{A}\utilde{D}+\utilde{B})(t-\xi)]\utilde{\tau}(h;\xi)\,d\xi,$$

and by using the perturbation lemma we end up with the error estimate

$$\|\utilde{E}(t)\| \leqq K(t)\cdot\|\utilde{E}(t=0)\| + \sup_{0\leqq\xi\leqq t}\|\utilde{\tau}(h;\xi)\|\cdot\int_{\xi=0}^{t} K(\xi)\,d\xi.$$

Thus, if an α-order accurate approximation starts with an α-order accurate initial data, i.e., both $\|\tau(h;\xi)\|$ and $\|\utilde{E}(t=0)\|$ are of order $O[h^\alpha]$, then stability will retain the α-order of convergence rate later on, $\|\utilde{E}(t)\| = O[h^\alpha]$. Consequently, verifying stability becomes our main objective in the rest of this section.

We start by noting the identity

$$\utilde{A}\utilde{D}t \equiv \tfrac{1}{2}(\utilde{A}\utilde{D}+\utilde{D}\utilde{A})t + \tfrac{1}{2}(\utilde{A}\utilde{D}-\utilde{D}\utilde{A})t.$$

We recall that $\underset{\sim}{A}$ is a symmetric matrix while $\underset{\sim}{D}$ is an antisymmetric one. It follows that the first parenthesis on the right of the above identity is antisymmetric, which therefore has a bounded exponent; more precisely, we have

$$\|\exp[\tfrac{1}{2}(\underset{\sim}{A}\underset{\sim}{D}+\underset{\sim}{D}\underset{\sim}{A})t]\| = 1.$$

We refer to the above identity once more, this time with the second parenthesis on its right viewed as a low-order perturbation of the first one; in order to estimate the exponent of their sum we invoke the perturbation lemma which yields the following bound:

$$\text{(2.6)} \qquad \|\exp[\underset{\sim}{A}\underset{\sim}{D}t]\| \leqq \exp[\|\tfrac{1}{2}(\underset{\sim}{A}\underset{\sim}{D}-\underset{\sim}{D}\underset{\sim}{A})\| \cdot T], \qquad 0 \leqq t \leqq T.$$

Thus, we are left with the task of finding a bound for the symmetric part of $\underset{\sim}{A}\underset{\sim}{D}$, $\operatorname{Re}(\underset{\sim}{A}\underset{\sim}{D}) \equiv \frac{1}{2}(\underset{\sim}{A}\underset{\sim}{D} - \underset{\sim}{D}\underset{\sim}{A})$. The (p, q) block entry of that part is given by

$$[\tfrac{1}{2}(\underset{\sim}{A}\underset{\sim}{D}-\underset{\sim}{D}\underset{\sim}{A})]_{pq} = \tfrac{1}{2}d_{[p-q]} \cdot [A(x_p)-A(x_q)], \qquad 0 \leqq p, q \leqq N-1.$$

Since $A(x)$ is a symmetric 2π-periodic matrix, we have,

$$\|A(x_p)-A(x_q)\| \leqq h \cdot \operatorname*{Max}_{0\leqq x\leqq 2\pi} \|A'(x)\| \cdot \operatorname{Min}[|p-q|, N-|p-q|].$$

Hence, the matrix in question, $\frac{1}{2}(\underset{\sim}{A}\underset{\sim}{D} - \underset{\sim}{D}\underset{\sim}{A})$, is dominated entrywise, and therefore in norm, by the matrix whose (p, q) block entry is given by

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

This latter matrix is a circulant one. Corollary (A.8) in the Appendix tells us that the norm of such a circulant matrix does not exceed the absolute value sum of the elements along the first $(p = 0)$ row,

$$\frac{h}{2} \cdot \operatorname{Max} \|A'(x)\| \cdot \sum_{q=0}^{N-1} \operatorname{Min}[q, N-q] \cdot |d_q|.$$

Recalling the antisymmetry of $\underset{\sim}{D}$, we have $d_k + d_{N-k} = 0$, and we finally end up with the desired bound

$$\text{(2.7)} \qquad \left\| \frac{1}{2}(\underset{\sim}{A}\underset{\sim}{D}-\underset{\sim}{D}\underset{\sim}{A}) \right\| \leqq h \cdot \sum_{k=1}^{n} k|d_k| \cdot \operatorname*{Max}_{0\leqq x\leqq 2\pi} \|A'(x)\|.$$

Inserting the last bound back into (2.6), we find that the following estimate holds:

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

The above estimate serves as a discrete analogue to the standard energy estimate one has in the differential case. An abstract version of the latter amounts to

$$\text{(2.9)} \qquad \begin{aligned} \|\exp([A(x)D_x t]\| &\leqq \exp \|\tfrac{1}{2}(A(x)D_x - D_x A(x))\| \cdot T) \\ &\leqq \exp(\tfrac{1}{2} \cdot \operatorname{Max} \|A'(x)\| \cdot T), \qquad 0 \leqq t \leqq T. \end{aligned}$$

Let us consider the case in which the coefficient matrix $A(x)$ is not a constant, i.e., $A'(x) \neq 0$. Then the two estimates, (2.8) and (2.9), differ—the term $h \cdot \sum_{k=1}^{n} k|d_k|$ appears only in the first of them. Therefore, in order to guarantee stability, we require this term to be bounded uniformly in h:

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

The essence of the (L) condition is that the differencing operator $\underset{\sim}{D}$ should be a *local* operator, thus reflecting the local nature of differentiation D_x. To summarize our conclusion, we have shown the stability of discrete approximations which employ rather general differencing operators. These are differencing operators which, like differentiation itself, are antisymmetric, periodic (= circulant form) and local. Examples of such stable differencing operators are provided by the centered divided differences D_{2s}, with any *fixed* value of s. These fixed, "low-order" accurate operators are clearly local, as they employ information extracted from a *fixed* number of neighboring gridvalues. The locality of these operators is also reflected in their matrix representation, $\underset{\sim}{D}_{2s}$, which has a finite width of s nonzero super diagonals. Thus, defining the *bandwidth* of the general circulant matrix $\underset{\sim}{D}$ in (1.5) as

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

yields $w(\underset{\sim}{D} = \underset{\sim}{D}_{2s}) = s$. Since in general, the differencing coefficients d_k do not exceed a constant times h^{-1}, it follows that finite-width operators are indeed local, i.e.,

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

This, in turn, implies that the corresponding local schemes are stable.

As s increases, however, $\underset{\sim}{D}_{2s}$ becomes a full matrix which fails to satisfy the locality condition (L). That is not to say that such highly accurate differencing results in unstable approximations, since the locality condition was only shown to be sufficient for stability. To the best of our knowledge, the necessity for such a condition is not known. One case in which stability can be verified independently of such locality restriction is when the coefficient matrix $A(x)$ is positive (or negative) definite. Indeed, multiplication by the definite matrix $A^{-1}(x)$ reduces our problem to that of the constant coefficient case where $\text{Max} \|A'(x)\| = 0$, and the estimate (2.8) yields a uniform bound of one, $\|\exp[\underset{\sim}{A}\underset{\sim}{D}t]\| \leqq 1$. Yet, regarding the general *variable*-coefficients problem, the above stability proof fails whenever the locality restriction is violated. To overcome the difficulty of controlling the high modes, which may arise with "nonlocal" methods, two standard types of remedies are usually employed. These are skew-symmetric differencing and the introduction of dissipation via appropriate smoothing. These topics will be discussed in the following section.

3. Skew-symmetric differencing and smoothing procedures. The spatial part of the differential system (0.1) is skew-symmetric apart from low order terms. This follows from the identity

$$A(x)D_x + B(x) \equiv \tfrac{1}{2}[A(x)D_x + D_xA(x)] + [B(x) - \tfrac{1}{2}A'(x)].$$

We begin by considering skew-symmetric differencing which is based on this formalism. It can be implemented for linear problems [20] as well as for a wide class of nonlinear ones [37].

Using the above identity, we can rewrite (0.1) in the equivalent form

$$\frac{\partial}{\partial t}u(x,t) = \left\{\frac{1}{2}\left[A(x)\frac{\partial}{\partial x}u(x,t) + \frac{\partial}{\partial x}(A(x)u(x,t))\right] + \left[B(x) - \frac{1}{2}A'(x)\right]\right\}u(x,t).$$

Replacing the spatial derivatives on the right by their discrete counterpart, we end up with the *skew-symmetric* approximation

$$\frac{\partial}{\partial t}\underset{\sim}{v}(t)=\left\{\frac{1}{2}[\underset{\sim}{A}\underset{\sim}{D}+\underset{\sim}{D}\underset{\sim}{A}]+\left[\underset{\sim}{B}-\frac{1}{2}\underset{\sim}{A}'\right]\right\}\underset{\sim}{v}(t).$$

The stability of the skew-symmetric approximation is immediate. The first term inside the curly brackets above is antisymmetric, and hence its exponent is bounded by 1. Therefore, according to the perturbation lemma, the exponent of the sum of both terms inside these curly brackets is bounded by the exponent of the second term:

$$\|\exp[\{\tfrac{1}{2}[\underset{\sim}{A}\underset{\sim}{D}+\underset{\sim}{D}\underset{\sim}{A}]+[\underset{\sim}{B}-\tfrac{1}{2}\underset{\sim}{A}']\}t]\|\leqq\exp(\|\underset{\sim}{B}-\tfrac{1}{2}\underset{\sim}{A}'\|\cdot T),\qquad 0\leqq t\leqq T.$$

This coincides with the exact energy estimate for the differential problem; compare with (2.9) in the special case $B=0$. Thus, the stability of skew-symmetric differencing is gained by retaining essential antisymmetry of the *whole* spatial operator, $A(x)D_x+B(x)$, rather than that of the differentiation itself. This is done, however, at the expense of doubling the total amount of work required.

A less expensive alternative to skew-symmetric differencing, which is also proved to be stable, is to apply appropriate smoothing procedures. This topic is discussed in the rest of this section. We start by going back to estimate (2.6) where we were left with the task of bounding the symmetric part of $\underset{\sim}{A}\underset{\sim}{D}$, $\operatorname{Re}(\underset{\sim}{A}\underset{\sim}{D})\equiv\frac{1}{2}(\underset{\sim}{A}\underset{\sim}{D}-\underset{\sim}{D}\underset{\sim}{A})$. By employing the spectral representation of $\underset{\sim}{D}$ (see (1.6)),

$$\underset{\sim}{D}=(N^{1/2}\mathbf{F})^*\underset{\sim}{\Lambda}(N^{1/2}\mathbf{F}),$$

we obtain the equality

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

Let us multiply this equality by $N^{1/2}\mathbf{F}$ on the left and by $(N^{1/2}\mathbf{F})^*$ on the right. By (A.6), this implies that the above matrix is unitarily similar and therefore equal in norm to

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

Next, we examine the matrix inside the curly brackets on the right of (3.2), whose (p,q) block entry is given by

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

Using the Fourier expansion

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

we can also express the (p,q) block entry given in (3.3) as

$$\begin{aligned}\frac{1}{N}\cdot\sum_{\nu=0}^{N-1}A(x_\nu)e^{i(q-p)\nu h}&=\frac{1}{N}\cdot\sum_{\nu=0}^{N-1}\left(\sum_{\omega=-\infty}^{\infty}\hat{A}(\omega)e^{i\omega x_\nu}\right)e^{i(q-p)\nu h}\\&=\sum_{\omega=-\infty}^{\infty}\hat{A}(\omega)\cdot\frac{1}{N}\cdot\sum_{\nu=0}^{N-1}e^{i(q-p+\omega)\nu h}=\sum_{j=-\infty}^{\infty}\hat{A}(p-q+jN).\end{aligned} \tag{3.4}$$

We now combine the representation of $\{(N^{1/2}\mathbf{F})\underset{\sim}{A}(N^{1/2}\mathbf{F})^*\}$ in (3.4) together with the diagonal structure of $\underset{\sim}{\Lambda}$ in (1.7). We then conclude, on account of (3.2), that the matrix norm of $\frac{1}{2}(\underset{\sim}{A}\underset{\sim}{D}-\underset{\sim}{D}\underset{\sim}{A})$ equals that of another matrix whose (p,q) block entry is given by

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

We note that the sufficiency of the locality condition (L) can be deduced again at this stage, if we are to proceed as follows: according to (1.7b) we have

$$(3.5b)\qquad \lambda^{(q)}-\lambda^{(p)}=2i\cdot\sum_{k=1}^{n}d_k\cdot(\sin(qkh)-\sin(pkh));$$

since

$$|\sin(qkh)-\sin(pkh)|\leqq k\cdot h\cdot \operatorname{Min}[|p-q|,N-|p-q|],$$

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

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

Consider the norm of the circulant matrix inside the above curly brackets. By Corollary (A.8) it does not exceed the absolute value sum of its elements along the first row, which in turn can be estimated in terms of the derivatives of $A(x)$. Thus, assuming that the locality condition holds, $h\cdot\sum_{k=1}^{n}k|d_k|\leqq$ Const, we conclude that $\frac{1}{2}(\underset{\sim}{A}\underset{\sim}{D}-\underset{\sim}{D}\underset{\sim}{A})$, and hence exp $[\underset{\sim}{A}\underset{\sim}{D}t]$, have bounded norms.

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

We write the matrix in (3.5) as the sum of two matrices. The first matrix singles out the index $j=0$ in the summation on the right-hand side of (3.5a),

$$(3.6a)\qquad -\left(\frac{\lambda^{(q)}-\lambda^{(p)}}{q-p}\right)\cdot(p-q)\cdot\hat{A}(p-q).$$

In the second matrix we include the rest of the j-indices,

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

It is a property of the finite difference methods that the first matrix in (3.6a) is bounded. Indeed, since the Fourier symbol, $\lambda^{(j)}=2i\cdot\sum d_k\sin(jkh)$ represents the discrete derivative of j mode, e^{ijx}, the difference $\lambda^{(q)}-\lambda^{(p)}$ should not exceed a constant times $|q-p|$. Hence, the matrix in (3.6a) is dominated entrywise and therefore in norm by the matrix whose (p,q) element is given by

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

According to Corollary (A.11), the norm of the above Toeplitz matrix does not exceed a constant times $\sum_{\omega=0}^{N-1}|\omega|\,\|\hat{A}(\omega)\|$, which in turn can be bounded by the derivatives of $A(x)$. Thus, it remains to verify the boundedness of the second matrix given in (3.6b). Here we observe that if $p-q$ is bounded away from jN, $j\neq 0$, e.g., $|p-q|\leqq\theta N$, $\theta<1$, we have $\|\sum_{j\neq 0}\hat{A}(p-q+jN)\|\leqq C_{\gamma,\theta}N^{-\gamma}$, and hence the corresponding (p,q) entries in (3.6b) are negligibly small. To guarantee the boundedness of the entries which correspond to the remaining indices where $|p-q|\sim N$, i.e., where $\pm p\sim\mp q\sim n$, we require $\lambda^{(q)}$ and $\lambda^{(p)}$ to be bounded. Thus the locality condition amounts to *the boundedness of the Fourier symbols, $\lambda^{(j)} \cdot I_m$, associated with the high frequencies $|j| \sim n$.* If this is the case, then 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 considerations are typical for a wide class of time-dependent discrete methods, whose accuracy is determined by the *exactness* of differentiating the low modes, $\lambda^{(j)} \sim ij$, while for their stability we need the *boundedness* of the Fourier symbols, $|\lambda^{(j)}|$, associated with the highest modes, $|j| \sim n$.[3] The combination of the two guarantees convergence, as the low modes carrying most of the information are accurately represented, while stability guarantees that the inaccurate highest modes are not amplified and hence rapidly tend to zero, just as is the case with the differential problem.

The two requirements of accuracy and stability are well accommodated in difference methods having fixed (= N independent) degree of accuracy. Consider, for example, the second-order differencing we started with in (1.1_2). According to (1.3_{2s}) we have $\lambda_2^{(j)} = ih^{-1}\sin(jh)$, and hence, $\lambda_2^{(j)} \sim ij + O(h^2)$ for $|j| \sim 0$, i.e., we identify the second-order accuracy. In addition we have $|\lambda_2^{(j)}| = |h^{-1}\sin(jh)| \leqq \text{Const}$ for $|j| \sim n$, which implies stability. The situation is less favorable, however, for highly accurate differencing methods (of order N or more): here the accuracy requirement for the highest modes, $\lambda^{(j)} \sim ij$, *contradicts* the locality restriction which requires the highest Fourier symbols to be bounded, $|\lambda^{(j)}| \leqq \text{Const}$. We observe that the latter contradiction still leaves us with a bound of order N, which corresponds to a familiar situation of "losing one derivative"[4] in the differential case.

The purpose of smoothing procedures is to resolve the above contradiction by bounding the Fourier symbols associated with the high modes while leaving the lower accurate modes unharmed. Consider, for example, the Shuman filter, where the smoothing transformation

$$v_\nu \rightarrow \tfrac{1}{4}(v_{[\nu+1]} + v_{[\nu-1]} + 2v_\nu)$$

is first applied to the right-hand side of (2.1a). In Fourier space, this amounts to the further multiplication of the j mode by $\frac{1}{2} \cdot (1 + \cos(jh))$. Thus, the following transformation takes place:

$$\hat{v}_j \rightarrow \tfrac{1}{2}(1 + \cos(jh)) \cdot \hat{v}_j.$$

The resulting *smoothed* discrete differentiation operator, $\underset{\sim}{D}_{\text{Shuman}}$, is of the form

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

where

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

In other words, we see that the original Fourier symbols $\lambda^{(j)}$ were replaced by $\lambda^{(j)} \cdot \frac{1}{2}(1 + \cos(jh))$. Consequently, we now achieve the desired boundedness of the highest modes where in fact we have $|\lambda^{(j)} \cdot \frac{1}{2}(1 + \cos(jh))| \sim 0$ for $|j| \sim n$. This is done, however, with the expense that the overall accuracy is now reduced to second order, i.e., $\lambda^{(j)} \cdot \frac{1}{2}(1 + \cos(jh)) \approx ij + O[h^2]$ for $|j| \sim 0$.

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

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

Let us discuss the general case. A linearly smoothed discrete differentiation operator, $\underset{\sim}{D}_*$, takes the form

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

where

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

Both the requirement of accuracy on the one hand and that of stability on the other hand can be formulated in terms of the smoothing coefficients, σ, in the following concise form:

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

Various smoothing procedures of the type described above have been advocated in [1], [15], [22], [24]–[26] and [29]. In these references, smoothing procedures which involve polynomial and exponential cutoff of the highest modes were introduced and were shown to guarantee stability for smooth as well as for nonsmooth data. We now demonstrate that in the smooth case, a first-degree polynomial cutoff will be sufficient for stability, compensating for the loss of *one* derivative mentioned earlier. To work out this case in some detail, we fix $\theta < 1$ and let the smoothing factors, σ, be given by

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

The new Fourier symbols are given now by $\lambda^{(j)}\sigma^{(j)}$. A *fixed* θ-portion of the first N modes remains the same so that the original order of accuracy is retained. To verify stability, we use the real symmetric part $\underset{\sim}{A}\underset{\sim}{D}_*$ in its unitarily equivalent form (3.5). As in (3.6), we consider it to be the sum of two matrices. The (p, q) block entry of the first is given by

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

As we argued before, this first matrix can be bounded by the norm of the derivatives of $A(x)$. We claim that the second matrix, whose (p, q) block entry is given by (compare (3.6b))

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

is likewise bounded. Indeed, for $|p-q| \leqq (1+\theta)N/2$, these entries are negligibly small as they are bounded by $N \cdot \sum_{j\neq 0} \|\hat{A}(p-q+jN)\| \leqq C_{\gamma,\theta} N^{-\gamma+1}$. For the remaining (p, q) indices, where $|p-q| > (1+\theta)N/2$, we either have $p > \theta n$ and $q < -\theta n$, or else the roles of p and q are reversed; in either case we conclude that $|p| > \theta n$ and $|q| > \theta n$. Consequently, the second matrix is bounded entrywise and therefore in norm by the matrix whose (p, q) entry is given by

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

A direct calculation shows that the latter matrix can be bounded in terms of the derivatives of $A(x)$.

Part II. The Pseudospectral Fourier Method.

4. The Fourier differencing operator. 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 $x_\nu = \nu h$, $h = 2\pi/N$, $\nu = 0, 1, \cdots, 2n$. To simplify the notation, we consider here the case of an odd number of gridpoints, $N = 2n + 1$, leaving for the appendix a similar treatment of the even case. By Fourier differentiation we mean differentiation of the trigonometric interpolant based on those gridvalues. That is, one constructs the *trigonometric interpolant*

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

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

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

The Fourier differentiation then takes the form

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

The above procedure consists of the following three basic steps. First, we transform the discrete space of gridvalues $\underset{\sim}{v} \equiv (v_0, \cdots, v_{2n})'$ into the Fourier space of amplitudes $\underset{\sim}{\hat{v}} \equiv (\hat{v}_{-n}, \cdots, \hat{v}_n)'$, or, in matrix notation,

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

Then we differentiate in Fourier space, which amounts to

$$\underset{\sim}{\hat{v}} \longrightarrow \underset{\sim}{\Lambda}_F \underset{\sim}{\hat{v}};$$

here $\underset{\sim}{\Lambda}_F$ denotes the block diagonal matrix which reflects the differentiation carried out in (4.2),

$$\underset{\sim}{\Lambda}_F = \operatorname{diag}\,[-in \cdot I_m, -i(n-1) \cdot I_m, \cdots, i(n-1) \cdot I_m, in \cdot I_m]. \tag{4.4a}$$

Finally, the differentiated amplitudes $\underset{\sim}{\Lambda}_F \underset{\sim}{\hat{v}}$ are transformed back into the discrete "physical" space thus arriving at the Fourier differentiated gridvalues, $\partial_F[\underset{\sim}{v}]$,

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

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

$$\underset{\sim}{F} = N\mathbf{F}^* \underset{\sim}{\Lambda}_F \mathbf{F}, \qquad \mathbf{F}^{-1} = N\mathbf{F}^*. \tag{4.4b}$$

This can be efficiently implemented by two FFT's and N scalar multiplications requiring $O(N \log N)$ operations.

An explicit representation of the Fourier differencing matrix $\underset{\sim}{F}$ can be obtained using the interpolant formula (cf. [42, Chap. X])

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

Differentiation of the right-hand side yields

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

Thus, the Fourier differencing operator belongs to the class of antisymmetric block circulant matrices discussed above in (1.5),

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

Being antisymmetric and circulant, the Fourier differencing matrix admits a spectral representation which can be read from (4.4). Fornberg [9] has shown that the matrix of Fourier symbols in this case, $\underset{\sim}{\Lambda}_F$, equals the increasing limit of the $2s$-order accurate finite-difference symbols studied in (1.3_{2s}),

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

That is, the Fourier differencing can be viewed as a special centered finite differencing based on an ever increasing number of periodic stencils

$$\underset{\sim}{F}=\lim_{s\to\infty}\underset{\sim}{D}_{2s}.$$

With this in mind, we may claim that while the Fourier differencing enjoys an "infinite order of accuracy"—a statement to be made precise below—it is a nonlocal one. Hence our previous discussion, at the end of §2, concerning the problem of stability versus high accuracy, is particularly relevant for the pseudospectral Fourier method. Here we intend to re-examine this problem in terms of the all-important aliasing phenomenon.

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

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

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

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

Given the sampled gridvalues $w(x_\nu)$, $\nu=0,1,\cdots,2n$, of the function $w(x)$, one can construct its interpolant, $\tilde{w}(x)$,

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

in terms of the discrete Fourier coefficients (see (4.1b))

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

Comparing this with the exact Fourier coefficients in (5.1b), we observe that the discrete coefficients are nothing but the trapezoidal rule applied to the integrals on the right of (5.1b). The precise relation between the two, 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 lemma.

ALIASING LEMMA (Poisson's summation formula). *Let $w(x)$ be as above. Then we have*

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

Proof. Inserting (5.1a) into (5.2b), we obtain

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

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

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

Indeed, the second summation 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 difference between the gridfunction $w(x)$ and its equidistant interpolant $\tilde{w}(x)$. We have $w(x) = [\sum_{|\omega| \leqq n} + \sum_{|\omega| > n}] \hat{w}(\omega) e^{i\omega x}$, and, with the help of the aliasing lemma, we rewrite

$$\tilde{w}(x) = \sum_{|\omega| \leqq n} \hat{w}(\omega) e^{i\omega x} + \sum_{|\omega| \leqq n} \left(\sum_{k \neq 0} \hat{w}(\omega + kN) \right) e^{i\omega x}.$$

This shows that the difference $w(x) - \tilde{w}(x)$ can be decomposed as the sum of two contributions. The first contribution, the *truncation error*, consists of the higher truncated modes, $|\omega| > n$,

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

the second one, the *aliasing error*, consists of the higher aliased modes folded back on the lower ones $|\omega| \leqq n$ due to the finite resolution of the grid

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

We observe that while the truncation error involves modes higher than n, the aliasing error involves modes less than or equal to n. Hence, these two kinds of errors are orthogonal to each other and the size of the difference $w(x) - \tilde{w}(x)$ is given by

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

Using Parseval's relation, the two squared terms on the right are found to equal

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

$$\| \text{Aliasing}\,(w) \|^2 = 2\pi \cdot \sum_{|\omega| \leqq n} \left| \sum_{k \neq 0} \hat{w}[\omega + k(2n+1)] \right|^2. \tag{5.5c}$$

In both cases only the high amplitudes, those associated with modes higher than n, participate in the summations on the right. It is well known that for smooth functions these high amplitudes rapidly tend to zero, i.e., integration by parts on the right-hand side of (5.1b) yields

$$| \hat{w}(\omega) | \leqq C_\gamma (1 + |\omega|)^{-\gamma} \quad \text{for any } \gamma > 0.$$

It then follows that the two squared terms appearing on the right of (5.5) have the same size, which is of order $C_\gamma \cdot N^{(-\gamma+1)}$. Likewise, we find that $(d/dx)w(x)$ differs from the differentiated interpolant, $(d/dx)\,\tilde{w}(x)$, by

$$\left\|\frac{d}{dx}w(x)-\frac{d}{dx}\tilde{w}(x)\right\|^2$$

$$=2\pi\cdot\sum_{|\omega|>n}|\omega|^2\cdot|\hat{w}(\omega)|^2+2\pi\cdot\sum_{|\omega|\leqq n}|\omega|^2\cdot\left|\sum_{k\neq 0}\hat{w}[\omega+k(2n+1)]\right|^2,$$

which is of order $C_\gamma N^{(-\gamma+2)}$. As pointed out above, the Fourier differencing of the function $w(x)$ is nothing else but the exact differentiation of its interpolant $\tilde{w}(x)$. We therefore conclude that the error committed by differentiating $\tilde{w}(x)$ rather than $w(x)$ is of the negligibly small order $C_\gamma h^\gamma$ for *any* $\gamma>0$. It is in this sense that we say the Fourier differencing has "infinite order accuracy."

We note in passing that the aliasing relation (5.3) can be used in order to show the isometry between the discrete and continuous space functions. To be more specific, consider the discrete space of 2π-periodic vector functions, $y(x)$, $z(x)$ equipped with the usual Euclidean inner product $(\cdot,\cdot)$,

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

The discrete analogue of this space consists of gridfunctions $\underset{\sim}{y}=(y_0,\cdots,y_{2n})'$, $\underset{\sim}{z}=(z_0,\cdots,z_{2n})'$ equipped with the discrete inner product $\langle\cdot,\cdot\rangle$

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

The above-mentioned isometry now takes the concise form

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

Indeed, if we let $w(x)$ be the 2π-periodic function $2\pi\cdot\tilde{z}^*(x)\,\tilde{y}(x)$, then by definition, the left-hand side of (5.7) equals $\hat{w}(\omega=0)$, while the right-hand side equals $\hat{w}_{\omega=0}$. According to the aliasing lemma, the two terms differ by the sum of aliased modes higher than $2n$, $\sum_{k\neq 0}\hat{w}[k(2n+1)]$. This sum vanishes, however, since $w(x)$ is a trigonometric polynomial of degree $2n$ which contains no modes higher than $2n$.

6. Stability of the Fourier method. In this section we examine the stability of the Fourier method, i.e., when spatial differentiation is carried out by Fourier differencing. According to the perturbation lemma we can neglect the low order term and assume that $B=0$. Hence, our Fourier approximation of (0.1) takes the form

$$(6.1)\qquad \frac{\partial}{\partial t}v_\nu(t)=L\tilde{v}_{|x=x_\nu},\qquad L=A(x)\frac{\partial}{\partial x}.$$

The stability question of the Fourier method can be answered similarly to our previous discussion on finite-difference methods in §3. That is, the unboundedness of the Fourier symbols in (4.4a), $\lambda_F^{(j)}=ij\cdot I_m$, requires smoothing of the highest modes, in agreement with the nonlocality of the method as evidenced from (4.6), $h\cdot\sum_{k=1}^n k|d_k^{(F)}|=O(1/h)$. The stability analysis of the Fourier method outlined below employs a somewhat different point of view; in fact, it is the one that motivated our discussion in §3 above.

To begin with, we multiply (6.1) by hv_ν^* and sum over all gridpoints, to obtain

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

Taking the real part of both sides of the last equality and using the isometry (5.7), we find

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

The next step, which is at the heart of the matter, involves the decomposition of the right-hand side into the sum of two terms. The first term accounts for the differential operator itself,

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

The second term accounts for the deviation due to interpolation of the differential operator,

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

We note that the above decomposition, reflected by the identity

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

is in complete analogy to the previous splitting of the matrix in (3.5a) into (3.6a) and (3.6b).

That the first term (6.3a) is bounded by a constant times $\|\tilde{v}\|^2$ is a property solely of the *differential* operator L, called semiboundedness. This can be easily verified in our case by integration by parts, yielding

$$\text{(6.5)}\qquad |([L+L^*]\tilde{v},\tilde{v})|\leqq \mathrm{Const}\cdot\|\tilde{v}\|^2.$$

The last estimate implies the usual exponential growth bound in complete agreement with the behavior indicated in (2.9). Thus we are left with the task of bounding the second term, the one given in (6.3b). It is exactly this term which measures by how much we deviate from the standard energy estimate whose abstract version was 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; these are 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 $\tilde{v}$. Hence the contribution of the truncation error to the deviation term under consideration, (6.3b), is completely suppressed. In other words, *it is solely the aliasing error which controls the stability of the discrete approximation* (6.1). To see this, we express the amplitudes of $L\tilde{v}$ as the convolution sums

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

By (5.4b) the aliasing error is then given by

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

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

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

or, after taking the real symmetric part of both sides of the last equality we get

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

Our purpose is to estimate the expression on the right of (6.6) in terms of $\|\tilde{v}\|^2$. By doing so we will end up with an energy estimate

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

which in turn will imply stability (see (2.5)):

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

To assert that the right-hand side of (6.6) does not exceed

$$\text{Const}\,\|\tilde{v}\|^2\equiv\text{Const}\cdot\sum_{|\omega|\leqq n}|\hat{v}_\omega|^2$$

for *all* possible amplitudes $\hat{v}_\omega$ is, by definition, equivalent to asserting the boundedness of the matrix given in the curly brackets above (6.6). Denoted by $\hat{\mathbf{A}}$, the (p, q) entry of this matrix equals

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

A similar expression was already obtained in (3.6b). These terms represent the effect of aliasing due to the presence of a variable coefficient matrix $A(x)$. In the constant coefficient case, for example, no aliasing occurs, $\hat{A}(\omega)=0$, $\omega\neq 0$, so that the terms in (6.8) and hence in (6.3b) vanish and stability follows, in agreement with the earlier conclusion of stability for the constant coefficient case. The situation with the variable coefficient case is more delicate, however. We examine the (p, q) entries we were left with in (6.8). For $|p-q|$ bounded away from $2n$, i.e., $|p-q|\leqq\theta\cdot 2n$, $\theta<1$, these entries are negligibly small, since by the smoothness of $A(x)$ we have

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

Yet when $|p-q|$ approaches $2n$, that is, either 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 $O(1)$, and hence these entries are of size $O(N=2n+1)$. In other words, we conclude that *the matrix* $\hat{\mathbf{A}}$ *given in* (6.8) *is unbounded, no matter how smooth* $A(x)$ *is*. For example, consider the case where $A(x)$ consists of only one mode; the only nonzero entries in (6.8) are $(p,q)=(\pm n,\mp n)$, given by $\mp 2n\hat{A}(\omega=\mp 1)$. The unboundedness of $\hat{\mathbf{A}}$ then follows. Another way of expressing this last conclusion is that in contrast to the local finite-difference methods studied in (2.7), here $\mathrm{Re}\,(\underset{\sim}{A}\underset{\sim}{F})=\frac{1}{2}(\underset{\sim}{A}F-\underset{\sim}{F}\underset{\sim}{A})$ is unbounded no matter how smooth $A(x)$ is. Indeed, up to unitary similarity the latter differ from $\hat{\mathbf{A}}$ by the bounded term (6.3a).

Nevertheless, the unboundedness encountered above 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 we try to estimate these (p, q) with $|p|, |q| \sim n$. These entries are multiplied in (6.6) by the amplitudes associated with the corresponding *high* modes $\hat{v}_p^*, \hat{v}_q$. The latter are *expected* to be of a negligible small size *provided* the Fourier method is indeed stable, which is what we are trying to prove. That is, despite the unboundedness of $\hat{\mathbf{A}}$ in (6.8) we can still bound the aliased terms in (6.6), provided a priori information on the decay rate of $|\hat{v}_\omega|$ is at our disposal. It is well known, however, that the L_2-norm $\|\tilde{v}\|^2$ is too weak to provide such a priori information on the decay rate.

With this in mind, smoothing may be viewed as a procedure which provides us the a priori information we seek. For example, consider the case where $A(x)$ consists of *fixed* number, say of r modes. Then smoothing by cutting off a *fixed* number of modes—only the last r ones, $\hat{v}_\omega = 0$, $|\omega| > n - r$—will guarantee stability. Indeed the aliasing term in (6.6) will vanish in this case. The particular case $r = 1$ requires the estimate of only the last amplitude $\hat{v}_n$. Such an estimate exists in the case of *even* number of gridpoints, since $\underset{\sim}{F}$, being an *even* order antisymmetric matrix, has a *double* zero eigenvalue. This then leads to H^1-stability (see the Appendix for details).

In closing this section, we remark that (6.6) can be rewritten in the form

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

The matrix in the last curly brackets is bounded; hence the expression on the right does not exceed a constant times

$$\|\tilde{v}\|^2_{H^{1/2}} \equiv \sum_{|\omega| \leqq n} (1+|\omega|^2)^{1/2} |\hat{v}_\omega|^2.$$

Together with (6.5) we are then led to the final estimate

$$\frac{d}{dt}\|\tilde{v}\|^2 \leqq \mathrm{Const}\, \|\tilde{v}\|^2_{H^{1/2}}. \tag{6.9}$$

That is, there is a loss of "one-half" derivative. If some dissipation is present in the system to begin with, e.g., with $L = A(x)D_x + D_x^2$, the *gain* of one derivative from the second-order spatial differentiation dominates and stability follows; see, e.g., [22].

Part III. The Fourier–Galerkin Method.

7. The Galerkin procedure. In this part we deal with the Fourier–Galerkin method. The literature on the subject is rather extensive; see [2], [6], [8], [11], [18], [27], [28], [32], to mention but a few. In the spirit of earlier remarks, we therefore confine ourselves to a stability study of this method, emphasizing its interplay with the other two methods—the finite-difference and pseudospectral ones.

The essential strategy behind the Galerkin-type methods is to reduce our infinite-dimensional differential equation by projecting the problem into a finite-dimensional subspace. Let the latter be spanned by a basis of 2π-periodic functions $\phi_p(x)$, $-n \leqq p \leqq n$. To solve (0.1),

$$\frac{\partial u}{\partial t} = Lu, \qquad L = A(x)\frac{\partial}{\partial x} + B(x),$$

an $N=2n+1$ degree approximation of the form

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

is sought. The coefficients $\hat{v}(q,t)$, $-n\leqq q\leqq n$, so-called generalized Fourier coefficients, are to be determined by projecting our problem

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

Substituting (7.1) into (7.2), we conclude that the vector of generalized Fourier coefficients, $\hat{\underset{\sim}{v}}(t)\equiv(\hat{v}(-n,t),\cdots,\hat{v}(n,t))'$, satisfies the following system of ordinary differential equations:

$$M\frac{\partial}{\partial t}\hat{\underset{\sim}{v}}(t)=G\hat{\underset{\sim}{v}}(t). \tag{7.3a}$$

Here, M and G are $(2n+1)$-dimensional matrices whose (p,q) block entries are given respectively by

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

The stability of the resulting system (7.3) is a direct consequence of the previously mentioned semiboundedness of the differential operator L, namely, we have, as in (6.5)

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

Indeed, multiplying (7.2$_p$) by $\hat{v}(p,t)$, summing up and taking real parts we find

$$\frac{1}{2}\frac{d}{dt}\|v(t)\|^2=\operatorname{Re}\left(\frac{\partial v}{\partial t},v\right)=\operatorname{Re}(Lv,v)\leqq\text{Const }\|v(t)\|^2,$$

and hence the asserted stability follows (compare (2.9)).

Unless chosen with care, however, the basis functions $\phi_k(x)$ may lead to an ill-conditioned mass matrix, M, whose required inversion in (7.3a) can be found numerically disastrous. To avoid such situations, two types of basis functions are usually employed. The first choice is *local* basis functions which induce sparse, well-behaved mass matrices. This in turn leads to either explicit or implicit, smoothed or unsmoothed finite-difference and finite-element local methods. Here locality is interpreted according to our discussion in §2 above. The second choice is *global, orthogonal* basis functions, where the mass matrix is reduced to the identity $M=I$. A universal example for the latter choice in the periodic case is the trigonometric system

$$\phi_p(x)=e^{ipx}, \qquad -n\leqq p\leqq n.$$

We thus arrive at the Fourier–Galerkin method which will occupy the rest of our discussion.

The expansion we seek in (7.1) now amounts to the usual truncated Fourier expansion. The corresponding Fourier coefficients $\hat{\underset{\sim}{v}}(t)=(\hat{v}(-n,t),\cdots,\hat{v}(n,t))'$, satisfy the ordinary differential equations

$$\frac{\partial}{\partial t}\hat{\underset{\sim}{v}}(t)=G\hat{\underset{\sim}{v}}(t). \tag{7.4a}$$

The coefficient matrix G is given here by

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

where as before (see (6.1b)) we have neglected the lower order term, assuming $L = A(x)D_X$. We note in passing that given the *exact* Fourier coefficients $\hat{A}(\omega)$, $|\omega| \leqq n$, the implementation of the Fourier–Galerkin method can be carried out *fast*, using $O(N \log N)$ operations. Indeed, the method consists of two basic steps: first, differentiation is carried out as a multiplication by the diagonal matrix $\underset{\sim}{\Lambda}_F$, $[\underset{\sim}{\Lambda}_F]_{qq} = iq \cdot I_m$, requiring $N = 2n + 1$ operations, and then multiplication by $A(x)$, reflected as a convolution sum in the Fourier space, follows. The latter can be accomplished by a *fast* multiplication of the *Toeplitz* matrix $\hat{A}(p-q)$. Details can be found in Appendix A (see Corollary (A.10)).

In order to evaluate the Fourier coefficients

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

various quadrature rules can be applied for the approximate evaluation of the integrals on the right. This in turn leads to a whole variety of discrete Fourier–Galerkin methods. In particular, the previously discussed finite-difference and Fourier methods are obtained as special cases.

8. Discretization. The Fourier–Galerkin method in component-wise form reads

$$(8.1\text{a}) \qquad \frac{\partial}{\partial t} \hat{v}(p,t) = \sum_{q=n}^{n} \hat{A}(p-q) \cdot iq \cdot \hat{v}(q,t)$$

where $\hat{A}(\omega)$ are the usual Fourier coefficients

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

To approximate the integrals on the right, we use the trapezoidal rule based on the $N = 2n + 1$ equidistant points $x_\nu = \nu h$, $h = 2\pi/N$,

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

Since $A(x)$ is assumed periodic, the trapezoidal rule serves our purpose as does any other high order quadrature rule—in fact, it is "infinite order accurate" in the precise sense discussed in §5 above (cf. [7, §2.9]).

If we use the trapezoidal discretization (8.2) we find that the terms on the right of (8.1a), $\hat{A}(p-q)$, are to be replaced by the corresponding discrete sums we have already met in (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 corresponding discretization of the Fourier–Galerkin method (8.1a) amounts to a system of ordinary differential equations which, using (8.3) and (4.4a), reads

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

The resulting system is exactly the Fourier method for the *discrete* Fourier amplitudes $\hat{\underset{\sim}{v}}(t) \equiv \mathbf{F}\underset{\sim}{v}(t) = (\hat{v}_{-n}(t), \cdots, \hat{v}_n(t))'$. Indeed, multipication of (8.4) by $\mathbf{F}^{-1}$ on the left brings this system back into its familiar form in the physical space (see (4.4b)),

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

In summary, we have seen that the equidistant discretization based on N gridpoints (8.3) reduces the N-degree Fourier–Galerkin method to the Fourier method. The difference between the two lies exactly in the aliasing term $\sum_{j\neq 0}\hat{A}(p-q+jN)$. Indeed, (3.4) tells us that the latter term is the exact difference between the right- and left-hand sides of (8.3). Since the Fourier–Galerkin method was shown to be stable, we thus shed a different light on our previous conclusion, namely, that stability of the Fourier method is solely determined by aliasing errors. To suppress those aliasing errors, one may smooth the highest modes. Smoothing procedures such as those discussed in §3 above can be interpreted within the framework of the discretized Fourier–Galerkin methods. In particular, the typical *cutting off* of the highest modes corresponds to equidistant discretization based on *more* than N gridpoints. The details are outlined below.

Let $M = (1+\varepsilon)N$ be the number of gridpoints, $x_\nu = \nu h$, $h = 2\pi/M$, $\nu = 0, 1, \cdots, M-1$, used by the trapezoidal rule, for the approximate evaluation of the Fourier integrals on the right of (7.4b),

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

If we substitute this into (8.1a), the resulting system is

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

Here, we adopt the subindex notation for the discrete Fourier coefficients of the computed amplitudes $\hat{v}_\omega(t)$. We note in passing that the matrix whose (p, q) block entry is given inside the right brackets is no longer a circulant matrix, as in the Fourier case where $M = N$. Yet, since it is a Toeplitz matrix, one can carry out a fast multiplication of such a matrix (see Corollary A.10).

To verify stability, we employ (3.4) rewriting the (p, q) entry inside these brackets in the form

$$\frac{\partial}{\partial t}\hat{v}_p(t) = \sum_{q=-n}^{n}\left[\sum_{j=-\infty}^{\infty}\hat{A}(p-q+jM)\right]\cdot iq\cdot\hat{v}_q(t). \tag{8.7b}$$

Similarly to our treatment of the finite-difference methods in (3.6a) and the Fourier method in (6.4), the second summation is decomposed into two contributions:

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

The first term on the right corresponds to the semi-bounded differential operator and can be estimated as before. The second summation term represents the pure effect of aliasing. Thanks to the smoothness of $A(x)$, this term is of a negligibly small size:

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

Indeed, a second look at (8.7b) reveals that the approximate system 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 $(1+\varepsilon)^{-1}M<|j|\leqq M$; 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 is frequently referred to in the discussion above.

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

$$\mathbf{T}\equiv\mathbf{T}(t_{1-N},\cdots,t_0,\cdots,t_{N-1})=\begin{bmatrix} t_0 & t_1 & t_2 & \cdots & t_{N-2} & t_{N-1}\\ t_{-1} & \ddots & \ddots & \ddots & \ddots & t_{N-2}\\ t_{-2} & \ddots & \ddots & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & \ddots & \ddots & t_2\\ t_{2-N} & \ddots & \ddots & \ddots & \ddots & t_1\\ t_{1-N} & t_{2-N} & \cdots & t_{-2} & t_{-1} & t_0\end{bmatrix}.\tag{A.1}$$

Thus, an $N\times N$ Toeplitz matrix is completely determined by a $(2N-1)$-dimensional vector $\underset{\sim}{t}\equiv(t_{1-N},\cdots,t_0,\cdots,t_{N-1})$; in our case the entries t_l, $-(N-1)\leqq l\leqq N-1$ are m-dimensional blocks.

In particular, if the vector $\underset{\sim}{t}$ is defined on its negative indices as the periodic extension of the positive ones, $t_{-l}=t_{N-l}$, $0<l\leqq N-1$, i.e., if $\mathbf{T}_{jk}$ depends on $(k-j)[\operatorname{mod} N]$ then the matrix $\mathbf{T}$ is defined as a block *circulant* one, $\mathbf{T}\equiv\mathbf{C}$, $[\mathbf{C}]_{jk}=c_{(k-j)[\operatorname{mod} N]}$,

$$\mathbf{C}\equiv\mathbf{C}(c_0,\cdots,c_{N-1})=\begin{bmatrix} c_0 & c_1 & c_2 & \cdots & c_{N-2} & c_{n-1}\\ c_{N-1} & \ddots & \ddots & \ddots & \ddots & c_{N-2}\\ c_{N-2} & \ddots & \ddots & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & \ddots & \ddots & c_2\\ c_2 & \ddots & \ddots & \ddots & \ddots & c_1\\ c_1 & c_2 & \cdots & c_{N-2} & c_{N-1} & c_0\end{bmatrix}.\tag{A.2}$$

Thus, a circulant matrix is completely determined by an N-dimensional vector $\underset{\sim}{c}\equiv(c_0,\cdots,c_{N-1})$; its entries, c_l, $0\leqq l\leqq N-1$, are again m-dimensional blocks in our case.

An essential ingredient in studying circulant matrices is their *spectral representation*, given by

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

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

$$[\mathbf{F}]_{jk}=\frac{1}{N}\cdot e^{-ijkh}\cdot I_m,\qquad -n\leqq j,k\leqq N-n-1,\quad n=\text{integral part of }\frac{n}{2},\tag{A.4}$$

where $\underset{\sim}{\Lambda}_c$ is a diagonal matrix whose diagonal block entries are given by the Fourier symbols

$$[\underset{\sim}{\Lambda}_c]_{jj}=\sum_{l=0}^{N-1}e^{ijlh}\cdot c_l,\qquad -n\leqq j\leqq N-n-1.\tag{A.5}$$

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

$$\begin{aligned}\{(N^{1/2}\mathbf{F})^*\underset{\sim}{\Lambda}_c(N^{1/2}\mathbf{F})\}_{jk} &= N\cdot\sum_{p=-n}^{N-n-1}[\mathbf{F}^*]_{jp}[\underset{\sim}{\Lambda}_c]_{pp}[\mathbf{F}]_{pk}\\ &= N\cdot\sum_{p=-n}^{N-n-1}\frac{1}{N}\cdot e^{ipjh}\cdot\left[\sum_{l=0}^{N-1}e^{iplh}\cdot c_l\right]\cdot\frac{1}{N}\cdot e^{-ipkh}\\ &= \frac{1}{N}\cdot\sum_{l=0}^{N-1}c_l\sum_{p=-n}^{N-n-1}e^{ip(l+j-k)h}.\end{aligned}$$

Since the second summation on the right vanishes unless $l+j-k=0[\mathrm{mod}\, N]$, i.e., unless $l=(k-j)[\mathrm{mod}\, N]$, we are finally left with the asserted term

$$\frac{1}{N}\cdot\sum_{l=0}^{N-1}c_l\sum_{p=-n}^{N-n-1}\exp\,(ip(l+j-k)h)_{|l=(k-j)[\mathrm{mod}\, N]}=c_{(k-j)[\mathrm{mod}\, N]}\equiv[\mathbf{C}]_{jk}.$$

If we view the block identity matrix I_N as a circulant one, generated by first row vector $\underset{\sim}{c}=(I_m, 0_m, \cdots, 0_m)$, then (A.3) and (A.5) give us its spectral representation as

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

That is to say that *the matrix* $N^{1/2}\mathbf{F}$ *is a unitary one.* Moreover, the spectral representation in (A.3) is then nothing but a *unitary* diagonalization of the circulant matrix $\mathbf{C}$. Since the spectum and the L_2-norm of a matrix are invariant under such unitary transformations, it follows that for general circulant matrices, $\mathbf{C}$, both are identical with those of the block diagonal matrix $\underset{\sim}{\Lambda}_c$—the Fourier symbol blocks. In particular we have the following lemma.

LEMMA A.7. *For a block circulant matrix* $\mathbf{C}(\underset{\sim}{c})$ *we have*

$$\|\mathbf{C}(\underset{\sim}{c})\|=\operatorname*{Max}_{-n\leqq j\leqq N-n-1}\left\|\sum_{l=0}^{N-1}e^{ijl(2\pi/N)}\cdot c_l\right\|. \tag{A.7}$$

Proof. The norm of a block diagonal matrix $\underset{\sim}{\Lambda}_c$ is given by the largest norm of its diagonal entries.

As an immediate corollary we have the following.

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

$$\|\mathbf{C}(\underset{\sim}{c})\|\leqq\sum_{l=0}^{N-1}\|c_l\|. \tag{A.8}$$

The corollary is just a restatement of that last inequality for the scalar case where $c_l=\mathrm{Const}_l\cdot I_m$.

Next, we employ the information just obtained for circulant matrices, and apply it to Toeplitz matrices with the help of the following lemma.

LEMMA A.9. *Any N-dimensional block Toeplitz matrix can be imbedded into a 2N-dimensional block circulant one.*

Proof. Consider the block Toeplitz matrix $\mathbf{T}=\mathbf{T}(\underset{\sim}{t})$ with $\underset{\sim}{t}=(t_{1-N}, \cdots, t_0, \cdots, t_{N-1})$. Denote $\underset{\sim}{t}^-=(t_{1-N}, \cdots, t_{-1})$, $\underset{\sim}{t}^+=(t_1, \cdots, t_{N-1})$ and define the associated

Toeplitz matrix $\mathbf{R} \equiv \mathbf{R_T} = T(\underline{t}^+, s, \underline{t}^-)$ with s being an arbitrary m-dimensional block. It is readily verified that

$$\text{(A.9a)} \qquad \mathbf{C} = \begin{pmatrix} \mathbf{T} & \mathbf{R_T} \\ \mathbf{R_T} & \mathbf{T} \end{pmatrix}$$

is a $2N$-dimensional block circulant

$$\text{(A.9b)} \qquad \mathbf{C} = \mathbf{C}(\underset{\sim}{c}), \qquad \underset{\sim}{c} = (t_0, \underline{t}^+, s, \underline{t}^-).$$

In entrywise form we have

$$\text{'A.9c)} \quad \mathbf{C} = \begin{bmatrix} t_0 & t_1 & \cdots & t_{N-2} & t_{N-1} & s & t_{1-N} & \cdots & t_{-2} & t_{-1} \\ t_{-1} & \ddots & & & t_{N-2} & t_{N-1} & \ddots & & & t_{-2} \\ \vdots & & \ddots & & \vdots & \vdots & & \ddots & & \vdots \\ t_{2-N} & & & \ddots & t_1 & & & & \ddots & t_{1-N} \\ t_{1-N} & t_{2-N} & \cdots & t_{-1} & t_0 & t_1 & & \cdots & t_{N-1} & s \\ s & t_{1-N} & \cdots & t_{-2} & t_{-1} & t_0 & t_1 & \cdots & t_{N-2} & t_{N-1} \\ t_{N-1} & \ddots & & & t_{-2} & t_{-1} & \ddots & & & t_{N-2} \\ \vdots & & \ddots & & \vdots & \vdots & & \ddots & & \vdots \\ t_2 & & & \ddots & t_{1-N} & t_{2-N} & & & \ddots & t_1 \\ t_1 & t_2 & \cdots & t_{N-1} & s & t_{1-N} & t_{2-N} & \cdots & t_{-1} & t_0 \end{bmatrix}.$$

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

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

COROLLARY A.10. *Multiplication by an N-dimensional block Toeplitz matrix can e implemented "fast," i.e., using $O(N \log N)$ block operations.*

Proof. Given an N-dimensional vector $\underset{\sim}{w}$, we want to compute $\underset{\sim}{z} = \mathbf{T}\underset{\sim}{w}$, where **T** s an N-dimensional Toeplitz matrix. To this end we imbed **T** into

$$\mathbf{C} = \begin{pmatrix} \mathbf{T} & \mathbf{R_T} \\ \mathbf{R_T} & \mathbf{T} \end{pmatrix}$$

nd compute $\underset{\sim}{z}_* = \mathbf{C}\underset{\sim}{w}_*$, $\underset{\sim}{w}_* = (\underset{\sim}{w}, \underset{\sim}{0}_N)'$. Since **C** is a circulant matrix, the last multilication can be efficiently implemented using the spectral representation (A.3) with wo FFT's requiring $O(N \log N)$ operations. The first N components of $\underset{\sim}{z}_*$ are then he desired vector $\underset{\sim}{z}$.

COROLLARY A.11. *For a block Toeplitz matrix $T(\underline{t})$ we have*

$$\text{A.10)} \qquad \|T(\underline{t})\| \leqq \operatorname*{Max}_{-N \leqq j \leqq N-1} \left\| \sum_{l=-(N-1)}^{N-1} e^{ijl(\pi/N)} \cdot t_l \right\|.$$

Proof. As before, we imbed $\mathbf{T}(\underline{t})$ into $\mathbf{C}(\underset{\sim}{c})$ with $\underset{\sim}{c} = (t_0, \underline{t}^+, 0, \underline{t}^-)$. Lemma A.7 tells s that

$$\|\mathbf{T}(\underline{t})\| \leqq \|\mathbf{C}(\underset{\sim}{c})\| = \operatorname*{Max}_{-N \leqq j \leqq N-1} \left\| \sum_{l=0}^{2N-1} e^{ijl(2\pi/2N)} \cdot c_l \right\|.$$

Inserting the values of the blocks c_l expressed in terms of the corresponding Toeplitz ones, t_l, shows that the expression on the right equals the asserted value

$$\underset{-N \leqq j \leqq N-1}{\mathrm{Max}} \left\| \sum_{l=0}^{N-1} e^{ijl(\pi/N)} \cdot t_l + (-1)^l \cdot 0 + \sum_{l=N+1}^{2N-1} e^{ijl(\pi/N)} \cdot t_{l-2N} \right\|$$

$$\equiv \underset{-N \leqq j \leqq N-1}{\mathrm{Max}} \left\| \sum_{l=-(N-1)}^{N-1} e^{ijl(\pi/N)} \cdot t_l \right\| .$$

Remark. Making use of the freedom in choosing an arbitrary block s along the main diagonal of the associated Toeplitz $\mathbf{R}_{\mathrm{T}}$, we similarly get

$$\|\mathrm{T}(\underset{\sim}{t})\| \leqq \inf_s \left[\underset{-N \leqq j \leqq N-1}{\mathrm{Max}} \left\| (-1)^j \cdot s + \sum_{l=-(N-1)}^{N-1} e^{ijl(\pi/N)} \cdot t_l \right\| \right] .$$

In Corollary A.11 the choice $s = 0$ was made. Corresponding to Corollary A.8 we now have the following corollary.

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.*

Appendix B. The Fourier method—the case of an even number of gridpoints. The Fourier method is usually implemented with trigonometric interpolants based on an even number of gridpoints, $N = 2n$. More precisely, when N is an integral power of two, then the Cooley–Tukey variants of the FFT are optimal. Here we record the slightly different formulas governing this case.

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

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

expressed in terms of the discrete Fourier coefficients $\hat{v}_\omega$

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

The double prime summation indicates the usual halving of the first and last terms.

An explicit representation of the Fourier differencing matrix $\underset{\sim}{F}$, transforming $\underset{\sim}{v} \equiv (v_0, \cdots, v_{2n-1})'$ into its differenced values $\partial_F[\underset{\sim}{v}] \equiv (D_x\tilde{v}|_{x_0}, \cdots, D_x\tilde{v}|x_{2n-1})'$, can be obtained by differentiating the interpolant formula, e.g., [42, Chap. X]

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

A straightforward calculation yields, e.g., [13]:

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

Being a block circulant matrix, $\underset{\sim}{F}$ admits the spectral representation

$$\text{(B.4a)} \qquad \underset{\sim}{F} = N\mathbf{F}^* \underset{\sim}{\Lambda}_F \mathbf{F}$$

whose Fourier symbols are given by

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

Observe that zero is a *double* eigenvalue in this case. This is necessarily so since $\underset{\sim}{F}$ being an antisymmetric *even* dimensional matrix must have the other, complex eigenvalues, in pairs. The left eigenvectors corresponding to the double zero eigenvalue are

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

and

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

The equalities (B.5) reflect the exactness of the differentiation in (B.3) for $\tilde{v}(x)=\text{Const}$ and $\tilde{v}(x)=\cos(nx)$ (compare [13, Lemma 1.1]).

The Fourier method (6.1) now reads

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

Stability analysis in this case is similar to that dealt with in §7 for the case of an odd number of gridpoints. That is, to estimate the real symmetric part of $(\widetilde{L\tilde{v}},\tilde{v})$ (6.2), we shall use the aliasing formula which still has the same format as in (5.3)

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

This, in turn, leaves us with the task of bounding an aliasing term similar to that encountered before in (6.6)

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

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

$$\frac{\partial}{\partial t}\underset{\sim}{w}(t)=\underset{\sim}{F}\underset{\sim}{A}\underset{\sim}{w}(t);$$

if we then multiply by $(\underset{\sim}{z}^{(n)})'$ on the left and use (B.5b) we conclude that

$$(\underset{\sim}{z}^{(n)})'\underset{\sim}{w}(t)=\sum_{\nu=0}^{2n-1}w_\nu\cos(nx_\nu)\equiv\hat{w}_{\pm n}(t)$$

remains constant in time, i.e., that $\hat{w}_{\pm n}(t)=\hat{w}_{\pm n}(t=0)=0$. Thus, returning to the aliasing term in (B.7), it is enough to sum only the first $(n-1)$ modes

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

In particular, if $A(x)$ contains only one mode, the vanishing right-hand side results in the desired energy estimate for $\underset{\sim}{w}=\underset{\sim}{F}\underset{\sim}{v}$, which in turn yields the H^1-stability derived in [13].

Acknowledgment. It is a pleasure to express my gratitude to Professor Eli Turkel who read the manuscript and sugested many valuable improvements.

REFERENCES

[1] S. Abarbanel, D. Gottlieb and E. Tadmor, *Spectral methods for discontinuous problems*, in Numerical Methods for Fluid Dynamics, K. W. Morton and M. J. Baines, eds., Oxford University Press, Oxford, 1986, pp. 129–153.

[2] A. N. Brooks and T. J. R. Hughes, *Streamline upwind/Petrov Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations*, Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.

[3] C. Canuto and A. Quarteroni, *Approximation results for orthogonal polynomials in Sobolev spaces*, Math. Comp., 38 (1982), pp. 67–86.

[4] ———, *Error estimates for spectral and pseudospectral approximations of hyperbolic equations*, SIAM J. Numer. Anal., 19 (1982), pp. 629–642.

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

[6] M. J. P. Cullen and K. W. Morton, *Analysis of evolutionary error in finite element and other methods*, J. Comput. Phys., 34 (1980), pp. 245–267.

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

[8] B. Engquist and H.-O. Kreiss, *Difference and finite element methods for hyperbolic differential equations*, Comput. Methods Appl. Mech. Engrg., 17/18 (1979), pp. 581–596.

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

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

[11] D. Gottlieb and S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, CBMS-NSF Regional Conference Series in Applied Mathematics 26, Society for Industrial and Applied Mathematics, Philadelphia, 1977.

[12] 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.

[13] 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.

[14] D. Gottlieb, M. Y. Hussaini and S. A. Orszag, *Theory and applications of spectral methods for partial differential equations*, in Spectral Methods for Partial Differential Equations, R. G. Voigt, D. Gottlieb and M. Y. Hussaini, eds., Society for Industrial and Applied Mathematics, Philadelphia, 1984, pp. 1–54.

[15] D. Gottlieb and E. Tadmor, *Recovering pointwise values of discontinuous data within spectral accuracy*, in Progress and Supercomputing in Computational Fluid Dynamics, E. M. Murman and S. S. Abarbanel, eds., Birkhauser, Boston, 1985, pp. 357–375.

[16] B. Gustafsson, H.-O. Kreiss and A. Sundström, *Stability theory of difference approximations for mixed initial-boundary value problems*. II., Math. Comp., 26 (1972), pp. 649–686.

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

[18] C. Johnson, V. Nävert and J. Pitkäranta, *Finite element methods for linear hyperbolic problems*, Comput. Methods Appl. Mech. Engrg., 45 (1984), pp. 285–312.

[19] H.-O. Kreiss, *Stability theory for difference approximations of mixed initial-boundary value problems*. I., Math. Comp., 22 (1968), pp. 703–714.

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

[21] ———, *Methods for the approximate solution of time dependent problems*, GARP Publications Series, No. 10, World Meteorological Organization, Geneva, 1973.

[22] ———, *Stability of the Fourier method*, SIAM J. Numer. Anal., 16 (1979), pp. 421–433.

[23] Y. Maday and A. Quarteroni, *Spectral and pseudospectral approximations of Navier–Stokes equations*, SIAM J. Numer. Anal., 19 (1982), pp. 761–780.

[24] A. Majda and S. Osher, *Propagation of error into regions of smoothness for accurate difference approximations to hyperbolic equations*, Comm. Pure Appl. Math., 30 (1977), pp. 671–705.

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

[26] M. S. Mock and P. Lax, *The computation of discontinuous solutions of linear hyperbolic equations*, Comm. Pure Appl. Math., 31 (1978), pp. 423–430.

[27] J. Oliger, *Methods for time-dependent partial differential equations*, Proc. Sympos. Appl. Math., 22 (1978), pp. 87–108.

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

[29] ———, *Spectral methods for problems in complex geometries*, J. Comput. Phys., 37 (1980), pp. 70–92.

[30] S. A. Orszag and M. Israeli, *Numerical simulation of viscous incompressible flows*, Annual Review of Fluid Mechanics, Annual Reviews, Palo Alto, CA, Vol. 6, 1974, pp. 281–312.

[31] S. Osher, *Systems of difference equations with general homogeneous boundary conditions*, Trans. Amer. Math. Soc., 137 (1969), pp. 177–210.

[32] A. Patera, *A spectral element method for fluid dynamics: laminar flow in a channel expansion*, J. Comput. Phys., 54 (1984), pp. 468–488.

[33] A. Quarteroni, *Theoretical and computational aspects of spectral methods*, Proc. 5th International Conference on Computational Methods in Applied Science and Engineering, Versailles, North-Holland, Amsterdam, New York, 1981.

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

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

[36] G. Strang and G. Fix, *An Analysis of the Finite-Element Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[37] E. Tadmor, *Skew-selfadjoint form for systems of conservation laws*, J. Math. Anal. Appl., 103 (1984), pp. 428–442.

[38] ———, *The exponential accuracy of Fourier and Chebyshev differencing methods*, SIAM J. Numer. Anal., 23 (1986), pp. 1–10.

[39] E. Turkel, *Numerical methods for large-scale time-dependent partial differential equations*, Computational Fluid Dynamics, 2 (1980), pp. 127–262.

[40] K. Yosida, *Functional Analysis*, 2nd ed., Springer-Verlag, Berlin, New York, 1968.

[41] T. A. Zang and M. Y. Hussaini, *Mixed spectral-finite difference approximations for slightly viscous flows*, Lecture Notes in Phys. 141, Springer-Verlag, Berlin, 1980, pp. 461–466.

[42] A. Zygmund, *Trigonometrical Series*, Vols. I and II, Cambridge Univ. Press, Cambridge, 1968.