SPECTRAL METHODS FOR HYPERBOLIC PROBLEMS

Eitan Tadmor

School of Mathematical Sciences
Tel-Aviv University
Tel-Aviv 69978 ISRAEL

and

Department of Mathematics
UCLA
Los Angeles CA 90095 USA
Email: *tadmor@math.ucla.edu*

Abstract

We review several topics concerning spectral approximations of time-dependent problems, primarily — the accuracy and stability of Fourier and Chebyshev methods for the approximate solutions of hyperbolic systems.

To make these notes self contained, we begin with a very brief overview of Cauchy problems. Thus, the main focus of the first part is on hyperbolic systems which are dealt with two (related) tools: the energy method and Fourier analysis.

The second part deals with spectral approximations. Here we introduce the main ingredients of spectral accuracy, Fourier and Chebyshev interpolants, aliasing, differentiation matrices ...

The third part is devoted to Fourier method for the approximate solution of periodic systems. The questions of stability and convergence are answered by combining ideas from the first two sections. In this context we highlight the role of aliasing and smoothing; in particular, we explain how the lack of resolution might excite small scales weak instability, which is avoided by high modes smoothing.

The forth and final part deals with non-periodic problems. We study the stability of the Chebyshev method, paying special attention to the intricate issue of the CFL stability restriction on the permitted time-step.

Lecture notes delivered at
Ecole des Ondes
"*Méthodes numériques d'ordre élevé pour les ondes en régime transitoire*"
Inria - Rocquencourt, January 24-28 1994

September 1996

Contents

1 TIME DEPENDENT PROBLEMS

1.1 Initial Value Problems of Hyperbolic Type

The wave equation,

$$w_{tt} = a^2 w_{xx}, \tag{1.1.1}$$

is the prototype for PDE's of hyperbolic type. We study the pure initial-value problem associated with (1.1.1), augmented with 2π-periodic boundary conditions and subject to prescribed initial conditions,

$$w(x,0) = f(x), \quad w_t(x,0) = g(x). \tag{1.1.2}$$

We can solve this equation using the *method of characteristics*, which yields

$$w(x,t) = \frac{f(x+at) + f(x-at)}{2} + \frac{1}{2a}\int_{x-at}^{x+at} g(s)ds. \tag{1.1.3}$$

We shall study *the manner in which the solution depends on the initial data*. In this context the following features are of importance.

1. Linearity: the principle of superposition holds.
2. Finite speed of propagation: influence propagates with speed $\leq$ a. This is the essential feature of hyperbolicity. In the wave equation it is reflected by the fact that the value of w at (x,t) is not influenced by initial values outside domain of dependence $(x-at, x+at)$.
3. Existence for large enough set of admissible initial data: arbitrary C_0^∞ initial data can be prescribed and the corresponding solution is C_0^∞.
4. Uniqueness: the solution is uniquely determined for $-\infty < t < \infty$ by its initial data.
5. Conservation of Energy. The wave equation (1.1.1) describes the motion of a string with kinetic energy, $\frac{1}{2}\rho\int w_t^2 dx$, and potential one, $\frac{1}{2}T\int w_x^2 dx$, $\quad (T/\rho = a^2)$. In order to show that the total energy

$$E_{\text{Total}} = \frac{1}{2}\rho\int (w_t^2 + a^2 w_x^2)dx,$$

is conserved in time we may proceed in one of two ways: either by the so called energy method or by Fourier analysis.

1.1.1 The wave equation — hyperbolicity by the energy method

Rewrite (1.1.1) as a first order system

$$\frac{\partial}{\partial t}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 0 & a^2 \\ 1 & 0 \end{bmatrix}\frac{\partial}{\partial x}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \quad \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} \dfrac{\partial w}{\partial t} \\ \dfrac{\partial w}{\partial x} \end{bmatrix}, \tag{1.1.4}$$

or equivalently,

$$\frac{\partial u}{\partial t} = A\frac{\partial u}{\partial x}. \tag{1.1.5}$$

The essential ingredient here is the existence of a positive symmetrizer, $H > 0$,

$$HA = \begin{bmatrix} 0 & a^2 \\ a^2 & 0 \end{bmatrix} \equiv A_s = A_s^T, \quad H = \begin{bmatrix} 1 & 0 \\ 0 & a^2 \end{bmatrix}, \tag{1.1.6}$$

so that multiplication by H on the left gives

$$Hu_t = A_s u_x. \tag{1.1.7}$$

Multiplying by u^T we are led to

$$(u, Hu_t) = (u, A_s u_x), \tag{1.1.8}$$

and the real part of both sides are in fact perfect derivatives, for by the symmetry of H,

$$\mathrm{Re}(u, Hu_t) = \frac{1}{2}(u, Hu_t) + \frac{1}{2}(Hu_t, u) =$$
$$\frac{1}{2}(u, Hu_t) + \frac{1}{2}(u_t, Hu) = \frac{\partial}{\partial t}\left[\frac{1}{2}(u, Hu)\right],$$

and similarly, by the symmetry of A_s, we have

$$\mathrm{Re}(u, A_s u_x) = \frac{1}{2}(u, A_s u_x) + \frac{1}{2}(A_s u_x, u) = \frac{\partial}{\partial x}\left[\frac{1}{2}(u, A_s u)\right].$$

Hence, by integration over the 2π-period we end up with energy conservation, asserting

$$\frac{d}{dt}\int_x (w_t^2 + a^2 w_x^2)dx = \frac{d}{dt}\int_x (u, Hu)dx = \int_x \frac{\partial}{\partial x}(u, A_s u)dx = 0. \tag{1.1.9}$$

We note that the positivity of H was not used in the proof and is assumed just for the sake of making (u, Hu) an admissible convex "energy norm."

1.1.2 The wave equation — hyperbolicity by Fourier analysis

Fourier transform (1.1.5) to get the ODE

$$\frac{\partial \hat{u}}{\partial t}(k,t) = ikA\hat{u}(k,t), \tag{1.1.10}$$

whose solution is

$$\hat{u}(k,t) = e^{ikAt}\hat{u}(k,0), \tag{1.1.11}$$

where $\hat{u}(k,0)$ is the Fourier transform of the initial data. Now, for

$$A = T\Lambda T^{-1}, \quad \Lambda = \begin{bmatrix} -a & \\ & a \end{bmatrix}, T = \begin{bmatrix} -a & a \\ 1 & 1 \end{bmatrix}, \tag{1.1.12}$$

we find

$$\hat{u}(k,t) = Te^{ik\Lambda t}T^{-1}\hat{u}(k,0); \tag{1.1.13}$$

put differently, we have

$$T^{-1}\hat{u}(k,t) = \begin{bmatrix} e^{-ikat} & 0 \\ 0 & e^{ikat} \end{bmatrix} T^{-1}\hat{u}(k,0) \tag{1.1.14}$$

and hence (since the diagonal matrix inside the brackets on the right is clearly unitary), the L^2-norm of $T^{-1}\hat{u}(k,t)$ is conserved in time,i.e.,

$$\|T^{-1}\hat{u}(k,t)\|^2 = \|T^{-1}\hat{u}(k,0)\|^2, \quad T^{-1} = -\frac{1}{2a}\begin{bmatrix} 1 & -a \\ -1 & -a \end{bmatrix}. \tag{1.1.15}$$

Summing over all modes and using Parseval's equality we end up with energy conservation

$$\int_x (w_t^2 + a^2 w_x^2)dx = 4a^2 \int_x \left(\frac{w_t - aw_x}{-2a}\right)^2 + \left(\frac{w_t + aw_x}{-2a}\right)^2 dx$$

$$= 4a^2 \int_x \|T^{-1}u\|^2 dx = 8\pi a^2 \sum_k \|T^{-1}\hat{u}(k,t)\|^2 = \text{Const.}$$

as asserted.
We note that the only tool used in the energy method was the existence of a positive symmetrizer for

A, while the only tool used in the Fourier method was the real diagonalization of A; in fact the two are related, for if $A = T\Lambda T^{-1}$, then with $H = (T^{-1})^* T^{-1} > 0$ we have

$$HA = (T^{-1})^* \Lambda T^{-1} = A_s \equiv A_s^T, \quad \Lambda \text{ real diagonal.} \tag{1.1.16}$$

Energy conservation implies (in view of linearity) uniqueness, and serves as a basic tool to prove existence. It will be taken as the definition of *hyperbolicity*. It implies and is implied by the qualitative properties (1)—(4) which opened our discussion on page 3.

We now turn to consider general PDE's of the form

$$\frac{\partial u}{\partial t} = P(x,t,D)u, \quad P(x,t,D) = \sum_{j=1}^{d} A_j(x,t)\frac{\partial}{\partial x_j}, \tag{1.1.17}$$

with 2π-periodic boundary conditions and subject to prescribed initial conditions, $u(x,0) = f(x)$. Motivated by the example of the wave equation, we make the definition of

Hyperbolicity: We say that the system (1.1.17) is *hyperbolic* if the following a priori energy estimate holds:

$$\|u(x,t)\|_{L^2(x)} \leq \text{Const}_T \cdot \|u(x,0)\|_{L^2(x)}, \quad -T \leq t \leq T. \tag{1.1.18}$$

As we shall see later on, this notion of hyperbolicity is equivalent with *energy conservation* (— measured with respect to an appropriate renormed weighted 'energy'), in analogy with what we have seen in the special case of the wave equation. Here are the basic facts concerning such systems.

1.1.3 Hyperbolic systems with constant coefficients

We consider the 2π-periodic constant coefficients system

$$\frac{\partial u}{\partial t} = P(D)u, \quad P(D) = \sum_{j=1}^{d} A_j \frac{\partial}{\partial x}, \quad A_j = \text{constant matrices.} \tag{1.1.19}$$

Define the Fourier symbol associated with $P(D)$:

$$\hat{P}(ik) = i\sum_{j=1}^{d} A_j k_j, \quad k = (k_1, k_2, \cdots, k_d)\epsilon R^d, \tag{1.1.20}$$

which arises naturally when we Fourier transform (1.1.19),

$$\frac{\partial}{\partial t}\hat{u}(k,t) = \hat{P}(ik)\hat{u}(k,t). \tag{1.1.21}$$

Solving the ODE (1.1.21) we find, as before, that hyperbolicity amounts to

$$\|e^{\hat{P}(ik)t}\| \leq \text{Const}_T, \quad -T \leq t \leq T, \quad \text{for} \quad \text{all } k\text{'s}. \tag{1.1.22}$$

For this to be true the necessary Gårding-Petrovski condition should hold, namely

$$|\text{Re}\lambda[\hat{P}(ik)]| \leq \text{Const}. \tag{1.1.23}$$

Example: For the wave equation, (1.1.4), $\lambda[\hat{P}(ik)] = \pm ika$.
But the Gårding-Petrovski condition is not sufficient for the hyperbolic estimate (1.1.18) as told by the counterexample

$$\frac{\partial}{\partial t}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} a & 1 \\ 0 & a \end{bmatrix} \frac{\partial}{\partial x}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}.$$

As before, in this case we have $\lambda[\hat{P}(ik)] = \pm ika$, hence the Gårding-Petrovski condition is fulfilled. Yet, Fourier analysis shows that we need both $\|u_1(x,0)\|_{L^2(x)}$ *and* $\|\frac{\partial u_2}{\partial x}(x,0)\|_{L^2(x)}$ in order to upperbound $\|u_1(x,t)\|_{L^2(x)}$. Thus, the best we can hope for with this counterexample is an a priori estimate of the form

$$\|u(x,t)\|_{L^2(x)} \leq \text{Const}_T \cdot \|u(x,0)\|_{H^1(x)}, \quad -T \leq t \leq T.$$

We note that in this case we have a "loss" of one derivative, and this brings us to the notion of Weak Hyperbolicity: We say that the system (1.1.17) is *weakly hyperbolic* if there exists an $s \geq 0$ such that the following a priori estimate holds:

$$\|u(x,t)\|_{L^2(x)} \leq \text{Const}_T \cdot \|u(x,0)\|_{H^s(x)}, \quad -T \leq t \leq T. \tag{1.1.24}$$

The Gårding-Petrovski condition is necessary and sufficient for the system (1.1.19) to be weakly hyperbolic. A necessary and sufficient characterization of hyperbolic systems is provided by the Kreiss matrix theorem: it states that (1.1.22) holds iff there exists a positive symmetrizer $\hat{H}(k)$ such that

$$\text{Re}[\hat{H}(k)\hat{P}(ik)] \equiv 0, \quad 0 < m \leq \hat{H}(k) \leq M, \tag{1.1.25}$$

and this yields the conservation of the L^2-weighted norm, $\|u(x,t)\|_H^2 = 2\pi \sum_k \|\hat{u}(k)\|^2_{\hat{H}(k)}$; that is,

$$2\pi \sum_k (\hat{u}(k,t)), \hat{H}(k)\hat{u}(k,t))$$

is conserved in time.

Remark: For an a priori estimate *forward* in time ($0 \leq t \leq T$), it will suffice to have

$$\text{Re}[\hat{H}(k)\hat{P}(ik)] = \frac{1}{2}[\hat{H}(k)\hat{P}(ik) + \hat{P}(ik)\hat{H}(k)] \leq 0. \tag{1.1.26}$$

Indeed, we have in this case

$$\frac{1}{2}\frac{d}{dt}(\hat{u}(k), \hat{H}(k)\hat{u}(k)) \leq (\text{Re}[\hat{H}(k)\hat{P}(ik)]\hat{u}(k), \hat{u}(k)) \leq 0,$$

and hence summing over all k's and using Parseval's equality

$$\|u(x,t)\|^2_{L^2(x)} \leq \frac{M}{m}\|u(x,0)\|^2_{L^2(x)}.$$

Two important subclasses of hyperbolic equations are the *strictly hyperbolic* systems — where $\hat{P}(ik)$ has distinct real eigenvalues so that $\hat{P}(ik)$ can be real diagonalized

$$\hat{P}(k) = iT(k)\Lambda(k)T^{-1}(k),$$

and as before, $\hat{H}(k) = (T^{-1}(k))^* T^{-1}(k)$ will do; the other important case consists of *symmetric hyperbolic* systems which can be symmetrizer in the physical space, i.e. there exists an $H > 0$ such that

$$HA_j = A_{js} = A_{js}^T.$$

Most of the physically relevant systems fall into these categories.

Example: Shallow water equations (linearized)

$$\frac{\partial}{\partial t}\begin{bmatrix} u \\ v \\ \phi \end{bmatrix} + A_1\frac{\partial}{\partial x}\begin{bmatrix} u \\ v \\ \phi \end{bmatrix} + A_2\frac{\partial}{\partial y}\begin{bmatrix} u \\ v \\ \phi \end{bmatrix} = 0,$$

with

$$A_1 = \begin{bmatrix} u_0 & 0 & 1 \\ 0 & u_0 & 0 \\ \phi & 0 & u_0 \end{bmatrix}, A_2 = \begin{bmatrix} v_0 & 0 & 0 \\ 0 & v_0 & 1 \\ 0 & \phi_0 & v_0 \end{bmatrix},$$

can be symmetrized with

$$H = \begin{bmatrix} \phi_0 & & \\ & \phi_0 & \\ & & 1 \end{bmatrix}.$$

1.1.4 Hyperbolic systems with variable coefficients

We want to extend our previous analysis to linear systems of the form

$$\frac{\partial u}{\partial t} = P(x,t,D)u. \tag{1.1.27}$$

This is the motivation for the definition of hyperbolicity (1.1.18) in the context of constant coefficient problems: freeze the coefficients and assume the hyperbolicity of the constant coefficient problem(s), $u_t = P(x_0,t_0,D)u$, uniformly for each (x_0,t_0); then – in contrast to the notion of weak hyperbolicity, the variable coefficients problem is also hyperbolic. This result is based on the invariance of the notion of hyperbolicity under low-order perturbations[1].

As before the study of the variable coefficients problem can be carried out by one of two ways:

- by the *Fourier method* – one characterize the hyperbolicity of (1.1.27) in terms of the algebraic properties of the pseudodifferential symbol, $\hat{P}(x,t,ik) = e^{-ikx}P(x,t,D)e^{ikx}$;
- alternatively, we can also work directly in physical space with the *energy method*. For example, if we assume that $P(x,t,D)$ is *semi-bounded*, i.e., if

$$-M\|u\|^2_{L^2(x)} \le \mathrm{Re}(u, P(x,t,D)u)_{L^2(x)} \le M\|u\|^2_{L^2(x)}, \quad 0 < M, \tag{1.1.28}$$

then we have hyperbolicity (1.1.18).

Example: The symmetric hyperbolic case $A_j(x,t) = A_j^T(x,t)$: we can rewrite such symmetric problems in the equivalent form

$$\frac{\partial u}{\partial t} = \frac{1}{2}\left[\sum_j A_j \frac{\partial u}{\partial x_j} + \sum_j \frac{\partial}{\partial x_j}(A_j u)\right] + Bu, \quad B = -\frac{1}{2}\sum_j \frac{\partial A_j}{\partial x_j}.$$

In this case the symmetry of the A_j's implies that $\frac{1}{2}\left[\sum jA_j\frac{\partial u}{\partial x_j} + \sum_j \frac{\partial}{\partial x_j}(A_j u)\right]$ is skew-adjoint, i.e., integration by parts gives

$$\left(u, \frac{1}{2}\left[\sum_j A_j \frac{\partial u}{\partial x_j} + \sum_j \frac{\partial}{\partial x_j}(A_j u)\right]\right)_{L^2(x)} \equiv 0.$$

Therefore we have

$$\mathrm{Re}(u, P(x,t,D)u)_{L^2(x)} \equiv \mathrm{Re}(Bu, u)_{L^2(x)},$$

and hence the semi-boundedness requirement (1.1.28) holds with $M = \|\mathrm{Re}B\|$. Consequently, if $A_j(x,t)$ are symmetric (or at least symmetrizable) then the system (1.1.17) is hyperbolic.

1.2 Initial Value Problems of Parabolic Type

The heat equation,

$$u_t = au_{xx}, \quad a > 0, \tag{1.2.1}$$

is the prototype for PDE's of parabolic type. We study the pure initial-value problem associated with (1.2.1), augmented with 2π-periodic boundary conditions and subject to initial conditions

$$u(x,0) = f(x). \tag{1.2.2}$$

We can solve this equation using *Fourier transform* which yields

$$\hat{u}(k,t) = e^{-ak^2t}\hat{f}(k). \tag{1.2.3}$$

[1] This is a rather strong notion of hyperbolicity; it restricts such hyperbolic system to be of first-order.

It reflects the *dissipative* effect (= the rapid decay of the amplitudes , $|\hat{u}(k,t)|$, as functions of the high wavenumbers, $|k| \gg 1$), which is the essential feature of parabolicity.

As before, we study the manner in which the solution depends on its initial data.

1. Linearity: the principal of superposition holds.

2. Uniqueness: the solution is uniquely determined for $t > 0$ by the explicit formula

$$u(x,t) = \int_{y=-\infty}^{\infty} Q(x-y,t)f(y)dy, \quad Q(z) = \frac{1}{\sqrt{4\pi a t}} e^{\frac{-z^2}{4at}} > 0. \tag{1.2.4}$$

3. Existence for large enough set of admissible initial data: bounded initial data $f(x)$ can be prescribed (and even f's with $|f(x)| \leq e^{\alpha x^2}$), and the corresponding solution is C^∞ – in fact $u(x, t > 0)$ is analytic because of exponential decay in Fourier space.

4. The maximum principle: follows directly from the representation of $u(x,t)$ as a convolution of $f(x)$ with the unit mass positive kernel $Q(z)$.

5. Energy decay: as in the hyperbolic case we may proceed in one of two ways: Fourier analysis and the energy method.

1.2.1 The heat equation — Fourier analysis and the energy method

We start with

$$\|\frac{\partial^s}{\partial x^s} u(x,t)\|_{L^2}^2 \leq 2\pi \sum_k |\hat{f}(k)|^2 \cdot \max_k [|k|^{2s} \cdot |e^{-ak^2 t}|^2] \leq \text{Const.} t^{-s} \cdot \|f\|_{L^2}^2, \tag{1.2.5}$$

The last a priori estimate shows that the parabolic solution becomes infinitely smoother than its initial data (– we "gain" infinitely many s-derivatives), and at the same time these higher derivatives decay faster as $t \uparrow \infty$.

Alternatively, we can work with the energy method. Multiply (1.2.1) by u and integrate to get

$$\frac{1}{2}\frac{d}{dt}\|u\|_{L^2(x)}^2 \leq -a\|u_x\|_{L^2(x)}^2, \tag{1.2.6}$$

and in general

$$\frac{1}{2}\frac{d}{dt}\|\frac{\partial^s u}{\partial x^s}\|_{L^2}^2 \leq -\text{Const.}\|\frac{\partial^{s+1} u}{\partial x^{s+1}}\|_{L^2}^2; \tag{1.2.7}$$

successive integration of (1.2.7) yields (1.2.5).

1.2.2 Parabolic systems

Turning to general case, we consider mth-order PDE's of the form,

$$\frac{\partial u}{\partial t} = P(x,t,D)u, \quad P(x,t,D) = \sum_{|j|=0}^{m} A_j(x,t)D^j. \tag{1.2.8}$$

We say that the system (1.2.8) is *weakly parabolic* of order α if

$$\|\frac{\partial^s}{\partial x^s} u(x,t)\|_{L^2} \leq \text{Const.} t^{-|s|/\alpha} \|u(x,0)\|_{L^2(x)}. \tag{1.2.9}$$

For problems with constant coefficients this leads to the Gårding-Petrovski characterization of parabolicity of order β, requiring

$$\mathrm{Re}\lambda\left[\hat{P}(ik)=\sum_{|j|=0}^{m}A_j(ik)^j\right]\le -C_1\cdot|k|^\beta+C_2.$$

Remark: Generically we have $\alpha=\beta=m$ the order of dissipation which is *necessarily* even.

The extension to problems with variable coefficients case (with Lipschitz continuous coefficients) may proceed in one of two ways. Either, we freeze the coefficients and Fourier analyze the corresponding constant coefficients problems; or we may use the energy method, e.g., integration by parts shows that for

$$P(x,t,D)=\sum_j\frac{\partial}{\partial x_j}\left(A_j(x,t)\frac{\partial u}{\partial x_j}\right)+B_j\frac{\partial u}{\partial x_j}+Cu,$$

with $A_j+A_j^*>\delta>0,\quad$ and $\quad B_j=B_j^*$, the corresponding systems (1.2.8) is parabolic of order 2.

Example: $u_t=au_{xx}+u_{xxx}$ is weakly parabolic of order two, yet it does not satisfy Petrovski parabolicity.

1.3 Well-Posed Time-Dependent Problems

Hyperbolic and parabolic equations are the two most important categories of time-dependent problems whose evolution process is well-posed. Thus, consider the initial value problem

$$\frac{\partial u}{\partial t}=P(x,t,D)u. \tag{1.3.1}$$

We assume that a large enough class of admissible initial data

$$u(x,t=0)=f(x) \tag{1.3.2}$$

there exists a unique solution, $u(x,t)$. This defines a solution operator, $E(t,\tau)$ which describes the evolution of the problem

$$u(t)=E(t,\tau)u(\tau). \tag{1.3.3}$$

Hoping to compute such solutions, we need that the solutions will depend continuously in their initial data, i.e.,

$$\|u(t)-v(t)\|\le \mathrm{Const}_T\|u(0)-v(0)\|_{H^s}\quad 0\le t\le T. \tag{1.3.4}$$

In view of linearity, this amounts to having the a priori estimate (boundedness)

$$\|u(t)\equiv E(t,\tau)u(\tau)\|\le \mathrm{Const}_T\|u(\tau)\|_{H^s},\quad 0\le t\le T, \tag{1.3.5}$$

which includes the hyperbolic and parabolic cases.

Counterexample: (Hadamard) By Cauchy-Kowalewski, the system

$$\frac{\partial u}{\partial t}+A\frac{\partial u}{\partial x}=0,\quad u=\begin{bmatrix}u_1\\u_2\end{bmatrix},A=\begin{bmatrix}0&+1\\-1&0\end{bmatrix},$$

has a unique solution for arbitrary *analytic* data, at least for sufficiently small time. Yet, with initial data

$$u_1(x,0)=\frac{\sin nx}{n},\quad u_2(x,0)=0, \tag{1.3.6}$$

we obtain the solution

$$u_1(x,t) = \frac{\cosh nt \sin nx}{n}, \quad u_2(x,t) = \frac{\sinh nt \cos nx}{n} \tag{1.3.7}$$

which tends to infinity $\|u(\cdot,t)\|_{n\to\infty} \to \infty$, while the initial data tend to zero. Thus, the Laplace equation, $\frac{\partial^2 u_1}{\partial t^2} + \frac{\partial^2 u_1}{\partial x^2} = 0$, is *not* well-posed as an initial-value problem.

Finally, we note that a well-posed problem is stable against perturbations of inhomogeneous data in view of the following

Duhammel's principle. The solution of the inhomogeneous problem

$$\frac{\partial u}{\partial t} = P(x,t,D)u + F(x,t) \tag{1.3.8}$$

is given by

$$u(t) = E(t,0)u(0) + \int_{\tau=0}^{t} E(t,\tau)F(\tau)d\tau. \tag{1.3.9}$$

Indeed, a straightforward substitution yields

$$\begin{aligned}
\frac{\partial}{\partial t}u(t) &= \frac{\partial}{\partial t}[E(t,0)u(0)] + \frac{\partial}{\partial t}\left[\int_{\tau=0}^{t} E(t,\tau)F(t)d\tau\right] \\
&= P(x,t,D)[E(t,0)u(0)] + E(t,t)F(t) + \int_{\tau=0}^{t} \frac{\partial}{\partial t}[E(t,\tau)F(t)]d\tau \\
&= P(x,t,D)[E(t,0)u(0) + \int_{\tau=0}^{t} E(t,\tau)F(\tau)d\tau] + F(t) = P(x,t,D)u(t) + F.
\end{aligned}$$

This implies the a priori stability estimate

$$\|u(t)\| \le \mathrm{Const}_T \|u(0)\|_{H^s} + \mathrm{Const}_T \int_{\tau=0}^{t} \|F(\tau)\|_{H^s} d\tau, \quad 0 \le t \le T, \tag{1.3.10}$$

as asserted.

2 SPECTRAL APPROXIMATIONS

2.1 The Periodic Problem — The Fourier Expansion

Consider the first order Sturm-Liouville (SL) problem

$$\frac{d}{dx}\phi = \lambda\phi(x), \quad 0 \le x \le 2\pi, \tag{2.1.1}$$

augmented with periodic boundary conditions

$$\phi(0) = \phi(2\pi). \tag{2.1.2}$$

It has an infinite sequence of eigenvalues, $\lambda_k = ik$, with the corresponding eigenfunctions $\phi_k(x) = e^{ikx}$. Thus, $(\lambda_k = ik,\ \phi_k = e^{ikx})$ are the *eigenpairs* of the differentiation operator $D \equiv \frac{d}{dx}$ in $L^2[0, 2\pi)$, and they form a complete system in this space — completeness in the sense described below.

Let the space $L^2[0, 2\pi)$ be endowed with the usual Euclidean inner product

$$(w_1(x), w_2(x)) \equiv \int_0^{2\pi} w_1(x)\overline{w_2}(x)dx. \tag{2.1.3}$$

Note that $\phi_k(x) = e^{ikx}$ are orthogonal with respect to this inner product, for

$$(e^{ikx}, e^{ijx}) = \begin{cases} 0 & j \neq k, \\ \|e^{ikx}\|^2 = 2\pi & j = k. \end{cases} \tag{2.1.4}$$

Let $w(x)\epsilon L^2[0, 2\pi)$ be associated with its *spectral representation* in this system, i.e., the Fourier expansion

$$w(x) \sim \sum_{k=-\infty}^{\infty} \hat{w}(k)\phi_k(x), \quad \hat{w}(k) = \frac{(w, \phi_k)}{\|\phi_k\|^2}, \tag{2.1.5}$$

or equivalently,

$$w(x) \sim \sum_{k=-\infty}^{\infty} \hat{w}(k)e^{ikx}, \quad \hat{w}(k) = \frac{1}{2\pi}\int_{\xi=0}^{2\pi} w(\xi)e^{-ik\xi}. \tag{2.1.6}$$

The truncated Fourier expansion

$$S_N w \equiv \sum_{k=-N}^{N} \hat{w}(k)e^{ikx}, \tag{2.1.7}$$

denotes the spectral-Fourier projection of $w(x)$ into π_N–the space of trigonometric polynomials of degree $\le N$: [2]

$$\begin{aligned} S_N w &= \hat{w}(0) + \sum_{k=1}^{N}[\hat{w}(k)e^{ikx} + \hat{w}(-k)e^{-ikx}] \\ &= \hat{w}(0) + \sum_{k=1}^{N}[\hat{w}(k) + \hat{w}(-k)]\cos kx + i[\hat{w}(k) - \hat{w}(-k)]\sin kx \\ &= \sum_{k=0}^{N}{}'\hat{a}_k \cos kx + \hat{b}_k \sin kx; \end{aligned} \tag{2.1.8}$$

[2] $\sum'$ (and respectively, $\sum''$) indicate summation with $\frac{1}{2}$ of the first (and respectively, the first and the last) terms.

here $\hat{a}_k$ and $\hat{b}_k$ are the usual Fourier coefficients given by

$$\hat{a}_k = \hat{w}(k) + \hat{w}(-k) = \frac{1}{\pi}\int_0^{2\pi} w(\xi)\cos k\xi d\xi,$$
$$\hat{b}_k = i[\hat{w}(k) - \hat{w}(-k)] = \frac{1}{\pi}\int_0^{2\pi} w(\xi)\sin k\xi d\xi. \tag{2.1.9}$$

Since $w - S_N w$ is orthogonal to the π_N-space:

$$(w - S_N w, e^{ikx}) = 2\pi\hat{w}(k) - 2\pi\hat{w}(k) = 0, \quad |k| \le N, \tag{2.1.10}$$

it follows that for any $p_N \epsilon \pi_N$ we have (see Figure 2.1)

$$\|w - p_N\|^2 = \|w - S_N w\|^2 + \|S_N w - p_N\|^2. \tag{2.1.11}$$

Hence, $S_N w$ solves the *least-squares problem*

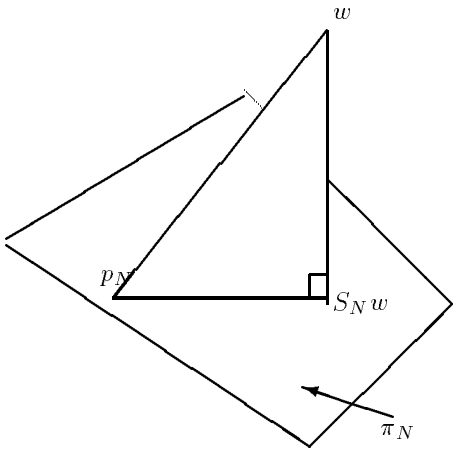


Figure 2.1: Least-squares approximation

$$\|w - S_N w\| = \min_{p_N \epsilon \pi_N} \|w - p_N\| \tag{2.1.12}$$

i.e., $S_N w$ is the best least-squares approximation to w. Moreover, (2.1.11) with $p_N = 0$ yields

$$\|S_N w\|^2 = \|w\|^2 - \|w - S_N w\|^2 \le \|w\|^2 \tag{2.1.13}$$

and by letting $N \to \infty$ we arrive at Bessel's inequality

$$2\pi \sum_{k=-\infty}^{\infty} |\hat{w}(k)|^2 \equiv \sum_{k=-\infty}^{\infty} |\hat{w}(k)|^2 \|\phi_k\|^2 \le \|w\|^2. \tag{2.1.14}$$

Remark: An immediate consequence of (2.1.14) is the Riemann-Lebesgue lemma, asserting that

$$\hat{w}(k) = \frac{1}{2\pi}\int_0^{2\pi} w(\xi)e^{-ik\xi}d\xi \longrightarrow_{k\to\infty} 0, \quad \text{for any } w\epsilon L^2[0,2\pi).$$

The system $\{\phi_k = e^{ikx}\}$ is complete in the sense that for any $w(x)\epsilon L^2[0,2\pi)$ we have Parseval's equality:

$$2\pi \sum_{k=-\infty}^{\infty} |\hat{w}(k)|^2 \equiv \sum_{k=-\infty}^{\infty} |\hat{w}(k)|^2 \|\phi_k\|^2 = \|w\|^2, \tag{2.1.15}$$

which in view of (2.1.13), is the same as

$$\lim_{N\to\infty} \|S_N w - w(x)\| = 0. \tag{2.1.16}$$

Thus *completeness* guarantee that the spectral projections 'fill in' the relevant space.
The last equality establishes the L^2 convergence of the spectral-Fourier projection, $S_N w(x)$, to $w(x)$, whose difference can be (upper-)bounded by the following

Error Estimate:

$$\|w - S_N w\|^2 = \|w\|^2 - \|S_N w\|^2 = \sum_{|k|>N} |\hat{w}(k)|^2 \|\phi_k\|^2 = 2\pi \sum_{|k|>N} |\hat{w}(k)|^2.$$

We observe that the RHS tends to zero as a tail of a converging sequence, i.e.,

$$\int_0^{2\pi} |w(x) - \sum_{k=-N}^{N} \hat{w}(k) e^{ikx}|^2 dx = 2\pi \sum_{|k|>N} |\hat{w}(k)|^2 \underset{N\to\infty}{\longrightarrow} 0. \tag{2.1.17}$$

The last equality tells us that the convergence rate depends on how fast the Fourier coefficients, $\hat{w}(k)$, decay to zero, and we shall quantify this in a more precise way below.

Remark. What about pointwise convergence? The L^2-convergence stated in (2.1.17) yields pointwise a.e. convergence for *subsequences*; one can show that in fact

$$\text{a.e.} \lim_{p\to\infty} |w(x) - S_{N_p} w(x)| = 0, \quad \inf_p \frac{N_{p+1}}{N_p} > 1. \tag{2.1.18}$$

The ultimate result in this direction states that , $w(x) = \text{a.e.}\lim_{N\to\infty} S_N w(x)$ (no subsequences) for all $w \epsilon L^2[0, 2\pi]$, though a.e. convergence may fail if $w(\cdot)$ is only $L^1[0, 2\pi]$-integrable.

The question of pointwise a.e. convergence is an extremely intricate issue for *arbitrary* L^2-functions. Yet, if we agree to assume sufficient smoothness, we find the convergence of spectral-Fourier projection to be very rapid, both in the L^2 and the pointwise sense. To this we proceed as follows.

2.1.1 Spectral accuracy

Define the Sobolev space $H^s[0, 2\pi)$ consisting of 2π-periodic functions for which their first s-derivatives are L^2-integrable; set the corresponding H^s-inner product as

$$(w_1, w_2)_{H^s} = \sum_{p=0}^{s} \int_0^{2\pi} D^p w_1(x) \overline{D^p w_2(x)} dx. \tag{2.1.19}$$

The essential ingredient here is that the system $\{e^{ikx}\}$ – which was already shown to be complete in $L^2[0, 2\pi) \equiv H^0[0, 2\pi)$, is also a complete system in $H^s[0, 2\pi)$ for any $s \geq 0$. For orthogonality we have

$$(e^{ikx}, e^{ijx})_{H^s} = \begin{cases} 0 & j \neq k, \\ 2\pi \sum_{p=0}^{s} k^{2p} & j = k. \end{cases} \tag{2.1.20}$$

The Fourier expansion now reads

$$w(x) \sim \sum_{k=-\infty}^{\infty} \hat{w}_s(k) e^{ikx} \tag{2.1.21}$$

where the Fourier coefficients, $\hat{w}_s(k)$, are given by

$$\hat{w}_s(k) = \frac{(w(x), e^{ikx})_{H^s}}{(e^{ikx}, e^{ikx})_{H^s}}. \tag{2.1.22}$$

We integrate by parts and use periodicity to obtain

$$\begin{aligned}(w(x),e^{ikx})_{H^s} &= \sum_{p=0}^{s}\int_0^{2\pi} D^p w(x)\overline{D^p e^{ikx}}dx = \\ &= \sum_{p=0}^{s}(-1)^p\int_0^{2\pi} w(x)\overline{D^{2p}e^{ikx}}dx \\ &= \sum_{p=0}^{s}(-1)^p(-ik)^{2p}\int_0^{2\pi} w(\xi)e^{-ik\xi}d\xi\end{aligned}$$

and together with (2.1.20) we recover the usual Fourier expansion we had before, namely

$$\hat{w}_s(k)\equiv\hat{w}(k)=\frac{1}{2\pi}\int_{\xi=0}^{2\pi} w(\xi)e^{-ik\xi}d\xi. \tag{2.1.23}$$

The completion of $\{e^{ikx}\}$ in $H^s[0,2\pi)$ gives us the Parseval's equality (compare (2.1.15)) which in turn implies

$$\begin{aligned}\|w-S_Nw\|^2_{H^s}=\sum_{|k|>N}|\hat{w}_s(k)|^2\|e^{ikx}\|^2_{H^s}=\sum_{|k|>N}\left[|\hat{w}(k)|^2\cdot 2\pi\sum_{p=0}^{s}k^{2p}\right]\leq \\ \sum_{p=0}^{s}N^{2p}\cdot 2\pi\cdot\sum_{|k|>N}|\hat{w}(k)|^2=\sum_{p=0}^{s}N^{2p}\cdot\|w-S_Nw\|^2.\end{aligned} \tag{2.1.24}$$

Since

$$\text{Const}_1(1+N^2)^{s/2}\leq\left(\sum_{p=0}^{s}N^{2p}\right)^{\frac{1}{2}}\leq\text{Const}_2(1+N^2)^{\frac{s}{2}}, \tag{2.1.25}$$

we conclude from (2.1.24), that for any $w\epsilon H^s[0,2\pi)$ we have

$$\|w-S_Nw\|\leq\text{Const}_s\cdot\frac{1}{N^s},\quad w\epsilon H^s[0,2\pi). \tag{2.1.26}$$

Note that $\text{Const}_s=\text{Const}_1\cdot\|w-S_Nw\|_{H^s}\underset{N\to\infty}{\longrightarrow}0$. This kind of estimate is usually referred to by saying that the Fourier expansion has *spectral accuracy*:

Spectral Accuracy — the error tends to zero faster than any *fixed* power of N, and is restricted only by the global smoothness of $w(x)$.

We note that as before, this kind of behavior is linked directly to the spectral decay of the Fourier coefficients. Indeed, by Cauchy-Schwartz inequality

$$\begin{aligned}|\hat{w}(k)|=|\hat{w}_s(k)| &\leq \frac{\|w\|_{H^s}\cdot\|e^{ikx}\|_{H^s}}{\|e^{ikx}\|^2_{H^s}}\leq\frac{1}{(2\pi\sum_{p=0}^{s}k^{2p})^{\frac{1}{2}}}\|w\|_{H^s} \\ &\leq \text{Const}\cdot\frac{1}{(1+|k|^2)^{\frac{s}{2}}}.\end{aligned} \tag{2.1.27}$$

In fact more is true. By Parseval's equality

$$\|w\|^2_{H^s}=\sum_{k=-\infty}^{\infty}|\hat{w}(k)|^2\|e^{ikx}\|^2_{H^s}=2\pi\sum_{k=-\infty}^{\infty}\left(\sum_{p=0}^{s}k^{2p}\right)|\hat{w}(k)|^2,$$

and hence by the Riemann-Lebesgue lemma, the product $(1+|k|^2)^{\frac{s}{2}}|\hat{w}(k)|$ is not only bounded (as asserted in (2.1.27), but in fact it tends to zero,

$$(1+|k|^2)^{\frac{s}{2}}|\hat{w}(k)| \underset{k\to\infty}{\longrightarrow} 0\,.$$

Thus, $\hat{w}(k)$ *tends to zero faster than* $|k|^{-s}$ *for all* $w(x)\epsilon H^s$. This yields spectral convergence, for

$$\|w - S_N w\|^2 = 2\pi \sum_{|k|>N} |\hat{w}(k)|^2 \leq \text{Const.} \sum_{|k|>N} \frac{1}{(1+|k|^2)^s} \leq \text{Const.} \frac{1}{N^{2s-1}}$$

i.e., we get slightly less than (2.1.26),

$$\|w - S_N w\| \leq \text{Const.} \frac{1}{N^{s-\frac{1}{2}}} \underset{N\to\infty}{\longrightarrow} 0 \quad s \geq 1.$$

Moreover, there is a rapid convergence for derivatives as well. Indeed, if $w(x)\epsilon H^s[0,2\pi)$ then for $0<\sigma<s$ we have

$$\begin{aligned}
\|w - S_N w\|^2_{H^\sigma} &= \sum_{|k|>N} (2\pi \sum_{p=0}^{\sigma} k^{2p}) |\hat{w}(k)|^2 \\
&\leq \text{Const.} \sum_{|k|>N} (1+|k|^2)^\sigma |\hat{w}(k)|^2 \\
&\leq \text{Const.} \sum_{|k|>N} \frac{(1+|k|^2)^s}{(1+N^2)^{s-\sigma}} |\hat{w}(k)|^2 \leq \\
&\leq \text{Const.} \sum_{|k|>N} \frac{(2\pi \sum_{p=0}^{s} k^{2p})}{(1+N^2)^{s-\sigma}} |\hat{w}(k)|^2 = \\
&\leq \text{Const.} \frac{\|w - S_N w\|^2_{H^s}}{N^{2(s-\sigma)}}.
\end{aligned}$$

Hence

$$\|w - S_N w\|_{H^\sigma} \leq \text{Const}_s \cdot \frac{1}{N^{s-\sigma}}, \quad \sigma \leq s, \quad w\epsilon H^s[0,2\pi) \tag{2.1.28}$$

with $\text{Const}_s \sim \|w - S_N w\|_{H^s} \underset{N\to\infty}{\longrightarrow} 0$. Thus, for each derivative we "lose" one order in the convergence rate.

As a corollary we also get uniform convergence of $S_N w(x)$ for $H^1[0,2\pi)$-functions $w(x)$, with the help of Sobolev-type estimate

$$\max_{0\leq x\leq 2\pi} |v(x)| \leq \text{Const.} \|v\|_{H^1}. \tag{2.1.29}$$

(Proof: Write $v(x) = \overline{v}(x_0) + \int_{x_0}^{x} v'(x)dx$ with $\overline{v}(x_0) \equiv \frac{1}{2\pi}\int_0^{2\pi} v(x)dx$, and use Cauchy-Schwartz to upper bound the two integrals on the right.)

Utilizing (2.1.29) with $v(x) = w(x) - S_N w(x)$ we find

$$\begin{aligned}
\max_{0\leq x\leq 2\pi} |w(x) - S_N w(x)| &\leq \text{Const.} \|w - S_N w\|_{H^1} \leq \\
&\leq \text{Const}_s \frac{1}{N^{s-1}} \underset{N\to\infty}{\longrightarrow} 0\,, \qquad w\epsilon H^s[0,2\pi),\ \text{Const}_s \to 0,\ s \geq 1.
\end{aligned} \tag{2.1.30}$$

In particular, we conclude that for any $w\epsilon H^s[0,2\pi), s>1$ we have, (in fact, $s>1/2$ will do – consult (2.5.22) below)

$$w(x) = \sum_{k=-\infty}^{\infty} \hat{w}(k) e^{ikx}, w\epsilon H^s[0,2\pi), \quad s>1. \tag{2.1.31}$$

In closing this section, we note that the spectral-Fourier projection, $S_N w(x)$, can be rewritten in the form

$$\begin{aligned} S_N w(x) = \sum_{k=-N}^{N} \hat{w}(k) e^{ikx} &= \frac{1}{2\pi} \int_{\xi=0}^{2\pi} w(\xi) \sum_{k=-N}^{N} e^{ik(x-\xi)} d\xi = \\ &= \int_{\xi=0}^{2\pi} D_N(x-\xi) w(\xi) d\xi \end{aligned} \tag{2.1.32}$$

where

$$D_N(x-\xi) = \frac{1}{2\pi} \sum_{k=-N}^{N} e^{ik(x-\xi)} = \frac{1}{2\pi} \frac{\sin\left(N+\frac{1}{2}\right)(x-\xi)}{\sin\left(\frac{x-\xi}{2}\right)}.$$

Thus, the spectral projection is given by a convolution with the so-called *Dirichlet kernel*,

$$D_N(x) = \frac{1}{2\pi} \frac{\sin\left(N+\frac{1}{2}\right)x}{\sin\frac{x}{2}}. \tag{2.1.33}$$

Now (2.1.30) reads

$$|w(x) - D_N(x) * w(x)| \le \text{Const}_s \cdot \frac{1}{N^{s-1}}, \quad \text{Const}_s \sim \|w\|_{H^s}. \tag{2.1.34}$$

2.2 The Periodic Problem — The Fourier Interpolant

We have seen that given the "moments"

$$\hat{w}(k) = \frac{1}{2\pi} \int_{\xi=0}^{2\pi} w(\xi) e^{-ik\xi} d\xi, \quad -N \le k \le N, \tag{2.2.1}$$

we can recover smooth functions $w(x)$ within spectral accuracy. Now, suppose we are given discrete data of $w(x)$: specifically, assume $w(x)$ is known at equidistant collocation points [3]

$$w_\nu = w(x_\nu), \quad x_\nu = r + \nu h, \quad \nu = 0, 1, \cdots, 2N. \tag{2.2.2}$$

Without loss of generality we can assume that r — which measures a fixed shift from the origin, satisfies

$$0 \le r < h \equiv \frac{2\pi}{2N+1}. \tag{2.2.3}$$

Given the equidistant values w_ν, we can approximate the above "moments," $\hat{w}(k)$, by the trapezoidal rule

$$\tilde{w}(k) = \frac{h}{2\pi} {\sum_{\nu=0}^{2N+1}}'' w_\nu e^{-ikx_\nu} \equiv \frac{1}{2N+1} \sum_{\nu=0}^{2N} w_\nu e^{-ikx_\nu}. \tag{2.2.4}$$

Using $\tilde{w}(k)$ instead of $\hat{w}(k)$ in (2.1.7), we consider now the pseudospectral approximation

$$\psi_N w = \sum_{k=-N}^{N} \tilde{w}(k) e^{ikx}. \tag{2.2.5}$$

The error, $w(x) - \psi_N w(x)$, consists of two parts:

$$w(x) - \psi_N w(x) = \sum_{|k|>N} \hat{w}(k) e^{ikx} + \sum_{|k|\le N} [\hat{w}(k) - \tilde{w}(k)] e^{ikx}.$$

[3] We treat here the case of an odd number of $2N+1$ collocation points. We get even in §2.2.3

The first contribution on the right is the *truncation error*

$$T_N w(x) \equiv (I - S_N)w(x) = \sum_{|k|>N} \hat{w}(k) e^{ikx}.$$

We have seen that it is spectrally small provided $w(x)$ is sufficiently smooth. The second contribution on the right is the *aliasing error*

$$A_N w(x) = \sum_{|k|\le N} [\hat{w}(k) - \tilde{w}(k)] e^{ikx}. \tag{2.2.6}$$

This is pure discretization error; to estimate its size we need the

Poisson's Summation Formula (Aliasing). *Assume* $w(x) \epsilon H^1[0, 2\pi)$. *Then we have*

$$\tilde{w}(k) = \sum_{p=-\infty}^{\infty} e^{ip(2N+1)r} \hat{w}(k + p(2N+1)). \tag{2.2.7}$$

The proof of (2.2.7) is based on the *pointwise* representation of $w(x) \epsilon H^1[0, 2\pi)$ by its Fourier expansion (2.1.31),

$$\tilde{w}(k) = \frac{1}{2N+1} \sum_{\nu=0}^{2N} w(x_\nu) e^{-ikx_\nu} = \frac{1}{2N+1} \sum_{\nu=0}^{2N} \left[\sum_{j=-\infty}^{\infty} \hat{w}(j) e^{ijx_\nu} \right] e^{-ikx_\nu}. \tag{2.2.8}$$

Since $w(x)$ is assumed to be in H^1, the summation on the right is absolutely convergent

$$\sum_{j=-\infty}^{\infty} |\hat{w}(j)| \le \left(\sum_j (1+j^2) |\hat{w}(j)|^2 \cdot \sum_j \frac{1}{1+j^2} \right)^{\frac{1}{2}} \le \text{Const.} \|w\|_{H^1}, \tag{2.2.9}$$

and hence we can interchange the order of summation

$$\tilde{w}(k) = \frac{1}{2N+1} \sum_{j=-\infty}^{\infty} \hat{w}(j) \sum_{\nu=0}^{2N} e^{i(j-k)x_\nu}. \tag{2.2.10}$$

Straightforward calculation yields

$$\begin{aligned} &\frac{1}{2N+1} \sum_{\nu=0}^{2N} e^{i(j-k)(r+\nu h)} = e^{i(j-k)r} \cdot \frac{1}{2N+1} \cdot \sum_{\nu=0}^{2N} e^{i(j-k)\nu \frac{2\pi}{2N+1}} = \\ &= e^{i(j-k)r} \cdot \frac{1}{2N+1} \begin{cases} \dfrac{e^{i(j-k)\frac{2\pi \cdot (2N+1)}{2N+1}} - 1}{e^{i(j-k)\frac{2\pi}{2N+1}} - 1} = 0 & j - k \neq 0 [\text{mod } 2N+1] \\ 2N+1, & j - k = p \cdot (2N+1). \end{cases} \end{aligned} \tag{2.2.11}$$

and we end up with the asserted equality

$$\tilde{w}(k) = \sum_{j=-\infty}^{\infty} \hat{w}(j) \cdot \frac{1}{2N+1} \sum_{\nu=0}^{2N} e^{i(j-k)x_\nu} = \sum_{p=-\infty}^{\infty} \hat{w}(k + p(2N+1)) \cdot e^{ip \cdot (2N+1)r}.$$

2.2.1 Aliasing and spectral accuracy

We note that once $w(x)$ is assumed to be smooth, it is completely determined (– in the pointwise sense) by its Fourier coefficients $\hat{w}(k)$; so are its equidistant values $w_\nu \equiv w(x_\nu)$ and so are its discrete Fourier coefficients $\tilde{w}(k)$. The aliasing formula shows that $\tilde{w}(k)$ are determined in terms of $\hat{w}(k)$, by folding back high modes on the lowest ones, due to the discrete resolution of the moments of $w(x)$: all modes $= k[\mathrm{mod} 2N+1]$ are aliased to the same place since they are equal on the gridpoints

$$e^{i(k+p(2N+1))x_\nu} = e^{ip(2N+1)r} \cdot e^{ikx_\nu}. \tag{2.2.12}$$

Let us rewrite (2.2.7) in the form

$$\tilde{w}(k) = \hat{w}(k) + \sum_{p\neq 0} e^{ip(2N+1)r} \cdot \hat{w}(k+p(2N+1)).$$

Returning to the aliasing error in (2.2.6), we now have

$$A_N w(x) = \sum_{|k|\leq N} \left[\sum_{p\neq 0} e^{ip\cdot(2N+1)r} \cdot \hat{w}(k+p\cdot(2N+1)) \right] e^{ikx}. \tag{2.2.13}$$

We note that the truncation error $T_N w(x)$ lies outside π_N, while the aliasing error $A_N w(x)$ lies in π_N, hence by H^s-orthogonality

$$\begin{aligned} &\|w(x) - \psi_N w(x)\|^2_{H^s} = \\ &= \overbrace{\sum_{|k|>N} (1+|k|^2)^s \cdot |\hat{w}(k)|^2}^{\text{truncation}} + \overbrace{\sum_{|k|\leq N} (1+|k|^2)^s \cdot |\sum_{p\neq 0} e^{ip(2N+1)r} \cdot \hat{w}(k+p(2N+1))|^2}^{\text{aliasing}}. \end{aligned} \tag{2.2.14}$$

Both contributions involve only the *high amplitudes* – higher than N in absolute value; in fact they involve precisely all of these high amplitudes. This leads us to aliasing estimate

$$\begin{aligned} &\sum_{|k|\leq N} (1+|k|^2)^s |\sum_{p\neq 0} e^{ip(2N+1)r} \cdot \hat{w}(k+p(2N+1))|^2 \leq \\ &\sum_{|k|\leq N} \sum_{p\neq 0} (1+|k+p(2N+1)|^2)^s |\hat{w}(k+p(2N+1))|^2 \cdot \cdot \max_{|k|\leq N} \sum_{p\neq 0} \left[\frac{1+|k|^2}{1+|k+p\cdot(2N+1)|^2} \right]^s \leq \\ &\|T_N w(x)\|^2_{H^s} \cdot \sum_{p\neq 0} \left[\frac{1+N^2}{1+4p^2N^2} \right]^s. \end{aligned} \tag{2.2.15}$$

We conclude that the aliasing error is dominated by the truncation error (at least for any $s > \frac{1}{2}$),

$$\|A_N w(x)\|_{H^s} \leq \mathrm{Const}_s \cdot \|T_N w(x)\|_{H^s}, \quad s > \frac{1}{2}. \tag{2.2.16}$$

Augmenting this with our previous estimates on the truncation error we end up with spectral accuracy as before, namely

$$\|w - \psi_N w\|_{H^\sigma} \leq \mathrm{Const}_s \cdot \frac{1}{N^{s-\sigma}}, \quad w \epsilon H^s[0,2\pi), \quad s \geq \sigma > \frac{1}{2}. \tag{2.2.17}$$

2.2.2 Fourier differentiation matrix

We observe that $\psi_N w(x)$ is nothing but the trigonometric interpolant of $w(x)$ at the equidistant points $x = x_\mu$:

$$\psi_N w(x)|_{x=x_\mu} = \sum_{k=-N}^{N} \left[\frac{1}{2N+1} \sum_{\nu=0}^{2N} w(x_\nu) e^{-ikx_\nu} \right] e^{ikx_\mu} = \tag{2.2.18}$$
$$= \sum_{\nu=0}^{2N} w(x_\nu) \cdot \frac{1}{2N+1} \sum_{k=-N}^{N} e^{ik(\mu-\nu)h} = w(x_\mu).$$

This shows that ψ_N is in fact a ψdospectral projection, which in the usual sin-cos formulation reads

$$\psi_N w = {\sum_{k=0}^{N}}' \tilde{a}_k \cos kx + \tilde{b}_k \sin kx \tag{2.2.19}$$
$$\begin{bmatrix} \tilde{a}_k \\ \tilde{b}_k \end{bmatrix} = \frac{2}{2N+1} \sum_{\nu=0}^{2N} w(x_\nu) \begin{bmatrix} \cos kx_\nu \\ \sin kx_\nu \end{bmatrix}.$$

Thus, trigonometric interpolation provides us with an excellent vehicle to perform approximate discretizations with high (= spectral) accuracy, of differential and integral operations. These can be easily carried out in Fourier space where the exponentials serve as eigenfunction. For example, suppose we are given the equidistant gridvalues, w_ν, of an underlying smooth (i.e., also periodic!) function $w(x), w(x) \epsilon H^s[0, 2\pi)$. A second-order accurate discrete derivative is provided by center differencing

$$\frac{dw}{dx}(x = x_\nu) = \frac{w_{\nu+1} - w_{\nu-1}}{2h} + \mathcal{O}(h^2).$$

Note that the error in this case is, $\mathcal{O}(h^2) \equiv w^{(3)}(\xi)h^2$, no matter how smooth $w(x)$ is. Similarly, fourth order approximation is given (via Richardson's extrapolation procedure) by

$$\frac{dw}{dx}(x = x_\nu) = \frac{8[w_{\nu+1} - w_{\nu-1}] - [w_{\nu+2} - w_{\nu-2}]}{12h} + \mathcal{O}(h^4).$$

The pseudospectral approximation gives us an alternative procedure: construct the trigonometric interpolant

$$\psi_N w(x) = \sum_{k=-N}^{N} \tilde{w}(k) e^{ikx}, \quad \tilde{w}(k) = \frac{1}{2N+1} \sum_{\nu=0}^{2N} w_\nu e^{-ikx_\nu}. \tag{2.2.20}$$

Differentiation in Fourier space amounts to simple multiplication, since the exponentials are eigenfunctions of differentiation,

$$\frac{d}{dx}\psi_N w(x) = \sum_{k=-N}^{N} \tilde{w}(k) ike^{ikx}, \tag{2.2.21}$$

and we approximate

$$\frac{dw}{dx}(x = x_\nu) = \frac{d}{dx}\psi_N w(x)|_{x=x_\nu} + \text{ spectrally small error.} \tag{2.2.22}$$

Indeed, by our estimates we have for $w(x) \epsilon H^s[0, 2\pi), s > 1$,

$$\max_{0 \le x \le 2\pi} |\frac{d}{dx} w(x) - \frac{d}{dx}\psi_N w(x)| \le \text{Const.} \|w(x) - \psi_N w(x)\|_{H^2} \le \frac{\text{Const}_s}{N^{s-2}} \tag{2.2.23}$$

which verifies the asserted spectral accuracy. Similar estimates are valid for higher derivatives. To carry out the above recipe, one proceeds as follows: starting with the vector of gridvalues, $\tilde{w} = (w_0, \cdots, w_{2N})$, one computes the discrete Fourier coefficients

$$\tilde{w}(k) = \frac{1}{2N+1} \sum_{\nu=0}^{2N} w_\nu e^{-ikx_\nu}, \quad -N \le k \le N, \tag{2.2.24}$$

or, in matrix formulation

$$\begin{bmatrix} \tilde{w}(-N) \\ \vdots \\ \tilde{w}(N) \end{bmatrix} = F \begin{bmatrix} w_0 \\ \vdots \\ w_{2N} \end{bmatrix}, \quad F_{k\nu} = \frac{1}{2N+1} e^{-ikx_\nu}; \tag{2.2.25}$$

then we differentiate

$$\tilde{w}(k) \to ik\tilde{w}(k), \tag{2.2.26}$$

or in matrix formulation

$$\begin{bmatrix} \tilde{w}(-N) \\ \vdots \\ \tilde{w}(N) \end{bmatrix} \to \Lambda \begin{bmatrix} \tilde{w}(-N) \\ \vdots \\ \tilde{w}(N) \end{bmatrix}, \quad \Lambda = \begin{bmatrix} -iN & & \\ & \ddots & \\ & & iN \end{bmatrix}, \tag{2.2.27}$$

and finally, we return to the "physical" space, calculating

$$\sum_{k=-N}^{N} ik\tilde{w}(k) e^{ikx_\nu}, \quad \nu = 0, 1, \cdots, 2N, \tag{2.2.28}$$

or in matrix formulation

$$\begin{bmatrix} \frac{dw}{dx}(x_0) \\ \vdots \\ \frac{dw}{dx}(x_{2N}) \end{bmatrix} = F^* \cdot (2N+1) \begin{bmatrix} -iN\tilde{w}(-N) \\ \vdots \\ iN\tilde{w}(N) \end{bmatrix}, \quad (2N+1)F^*_{\nu k} = e^{ikx_\nu}. \tag{2.2.29}$$

The summary of these three steps is

$$\begin{bmatrix} w'(x_0) \\ \vdots \\ w'(x_{2N}) \end{bmatrix} = \psi D \begin{bmatrix} w_0 \\ \vdots \\ w_{2N} \end{bmatrix}, \quad \psi D \equiv (2N+1)F^*\Lambda F, \tag{2.2.30}$$

where ψD represents the discrete differentiation matrix, and similarly ψD^s for higher derivatives.

Note: Since $(2N+1)F^*F = I_{2N+1}$ (interpolation!) we apply $\psi D^s = (2N+1)F^*\Lambda^s F$. How does this compare with finite differences and finite-element type differencing?

In periodic second-order differencing we have

$$FD_2 = \frac{1}{2h} \begin{bmatrix} 0 & 1 & \cdots & 0 & -1 \\ -1 & 0 & & & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & & & 0 & 1 \\ 1 & 0 & \cdots & -1 & 0 \end{bmatrix};$$

fourth order differencing yields

$$FD_4 = \frac{1}{12h} \begin{bmatrix} 0 & 8 & -1 & \cdots & 1 & -8 \\ -8 & 0 & & & & 1 \\ 1 & & \ddots & & & \vdots \\ \vdots & & & & & -1 \\ -1 & & & & 0 & 8 \\ 8 & -1 & \cdots & 1 & -8 & 0 \end{bmatrix}.$$

In both cases the second and fourth order differencing takes place in the physical space. The corresponding differencing matrices have *finite* bandwidth and this reflects the fact that these differencing methods are *local*. Similarly, finite-element differencing,

$$\frac{1}{6}w'_{\nu-1}+\frac{4}{6}w'_{\nu}+\frac{1}{6}w'_{\nu+1}=\frac{w_{\nu+1}-w_{\nu-1}}{2h}$$

corresponds to a differencing matrix

$$FE_4=\begin{bmatrix}\frac{4}{6}&\frac{1}{6}&\cdots&\frac{1}{6}\\ \frac{1}{6}&\ddots&&\frac{1}{6}\\ \frac{1}{6}&\cdots&\frac{1}{6}&\frac{4}{6}\end{bmatrix}^{-1}\cdot\frac{1}{2h}\begin{bmatrix}0&1&\cdots&0&-1\\ -1&0&&&0\\ \vdots&&\ddots&&\vdots\\ 0&&&0&1\\ 1&0&\cdots&-1&0\end{bmatrix}.$$

We still operate in physical space with $\mathcal{O}(N)$ operations (tridiagonal solver) and locality is reflected by a very rapid (exponential decay) away from main diagonal. Nevertheless, if we increase the periodic center differences stencil to its limit then we end up with *global* pseudospectral differentiation

$$\frac{d}{dx}\psi_N w(x_\nu)=\sum_{k=-N}^{N}\left(\frac{ik}{2N+1}\sum_{\mu=0}^{2N}w_\mu e^{-ikx_\mu}\right)e^{ikx_\nu};\tag{2.2.31}$$

recall the Dirichlet kernel (2.1.33)

$$\sum_{k=-N}^{N}e^{ikx}=e^{-iNx}\frac{e^{i(2N+1)x}-1}{e^{ix}-1}=\frac{\sin(N+\frac{1}{2})x}{\sin\frac{x}{2}},\tag{2.2.32}$$

and its derivative,

$$\sum_{k=-N}^{N}ike^{ikx}=\frac{d}{dx}\frac{\sin(N+\frac{1}{2})x}{\sin\frac{x}{2}}=\frac{(N+\frac{1}{2})\cos(N+\frac{1}{2})x\sin\frac{x}{2}-\frac{1}{2}\cos\frac{x}{2}\sin(N+\frac{1}{2})x}{\sin^2\frac{x}{2}}\tag{2.2.33}$$

so that

$$\sum_{k=-N}^{N}ike^{ik(\nu-\mu)h}=\frac{(N+\frac{1}{2})\cos[(N+\frac{1}{2})(\nu-\mu)h]}{\sin(\frac{x_\nu-x_\mu}{2})}.\tag{2.2.34}$$

Hence (2.2.31), (2.2.34) give us

$$w'(x_\nu)\equiv\frac{d}{dx}\psi_N w(x_\nu)=\sum_{\mu=0}^{2N}\frac{1}{2}\frac{(-1)^{\nu-\mu}}{\sin(\frac{x_\nu-x_\mu}{2})}\cdot w_\mu,\quad [\psi D]_{\nu\mu}=\delta_{\nu\mu}\frac{(-1)^{\nu-\mu}}{2\sin(\frac{x_\nu-x_\mu}{2})}.\tag{2.2.35}$$

In this case ψD is a full $(2N+1)\times(2N+1)$ matrix whose multiplication requires $\mathcal{O}(N^2)$ operations; however, we can multiply $\psi D[w]$ efficiently using its *spectral representation* from (2.2.30),

$$\psi D=(2N+1)F^*\Lambda F.$$

Multiplication by F and F^* can be carried out by FFT which requires only $\mathcal{O}(N\log N)$ operating and hence the total cost here is almost as good as standard "local" methods, and in addition we maintain spectral accuracy.

We have seen how the pseudospectral differentiation works in the physical space. Next, let's examine how the standard finite-difference/element differencing methods operate in the Fourier space. Again,

the essential ingredient is that exponentials play the role of *eigenfunctions* for this type of differencing. To see this, consider for example the usual second order centered differencing, $D_2(h)$, for which we have

$$D_2(h)e^{ikx}|_{x=x_\nu} = \frac{e^{ikx_{\nu+1}} - e^{-ikx_{\nu-1}}}{2h} = \frac{i\sin(kh)}{h}e^{ikx}|_{x=x_\nu}, \tag{2.2.36}$$

The term $\frac{i\sin(kh)}{h}$ is called the "symbol" of center differencing. By superposition we obtain for arbitrary grid function (represented here by its trigonometric interpolant)

$$\psi_N w(x) = \sum_{k=-N}^{N} \tilde{w}(k)e^{ikx} \tag{2.2.37}$$

that

$$\begin{aligned} \frac{w_{\nu+1} - w_{\nu-1}}{2h} = D_2(h)\psi_N w &= \sum_{k=-N}^{N} \tilde{w}(k)D_2(h)e^{ikx}|_{x=x_\nu} \\ &= \sum_{k=-N}^{N} \frac{i\sin(kh)}{h}\tilde{w}(k)e^{ikx}|_{x=x_\nu}. \end{aligned} \tag{2.2.38}$$

It is second-order accurate differencing since its symbol satisfies

$$\frac{i\sin(kh)}{h} = ik + \mathcal{O}(k^3h^2). \tag{2.2.39}$$

Note that for the low modes we have $\mathcal{O}(h^2)$ error (the less significant high modes are differenced with $\mathcal{O}(1)$ error but their amplitudes tend rapidly to zero). Thus we have

$$\begin{aligned} \|\frac{d}{dx}\psi_N w - D_2(h)\psi_N w\|^2 &= \sum_{k=-N}^{N} |k - \frac{\sin(kh)}{h}|^2|\tilde{w}(k)|^2 \\ \leq \text{Const.}h^4 \sum_{|k|\leq N} (1+|k|^2)^3|\tilde{w}(k)|^2 &\leq \text{Const.}h^4 \cdot \|\psi_N w\|^2_{H^3}, \end{aligned} \tag{2.2.40}$$

and this estimate should be compared with the usual

$$\left|\frac{d}{dx}w(x_\nu) - \frac{w_{\nu+1} - w_{\nu-1}}{2h}\right| \leq \text{Const.}h^2 \cdot \max_{x_{\nu-1}\leq x\leq x_{\nu+1}} \left|w^{(3)}(x)\right|.$$

The main difference between these two estimates lies in the fact that the last estimate is *local*, i.e., we need the smoothness of $w(x)$ only in the neighborhood of $x = x_\nu$, and not in the whole interval, $x_{\nu-1} \leq x \leq x_{\nu+1}$. The analogue localization in the Fourier space will be dealt later.

Similarly, we have for fourth order differencing the symbol

$$i\frac{1}{3}\left[4\frac{\sin kh}{h} - \frac{\sin 2kh}{2h}\right] = ik + \mathcal{O}(k^5h^4).$$

In general, we encounter difference operators whose matrix representation, D,

$$D = [d_{jk}] \quad -N \leq j,k \leq N, \tag{2.2.41}$$

is periodic and antisymmetric (here $[\ell] \equiv \ell[\text{mod } 2N+1]$),

$$\begin{aligned} &\text{(i)} \quad \text{periodicity}: \ d_{jk} = d_{[k-j]} \\ &\text{(ii)} \quad \text{antisymmetry}: \ d_{jk} = -d_{kj}. \end{aligned} \tag{2.2.42}$$

Matrices satisfying the periodicity property are called circulant, and they all can be diagonalized by the unitary Fourier matrix

$$D = U^* \Lambda U, \quad U = (2N+1)^{\frac{1}{2}} \cdot F, \quad U^* U = I_{2N+1}. \tag{2.2.43}$$

Indeed, with $p - q = \ell$ we have

$$\begin{aligned}
[U^* D U]_{j\,k} &= \frac{1}{2N+1} \sum_{p,q=-N}^{N} e^{ij\,x_p} \cdot d_{[p-q]} e^{-ik\,x_q} = \\
&= \frac{1}{2N+1} \sum_{\ell,q=-N}^{N} e^{ij[r+(q+\ell)h]} d_{[\ell]} e^{-ik(r+qh)} \\
&= \frac{1}{2N+1} \sum_{\ell,q=-N}^{N} e^{ij\ell h} d_{[\ell]} \cdot \sum_{q=-N}^{N} e^{-i(k-j)\cdot(r+qh)} \\
&= \begin{cases} 0 & j \neq k, \\ \displaystyle\sum_{\ell=-N}^{N} e^{ik\ell h} d_{[\ell]} & j = k, \end{cases}
\end{aligned} \tag{2.2.44}$$

and using the antisymmetry we end up with symbols λ_k

$$\Lambda = \mathrm{diag}(\lambda_{-N}, \cdots, \lambda_N), \quad \lambda_k = 2i \sum_{\ell=1}^{N} d_\ell \sin(k\ell h). \tag{2.2.45}$$

As an example, we obtain for the (linear) finite-element differencing system

$$\begin{aligned}
\lambda_k &= i\frac{\sin kh}{h} \left(\frac{4}{6} + \frac{1}{6} e^{ikh} + \frac{1}{6} e^{-ikh} \right) = \\
&= \frac{6i}{h} \cdot \frac{\sin(kh)}{4 + 2\cos(kh)} = ik + \mathcal{O}(h^4).
\end{aligned} \tag{2.2.46}$$

This corresponds to differentiation of the forth-order Padé expansion.

In general, the symbols are trigonometric polynomials or rational functions in the "dual variable," kh, which has "exact" representation on the grid in terms of translation operator (polynomials or rational functions), and accuracy is determined by the ability to approximate the exact differentiation symbol, ik, for $|kh| \sim 1$, consult Figure 2.2.

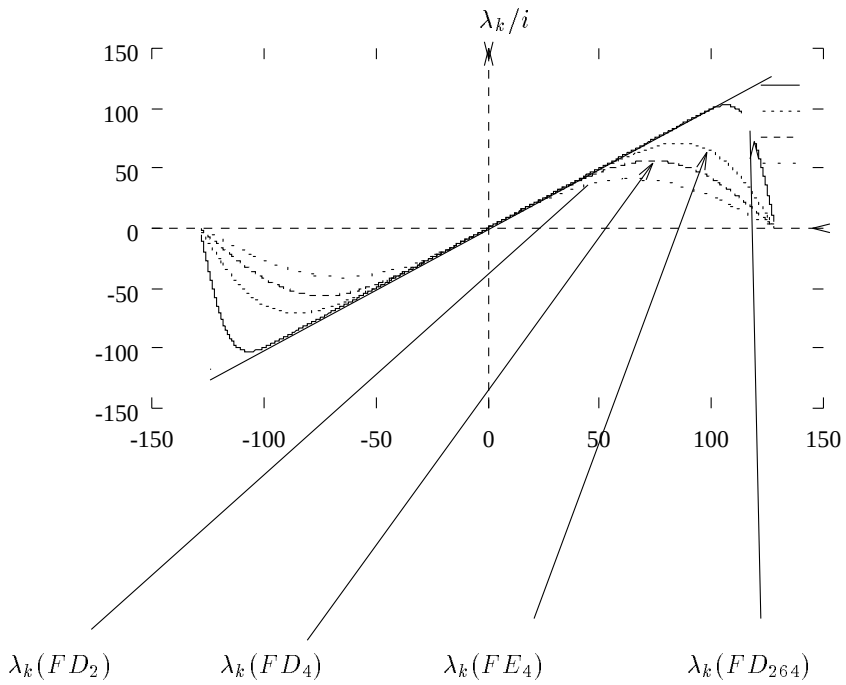


Figure 2.2: The symbols of center differencing

2.2.3 Fourier interpolant revisited on an even number of gridpoints

We assume $w(x)$ is known at the $2N$ gridpoints $x_\nu = r + \nu h \quad \nu = 0, 1, \cdots, 2N-1$,

$$w_\nu = w(x_\nu) \quad \nu = 0, 1, \cdots, 2N-1 \tag{2.2.47}$$

with $h \equiv \frac{2\pi}{2N} = \frac{\pi}{N}$, and $0 \le r < h$ is fixed. We use the trapezoidal rule to approximate the Fourier coefficients $\tilde{w}(k)$ in (2.2.1)

$$\tilde{w}(k) = \frac{1}{2\pi} \sum_{\nu=0}^{2N}{}'' w_\nu e^{-ikx_\nu} h = \frac{1}{2N} \sum_{\nu=0}^{2N-1} w_\nu e^{-ikx_\nu} \tag{2.2.48}$$

to obtain the pseudospectral approximation

$$\psi_N w = \sum_{k=-N}^{N}{}'' \tilde{w}(k) e^{ikx}. \tag{2.2.49}$$

Note: We now have only $2N$ pieces of discrete data at the different $2N$ grid points $x_0, x_1, \cdots, x_{2N-1}$ and they correspond to $2N$ waves, as we have a "silent" last mode, i.e., with $r = 0$, $k = N$, $Im[e^{ikx}]_{x=x_\nu} = i \sin \nu\pi = 0$. Thus $\psi_N w$ is well-defined; in view of (2.2.49) it is the unique interpolant of $w(x)$ at the $2N$ gridpoints $x = x_\nu$:

$$\begin{aligned} \psi_N w(x)|_{x=x_\mu} &= \sum_{k=-N}^{N}{}'' \left[\frac{1}{2N} \sum_{\nu=0}^{2N-1} w(x_\nu) e^{-ikx_\nu} \right] e^{ikx_\mu} \\ &= \sum_{\nu=0}^{2N-1} w(x_\nu) \cdot \frac{1}{2N} \sum_{k=-N}^{N}{}'' e^{ik(\mu-\nu)h} = w(x_\mu). \end{aligned} \tag{2.2.50}$$

The aliasing relation in this case reads – compare (2.2.7)

$$\tilde{w}(k) = \sum_{p=-\infty}^{\infty} e^{ip2Nr}\hat{w}(k+2pN) \tag{2.2.51}$$

and spectral convergence follows – compare with (2.2.16)

$$\|A_N w(x)\|_{H^s} \leq \text{Const}_s \cdot \|T_N w(x)\|_{H^s}, \quad s > \frac{1}{2}. \tag{2.2.52}$$

In the usual sin-cos formulation it takes the form

$$\psi_N w = \sum_{k=0}^{N}{}'' \tilde{a}_k \cos kx + \tilde{b}_k \sin kx, \quad \begin{bmatrix} \tilde{a}_k \\ \tilde{b}_k \end{bmatrix} = \frac{1}{N}\sum_{\nu=0}^{2N+1} w(x_\nu)\begin{bmatrix} \cos kx_\nu \\ \sin kx_\nu \end{bmatrix}, \quad 0 \leq k \leq N. \tag{2.2.53}$$

Noting that $\tilde{b}_N = 0$ we have $2N$ free parameters $\{\tilde{a}_0,\ \{\tilde{a}_k, \tilde{b}_k\}_{k=1}^{N-1},\ \tilde{a}_N\}$ to match our data at $\{x_\nu\}_{\nu=0}^{2N-1}$.

2.3 The (Pseudo)Spectral Fourier Expansions – Exponential Accuracy

We have seen that the spectral and the pseudospectral approximations enjoy what we called "spectral accuracy" – that is, the convergence rate is restricted solely by the global smoothness of the data. The statement about "infinite" order of accuracy for C^∞ functions is an *asymptotic* statement. Here we show that in the analytic case the error decay rate is in fact *exponential*.

To this end, assume that

$$w(z) = \sum_{k=-\infty}^{\infty} \hat{w}(k)e^{ikz}, \quad |Im\ z| \leq \eta < \eta_0, \tag{2.3.1}$$

is 2π-periodic analytic in the strip $-\eta_0 < Im\ z < \eta_0$. The error decay rate in both the spectral and pseudospectral cases is determined by the decay rate of the Fourier coefficients $\hat{w}(k)$. Making the change of variables $\zeta = e^{iz}$ we have for

$$v(\zeta) = w(z = +i\ell n\zeta), \tag{2.3.2}$$

the power series expansion

$$v(\zeta) = \sum_{k=-\infty}^{\infty} \hat{w}(k)\zeta^k. \tag{2.3.3}$$

By the periodic analyticity of $w(z)$ in the strip $|Imz| \leq \eta < \eta_0$, $v(\zeta)$ is found to be single-valued analytic in the corresponding annulus

$$e^{-\eta_0} < |\zeta| < e^{\eta_0}, \tag{2.3.4}$$

whose Laurent expansion is given in (2.3.3):

$$\hat{w}(k) = \frac{1}{2\pi i}\int_{|\zeta|=r} v(\zeta)\zeta^{-(k+1)}d\zeta, \quad e^{-\eta_0} < r < e^{\eta_0}. \tag{2.3.5}$$

This yields exponential decay of the Fourier coefficients

$$|\hat{w}(k)| \leq M(\eta)e^{-k\eta}, \quad M(\eta) = \max_{|Imz|\leq\eta} |w(z)|, \quad 0 < \eta < \eta_0. \tag{2.3.6}$$

We note that the inverse implication is also true; namely an exponential decay like (2.3.6) implies the analyticity of $w(z)$. Inserting this into (2.1.17) yields

$$\begin{aligned} \|w - S_N w\|^2 &= 2\pi \sum_{|k|>N} |\hat{w}(k)|^2 \leq \\ &\leq 2\pi \cdot M^2(\eta) \cdot \sum_{|k|>N} e^{-2k\eta} = 2\pi\frac{M^2(\eta)}{e^{2\eta}-1}\cdot e^{-2N\eta} \end{aligned} \tag{2.3.7}$$

and similarly for the pseudospectral approximation

$$\|w - \psi_N w\|^2 \leq \text{Const.} \frac{M^2(\eta)}{e^{2\eta} - 1} \cdot e^{-2N\eta}. \tag{2.3.8}$$

Note that in either case the exponential factor depends on the distance of the singularity (lack of analyticity) from the real line. For higher derivatives we likewise obtain

$$\|w - S_N w\|^2_{H^\sigma} + \|w - \psi_N w\|^2_{H^\sigma} \leq \text{Const.} N^{2\sigma} \cdot M^2(\eta) \cdot \frac{e^{-2N\eta}}{e^{2n} - 1}. \tag{2.3.9}$$

We can do even better, by taking into account higher derivatives, e.g.,

$$k\hat{w}(k) = \frac{1}{2\pi i} \int_{|\zeta|=r} \frac{dv}{d\zeta}(\zeta)\zeta^{-k} d\zeta, \tag{2.3.10}$$

so that with

$$M_s(\eta) = e^{2s\eta} \sum_{j=0}^{s} \max_{|\zeta|=e^\eta} |v^{(j)}(\zeta)|, \tag{2.3.11}$$

we have

$$k|\hat{w}(k)| \leq M_1(\eta) e^{-k\eta}, \tag{2.3.12}$$

and hence

$$\|w - S_N w\|^2_{H^\sigma} + \|w - \psi_N w\|^2_{H^\sigma} \leq \text{Const.} M^2_\sigma(\eta) \frac{e^{-2N\eta}}{e^{2\eta} - 1}. \tag{2.3.13}$$

2.4 The Non-Periodic Problem — The Chebyshev Expansion

We start by considering the second order Chebyshev ODE

$$-\sqrt{1-x^2}\frac{d}{dx}(\sqrt{1-x^2}\frac{d}{dx}\psi) = \lambda\psi(x), \quad -1 \leq x \leq 1. \tag{2.4.1}$$

This is a special case of the general Sturm-Liouville (SL) problem

$$L\psi = -\frac{1}{\omega(x)}\frac{d}{dx}\left(p(x)\frac{d\psi}{dx}\right) + \frac{1}{\omega(x)}q(x)\psi(x) = \lambda\psi(x), \quad p, q, \omega \geq 0. \tag{2.4.2}$$

Noting the Green identity

$$(L\psi, \phi)_{\omega(x)} = \int_a^b -(p\psi')'\phi + q\psi\phi = p(x)[\psi, \phi]|_a^b + (\psi, L\phi)_{\omega(x)}, \quad [\psi, \phi] \equiv \psi\phi' - \phi\psi', \tag{2.4.3}$$

we find that L is (formally) self-adjoint provided certain auxiliary conditions are satisfied. In the nonsingular case where $p(a) \cdot p(b) \neq 0$, we augment (2.4.2) with homogeneous boundary conditions,

$$\psi(a) = \phi(a) = 0, \quad \psi(b) = \phi(b) = 0. \tag{2.4.4}$$

Then L is self-adjoint in this case with a complete eigensystem $(\lambda_k, \psi_k(x))$: each $w(x) \epsilon L_{\omega(x)}[a, b]$ has the "generalized" Fourier expansion

$$w(x) \sim \sum_{k=0}^{\infty} \hat{w}(k)\psi_k(x), \quad \hat{w}(k) = \frac{(w(x), \psi_k(x))_\omega}{\|\psi_k(x)\|^2_\omega} \tag{2.4.5}$$

with Fourier coefficients

$$\hat{w}(k) = \frac{1}{\|\psi_k\|^2_\omega} \int_a^b w(x)\psi_k(x)\omega(x)dx. \tag{2.4.6}$$

The decay rate of the coefficients is algebraic: indeed

$$\begin{aligned}
\hat{w}(k) &= \frac{1}{\|\psi_k\|_\omega^2}\cdot\frac{1}{\lambda_k}(L\psi_k, w)_\omega = \\
&= \frac{1}{\|\psi_k\|_\omega^2}\cdot\frac{1}{\lambda_k}\left[p(x)\cdot[\psi_k, w]\right]_a^b + (\psi_k, Lw)_\omega\big] = \\
&= \frac{1}{\|\psi_k\|_\omega^2}\cdot\frac{1}{\lambda_k}\left[p(x)\cdot[\psi_k, w]\right]_a^b + \frac{1}{\lambda_k}(L\psi_k, Lw)_\omega\Big] = \ldots \\
&= \frac{1}{\|\psi_k\|_\omega^2}\left\{p(x)\cdot\sum_{j=0}^{s-1}\frac{1}{\lambda_k^{s+1}}[\psi_k, L^{(j)}w]\right]_a^b + \frac{1}{\lambda_k^j}(\psi_k, L^{(s)}w)_\omega\Big\}, \quad \psi_k'(x)|_{x=a,b} < \infty.
\end{aligned} \tag{2.4.7}$$

The asymptotic behavior of the eigenvalues for *nonsingular* SL problem is

$$\lambda_k \sim \left[\frac{\pi k}{\int_a^b\sqrt{\frac{\omega(x)}{p(x)}}dx}\right]^2 \sim \text{Const.}k^2$$

and hence, unless $w(x)$ satisfies an infinite set of boundary restrictions, we end with algebraic decay of $\hat{w}(k)$

$$\hat{w}(k) \sim \frac{1}{\|\psi_k\|_\omega^2}\cdot -\frac{p(x)}{\lambda_k}\cdot\psi_k'(x)w(x)\Big]_a^b \sim \frac{\text{Const.}}{k^2}.$$

This leads to algebraic convergence of the corresponding spectral and pseudospectral projections.

In contrast, the singular case is characterized by, $p(a) = p(b) = 0$; in this case L is self-adjoint independent of the boundary conditions (since the Poisson brackets $[\ ,\ ]$ drop), and we end up with the spectral decay estimate — compare (2.1.22)

$$\hat{w}(k) = \frac{1}{\|\psi_k\|_\omega^2}\cdot\frac{1}{\lambda_k^s}\cdot(\psi_k, L^{(s)}w)_\omega \le \frac{1}{\lambda_k^s}\frac{\|L^{(s)}w\|_\omega}{\|\psi_k\|_\omega}; \tag{2.4.8}$$

Thus, the decay of $\hat{w}(k)$ is as rapid as the smoothness of $w(x)$ permits.

As a primary example for this category of singular SL problems we consider the Jacobi equation associated with weights of the form $(1-x)^\alpha(1+x)^\beta,\ \alpha,\beta > -1$,

$$-\frac{d}{dx}\left((1-x^2)\omega(x)\frac{d\psi}{dx}\right) = \lambda\omega(x)\psi(x), \quad \omega = (1-x)^\alpha(1+x)^\beta,\ -1 \le x \le 1. \tag{2.4.9}$$

We now focus our attention on the Chebyshev-SL problem (2.4.1) corresponding to $\alpha = \beta = -1/2$. The transformation

$$x = \cos\theta, \quad \frac{d}{dx} = \frac{1}{\frac{dx}{d\theta}}\cdot\frac{d}{d\theta} = -\frac{1}{\sqrt{1-x^2}}\frac{d}{d\theta} \tag{2.4.10}$$

yields

$$-\frac{d^2}{d\theta^2}\phi(\theta) = \lambda\phi(\theta), \quad \phi(\theta) \equiv \psi(\cos\theta), \tag{2.4.11}$$

and we obtain the two sets of eigensystems

$$(\lambda_k = k^2, \phi_k = \cos k\theta), \tag{2.4.12}$$

and

$$(\lambda_k = k^2, \phi_k = \sin k\theta).$$

The second set violates the boundedness requirement which we now impose

$$|\psi_k'(\pm 1)| \le \text{ Const.}, \tag{2.4.13}$$

and so we are left with

$$(\lambda_k = k^2, \psi_k(x) = \cos(k \cos^{-1} x)). \tag{2.4.14}$$

The trigonometric identity

$$\cos(k+1)\theta = 2\cos\theta\cos k\theta - \cos(k-1)\theta$$

yields the recurrence relation

$$\psi_{k+1}(x) = 2x\psi_k(x) - \psi_{k-1}(x), \quad \psi_0(x) \equiv 1, \psi_1(x) = x, \tag{2.4.15}$$

hence, $\psi_k(x)$ are polynomials of degree k – these are the Chebyshev polynomials

$$T_k(x) = \cos(k \cos^{-1} x) \tag{2.4.16}$$

which are orthonormal w.r.t. Chebyshev weight $\omega(x) = (1-x^2)^{-\frac{1}{2}}$,

$$(T_k(x), T_j(x))_\omega = \int_{-1}^{1} \frac{T_k(x)T_j(x)}{\sqrt{1-x^2}}dx = \begin{cases} 0 & j \neq k, \\ \|T_k\|_\omega^2 = \frac{\pi}{2} & j = k > 0, \\ \|T_0\|_\omega^2 = \pi & j = k = 0. \end{cases} \tag{2.4.17}$$

In analogy with what we had done before, we consider now the Chebyshev-Fourier expansion

$$w(x) \sim \sum_{k=0}^{\infty} \hat{w}(k) T_k(x), \quad \hat{w}(k) = \frac{(w(x), T_k(x))_\omega}{\|T_k\|_\omega^2}. \tag{2.4.18}$$

To get rid of the factor $\frac{1}{2}$ for $k = 0$ we may also write this as

$$\begin{aligned} w(x) &\sim {\sum_{k=0}^{\infty}}' \hat{w}(k) T_k(x), \\ \hat{w}(k) = \frac{(w(x), T_k(x))_\omega}{\pi/2} &= \frac{2}{\pi}\int_{-1}^{1} \frac{w(x)\cos(k\cos^{-1}x)dx}{\sqrt{1-x^2}} = \frac{2}{\pi}\int_{\xi=0}^{\pi} w(\cos\xi)\cos k\xi \; d\xi. \end{aligned} \tag{2.4.19}$$

Thus, we go from the interval $[-1,1]$ into the 2π-periodic circle by *even extension*, with Fourier expansion of $w(\cos\theta)$, compare (2.1.9),

$$\hat{w}(k) = \frac{1}{\pi}\int_{\xi=0}^{2\pi} w(\cos\xi)\cos k\xi d\xi = \frac{2}{\pi}\int_{\xi=0}^{\pi} w(\cos\xi)\cos k\xi d\xi.$$

Another way of writing this employs a symmetric doubly infinite Fourier-like summation, where

$$w(x) \sim \frac{1}{2}\sum_{k=-\infty}^{\infty} \hat{w}(k) T_k(x) \tag{2.4.20}$$

with $T_{-k}(x) \equiv T_k(x)$ and

$$\hat{w}(k) = \frac{2}{\pi}\int_{-1}^{1} \frac{w(x)T_k(x)}{\sqrt{1-x^2}}dx, \quad -\infty < k < \infty. \tag{2.4.21}$$

The Parseval identity reflects the completeness of this system

$$\begin{aligned} \|w(x)\|_T^2 \equiv \int \frac{w^2(x)dx}{\sqrt{1-x^2}} &= \frac{1}{4}\left[\pi|\hat{w}(0)|^2 + \frac{\pi}{2}\sum_{k\neq 0}|\hat{w}(k)|^2\right] \\ &= \frac{\pi}{4}\sum_{k=0}^{\infty}|\hat{w}(k)|^2 \end{aligned} \tag{2.4.22}$$

which yields the error estimate

$$\|w - S_N w\|_T^2 = \frac{\pi}{4} \sum_{k>N} |\hat{w}(k)|^2 .$$

2.4.1 Spectral accuracy

In order to measure the spectral convergence of Chebyshev expansion, we have to estimate the decay rate of Chebyshev coefficients in terms of the smoothness of $w(x)$ and its derivatives; to this end we need Sobolev like norms. Unlike the Fourier case, $\{T_k(x)\}$ is not complete with respect to H^s – orthogonality is lost because of the Chebyshev weight. So we can proceed formally as before, see (2.1.24),

$$\|w - S_N w\|_T^2 = 2\pi \sum_{k>N} |\hat{w}(k)|^2 \leq \sum_{k>N} \frac{(1+|k|^2)^s}{(1+N^2)^s} |\hat{w}(k)|^2 \tag{2.4.23}$$

i.e., if we define the Chebyshev-Sobolev norm

$$\|w\|_{H_T^s}^2 = \sum_{k=0}^{\infty} (1+|k|^2)^s |\hat{w}(k)|^2,$$

then we have spectral accuracy

$$\|w - S_N w\|_T \leq \text{ Const}_s \cdot \frac{1}{N^s}, \quad w \epsilon H_T^s[-1,1].$$

In fact the H_T^s space can be derived from an appropriate inner product in the real space as done in Fourier expansion. The correct inner product — expressed in terms of $L = -\sqrt{1-x^2}\frac{d}{dx}(\sqrt{1-x^2}\frac{d}{dx}\cdot)$, is given by (in analogous manner to (2.1.19))

$$(w_1, w_2)_{H_T^{2s}} = \sum_{p=0}^{s} (L^p w_1, L^p w_2)_T \underbrace{=}_{x=cos\theta} \sum_{p=0}^{s} \int_{\theta=0}^{2\pi} \frac{d^{2p}}{d\theta^{2p}} w_1(\cos\theta) \frac{d^{2p}}{d\theta^{2p}} w_2(\cos\theta) d\theta, \tag{2.4.24}$$

so that

$$(T_k, T_j)_{H_T^{2s}} = \begin{cases} 0 & j \neq k, \\ \\ \frac{\pi}{2} \sum_{p=0}^{s} k^{4p}, & j = k \text{ (with } \pi \text{ factor at } j = k = 0). \end{cases} \tag{2.4.25}$$

Hence the Fourier coefficients in this Hilbert space behave like

$$(w(x), T_k)_{H_T^{2s}} \sim \sum_{k=0}^{\infty} (1+k^2)^{2s} \hat{w}(k), \tag{2.4.26}$$

and the corresponding norm is equivalent to

$$\|w\|_{H_T^{2s}}^2 \sim \sum_{k=0}^{\infty} (1+k^2)^{2s} |\hat{w}(k)|^2. \tag{2.4.27}$$

The reason for the squared factors here is due to the fact that L is a *second order* differential operator, unlike the first-order $D = \frac{d}{dx}$ in the Fourier case, i.e.,

$$\sum_{k=0}^{\infty} (1+|k|^2)^{2s} |\hat{w}(k)|^2 \sim \sum_{p=0}^{s} \|L^p w\|_T^2 \tag{2.4.28}$$

involves the first 2s-derivatives of $w(x)$ – appropriately weighted by Chebyshev weight. This completes the analogy with the Fourier case, and enables us to estimate derivative as well-compare (2.1.28),

$$\|w - S_N w\|_{H_T^\sigma} \leq \text{ Const}_s \frac{1}{N^{s-\sigma}}, \quad \sigma \leq s, \quad w \epsilon H_T^s[-1,1]. \tag{2.4.29}$$

2.5 The Non-Periodic Problem — The Chebyshev Interpolant

Next, let's discuss the discrete setup. Since we seek an *even* extension of the upper semi-circle we consider the case of *even* number of grid points – equally distributed along the unit circle. There are two main choices: one choice is to consider only the *interior* points, $\theta_\nu = (\nu + \frac{1}{2})\frac{\pi}{N+1}$ (here, $h = \frac{\pi}{N+1}$ and $r = \frac{h}{2}$). This yields the so called Gauss-Chebyshev points, consult §2.5.1 below,

$$x_\nu = \cos(\frac{(\nu + 1/2)\pi}{N+1}),\ \nu = 0, 1, \ldots, N. \tag{2.5.1}$$

The second choice takes into account also the ± 1-boundaries, considering $\theta_\nu = \nu\frac{\pi}{N}$ (here, $h = \frac{2\pi}{2N}$ and $r = 0$), which yield the so called Gauss-Lobatto-Chebyshev points – consult §2.5.2 below,

$$x_\nu = \cos(\frac{\nu\pi}{N}), \quad \nu = 0, 1, \cdots, N. \tag{2.5.2}$$

2.5.1 Chebyshev interpolant at Gauss gridpoints

We consider the Chebyshev-Fourier expansion, (2.4.19)

$$w(x) \sim {\sum_{k=0}^{\infty}}' \hat{w}(k) T_k(x), \quad \hat{w}(k) = \frac{2}{\pi}\int_{-1}^{1} \frac{w(x)T_k(x)dx}{\sqrt{1-x^2}}. \tag{2.5.3}$$

We want to collocate the Chebyshev-Fourier coefficients at the Gauss quadrature points. Here we invoke the

Gauss quadrature rule. Let $\phi_k(x)$ be an orthogonal family of k-degree polynomials in $L^2_\omega[-1,1]$, where $\omega(x) = (1-x)^\alpha(1+x)^\beta$ with $\alpha, \beta > -1$ [4]. Let $-1 < x_1 < x_2 < \ldots x_N < 1$ be the N zeros of $\phi_N(x)$. Then, there exist positive weights, $\{\omega = \omega^G\}_{j=1}^N$ such that for all polynomials $p(x)$ of degree$\leq 2N-1$ we have

$$\int_{-1}^{1} \omega(x)p(x)dx = \sum_{j=1}^{N} \omega_j p(x_j), \quad \omega_j = \omega_j^G. \tag{2.5.4}$$

Remark. To compute the Gauss weights we set $p(x) = \frac{\phi_N(x)}{x-x_k}$ in (2.5.4). Since $p(x_j) = 0\ \forall j \neq k$, (2.5.4) yields

$$\omega_k = \frac{1}{\phi_N'(x_k)}\int_{-1}^{1} \omega(x)\frac{\phi_N(x)}{x-x_k}dx, \quad 1 \leq k \leq N. \tag{2.5.5}$$

Equivalently, the corresponding weights are given by

$$\omega_j = \frac{-A_{N+1}\|\phi_N\|^2_{\omega(x)}}{A_N\phi_{N+1}(x_j)\phi_N'(x_j)}, \quad \phi_N(x) = A_N x^N + \ldots, \quad j = 1, 2, \ldots, N. \tag{2.5.6}$$

To verify (2.5.4) we express $p(x)$ as $p(x) = t(x)\phi_N(x) + r(x)$ for some $(N-1)$-degree polynomials, $t(x)$ and $r(x)$. The choice of weights in (2.5.5) guarantees that (2.5.4) is valid for all polynomials of degree$\leq N-1$, since the latter are spanned by $\{\frac{\phi_N(x)}{x-x_k}\}_{k=1}^N$. This, together with the fact that $\phi_N(x)$ is $L^2_{\omega(x)}$-orthogonal to all polynomials of degree$\leq N-1$, implies

$$\int_{-1}^{1} \omega(x)p(x)dx = \int_{-1}^{1} \omega(x)r(x)dx = \sum_{j=1}^{N} \omega_j r(x_j) = \sum_{j=1}^{N} \omega_j p(x_j). \tag{2.5.7}$$

Example. The N-degree Gauss-Chebyshev quadrature rule (based on the $N+1$ collocation points, $x_\nu = \cos(\frac{\nu+1/2)\pi}{N+1}),\ \nu = 0, 1, \ldots, N$) reads

$$\int_{-1}^{1} \frac{f(x)dx}{\sqrt{1-x^2)}} = \frac{\pi}{N+1}\sum_{\nu=0}^{N} f(x_\nu) + E, \quad x_\nu := \cos(\frac{(\nu+\frac{1}{2})\pi}{N+1}), \nu = 0, 1, \ldots, N, \tag{2.5.8}$$

[4] $\alpha = \beta = -1/2$ correspond to Chebyshev family, $\alpha = \beta = 0$ correspond to Legendre, etc.

with an error term, $E = \frac{2\pi}{2^{2N+2}(2N+2)!} f^{(2N+2)}(\eta)$, which vanishes for all polynomials of degree$\leq 2N+1$. Applying the latter to the Fourier-Chebyshev coefficients in (2.5.3) we arrive at discrete Chebyshev coefficients, $\tilde{w}(k)$ which yield

$$\psi_N w(x) = \sum_{k=0}^{N}{}'\tilde{w}(k)T_k(x), \quad \tilde{w}(k) = \frac{2}{N+1}\sum_{\nu=0}^{N} w(x_\nu)T_k(x_\nu). \tag{2.5.9}$$

We claim that $\psi_N w(x)$ is the N-degree algebraic interpolant of $w(x)$ at Chebyshev points $\{x_\nu\}_{\nu=0}^N$. To see this we employ the

Christoffel-Darboux identity. There holds

$$\sum_{k=0}^{N}\frac{\phi_k(x)\phi_k(y)}{\|\phi_k(x)\|_\omega^2} = \frac{A_{N+1}}{A_N\|\phi_N(x)\|_\omega^2}\frac{\phi_{N+1}(x)\phi_N(y)-\phi_N(x)\phi_{N+1}(y)}{x-y}. \tag{2.5.10}$$

We omit the straightforward proof of the general case (— which is based on the three step recurrence relations for orthogonal polynomials), and concentrate on the Chebyshev expansion in which case Christoffel-Darboux formula reads

$$\sum_{k=0}^{N}{}'T_k(x)T_k(y) = \frac{T_{N+1}(x)T_N(y)-T_N(x)T_{N+1}(y)}{2(x-y)}. \tag{2.5.11}$$

Using this we find that $\psi_N w(x)$ interpolates $w(x)$ at Chebyshev points as asserted. Indeed we have

$$\begin{aligned}
\psi_N w(x_i) &= \sum_{k=0}^{N}{}'\frac{2}{N}\sum_{\nu=0}^{N} w(x_\nu)T_k(x_\nu)T_k(x_i) = \\
&= \frac{2}{N+1}\sum_{\nu=0}^{N} w(x_\nu)\sum_{k=0}^{N}{}'T_k(x_\nu)T_k(x_i) = \\
&= \frac{2}{N+1}\sum_{\nu=0}^{N} w(x_\nu)\left\{\begin{array}{ll}\dfrac{T_{N+1}(x_\nu)T_N(x_i)-T_N(x_\nu)T_{N+1}(x_i)}{2(x_\nu-x_i)} = 0, & \nu\neq i\\ \dfrac{T'_{N+1}(x_i)T_N(x_i)}{2} = \dfrac{N+1}{2}, & \nu = i\end{array}\right\} = w(x_i).
\end{aligned} \tag{2.5.12}$$

We want to estimate the error between $w(x)$ and its Chebyshev interpolant $\psi_N w(x)$. As in the periodic Fourier case, we use here the *aliasing relation*

$$\tilde{w}(k) = \sum_{p=-\infty}^{\infty}\hat{w}(k+2p(N+1)), \tag{2.5.13}$$

which follows from the straightforward computation. One concludes that the aliasing errors are dominated by the spectrally small truncation error (2.4.29), and spectral convergence follows.

2.5.2 Chebyshev interpolant at Gauss–Lobatto gridpoints

The starting point is the Gauss-Lobatto quadrature rule. We make a short intermezzo on this issue. If $\{\phi_k\}$ is an L^2_ω-orthogonal family of k-degree polynomials, then by utilizing [5] Jacobi equation (2.4.9), one finds that $\{\phi'_{k+1}\}$ is k-degree family which is orthogonal with respect to the weight $(1-x^2)\omega(x)$. Applying Gauss rule to the latter we find that there exist discrete gauss weights ω_j^G such that

$$\int_{-1}^{1}(1-x^2)w(x)r(x)dx = \sum_{j=1}^{N} w_j^G r(x_j), \text{ for all } r\epsilon\pi_{2N-1}.$$

[5] Utilizing = integration by parts in this case.

This is in fact a special case of the Gauss-Lobatto-Jacobi quadrature rule which is exact for all $p\epsilon\pi_{2N+1}$. Indeed, all such p's can be expressed as $p(x) = (1-x^2)r(x) + \ell(x)$ with $r(x)$ in π_{2N-1}, and a linear $\ell(x) = p(-1)\frac{1-x}{2} + p(1)\frac{1+x}{2}$. The last equality tells us that

$$\begin{aligned}\int_{-1}^{1} w(x)p(x)dx &= \sum_{j=1}^{N} w_j^G r(x_j) + \int_{-1}^{1} w(x)\ell(x) = \\ &= \sum_{j=1}^{N} \frac{w_j^G}{1-x_j^2} p(x_j) + \int_{-1}^{1} w(x)\ell(x) - \sum_{j=1}^{N} \frac{w_j^G}{1-x_j^2}\ell(x_j) = I + II + III.\end{aligned}$$

Thus, we have

$$I = \sum_{j=1}^{N} w_j^L p(x_j), \quad w_j^L \equiv \frac{w_j^G}{1-x_j^2}, \tag{2.5.14}$$

and the two expressions, $II + III$, amount to a linear combination of $p(-1)$ and $p(1)$,

$$II + III = w_0^L p(x_0) + w_{N+1}^L p(x_{N+1}), \quad x_0 \equiv -1 < x_1 < \ldots < x_N < 1 \equiv x_{N+1}. \tag{2.5.15}$$

We conclude with

Gauss-Lobatto quadrature rule. Let $\phi_k(x)$ be an orthogonal family of k-degree polynomials in $L^2_\omega[-1,1]$, where $\omega(x) = (1-x)^\alpha(1+x)^\beta$ with $\alpha, \beta > -1$ Let $-1 = x_0 < x_1 < x_2 < \ldots x_N < x_{N+1} = 1$ be the $N+2$ extrema of $\phi_{N+1}(x)$. Then, there exist positive weights $\{w_j = w_j^L\}_{j=0}^{N+1}$ such that

$$\int_{-1}^{1} w(x)p(x)dx = \sum_{j=0}^{N+1} w_j p(x_j), \quad \text{for all } p\epsilon\pi_{2N+1}, \quad \omega_j = \omega_j^L. \tag{2.5.16}$$

Example. The Gauss-Lobatto-Chebyshev quadrature rule (corresponding to $\omega(x) = \sqrt{1-x^2}$ and $x_\nu = \cos(\nu h), \nu = 0, 1, \ldots, N$) is nothing but the familiar trapezoidal rule — indeed starting with (2.4.19), we have

$$\hat{w}(k) = \frac{2}{\pi}\int_{\xi=0}^{\pi} w(\cos\xi)\cos k\xi d\xi \to \frac{2}{\pi}{\sum_{\nu=0}^{N}}'' w(\cos\theta_\nu)\cos k\theta_\nu \cdot \frac{\pi}{N}, \tag{2.5.17}$$

and we end up with the discrete Chebyshev coefficients

$$\tilde{w}(k) = \frac{2}{N}{\sum_{\nu=0}^{N}}'' w_\nu T_k(x_\nu), \quad 0 \le k \le N. \tag{2.5.18}$$

This corresponds to the Fourier interpolant with an even number of equidistant gridpoints θ_ν (consult (2.2.48)), for

$$\begin{aligned}\tilde{w}(k) = \frac{1}{2\pi}{\sum_{\nu=0}^{2N}}'' w_\nu e^{-ik\theta_\nu}h &= \frac{1}{\pi}\sum_{\nu=0}^{N} w_\nu\left[e^{-ik\theta_\nu} + e^{ik\theta_\nu}\right]\frac{2\pi}{2N} = \\ &= \frac{2}{N}{\sum_{\nu=0}^{N}}'' w_\nu\cos(k\theta_\nu).\end{aligned}$$

Then one may construct the Chebyshev interpolant at these $N+1$ gridpoints

$$\psi_N w(x) = {\sum_{k=0}^{N}}'' \tilde{w}(k)T_k(x). \tag{2.5.19}$$

We have an identical aliasing relation (compare (2.2.51)),

$$\tilde{w}(k) = \sum_{p=-\infty}^{\infty} \hat{w}(k+2pN). \tag{2.5.20}$$

(Verification: insert the Chebyshev expansion evaluated at x_ν into (2.5.18),

$$\tilde{w}(k) = \frac{2}{N}\sum_{\nu=0}^{N}{}'' \left[\sum_{j=0}^{\infty}{}' \hat{w}(j) T_j(x_\nu)\right] T_k(x_\nu) = \frac{2}{N}\sum_{j=0}^{\infty}{}' \hat{w}(j) \left[\sum_{\nu=0}^{N}{}'' T_j(x_\nu) T_k(x_\nu)\right];$$

to calculate the summation on the right we employ the identity $2T_j(x)T_k(x) \equiv T_{j+k}(x) + T_{|j-k|}(x)$ which yields

$$\tilde{w}(k) = \sum_{j=0}^{\infty}{}' \hat{w}(j) \left[\delta_{jk} + \delta_{j0}\delta_{k0} + \sum_{p=1}^{\infty} \delta_{j,2pN\pm k}\right],$$

and (2.5.20) follows.) The spectral Chebyshev estimate (2.4.29) together with the aliasing relation (2.5.20) yield the ψdospectral convergence estimate, (compare (2.2.17))

$$\|w(x) - \psi_N w(x)\|_{H_T^\sigma} \leq \text{Const}_s \cdot \frac{1}{N^{s-\sigma}}, \quad w \epsilon H_T^s, \ s \geq \sigma, \tag{2.5.21}$$

where $\text{Const}_s \sim \|w\|_{H_T^s}$.

Example: We have the Sobolev embedding of $H_T^\sigma \subset L^\infty$ with $\sigma > 1/2$,

$$\begin{aligned} |w(x)| \leq \frac{1}{2}\sum_{k=-\infty}^{\infty} |\hat{w}(k)| &\leq \frac{1}{2}\left(\sum_k (1+k^2)^\sigma |\hat{w}(k)|^2 \cdot \sum_k \frac{1}{(1+k^2)^\sigma}\right)^{\frac{1}{2}} \\ &\leq \text{Const}_\sigma \cdot \|w\|_{H_T^\sigma}, \quad \sigma > \tfrac{1}{2}. \end{aligned} \tag{2.5.22}$$

Consequently,

$$\max_x |w(x) - \psi_N w(x)| \leq \text{Const}_s \cdot \frac{1}{N^{s-\sigma}}, \quad \text{Const}_s \sim \|w\|_{H_T^s}, \quad s \geq \sigma > \frac{1}{2}.$$

In particular, with $s = N+1$ we obtain an improved estimate[6] for the near min-max approximation collocated at $x_\nu = \cos\left(\left(\nu + \frac{1}{2}\right)\frac{\pi}{N}\right)$,

$$\max_x |w(x) - \psi_N w(x)| \leq \text{Const.} \|w\|_{H_T^{N+1}} \cdot \frac{e^{-N}}{(N+1)!}.$$

2.5.3 Exponential convergence of Chebyshev expansions

We briefly mention the exponential convergence in the analytic case. To this end we employ Bernstein's regularity ellipse, E_r, with foci ± 1 and sum of its semi axis $= r$. Denoting

$$M(\eta) = \max_{z \epsilon E_r} |w(z)|, \quad r = e^\eta. \tag{2.5.23}$$

We have

Theorem 2.1 Assume $w(x)$ is analytic in *[-1,1]* with regularity ellipse whose sum of semiaxis $= r_0 = e^{\eta_0} > 1$. Then

$$\|w(x) - \psi_N w(x)\|_{H^\sigma}^2 + \|w(x) - S_N w(x)\|_{H^\sigma}^2 \leq \text{Const.} \frac{M^2(\eta)}{e^{2\eta} - 1} \cdot N^{2\sigma} e^{-2N\eta}.$$

[6] This should be compared with the straightforward 'familiar' bound $\|w^{(N+1)}\|_{L^\infty} \frac{2^{-N}}{(N+1)!}$.

Proof: The transformation $z = (\zeta + \zeta^{-1})/2$ takes E_{r_0} from the z-plane into the annulus $r_0^{-1} < |\zeta| < r_0$ in the ζ-plane. Hence, $v(\zeta) = 2w\left(z = \frac{\zeta+\zeta^{-1}}{2}\right)$ admits the power expansion

$$v(\zeta) = 2w\left(\frac{\zeta+\zeta^{-1}}{2}\right) = \sum_{k=-\infty}^{\infty} \hat{w}(k)\zeta^k, \quad r_0^{-1} < |\zeta| < r_0 = e^{\zeta_0}; \tag{2.5.24}$$

indeed, setting $\zeta = e^{i\theta}$ and recalling $\hat{w}(-k) = \hat{w}(k)$, the above expansion clearly describes the real interval [-1,1]

$$w(z = \cos\theta) = \sum_{k=0}^{\infty}{}' \hat{w}(k)\cos k\theta. \tag{2.5.25}$$

Using the Laurent expansion in (2.5.24)

$$\hat{w}(k) = \frac{1}{2\pi i}\int_{|\zeta|=r} \frac{\nu(\zeta)}{\zeta^{k+1}} d\zeta, \quad e^{-\eta_0} < r < e^{\eta_0}, \tag{2.5.26}$$

hence

$$|\hat{w}(k)| \le M(\eta)e^{-k\eta} \tag{2.5.27}$$

and the result follows along the lines of (2.3.7)-(2.3.8).

2.5.4 Chebyshev differentiation matrix

We conclude with a discussion on Chebyshev differencing. Starting with grid values w_ν at Chebyshev points $x_\nu = \cos\left(\nu\frac{\pi}{N}\right)$, one constructs the Chebyshev interpolant

$$\psi_N w(x) = \sum_{k=0}^{N}{}'' \tilde{w}(k)T_k(x), \quad \tilde{w}(k) = \frac{2}{N}\sum_{\nu=0}^{N}{}'' w_\nu \cos(k\cos^{-1}x_\nu). \tag{2.5.28}$$

One can compute $\tilde{w}(k), 0 \le k \le N$, efficiently via the cos-FFT with $\mathcal{O}(N\log N)$ operations. Next, we differentiate in Chebyshev space

$$\frac{d}{dx}\psi_N w(x) = \sum_{k=0}^{N}{}'' \tilde{w}(k)\frac{d}{dx}T_k(x). \tag{2.5.29}$$

In this case, however, $T_k(x)$ is not an eigenfunction of $\frac{d}{dx}$; instead $\frac{d}{dx}T_k(x)$ – being a polynomial of degree $\le k-1$, can be expressed as a linear combination of $\{T_j(x)\}_{j=0}^{k-1}$ (in fact $T_k(x)$ is even/odd for even/odd $k's$): with $c_0 = 2, c_{k>0} = 1$ we obtain

$$\frac{d}{dx}T_k(x) = k\sum_{\substack{0\le j<k \\ k-j \text{ odd}}} \frac{2}{c_j}T_j(x), \tag{2.5.30}$$

and hence

$$\frac{d}{dx}\psi_N w(x) = \sum_{k=0}^{N}{}'' k\tilde{w}(k)\sum_{\substack{0\le j<k \\ k-j \text{ odd}}} \frac{2}{c_j}T_j(x). \tag{2.5.31}$$

Rearranging we get (here, $\sum'$ indicates halving the *last* term)

$$\frac{d}{dx}\psi_N w(x) = \sum_{k=0}^{N-1} \tilde{w}'(k)T_k(x), \quad \tilde{w}'(k) = \frac{2}{c_k}\sum_{\substack{p\ge k+1 \\ p+k \text{ odd}}}^{N}{}' p\tilde{w}(p) \tag{2.5.32}$$

and similarly for the second derivative

$$\tilde{w}''(k) = \frac{2}{c_k}\sum_{\substack{p\ge k+2 \\ p+k \text{ even}}} p(p^2-k^2)\tilde{w}(p). \tag{2.5.33}$$

The amount of work to carry out the differentiation in this form is $\mathcal{O}(N^2)$ operations which destroys the $N \log N$ efficiency. Instead, we can employ the recursion relation which follows directly from (2.5.32)

$$\tilde{w}'(k+1) = \tilde{w}'(k-1) \cdot c_{k-1} - 2k\tilde{w}(k). \tag{2.5.34}$$

To see this in a different way we note that

$$\sin(k+1)\theta = \sin(k-1)\theta + 2\sin\theta\cos k\theta,$$

which leads to

$$\frac{1}{k+1}\frac{dT_{k+1}}{dx} = \frac{1}{k-1}\frac{dT_{k-1}}{dx} + 2T_k(x),$$

and hence

$$\begin{aligned}
\frac{d}{dx}\psi_N w(x) &= {\sum_{k=0}^{N}}'' k\tilde{w}(k)\frac{1}{k}T_k'(x) = \\
&= \frac{1}{2}{\sum_{k=0}^{N}}''(\tilde{w}'(k-1) - \tilde{w}'(k+1))\frac{1}{k}T_k'(x) = \longleftarrow \text{summation by parts} \\
&= \frac{1}{2}{\sum_{k=0}^{N}}''2\tilde{w}'(k)T_k(x) = {\sum_{k=0}^{N}}''\tilde{w}'(k)T_k(x)
\end{aligned}$$

as asserted. In general we have

$$\tilde{w}^{(s)}(k+1) = \tilde{w}^{(s)}(k-1)c_{k-1} - 2k\tilde{w}^{(s-1)}(k). \tag{2.5.35}$$

With this, $\tilde{w}(k)$ can be evaluated using $\mathcal{O}(N)$ operations, and the differentiated polynomial at the grid points is computed using another cos-FFT employing $\mathcal{O}(N \log N)$ operations

$$\frac{d}{dx}\psi_N w(x)|_{x=x_\nu} = {\sum_{k=0}^{N}}''\tilde{w}'(k)\cos kx_\nu, \tag{2.5.36}$$

with spectral/exponential error

$$\max_{x=x_\nu}|\frac{d}{dx}w(x) - \frac{d}{dx}\psi_N w(x)| \leq \begin{cases} \text{Const}_s \cdot \dfrac{1}{N^{s-\sigma}} & \dfrac{3}{2} < \sigma < s, \\ \text{Const}_\eta \cdot e^{-N\eta} & \end{cases}. \tag{2.5.37}$$

The matrix representation of Chebyshev differentiation, D_T, takes the almost antisymmetric form (here $c_k = 1$ except for $c_0 = c_N = 2$)

$$(D_T)_{j\,k} = \begin{cases} \dfrac{c_j}{c_k}\dfrac{(-1)^{j+k}}{x_j - x_k} & j \neq k, \\ -\dfrac{x_j}{2(1-x_j^2)} & j = k \neq (0, N), \\ \dfrac{2N^2+1}{6} & j = k = 0, \\ -\dfrac{2N^2+1}{6} & j = k = N. \end{cases}$$

3 THE FOURIER METHOD

3.1 The Spectral Fourier Approximation

We begin with the simplest hyperbolic equation – the scalar constant-coefficients wave equation

$$\frac{\partial u}{\partial t} = a\frac{\partial u}{\partial x} \tag{3.1.1}$$

subject to initial conditions

$$u(x,0) = f(x), \tag{3.1.2}$$

and periodic boundary conditions.

This *Cauchy problem* can be solved by the Fourier method: with $f(x) = \sum_{-\infty}^{\infty} \hat{f}(k)e^{ikx}$ we obtain after integration of (3.1.1),

$$\frac{\partial}{\partial t}\hat{u}(k,t) = ika\hat{u}(k,t), \tag{3.1.3}$$

with solution

$$\hat{u}(k,t) = e^{ikat}\hat{f}(k), \tag{3.1.4}$$

and hence

$$u(x,t) = \sum_k e^{ikat}\hat{f}(k)e^{ikx} = \sum_k \hat{f}(k)e^{ik(x+at)} = f(x+at). \tag{3.1.5}$$

Thus the solution operator in this case amounts to a simple translation

$$E(t,\tau)u(x,\tau) = u(x+a(t-\tau),t), \quad \|E(t,\tau)\| = 1. \tag{3.1.6}$$

This is reflected in the Fourier space, see (3.1.4), where each of the Fourier coefficients has the same change in phase and no change in amplitude; in particular, therefore, we have the a priori energy bound (conservation)

$$\|u(\cdot,t)\|^2 = 2\pi\sum_k |\hat{u}(k,t)|^2 = 2\pi\sum_k |\hat{f}(k)|^2 = \|f(\cdot)\|^2. \tag{3.1.7}$$

We want to solve this equation by the spectral Fourier method. To this end we shall approximate the spectral Fourier projection of the exact solution $Su_N \equiv S_N u(x,t)$. Projecting the equation (3.1.1) into the N-space we have

$$\frac{\partial u_N}{\partial t} = S_N\left[a\frac{\partial u}{\partial x}\right]. \tag{3.1.8}$$

Since S_N commutes with multiplication by a constant and with differentiation we can write this as

$$\frac{\partial u_N}{\partial t} = a\frac{\partial u_N}{\partial x}. \tag{3.1.9}$$

Thus $u_N = S_N u$ satisfies the same equation as the exact solution does, subject to the approximate initial data

$$u_N(t=0) = S_N f. \tag{3.1.10}$$

The resulting equations amount to $2N+1$ ordinary differential equations (ODEs) for the amplitudes of the projected solution

$$\frac{d}{dt}\hat{u}_N(k,t) = ika\hat{u}_N(k,t), \quad -N \le k \le N, \tag{3.1.11}$$

subject to the initial conditions

$$\hat{u}_N(k,0) = \hat{f}(k). \tag{3.1.12}$$

Since these equations are independent of each other, we can solve them directly, obtaining

$$\hat{u}_N(k,t) = e^{ikat}\hat{f}(k) \tag{3.1.13}$$

and the approximate solution takes the form

$$u_N(x,t) = \sum_{k=-N}^{N} \hat{f}(k) e^{ik(x+at)}. \tag{3.1.14}$$

Hence, the approximate solution $u_N(x,t) = f_N(x+at)$ satisfies

$$u(x,t) - u_N(x,t) = E(t,0)f(x) - E(t,0)S_N f(x) \tag{3.1.15}$$

and therefore, it converges spectrally to the exact solution, compare (2.1.26),

$$\begin{aligned} \|u(t) - u_N(t)\| &\leq \|E(t,0)(I - S_N)f(x)\| \leq \\ &\leq \|(I - S_N)f(x)\| \leq \text{Const}\|f\|_{H^s} \cdot \frac{1}{N^s}. \end{aligned} \tag{3.1.16}$$

Similar estimates holds for higher Sobolev norms; in fact if the initial data is analytic then the convergence rate is exponential. In this case the only source of error comes from the initial data, that is we have the error equation

$$\frac{\partial}{\partial t}[u - u_N] = a\frac{\partial}{\partial x}[u - u_N] \tag{3.1.17}$$

subject to initial error

$$u - u_N(t=0) = f - f_N. \tag{3.1.18}$$

Consequently, we have the a priori estimate of this constant coefficient wave equation

$$\|u - u_N(t)\| \leq \text{Const}_T\|f - f_N\| \leq \text{Const.}\|f\|_{H^s} \cdot \frac{1}{N^s} \quad \text{Const}_T = 1. \tag{3.1.19}$$

Now let us turn to the scalar equation with variable coefficients

$$\frac{\partial u}{\partial t} = a(x,t)\frac{\partial u}{\partial x}, \quad a(x,t) = 2\pi - \text{periodic}. \tag{3.1.20}$$

This hyperbolic equation is well-posed: by the energy method we have

$$\frac{1}{2}\frac{d}{dt}\int_x u^2(x,t)dx = \overbrace{\int ua(x,t)u_x dx}^{=\int a(x,t)(u^2/2)_x dx} = -\frac{1}{2}\int a_x(x,t)u^2(x,t)dx, \tag{3.1.21}$$

and hence

$$\|u(x,t)\|_{L^2(x)} \leq \text{Const}_T \cdot \|f(x)\| \tag{3.1.22}$$

with

$$\text{Const}_T = e^{MT}, \quad M = \max_{x,t}[-a_x(x,t)]. \tag{3.1.23}$$

In other words, we have for the solution operator

$$\|S(t,\tau)u(\tau)\|_{L^2(x)} \leq e^{M(t-\tau)}\|u(\tau)\|_{L^2(x)} \tag{3.1.24}$$

and similarly for higher norms. As before, we want to solve this equation by the spectral Fourier method. We consider the spectral Fourier projection of the exact solution $u_N = S_N u(x,t)$; projecting the equation (3.1.20) we get

$$\frac{\partial}{\partial t}u_N = S_N\left[a(x,t)\frac{\partial u}{\partial x}\right]. \tag{3.1.25}$$

Unlike the previous constant coefficients case, now S_N does not commute with multiplication by $a(x,t)$, that is, for arbitrary smooth function $\rho(x,t)$ we have (suppressing time dependence)

$$S_N a(x)\rho(x) = \sum_{k=-N}^{N}\left(\sum_{j=-\infty}^{\infty}\hat{a}(k-j)\hat{\rho}(j)\right)e^{ikx} \tag{3.1.26}$$

while

$$a(x)S_N\rho(x) = \sum_{k=-\infty}^{\infty}\left(\sum_{j=-N}^{N}\hat{a}(k-j)\hat{\rho}(j)\right)e^{ikx}. \tag{3.1.27}$$

Thus, if we exchange the order of operations we arrive at

$$\frac{\partial u_N}{\partial t} = a(x,t)\frac{\partial u_N}{\partial x} - [a(x,t)S_N - S_N a(x,t)]\frac{\partial u}{\partial x}. \tag{3.1.28}$$

While the second term on the right is not zero, this commutator between multiplication and Fourier projection is spectrally small, i.e.,

$$\begin{gathered}\|S_N a(x)\rho(x) - a(x)S_N\rho(x)\|_{L^2(x)} = \\ \|(S_N - I)a(x)\rho(x) + a(x)(I - S_N)\rho(x)\|_{L^2(x)} \leq \\ \leq \text{Const.}\|a(x)\rho(x)\|_{H^s}\cdot\frac{1}{N^s} + \text{Const.}\|a(x)\|_{L_\infty(x)}\cdot\|\rho(x)\|_{H^s}\cdot\frac{1}{N^s}\end{gathered} \tag{3.1.29}$$

and so we intend to neglect this spectrally small contribution and to set as an *approximate* model equation for the Fourier projection of $u(x,t)$

$$\frac{\partial v_N}{\partial t} = a(x,t)\frac{\partial v_N}{\partial x}. \tag{3.1.30}$$

The second term may lie outside the N-space, and so we need to project it back, thus arriving at our final form for the spectral Fourier approximation of (3.1.20)

$$\frac{\partial v_N}{\partial t} = S_N\left(a(x,t)\frac{\partial v_N}{\partial x}\right). \tag{3.1.31}$$

Again, we commit here a spectrally small deviation from the previous model, for

$$\|(I - S_N)a\rho(x)\|_{L^2(x)} \leq \text{Const}\|a(x)\rho(x)\|_{H^s}\cdot\frac{1}{N^s}. \tag{3.1.32}$$

The Fourier projection of the exact solution does not satisfy (3.1.22)-(3.1.23), but rather a near-by equation,

$$\frac{\partial u_N}{\partial t} = S_N\left(a(x,t)\frac{\partial u_N}{\partial x}\right) + F_N(x,t) \tag{3.1.33}$$

where the *local truncation error*, $F_N(x,t)$ is given by

$$F_N(x,t) = S_N\left[a(x,t)(I - S_N)\frac{\partial u}{\partial x}\right]. \tag{3.1.34}$$

The local truncation error is the amount by which the (projection of) the exact solution misses our approximate mode (3.1.31); in this case it is spectrally small by the errors committed in (3.1.29) and (3.1.19). More precisely we have

$$\|F_N(x,t)\|_{L^2(x)} \leq \|a(x,t)\|_{L^2(x)}\cdot\|u\|_{H^{s+1}}\frac{1}{N^s}, \tag{3.1.35}$$

depending on the degree of smoothness of the exact solution. We note that by hyperbolicity, the later is exactly the degree of smoothness of the initial data, i.e., by the hyperbolic differential energy estimate

$$\|F_N(x,t)\|_{L^2(x)} \leq \|a(x,t)\|_{L^2(x)}\cdot\|f\|_{H^{s+1}}\cdot\frac{1}{N^s} \tag{3.1.36}$$

and in the particular case of analytic initial data, the truncation error is exponentially small.

From this point of view, the spectral approximation (3.1.31) satisfies an evolution model which deviates by a spectrally small amount from the equation satisfied by the Fourier projection of the *exact* solution (3.1.33). This is in addition to the spectrally small error we commit initially, as we had before

$$v_N(t=0) = S_N f \equiv f_N. \tag{3.1.37}$$

3.1.1 Stability and convergence

We now raise the question of convergence. That is, whether the accumulation of spectrally small errors while integrating (3.1.31) rather than (3.1.33), give rise to an approximate solution $v_N(x,t)$ which is only spectrally away from the exact projection $u_N(x,t)$. We already know that the distance between $u_N(x,t)$ and the exact solution $u(x,t)$ – due to the spectrally small initial error – is spectrally small as we have seen in the previous constant coefficient case.

To answer this convergence question we have to require the *stability* of the approximate model (3.1.31). That is, we say that the approximation (3.1.31) is stable if it satisfies an a priori energy estimate analogous to the one we have for the differential equation

$$\|v_N(t)\| \leq \text{Const.} e^{Mt} \|v_N(0)\|. \tag{3.1.38}$$

Clearly, such a stability estimate is necessary in any computational model. Otherwise, the evolution model does not depend continuously on the (initial) data, and small rounding errors can render the computed solution useless. On the positive side we will show that the stability implies the spectral convergence of an approximate solution $u_N(x,t)$.[7] Indeed the error equation for $e_N(t) = u_N(t) - v_N(t)$ takes the form

$$\frac{\partial e_N}{\partial t} = S_N \left[a(x,t) \frac{\partial e_N}{\partial x} + F_N(x,t) \right]. \tag{3.1.39}$$

Let $E_N(t,\tau)$ denote the evolution operator solution associated with this approximate model. By the stability estimate (3.1.38),

$$\|E_N(t,\tau) v_N(\tau)\| \leq \text{Const} e^{M(t-\tau)} \|v_N(\tau)\|. \tag{3.1.40}$$

Hence, by (3.1.40) together with Duhammel's principle we get for the inhomogeneous error equation (3.1.39)

$$e_N(t) = E_N(t,0) e_N(0) + \int_{\tau=0}^{t} E_N(t,\tau) F_N(\tau) d\tau \tag{3.1.41}$$

and

$$\|e_N(t)\| \leq \text{Const.} e^{Mt} \left[\|e_N(0)\|_{L^2(x)} + \int_{\tau=0}^{t} \|F_N(x,\tau)\|_{L^2(x)} d\tau \right]. \tag{3.1.42}$$

In our case $e_N(0) = f_N - S f_N = 0$, and the truncation error $F_N(x,\tau)$ is spectrally small; hence

$$\|e_N \equiv u_N(t) - v_N(t)\| \leq \text{Const.} e^{Mt} \cdot \frac{1}{N^s} \tag{3.1.43}$$

where the constant depends on $\|a(x,t)\|_{L_\infty}$(!) and $\|f\|_{H^{s+1}}$, i.e., restricted solely by the smoothness of the data. In the particular case of analytic data we have exponential convergence

$$\|e_N(t) \equiv u_N(t) - v_N(t)\| \leq \text{Const.} e^{Mt} \cdot e^{-\eta N}. \tag{3.1.44}$$

Adding to this the error between $u_N(t)$ and $u(t)$ (– which is due to the spectrally small error in the initial data between f_N and f) we end up with

$$\|u(t) - v_N(t)\| \leq \text{Const.} e^{Mt} \cdot \begin{cases} \frac{1}{N^s} \text{ for } H^{s+1} \text{ initial data} \\ e^{-\eta N} \text{ for analytic initial data} \end{cases}. \tag{3.1.45}$$

To summarize, we have shown that our spectral Fourier approximation converges spectrally to the exact solution, *provided the approximation (3.1.31) is stable.*

Is the approximation (3.1.31) stable? That is, do we have the a priori estimate (3.1.38)? To show this we try to follow the steps that lead to the analogue estimate in the differential case, compare (3.1.21). Thus, we multiply (3.1.31) by $v_N(x,t)$ and integrate over the 2π-period, obtaining

$$\frac{1}{2} \frac{d}{dt} \int_x v_N^2(x,t) dx = + \int_x v_N(x,t) S_N \left(a(x,t) \frac{\partial v_N}{\partial x} \right) dx. \tag{3.1.46}$$

[7] We note that in the previous constant coefficient case, the approximate model coincides with the differential case, hence the stability estimate was nothing but the a priori estimate for the differential equation itself.

But $v_N(x,t)$ is orthogonal to $(I-S_N)\left[a(x,t)\frac{\partial v_N}{\partial x}\right]$ so adding this to the right-hand side of (3.1.46) we arrive at

$$\frac{1}{2}\frac{d}{dt}\int_x v_N^2(x,t) = \int_x v_N(x,t)a(x,t)\frac{\partial v_N}{\partial x}dx \tag{3.1.47}$$

and we continue precisely as before to conclude, similarly to (3.1.22)-(3.1.23), that the stability estimate (3.1.38) holds

$$\|v_N(t)\| \leq \text{Const.}e^{Mt}\|v_N(0)\|, \quad M = \max_{x,t}[-a_x(x,t)]. \tag{3.1.48}$$

In the constant coefficient case the Fourier method amounts to a system of $(2N+1)$ decoupled ODE's for the Fourier coefficients of $v_N = u_N$ which were integrated explicitly. Let's see what is the case with problems having variable coefficients say, for simplicity, $a \equiv a(x)$. Fourier transform (3.1.22)-(3.1.23) we obtain for $\hat{v}(k,t) = \hat{v}_N(k,t)$ – the kth-Fourier coefficient of $v_N(x,t) = \sum_{k=-N}^{N}\hat{v}(k,t)e^{ikx}$,

$$\frac{d\hat{v}(k,t)}{dt} = \sum_{j=-N}^{N}\hat{a}(k-j)ij\hat{v}(j,t), \quad -N \leq k \leq N. \tag{3.1.49}$$

In this case we have a $(2N+1)\times(2N+1)$ *coupled* system of ODE's written in the matrix-vector form, consult (2.2.46)

$$\frac{d}{dt}\hat{v}(t) = \hat{A}\Lambda\hat{v}(t), \quad \hat{v}(t) = \begin{bmatrix} \hat{v}(-N,t) \\ \vdots \\ \hat{v}(N,t) \end{bmatrix} \quad \hat{A}_{kj} = \hat{a}(k-j), \Lambda = \text{diag}(ik). \tag{3.1.50}$$

We can solve this system explicitly (since a $(\cdot)$ was assumed not to depend on time)

$$\hat{v}(t) = e^{\hat{A}\Lambda t}\hat{v}(0); \tag{3.1.51}$$

that is, we obtain an explicit representation of the solution operator

$$E_N(t,\tau) = F_N^{-1}e^{\hat{A}\Lambda(t-\tau)}F_N, \qquad \hat{A} = \hat{A}_N, \Lambda = \Lambda_N \tag{3.1.52}$$

where F_N denote the spectral Fourier projection

$$F_N v_N(x) = \begin{bmatrix} \hat{v}(-N) \\ \vdots \\ \hat{v}(N) \end{bmatrix}. \tag{3.1.53}$$

We note that in view of Parseval's identity $\|F_N v_N(x)\|_2 = \|v_N(x)\|_{L^2(x)}$ (modulo factorization factor), hence, stability amounts to having the a priori estimate on the discrete symbol $\hat{E}_N(t,\tau) = e^{\hat{A}_N\Lambda(t-\tau)}$, requiring

$$\|e^{\hat{A}_N\Lambda(t-\tau)}\| \leq \text{Const.}e^{M(t-\tau)}. \tag{3.1.54}$$

The essential point of stability here, lies in having a uniform bound for the RHS of (3.1.54) — a bound which is independent of the *order* of the system; for example, the 'naive' straightforward estimate of the form

$$\|e^{\hat{A}_N\Lambda(t-\tau)}\| \leq e^{\|\hat{A}_N\|\cdot\|\Lambda\|(t-\tau)} \tag{3.1.55}$$

will not suffice for that purpose because $\|\Lambda_N\|_{N\to\infty}^{\uparrow}\infty$. The essence of the a priori estimate we obtained in (3.1.22)-(3.1.23), and likewise in (3.1.47), was that the (unbounded) operator $P(x,t,D) \equiv a(x,t)\partial_x$ is *semi-bounded*, i.e.,

$$\text{Re}\left[a(x,t)\frac{\partial}{\partial x}\right] = \frac{1}{2}\left[a(x,t)\frac{\partial}{\partial x} - \frac{\partial}{\partial x}(a(x,t)\cdot)\right] = -\frac{1}{2}a_x(x,t); \tag{3.1.56}$$

namely, (compare (1.1.28))

$$\left(\mathrm{Re}\left[a(x,t)\frac{\partial}{\partial x}\right]u,u\right)_{L^2(x)} \le M\|u\|^2_{L^2(x)} \tag{3.1.57}$$

and likewise for $\mathrm{Re}\left(S_N\left[a(x,t)\frac{\partial}{\partial x}\right]\right)$. In the present form this is expressed by the sharper estimate of the matrix exponent,[8] compare (3.1.55)

$$\|e^{\hat{A}_N\Lambda(t-\tau)}\| \le e^{\|\mathrm{Re}\hat{A}_N\Lambda\|\cdot(t-\tau)}. \tag{3.1.58}$$

This time, $\|\mathrm{Re}\hat{A}_N\Lambda\|$ like the $\mathrm{Re}[P(x,t,D)]$, is bounded. Indeed, $[\mathrm{Re}\hat{A}\Lambda]_{kj} = \frac{1}{2}[\hat{a}(k-j)ij + \overline{\hat{a}(j-k)ik}]$, and since $a(x,t)$ is real (hyperbolicity!) then $\overline{\hat{a}(p)} = \hat{a}(-p)$, i.e.,

$$[\mathrm{Re}\hat{A}\Lambda]_{kj} = \frac{1}{2}i(j-k)\hat{a}(k-j) \quad -N \le j,k \le N. \tag{3.1.59}$$

Thus, $\mathrm{Re}\hat{A}\Lambda$ is a (possibly complex-valued) Toeplitz matrix, namely its (k,j) entry depends solely on its distance from the main diagonal $k-j$; we leave it as an exercise (utilizing our previous study on circulant matrices in (2.2.43)) – to see that its norm does not exceed the sum of the absolute values along the, say, zeroth $(j=0)$ row, i.e.,

$$\|\mathrm{Re}\hat{A}_N\Lambda\| \le \frac{1}{2}\sum_{k=N}^{N}|k\hat{a}(k)| \tag{3.1.60}$$

which is bounded, *uniformly with respect to* N, provided $a(x,t)$ is sufficiently smooth, e.g., we can take the exponent M to be

$$\begin{aligned} M = \frac{1}{2}\sum_{k=-N}^{N}|k\hat{a}(k)| &\le \frac{1}{2}\sqrt{\sum k^4|a(k)|^2}\cdot\sum_{k=-N}^{N}\frac{1}{k^2} \le \\ &\le \frac{\pi}{6}\cdot\|a_{xx}(x,t)\|_{L^2(x)} \end{aligned} \tag{3.1.61}$$

which is only slightly worse than what we obtained in (3.1.48).

A similar analysis shows the convergence of the spectral-Fourier method for hyperbolic systems. For example, consider the $N\times N$ symmetric hyperbolic problem

$$\frac{\partial u}{\partial t} = A(x,t)\frac{\partial u}{\partial x} + B(x,t)u, \quad \text{with symmetric} \quad A(x,t). \tag{3.1.62}$$

We note that if the system is not in this symmetric form, then (in the 1-D case) we can bring it to the symmetric form by a change of variables, i.e., the existence of a smooth symmetric $H(x,t)$ such that $H(x,t)A(x,t)$ is symmetric, implies that for $w(x,t) = T^{-1}(x,t)u(x,t)$ with $H = (T^{-1})^*T^{-1}$ we have, compare (1.1.16)

$$\frac{\partial w}{\partial t} = T^{-1}(x,t)A(x,t)T(x,t)\frac{\partial w}{\partial x} + C(x,t)w(x,t) \tag{3.1.63}$$

where $T^{-1}(x,t)A(x,t)T(x,t) \equiv T^*(x,t)H(x,t)A(x,t)T(x,t)$ is symmetric, and $C(x,t) = B(x,t) + \frac{\partial T^{-1}}{\partial t}(x,t) - T^{-1}(x,t)A(x,t)\frac{\partial T}{\partial x}(x,t)$. The spectral Fourier approximation of (3.1.62) takes the form

$$\frac{\partial v_N}{\partial t} = S_N\left(A(x,t)\frac{\partial u_N}{\partial x}\right) + S_N B(x,t)v_N(x,t). \tag{3.1.64}$$

Its stability follows from integration by parts, for by orthogonality

$$\frac{1}{2}\frac{d}{dt}\int_x v_N^2(x,t)dx = \int v_N A(x,t)\frac{\partial v_N}{\partial x}dx + \int u_N B(x,t)u_N dx \le M\int_x v_N^2(x,t)dx \tag{3.1.65}$$

[8] To see this, use Duhammel's principle for $\frac{d\hat{v}}{dt} = \mathrm{Re}\hat{A}\Lambda\hat{v}(t) + F(t)$ where $F(t) = iIm\hat{A}\Lambda e^{\hat{A}\Lambda t}$ or integrate directly.

where

$$M = \max_{x,t}\left[-\frac{\partial A(x,t)}{\partial x} + \mathrm{Re}B(x,t)\right] \tag{3.1.66}$$

and hence

$$\|v_N(t)\|_{L^2(x)} \le e^{Mt}\|v_N(0)\|. \tag{3.1.67}$$

The approximation (3.1.64) is spectrally accurate with (3.1.62) and hence spectral convergence follows. The solution of (3.1.64) is carried out in the Fourier space, and takes the form

$$\frac{d}{dt}\hat{v}(k,t) = \sum_{j=-N}^{N} \hat{A}(k-j,t)ij\hat{v}(j,t), \quad -N \le k \le N, \tag{3.1.68}$$

which form a coupled $(2N+1)\times(2N+1)$ system of ODE's for the $(2N+1)$-vectors of Fourier coefficients $\hat{v}(k,t)$.

There are two difficulties in carrying out the calculation with the spectral Fourier method. First, is the time integration of (3.1.68); even in the constant coefficient case, it requires to the computation of the exponent $e^{\hat{A}\cdot\Lambda t}$ which is expensive, and in the time-dependent case we must appeal to approximate numerical methods for time integration. Second, to compute the RHS of (3.1.68) we need to multiply an $(2N+1)\times(2N+1)$ matrix, $\hat{A}\Lambda$ by the Fourier coefficient vector which requires $\mathcal{O}(N^2)$ operations. Indeed, since $\hat{A}$ is a Toeplitz matrix and Λ is diagonal, we can still carry out this multiplication efficiently, i.e., using two FFT's which requires $\mathcal{O}(N\log N)$ operations. Yet, it still necessitates carrying out the calculation in the Fourier space. We can overcome the last difficulty with the pseudospectral Fourier method.

Before leaving the spectral method, we note that its spectral convergence equally applies to any PDE

$$\frac{\partial u}{\partial t} = P(x,t,D)u \tag{3.1.69}$$

with semi-bounded operator $P(x,t,D)$, e.g., the symmetric hyperbolic as well as the parabolic operators. Indeed, the spectral approximation of (3.1.69) reads

$$\frac{\partial v_N}{\partial t} = S_N P(x,t,D)v_N. \tag{3.1.70}$$

Multiply by v_N and integrate – by orthogonality and semi-boundedness we have

$$\frac{1}{2}\frac{d}{dt}\int_x v_N^2(x,t)dx = \mathrm{Re}(v_N, P(x,t,D)v_N) \le M\int_M v_N^2(x,t)dx. \tag{3.1.71}$$

Hence stability follows and the method converges spectrally.

3.2 The Pseudospectral Fourier Approximation

We return to the scalar constant coefficient case

$$\frac{\partial u}{\partial t} = a\frac{\partial u}{\partial x} \tag{3.2.1}$$

subject to periodic boundary conditions and prescribed initial data

$$u(x,0) = f(x). \tag{3.2.2}$$

To solve this problem by the pseudospectral Fourier method, we proceed as before, this time projecting (3.2.1) with the pseudospectral projection ψ_N, to obtain for $u_N = \psi_N u(x,t)$

$$\frac{\partial u_N}{\partial t} = \psi_N\left(a\frac{\partial u_N}{\partial x}\right). \tag{3.2.3}$$

Here, ψ_N commutes with multiplication by a constant, but unlike the spectral case, it does not commute with differentiation, i.e., by the aliasing relation (2.2.3) we have

$$\psi_N \frac{\partial \rho}{\partial x} = \sum_{k=-N}^{N} (k\tilde{\rho(k)}) e^{ikx} = \sum_{k=-N}^{N} \sum_{j} i[k + j(2N+1)]\hat{\rho}[k + j(2N+1)] e^{ikx}$$

where as

$$\frac{\partial}{\partial x}\psi_N \rho = \sum_{k=-N}^{N} (k\tilde{\rho}(k) e^{ikx} = \sum_{k=-N}^{N} ik \sum_{j} \hat{\rho}[k + j(2N+1)] e^{ikx}.$$

The difference between these two expressions is a pure aliasing error, i.e., we have for $\psi_N = S_N + A_N$, see (2.2.13)

$$\psi_N \frac{d\rho}{dx} - \frac{d}{dx}(\psi_N \rho) \equiv \left[A_N, \frac{d}{dx}\right] \rho = \sum_{k=-N}^{N} \sum_{j \neq 0} i[k + j(2N+1)]\hat{\rho}(k + j(2N+1)) e^{ikx}$$

which is spectrally small. Sacrificing such spectrally small errors, we are led to the pseudospectral approximation of (3.2.1)

$$\frac{\partial v_N}{\partial t} = a \frac{\partial v_N}{\partial t} \tag{3.2.4}$$

subject to initial conditions

$$v_N(t=0) = \psi_N f. \tag{3.2.5}$$

Here, $v_N = v_N(x,t)$ is an N-degree trigonometric polynomial which satisfies a nearby equation satisfied by the interpolant of the exact solution $\psi_N u(x,t)$. That is, $u_N \equiv \psi_N u(x,t)$ satisfies (3.2.4) modulo spectrally small truncation error

$$\frac{\partial u_N}{\partial t} = a \frac{\partial u_N}{\partial x} + F_N(x,t), \quad F_N(x,t) = a\psi_N \left[\frac{\partial}{\partial x}(I - \psi_N)u\right] \tag{3.2.6}$$

where by (3.2.3), $F_N(x,t) = a\left[\psi_N \frac{\partial u}{\partial x} - \frac{\partial}{\partial x}(\psi_N u)\right]$, and by (2.2.17) it is indeed spectrally small

$$\|F_N(x,t)\| \leq |a| \; \|\frac{\partial}{\partial x}[(I - \psi_N)u]\|] \leq |a| \; \|u\|_{H^{s+1}} \frac{1}{N^s}. \tag{3.2.7}$$

The stability proof of (3.2.4) follows along the lines of the spectral stability, and spectral convergence follows using Duhammel's principle for the stable numerical solution operator. That is, the error equation for $e_N = u_N - v_N$ is

$$\frac{\partial e_N}{\partial t} = a \frac{\partial e_N}{\partial x} + F_N(x,t) \tag{3.2.8}$$

whose solution is

$$e_N(t) = E_N(t,0)(f_N - \psi_N f) + \int_{\tau=0}^{t} E_N(t,\tau) F_N(x,\tau) d\tau. \tag{3.2.9}$$

Hence, by stability

$$\|e_N(t)\| \leq \text{Const.} e^{Mt} \cdot \|u\|_{H^{s+1}} \frac{1}{N^s} \leq \text{Const.} e^{Mt} \|f\|_{H^{s+1}} \cdot \frac{1}{N^s}; \tag{3.2.10}$$

this together with the estimate of the pseudospectral projection yields

$$\|u(t) - v_N(t)\| \leq \text{Const.} e^{Mt} \cdot \begin{cases} \frac{1}{N^s} & \text{for } H^{s+1} \text{ initial data} \\ \\ e^{-\eta N} & \text{for analytic initial data} \end{cases}. \tag{3.2.11}$$

To carry out the calculation of (3.2.4) we can compute the discrete Fourier coefficients $\tilde{v}(k,t)$ which obey the ODE,

$$\frac{d\hat{v}}{dt}(k,v) = ika\hat{v}(k,t), \tag{3.2.12}$$

as was done with the spectral case; alternatively, we can realize our approximate interpolant $v_N(x,t)$ at the $2N+1$ equidistant points $x_\nu = \nu h$, and (3.2.4) amounts to a coupled $(2N+1)$ – ODE system in the real space

$$\frac{dv_N}{dt}(x_\nu,t) = a\frac{\partial v_N}{\partial x}(x=x_\nu,t) \quad \nu = 0,1,\cdots,2N. \tag{3.2.13}$$

$$v_N(x_\nu,0) = f(x_\nu). \tag{3.2.14}$$

3.2.1 Is the pseudospectral approximation with variable coefficients stable?

Let us turn to the variable coefficient case,

$$\frac{\partial u}{\partial t} = a(x,t)\frac{\partial u}{\partial x}. \tag{3.2.15}$$

The pseudospectral approximation takes the form

$$\frac{\partial v_N}{\partial t} = \psi_N\left[a(x,t)\frac{\partial v_N}{\partial x}\right] \tag{3.2.16}$$

subject to initial conditions

$$v_N(x_\nu,0) = f(x_\nu).$$

It can be solved as a coupled ODE system in the Fourier space, and at the same time it can be realized at the $2N+1$ so-called collocation points

$$\frac{dv_N(x_\nu,t)}{dt} = a(x_\nu,t)\frac{\partial v_N}{\partial x}(x=x_\nu,t), \tag{3.2.17}$$

with initial conditions

$$v_N(x_\nu,t=0) = f(x_\nu).$$

The truncation error of this model is spectrally small in the sense that $u_N = \psi_N u$ satisfies

$$\frac{\partial u_N}{\partial t} = \psi_N\left[a(x,t)\frac{\partial u_N}{\partial x}\right] + F_N(x,t) \tag{3.2.18}$$

where

$$F_N(x,t) = \psi_N\left[a(x,t)\frac{\partial u}{\partial x}\right] - \psi_N\left[a(x,t)\frac{\partial}{\partial x}(\psi_N u)\right] \tag{3.2.19}$$

is spectrally small

$$\begin{aligned} \|F_N(x,t)\| &\leq \|\psi_N\left[a(x,t)\frac{\partial}{\partial x}\left[(I-\psi_N)u\right]\right\| \leq \| \\ &\leq e^{C_s t}\cdot\|f\|_{H^{s+1}}\cdot\frac{1}{N^s}, \quad C_s \sim \|\partial_x^{s+2}a(x,t)\|_{L^\infty}. \end{aligned} \tag{3.2.20}$$

Hence, *if* the approximation (3.2.11) is stable then spectral convergence follows. Is the approximation (3.2.11) stable? The presence of aliasing errors makes this stability question an intricate one – here is a brief explanation.

Trying to follow the differential and spectral setup, we should multiply by $v_N(x,t)$, integrate by parts and hope for the best. However, here $v_N(x,t)$ is *not* orthogonal to $(I-\psi_N)[\cdots]$ (— otherwise this would enable us to estimate $\int v_N(x,t)a(x,t)\frac{\partial v_N}{\partial x}(x,t)dx$ in terms of $\int_x v_N^2(x,t)dx$ and we are done); more precisely, for $I-\psi_N = I-S_N-A_N$ we only have that $\int v_N(I-S_N)[\cdots]dx = 0$; yet $\int v_N A_N[\cdots]dx$ leaves us with an additional contribution which is not necessarily bounded in terms of $\int_x v_N^2(x,t)dx$, and this argument fails short of a straightforward stability proof by Gronwall's inequality. To shed a different light on this difficulty, we can turn to the Fourier space; we write (3.2.16) in the form

$$\frac{\partial v_N}{\partial t} = a(x)\frac{\partial v_N}{\partial x} \tag{3.2.21}$$

and Fourier transform to get for the kth Fourier coefficient

$$\frac{d}{dt}\tilde{v}(k,t) = \sum_{j=-N}^{N} \tilde{a}(k-j,t)ij\tilde{v}(j,t) \tag{3.2.22}$$

i.e.,

$$\frac{d}{dt}\tilde{v}(t) = \tilde{A}_N\Lambda\tilde{v}(t) \quad \tilde{A}_{kj} = \sum_p \hat{a}[k-j+p(2N+1)]. \tag{3.2.23}$$

This time, $\mathrm{Re}\tilde{A}_N\Lambda$ is unbounded. This difficulty appears when we confine ourselves to the discrete framework: multiplying (3.2.17) by $v(x_v,t)$ and trying to sum by parts we arrive at

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\sum_\nu v_N^2(x_\nu,t) &= \sum_\nu a(x_\nu,t)v(x_\nu,t)\frac{\partial v}{\partial x}(x_\nu,t)\\ &= \sum_\nu \frac{\partial}{\partial x}\left[\frac{1}{2}a(x,t)v^2(x,t)\right]\Big|_{x=x_\nu} - \sum_\nu \frac{1}{2}a'(x_\nu,t)v_N^2(x_\nu,t);\end{aligned} \tag{3.2.24}$$

but the first term on the right does not vanish in this case – it equals, by the aliasing relation, to

$$\frac{2\pi}{2N+1}\sum_\nu \frac{\partial}{\partial x}\left[\frac{1}{2}a(x,t)v^2(x,t)\right]\Big|_{x=x_\nu} = \int \frac{\partial}{\partial x}[\cdots] + \sum_{p\neq 0} ip\cdot(2N+1)\frac{1}{2}\hat{av^2}[p\cdot(2N+1)] \tag{3.2.25}$$

and a loss of one derivative is reflected by the factor $2N+1$ inside the right summation. This does not prove an instability as much as it shows the failure of disproving it along these lines.

3.3 Aliasing, Resolution and (weak) Stability

3.3.1 Weighted L^2-stability

We now turn to consider the intriguing case where $a(x)$ may change sign[9]. In this section we take a rather detailed look at the prototype case of $a(x) = \sin(x)$:

$$\frac{\partial}{\partial t}u_N(x,t) = \frac{\partial}{\partial x}\psi_N\left[\sin(x)u_N(x,t)\right]. \tag{3.3.1}$$

We shall show that the solution operator associated with (3.3.1) is also similar to a unitary matrix — consult (3.3.17) below for the precise statement. This in turn leads to the announced weighted L^2-stability. It should be noted, however, that the similarity transformation in this case involves the ill-conditioned $N\times N$ Jordan blocks; as the condition number of the latter may grow linearly with N, this in turn implies *weak* L^2-instability.

We begin by noting that the Fourier approximation (3.3.1) admits a rather simple representation in the Fourier space, using the $(2N+1)$-vector of its Fourier coefficients, $\hat{u}(t) := (\hat{u}_{-N}(t),\ldots,\hat{u}_N(t))$. With the periodic extension of $\hat{u}_k(t)$ $\forall k\in Z$ in mind we are able to express the interpolant of $\sin(x)u_N(x,t)$ as

$$\psi_N\left[\sin(x)u_N(x,t)\right] = \sum_{k=-N}^{N}\frac{1}{2i}[\hat{u}_{k-1}(t)-\hat{u}_{k+1}(t)]e^{ikx},$$

so that the Fourier approximation (3.3.1) then reads

$$\frac{d}{dt}\hat{u}_k(t) = \frac{k}{2}[\hat{u}_{k-1}(t)-\hat{u}_{k+1}(t)], \quad -N\le k\le N, \tag{3.3.2}$$

[9] If $a(x) > 0$, then (3.2.21) is semi-bounded (and hence stable) in the *weighted* $L^2_{A^{-1}}$-norm, with $A = diag\{a(x_0),\ldots,a(x_{2N})\}$.

augmented by the aliasing boundary conditions,

$$\hat{u}_{-(N+1)}(t) = \hat{u}_N(t) \equiv \overline{\hat{u}}_{-N}(t), \qquad \hat{u}_{N+1}(t) = \hat{u}_{-N}(t) \equiv \overline{\hat{u}}_N(t). \tag{3.3.3}$$

Thus, in the Fourier space, our approximation is converted into the system of ODE's

$$\frac{d}{dt}\hat{u}(t) = \Lambda\hat{A}\hat{u}(t), \quad \Lambda_{jk} = k\delta_{jk}, \quad \hat{A} = \frac{1}{2}\begin{bmatrix} 0 & -1 & 0 & \dots & -1 \\ 1 & 0 & -1 & & 0 \\ 0 & 1 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 0 & -1 \\ -1 & 0 & \dots & 1 & 0 \end{bmatrix}. \tag{3.3.4}$$

We shall study the stability of (3.3.1) in terms of its unitarily equivalent Fourier representation in (3.3.4), which is decoupled into its real and imaginary parts, $\hat{u}(t) = a(t) + ib(t)$. According to (3.3.2)-(3.3.3), the real part of the Fourier coefficients, $a_k(t) := I\!R e\ \hat{u}_k(t)$, satisfies

$$\frac{d}{dt}a_k(t) = \frac{k}{2}\left[a_{k-1}(t) - a_{k+1}(t)\right], \quad -N \le k \le N, \tag{3.3.5}$$

augmented with the boundary conditions

$$a_{-(N+1)}(t) = a_{-N}(t), \quad a_{N+1}(t) = a_N(t). \tag{3.3.6}$$

The imaginary part of the Fourier coefficients, $b_k(t) := I\!m\ \hat{u}_k(t)$, satisfy the same recurrence relations as before

$$\frac{d}{dt}b_k(t) = \frac{k}{2}\left[b_{k-1}(t) - b_{k+1}(t)\right], \quad -N \le k \le N, \tag{3.3.7}$$

the only difference lies in the augmenting boundary conditions which now read

$$b_{-(N+1)}(t) = -b_{-N}(t), \quad b_{N+1}(t) = -b_N(t). \tag{3.3.8}$$

The weighted stability of the ODE systems (3.3.5) and (3.3.7) is revealed upon change of variables. For the real part in (3.3.5) we introduce the local differences,

$$\rho_k^-(t) := a_k(t) - a_{k+1}(t);$$

for the imaginary part in (3.3.7) we consider the local averages,

$$\rho_k^+(t) := b_k(t) + b_{k+1}(t).$$

Differencing consecutive terms in (3.3.5) while adding consecutive terms in (3.3.7) we find

$$\frac{d}{dt}\rho_k^\pm(t) = \frac{k}{2}\rho_{k-1}^\pm(t) - \frac{k+1}{2}\rho_{k+1}^\pm(t) \pm \frac{1}{2}\rho_k^\pm(t), \quad -N \le k \le N-1. \tag{3.3.9}$$

The motivation for considering this specific change of variables steams from the side conditions in (3.3.6) and (3.3.8), which are now translated into zero boundary values

$$\rho_{-(N+1)}^\pm(t) = \rho_N^\pm(t) = 0. \tag{3.3.10}$$

Observe that (3.3.9),(3.3.10) amount to a fixed translation of *antisymmetric* ODE systems for $\rho^-(t) := (\rho_{-N}^-(t), \dots, \rho_{N-1}^-(t))$ and $\rho^+(t) := (\rho_{-N}^+(t), \dots, \rho_{N-1}^+(t))$, that is, we have

$$\frac{d}{dt}\rho^\pm(t) = \frac{1}{2}(\pm I + \mathcal{S})\rho^\pm(t), \tag{3.3.11}$$

where $\mathcal{S}$ denotes the antisymmetric matrix

$$\mathcal{S} = \begin{bmatrix} 0 & N-1 & 0 & \dots \\ 1-N & 0 & \ddots & 0 \\ 0 & \ddots & \ddots & 1 \\ \vdots & 0 & -1 & 0 \end{bmatrix} \oplus \begin{bmatrix} 0 & -1 & 0 & \dots \\ 1 & 0 & \ddots & 0 \\ 0 & \ddots & \ddots & 1-N \\ \vdots & 0 & N-1 & 0 \end{bmatrix}.$$

The solution of these systems is expressed in terms of the *unitary* matrix $U(t) = e^{\frac{1}{2}\mathcal{S}t}$,

$$\rho^{\pm}(t) = e^{\pm t/2} U(t)\rho^{\pm}(0), \quad U^*(t)U(t) = I_{2N}. \tag{3.3.12}$$

The explicit solution given in (3.3.12) shows that our problem — when expressed in terms of the new variables $\rho^{\pm}(t)$, is clearly L^2-stable,

$$\|\rho^{\pm}(t)\| = e^{\pm t/2}\|\rho^{\pm}(0)\|.$$

Remark. We note that this L^2-type argument carries over for higher derivatives, that is, the W^{α}-norms of $\rho^{\pm}(t)$ remain bounded,

$$\|\rho\|_{W^{\alpha}} := \|\Lambda^{\alpha}\rho\| = \left(\sum_k |k|^{2\alpha}|\rho_k|^2\right)^{\frac{1}{2}}, \quad \Lambda_{jk} = k\delta_{jk}. \tag{3.3.13}$$

We want to interpret these L^2-type stability statements for the $\rho^{\pm}$-variables in term of the original variables — the real and imaginary parts of the system (3.3.4). This will be achieved in term of simple linear transformations involving the $N \times N$ Jordan blocks

$$J_{\pm} = \begin{bmatrix} 1 & \pm 1 & \dots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & & \ddots & \pm 1 \\ 0 & \dots & 0 & 1 \end{bmatrix}.$$

To this end, let us assume temporarily that the initial conditions have zero average, i.e., that

$$a_0(0) \equiv \frac{1}{2N+1}\sum_{\nu} u(x_\nu, 0) = 0. \tag{3.3.14}$$

According to (3.3.5), $a_0(t)$ remains zero $\forall t$, and so will be temporarily ignored. Then, if we let

$$\tilde{a}(t) := (a_{-N}(t), \dots, a_{-1}(t), a_1(t), \dots, a_N(t))$$

denote the 'punctured' $2N$-vector of real part associated with (3.3.4), it is related to the $2N$-vector of local differences, $\rho^{-}(t)$, through

$$\rho^{-}(t) = T_{-}\tilde{a}(t), \quad T_{-} := J_{-} \ominus J_{-}^{t}.$$

This enables us to rewrite the solution given in $(3.3.12)_{-}$ as

$$T_{-}\tilde{a}(t) = e^{-t/2} U(t) T_{-}\tilde{a}(0). \tag{3.3.15}$$

Similarly, since $b_0(t) \equiv \mathit{Im}\ \frac{1}{2N+1}\sum_{\nu} u(x_\nu, t) = 0$ in the real case, it will be temporarily ignored. Then, the 'punctured' $2N$-vector of imaginary part associated with (3.3.4),

$$\tilde{b}(t) := (b_{-N}(t), \dots, b_{-1}(t), b_1(t), \dots, b_{N-1}(t)),$$

is related to the $2N$-vector of local averages, $\rho^+(t)$, through

$$\rho^+(t) = T_+\tilde{b}(t), \quad T_+ := J_+ \oplus J_+^t,$$

which enables us to rewrite the solution given in $(3.3.12)_+$ as

$$T_+\tilde{b}(t) = e^{t/2}U(t)T_+\tilde{b}(0). \tag{3.3.16}$$

The equalities (3.3.15) and (3.3.16) confirm our assertion in the beginning of this section, namely,

Assertion. *The solution operator associated with the Fourier approximation*, (3.3.1),(3.3.14), *is similar to the unitary matrix* $\tilde{U}(t) := U(t) \oplus U(t)$, *in the sense that*

$$\tilde{u}(t) = T^{-1}\tilde{U}(t)T\tilde{u}(0), \qquad \tilde{u}(t) := e^{t/2}\tilde{a}(t) \oplus e^{-t/2}\tilde{b}(t), \quad T := T_- \oplus T_+. \tag{3.3.17}$$

We are now in a position to translate this similarity into an appropriate *weighted* L^2-stability. On the left of (3.3.15) we have a weighted L^2-norm of $\tilde{a}(t)$, $\|T_-\tilde{a}(t)\| \equiv \|\tilde{a}(t)\|_{T_-^tT_-}$. Also, $U(t)$ being a unitary matrix has an L^2-norm $= 1$, hence the right hand side of (3.3.15) does not exceed, $e^{-t/2}\|T_-\tilde{a}(0)\| \equiv e^{-t/2}\|\tilde{a}(0)\|_{T_-^tT_-}$, and therefore $\tilde{a}(t) = I\!R e\,(\hat{u}_{-N}(t), \ldots, \hat{u}_{-1}, \hat{u}_1(t), \ldots, \hat{u}_N(t))$ satisfies

$$\|\tilde{a}(t)\|_{T_-^tT_-} \le e^{-t/2}\|\tilde{a}(0)\|_{T_-^tT_-}, \quad T_-^tT_- = J_-^tJ_- \oplus J_-J_-^t.$$

Expanding the last inequality by augmenting it with the zero value of $a_0(t)$ we find the weighted L^2-stability of the real part

$$\|a(t)\|_{H_-} \le e^{-t/2}\|a(0)\|_{H_-}, \quad H_- = J_-^tJ_- \oplus 1 \oplus J_-J_-^t. \tag{3.3.18}$$

Similarly, (3.3.16) gives us the weighted stability of the imaginary part

$$\|b(t)\|_{H_+} \le e^{t/2}\|b(0)\|_{H_+}, \quad H_+ = J_+^tJ_+ \oplus 1 \oplus J_+J_+^t. \tag{3.3.19}$$

Summarizing (3.3.18) and (3.3.19) we have shown

Theorem 3.1 (Weighted stability) *Consider the Fourier method (3.3.1) subject to initial conditions with zero mean, (3.3.14). Then the following weighted* L^2*-stability estimate holds*

$$|||u_N(t)|||_H \le e^{t/2}|||u_N(0)|||_H. \tag{3.3.20}$$

Here $|||u_N(t)|||_H$ *denotes the weighted* L^2*-norm*

$$|||u_N(t)|||_H := \|I\!R e\ \hat{u}(t) \oplus I\!m\ \hat{u}(t)\|_H, \tag{3.3.21}$$

where the weighting matrix $H := H_- \oplus H_+ > 0$ *is given by*

$$H_\pm := J_\pm^tJ_\pm \oplus 1 \oplus J_\pm J_\pm^t = \left[\begin{array}{ccccccccccc}
1 & \pm 1 & & & & & & & & & \\
\pm 1 & 2 & \pm 1 & & & & & & & & \\
 & \ddots & \ddots & \ddots & & & & & & & \\
 & & \ddots & \ddots & \pm 1 & & & & & & \\
 & & & \pm 1 & 2 & & & & & & \\
\hline
 & & & & & 1 & & & & & \\
 & & & & & & 2 & \pm 1 & & & \\
 & & & & & & \pm 1 & 2 & \pm 1 & & \\
 & & & & & & & \ddots & \ddots & \ddots & \\
 & & & & & & & & \ddots & \ddots & \pm 1 \\
 & & & & & & & & & \pm 1 & 1
\end{array}\right].$$

We close this section by noting three possible extensions of the last weighted stability result. Duhammel's principle gives us

1. Inhomogeneous terms. *Let $u_N(t) \equiv u_N(\cdot,t)$ denote the solution of the inhomogeneous Fourier method*

$$\frac{\partial}{\partial t}u_N(x,t) = \frac{\partial}{\partial x}\psi_N\left[\sin(x)u_N(x,t)\right] + F_N(x,t). \tag{3.3.22}$$

Then there exists a constant, $\mathrm{C}(t)$, such that the following weighted L^2-stability estimate holds

$$|||u_N(t)|||_H \leq \mathrm{C}(t)\left[|||u_N(0)|||_H + \max_{0\leq\tau\leq t}|||F_N(\tau)|||_H\right]. \tag{3.3.23}$$

Our second corollary shows that the weighted L^2-stability of the Fourier method is invariant under low order perturbations.

2. Low order terms. *Let $u_N(t) \equiv u_N(\cdot,t)$ denotes the solution of the Fourier method*

$$\frac{\partial}{\partial t}u_N(x,t) = \frac{\partial}{\partial x}\psi_N\left[\sin(x)u_N(x,t)\right] + \psi_N\left[p(x)u_N(x,t)\right], \qquad p\epsilon L^\infty[0,2\pi). \tag{3.3.24}$$

Then there exists a constant, $\mathrm{C}(t)$, such that the following weighted L^2-stability estimate holds

$$|||u_N(t)|||_H \leq \mathrm{C}(t)|||u_N(0)|||_H. \tag{3.3.25}$$

In our third corollary we note that the last two weighted L^2-stability results apply equally well to higher order derivatives, which brings us to

3. Weighted W^α-Stability. *Let $u_N(t) \equiv u_N(\cdot,t)$ denote the solution of the Fourier method*

$$\frac{\partial}{\partial t}u_N(x,t) = \frac{\partial}{\partial x}\psi_N\left[\sin(x)u_N(x,t)\right]. \tag{3.3.26}$$

Then there exist positive definite matrices, $H_\pm^{(\alpha)}$, and a constant C_α, such that the following weighted W^α-stability estimate holds

$$|||u_N(t)|||_{W_H^\alpha} \leq \mathrm{C}_\alpha(t)|||u_N(0)|||_{W_H^\alpha} \tag{3.3.27}$$

Here $|||u_N(t)|||_{W_H^\alpha}$ denotes the weighted W^α-norm

$$|||u_N(t)|||^2_{W_H^\alpha} := \|\Lambda^\alpha I\!R e\ \hat{u}(t)\|^2_{H_-^{(\alpha)}} + \|\Lambda^\alpha I\!m\ \hat{u}(t)\|^2_{H_+^{(\alpha)}}. \tag{3.3.28}$$

The last results enable to put forward a complete weighted L^2-stability theory. The following assertion contains the typical ingredients.

Assertion. *The Fourier method*

$$\frac{\partial}{\partial t}u_N(x,t) = \psi_N[\sin(x)\frac{\partial}{\partial x}u_N(x,t)], \tag{3.3.29}$$

satisfies the following weighted W^α-stability estimate

$$\|u_N(\cdot,t)\|_{W_H^\alpha} \leq \mathrm{C}_\alpha(t)\|u_N(\cdot,0)\|_{W_H^\alpha}. \tag{3.3.30}$$

This last assertion confirms the weighted stability of the Fourier method in its non-conservative transport form.

Sketch of the Proof. We rewrite (3.3.29) in the 'conservative form'

$$\frac{\partial}{\partial t}u_N(x,t) = \frac{\partial}{\partial x}\psi_N[\sin(x)u_N(x,t)] + \left[\psi_N\sin(x), \frac{\partial}{\partial x}\right]u_N(x,t),$$

where $[\psi_N\sin(x), \frac{\partial}{\partial x}] := \psi_N(\sin(x)\frac{\partial}{\partial x}\cdot) - \frac{\partial}{\partial x}(\psi_N\sin(x)\cdot)$ denotes the usual commutator between interpolation and differentiation. The weighted L^2-stability stated in Theorem 2.1 tells us that this commutator is bounded in the corresponding weighted operator norm. Therefore, we may treat the right hand side of (3.3.29) as a low order term and weighted L^2-stability ($\alpha = 0$) follows in view of the second corollary above. The case of general $\alpha > 0$ follows with the help of the third corollary. ■.

2N+1	65	129	257	513	1205
$\frac{\|u_N(t)\|}{\|u_N(0)\|}$	570	2003	5535	15028	39798

Table 3.1: *Amplification of* $\|u_N(t)\|$ *at* $t = 10$, *subject to initial data* $\hat{u}_k(0) = i\sin(k\pi/N)$.

3.3.2 Algebraic stability and weak L^2-instability

In this section we turn our attention to the behavior of the Fourier method (3.3.1) in terms of the L^2-norm. Table 3.1 suggests that when measured with respect to the standard (weight-free) L^2-norm, the Fourier approximation may grow linearly with the number of gridpoints N.

The main result of this section asserts that this is indeed the case.

Theorem 3.2 (Weak instability) *There exist constants,* $\mathrm{C}_1(t)$ *and* $\mathrm{C}_2(t)$, *such that the following estimate holds*

$$\mathrm{C}_1(t)N \leq \|e^{DAt}\| \leq \mathrm{C}_2(t)N. \tag{3.3.31}$$

The right hand side of (3.3.31) tells us that the Fourier method may amplify the L^2-size of its initial data by an amplification factor $\leq \mathcal{O}(N)$ — that is , the Fourier method is *algebraically stable*. The left hand side of (3.3.31) asserts that this estimate is sharp in the sense that there exist initial data for which this $\mathcal{O}(N)$ amplification is attained — that is, the Fourier method is *weakly* L^2*- unstable*.

We turn to the proof of the algebraic stability. Let $u_N(t)$ denote the solution of the Fourier method (3.3.1) subject to arbitrary initial data, $u_N(0)$. We claim that we can bound the ratio $\|u_N(t)\|/\|u_N(0)\|$ in terms of the *condition number*, $\kappa(H)$, of the weighting matrix H, $\kappa(H) := \|H\| \cdot \|H^{-1}\|$. Indeed

$$\begin{aligned}
\|u_N(t)\| &= \|I\!Re\ \hat{u}(t) \oplus I\!I m\ \hat{u}(t)\| \leq \sqrt{\|H^{-1}\|} \cdot |||u_N(t)|||_H \leq \\
&\leq \mathrm{C}(t)\sqrt{\|H^{-1}\|} \cdot |||u_N(0)|||_H \leq \\
&\leq \mathrm{C}(t)\sqrt{\|H\| \cdot \|H^{-1}\|} \cdot \|I\!Re\ \hat{u}(0) \oplus I\!I m\ \hat{u}(0)\| = \\
&= \mathrm{C}(t)\sqrt{\kappa(H)} \cdot \|u_N(0)\|.
\end{aligned} \tag{3.3.32}$$

Here, the first and last equalities are Parseval's identities; the second and forth inequalities are straightforward by the definition of a weighted norm; and the third is a manifestation of the weighted L^2-stability stated in Theorem 3.1.

The estimate (3.3.32) requires to upper-bound the condition number of the weighting matrix H. We recall that the weighting matrix H is the direct sum of the matrices $H_\pm$ given in (3.3.18)-(3.3.19), whose L^2-norms equal the squared L^2-norms of the corresponding Jordan blocks, $\|H_\pm\| \equiv \|J_\pm\|^2$, $\|H_\pm^{-1}\| \equiv \|J_\pm^{-1}\|^2$. Inserting this into (3.3.32) we arrive at

$$\|e^{DAt}\| := \sup_{u_N(0)\neq 0} \frac{\|u_N(t)\|}{\|u_N(0)\|} \leq \mathrm{C}(t)\kappa(J), \quad J := J_- \oplus J_+. \tag{3.3.33}$$

Thus it remains to upper bound the condition number of the Jordan blocks, $J_\pm$. For the sake of completeness we include a brief calculation of the latter. The inverse of $J_\pm$ are upper-triangular Toeplitz matrices,

$$\left(J_\pm^{-1}\right)_{jk} = \begin{cases} (\mp 1)^{j-k} & k \geq j, \\ 0 & k < j, \end{cases} \tag{3.3.34}$$

for which we have,

$$\|J_{\pm}^{-1}w\|^2 = \sum_{j=-N}^{N} |\sum_{k\geq j}(\mp 1)^{j-k}w_k|^2 \leq \sum_{j=-N}^{N}\sum_{k}|w_k|^2\sum_{k\geq j}1 \sim 2N^2\|w\|^2. \tag{3.3.35}$$

This means that $\|J_{\pm}^{-1}\| \leq \sqrt{2}N$, and together with the straightforward upper-bound, $\|J_{\pm}\| \leq 2$, the right hand side of the inequality (3.3.31) now follows with $\mathrm{C}_2(t) = 2\sqrt{2}\mathrm{C}(t)$. ■

The above $\mathcal{O}(N)$-algebraic stability is essentially due to the $\mathcal{O}(N)$ upper-bound on the size of the inverses of Jordan blocks stated in (3.3.35). Can this upper-bound be improved? an affirmative answer to this question depends on the regularity of the data, as shown by the estimate

$$\|J_{\pm}^{-1}w\|^2 = \sum_{j=-N}^{N} |\sum_{k\geq j}(\mp 1)^{j-k}w_k|^2 \leq \sum_{j=-N}^{N}\sum_{k}|k|^{2\alpha}|w_k|^2\sum_{k\geq j}|k|^{-2\alpha},$$

which yields an $\mathcal{O}(N^{(1-\alpha)_+})$ bound for W^{α}-data,

$$\|J_{\pm}^{-1}w\| \leq \mathrm{C}_{N,\alpha}N^{(1-\alpha)_+}\|w\|_{W^{\alpha}}, \quad \|w\|_{W^{\alpha}} := \left(\sum_{k}|k|^{2\alpha}|w_k|^2\right)^{1/2}.$$

Noting that the rest of the arguments in the proof of algebraic stability are invariant with respect to the W^{α}-norm (— in particular, the weighted W^{α}-stability stated above), we conclude the following extension of the right inequality in (3.3.31).

Corollary 3.1 (Weak W^{α}-stability estimate) *There exist constants $\mathrm{C}_{s,\alpha}$, $s,\alpha \geq 0$, such that the following estimate holds*

$$\|u_N(\cdot,t)\|_{W^s} \leq \mathrm{C}_{N,s,\alpha}N^{(1-\alpha)_+}\|u_N(\cdot,0)\|_{W^{s+\alpha}}. \tag{3.3.36}$$

Here $\mathrm{C}_{N,s,\alpha} = \begin{cases} \mathrm{Const}\cdot\sqrt{\log N} & \alpha = \frac{1}{2}, 1, \\ \leq \mathrm{C}_{s,\alpha} & \text{otherwise.} \end{cases}$

Corollary 3.1 tells us how the smoothness of the initial data is related to the possible algebraic growth; actually, for W^{α}-initial data with $\alpha > 1$, there is no L^2-growth. However, for arbitrary L^2 data ($s = \alpha = 0$) we remain with the $\mathcal{O}(N)$ upper bound (3.3.35), and this bound is indeed sharp for, say, $w_k \sim (-1)^k$. (In fact, the latter is reminiscent of the unstable oscillatory boundary wave we shall meet later in (3.3.54)).

These considerations lead us to the question whether the *linear* L^2-growth upper-bound offered by the right hand side of (3.3.31) is sharp. To answer this question we return to take a closer look at the real and imaginary parts of our system (3.3.2).
We recall that according to (3.3.5) the real part, $a_k(t) = I\!R e\ \hat{u}_k(t)$, satisfies,

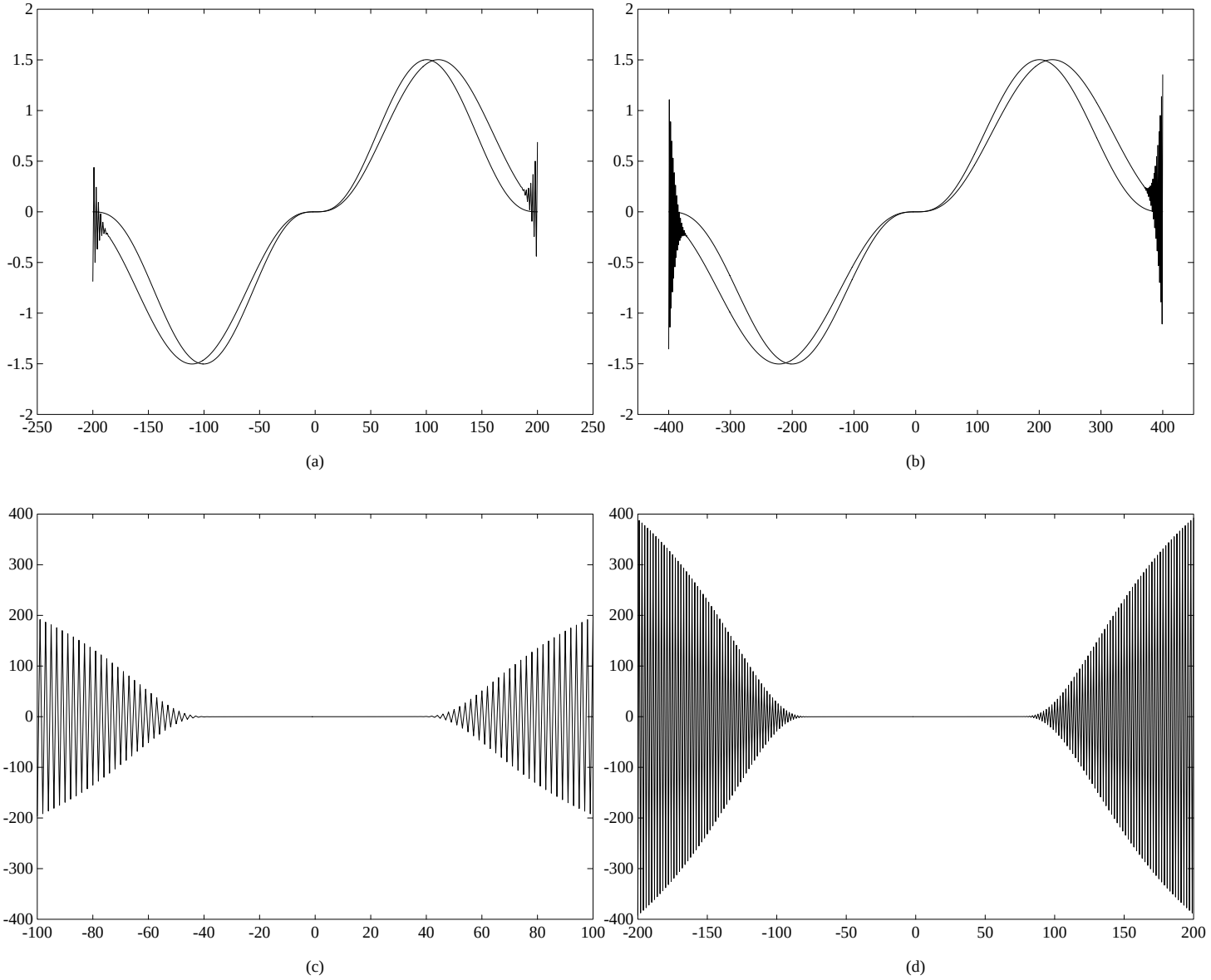
$$\frac{d}{dt}a_k(t) = \frac{k}{2}\left[a_{k-1}(t) - a_{k+1}(t)\right], \quad -N \leq k \leq N.$$

Summing by parts against $a_k(t)$ we find

$$\frac{1}{2}\frac{d}{dt}\sum_{k=-N}^{N}a_k^2(t) = \frac{1}{2}\sum_{k=-N+1}^{N}a_k(t)a_{k-1}(t) - \frac{N}{2}[a_{-(N+1)}(t)a_{-N}(t) + a_{N+1}(t)a_N(t)].$$

The boundary conditions (3.3.6), $a_{-(N+1)}(t) - a_{-N}(t) = a_{N+1}(t) - a_N(t) = 0$, imply that the second term on the right is positive; using Cauchy-Schwartz to upper bound the first term yields $\frac{d}{dt}\|a(t)\|^2 \leq \|a(t)\|^2$, which in turn implies that the real part of the system (3.3.2) is L^2-stable

$$\|a(t)\| \leq e^{t/2}\|a(0)\|, \quad a(t) = I\!R e\ \hat{u}(t).$$



Imaginary part of Fourier coefficients, $\mathit{Im}\ \hat{u}_k(t)$, computed with $\Delta t = \frac{1}{5N}$ at
(a) $t = 0$ and $t = 0.1$ with $N = 200$ (b) $t = 0$ and $t = 0.1$ with $N = 400$
(c) $t = 1.$ with $N = 100$ (d) $t = 1.$ with $N = 200$

Figure 3.1: Fourier Solution of $u_t = (\sin(x)u)_x, \quad \hat{u}_k(0) = \xi_k^3(\pi - \xi_k)^3/20, \xi_k = k\pi\Delta\xi.$

In contrast to the L^2-bounded real part, it will be shown below that the imaginary part of our system experiences an L^2 linear growth, which is responsible for the algebraically weak L^2-instability of the Fourier method.

The imaginary part of our system, $b(t) = \mathbb{I}m\ \hat{u}_k(t)$, satisfies the same recurrence relations as before

$$\frac{d}{dt}b_k(t) = \frac{k}{2}\left[b_{k-1}(t) - b_{k+1}(t)\right], \quad -N \le k \le N, \tag{3.3.37}$$

the only difference lies in the augmenting boundary conditions which now read

$$b_{-(N+1)}(t) = -b_{-N}(t), \quad b_{N+1}(t) = -b_N(t) = 0. \tag{3.3.38}$$

Trying to repeat our argument in the real case, we sum by parts against $b_k(t)$,

$$\frac{1}{2}\frac{d}{dt}\sum_{k=-N}^{N} b_k^2(t) = \frac{1}{2}\sum_{k=-N+1}^{N} b_k(t)b_{k-1}(t) - \frac{N}{2}[b_{-(N+1)}(t)b_{-N}(t) + b_{N+1}(t)b_N(t)], \tag{3.3.39}$$

but unlike the previous case, the judicious minus sign in the augmenting boundary conditions (3.3.38) leads to the *lower* bound

$$\frac{d}{dt}\|b(t)\|^2 \ge -\|b(t)\|^2 + N[b_{-N}^2(t) + b_N^2(t)]. \tag{3.3.40}$$

This lower bound indicates (but does not prove!) the possible L^2-growth of the imaginary part. Figure 3.1 confirms that unlike the L^2-bounded real part, the behavior of the imaginary part is indeed markedly different — it consists of binary oscillations which form a growing modulated wave as $|k| \uparrow N$. These *binary* oscillations suggest to consider $v_k(t) := (-1)^k b_k(t)$, in order to gain a better insight into the growth of the underlying modulated wave. Observe that (3.3.37)-(3.3.38) then recasts into the centered difference scheme

$$\frac{d}{dt}v_k(t) = \xi_k\frac{v_{k+1}(t) - v_{k-1}(t)}{2\Delta\xi}, \quad \xi_k := k\Delta\xi, \quad 0 \le k \le N, \quad \Delta\xi := \frac{1}{N+\frac{1}{2}}, \tag{3.3.41}$$

which is augmented with first order homogeneous extrapolation at the 'right' boundary

$$v_{N+1}(t) - v_N(t) = 0. \tag{3.3.42}$$

We note in passing that {i} The $b_k(t)$'s, and hence the $v_k(t)$'s, are symmetric — in this case they have an odd extension for $-N \le k \le 0$; {ii} No additional boundary condition is required at the left characteristic boundary $\xi_0 = 0$; and finally, {iii} Though (3.3.41)-(3.3.42) are independent of the frequency spacing — in fact any $\Delta\xi = \mathcal{O}(1/N)$ will do, yet the choice of $\Delta\xi = (N+\frac{1}{2})^{-1}$ will greatly simplify the formulae obtained below. These simplifications will be advantageous throughout the rest of this section.

Clearly, the centered difference scheme (3.3.41) could be viewed as a consistent approximation to the linear wave equation

$$\frac{\partial}{\partial t}v(\xi,t) = \xi\frac{\partial}{\partial\xi}v(\xi,t), \quad 0 \le \xi \le 1.$$

The essential point is that $\xi = 1$ is an *inflow* boundary in this case, and that the boundary condition (3.3.42) is *inflow-dependent* in the sense that it is consistent with the interior inflow problem. Such inflow-dependent boundary condition renders the related constant coefficient approximation *unstable*.

To show that there is an $\mathcal{O}(N)$-growth in this case requires a more precise study, which brings us to the proof of the weak L^2-instability. We decompose the imaginary components, $b_k(t)$, as the sum of two contributions — a stable part, $s_k(t)$, associated with the evolution of the initial data; and an unstable part, $\omega_k(t)$, which describes the unstable binary oscillations propagating from the boundaries into the interior domain,

$$b_k(t) \equiv s_k(t) + \omega_k(t).$$

Here, $s(t) := (s_1(t), \ldots, s_N(t))$ is governed by an *outflow* centered difference scheme which is complemented by *stable* boundary extrapolation,

$$\left\{ \begin{array}{lcl} \frac{d}{dt}s_k(t) & + & \xi_k \frac{s_{k+1}(t) - s_{k-1}(t)}{2\Delta\xi} = 0, \quad 0 \le k \le N, \quad \Delta\xi := \frac{1}{N+\frac{1}{2}} \\ s_k(0) & = & b_k(0), \\ s_{N+1}(t) & = & s_N(t). \end{array} \right. \tag{3.3.43}$$

As before, we exploit symmetry to confine our attention to the 'right half' of the problem, $0 \le k \le N$.

A straightforward L^2-energy estimate confirms that this part of the imaginary components is L^2-stable, $\|s(t)\| \le e^{-t}\|b(0)\|$. In fact, the scheme (3.3.43) retains high-order stability in the sense that

$$\|s(t)\|_{W^\alpha} = \left(\sum_{k=0}^{N} |k|^{2\alpha} |s_k(t)|^2 \right)^{1/2} \le \mathrm{Const}_{\alpha,t} \cdot \|b(0)\|_{W^\alpha}, \quad \forall \alpha \ge 0. \tag{3.3.44}$$

We close our discussion on the so called "s"-part by noting that (3.3.43) is a second-order accurate approximation to the initial-value problem

$$\left\{ \begin{array}{l} \frac{\partial}{\partial t}s(\xi,t) = \xi \frac{\partial}{\partial \xi}s(\xi,t), \qquad \xi \ge 0, \\ s(\xi,0) = b(\xi), \qquad\qquad b(\xi) := \frac{-1}{2N+1} \sum_{\nu=0}^{2N} u_N(x_\nu, 0) \sin(\pi\nu\xi); \end{array} \right. \tag{3.3.45}$$

Observe that the initial condition $b(\xi)$ is nothing but a trigonometric interpolant in the frequency 'ξ-space', which coincides with the initial value of the imaginary components, $b(\xi_k) = \mathit{Im}\ \hat{u}_k(0) \equiv b_k(0)$. Using the explicit solution of this initial value problem, we end up with a second order convergence statement which reads[10]

$$s_k(t) = b(\xi_k e^{-t}) + \mathcal{O}(\Delta\xi)^2, \quad t \ge 0. \tag{3.3.46}$$

We now turn our attention to the unstable oscillatory part, $\omega_k(t) = (-1)^{N-k} v_k(t)$. It is governed by an *inflow* centered difference scheme,

$$\left\{ \begin{array}{lcl} \frac{d}{dt}v_k(t) & = & \xi_k \frac{v_{k+1}(t) - v_{k-1}(t)}{2\Delta\xi}, \quad 0 \le k \le N, \\ v_k(0) & \equiv & 0, \end{array} \right. \tag{3.3.47}$$

which is coupled to the previous stable "s"-part (3.3.43), through the boundary condition

$$v_{N+1}(t) - v_N(t) = s_{N+1}(t) + s_N(t). \tag{3.3.48}$$

The boundary condition (3.3.48) is the first-order accurate extrapolation we met earlier in (3.3.42) — but this time, with the additional inhomogeneous boundary data. And as before, a key ingredient in the L^2-instability is the fact that such boundary treatment is inflow-dependent.

Specifically, we claim: *the inflow-dependent extrapolation on the left of* (3.3.48) *reflects the boundary values on the right of* (3.3.48), *which 'inflow' into the interior domain with an amplitude amplified by a factor of order* $\mathcal{O}(N)$.

To prove this claim we proceed as follows. Forward differencing of (3.3.47) implies that $r_{k+\frac{1}{2}}(t) := v_{k+1}(t) - v_k(t)$ satisfy the stable difference scheme

$$\left\{ \begin{array}{lcl} \frac{d}{dt}r_{k+\frac{1}{2}}(t) & = & \frac{\xi_{k+\frac{3}{2}} r_{k+\frac{3}{2}}(t) - \xi_{k-\frac{1}{2}} r_{k-\frac{1}{2}}(t)}{2\Delta\xi} - \frac{r_{k+\frac{3}{2}}(t) - 2r_{k+\frac{1}{2}}(t) + r_{k-\frac{1}{2}}}{4}, \quad k \le N-1, \\ r_{k+\frac{1}{2}}(0) & \equiv & 0, \\ r_{N+\frac{1}{2}}(t) & = & s_{N+1}(t) + s_N(t) \equiv 2s_N(t). \end{array} \right. \tag{3.3.49}$$

[10] The last equality should be interpreted of course in the W^α-sense, with α limited by the initial W^α-smoothness of $b_k(0)$.

Clearly, this difference scheme is consistent with, and hence convergent to the solution of the initial-boundary value problem

$$\left\{ \begin{array}{lcl} \frac{\partial}{\partial t} r(\xi,t) & = & \frac{\partial}{\partial \xi}(\xi r(\xi,t)), \quad 0 \le \xi \le 1, \\ r(\xi,0) & \equiv & 0 \\ r(1,t) & = & 2 s_N(t). \end{array} \right. \tag{3.3.50}$$

Observe that $r(\xi,t)$ describes a boundary wave which is prescribed on the $\xi_{N+\frac{1}{2}} = 1$ boundary of the computed spectrum, $r(1,t) = 2s_N(t)$, and propagates into the interior domain of lower frequencies $\xi < 1$,

$$r(\xi,t) = \left\{ \begin{array}{ll} \frac{2}{\xi} s_N(t + \ln \xi), & t + \ln \xi \ge 0, \\ 0, & t + \ln \xi \le 0. \end{array} \right. \tag{3.3.51}$$

We conclude that the forward differences, $r_{k+\frac{1}{2}}(t) = v_{k+1}(t) - v_k(t)$, form a second-order accurate approximation of this boundary wave,

$$r_{k+\frac{1}{2}}(t) = r(\xi_{k+\frac{1}{2}}, t) + \mathcal{O}(\Delta\xi)^2, \qquad \xi_{k+\frac{1}{2}} = (k + \frac{1}{2})\Delta\xi.$$

Returning to the original variables, $\omega_k(t) \equiv (-1)^k \sum_{j=0}^{k-1} r_{j+\frac{1}{2}}(t)$, the latter equality reads

$$\begin{array}{rcl} \omega_k(t) & = & (-1)^k \displaystyle\sum_{j=0}^{k-1} r(\xi_{j+\frac{1}{2}}, t) + \mathcal{O}(k(\Delta\xi)^2) = \\ & = & \frac{(-1)^k}{\Delta\xi} R(\xi_k, t) + \mathcal{O}(\Delta\xi), \qquad R(\xi_k,t) := \displaystyle\int_{e^{-t}}^{\xi_k} r(\xi,t) d\xi, \end{array} \tag{3.3.52}$$

which confirms our above claim regarding the amplification of a boundary wave by a factor of $\mathcal{O}(1/\Delta\xi \sim N)$.

The a priori estimates (3.3.44) and (3.3.52) provide us with precise information on the behavior of the imaginary components, $b(t) = s(t) + \omega(t)$: their initial value at $t = 0$ propagate by the stable "s"-part and reaches the boundary of the computed spectrum at $\xi_{N+\frac{1}{2}} = 1$ with the approximate boundary values of (3.3.46), $s_N(t) = b(e^{-t}) + \mathcal{O}(\Delta\xi)$; the latter propagate into the interior spectrum as a boundary wave of the form (3.3.51), $r(\xi,t) = \frac{2}{\xi} b(\frac{1}{\xi e^t})$, whose *primitive* in (3.3.52) describes the unstable oscillatory "ω"-part of the solution. Added all together we end up with

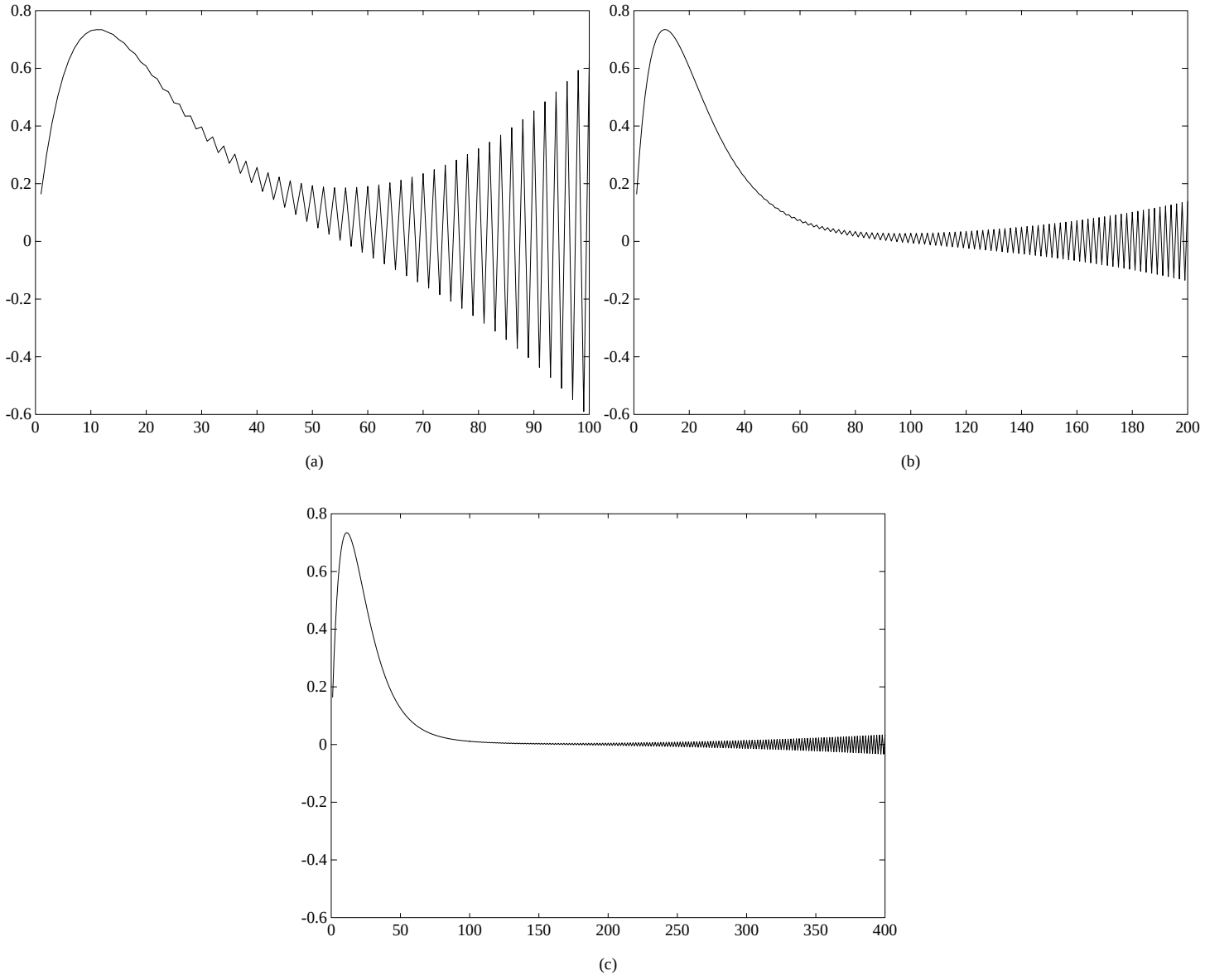
$$b_k(t) = b(\xi_k e^{-t}) + \left\{ \begin{array}{ll} \frac{2(-1)^k}{\Delta\xi} \displaystyle\int_{\xi \ge e^{-t}/\xi_k}^{1} b(\xi) \frac{d\xi}{\xi}, & e^{-t} \le \xi_k \le 1 \\ 0 \quad , & 0 \le \xi_k \le e^{-t} \end{array} \right\} + \mathcal{O}(\Delta\xi). \tag{3.3.53}$$

Thus, the unstable "ω"-part contributes a wave which is modulated by binary oscillations; the amplitude of these oscillations start with $\mathcal{O}(1/\Delta\xi \sim N)$ amplification near the boundary of the computed spectrum, $\xi_N \sim 1$, and decreases as they propagate into the interior domain of lower frequencies. Moreover, for any fixed $t > 0$, only those modes with wavenumber k such that $e^{-t} < |k|/N \le 1$, are affected by the unstable "ω" part. Put differently, we state this as

Corollary 3.2 (Weak instability revisited) *For any fixed $t > 0$, the Fourier method (3.3.1) experiences a weak instability which affects only a fixed fraction of the computed spectrum. Yet, the size of this fixed fraction, $1 - e^{-t}$, approaches unity exponentially fast in time.*

There are two different cases to be considered, depending on the smoothness of the initial data.

1. Smooth initial data. If the initial data $u_N(x,0)$ are sufficiently smooth, then $b_k(0) = \mathbb{I}m\ \hat{u}_k(0)$ are rapidly decaying as $|k| \uparrow N$, and hence — by the W^α-stability of the "s"-part in (3.3.44), this

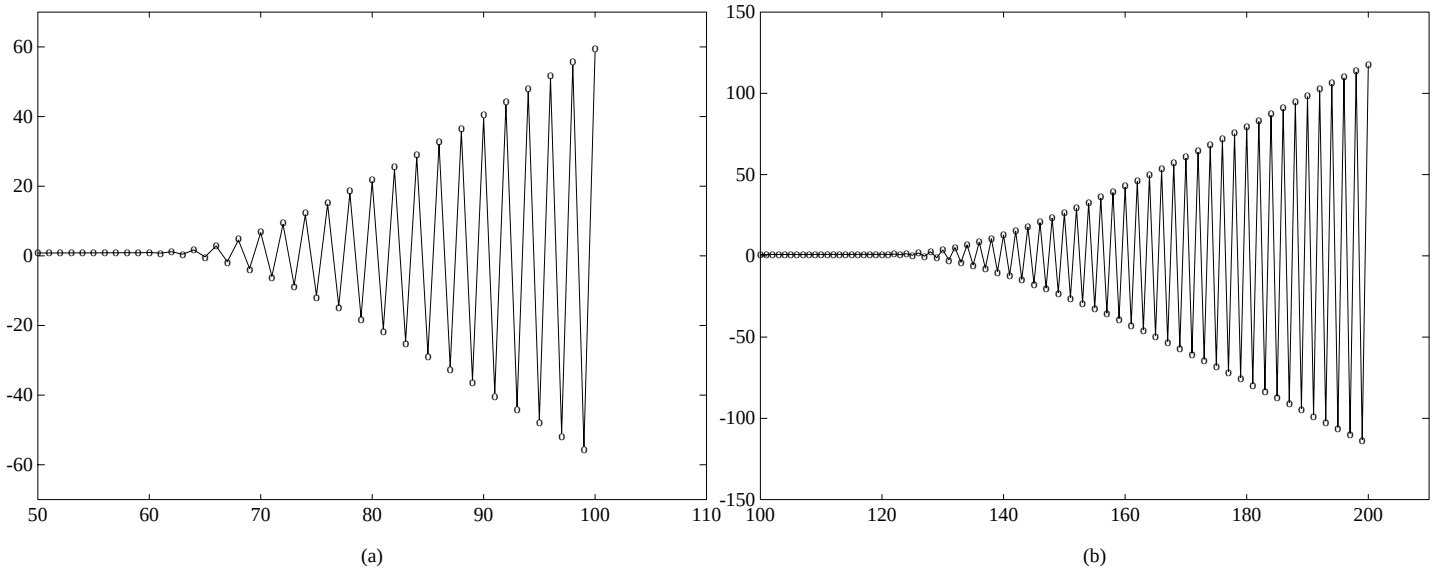


Imaginary part of Fourier coefficients, $\mathit{Im}\ \hat{u}_k(t)$, computed at $t = 3$ with $\Delta t = \frac{1}{10N}$ and
(a) with $N = 100$ (b) with $N = 200$ (c) with $N = 800$

Figure 3.2: Fourier solution of $u_t = (\sin(x)u)_x, \quad \hat{u}_k(0) \sim \frac{i}{k^3}$.

rapid decay is retained later in time for $s_k(t), t > 0$. This implies that the discrete boundary wave — governed by the stable scheme (3.3.49), is negligibly small, $r_{k+\frac{1}{2}}(t) \approx 0$, because its boundary values are, $2s_N(t) \approx 0$. We conclude that in the smooth case, $\|b(t)\| \sim \|b(0)\| + \mathcal{O}(1)$ remains of the same size as its initial data, $\|b(0)\|$.

Figure 3.2 demonstrates this result for a prototype case of smooth initial data in Besov $B^3_\infty(L^\infty)$ — in this case, initial data with cubically decaying imaginary components, $b_k(0) \sim |k|^{-3}$. As told by (3.3.53), the temporal evolution of these components should include an amplified oscillatory boundary wave, $\omega_k(t) \sim (-1)^k k^3 N^{-5}$, consult Remark 3 below. This $\mathcal{O}(N)$ amplification is confirmed by the *quadratic* decay of the boundary amplitudes, $\omega_N(t)$. Note that despite this amplification, the boundary wave and hence the whole Fourier solution remain L^2 bounded in this smooth case.



Imaginary part of Fourier coefficients, $\mathit{Im}\ \hat{u}_k(t), (---)$ computed at $t = 0.5$ vs. $s_k(t) + \omega_k(t), (ooo)$,
(a) with $N = 100$ (b) with $N = 200$

Figure 3.3: Fourier solution of $u_t = (\sin(x)u)_x, \quad \hat{u}_k(0) = i\sin(\xi_k),\ \xi_k = k\pi\Delta\xi$.

2. Nonsmooth initial data. We consider initial data $u_N(x, 0)$ with very low degree of smoothness beyond their mere L^2-integrability, e.g., for $b(\xi) = N^{-1/2}(1 - \xi)$, the corresponding components of $\mathit{Im}\ \hat{u}_k(0) = N^{-1/2}(1 - \frac{k}{N})$, are square summable but slowly decaying as $|k| \uparrow N$. Since $b(0)$ serves as initial data for the stable "s"-part in (3.3.43), the components of $s_k(t)$ will remain square summable for $t > 0$, but will remain slowly decaying as $|k| \uparrow N$. In particular, this means that $s_N(t) = \mathcal{O}(N^{-1/2})$ can be used to create the $\mathcal{O}(N^{-1/2})$ boundary wave $r(\xi, t)$ dictated by (3.3.50). According to (3.3.52), the amplified primitive of this boundary wave, $(-1)^k R(\xi_k, t)/\Delta\xi \sim N^{1/2}$, will serve as the leading order term of the unstable part. We conclude that the imaginary part $\|b(t)\|$ will be amplified by a factor of $\mathcal{O}(N)$ relative to the size of its nonsmooth initial data $\|b(0)\|$, which confirms the left hand side of the inequality (3.3.31).

Figure 3.3 demonstrates this result for a prototype case of nonsmooth initial data with imaginary components given by, $b_k(0) = \sin(\xi_k)$, that is, initial data represented by a strongly peaked dipole at $x_{\pm 1}, \quad u_N(x_\nu, 0) = (2N + 1)\delta_{|\nu|,1}$. According to (3.3.53), the evolution of these components in time yields

$$b_k(t) \sim \sin(\xi_k e^{-t}) + \mathrm{Const}_k \frac{(-1)^k}{\Delta\xi}\left(1 - \frac{1}{\xi_k e^{-t}}\right)_+ + \mathcal{O}(\Delta\xi). \tag{3.3.54}$$

In this case the $\mathcal{O}(N)$ oscillatory boundary wave, $\frac{(-1)^k}{\Delta\xi}\left(1 - \frac{1}{\xi_k e^{-t}}\right)_+$, is added to the $\mathcal{O}(1)$-initial

conditions, $\sin(\xi_k)$, which is responsible for the L^2-growth of order $\mathcal{O}(N)$. This linear L^2-growth is even more apparent with the 'rough' initial data we met earlier in Figure 3.1. ■

Remarks

1. Smoothing. The last Theorem confirms the L^2-instability indicated previously by the lower bound (3.3.40),

$$\frac{d}{dt}\|b(t)\|^2 \geq -\|b(t)\|^2 + N\left[b_{-N}^2 + b_N^2\right].$$

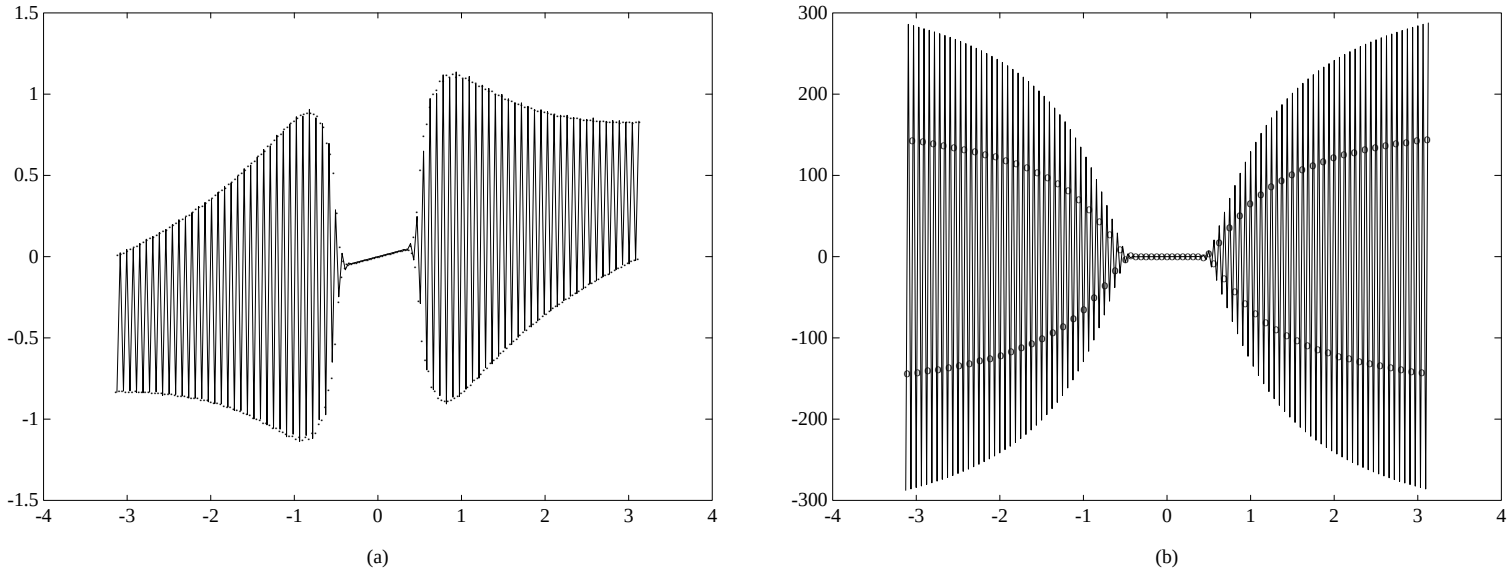
By the same token, summation by parts of the imaginary part (3.3.39), leads to the *upper* bound

$$\frac{d}{dt}\|b(t)\|^2 \leq \|b(t)\|^2 + N\left[b_{-N}^2 + b_N^2\right],$$

which shows that *had* the boundary values of the computed spectrum — which in this case consist of the last single mode $b_{\pm N}(t)$, were to remain relatively small, then the imaginary part – and consequently the whole Fourier approximation would have been L^2-stable. For example, the rather weak *a priori* bound will suffice

$$|b_{\pm N}(t)| \leq \frac{\mathrm{C}}{\sqrt{N}}\|b(0)\| \Longrightarrow \|b(t)\| \leq e^{(1/2+\mathrm{C}^2)t}\|b(0)\|. \tag{3.3.55}$$

What we have shown (in the second part of Theorem 3.2) is that such an a priori bound does not hold for general nonsmooth L^2-initial data, where according to (3.3.53), $b_N(t) \sim \mathcal{O}(N)\|b(t)\|$.



Imaginary part of Fourier coefficients, $\mathit{Im}\ \hat{u}_k(t)$ vs. $k\pi\Delta\xi$, computed at $t = 2$
(a) with de-aliasing ($N = 80$ and $N = 160$) (b) without de-aliasing ($N = 50$ and $N = 100$)

Figure 3.4: Fourier solution of $u_t = (\sin(x)u)_x$, $\hat{u}(\xi, 0) = \sin(\xi)$.

We recall that there are various procedures which enforce stability of the Fourier method, without sacrificing its high order accuracy. One possibility is to use the skew-symmetric formulation – consult §3.4 below. Another possibility is based on the observation that the current instability is due to the inflow-dependent boundary conditions (3.3.42) — or equivalently (3.3.38), and the origin of the latter could be traced back to the aliasing relations (2.2.7). We can therefore de-alias and hence by (3.3.55) stabilize the Fourier method by setting $b_{\pm N}(t) \equiv 0$, or more generally, $\hat{u}_{\pm N}(t) \equiv 0$. De-aliasing could be viewed as a robust form of high-frequency smoothing. This issue is dealt in §3.5 below. Figure 3.4a shows how the de-aliasing procedure (— setting $b_{\pm N}(t) \equiv 0$), stabilizes the Fourier method which otherwise experiences the unstable linear growth in Figure 3.4b. With (3.3.55) in mind, we may interpret

2N	64	128	256	512
$\frac{\|u_N(t)\|}{\|u_N(0)\|}$	366	712	1906	5152

Table 3.2: *Amplification of* $\|u_N(t)\|$ *at* $t = 5$ *with even number of gridpoints. Here,* $\frac{\partial}{\partial t}u_N(x,t) = \frac{\partial}{\partial x}\psi_N(\sin(2x)u_N(x,t)), \quad u_N(x,0) = \sin(x)$.

these procedures as a mean to provide the missing *a priori* decaying bounds on the highest mode(s) of the computed spectrum, which in turn guarantee the stability of the whole Fourier approximation.

2. Smoothing cont'd – even number of gridpoints. The situation described in the previous remark is a special case of the following assertion: *Assume that* $a(x)$ *consists of a finite number, say* m *modes. Then the corresponding Fourier approximation* (3.2.16) *is* L^2*-stable, provided the last* m *modes were filtered so that the following a priori bound holds*

$$\sum_{|k|>N-m}^{N} |\hat{u}_k(t)|^2 \leq \frac{1}{N}\|b(0)\|^2.$$

It should be noted that our present discussion of $a(x)$ with $m = 1$ modes is a prototype case for the behavior of the Fourier method, as long as the corresponding Fourier approximation is based on an *odd* number of $2N + 1$ gridpoints; otherwise the case of an even number of gridpoints is L^2-stable. The unique feature of this L^2-stability is due to the fact that Fourier differentiation matrix in this case, $D_{jk} = \frac{(-1)^{j-k}}{2}\cot(\frac{x_j - x_k}{2})(1 - \delta_{jk})$ — being *even* order antisymmetric matrix, must have zero as a *double* eigenvalue, which in turn inflicts a 'built-in' smoothing of the last mode in this case, namely,

$$b_{\pm N}(t) \equiv 0. \tag{3.3.56}$$

Table 3.2 confirms the usual linear weak L^2-instability already for a 2-wave coefficient.

3. W^α-initial data. Consider the case of sufficiently smooth initial data so that the imaginary components decay of order α,

$$b_k(0) \sim |k|^{-\alpha}, \quad \alpha > \frac{1}{2}.$$

In this case, we may approximate the corresponding initial interpolant $b(\xi) \sim (\Delta\xi/\xi)^\alpha$, and (3.3.53) tells us the Fourier approximation takes the approximate form

$$b_k(t) = \frac{e^{\alpha t}}{k^\alpha} + \frac{2(-1)^k}{\Delta\xi}\int_{\xi \geq e^{-t}/\xi_k}^{1} \frac{(\Delta\xi)^\alpha}{\xi^{\alpha+1}}d\xi + \mathcal{O}(\Delta\xi) \sim \frac{e^{\alpha t}}{k^\alpha} + \frac{(-1)^k}{N^{\alpha-1}}\left[\left(\frac{ke^t}{N}\right)^\alpha - 1\right]_+ + \mathcal{O}(\Delta\xi).$$

Observe that $\|b(t)\| \sim \mathrm{C}_\alpha N^{\frac{3}{2}-\alpha}$, (with $\mathrm{C}_\alpha \sim (e^{2\alpha t} - 1)/(2\alpha + 1)$), where as $\|b(0)\|_{W^\alpha} \sim \sqrt{N}$. This lower bound is found to be in complete agreement with the W^α-stability statement of Corollary 3.1 (apart from the $\log N$ factor for $\alpha = 1$) — an enjoyable sharpness.

3.3.3 Epilogue

On previous subsections we analyzed the stability of Fourier method in terms of two main ingredients: weighted L^2-stability on the one hand, and high frequencies instability on the other hand. Here we would like to show how *both* of these ingredients contribute to the actual performance of the Fourier method.

We first address the issue of *resolution*. We were left with the impression that the weak L^2-instability is a rather 'rare occurrence', as it is excited only in the presence of nonsmooth initial data. But in fact, the mechanism of this weak L^2-instability will be excited whenever the Fourier method lacks enough resolution.
In this context let us first note that the solution of the underlying hyperbolic problem may develop large spatial gradients due to the almost impinging characteristics along the zeroes of the increasing part of $a(x)$. Consequently, the Fourier method might not have enough modes to resolve these large gradients as they grow in time. This tells us that independent whether the initial data are smooth or not, the computed approximation will then 'see' the underlying solution as a nonsmooth one, and this lack of resolution will be recorded by a slower decay of the computed Fourier modes. The latter will experience the high-frequency instability discussed earlier and this in turn will lead to the linear L^2-growth. Our prototype example of $a(x) = \sin(x)$ is case in point: according to Corollary 3.2, one needs here at least $N >> e^t$ modes in order to resolve the solution, for otherwise, (3.3.53) shows that spurious $\mathcal{O}(N)$ oscillations will contaminate the whole computed spectrum.

We conclude that the lack of resolution manifests itself as a weak L^2-instability. This phenomenon is demonstrated in Figures 3.5-3.9, describing the Fourier method (3.3.1) subject to (the perfectly smooth ...) initial condition, $u(x,0) = \sin(x)$. Figure 3.5 shows how the Fourier method with fixed number of $N = 64$ modes propagates information regarding the steepening of the Fourier solution in physical space, from low modes to the high ones. And, as this information is being transferred to the high modes, their $\mathcal{O}(N)$ amplification become more noticeable as time progresses in Figures 3.5a-3.5d. Consequently, though $N = 64$ modes are sufficient to resolve the exact solution at $t \leq 2.7$, Figure 3.6c-d shows that at later time, $t = 3$ and in particular $t = 5$, the under resolved Fourier solution with 64-modes will be completely dominated by the spurious centered spike. This loss of resolution requires more modes as time progresses. Figure 3.7 shows how the Fourier method is able to resolve the exact solution at $t = 3.5$, once 'sufficiently many' modes, $N >> e^{3.5}$ are used, in agreement with Corollary 3.3. According to Figures 3.8 and 3.9, $N = 512 >> e^4$ modes are required to correctly resolve the two strong boundary dipoles at $t = 4$, yet at $t = 8$ the Fourier solution will be completely dominated by the spurious centered spike.

Assuming that the Fourier method contains sufficiently many modes dictated by the requirement of resolution, we now turn to the second issue of this section concerning the *convergence of the Fourier method*.

Theorem 3.3 (Convergence rate estimate) *Let $u_N(x,t)$ denotes the N-degree Fourier approximation of the corresponding exact solution $u(x,t)$. Then the following error estimate holds*

$$\|u_N(\cdot,t) - u(\cdot,t)\|_{W^s} \leq \text{Const}_{s,\alpha} N^{2-\alpha} \|u(\cdot,0)\|_{W^{s+\alpha}}, \quad \forall s + \alpha > \frac{1}{2}. \tag{3.3.57}$$

Remark. The requirement from the initial data to have at least $W^{1/2}$-regularity is clearly necessary in order to make sense of its *pointwise* interpolant.

3.4 Skew-Symmetric Differencing

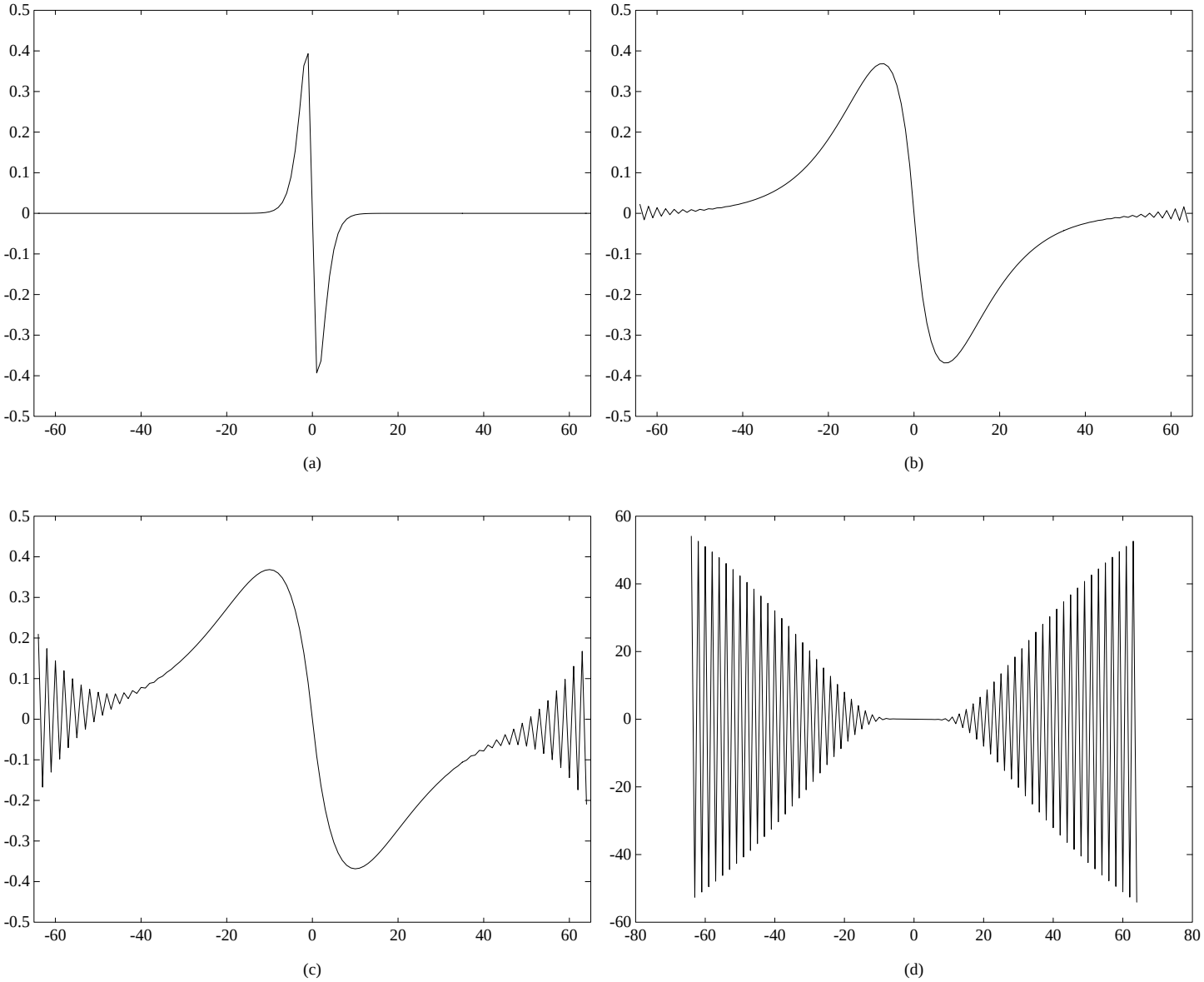
There are two main approaches to enforce stability at this point: skew-symmetric differencing and smoothing. We discuss these issues in the next two subsections.

The essential argument of well-posedness for symmetric hyperbolic systems with constant coefficients is the fact that (say in the 1-D case) $P(D) = A\frac{\partial}{\partial x}$ is a skew-adjoint operator. With variable coefficients this is also true, modulo low-order bounded terms, i.e.,

$$P(x,t,D) \equiv A(x,t)\frac{\partial}{\partial x} = \frac{1}{2}\left[A(x,t)\frac{\partial}{\partial x} + \frac{\partial}{\partial x}(a(x,t)\cdot)\right] - \frac{1}{2}A_x(x,t). \tag{3.4.1}$$

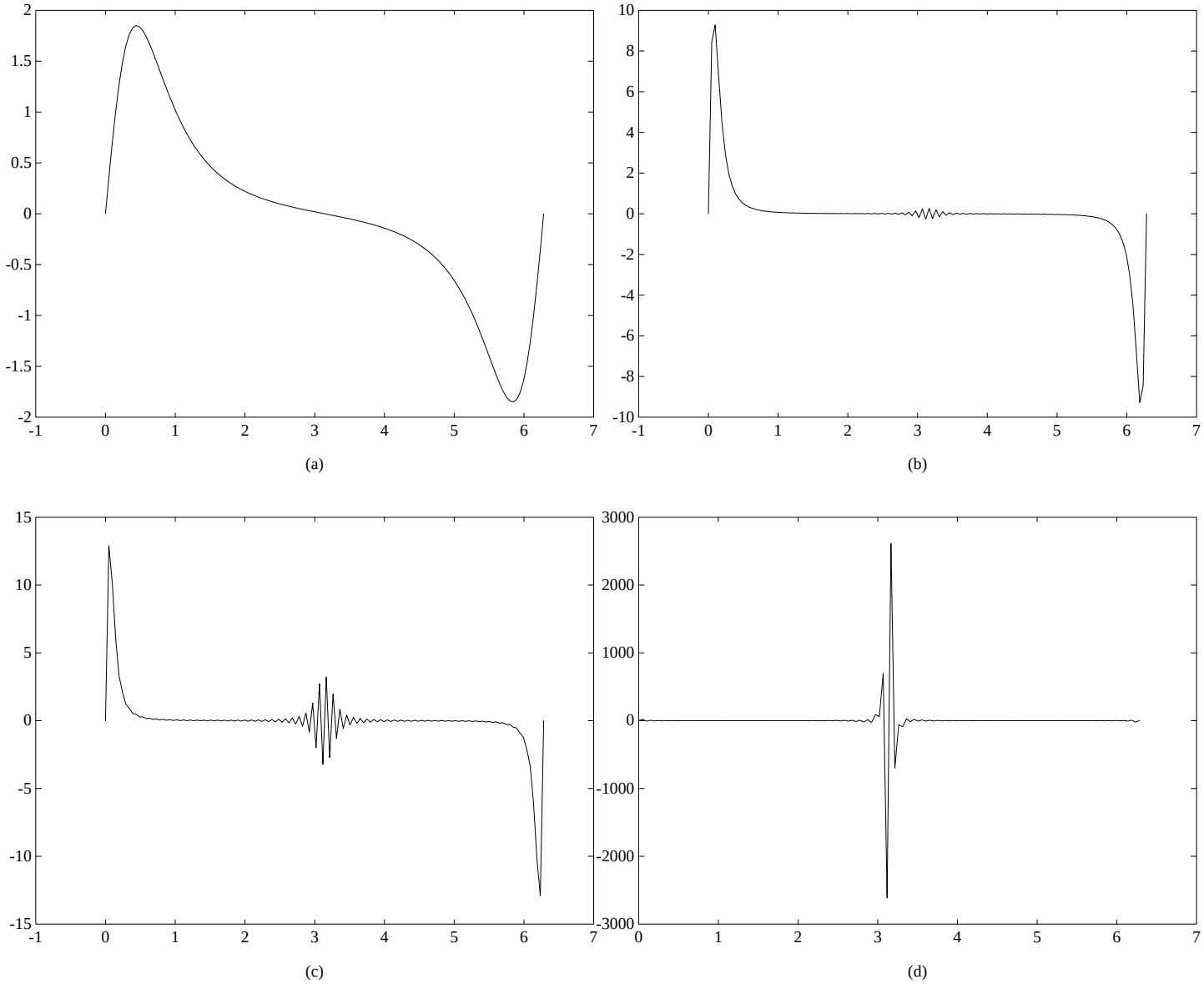
The stability proofs of spectral methods follow the same line, i.e., we have in the Fourier space, compare (3.1.50),

$$\hat{A}_N\Lambda = \frac{1}{2}\left[\hat{A}_N\Lambda - \Lambda\hat{A}_N\right] + \frac{1}{2}\left[\hat{A}_N\Lambda + \Lambda^*\hat{A}_N\right] \tag{3.4.2}$$



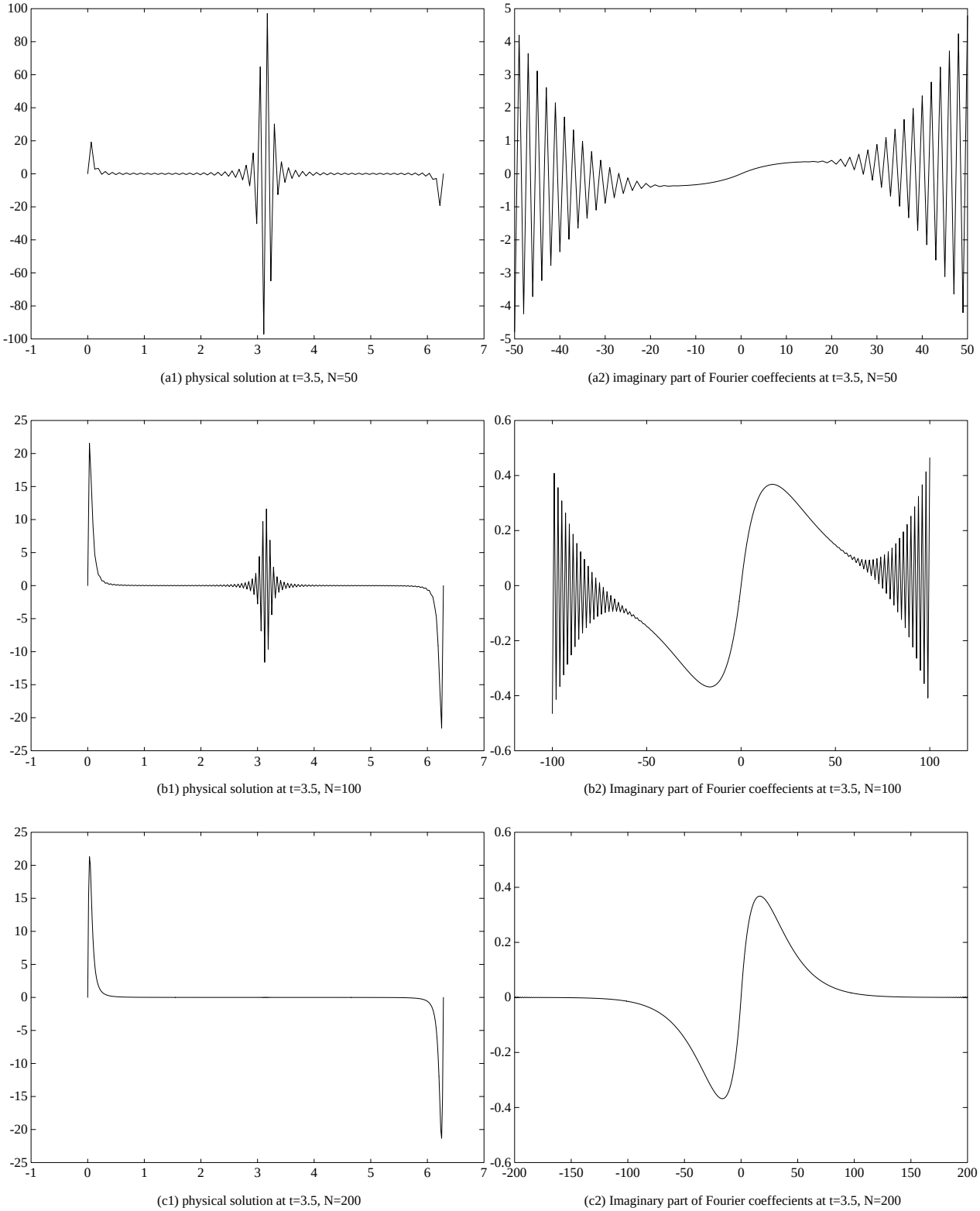
Imaginary part of Fourier coefficients, $\mathbb{I}m\ \hat{u}_k(t)$, computed with $N = 64$ modes at
(a) $t = 1.0$ (b) $t = 2.7$
(c) $t = 3.0$ (d) $t = 5.0$

Figure 3.5: Fourier solution of $u_t = (\sin(x)u)_x, \quad u(x,0) = \sin(x)$.



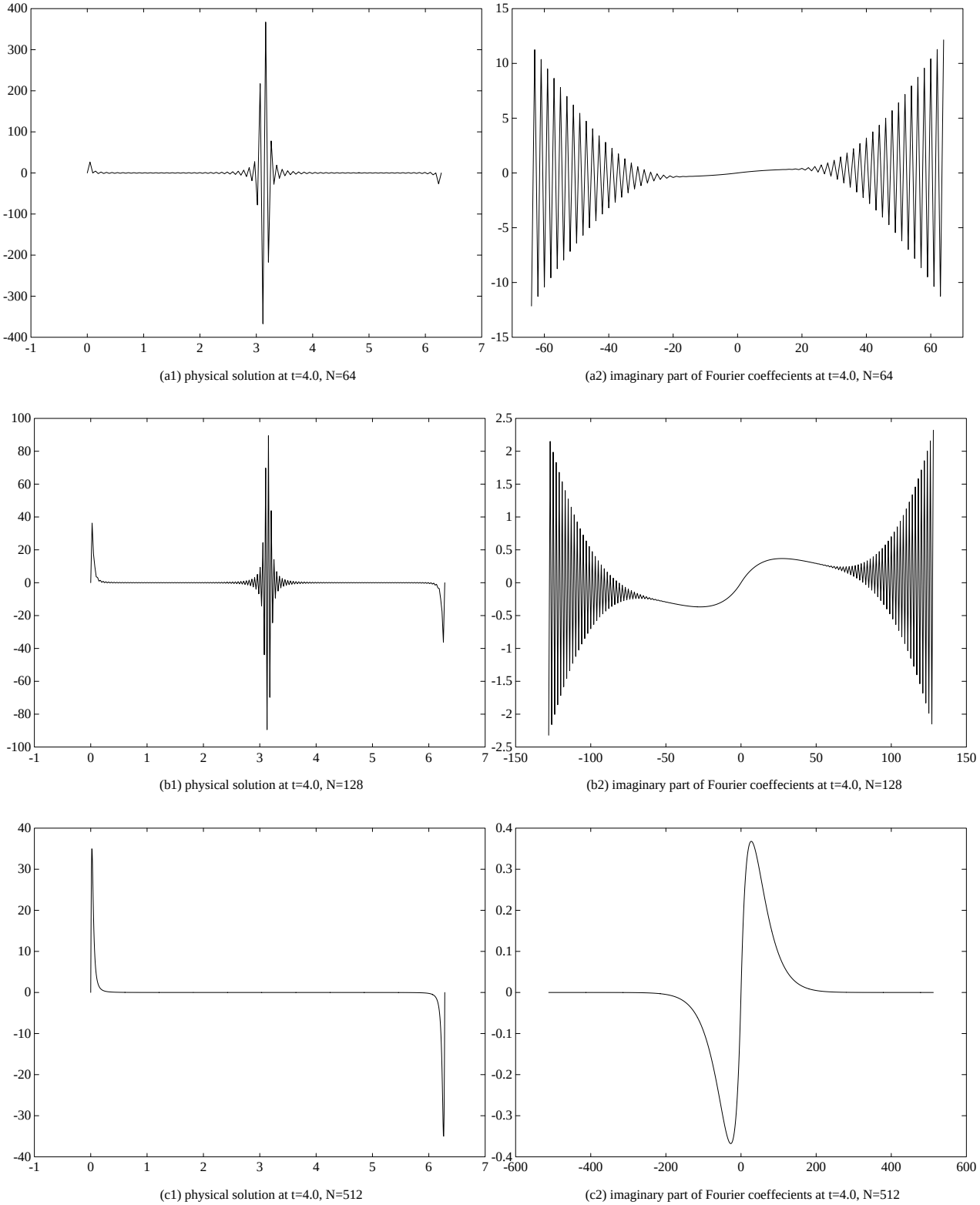
Computed solution, $u_N(\cdot,t)$, with $N = 64$ modes at
(a) $t = 1.0$ (b) $t = 2.7$
(c) $t = 3.0$ (d) $t = 5.0$

Figure 3.6: Fourier solution of $u_t = (\sin(x)u)_x, u(x,0) = \sin(x)$.



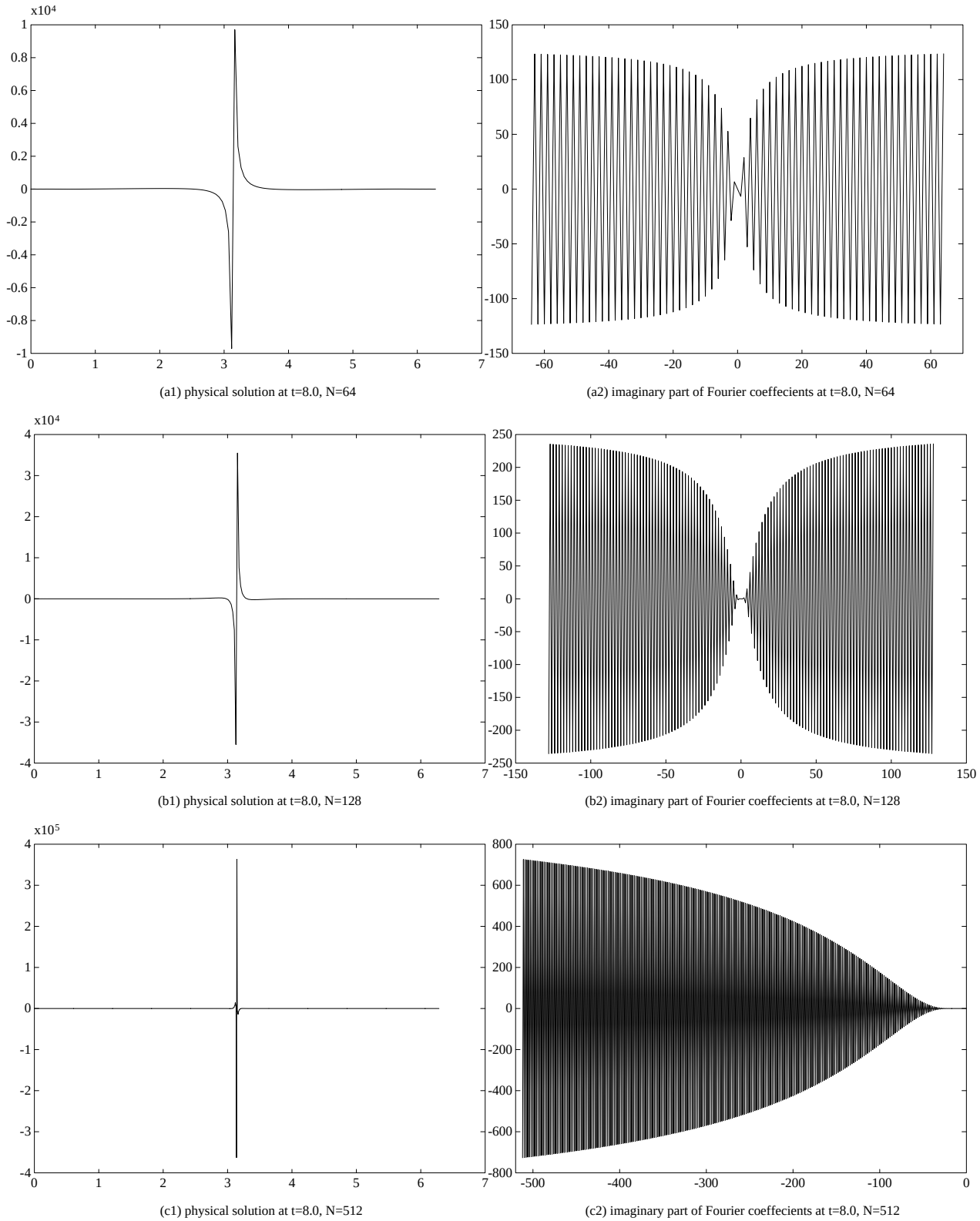
Approximate solution, $u_N(\cdot,t)$ and imaginary part of its Fourier coefficients, $\mathit{Im}\ \hat{u}_k(t)$ at $t = 3.5$
(a) with $N = 50$ (b) with $N = 100$ (c) with $N = 200$

Figure 3.7: Fourier solution of $u_t = (\sin(x)u)_x, \quad u(x,0) = \sin(x)$.



Approximate solution, $u_N(\cdot,t)$ and imaginary part of its Fourier coefficients, $\mathit{Im}\ \hat{u}_k(t)$ at $t = 4.0$
(a) with $N = 64$ (b) with $N = 128$ (c) with $N = 512$

Figure 3.8: Fourier solution of $u_t = (\sin(x)u)_x, \quad u(x,0) = \sin(x)$.



(a1) physical solution at t=8.0, N=64

(a2) imaginary part of Fourier coeffecients at t=8.0, N=64

(b1) physical solution at t=8.0, N=128

(b2) imaginary part of Fourier coeffecients at t=8.0, N=128

(c1) physical solution at t=8.0, N=512

(c2) imaginary part of Fourier coeffecients at t=8.0, N=512

Approximate solution, $u_N(\cdot,t)$, and imaginary part of its Fourier coefficients, $\mathit{Im}\ \hat{u}_k(t)$ at $t = 8.0$
(a) with $N = 64$ (b) with $N = 128$ (c) with $N = 512$

Figure 3.9: Fourier solution of $u_t = (\sin(x)u)_x, \quad u(x,0) = \sin(x)$.

and stability amounts to show that the second term in (3.4.2) is bounded: for then we have in (3.4.2) (as in ((3.4.1)) a skew-adjoint term with an additional bounded operator. The difficulty with the stability of pseudo-spectral methods arises from the fact that the second term on the right of (3.4.2) is unbounded,

$$\lim_{N\to\infty}\|\frac{1}{2}(\tilde{A}_N\Lambda+\Lambda^*\tilde{A}_N)\|\uparrow\infty. \tag{3.4.3}$$

To overcome this difficulty, we can discretized the symmetric hyperbolic system (again, say the 1-D case)

$$\frac{\partial u}{\partial t}=A(x,t)\frac{\partial u}{\partial x} \tag{3.4.4}$$

when the spatial operator is already put in the "right" skew-adjoint form, compare (3.4.1),

$$\frac{\partial u}{\partial t}=\frac{1}{2}\left[A(x,t)\frac{\partial u}{\partial x}+\frac{\partial}{\partial x}(A(x,t)u)\right]-\frac{1}{2}A_x(x,t)u.$$

The pseudospectral approximation takes the form

$$\frac{\partial u_N}{\partial t}=\frac{1}{2}\left[\psi_N\left(A(x,t)\frac{\partial u_N}{\partial x}\right)+\frac{\partial}{\partial x}\psi_N(A(x,t)u_N)\right]-\frac{1}{2}\psi_N(A_x(x,t)u_N). \tag{3.4.5}$$

In the Fourier space, this gives us

$$\frac{d\tilde{v}}{dt}=\frac{1}{2}[\tilde{A}_N\Lambda+\Lambda\tilde{A}_N]\tilde{v}-\frac{1}{2}\frac{\partial A_N}{\partial x}\tilde{v}. \tag{3.4.6}$$

Now, $\tilde{A}_N\Lambda+\Lambda\tilde{A}_N$ is symmetric because Λ is, $\frac{\partial A_N}{\partial x}$ is bounded and stability follows.

3.5 Smoothing

We have already met the process of smoothing in connection with the heat equation: starting with bounded initial data, $f(x)$, the solution of the heat equation (1.2.1)

$$u(x,t)=Q*f(x),\quad Q(x)=\frac{1}{\sqrt{4\pi a}}e^{\frac{-x^2}{4at}},\quad t>0 \tag{3.5.1}$$

represents the effect of smoothing $f(x)$, so that $u(\cdot,t>0)\epsilon C^\infty$ (in fact analytic) and $u(x,t\downarrow 0)=f(x)$.

A general process of smoothing can be accomplished by convolution with appropriate smoothing kernel $Q(x)$

$$f_\varepsilon(x)=Q_\varepsilon(x)*f(x) \tag{3.5.2}$$

such that $Q_\varepsilon(x)*f(x)$ is sufficiently smoother than $f(x)$ is, and

$$Q_\varepsilon(x)*f(x)\underset{\varepsilon\to 0}{\longrightarrow}f(x). \tag{3.5.3}$$

With the heat kernel, the role of ε is played by time $t>0$. A standard way to construct such filters is the following. We start with a C^s-function supported on, say, (-1,1), such that it has a unit mass and zero first r moments,

$$\int_{-1}^{1}Q(x)dx=1,\quad \int_{-1}^{1}x^j\phi(x)dx=0,\quad j=1,2,\cdots,r. \tag{3.5.4}$$

Then we set $Q_\varepsilon(x)=\frac{1}{\varepsilon}Q\left(\frac{x}{\varepsilon}\right)$ and consider

$$f_\varepsilon(x)=Q_\varepsilon(x)*f(x),\quad \varepsilon>0. \tag{3.5.5}$$

Now, assume f is $(r+1)$ – differentiable in the ε neighborhood of x; then, since $Q_\varepsilon(x)$ is supported on $(-\varepsilon,\varepsilon)$ and satisfies (3.5.4) as well, we have by Taylor expansion

$$\begin{aligned} f(x) - Q_\varepsilon(x) * f(x) &= \int_{|y|\le\varepsilon} Q_\varepsilon(y)[f(x) - f(x-y)]dy = \\ &= \int_{|y|\le\varepsilon} Q_\varepsilon(y) \left[\sum_{j=1}^{r} \frac{(-y)^j}{j!} f^{(j)}(x) + \frac{(-y)^{r+1}}{(r+1)!} f^{r+1}(\xi)\right] dy. \end{aligned} \tag{3.5.6}$$

The first r moments of $Q_\varepsilon(y)$ vanish and we are left with

$$|f(x) - Q_\varepsilon(x) * f(x)| \le \text{Const.} \max_{|y-x|\le\varepsilon} |f^{(r+1)}(y) \cdot \varepsilon^{r+1}, \tag{3.5.7}$$

i.e., $f_\varepsilon(x)$ converges to $f(x)$ with order $r+1$ as $\varepsilon \to 0$. Moreover, $f_\varepsilon(x)$ is as smooth as $\phi(x)$ is, since

$$f_\varepsilon(x) = \int_y \frac{1}{\varepsilon} Q\left(\frac{x-y}{\varepsilon}\right) f(y)dy \tag{3.5.8}$$

has many bounded derivatives as Q has, i.e., starting with differentiable function f of order $r+1$ in the neighborhood of x, we end up with regularized function $f_\varepsilon(x)$ in $C^s, s > r$.

Example: For C^∞ regularization – choose a unit mass C^∞ kernel, see Figure 3.10,

$$Q(x) = \begin{cases} Q_0 e^{-\frac{1}{1-x^2}}, & |x| < 1 \\ 0, & |x| \ge 1 \end{cases} \quad \text{with } Q_0 \text{ such that } \int Q(x)dx = 1. \tag{3.5.9}$$

Then $f_\varepsilon(x) = Q_\varepsilon(x) * f(x)$ is a C^∞ regularization of $f(x)$ with first order convergence rate

$$|f(x) - f_\varepsilon(x)| \le \text{Const.} \max_{|y-x|\le\varepsilon} |f'(y)| \cdot \varepsilon \to 0.$$

To increase the order of convergence, one requires more vanishing moments, (3.5.4),(which yield more oscillatory kernels). We note that this smoothing process is purely local — it involves ε-neighboring values of C^{r+1} function f, in order to yield a C^s-regularized function $f_\varepsilon(x)$ with $f_\varepsilon(x) \longrightarrow f(x)$. The convergence rate here is $r+1$.

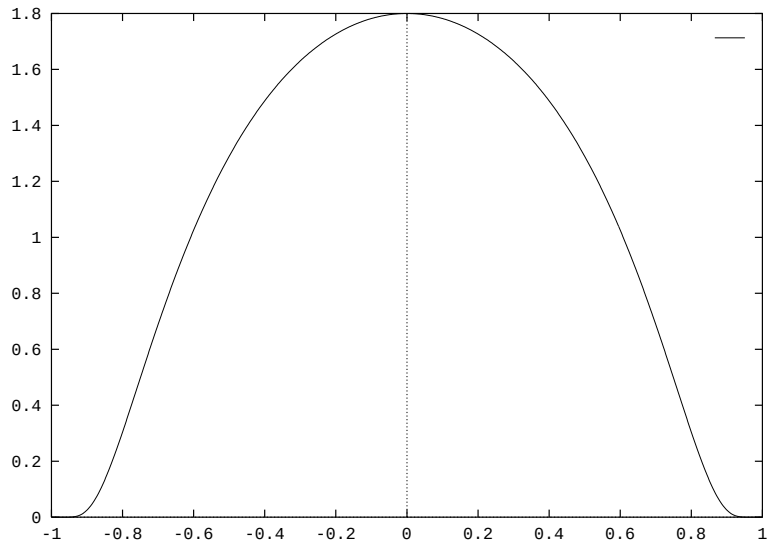


Figure 3.10: Unit mass mollifiers

We can also achieve local regularization with spectral convergence. To this end we set

$$Q_N(x) = \frac{1}{\theta}\rho(\frac{x}{\theta})D_m(\frac{x}{\theta}), \quad \rho(0) = 1, \tag{3.5.10}$$

where $\rho(x)$ is a C^∞-function supported on $(-\pi, \pi)$. Figure 3.11 demonstrates such a mollifier. In this case the support of the mollifier is kept fixed; instead, by increasing m — particularly, by allowing $m = m_N$ to increase together with N, we obtain a highly oscillatory kernel whose monomial moments satisfy (3.5.4) modulo a spectrally small error.

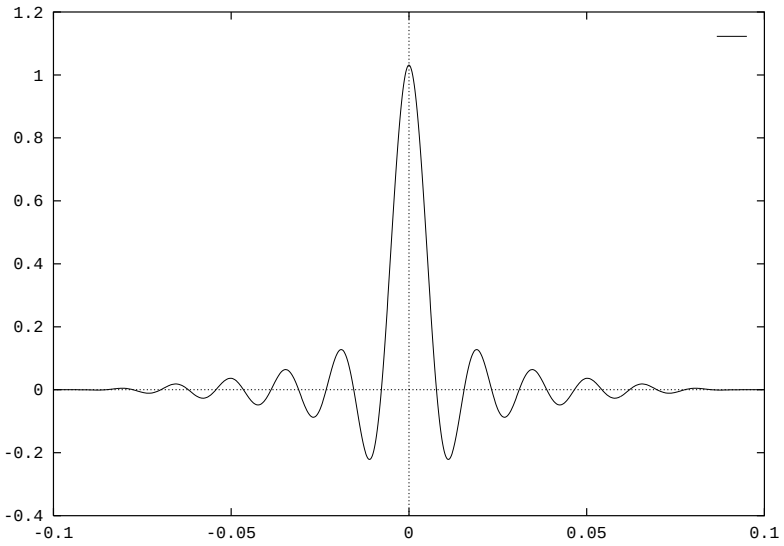


Figure 3.11: A spectral unit mass mollifier

Consider now

$$f_N(x) = Q_N * f(x). \tag{3.5.11}$$

Then $f_N(x)$ is C^∞ because Q_N is; and the convergence rate is spectral, since by (2.1.34)

$$\begin{aligned} f(x) - Q_N * f(x) &= f(x) - \int_{|y|\le\pi} D_m(y)\rho(y)f(x-\theta y)dy \\ &= f(x) - \rho(y)f(x-\theta y)|_{y=0} + \text{residual}, \end{aligned} \tag{3.5.12}$$

and since $\rho(0)$ was chosen as $\rho(0) = 1$ we are left with a residual term which does not exceed

$$|\text{residual}| \le \text{Const.}\|\rho(\cdot)f(x-\theta\cdot)\|_{H^s(-\pi,\pi)}\frac{1}{m^{s-1}}, \quad \forall s > 0.$$

Thus, the convergence rate is as fast as the *local* smoothness of f permits;(in this case – the local neighborhood $[x - \pi\theta, x + \pi\theta]$). Of course, with $\theta = \rho \equiv 1$ we recover the *global* C^∞-regularization due to the spectral projection. The role of ρ was to *localize* this process of spectral smoothing.

We can as easily implement such smoothing in the Fourier space: For example, with the heat kernel we have

$$\hat{u}(k,t) = e^{-ak^2t}\hat{f}(k) \tag{3.5.13}$$

so that $\hat{u}(k,t)$ for any $t > 0$ decay faster than exponential and hence $u(x, t > 0)$ belong to H^s for any s (and by Sobolev embedding, therefore, is in C^∞ and in fact analytic). In general we apply,

$$f_\varepsilon(x) = \sum_{k=-\infty}^{\infty}\hat{Q}_\varepsilon(k)\hat{f}(k)e^{ikx} \tag{3.5.14}$$

such that for $f_\varepsilon(x)$ to be in H^s we require

$$\sum_{k=-\alpha}^{\infty}(1+|k|^2)^s|Q_\varepsilon(k)|^2(\hat{f}(k)|^2 \le \text{Const.}$$

and $r+1$ order of convergence follows with

$$|\hat{\phi}_\varepsilon(k)-1| \le \text{Const.}(\varepsilon k)^{r+1}. \tag{3.5.15}$$

Indeed, (3.5.15) implies

$$\begin{aligned} |f(x)-f_\varepsilon(x)| &\le \text{Const.}\varepsilon^{r+1}\sum_{k=-\alpha}^{\infty}|k^{r+1}\hat{f}(k)e^{ikx}| \\ &\le \text{Const. } \max|f^{(r+1)}|\cdot\varepsilon^{r+1}. \end{aligned} \tag{3.5.16}$$

Note: Since $\hat{\phi}_\varepsilon(k)\downarrow 0$ we can deal with any unbounded f by splitting $\sum_{|k|\le|k_0|}+\sum_{|k|>|k_0|}$. To obtain spectral accuracy we may use

$$\hat{Q}_N(k)=\begin{cases}\equiv 1, & |k|<m_N \\ \sim \text{smoothly decay to zero} & m_N\le|k|\le N\end{cases}. \tag{3.5.17}$$

Clearly $Q_N * f(x)$ is C^∞ and the familiar Fourier estimates give us

$$|f(x)-Q_N*f(x)| \le \sum_{|k|>m_N}|\hat{f}(k)e^{ikx}| \le \text{Const.}\|f\|_{H^s}\cdot\frac{1}{m_N^{s-1}}.$$

We emphasize that this kind of smoothing in the Fourier space need not be local; rather $Q_\varepsilon(x)$ or $\phi_N(x)$ are negligibly small away from a small interval centered around the origin depending on ε or $\frac{1}{N}$. (This is due to the uncertainty principle.)

The smoothed version of the pseudospectral approximation of (3.2.15) reads

$$\frac{\partial v_N}{\partial t}=\psi_N(a(x,t)\frac{\partial}{\partial x}(Q*v_N)) \tag{3.5.18}$$

i.e., in each step we smooth the solution either in the real space (convolution) or in the Fourier space (cutting high modes).[11] We claim that this smoothed version is stable hence convergent under very mild assumptions on the smoothing kernel $Q_N(x)$. Specifically, (3.5.18) amounts in the Fourier space, compare (3.2.3)

$$\frac{\partial\tilde{v}}{\partial t}=\tilde{A}_N\Lambda Q_N\tilde{v}. \tag{3.5.19}$$

The real part of the matrix in question is given by

$$[\text{Re}\tilde{A}_N\Lambda Q_N]_{kj}=i(\lambda_k-\lambda_j)\sum_p\hat{a}[k-j+p(2N+1)], \quad -N\le k,j\le N \tag{3.5.20}$$

where $\Lambda Q_N=\text{diag}_k(i\lambda_k)$

$$i\lambda_k=ik\hat{Q}_N(k)$$

is interpreted as the smoothed differentiation operator. Now, looking at (3.5.20) we note:

1. For $p=0$ we are back at the spectral analysis, compare (3.1.59), (3.1.60) and the real part of the matrix in (3.5.20) – the aliasing free one – is bounded.

[11] Either one can be carried out efficiently by the FFT.

2. We are left with $|p| = 1$: in the unsmoothed version, these terms were unbounded since $|\lambda_k - \lambda j| \uparrow \infty$ as $k \downarrow -N$ or $j \uparrow N$. With the smoothed version, these terms are bounded (and stability follows), provided we have

$$|\lambda_k = ik\hat{Q}_N(k)| \underset{|k|\uparrow N}{\longrightarrow} 0 \ . \tag{3.5.21}$$

For example, consider the smoothing kernel $Q_N(x)$ where

$$\hat{Q}_N(k) = \frac{\sin kh}{kh}, \quad h = \frac{2\pi}{2N+1}.$$

This yields the smoothed differentiation symbols

$$\lambda_k = i \sin \frac{kh}{h} \tag{3.5.22}$$

which corresponds to the second order center differencing in (2.2.36); stability is immediate by (3.5.21) for

$$|\lambda_k \equiv \frac{\sin \frac{2\pi k}{2N+1}}{\frac{2\pi k}{2N+1}}| \underset{|k|\uparrow N}{\longrightarrow} 0 \ . \tag{3.5.23}$$

Yet, this kind of smoothing reduces the overall spectral accuracy to a second one; a fourth order smoothing will be

$$\begin{aligned} \lambda_k &= i\tfrac{1}{3}\left[4\sin\tfrac{kh}{h} - \sin\tfrac{2kh}{2h}\right], \quad \longleftrightarrow \quad \hat{Q}_N(k) = \tfrac{\lambda_k}{ik} \\ \lambda_k &= \tfrac{6i}{h}\tfrac{\sin kh}{4+2\cos kh}, \qquad\qquad \longleftrightarrow \quad \hat{Q}_N(k) = \tfrac{\lambda_k}{ik}. \end{aligned} \tag{3.5.24}$$

In general, the accuracy is determined by the low modes while stability has to do with high ones. To entertain spectral accuracy we may consider smoothing kernels other than trigonometric polynomials ($\equiv$ finite difference), but rather, compare (3.5.17)

$$\hat{Q}_N(k) = \begin{cases} \equiv 1, & |k| \le m_N \\ \\ \sim \text{smoothly decay to zero} & m_N < |k| \le N. \end{cases} \tag{3.5.25}$$

An increasing portion of the spectrum is differentiated *exactly* which yields spectral accuracy; the *highest* modes are not amplified because of the smoothing effect in this part of the spectrum.

We close this section noting that if the *differential* model contains some dissipation – e.g., the parabolic equation

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(a(x,t)\frac{\partial u}{\partial x}\right), \quad a(x,t) \ge \alpha > 0, \tag{3.5.26}$$

then stability follows with no extra smoothing. The parabolic dissipation compensates for the loss of "one derivative" due to aliasing in first order terms. To see this we proceed as follows: multiply

$$\frac{\partial v_N}{\partial t}(x_\nu, t) = \frac{\partial}{\partial x}\left[a(x_\nu, t)\frac{\partial v_N}{\partial x}(x_\nu, t)\right] \tag{3.5.27}$$

by $v_N(x_\nu t)$ and sum to obtain

$$\frac{1}{2}\frac{d}{dt}\sum_\nu v_N^2(x_\nu, t) = \sum_\nu v_N(x_\nu, t)\frac{\partial}{\partial x}(a(x_\nu, t)\frac{\partial v_N}{\partial x}(x_\nu, t)). \tag{3.5.28}$$

Suppressing excessive indices, $v_N(x_\nu, t) \equiv v_\nu(t)$, we have for the RHS of (3.5.28)

$$\sum_\nu v_\nu \frac{\partial}{\partial x}\left(a_\nu(t)\frac{\partial v_\nu}{\partial x}\right) = \frac{1}{2}\sum_\nu \frac{\partial}{\partial x}\left(a_\nu(t)\frac{\partial v_\nu^2}{\partial x}\right) - \sum_\nu a_\nu(t)\left(\frac{\partial v_\nu}{\partial x}\right)^2. \tag{3.5.29}$$

Now, the first sum on the right gives us the usual loss of one derivative and the second are compensates with gain of such quantity. Petrovski type stability (gain of derivatives) follows. We shall only sketch the details here. Starting with the first term on the right of (3.5.29) we have

$$\frac{\pi}{2N+1}\sum \frac{\partial}{\partial x}\left(a_\nu v_\nu \frac{\partial v_\nu}{\partial x}\right) = \frac{1}{2}\int \frac{\partial}{\partial x}[\cdots] + \frac{1}{2}\cdot \text{ [aliasing errors]} \tag{3.5.30}$$

while for the second term

$$-\sum a_\nu(t)\left(\frac{\partial v_\nu}{\partial x}\right)^2 \le -\alpha \int \left[\frac{\partial v_N}{\partial x}(x,t)\right]^2 dx \tag{3.5.31}$$

and this last term dominates the RHS of (3.5.30).

4 THE CHEBYSHEV METHOD

4.1 Forward Euler — the CFL Condition

We are concerned here with *fully-discrete* spectral/pseudospectral approximations to initial-boundary value problems associated with hyperbolic equations. In this context, the spectral (and respectively, the pseudospectral) approximations consist of truncation (and, respectively, collocation) of N-term spatial expansions, which are expressed in terms of general Jacobi polynomials; Chebyshev and Legendre expansions are the ones most frequently found in practice. We will show that such N-term approximations are stable, provided their time step, Δt, fulfills the *CFL-like condition*, $\Delta t \leq \text{Const} \cdot N^{-2}$.

To clarify the origin of such a CFL-like condition in our case, we recall that the Jacobi polynomials are in fact the eigenfunctions of second-order singular Sturm-Liouville problems. Our arguments show that the main reason for the above CFL limitation is the $O(N^2)$ growth of the Nth eigenvalue associated with these Sturm-Liouville problems.

We start with the scalar constant-coefficient hyperbolic equation,

$$u_t = a u_x, \quad (x,t)\epsilon[-1,1]\times[0,\infty), \quad a>0, \tag{4.1.1}$$

which is augmented with homogeneous conditions at the inflow boundary,

$$u(1,t) = 0, \quad t>0. \tag{4.1.2}$$

To approximate (4.1.1), we use forward Euler time-differencing on the left, and either spectral or ψdospectral differencing on the right. Thus, we seek a temporal sequence of spatial π_N-polynomials, $v^m = v_N(x, t^m = m\Delta t)$, such that

$$v_N(x,t^m+\Delta t) = v_N(x,t^m) + \Delta t \cdot v_N'(x,t^m) + \Delta t \cdot \tau(t^m) q_N(x). \tag{4.1.3}$$

Here, $q_N(x)$ is a π_N-polynomial which characterizes the specific (pseudo)spectral method we employ, v' denotes spatial differentiation, and $\tau = \tau(t^m)$ is a free scalar multiplier to be determined by the boundary constraint

$$v_N(x=1,t^m) = 0. \tag{4.1.4}$$

We shall study the so called spectral tau method associated with *general* Jacobi polynomials $P_N^{(\alpha,\beta)}(x)$, $\alpha,\beta\epsilon(-1,1)$,

$$v_N(x,t^m+\Delta t) = v_N(x,t^m) + \Delta t \cdot a v_N'(x,t^m) + \Delta t \cdot \tau(t^m) q_N(x), \quad q_N(x) = P_N^{(\alpha,\beta)}(x). \tag{4.1.5}$$

Remark. The generality of our spectral formulation includes as a special case, the ψdospectral Jacobi methods which are collocated at the interior extrema of $P_{N+1}^{(\alpha,\beta)}$, $\alpha,\beta\epsilon(-1,0)$, i.e.,

$$v_N(x,t^m+\Delta t) = v_N(x,t^m) + \Delta t \cdot a v_N'(x,t^m) + \Delta t \cdot \tau(t^m) q_N(x), \quad q_N(x) = P_{N+1}^{(\alpha,\beta)'}(x). \tag{4.1.6}$$

Indeed, the spectral and ψdospectral Jacobi methods are closely related since $P_{N+1}^{(\alpha,\beta)'}(x)$ is a scalar multiple of $P_N^{(\alpha+1,\beta+1)}(x)$. For example, $\alpha=\beta=\frac{1}{2}$ and $\alpha=\beta=-\frac{1}{2}$ correspond to Chebyshev spectral and *psi*dospectral methods, respectively.

Let $-1<x_1<x_2<\ldots<x_N<1$ be the N distinct zeros of the forcing polynomial $q_N(x)$. For Jacobi type methods, (4.1.5) and (4.1.6), the nodes $\{x_j\}_{j=1}^N$ are the zeros of Jacobi polynomials associated with the Gauss and Gauss-Lobatto quadrature rules, with minimal gridsize of order

$$\Delta x_{\min} = \min(1+x_1, 1-x_N). \tag{4.1.7}$$

The spectral approximation (4.1.3) restricted to these points reads

$$v_N(x_j,t^{m+1}) = v_N(x_j,t^m) + \Delta t \cdot a v_N'(x_j,t^m), \quad 1\leq j\leq N, \tag{4.1.8}$$

and is augmented with the homogeneous boundary conditions

$$v_N(1,t^m)=0. \tag{4.1.9}$$

Equations (4.1.8), (4.1.9) furnish a complete equivalent formulation of the spectral approximation (4.1.3), (4.1.4). An essential ingredient in a stability theory of such approximations lies in the choice of appropriate L^2-*weighted* norms

$$\|f(x)\|_\omega^2 =< f(x),f(x) >, \quad < f(x),g(x) >= \sum_{j=1}^{N} \omega_j f(x_j)g(x_j). \tag{4.1.10}$$

We now make the definition of

Stability. We say the approximation (4.1.8), (4.1.9) is *stable* if there exist discrete weights, $\{\omega_j>0\}_{j=1}^N$, and a constant η_0 independent of N, such that

$$\|v_N(\cdot,t)\|_\omega \le \text{Const}\cdot e^{\eta_0 t}\|v_N(\cdot,0)\|_\omega, \tag{4.1.11}$$

and it is *strongly stable* if (4.1.11) holds with Const $=1$ and $\eta_0\le 0$,

$$\|v_N(\cdot,t)\|_\omega \le \|v_N(\cdot,0)\|_\omega. \tag{4.1.12}$$

With this in mind we turn to our main stability result stating

Theorem 4.1 (Stability of the spectral and ψdospectral Jacobi methods) *Consider the spectral approximations (4.1.8), (4.1.9), associated with the Jacobi tau method (4.1.5), or the ψdospectral Jacobi method (4.1.6). There exists a positive constant $\eta_0\equiv\eta_0(\alpha,\beta)>0$ independent of N such that if the following CFL condition holds:*

$$\Delta t\cdot a\left(\lambda_{N-1}+\frac{2}{\Delta x_{\min}}\right)\le\eta_0, \tag{4.1.13}$$

then the approximation (4.1.8), (4.1.9) is strongly stable, and the following estimate is fulfilled:

$$\|v_N(\cdot,t)\|_\omega \le e^{-\eta_0 a t}\|v_N(\cdot,0)\|_\omega. \tag{4.1.14}$$

Notes.

1. The choice of L^2-weighted norms. Theorem 4.1 deals with the stability of both the spectral tau methods associated with $P_N^{(\alpha,\beta)}(x)$, $\alpha,\beta\epsilon(-1,1)$, and the closely related ψdospectral methods associated with $P_{N+1}^{(\alpha,\beta)'}(x)$, $\alpha,\beta\epsilon(-1,0)$. In each case, there are (at least two) different weighted stability results, based on different choices of discrete L^2-weighted norms; these discrete weights $\{\omega_j\}_{j=1}^N$ are given by

$$\omega_j=\frac{1+x_j}{1-x_j}w_j^G, \quad \{w_j^G\}_{j=1}^N = \text{ Gauss}-\text{Jacobi weights in (2.5.5)}, \tag{4.1.15}$$

$$\omega_j=(1+x_j)w_j^L, \quad \{w_j^L\}_{j=1}^N = \text{ (interior) Gauss}-\text{Lobatto Jacobi weights in (2.5.14, 2.5.15)}. \tag{4.1.16}$$

2. The CFL condition. The CFL condition (4.1.13) places an $O(N^{-2})$ stability restriction on the time step Δt. Indeed, this stability restriction involves two factors : the eigenvalues associated with Jacobi equation (2.4.9),

$$\lambda_{N-1}\equiv\lambda_{N-1}(\alpha,\beta)<(N+1)^2, \quad \alpha,\beta\epsilon(-1,1), \tag{4.1.17}$$

and the collocated Gauss nodes, which accumulate within $O(N^{-2})$ neighborhoods near the boundaries,

$$\frac{1}{\Delta x_{\min}}\le\text{Const}\cdot N^2. \tag{4.1.18}$$

Thus, the CFL condition (4.1.13) boils down to

$$\Delta t \cdot aN^2 \leq \text{Const}_{\alpha,\beta}. \tag{4.1.19}$$

(For the practical range of parameters, $\alpha, \beta \epsilon [-\frac{1}{2}, \frac{1}{2}]$, we have $\text{Const}_{\alpha,\beta} \sim \frac{1}{5}\eta_0(\alpha,\beta)$).

3. The choice of a stability norm. The stability statement asserted in theorem (4.1) is formulated in terms of *discrete* seminorms, $\|\cdot\|_\omega$, which are ω-weighted by either (4.1.15) or (4.1.16). We note that $\|\cdot\|_\omega$ are in fact well-defined norms on the space of π_N-polynomials satisfying the vanishing boundary condition (4.1.9), i.e., corresponding to (4.1.15) or (4.1.16) we have[12]

$$\|v_N(\cdot,t)\|_\omega \geq \int_{-1}^{1} w(x)\frac{1+x}{1-x}v_N^2(x,t)dx, \quad v_N(1,t) = 0, \tag{4.1.20}$$

and in view of (2.5.16),

$$\|v_N(\cdot,t)\|_\omega = \int_{-1}^{1} w(x)(1+x)v_N^2(x,t)dx, \quad v_N(1,t) = 0. \tag{4.1.21}$$

Moreover, in view of (4.1.18), one may convert the stability statement (4.1.14) into the usual L_2-type stability estimate at the expense of possible algebraic growth which reads

$$\|v_N(\cdot,t)\|_{w(x)} \leq \text{Const} \cdot N^2 e^{-\eta_0 at}\|v_N(\cdot,0)\|_{w(x)}, \quad \|v_N(\cdot,t)\|^2_{w(x)} = \int_{-1}^{1} w(x)v_N^2(x,t)dx. \tag{4.1.22}$$

4. Exponential time decay. Let us integrate by parts the differential equation (4.1.1) against $(1+x)u$. Thanks to the homogeneous boundary condition (4.1.2) we find

$$\frac{d}{dt}\int_{-1}^{1}(1+x)u^2(x,t)dx \leq -\frac{a}{2}\int_{-1}^{1}(1+x)u^2(x,t)dx, \tag{4.1.23}$$

and therefore,

$$\|u(\cdot,t)\|_{1+x} \leq e^{-\frac{1}{4}at}\|u(\cdot,0)\|_{1+x}. \tag{4.1.24}$$

This estimate corresponds to the special case of the stability statement (4.1.14) for the spectral Legendre tau method ($\alpha = \beta = 0$) weighted by (4.1.16). The exponential time decay indicated in (4.1.24), and more generally in (4.1.14), is due to the special choice of ω-weighted stability norms. The weights $\{w_j\}_{j=1}^N$ in (4.1.15), (4.1.16) involve the essential factors $1+x_j$ or $\frac{1+x_j}{1-x_j}$ which amplify the inflow boundary values in comparison to the outflow ones. Since in the current homogeneous case, vanishing inflow data is propagating into the domain, this results in the exponential time decay indicated in (4.1.24) and likewise in the stability statement (4.1.14).

5. The inflow problem. A stability statement similar to theorem 4.1 is valid in the inflow case where $a < 0$. Assume that the CFL condition (4.1.13) holds with $\eta_0 = \eta_0(\beta,\alpha)$, then (4.1.14) follows with discrete weights $\omega_j = \frac{1-x_j}{1+x_j}w_j$ or $\omega_j = (1-x_j)w_j$.

As we noted before, there are several variants of theorem 4.1; we quote below two of these variants.

6. Stability of the spectral tau method. *The spectral Jacobi method* (4.1.5) *satisfies the stability estimate* (4.1.14) *with*

$$\omega_j = \frac{1+x_j}{1-x_j}w_j, \quad \{w_j = w_j^G(\alpha,\beta)\}_{j=1}^N = \textit{Gauss} - \textit{Jacobi weights}, \tag{4.1.25}$$

[12] Here we utilize the fact that the error term in Gauss quadrature (2.5.4) is proportional to an intermediate value of the $2N$-th derivative, $w^{(2N)}$ (— e.g. consult (2.5.8)) in the present context the inequality follows, $\frac{d^{(2N)}}{dx^{(2N)}}(\frac{1+x}{1-x}v_N^2(x,t)) > 0$.

$$\eta_0 \equiv \eta_0(\alpha,\beta) = \begin{cases} \frac{1}{2}(1+\beta), & \alpha+\beta \le 0, \\ \\ \frac{1}{2}(1-\alpha), & \alpha+\beta \ge 0. \end{cases} \qquad \alpha,\beta\epsilon(-1,1). \tag{4.1.26}$$

we proceed as follows. Squaring of (4.1.8) yields

$$\begin{aligned} \|v_N(\cdot,t^{m+1})\|_\omega^2 &= \|v_N(\cdot,t^m)\|_\omega^2 + \\ &+2\Delta t\cdot a < v_N(\cdot,t^m), v_N'(\cdot,t^m) > +(\Delta t\cdot a)^2\|v_N'(\cdot,t^m)\|_\omega^2 = \\ &= \|v_N(\cdot,t^m)\|_\omega^2 + 2\Delta t\cdot a I + (\Delta t\cdot a)^2 II, \end{aligned} \tag{4.1.27}$$

and we turn to estimate the two expressions, I and II, on the right of (4.1.27).

First let us note that since the π_N-polynomial $v_N(x,t^m)$ vanishes at the inflow boundary, (4.1.4), we have

$$v_N(x,t^m) = (1-x)p(x) \quad \text{for some } p(x) \equiv p_{N-1}(x)\epsilon\pi_{N-1}. \tag{4.1.28}$$

Also, a straightforward computation shows that

$$\left(w(x)\frac{1+x}{1-x}\right)'(1-x)^2 = [(\beta-\alpha+2)-(\beta+\alpha)x]w(x) \ge 4\eta_0 w(x), \quad |x|\le 1, \tag{4.1.29}$$

where $\eta_0 = \eta_0(\alpha,\beta)$ is given in (4.1.26).

Now, since $\frac{1+x}{1-x}v_N(x,t^m)v_N'(x,t^m)\epsilon\pi_{2N-1}$, the Gauss quadrature rule (2.5.4) implies

$$I \equiv \sum_{j=1}^N w_j \frac{1+x_j}{1-x_j} v_N(x_j,t^m)v_N'(x_j,t^m) = \int_{-1}^1 w(x)\frac{1+x}{1-x}v_N(x,t^m)v_N'(x,t^m)dx.$$

We integrate by parts the right-hand side of I, substitute $v_N(x,t^m) = (1-x)p(x)$ from (4.1.28), and in view of (4.1.29) we obtain

$$I = -\frac{1}{2}\int_{-1}^1 \left(w(x)\frac{1+x}{1-x}\right)'(1-x)^2p^2(x)dx \le -2\eta_0\|p\|_{w(x)}^2. \tag{4.1.30}$$

Next, let us consider the second expression, II, on the right of (4.1.27). As before, we substitute $v_N(x,t^n) = (1-x)p(x)$ from (4.1.28) and obtain

$$II \equiv \|v_N'(\cdot,t^m)\|_\omega^2 = \sum_{j=1}^N w_j\frac{1+x_j}{1-x_j}[(1-x_j)p'(x_j)-p(x_j)]^2 \le$$

$$leq2\sum_{j=1}^N w_j(1-x_j^2)(p'(x_j))^2 + 2\sum_{j=1}^N w_j\frac{1+x_j}{1-x_j}p^2(x_j) = II_1 + II_2.$$

To proceed we invoke the following

- Inverse inequality. *For all $p\epsilon\pi_N$ we have*

$$\|p'\|_{(1-x^2)w(x)} \le \sqrt{\lambda_N}\|p\|_{w(x)}, \quad p\epsilon\pi_N. \tag{4.1.31}$$

Here, $w(x)$ is any Jacobi weight, and λ_N is the corresponding Nth eigenvalue.

To verify (4.1.31): one expands $p(x) = \sum_{k=0}^N a_k P_k^{(\alpha,\beta)}(x)$ and $p'(x) = \sum_{k=0}^N a_k P_k^{(\alpha,\beta)'}(x)$; starting with the left-hand side of (4.1.31) and using the orthogonality of $P_k^{(\alpha,\beta)}$ w.r.t. $(1-x^2)w(x)$ we conclude

$$(LHS)^2 = \sum_{k=0}^N a_k^2\|P_k^{(\alpha,\beta)'}\|_{(1-x^2)w(x)}^2 = \sum_{k=0}^N \lambda_k a_k^2\|P_k^{(\alpha,\beta)}\|_{w(x)}^2 \le \lambda_N(RHS)^2,$$

and the assertion (4.1.31) follows.

The inverse inequality (4.1.31) preceded by Gauss rule (2.5.4), imply

$$II_1 \equiv 2\sum_{j=1}^{N} w_j(1-x_j^2)(p'(x_j))^2 = 2\|p'\|^2_{(1-x^2)w(x)} \le 2\lambda_{N-1}\|p\|^2_{w(x)}, \quad p\epsilon\pi_{N-1},$$

and this together with the obvious upper bound

$$II_2 \equiv 2\sum_{j=1}^{N} w_j \frac{1+x_j}{1-x_j}p^2(x_j) \le \frac{4}{\Delta x_{\min}}\|p\|^2_{w(x)},$$

give us

$$II \le \left(2\lambda_{N-1} + \frac{4}{\Delta x_{\min}}\right)\|p\|^2_{w(x)}. \tag{4.1.32}$$

Equipped with (4.1.30) and (4.1.32), we return to (4.1.27) to find

$$\|v_N(\cdot,t^{m+1})\|^2_\omega \le \|v_N(\cdot,t^m)\|^2_\omega - 2\Delta t\cdot a\left[2\eta_0 - \Delta t\cdot a\left(\lambda_{N-1}+\frac{2}{\Delta x_{\min}}\right)\right]\|p\|^2_{w(x)}. \tag{4.1.33}$$

The CFL condition (4.1.26) implies that the expression in square brackets on the right is nonnegative,

$$\left[2\eta_0 - \Delta t\cdot a\left(\lambda_{N-1}+\frac{2}{\Delta x_{\min}}\right)\right] \ge \eta_0 > 0, \tag{4.1.34}$$

and hence strong stability holds.

In fact, one more application of Gauss quadrature yields

$$\begin{aligned} \|p\|^2_{w(x)} = \textstyle\sum_{j=1}^N w_j p^2(x_j) &= \textstyle\sum_{j=1}^N w_j \frac{v_N^2(x_j,t^m)}{(1-x_j)^2} \ge \\ &\ge \textstyle\sum_{j=1}^N w_j \frac{1+x_j}{1-x_j} v_N^2(x_j,t^m) = \|v_N(\cdot,t^m)\|^2_\omega . \end{aligned} \tag{4.1.35}$$

The inequalities (4.1.35), (4.1.34) together with (4.1.33) imply

$$\|v_N(\cdot,t^{m+1})\|^2_\omega \le (1-2\eta_0\Delta t\cdot a)\|v_N(\cdot,t^m)\|^2_\omega, \tag{4.1.36}$$

and the result (4.1.14) follows. ■

Since $P_{N+1}^{(\alpha,\beta)'}$ is proportional to $P_N^{(\alpha+1,\beta+1)}$, we conclude the stability of the ψdospectral method (4.1.6), with $\omega_j = \frac{1+x_j}{1-x_j}w_j^G(\alpha+1,\beta+1)$ and $\eta_0 \equiv \eta_0(\alpha,\beta) = -\frac{\alpha}{2} > 0$.

As mentioned before, alternative variants of theorem 4.1 are possible. For example, one may employ a stable norm weighted by $\omega_j = (1+x_j)w_j$ (instead of the $\omega_j = \frac{1+x_j}{1-x_j}w_j$ weights used before. This yields the

Stability of the spectral-tau method revisited – *The spectral Jacobi tau method* (4.1.5). *satisfies the stability estimate* (4.1.14) *with* $\omega_j = (1+x_j)w_j^G$ and

$$\eta_0 = \eta_0(\alpha,\beta) = \begin{cases} -\frac{\alpha}{2}, & \alpha+\beta+1 \ge 0, \\ \frac{1}{2}(1-\beta), & \alpha+\beta+1 \le 0, \end{cases} \quad \alpha,\beta\epsilon(-1,0). \tag{4.1.37}$$

we omit the detailed derivation (— which as before, hinges on the exactness of Gauss quadrature rule for $2N$-polynomials), consult (2.5.4). If we replace the Gauss quadrature rule by the Gauss-Lobatto one, we are led to stability of the ψdospectral method (4.1.6) with $\omega_j = (1+x_j)w_j^L(\alpha,\beta)\}$ and with the same η_0 given in (4.1.37).

4.1.1 Problems with inhomogeneous initial-boundary conditions

We consider the inhomogeneous scalar hyperbolic equation

$$u_t = au_x + F(x,t), \quad (x,t)\epsilon[-1,1]\times[0,\infty), \quad a>0, \tag{4.1.38}$$

which is augmented with inhomogeneous data prescribed at the inflow boundary

$$u(1,t) = g(t), \quad t>0. \tag{4.1.39}$$

Using forward Euler time-differencing, the spectral approximation of (4.1.38) reads, at the N zeros of $q_N(x)$,

$$v_N(x_j,t^{m+1}) = v_N(x_j,t^m) + \Delta t\cdot a v_N'(x_j,t^m) + \Delta t F(x_j,t^m), \quad q_N(x_j)=0, \tag{4.1.40}$$

and is augmented with the boundary condition

$$v_N(1,t^m) = g(t^m). \tag{4.1.41}$$

In this section, we study the stability of (4.1.40), (4.1.41) in the two cases of

$$\text{Spectral Jacobi tau method}: \quad q_N(x) = P_N^{(\alpha,\beta)}(x), \quad \alpha,\beta\epsilon(-1,1), \tag{4.1.42}$$

and the closely related

$$\psi\text{dospectral Jacobi method}: \quad q_N(x) = P_{N+1}^{(\alpha,\beta)'}(x), \quad \alpha,\beta\epsilon(-1,0). \tag{4.1.43}$$

To deal with the inhomogeneity of the boundary condition (4.1.41), we consider the π_N-polynomial

$$V_N(x,t) = v_N(x,t) - \frac{q_N(x)}{q_N(1)}g(t). \tag{4.1.44}$$

If we set

$$\tilde{F}(x,t) = F(x,t) + a\frac{q_N'(x)}{q_N(1)}g(t), \tag{4.1.45}$$

then $V_N(x,t)$ satisfies the inhomogeneous equation

$$V_N(x_j,t^{m+1}) = V_N(x_j,t^m) + \Delta t\cdot a V_N'(x_j,t^m) + \Delta t\tilde{F}(x_j,t^m), \tag{4.1.46}$$

which is now augmented by the homogeneous boundary condition

$$V_N(1,t^m) = 0. \tag{4.1.47}$$

theorem 4.1 together with Duhammel's principle provide us with an a priori estimate of $\|V_N(\cdot,t)\|_\omega$ in terms of the initial and the inhomogeneous data, $\|V_N(\cdot,0)\|_\omega$ and $\|\tilde{F}(\cdot,t)\|_\omega$. Namely, if the CFL condition (4.1.13) holds, then we have

$$\|V_N(\cdot,t)\|_\omega \le e^{-\eta_0 a t}\|V_N(\cdot,0)\|_\omega + \sum_{0<t^m\le t}\Delta t\cdot e^{-\eta_0 a(t-t^m)}\|\tilde{F}(\cdot,t^m)\|_\omega. \tag{4.1.48}$$

Since the discrete norm $\|\cdot\|_\omega$ is supported at the zeros of $q_N(x)$, where $V_N(x_j,t) = v_N(x_j,t)$, we conclude

Theorem 4.2 (Stability with inhomogeneous terms) *Consider the spectral approximation (4.1.40), (4.1.41) associated with the Jacobi tau method (4.1.42) or the ψdospectral Jacobi method (4.1.43). There exists a positive constant $\eta_0 = \eta_0(\alpha,\beta) > 0$ independent of N, such that if the following CFL condition holds (consult (4.1.13)):*

$$\Delta t\cdot a\left(\lambda_{N-1} + \frac{2}{\Delta x_{min}}\right) \le \eta_0, \tag{4.1.49}$$

then the approximation (4.1.40), (4.1.41) satisfies the stability estimate

$$\|v_N(\cdot,t)\|_\omega \le e^{-\eta_0 a t}\|v_N(\cdot,0)\|_\omega + \sum_{0<t^m\le t}\Delta t\cdot e^{-\eta_0 a(t-t^m)}\left[\|F(\cdot,t^m)\|_\omega + a\frac{\|q_N'(\cdot)\|_\omega}{|q_N(1)|}|g(t^m)|\right]. \tag{4.1.50}$$

The last theorem provides us with an a priori stability estimate in terms of the initial data, $v_N(\cdot,0)$, the inhomogeneous data, $F(\cdot,t)$, and the boundary data $g(t)$. The dependence on the boundary data involves the factor of $\frac{\|q_N'(\cdot)\|_\omega}{|q_N(1)|}$, which grows linearly with N, so that we end up with the stability estimate

$$\|v_N(\cdot,t)\|_\omega \le e^{-\eta_0 at}\|v_N(\cdot,0)\|_\omega + \sum_{0<t^m\le t} \Delta t\cdot e^{-\eta_0 a(t-t^m)}\left[\|F(\cdot,t^m)\|_\omega + \text{Const}\cdot N|g(t^m)|\right]. \quad (4.1.51)$$

An inequality similar to (4.1.51) is encountered in the stability study of finite difference approximations to mixed initial-boundary hyperbolic systems. We note in passing that the stability estimate (4.1.51) together with the usual consistency requirement guarantee the spectrally accurate convergence of the spectral approximation.

4.2 Multi-level and Runge-Kutta Time Differencing

We extend our forward Euler stability result for certain second- and third-order accurate multi-level and Runge-Kutta time-differencing.

To this end, we view our π_N-approximate solution at time level t, $v(\cdot,t)$, as an $(N+1)$-dimensional column vector which is uniquely realized at the Gauss collocation nodes $(v(x_1,t),\dots v(x_N,t),v(1,t))$.

The forward Euler time-differencing (4.1.8) with homogeneous boundary conditions (4.1.9), reads

$$v(t^m+\Delta t) = [I+\Delta t\cdot aL]v(t^m), \quad a>0, \quad (4.2.1)$$

where L is an $(N+1)\times(N+1)$ matrix which accounts for the spatial spectral differencing together with the homogeneous boundary conditions,

$$Lv(t^m) = (v'(x_1,t^m),\dots,v'(x_N,t^m),0). \quad (4.2.2)$$

Theorem 4.1 tells us that if the CFL condition (4.1.13) holds, i.e., if

$$\Delta t\cdot a\left(\lambda_{N-1}+\frac{2}{\Delta x_{\min}}\right)\le\eta_0, \quad (4.2.3)$$

then $I+\Delta t\cdot aL$ is bounded in the ω-weighted induced operator norm,

$$\|I+\Delta t\cdot aL\|_\omega \le e^{-\eta_0 a\cdot\Delta t}. \quad (4.2.4)$$

Let us consider an $(s+2)$-level time differencing method of the form

$$v(t^m+\Delta t) = \sum_{k=0}^{s}\theta_k[I+c_k\Delta t\cdot aL]v(t^{m-k}), \quad c_k\ge 0, \quad \theta_k\ge 0, \quad \sum_{k=0}^{s}\theta_k = 1. \quad (4.2.5)$$

In this case, $v(t^m+\Delta t)$ is given by a *convex* combination of stable forward Euler differencing, and we conclude

Multi-level time differencing. *Assume that the following CFL condition holds,*

$$\Delta t\cdot a\left(\lambda_{N-1}+\frac{2}{\Delta x_{\min}}\right)\le\frac{\eta_0(\alpha,\beta)}{c_k}, \quad c_k\ge 0, \quad k=0,1,\dots,s. \quad (4.2.6)$$

Then the spectral approximation (4.2.5) *is strongly stable, and the following estimate holds*

$$\|v_N(\cdot,t)\|_\omega \le e^{-\eta_* at}\|v_N(\cdot,0)\|_\omega, \quad \eta_* = \min_k\frac{\eta_0}{c_k}>0. \quad (4.2.7)$$

Second and third-order accurate multi-level time differencing methods of the positive type (4.2.5) take the particularly simple form

$$v(t^m+\Delta t) = \theta[I+c_0\Delta t\cdot aL]v(t^m)+(1-\theta)[I+c_s\Delta t\cdot aL]v(t^{m-s}), \quad (4.2.8)$$

Second-order time differencing	θ	c_0	c_s
4-level method (s = 2)	$\frac{3}{4}$	2	0
5-level method (s = 3)	$\frac{8}{9}$	$\frac{3}{2}$	0
Third-order time differencing			
5-level method (s = 3)	$\frac{16}{27}$	3	$\frac{12}{11}$
6-level method (s = 4)	$\frac{25}{32}$	2	$\frac{10}{7}$
7-level method (s = 5)	$\frac{108}{125}$	$\frac{5}{3}$	$\frac{30}{17}$

Table 4.1: Multi-level methods

Second order time differencing	θ_2	θ_3
Two-step modified Euler ($s = 2$)	$\frac{1}{2}$	–
Third order time differencing		
Three-step method ($s = 3$)	$\frac{3}{4}$	$\frac{1}{3}$

Table 4.2: Runge-Kutta methods

with positive coefficients, θ, c_0, c_s, given in Table 4.1

Similar arguments apply for Runge-Kutta time-differencing methods. In this case the resulting positive type Runge-Kutta methods take the form

$$v^{(1)}(t^{m+1}) = [I + \Delta t \cdot aL]v(t^m), \tag{4.2.9}$$

$$v^{(k)}(t^{m+1}) = \theta_k v(t^m) + (1-\theta_k)[I + \Delta t a L]v^{(k-1)}(t^{m+1}), \quad k = 2, \cdots, s, \tag{4.2.10}$$

$$v(t^{m+1}) = v^{(s)}(t^{m+1}). \tag{4.2.11}$$

We arrive at

Runge-Kutta time-differencing. *Assume that the CFL condition* (4.1.13) *holds. Then the spectral approximation* (4.2.9)–(4.2.11) *with* $0 \le \theta_k < 1$ *is strongly stable and the stability estimate* (4.1.14) *holds.*

Table 4.2 quotes second and third-order choices of positive-type Runge-Kutta method.

4.3 Scalar Equations with Variable Coefficients

When dealing with finite difference approximations which are *locally supported*, i.e., finite difference schemes whose stencil occupy a *finite* number of neighboring grid cells each of which of size Δx, then one encounters the hyperbolic CFL stability restriction

$$\frac{\Delta t}{\Delta x}|a| \le \text{ Const.} \tag{4.3.1}$$

With this in mind, it is tempting to provide a heuristic justification for the stability of spectral methods, by arguing that a CFL stability restriction similar to (4.3.1) should hold. Namely, when Δx

is replaced by the minimal grid size, $\Delta x_{min} = \min_j |x_{j+1} - x_j| = O(N^{-2})$, then (4.3.1) leads to

$$\Delta t \cdot |a| N^2 \leq \text{ Const.} \tag{4.3.2}$$

Although the final conclusion is correct (consult (4.1.19)), it is important to realize that this "hand-waving" argument is not well-founded in the case of spectral methods. Indeed, since the spectral stencils occupy the whole interval (-1,1), spectral methods do not lend themselves to the stability analysis of locally supported finite difference approximations. Of course, by the same token, this explains the existence of *unconditionally* stable fully implicit (and hence globally supported) finite difference approximations.

As noted earlier, our stability proof (in Theorem (4.1)) shows that the CFL condition (4.3.2) is related to the following two points:

#1. The size of the corresponding Sturm-Liouville eigenvalues, $\lambda_{N-1} = O(N^2)$.

#2. The minimal gridsize, $\frac{1}{\Delta x_{\min}} = O(N^2)$.

The second point seems to support the fact that $\Delta x_{\min}$ plays an essential role in the CFL stability restriction for the global spectral methods, as predicted by the local heuristic argument outlined above. To clarify this issue we study in this section the stability of spectral approximations to scalar hyperbolic equations with variable coefficients. The principal raison d'$\hat{e}$tre, which motivates our present study, is to show that our stability analysis in the constant coefficients case is versatile enough to deal with certain variable-coefficient problems.

We now turn to discuss scalar hyperbolic equations with positive variable coefficients,

$$u_t = a(x) u_x, \quad 0 < a(x) < a_\infty, \quad (x,t) \epsilon [-1,1] \times [0,\infty), \tag{4.3.3}$$

which are augmented with homogeneous conditions at the inflow boundary

$$u(1,t) = 0. \tag{4.3.4}$$

We consider the ψdospectral Jacobi method collocated at the N zeros of $P_{N+1}^{(\alpha,\beta)'}(x)$. Using forward Euler time- differencing, the resulting approximation reads

$$v_N(x_j, t^{m+1}) = v_N(x_j, t^m) + \Delta t \cdot a(x_j) v_N'(x_j, t^m), \quad P_{N+1}^{(\alpha,\beta)'}(x_j) = 0, \tag{4.3.5}$$

together with the boundary condition

$$v_N(1, t^m) = 0. \tag{4.3.6}$$

Arguing along the lines of Theorem (4.1), we have

Theorem 4.3 (Stability of the ψdospectral Jacobi method with variable coefficients) *Consider the ψdospectral Jacobi approximation (4.3.5), (4.3.6). There exists a constant $\eta_0 \equiv \eta_0(\alpha,\beta)$,*

$$\eta_0 \equiv \eta_0(\alpha,\beta) = \begin{cases} -\frac{\alpha}{2}, & \alpha + \beta + 1 \geq 0, \\ \\ \frac{1}{2}(1-\beta), & \alpha + \beta + 1 \leq 0, \end{cases} \qquad \alpha, \beta \epsilon (-1, 0), \tag{4.3.7}$$

such that if the following CFL condition holds:

$$\Delta t \left(a_\infty \lambda_{N-1} + 2 \max_{1 \leq j \leq N} \frac{a(x_j)}{1 - x_j} \right) \leq \eta_0, \tag{4.3.8}$$

then the approximation (4.3.5), (4.3.6) is strongly stable, i.e., there exist discrete weights

$$\omega_j = (1 + x_j) \frac{w_j}{a(x_j)}, \quad \left\{ w_j = w_j^L(\alpha,\beta) \right\}_{j=1}^N = \text{ Gauss} - \text{Lobatto weights}, \tag{4.3.9}$$

such that

$$\|v_N(\cdot,t)\|_\omega \leq \|v_N(\cdot,0)\|_\omega. \tag{4.3.10}$$

PROOF. We divide (4.3.5) by $\sqrt{a(x_j)}$,

$$\frac{1}{\sqrt{a(x_j)}}v_N(x_j,t^{m+1}) = \frac{1}{\sqrt{a(x_j)}}v_N(x_j,t^m) + \Delta t\cdot\sqrt{a(x_j)}\cdot v_N'(x,t^m),$$

and, proceeding as before, we square both sides to obtain

$$\begin{aligned}\|v_N(\cdot,t^{m+1})\|_\omega^2 &= \|v_N(\cdot,t^m)\|_\omega^2 + \\ &+2\Delta t < v_N(\cdot,t^m), v_N'(\cdot,t^m) > +(\Delta t)^2\|a(\cdot)v_N'(\cdot,t^m)\|_\omega^2 \qquad (4.3.11)\\ &= \|v_N(\cdot,t^m)\|_\omega^2 + 2\Delta t\cdot I + (\Delta t)^2\cdot II.\end{aligned}$$

The first expression, I, involves discrete summation of the π_{2N}-polynomial $f(x) = (1+x)v_N(x,t^m)v_N'(x,t^m)$ and since $f(\pm 1) = 0$ (in view of (4.3.6)), the N-nodes Gauss-Lobatto quadrature rule yields

$$I \equiv \sum_{j=0}^{N+1} w_j^L(1+x_j)v_N(x_j,t^m)v_N'(x_j,t^m) = \int_{-1}^{1} w(x)(1+x)v_N(x,t^m)v_N'(x,t^m)dx.$$

We integrate by parts the right-hand side of I, substitute $v_N(x,t^m) = (1-x)p(x)$ with $p \equiv p_{N-1}\epsilon\pi_{N-1}$ and a straightforward integration by parts yields

$$I \le -2\eta_0\|p\|_{(1-x)w(x)}^2. \qquad (4.3.12)$$

The second expression, II, gives us

$$\begin{aligned}II &= \textstyle\sum_{j=1}^N w_j a(x_j)(1+x_j)[(1-x_j)p'(x_j) - p(x_j)]^2 \le \\ &\le 2a_\infty\textstyle\sum_{j=1}^N w_j(1-x_j^2)(1-x_j)(p'(x_j))^2 + 2\sum_{j=1}^N w_j a(x_j)(1+x_j)p^2(x_j) \qquad (4.3.13)\\ &= 2a_\infty II_1 + 2\cdot II_2.\end{aligned}$$

The inverse inequality (4.1.31) with weight $\omega(x) = (1-x)w(x)$ implies

$$II_1 = \|p'\|_{(1-x^2)(1-x)w(x)}^2 \le \lambda_{N-1}\|p\|_{(1-x)w(x)}^2, \quad \lambda_{N-1} = \lambda_{N-1}(\alpha+1,\beta)$$

and the expression II_2 does not exceed

$$II_2 \le \max_{1\le j\le N}[a(x_j)\frac{1+x_j}{1-x_j}]\cdot\sum_{j=0}^{N+1} w_j(1-x_j)p^2(x_j) \le 2\cdot\max_{1\le j\le N}\frac{a(x_j)}{1-x_j}\cdot\|p\|_{(1-x)w(x)}^2.$$

Consequently, we have

$$II \le 2\left(a_\infty\lambda_{N-1} + 2\cdot\max_{1\le j\le N}\frac{a(x_j)}{1-x_j}\right)\|p\|_{(1-x)w(x)}^2. \qquad (4.3.14)$$

Equipped with (4.3.12) and (6.19) we return to (6.16) to find

$$\|v_N(\cdot,t^{m+1})\|_\omega^2 \le \|v_N(\cdot,t^m)\|_\omega^2 - 2\Delta t\left[2\eta_0 - \Delta t\left(a_\infty\lambda_{N-1} + 2\max_{1\le j\le N}\frac{a(x_j)}{1-x_j}\right)\right]\|p\|_{(1-x)w(x)}^2, \quad (4.3.15)$$

and (4.3.10) follows in view of the CFL condition (6.14b). ■

Notes.

1. The case $a(x_j) \equiv a = \text{Const} > 0$ corresponds to one variant of the stability statement of theorem 4.1. Similar stability statements with the appropriate weights which correspond to various alternatives

of theorem 4.1, namely, with $\omega_j = \frac{1+x_j}{1-x_j}\frac{w_j^G}{a(x_j)}$, and $\omega_j = (1+x_j)\frac{w_j^G}{a(x_j)}$, hold. These statements cover the stability of the corresponding spectral and ψdospectral Jacobi approximations with variable coefficients.

2. We should highlight the fact that the stability assertion stated in theorem 4.3 depends solely on the uniform bound of $a(x_j)$ but otherwise is *independent* of the smoothness of $a(x)$.

3. The proof of theorem 4.3 applies mutatis mutandis to the case of variable coefficients with $a = a(x,t)$. If $a(x_j,t)$ are C^1-functions in the time variable, then (4.3.15) is replaced by

$$\|_N(\cdot,t^{m+1})\|_{\omega^{m+1}} \le (1 + \text{ Const.}\Delta t)\|v(\cdot,t^m)\|_{\omega^m}, \quad \omega_j^m = (1+x_j)\frac{w_j^L}{a(x_j,t^m)},$$

and stability follows.

4. We conclude by noting that the CFL condition (4.3.8) depends on the quantity $\max_{1\le j\le N}\frac{a(x_j)}{1-x_j}$, rather than the minimal grid size, $\frac{1}{\Delta x_{\min}}$, as in the constant-coefficient case (compare (4.1.13)). This amplifies our introductory remarks at the beginning of this section, which claim that the $O(N^{-2})$ stability restriction is essentially due to the size of the Sturm-Liouville eigenvalues, $\lambda_{N-1} = O(N^2)$. Indeed, the other portion of the CFL condition, requiring

$$\Delta t \cdot 2 \max_{1\le j\le N}\frac{a(x_j)}{1-x_j} \le \eta_0, \tag{4.3.16}$$

guarantees the *resolution* of waves entering through the inflow boundary $x = 1$. In the constant-coefficient case this resolution requires time steps Δt of size $\frac{1}{\Delta x_{\min}}$. However, when the inflow boundary is almost characteristic, i.e., when $a(1) \sim 0$, then the CFL condition is essentially independent of $\Delta x_{\min}$, for (4.3.16) boils down to $\Delta t \cdot 2a'(1) \le \eta_0$. In purely outflow cases the time step is independent of any resolution requirement at the boundaries, and we are left with the CFL condition restricted *solely* by the size of the corresponding SL eigenvalues.

We close this section with the particular example

$$u_t = -x u_x, \quad (x,t)\epsilon[-1,1]\times[0,\infty).$$

Observe that no augmenting boundary conditions are required, since both boundaries, $x = \pm 1$, are outflow ones. Consequently, the various forward Euler π_N-spectral approximations in this case amount to

$$v_N(x,t^m+\Delta t) = v_N(x,t^m) - \Delta t \cdot x v_N'(x,t^m). \tag{4.3.17}$$

The CFL stability restriction in this case is related to the $O(N^2)$-size of the Sturm-Liouville eigenvalues (point #1 above), but otherwise it is independent of the minimal grid size mentioned in point #2 above. We have

Outflow stability. *Assume that the following CFL condition holds*:

$$\Delta t \cdot \lambda_N \le 1, \quad \lambda_N = N(N+1).$$

Then the spectral approximation (4.3.17) *is stable, and the following estimate is fulfilled*:

$$\|v_N(\cdot,t)\|_{1-x^2} \le e^t \|v_N(\cdot,0)\|_{1-x^2}.$$

