The Computation of Equilibrium Solutions of Forced Hyperbolic Partial Differential Equations

Simon Wardrop
Magdalen College



Thesis submitted for the degree of
Doctor of Philosophy at the University of Oxford
Michaelmas Term 1990

This work was funded by a Scholarship from the
Royal Commission for the Exhibition of 1851

To my parents and Sarah

By reason of these things, then, the whaling voyage was welcome;
... and in the wild conceits that swayed me to my purpose, two and two there
floated into my inmost soul, endless processions of the whale, and, mid most
of them all, one grand hooded phantom, like a snow hill in the air.

Moby Dick

Acknowledgements

I would like to express my gratitude to my supervisor, Professor K.W. Morton of the Oxford University Computing Laboratory, for his great patience with me, and for his unfailing support.

I would also like to thank the Royal Commission for the Exhibition of 1851, for funding this thesis, and extending my Scholarship when illness delayed its submission.

Finally, I would like to thank Sarah Cannon, Richard Major and my parents for being such good friends. I would like to thank Sarah Cannon again for her careful proof-reading.

A thesis submitted for the degree
of Doctor of Philosophy

Simon Wardrop
Magdalen College
Michaelmas 1990

The Computation of Equilibrium Solutions of Forced Hyperbolic Partial Differential Equations

Abstract

This thesis investigates the convergence of numerical schemes for the computation of *equilibrium solutions*. These are solutions of evolutionary PDEs that arise from (bounded, non-decaying) boundary forcing after the dissipation of any (initial data dependent) *transients*. A rigorous definition of the term 'equilibrium solution' is given. Classes of evolutionary PDEs for which equilibrium solutions exist uniquely are identified. The *uniform well-posedness* of equilibrium problems is also investigated.

Equilibrium solutions may be approximated by *evolutionary initialization*: that is, by finding the solution of an initial boundary value problem, with arbitrary initial data, over a period of time $t \in [0,T]$. If T is chosen large enough, the analytic transient will be small, and the analytic solution over $t \in [T, T+T_0]$ will be a good approximation to the analytic equilibrium solution.

However, in numerical computations, T must be chosen so that the analytical transient is small in comparison with the numerical error E_h, which depends on the fineness of the grid h. Thus $T = T_h$, and, in general, $T_h \to \infty$ as $h \to 0$. Convergence is required over $t \in [T_h, T_h + T_0]$.

The existing Lax–Richtmyer and GKS convergence theories cannot ensure convergence over such increasing periods of time. Furthermore, neither of these theories apply when the forcing does not decay. Consequently, these theories are of little help in predicting the convergence of finite difference methods for the computation of equilibrium solutions.

For these reasons, a new definition of stability — *uniform stability* — is proposed. Uniformly stable, consistent, finite difference schemes, for uniformly well posed problems, converge uniformly over $t \geq 0$. Uniformly convergent schemes converge for bounded and non-decaying forcing.

Finite difference schemes for hyperbolic PDEs may admit waves of zero group velocity, even when the underlying analytic problem does not. Such schemes may be GKS convergent,

provided that the boundary conditions exclude these waves. The deficiency of the GKS theory for equilibrium computations is traced to this fact. However, uniform stability finds schemes that admit waves of zero group velocity to be (weakly) unstable, regardless of the boundary conditions.

It is also shown that weak uniform instabilities are the result of time-dependent analogues of the 'spurious modes' that occur in steady-state calculations. In addition, uniform stability theory sheds new light on the phenomenon of spurious modes.

Contents

1 Introduction **6**

1.1 Introduction to the Problem 6
1.2 Equilibrium Solutions 7
1.3 Examples of Equilibrium Solutions 9
1.3.1 Forced ODEs 9
1.3.2 One-Dimensional Linear Advection Equation 9
1.3.3 St Venant Equations 10
1.3.4 Telegrapher's Equations 11
1.3.5 Two-Dimensional Shallow Water Equations 12
1.3.6 One-dimensional Klein–Gordon Equations 16
1.3.7 Other Problems 16
1.4 Reaching Equilibrium 17
1.4.1 Direct Initialization Methods 17
1.4.2 Evolutionary Initialization Methods 19
1.5 Numerical Methods for Evolutionary Problems 19
1.5.1 Desirable Characteristics of Schemes for Equilibrium Problems 20
1.5.2 Time-Accurate Implicit Schemes for HPDEs 22
1.5.3 Other Methods 23
1.6 The Need for a New Convergence Theory 23
1.7 An Outline of this Thesis 24

2 Schemes and Notions of Convergence **26**

2.1 Notation for Finite Difference Methods 26
2.1.1 Two-Level Schemes 27
2.1.2 Two-Level Unitary Schemes 28
2.2 Outline of Convergence Theory 34
2.2.1 Lax–Richtmyer Stability for Two-Level Schemes 38
2.2.2 Extended Lax–Richtmyer Stability for Two-Level Schemes 39
2.2.3 Stability of Two-Level Implicit Methods 40
2.2.4 The GKS Theory 41
2.3 Uniform Invertibility 43
2.3.1 Uniform Invertibility for IVPs and PBCPs 43
2.3.2 Uniform Invertibility for Homogeneous IBVPs 44
2.3.3 Uniform Invertibility for General Homogeneous IBVPs 47
2.3.4 Uniform Invertibility in Two Dimensions 49
2.3.5 The Godunov–Ryabenkii Theory 50
2.4 Conclusions 51

3 Solutions of Forced Evolution Equations 52
3.1 Introduction . . . 52
3.2 Semigroups and Processes . . . 53
3.2.1 Properties of Linear C^0-Semigroups . . . 53
3.2.2 Autonomous Linear Evolutionary Problems . . . 54
3.2.3 Non-Autonomous Linear Problems . . . 58
3.2.4 More General Problems: Processes . . . 60
3.3 IVP Stability of Solutions . . . 61
3.3.1 Autonomous Linear Problems . . . 61
3.3.2 Non-Autonomous Linear Periodic Problems . . . 66
3.4 Generalized Solutions and Well-Posedness . . . 66
3.5 Equilibrium Solutions . . . 67
3.5.1 Equilibrium Solutions of Autonomous Linear Evolution Equations . . . 68
3.5.2 Steady and Periodic Equilibrium Solutions . . . 69
3.5.3 Equilibrium Solutions of Non-Autonomous Linear Equations . . . 70
3.5.4 Equilibrium Solutions of General FDEs: Attractors . . . 71
3.6 BVP Stability of Equilibrium Solutions . . . 72
3.6.1 BVP Stability of Periodic Equilibrium Solutions . . . 72
3.7 Conclusions . . . 73

4 Solutions of Forced Difference Equations 75
4.1 Introduction . . . 75
4.2 Discrete Semigroup Theory . . . 77
4.2.1 Discrete Spaces . . . 78
4.2.2 A Battery of Methods for the Linear Advection Equation . . . 78
4.3 IVP Stability of Solutions . . . 80
4.3.1 Numerical Stability Definitions for Autonomous Linear Problems . . . 80
4.3.2 Ensuring n.s. . . . 82
4.3.3 Ensuring n.a.s. . . . 82
4.3.4 Ensuring u.n.a.s. . . . 83
4.3.5 P-Stability . . . 89
4.4 Discrete Equilibrium Solutions . . . 91
4.5 BVP Stability of Equilibrium Solutions . . . 92
4.5.1 BVP Stability of Periodic and Steady Equilibrium Solutions . . . 92
4.5.2 BVP Stability . . . 95
4.6 Uniform Stability . . . 96
4.7 Checking the Near-Uniform Invertibility Condition . . . 100
4.7.1 Near-Uniform Invertibility of One-Dimensional Toeplitz Matrices . . . 100
4.7.2 The Scalar One-Dimensional Linear Advection Equation . . . 102
4.7.3 Group Velocity . . . 106
4.7.4 Near-Uniform Invertibility of Two-Dimensional Toeplitz Operators . . . 108
4.7.5 Steady Problems and Spurious Modes . . . 110
4.8 Multi-Step Schemes . . . 112
4.9 Conclusions . . . 113

5 Convergence **116**
5.1 Introduction 116
5.2 Convergence Orders 116
5.2.1 Equilibrium Order 116
5.2.2 Uniform Convergence 117
5.2.3 BVP Convergence and u.n.a.s. 117
5.2.4 Global Equilibrium Order 117
5.3 A Lax Equivalence Theorem for Uniform Stability 118
5.4 A Lax Equivalence Theorem for Periodic Problems 119
5.5 Analysis in Sobolev Spaces 120
5.6 Conclusions 120

6 Empirical Results **122**
6.1 Introduction 122
6.2 One-Dimensional Problems 123
6.2.1 Linear Advection Equation 123
6.2.2 St Venant Equations 127
6.3 Linear Advection Equation and Spurious Modes 129
6.3.1 ℓ_h^2-stability 129
6.3.2 Generalized Solutions 131
6.4 Conclusions 133

7 Conclusions **134**

List of Figures

1.1 A Cable 12
1.2 Dispersion relation for the Telegrapher's equations 13
1.3 Dispersion relation for 2-D shallow water equations 14
1.4 A stage of a turbomachine 16

2.1 Exact dispersion relation 31
2.2 Box scheme dispersion relation 31
2.3 Crank–Nicolson scheme dispersion relation 32
2.4 Angled-Derivative scheme dispersion relation 32
2.5 Sloan's scheme dispersion relation 33
2.6 Relationship between the discretization and truncation errors 36

3.1 Decay of the norm of a semigroup 56

4.1 Decay of solutions produced by n.a.s. methods 81
4.2 Blowup of solutions 83
4.3 Decay envelope of u.n.a.s. methods 84
4.4 $d_h(t)$ for the Box method for $\nu = 0.5$ 86
4.5 $d_h(t)$ for the Box method for $\nu = 1.5$ 87
4.6 $d_h(t)$ for the Box method for $\nu = 4.0$ 87
4.7 $d_h(t)$ for the CNBOX method for $\nu = 0.5$ 88
4.8 $d_h(t)$ for the CNBOX method for $\nu = 1.5$ 88
4.9 $d_h(t)$ for the CNBOX method for $\nu = 4.0$ 89
4.10 Roots of a polynomial assosciated with the Crank–Nicolson scheme 104

6.1 Error settling 126

List of Tables

2.1 Some unitary schemes 29
2.2 Dispersion relations 30

4.1 Numerical determination of the u.n.a.s. of various methods 90
4.2 Computed spectral radius exponents $\nu = 0.5$ 90
4.3 Computed spectral radius exponents $\nu = 1.5$ 90
4.4 Computed spectral radius exponents $\nu = 4.0$ 91
4.5 Summary of types of stability introduced. 114

6.1 L-Convergence for CNBOX and Box Methods 124
6.2 L-Convergence for CNBOX and Box Methods 125
6.3 L-Convergence for CNBOX and Box Methods 125
6.4 L-Convergence for CNBOX and Box Methods 125
6.5 Equilibrium orders for the St Venant equations 129
6.6 Stability of cell-centred and cell-vertex methods 130
6.7 Convergence orders for smooth forcing 132
6.8 Convergence orders with rough forcing 132

Chapter 1

Introduction

1.1 Introduction to the Problem

Several years ago, the Canadian Government proposed building a dam in the northern reaches of the Bay of Fundy, with the aim of exploiting the large tides in the bay for hydroelectric power generation. The tides in the Bay of Fundy are exceptionally large because the natural frequency of the bay differs from the forcing frequency offered by the tides of the Atlantic Ocean by only ten minutes or so. Unfortunately, constructing the dam would have increased the natural frequency of the bay, rendering it almost exactly coincident with the frequency of the Atlantic tides, and this would have resulted in still larger tides. Numerical predictions indicated that the result for Boston, towns on the coast of Maine, and the Nantucket Shoals, would have been catastrophic. The plan has now been scrapped.

At the heart of these predictions was a mathematical model [41]. This model covered only the bay area; the Atlantic tides were modelled by prescribing the height and flow of the ocean along an artificial boundary running from the southern tip of Nova Scotia, through the Gulf of Maine, towards Boston. Since the exact disposition of the bay's waters on any particular day were not known, they were assumed to be initially still. This model was, of course, mathematically intractable; no *analytic solution* could be found. But from studies of simpler models [11] we anticipate that the essentially arbitrary initial disposition of the bay (in this case, still) would clash with the tidal forcing, giving totally artificial results for some period of time. After these *transients* had dissipated to a negligible size, good results would be expected; the bay would settle down to a periodic *equilibrium solution* with (perhaps) the same period as the tidal forcing.

In flight, an aircraft wing can begin to flutter as a result of mechanical vibration or turbulent air flow over its surface. In the 1950s, linear perturbation theory [9,146,34] was applied to the problem of predicting the lift produced by fluttering thin aerofoils. The central conclusion of this work was that lift can be lost as the frequency of oscillation increases. Such losses of lift can be catastrophic during the take-off and landing of heavily laden aeroplanes, and their prediction is crucial to calculating the lift of helicopters in forward flight [131].

Since those first predictions, many researchers have used computer models to predict the lift of more realistic aerofoils, and the effect of flutter on this lift [5,111,138,131,125,137]. These predictions are obtained by modelling the air flow using the time-dependent Euler equations, and the flutter by varying the aerofoil's angle of attack in a periodic fashion. Since it is usually impossible to establish the actual flow, at a given instant, around a real fluttering aerofoil, the air is initially assumed to be streaming steadily.

It could be anticipated that the essentially arbitrary initial flow would clash in some way with the fluttering wing, resulting in totally artificial results for some period of time. Sidès [131, page 20] reports that

> In the vicinity of the airfoil, the periodic regime is established after three cycles of oscillatory motion.

Thus, after the transients have dissipated to a negligible size, the flow settles down to a periodic *equilibrium* solution, representing a dynamic equilibrium between the continual introduction of energy through forcing and the continual loss of energy through advection.

These examples are typical of the mathematical problems that motivate this thesis. The problems are essentially concerned with wave propagation, so the partial differential equations (PDEs) are hyperbolic. They are posed on finite domains, so, in addition to initial data, we must prescribe boundary data: that is, we are dealing with evolutionary initial boundary value problems (IBVPs). The examples are forced, and the forcing need not decay in time.

However, the aspect of these examples that distinguishes them from general hyperbolic IBVPs is that only the equilibrium solution is of interest; this equilibrium depends solely upon the boundary forcing. The transients that arise from the (arbitrary) initial data are not of interest.

Equilibrium solutions might be viewed simply as special hyperbolic IBVPs, if special 'equilibrium' initial data could be found. For example, in the case of the fluttering aerofoil problem, special equilibrium initial data could presumably be found by carefully measuring the flow at some initial time. Similarly, careful surveys of the Bay of Fundy at a particular instant might give equilibrium initial data. This initial data could be evolved using standard techniques to obtain the sought equilibrium solution. However, such measurements are rarely feasible, and are inevitably contaminated with error. For this reason, equilibrium solutions cannot be viewed simply as hyperbolic IBVPs.

Equilibrium initial data can also be prepared by integrating some arbitrary data through a large period of time. This is by far the most common procedure for finding steady (equilibrium) solutions of PDEs that are subject to steady forcing. However, equilibrium solutions need not be steady.

These considerations show that equlibrium solutions have features of both steady and wave-like solutions, but do not fit neatly into either category. To the author's knowledge, no clear notion of 'equilibrium solutions' exists in the literature. This thesis aims to give a rigorous definition of equilibrium solutions, and investigates numerical methods for their approximation.

1.2 Equilibrium Solutions

This section aims to clarify what we mean by 'equilibrium solutions' of forced evolution equations. This heuristic discussion will be refined into a rigorous definition in Chapter 3, but we must emphasize from the start that our definition is quite different from that of dynamical systems theory [50]. In Chapter 3, questions of existence, and properties of such solutions are also discussed. Furthermore, the definition is broadened to include general (possibly non-linear) functional differential equations.

The *equilibrium solution* of an evolutionary equation is intended to be the bounded solution of the evolution equation, defined for all time, that would result from an infinite

history of forcing, and which is uninterrupted by the imposition of (what we consider to be) arbitrary initialization. The equilibrium solution is a solution of the evolutionary equation alone, and not of a particular initial value problem (IVP). However, the solution of any particular IVP must also tend towards the equilibrium solution, at large time.

For example, the Bay of Fundy may be assumed to have been forced by the tides at all times in the past. The 'initial configuration' of the bay has been 'forgotten'. Furthermore, when the state of the bay is perturbed, say, by a storm, the bay eventually settles back to equilibrium.

Equilibrium Solutions of Autonomous Linear Problems

To be more precise, consider an autonomous linear evolutionary problem. An unforced linear evolutionary IVP may be written abstractly as $\dot{u} = Au$ with the data $u(0) = u^0$. Here A is some operator, which may possibly involve spatial derivatives. The (unforced, outflow) boundary conditions are enforced through the definition of the solution space. The solution of this problem is written $u(t) = G(t)u^0$, and it is assumed that $u(t) \to 0$ as $t \to \infty$ for any initial data. Thus, it is supposed that $G(t) \to 0$ as $t \to \infty$.

The solution of the forced IVP

$$\dot{u} = Au + f(t), \quad u(t^0) = u^0 \tag{1.1}$$

is given by the variation of parameters formula

$$u(t; t^0, u^0, f) = G(t - t^0)u^0 + \int_{t^0}^{t} G(t - s)f(s)\, ds. \tag{1.2}$$

If the initial data u^0 is constrained to lie in some (arbitrary) bounded set C, and the limit as $t^0 \to -\infty$, with $u^0 \in C$, is taken, then the equilibrium solution is

$$u_{eq}(t; f) \equiv \int_{-\infty}^{t} G(t - s)f(s)\, ds. \tag{1.3}$$

In fact, we shall take equation (1.3) as the ***definition*** of an equilibrium solution for forced, autonomous linear problems. This definition is a generalization of the expression given in reference [20, page 68] for periodic solutions of periodically forced ODEs.

Two important special cases of equilibrium solutions must be distinguished. First, when the forcing f is independent of time, u_{eq} is a (time-independent) ***steady state***. Secondly, when f is T-periodic, under suitable hypotheses u_{eq} will also be T-periodic. These equilibrium solutions will be called ***periodic states***. For general forcing, the corresponding equilibrium solution is also called an ***equilibrium state***. Equilibrium states generalize steady states to unsteadily forced problems.

Transients

In practice, $G(t)$ can rarely be computed explicitly. Only in rare instances can the equilibrium solution be found using equation (1.3). Instead, equation (1.1) must be solved with arbitrary initial data $u(t^0) = u^0$. Consequently, a ***transient*** is superimposed on the equilibrium solution,

$$u_{trans}(t; t^0, u^0, f) \equiv u(t; t^0, u^0, f) - u_{eq}(t; f). \tag{1.4}$$

In fact, it is easily seen that

$$u_{trans}(t;t^0,u^0,f) = G(t-t^0)\left[u^0 - u_{eq}(t^0;f)\right]. \tag{1.5}$$

The assumption that $G(t)\rightarrow 0$ at large time implies that the 'transient' truly decays with time. This equality also implies that there is unique initial data that eliminates the transient, namely $u^0 = u_{eq}(t^0)$.

1.3 Examples of Equilibrium Solutions

Singling out the equilibrium solution as being the appropriate description of some physical system is a modelling assumption. Ultimately, this assumption can only be evaluated by comparison with measurements. However, for many physical systems, the solution that results from the imposition of initial data seems quite unnatural compared with an equilibrium solution. In this section we discuss a number of such systems.

1.3.1 Forced ODEs

Consider the scalar ODE

$$\dot{u} = au + \dot{f}(t) - af(t), \quad u(0) = u^0,$$

where $a < 0$, and f is a forcing term. This is the model problem of Prothero and Robinson [27, page 196], which led to the concept of B-convergence of discretizations of ODEs. The solution is

$$u(t) = f(t) + e^{at}\left[u^0 - f(0)\right].$$

For $a < 0$, the equilibrium solution is given by $u_{eq}(t) = f(t)$, whilst the transient is $u_{trans}(t) = e^{at}\left[u^0 - f(0)\right]$.

1.3.2 One-Dimensional Linear Advection Equation

The one-dimensional scalar linear advection is of little practical interest, but does provide the simplest imaginable example of an IBVP for a hyperbolic PDE (HPDE) that has an equilibrium solution. Consider the scalar linear advection equation

$$u_t + au_x = 0, \quad a > 0,$$

posed on $(x,t) \in I \times \Re^+$, where $I \equiv [0,1]$ and $\Re^+ \equiv [0,\infty)$. The initial and boundary data are

$$u(x,0) = u^0(x), \quad u(0,t) = f(t).$$

The solution is given by

$$u(x,t) = \begin{cases} u^0(x-at) & \text{if } x-at \geq 0 \\ f(t-x/a) & \text{if } x-at < 0. \end{cases}$$

It is clear that for $t > T \equiv 1/a$, the solution does not depend on the initial data: the transient depending on u^0 has left the domain, and the equilibrium solution depends only on the boundary forcing.

1.3.3 St Venant Equations

Consider a one-dimensional channel (or estuary) open to the ocean. Tidal forcing at the end of the channel is modelled by prescribing sinusoidally varying flows at the outlet. The flow in the channel is given by the St Venant (or long wave) equations.

If x is the distance upstream, and t represents time, the St Venant equations are written in terms of

$h(x,t)$, the water depth from a fixed reference, and

$u(x,t)$, the water velocity.

The reference curve is supposed linear, with constant slope S_0. Writing g for the gravitational constant, the continuity equation is

$$h_t + uh_x + hu_x = 0. \tag{1.6}$$

Under the hypotheses of [84, page 292], the momentum equation is

$$u_t + uu_x = g\left(S_0 - h_x - \frac{u^2}{C^2 h}\right), \tag{1.7}$$

where C is the Chezy friction factor. Further details may be found in references [11,108,84, 154].

The boundary conditions are

$$h(0,t) = H_0 + \hat{a}\sin(\omega t), \quad u(L,t) = 0, \tag{1.8}$$

the first representing tidal forcing, and the second the closed end of the estuary at $x = L$.

For undisturbed flow, it is assumed that the height and velocity take the constant values h_0 and u_0. These are related by $u_0 = C\sqrt{h_0 S_0}$. Setting $h = h_0 + \hat{h}$ and $u = u_0 + \hat{u}$, and linearizing gives

$$\hat{h}_t + u_0\hat{h}_x + h_0\hat{u}_x = 0, \tag{1.9}$$

$$\hat{u}_t + u_0\hat{u}_x + g\hat{h}_x + gS_0\left(\frac{2\hat{u}}{u_0} - \frac{\hat{h}}{h_0}\right) = 0. \tag{1.10}$$

In fact, these are the Telegrapher's equations. If these equations are written as a linear system, in terms of $w = [\hat{h}, \hat{u}]^T$,

$$w_t = \tilde{A}w_x + \tilde{B}w, \tag{1.11}$$

the spectrum of A is real and given by $\sigma(\tilde{A}) = \{u_0 \pm \sqrt{gh_0}\}$. Thus, the system is hyperbolic, with characteristic speeds $u_0 \pm \sqrt{gh_0}$. When $|u_0| < \sqrt{gh_0}$, the flow is said to be *subcritical*, which is why one boundary condition is needed at each end of the channel.

Where u_0 and $S_0 = 0$ are small, but $r = 2gS_0/u_0$ is not, then (dropping the carets) [154, page 83] [11]

$$h_t + h_0 u_x = 0, \tag{1.12}$$

$$u_t + g h_x + r u = 0, \tag{1.13}$$

The choice of initial conditions presents a problem. Bode and Sobey [11] chose

$$h(x,0) = H_0, \quad u(x,0) = 0. \tag{1.14}$$

Using such 'impulsive' initial conditions is almost always a bad idea, and much of reference [11] is devoted to improving on (1.14). Discretizations of non-linear problems can become 'unstable' when used with wildly wrong initial conditions. This has been observed for exterior problems for the unsteady Euler equations [109, page 270]. We surmise that these instabilities may be due to the initial data lying outside the attractor of the difference equations, a situation that may occur for even the simplest non-linear iterations (see the discussion of Julia sets in reference [107]), and which may be allied to the idea of stability thresholds in non-linear stability definitions [89].

For discretizations of the linearized problem, such non-linear effects cannot occur. The stability of any orbit is the same as that of the zero solution, and the basin of attraction of zero is either empty, or it covers the whole space of initial data. Even so, impulsive initial data leads to undesirable transients. Bode and Sobey comment that

> Initial transients are encountered in all numerical solutions of the long wave equations; they are free mode responses and cannot be avoided. Their origin is physical, not numerical ... For harmonic forcing (astronomical tide) a 'settling down' time is always required to allow the initial transients to be dissipated by bottom friction ... The settling-down time is judged to be sufficient when a 'smooth' solution is achieved.

This seems to be a clear case of the (periodic) equilibrium state being the most physically appropriate solution.

1.3.4 Telegrapher's Equations

The Telegrapher's equations are of interest not only as linearizations of the St Venant equations, but also in their own right. They are known as the Telegrapher's equations because they model the transmission of signals in telegraphic cables. If we write $i(x,t)$ for the current and $v(x,t)$ for the voltage in the one-dimensional cable, and set

$L(x,t)$ to be the self-inductance,

$C(x,t)$ to be the capacitance,

$G(x,t)$ to be the leakage conductance, and

$R(x,t)$ to be the resistance,

all *per unit length*, then the Telegrapher's equations are

$$i_t + \frac{1}{L} v_x + \frac{R}{L} i = 0, \tag{1.15}$$

$$v_t + \frac{1}{C} i_x + \frac{G}{C} v = 0. \tag{1.16}$$

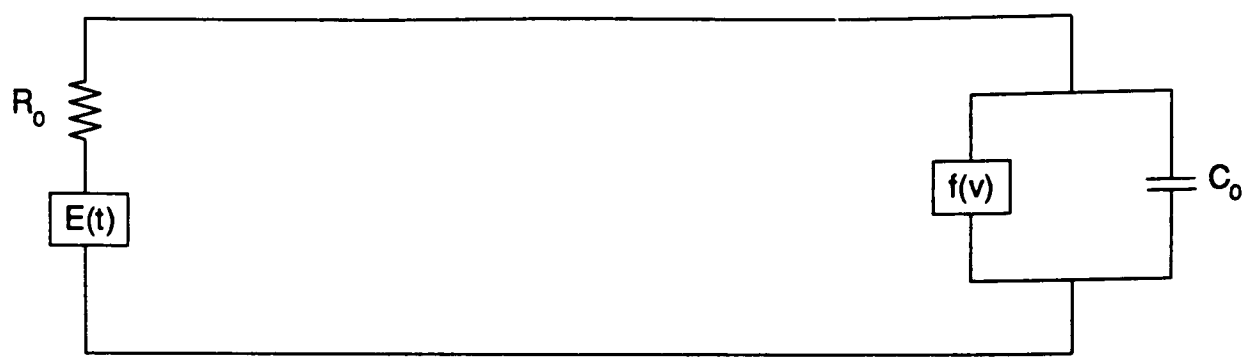


Figure 1.1: A cable.

For further physical details, see references [10,7].

If the linear system (1.15–1.16) is written in terms of $w = [i, v]^T$, then an equation of the form (1.11) is obtained. Here $\sigma(\tilde{A}) = \{\pm 1/\sqrt{LC}\}$. The system is hyperbolic, and one boundary condition is needed at each end of the cable.

Now consider the cable of Figure 1.1. The boundary condition at $x = 0$ is $E_0(t) = v(0,t) + i(0,t)R_0$. At $x = 1$, suppose that the non-linear element passes a current $i = f(v)$. The capacitor passes a current $i = C_1 \frac{dv}{dt}$. Thus, at $x = 1$, the boundary condition is $C_1 \frac{dv}{dt}(1,t) + f(v(1,t)) = i(1,t)$. In sum,

$$v(0,t) + R_0 i(0,t) = \quad E_0(t), \quad \text{at } x = 0, \tag{1.17}$$

$$C_1 \frac{d}{dt} v(1,t) + f(v(1,t)) = \quad i(1,t), \quad \text{at } x = 1. \tag{1.18}$$

In general, $f(v)$ represents a non-linear element. Brayton and Miranker [13] consider a tunnel diode. Schmitt [132] has used a similar model, to describe a flip-flop. When $f(v) = -R_1 v$, we recover the linear boundary condition investigated in [100, page 343]. These boundary conditions will be investigated in Chapter 3.

The dispersion relation for the Telegrapher's equations is given by

$$(\omega - ia)^2 = c^2 k^2 - b^2,$$

where $c \equiv 1/\sqrt{LC}$, $a \equiv (RC + LG)/2LC$ and $b \equiv (RC - LG)/2LC$.

The Telegrapher's equations also model a string under constant tension, which is repelled from the zero-displacement line by a force proportional to the transverse displacement [55, page 79]. In these circumstances, it is physically possible to have $a = 0$, but $b \neq 0$. The dispersion relation is then

$$\omega^2 = c^2 k^2 - b^2,$$

which is graphed in Figure 1.2.

1.3.5 Two-Dimensional Shallow Water Equations

Not all IBVPs for HPDEs will asymptotically approach a state that is independent of the initial data. One example is given by the two-dimensional linearized shallow water equations. These equations govern the flow in a shallow rotating sea of constant depth. We assume that the flow velocities in the x and y directions are $u(x,y,t)$ and $v(x,y,t)$, respectively. The elevation about some reference point is denoted $\eta(x,y,t)$. We write

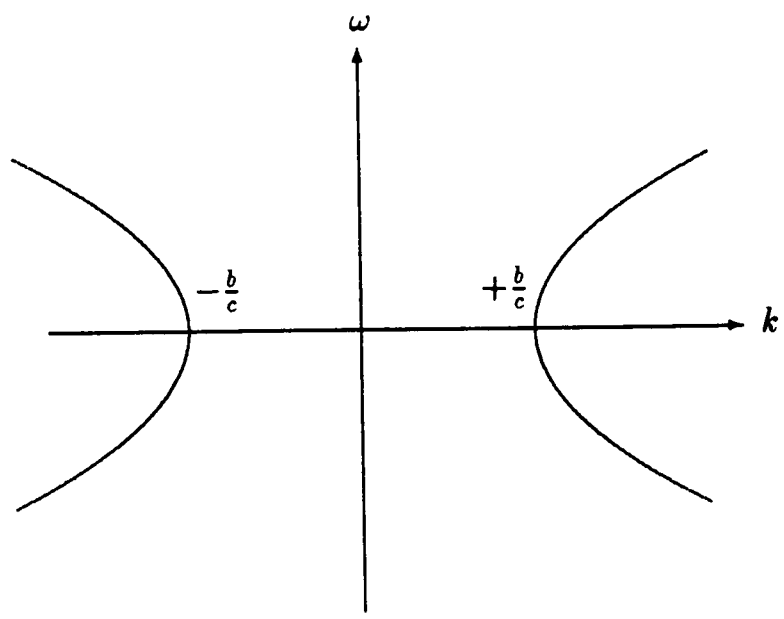


Figure 1.2: The dispersion relation for the Telegrapher's equations with $a = 0$.

H for the mean depth below the reference point,

U for the mean flow in the x direction,

V for the mean flow in the y direction,

f for the Coriolis parameter (half the rotation rate [37, page 191]), and

g for the gravitational acceleration.

Defining the (linearized) *substantial derivative* by

$$\frac{D}{Dt} \equiv \partial_t + U\partial_x + V\partial_y,$$

the linearized two-dimensional shallow water equations can be written [37,154]

$$\frac{Du}{Dt} - fv = -g\eta_x, \tag{1.19}$$

$$\frac{Dv}{Dt} + fu = -g\eta_y, \tag{1.20}$$

$$\frac{D\eta}{Dt} + H(u_x + v_y) = 0. \tag{1.21}$$

These are the two-dimensional Klein–Gordon equations. They form a hyperbolic system, with phase speeds $U + c$, $U - c$, U, where $c \equiv \sqrt{gH}$ [18, page 108].

Asymptotically, the solution of these equations is not one of rest, but rather one of *geostrophic balance* [37, page 196], with non-zero steady height and velocity distributions. This balance is *not* uniquely determined by the steady counterparts of equations (1.19)–(1.21): another piece of information is needed, usually the value of the potential vorticity at some instant. Conservation of potential vorticity Q,

$$\frac{DQ}{Dt} = 0, \quad Q \equiv \frac{\zeta}{f} - \frac{\eta}{H}, \quad \zeta \equiv v_x - u_y,$$

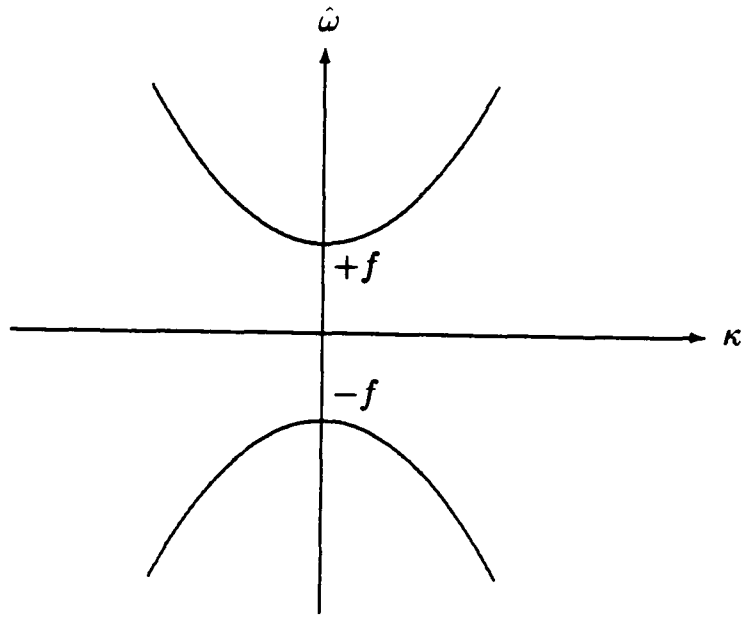


Figure 1.3: The dispersion relation of the two-dimensional shallow water equations. Note that $\kappa^2 = k^2 + l^2$.

together with the steady equations, then yields a unique solution — the geostrophic balance $u_{geo}(x, y; u^0)$ — at all subsequent times.

However, this implies that the 'memory' of the time-dependent system is infinite. The concept of an equilibrium state (as we have defined it) does not apply, since the equilibrium state cannot depend solely on the forcing. The geostrophic solution could be regarded as a kind of equilibrium state. In fact, over any domain Ω, it is easily shown [37, page 196] that

$$\|u(t) - u_{geo}(t)\|_\Omega \to 0 \quad \text{as} \quad t \to \infty.$$

However, we choose not to regard $u_{geo}(x, y; u^0)$ as an equilibrium state, since it depends on the initial data. Similar conclusions hold for the corresponding IBVPs, since an IVP with the solution measured over a finite domain is an IBVP with the most perfect radiative boundary conditions imaginable.

To explain this lack of decay, it is helpful to calculate the group velocity of waves admitted by the IVP for the two-dimensional shallow water equations. Writing $w = [u, v, \eta]^T$, and seeking solutions of the form $w = \bar{w} \exp i(kx + ly + \omega t)$, gives the dispersion relation ($c^2 = gH$)

$$\hat{\omega}\left[\hat{\omega}^2 - c^2(k^2 + l^2) - f^2\right] = 0, \quad \hat{\omega} \equiv \omega + Uk + Vl.$$

The solutions are $\hat{\omega} = 0$ (plane waves), and

$$\hat{\omega}^2 = c^2\kappa^2 + f^2, \quad \kappa^2 \equiv k^2 + l^2. \tag{1.22}$$

(see Figure 1.3). Clearly, the system is dispersive.

The group velocity is defined as

$$c_g(k, l) \equiv \left[\frac{\partial \omega}{\partial k}, \frac{\partial \omega}{\partial l}\right],$$

and this describes the (asymptotic) velocity of energy propagation of waves of wave number (k, l) [83,8], so long as that velocity is less than the maximum characteristic speed. (When

c_g exceeds the maximum characteristic speed, energy propagates at a different *signal speed* [85,136,55,14].) This remains true whenever k, l, ω and components of the group velocity are all real, even if neighbouring values $[k,l]$ result in complex ω [149]. For dispersive problems such as the present one, the group velocity differs from the *phase velocity*

$$c_p(k,l) \equiv \left[\frac{\omega}{k}, \frac{\omega}{l}\right].$$

The maximum characteristic speed gives an upper bound on the signal (and phase) speeds [136,85].

For the two-dimensional shallow water equations, the group velocity of the non-planar waves is given by

$$c_g(k,l) = \frac{c^2}{\hat{\omega}}[k,l] - [U,V],$$

so that waves of wave number $[k,l] = [U,V]\hat{\omega}/c^2$ have zero group velocity. A consequence of this observation is that the rate of decay of the solution, measured over some finite domain Ω, is slower than that of equations that do not admit zero group velocities [37, page 202], such as the two-dimensional shallow water equations in the absence of rotation, $f = 0$ (the wave equation). Rauch [116] has shown that solutions of general hyperbolic IVPs *decay with a rate* provided that the group velocity is bounded away from zero: that is, provided there exists $f(t)\rightarrow 0$ such that

$$\|u(t)\|_\Omega \leq f(t)\|u^0\|_\Omega.$$

It might be argued that these difficulties could be avoided simply by introducing bottom friction terms into the equations, at least where the water is shallow. However this is not a complete answer. In ocean models, such terms introduce a parameter which must be 'tuned' and which is usually chosen to be as small as possible. Also, when the two-dimensional shallow water equations are used to model atmospheric processes, the friction terms are negligible.

This example suggests that some solutions that might be regarded as being 'equilibrium' states will not fit our definition. We do not see this as a serious deficiency. There are clearly two kinds of 'equilibrium' state: those that are independent of the initial data, and those that are not. Our definition makes this distinction clear, and chooses to exclude the latter.

For IBVPs, the slowly decaying modes could possibly be excluded by boundary conditions. However, these modes are a real and important feature of the IVP's solution. Since the limited area model is introduced only to help with computer calculations, we should be wary of prescribing boundary conditions that change the physics so drastically. Interestingly, Majda and Osher [91] claim that the two-dimensional shallow water problem is representative of a class of problems — those with a uniformly characteristic boundary — for which Kreiss' well-posedness theory is inapplicable.[1] Until further analysis has been done on the decay of solutions for two-dimensional hyperbolic IBVPs, we cannot definitely say whether or not the shallow water equation IBVP has equilibrium solutions. We shall investigate a one-dimensional analogue of this problem — the one-dimensional Klein–Gordon equations — but the results do not necessarily carry over to higher dimensions.

[1] We must mention that other authors do not seem to agree: see the discussion in reference [18, page 111].

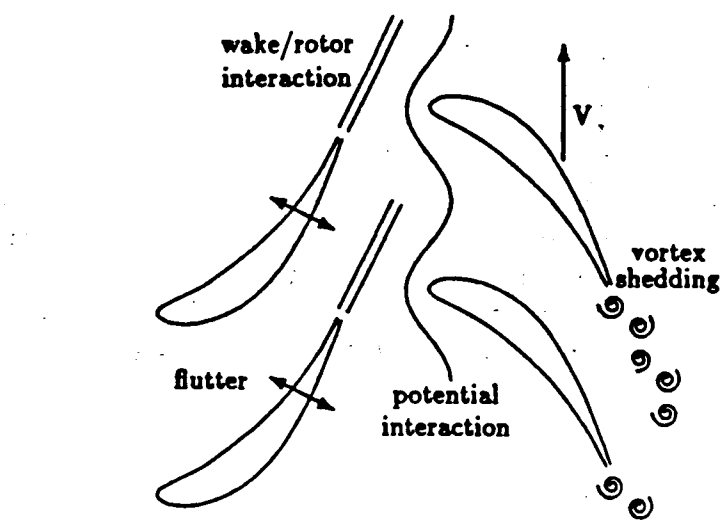


Figure 1.4: A stage of a turbomachine. From [141].

1.3.6 One-dimensional Klein–Gordon Equations

The one-dimensional Klein–Gordon equations model a variety of phenomena, including a stretched string in a rubberized medium [3] and the relativistic evolution of a free particle in quantum mechanics [26, page 191].

The dispersion relation is given by $\omega^2 = c^2k^2 + f^2$ (c.f. Figure 1.3), so we may expect the sort of decay problems that the two-dimensional shallow water equations exhibit.

1.3.7 Other Problems

There are a variety of physical systems whose behaviour is best described by the equilibrium state. The periodically forced Euler equations modelling fluttering aerofoils, introduced earlier, are one example. See references [5,111,138,131,125,137].

Gas turbines offer another example. A single stage of a turbomachine consists of two rows of blades, the first being the *stators* and the second the *rotors*: see Figure 1.4. Our concern is with the lift and moment of the rotors. The wakes from the stators may be considered to be steady in the stator's frame, and are generated by viscous effects [36]. Since the stators and the rotors are rotating relative to one another, the stator wakes are periodically cut by the rotors. The subsequent interaction between wake and rotors is largely inviscid.

Wake/rotor interaction is usually modelled in the rotor's frame. The inviscid flow through the rotor blades is modelled using the two-dimensional Euler equations between a pair of identical aerofoils: periodic boundary conditions are prescribed vertically, where the aerofoils do not provide a physical boundary. The stator wake is modelled as a time-periodically varying inflow to the gap between the aerofoils. See references [36,35,32,141,146] for further details.

It is clear that the periodic equilibrium solution is the one required. Typically, the model must be started impulsively, but the transient is of no interest. The calculation is simply continued until a time-periodic solution is reached [36, page 56]. When this happens, the lift and moment of one rotor blade (and hence of the whole rotor) may be calculated.

Finally, Sobey [133,140,134] has studied a blood oxygenator devised by Bellhouse. This device consists of a furrowed channel through which blood is passed back and forth. The walls of the furrowed channel are made of a gas-permeable membrane, which passes oxygen to the blood in the channel. The Bellhouse oxygenator is remarkably efficient, and Sobey sought to explain this efficiency by calculating the blood flow in the device.

This was done by modelling the oxygenator as a cavity, with periodic furrowed walls in the x direction. The blood flow was described by the Navier–Stokes equations, with the usual zero velocity boundary conditions at the walls. The flow was assumed to be periodic in x, with the same spatial period as the furrowing. Oscillatory flow was ensured by requiring that the stream function vary sinusoidally (in time) at the walls.

The calculation was begun with the fluid at rest, and forced with period $T = 1$. Sobey reports that

> The solutions were obtained on a PDP 11–34 computer and a typical run from $t = 0$ until $t = 1.5$ would take about 5–10 hours. In those cases where the flow was inertially dominated the solution for $t = 0.5$ until $t = 1.5$ would be an accurate sample of a general cycle.

In other words, the periodic state was contaminated by transients of substantial magnitude for $t \in [0, 0.5]$, and consequently the first 1–3 hours of (PDP 11–34) computation had to be discarded.

1.4 Reaching Equilibrium

There are two possible approaches to equilibrium. The transient-free initial data $u_{eq}(t^0)$ of equation (1.5) can be determined directly in some way. The equilibrium solution is then $u(t) = u_{eq}(t)$ for $t \geq t^0$. We shall call this a *direct initialization method.* Alternatively, some arbitrary data u^0 can be chosen. This data is then integrated through some period of time T, chosen to ensure that $u_{trans}(T)$ is acceptably small. Since $u(t) = u_{eq}(t) + u_{trans}(t)$, clearly, $u(t) \approx u_{eq}(t)$ for $t \geq T$. We shall call this an *evolutionary initialization method.*

Direct initialization methods have been in use in meteorological circles since Lewis Fry Richardson's time (see references [119] and [37, page 245]), and many refined techniques are now available. Bode and Sobey [11, page 1393] have suggested adapting these methods in order to eliminate transients, wherever these present a problem. Since our search is for transient-free solutions, we now investigate whether these can help in choosing equilibrium initial data.

1.4.1 Direct Initialization Methods

Direct initialization methods were developed mainly as a result of difficulties encountered with the numerical solution of various problems. However, these methods essentially concern analytic solutions, and may be discussed without reference to numerical techniques.

For many physical problems, it is known empirically that waves do not occur at the maximum possible characteristic speed with appreciable amplitudes. If a_{typ} is the maximum wave speed *observed* in a hyperbolic system, and a_{max} is the maximum characteristic speed admitted by the HPDE modelling the system, then we can define a smoothness ratio as

$$S \equiv \frac{a_{typ}}{a_{max}}.$$

When $S = 0$ we have a steady solution of an evolutionary problem; when $S = 1$ we have a wave-like solution. Intermediate values of S result in intermediate behaviour. When S is small, we shall say the solution is *super-smooth*.

Examples of Super-Smooth Solutions

The two-dimensional shallow water equations are sometimes used as approximate meteorological forecast equations [37, page 191]. The dispersion relation for these equations is given by equation (1.22): we assume $U = V = 0$. There are slow-moving long waves for which $\omega \approx \pm f$, and fast-moving short waves for which $\omega \approx \pm c(k^2 + l^2)^{1/2}$. In an undisturbed region, slow-moving waves predominate: the atmosphere is usually in approximate geostrophic balance. Consequently, observed velocity fields are usually super-smooth. However, measured velocity fields are always contaminated with errors, and so are always out of geostrophic balance. Measured velocity fields therefore result in physically unrealistic fast-moving transients, unless the initial data is carefully 'treated' in some way.

The primitive equations of meteorology present similar difficulties. These equations admit Rossby (or planetary) waves, which owe their existence to variation of the Coriolis force with latitude (the so-called β-effect, [61, page 165][37, page 498]), and typically have a horizontal speed of about 10ms^{-1} [61, pages 197,166]. However, the equations also admit faster inertia-gravity waves, which owe their existence to hydrostatic pressure variations. The velocity of these waves is mainly vertical, and an order of magnitude larger. It is known empirically that the fast inertia-gravity waves usually have negligible amplitude in nature, and they appear to play a minor role in determining the weather over a period of a few days [4][61, pages 175,205]). However, in weather models, errors in the measurement of the initial data field often result in these waves subsequently having artificially large amplitudes.

Ensuring Super-Smoothness

It is highly unlikely that super-smooth solutions will arise from arbitrary initial data. Indeed, initial data can always be contrived that will be advected at the maximum characteristic speed, giving $S \approx 1$. Direct initialization methods ensure the super-smoothness of solutions by contriving special initial data that suppresses fast moving waves.

A common initialization procedure in meteorology consists of integrating the measured initial data back and forth in time, using some (usually heavily damped) numerical method. This is called an *iteration method* [52, page 259]. Further discussion of initialization procedures for meteorological problems may be found in references [37, page 498], [52] and the references cited therein.

A second initialization procedure is Kreiss' *bounded derivative method* for linear hyperbolic IBVPs, introduced in references [76,44,45]. The crux of this procedure is the identification of solutions containing fast-moving waves with solutions having large time derivatives. It is then shown that, if the initial data is prepared so that the time derivative of the solution is initially small, then the time derivative will remain small over any subsequent (finite) period of time. This method is applied to the shallow water equations problem in references [15,16], and extensions of this work to some non-linear problems may be found in references [17,4]. We also note Higdon's work [58]. This deals with problems where the initial data is known accurately, but the boundary data is not. Boundary conditions are devised which ensure that the solution will be super-smooth.

Relevance to Equilibrium Computations

For these methods to be helpful in the determination of equilibrium initial data, it would have to be shown that equilibrium solutions are always super-smooth. The iteration and bounded derivative methods could then be used as direct initialization procedures.

Unfortunately, equilibrium solutions need not be super-smooth. The linear advection equation of Section 1.3.2 is one example. The boundary forcing is advected at speed a, so the equilibrium solution has $S = 1$. This observation calls into question the remarks quoted on page 11, which propose that the 'settling down' time be judged sufficient when a smooth solution is achieved. This claim may well hold true in the presence of friction or dissipation, but no proof is known to the author.

1.4.2 Evolutionary Initialization Methods

For these reasons, only evolutionary initialization methods will be considered in this thesis. Of course, evolutionary initialization methods can only be used when discarding the transient is an option, and this is not always the case. For example, in operational meteorological forecasting, the forecaster cannot discard a three-day forecast and refuse to comment on the weather over that period!

1.5 Numerical Methods for Evolutionary Problems

Evolutionary initialization methods involve the integration of evolutionary problems through large periods of time. Before criteria are given which ensure that a numerical method is particularly well suited to equilibrium computations, some general classes of numerical methods are considered.

Consider finite difference methods for the solution of hyperbolic IBVPs — a *method* being an interior *scheme* together with suitable numerical boundary conditions. Finite difference schemes for one-dimensional hyperbolic problems fall into three broad classes.

(i) There are explicit schemes with fixed stencils. By the CFL condition [120], these schemes are necessarily conditionally stable for IVPs, in the Lax–Richtmyer sense [120].

(ii) There are explicit schemes with stencils which adapt to the CFL number, such as the characteristic Galerkin methods described in references [95,94]. These schemes may be unconditionally LR-stable, but they require $O(\nu)$ boundary conditions at large CFL numbers ν.

(iii) There are implicit schemes, which may or may not be unconditionally LR-stable.

In practice, only explicit fixed stencil and implicit unconditionally stable schemes are widely used.

There is an enormous body of theory to help with the selection of a suitable finite difference scheme for the computation of wave-like solutions. The history of finite difference schemes for wave-like solutions can be viewed as a series of attempts to devise schemes whose physics matches that of the underlying HPDE. Work began with accuracy and stability investigations, and has been carried through to phase speed and dissipation studies for linear problems. For non-linear conservation laws, TVD schemes that eliminate spurious

oscillations near shocks, and shock fitting, have reached a high state of development. The preferred schemes are nearly always explicit conditionally stable schemes. This is reasonable since (for IVPs) explicit schemes have a finite speed of signal propagation, and, correspondingly, a finite domain of dependence; this reflects the physics of exact hyperbolic IVPs. Unconditionally stable (implicit) schemes, on the other hand, have an unbounded domain of dependence. One must put up a good argument for using methods whose physics differs so radically from that of the analytical problem.

When computing truly steady solutions, many of these physical criteria become irrelevant. Time dependency is introduced into the equations artificially, merely to provide a heuristic for the development of iterative methods for solving the systems that result from discretizing the time-independent problem. Time accuracy is of little importance, and phase accuracy concerns are irrelevant. In fact, the (time-dependent) analytical equations are often *changed* to help obtain steady solutions more quickly. Viviand's H-system for the Euler equations [153] and the artificial compressibility method for the incompressible Navier–Stokes equations [109] are just two examples.

Recently, however, it has been more common to retain the time-dependent analytical equations, and integrate initial data through large periods of time, obtaining an approximation to the steady state. Throughout the 1980s there was controversy over whether implicit or explicit methods should be used. Many researchers favoured damped explicit time-stepping methods (such as Lax–Wendroff time-stepping). These migrate easily to vector computers, and deal well with shocks (shock capturing). More recently, unconditionally stable implicit methods have been used [110], on the grounds that large time steps (relative to the spatial mesh size) may be able to skip over transients, reaching the steady state more rapidly. This change has been largely the result of greater experience with transferring implicit algorithms to parallel and vector computers [121]. One recent example is the scheme proposed by Lerat and Sidès [81] for the steady Euler equations.

Equilibrium problems have both wave-like and steady features, but do not fit neatly into either category. Time accuracy, inessential for steady problems, *must* be maintained for equilibrium problems. On the other hand, evolutionary initialization methods have much in common with steady state computations. This raises the question of what sort of finite difference scheme is most suitable for the computation of equilibrium solutions.

1.5.1 Desirable Characteristics of Schemes for Equilibrium Problems

When a discretization that employs evolutionary initialization is used to compute an approximation to an equilibrium solution, the discretization will produce a *discrete transient* superimposed upon a *discrete equilibrium solution.*

As with steady state calculations, the aim of evolutionary initialization is to ensure that the discrete transient is negligible. (For a steady calculation, that is all there is to the solution procedure.) This suggests the notion of *damping efficiency*: that is, the reduction in the size of the transient per unit of computational work.

However, the approximate equilibrium solution is only complete when the equilibrium initial data has been integrated through the final period of time. This suggests the idea of *error efficiency*: that is, the reduction in equilibrium error per unit of computational work.

Implicit Methods for Evolutionary Initialization

It would seem best to use an implicit finite difference scheme with a large CFL number to skip over the transients. These schemes appear to have the best damping efficiency [110]. Furthermore, the time accuracy requirement means that local time-stepping cannot be used. On distorted meshes with small elements, implicit methods must be used at large CFL numbers to ensure reasonable time steps. However, implicit methods are frequently criticised for failing to observe the physics of HPDEs. It is claimed that the CFL limit is necessary for accuracy (see especially Löhner [86]). These charges must be answered.

Accuracy at Large CFL Numbers

Methods can only be used successfully with large CFL numbers in special circumstances that depend on how well the finite difference solution approximates the exact solution. This raises problems, since the discrete solution usually lies in a different space from the exact solution. These problems will be discussed in the next chapter, but for the time being we offer the following heuristic arguments.

Envisage a (fixed size) grid, with space steps of size h, and time steps of size k. The CFL number is then

$$\nu \equiv a_{max}\frac{k}{h}.$$

Now *on approximation grounds* we would expect to need to dispose the grid points so that

$$1 \approx a_{typ}\frac{k}{h} = S\nu \implies \nu \approx S^{-1};$$

otherwise computational effort is wasted. Thus, when $S \ll 1$, on *approximation grounds* we could expect to use suitable schemes with large CFL numbers and still obtain an accurate discrete solution. At the other extreme, of $S \approx 1$, again on approximation grounds, we would expect $\nu \approx 1$ to be necessary for the discrete solution to be accurate.

This sort of reasoning can be seen behind the work of several authors. For example, Beam and Warming [5, page 88] remark that

> ... the use of an implicit scheme usually makes the tacit assumption of a priori knowledge of the solution — in particular that the features which determine the explicit CFL condition are 'parasitic stiff' and need not be accurately resolved.

Again (on page 101), in the context of the two-dimensional Euler equations for a fluttering aerofoil at near-critical flow rates of $u \approx a$, we have

> The numerical stability of an explicit scheme would be governed by the largest eigenvalue $u + a$; however, for low-frequency airfoil motion, the predominant unsteady aerodynamic forces are governed by the smallest eigenvalue $u - a$. Thus the use of an implicit algorithm seems justified in order to set the time step to correspond with the desired accuracy.

Also, in reference [45, page 209],

> If an implicit scheme is used with a time step which is significantly longer than what would be allowed by the Courant–Friedrichs–Lewy condition, then we can expect accurate solutions only if the slow waves are the dominant part of the solution.

In the parlance introduced previously, good error efficiency can only be obtained at large CFL numbers when the exact solution is super-smooth (that is, S is small).

However, for evolutionary initialization, good error efficiency is not required; good damping efficiency is all that is needed. Furthermore, there is substantial evidence [110,121] that implicit methods do provide good damping efficiency. For this reason, implicit methods at large CFL numbers are advocated for evolutionary initialization.

Implicit Methods for the Final Integration

Ideally, the same numerical scheme should be used for both the initialization and the final integration phases of the calculation. This implies that an implicit method should be used for the final integration phase. However, as discussed above, good accuracy cannot be expected unless either the exact solution is super-smooth, or a smaller CFL number ($\nu \approx 1$) is used. Thus implicit schemes, with CFL numbers of approximately 1, are advocated for the final integration.

There are other reasons for advocating implicit schemes. Such schemes are frequently energy conserving. For many problems this property is crucial [112].

1.5.2 Time-Accurate Implicit Schemes for HPDEs

For the reasons given above, time-accurate implicit finite difference methods are advocated for the computation for equilibrium solutions. This section surveys some suitable schemes.

A number of time-accurate implicit schemes have been devised for hyperbolic problems. Beam and Warming [5] have proposed a scheme which reduces to the Crank–Nicolson scheme for one-dimensional linear advection problems: trapezoidal time-stepping is used in time, together with central differences in space. Normally, application of such a scheme to non-linear problems results in a non-linear system which must be solved by Newton's method. To avoid such systems, Beam and Warming incorporate one Newton iteration into their scheme. Consequently, their scheme can integrate a non-linear problem merely by inverting a tridiagonal linear system at each time step. For multi-dimensional problems, the ADI factorization is used.

This scheme has been widely used for aerodynamic flutter problems [131]. It is also popular for two dimensional hydraulics problems, where it goes by the name of the Leendertse scheme [154]. The scheme migrates easily to vector computers, and even to massively parallel machines such as the Connection Machine [1]. Schemes similar to Beam and Warming's are described in references [137,81].

Weare [158] has noted that the ADI factorization can cause problems when the computational grid is not aligned with boundaries. To correct this deficiency, variations of Beam and Warming's scheme have been proposed [159,150], based on the analysis of Stelling et al. [139] of the shortcomings of the ADI approach. (See also Casulli's scheme [21, page 74] for the two-dimensional shallow water equations, and his comments on the ADI factorization.) Pulliam [110] objects to ADI factorizations on other grounds, as well as having general doubts about the suitability of trapezoidal time-stepping for the computation of steady states.

Casulli and Greenspan [22] have developed an implicit scheme for the unsteady Euler equations. This SOLA–ICE method is an extension of the earlier ICE [53] and modified ICE [54] methods. Finally, various finite volume methods with implicit time-stepping have been proposed. See reference [131].

1.5.3 Other Methods

A number of other non-implicit time-accurate methods may also be suitable for equilibrium computations. Jespersen [69] has proposed a time-accurate multigrid scheme. This scheme uses an (implicit or explicit) relaxation scheme that is consistent with the time-dependent equation. For a hierarchy of g grids, with mesh sizes $h_0, \ldots, h_g$, the time step for each sawtooth cycle is $k = k_0 + \cdots + k_g$. For hyperbolic equations for which $k_i = \text{const.} h_i$, this results in efficient evolution even with explicit smoothers. Jespersen reports good oscillating airfoil lift results with CFL numbers of order 10.

Next we must mention the semi-implicit schemes developed initially by Robert [122], and later refined by Kwizac and Robert [78]. These schemes were designed to overcome problems with the integration of the primitive equations of weather prediction. As we have discussed, initial data for these equations is often specially prepared to ensure that the resulting solutions are super-smooth. Implicit schemes cannot be used, since these result in enormous systems of non-linear equations which must then be solved by iterative means. Semi-implicit methods, however, employ implicitness only in those parts of the primitive equations that support the fast acoustic and inertia-gravity waves. Thus, the important slow waves are essentially dealt with using explicit methods, but the fast waves, of negligible amplitude, are dealt with implicitly. Consequently, the time step is constrained by the less restrictive CFL condition $a_{slow} k/h < 1$, where a_{slow} is the important slower wave speed, and the linear systems to be solved at each time step are more manageable. See references [157,52,42] for further details. Various semi-implicit schemes have been proposed for the computation of super-smooth solutions [45,42,44]. LeVeque's 'time split' schemes [82] are related.

1.6 The Need for a New Convergence Theory

In the previous sections we have surveyed a range of schemes that might be suitable for the computation of equilibrium solutions. Our task now is to devise criteria for deciding which schemes are best; or, indeed, whether a given scheme works at all. At first glance it would seem that all we require in a numerical method for implementing evolutionary initialization is convergence over an appropriate time interval $[0, T]$: this would ensure that $u(T) \approx u_{eq}(T)$, and that the transient was suitably small. However, there are two difficulties with this reasoning.

Suppose that the exact problem is posed on some domain Ω, and that the solution is of the form $u(t) = u_{eq}(t) + u_{trans}(t)$, where u_{eq} is the equilibrium component of the solution, and u_{trans} is a decaying transient. Denote the corresponding numerical solution by $u_h(t) = u_{eq,h}(t) + u_{trans,h}(t)$. Convergence over a fixed time interval merely shows that, in some norm,

$$\|u - u_h\|_{\Omega \times [0,T]} \to 0 \tag{1.23}$$

as $h \to 0$. The first problem is whether such an estimate allows the conclusion

$$\|u(T) - u_h(T)\|_{\Omega} \to 0. \tag{1.24}$$

Estimates like this are needed because, once the initialization procedure is complete, the prepared initial data must be integrated over some additional period of time; and bounding this evolutionary error requires pointwise estimates of the accuracy of the initial data.

In Chapter 2 we shall see that for the usual GKS convergence theory we cannot conclude (1.24) from (1.23), since the GKS norm uses an exponentially weighted L^2 norm in time. Thus, GKS convergence does not ensure the pointwise convergence that we demand. However, it will be shown that a modification of the Lax–Richtmyer (LR) theory, which employs a time-uniform norm, can circumvent this problem.

The second difficulty with the estimate of equation (1.24) is that, for sufficiently fine grids, the transient component of the solution $u_{trans}(T)$ may dominate any numerical errors. Whilst the modified LR theory can ensure that (1.23) implies pointwise convergence (1.24), it is not necessarily the case that this latter result will ensure that

$$\|u_{eq}(T) - u_h(T)\|_\Omega \rightarrow 0. \tag{1.25}$$

To ensure equation (1.25), T must be increased as the mesh is refined. Thus, $T = T_h$, and in general $T_h \rightarrow \infty$ as $h \rightarrow 0$. However, the modified LR convergence theory does not apply under these conditions.

A new theory, that does not suffer from these difficulties, is required. To the author's knowledge, the only existing works with a similar aim to that of this thesis, are Young's chapter [160, chapter 9] concerning 'non-stationary linear iterative methods' Larsson's work [80] concerning finite element methods for parabolic problems, Sanz-Serna and Stuart's [128] work on the convergence of finite difference schemes for the reaction-diffusion equation, and Heywood and Rannacher's paper [57] on uniform in time error estimates for schemes for the Navier–Stokes equations.

1.7 An Outline of this Thesis

This chapter has argued the following points.

(i) A heuristic definition of equilibrium solutions has been given.

(ii) In many physical situations, the equilibrium solution is of primary interest.

(iii) The equilibrium solution should be calculated using an evolutionary initialization method.

(iv) Unconditionally stable finite difference methods, at large CFL numbers, should be used for evolutionary initialization.

(v) Existing convergence theories, such as the GKS and Lax–Richtmyer theories, are not helpful for predicting the convergence of numerical methods for the computation of equilibrium solutions.

(vi) A new convergence theory, that allows non-decaying forcing and ensures uniform convergence over $t \geq 0$ is required.

This thesis consistently uses Sanz-Serna's framework for the well-posedness of analytic problems, and for the stability and convergence of discretizations. This framework is reviewed in Chapter 2, where various additional notations are established.

Points (i)–(ii) imply the need for a precise definition of equilibrium solutions. Towards such a definition, Chapter 3 reviews some semigroup theory. The concept of an equilibrium solution relies on the decay of the semigroup, which ensures that the solution is asymptotically independent of the initial data. Essentially, this means that the equilibrium solution

must be asymptotically stable. Chapter 3 discusses various notions of asymptotic stability, and reviews some recent results concerning the asymptotic stability of solutions of hyperbolic IBVPs. The connection between asymptotic stability and well-posedness of the corresponding IBVP is also discussed. Finally, new definitions of equilibrium solutions are offered.

Numerical methods are considered in Chapter 4. Point (iv) suggests that unconditionally stable methods should be favoured. As with the analytic problem, convergence requires the numerical transients to decay. Thus, in Chapter 4, the definition of asymptotic stability is extended to admit discrete solutions, dependent on a grid parameter. Furthermore, necessary conditions are derived for numerical asymptotic stability in terms of the group velocities of waves admitted by the scheme — irrespective of the numerical boundary conditions. As with the exact problem, there is a close connection between this numerical asymptotic stability and stability in a suitable space/time norm. Thus, the necessary conditions for numerical asymptotic stability are also necessary for the stability of the scheme.

Chapter 5 unites the well-posedness, stability and consistency investigations of the previous chapters, and discusses convergence using the generalized Lax equivalence theorem. Stability, and hence convergence, is not possible for schemes that admit waves with zero group velocity. This assertion is proved for one-dimensional linear problems, but remains a plausible conjecture for more general problems.

Chapter 6 is an empirical study to test whether the conjecture carries over to 'real world' problems, such as the St Venant equations of hydraulics. Chapter 7 presents conclusions.

Chapter 2

Schemes and Notions of Convergence

This chapter establishes some notation for finite difference schemes, and reviews a number of schemes for the linear advection equation. Sanz-Serna's theory of discretizations is also reviewed, and various kinds of stability are deduced from the general stability definition. This approach clarifies why existing stability and convergence theories are unsuitable for equilibrium computations.

2.1 Notation for Finite Difference Methods

Consider finite difference methods for the IVP for the scalar linear advection equation

$$u_t + au_x = 0, \quad u(0) = u^0. \tag{2.1}$$

Choose a fixed grid in space $x_i = ih$, $i \in Z$, of width h, and a constant time step k giving the time grid $t^n = nk$, $n = 0, 1, \ldots$ The CFL number is defined as $\nu \equiv ak/h$, and is normally kept fixed throughout a sequence of computations with diminishing mesh sizes h; in which case, we write $k = k(h) = \nu h/a$. Each u_j^n is intended to be an approximation to $u(jh, nk)$. The notation

$$u_h^n = [\ldots, u_{-1}^n, u_0^n, u_1^n, \ldots] \tag{2.2}$$

will be used to denote an approximation to $u(nk)$. Also, $u_h = [u_h^0, u_h^1, u_h^2, \ldots]$ denotes an approximation to u over all of $x \in \Re$, $t \geq 0$.

Finite difference schemes for (2.1) result from the replacement of derivatives by finite differences. For example, the Box scheme uses the approximations to the derivatives evaluated at $(x,t) = ((j-1/2)h, (n-1/2)k)$,

$$\begin{aligned} u_t &\approx \frac{1}{2}\left\{\frac{u_j^{n+1} - u_j^n}{k} + \frac{u_{j-1}^{n+1} - u_{j-1}^n}{k}\right\}, \\ u_x &\approx \frac{1}{2}\left\{\frac{u_j^{n+1} - u_{j-1}^{n+1}}{h} + \frac{u_j^n - u_{j-1}^n}{h}\right\}. \end{aligned}$$

For the linear advection equation, these approximations yield the scheme

$$q(\nu)u_{j-1}^{n+1} + r(\nu)u_j^{n+1} = r(\nu)u_{j-1}^n + q(\nu)u_j^n,$$

where

$$q(\nu) = 1 - \nu, \quad r(\nu) = 1 + \nu.$$

The specification of the discrete IVP is completed by specifying $u_h^0 \approx u(0)$.

Of course, this scheme cannot actually be used *on a computer* to obtain a discrete solution to the discrete IVP, since progressing from u_h^n to u_h^{n+1} requires the solution of an infinite system of linear equations. It can, however, be used for the computer solution of an IBVP on a finite domain.

Interest in the IVP is mainly theoretical. The IVP is simple to analyse, since there are few mathematical difficulties involved in solving the infinite linear system. Our hope is that the IVP will yield information that also applies to IBVPs. In this context, note that schemes for the linear advection equation are often studied with spatially periodic boundary conditions (the periodic boundary condition problem, or PBCP), since this is similar to the IVP, yet methods for this problem can actually be tested on computers.

2.1.1 Two-Level Schemes

A general two-level finite difference scheme for the linear advection equation can be cast in the form

$$\sum_{j=-l}^{r} a_j(\nu) u_{i+j}^{n+1} = \sum_{j=-L}^{R} b_j(\nu) u_{i+j}^{n}, \quad i \in Z. \tag{2.3}$$

To avoid notational clutter, we shall write this scheme in the following form:

$$\begin{matrix} a_{-l} & \cdots & a_0 & \cdots & a_r \\ = & & & & \\ b_{-L} & \cdots & b_0 & \cdots & b_R \end{matrix} \tag{2.4}$$

The coefficients a_{-l}, a_r, b_{-L} and b_R are all assumed to be non-zero for the CFL number being considered. The bounds l, L, r and R are (possibly negative) integers, which may change with ν. When there is only one non-zero a_i, the scheme is said to be *explicit*. Otherwise, it is known as an *implicit* scheme, and time-marching usually requires the solution of a linear system at each time step. However, numerical methods are commonly designed so that the resulting linear systems are sufficiently structured for efficient solution.

In Section 2.3.1 it will be shown that, for IVPs and PBCPs, the choice of *centre* a_0 in equation (2.3) is arbitrary; the choice of centre does not affect the convergence of a scheme for IVPs/PBCPs. Any coefficient a_i could be deemed the centre, provided l, r, L and R are adjusted appropriately. For this reason, the position of the equals sign in equation (2.4) is not significant.

However, we shall see later that the choice of centre is important for the stability of methods for IBVPs, because it amounts to a choice concerning the disposition of numerical boundary conditions.

Corresponding to equation (2.3), we have the *symbols*

$$a(z,\nu) \equiv \sum_{j=-l}^{r} a_j(\nu) z^j, \quad b(z,\nu) \equiv \sum_{j=-L}^{R} b_j(\nu) z^j. \tag{2.5}$$

Equivalently, the coefficients a_j are the Fourier coefficients of the continuous function $a(z,\nu)$. Wherever $a(z,\nu)$ is non-zero, define

$$\lambda(z,\nu) \equiv \frac{b(z,\nu)}{a(z,\nu)}. \tag{2.6}$$

The function $\lambda(e^{i\xi}, \nu)$ is known as the scheme's *amplification factor*, because it describes the amplitude and phase changes undergone by a mode $u_j = \exp(ij\xi)$ when a single step of the scheme is applied. The amplification factor is extremely useful for the analysis of the stability of schemes for IVPs and PBCPs.

For schemes for IVPs/PBCPs for the scalar linear advection equation over $t \gtrsim 0$ or over some finite time interval, it can be shown that Lax–Richtmyer stability is equivalent [135,120] to the *von Neumann condition* (or VNC)

$$|\lambda(e^{i\xi}, \nu)| \leq 1, \quad \forall \xi \in (-\pi, \pi].$$

For non-scalar IVPs/PBCPs (i.e. systems), this condition remains necessary, but is no longer sufficient [99,135]. One aim of this thesis is to develop an alternative to the Lax–Richtmyer theory that is better suited to equilibrium computations. However, it will be shown that this new stability definition, even for IBVPs, always requires the satisfaction of the VNC. Thus the satisfaction of the VNC, along with consistency, are considered minimum requirements in a scheme.

A necessary condition for stability is that the numerical domain of dependence include the analytic domain of dependence. This is the Courant–Friedrichs–Lewy condition, or *CFL condition*, for IVPs. Explicit finite difference schemes with bounded stencils can only be *conditionally* stable: that is, they can only be stable for the finite range of CFL numbers that satisfies the CFL condition. Implicit finite difference schemes, on the other hand, can be *unconditionally* stable. It must be remembered that 'stable' here means stability in the sense of the VNC: that is, Lax–Richtmyer IVP stability for scalar problems. Note that if unbounded stencils are allowed, then explicit schemes may also be unconditionally stable. For example, the Euler–Characteristic–Galerkin (ECG) scheme [95,94] on a uniform grid, with piecewise constant elements, may be written as

$$u_j^{n+1} = (1 - \hat{\nu})u_{j-p}^n + \hat{\nu}u_{j-p-1}^n, \quad \hat{\nu} \equiv \nu - p \in [0, 1),$$

where p is a suitable integer. This scheme is unconditionally stable in the VNC sense. However, this scheme requires $p+1$ left-hand boundary conditions for the solution of IBVPs, which may be unwieldy for large ν.

2.1.2 Two-Level Unitary Schemes

A scheme with

$$|\lambda(z, \nu)| = 1, \quad \forall |z| = 1$$

conserves the $\ell_h^2(\Re)$ norm of the initial data, and so may be termed an *energy-conserving* or *unitary* scheme. These schemes satisfy the VNC unconditionally. They are particularly interesting for hyperbolic problems, since they have wave-like solutions.

Next, some examples of two-level unitary schemes for the linear advection equation are presented. Consider the following four schemes (names to be abbreviated as indicated):

- *Box* – Box scheme
- *CN* – Crank–Nicolson scheme
- *AD* – Angled-Derivative scheme
- *SL* – Sloan's scheme

Scheme	Coefficients	
Box scheme	$\begin{matrix} q & & r \\ & = & \\ r & & q \end{matrix}$	$q = 1 - \nu$ $r = 1 + \nu$
Crank–Nicolson scheme	$\begin{matrix} q & r & s \\ & = & \\ s & r & q \end{matrix}$	$q = -\nu/4$ $r = 1$ $s = \nu/4$
Angled-Derivative scheme	$\begin{matrix} q & r & \\ & = & \\ & r & q \end{matrix}$	$q = -\nu$ $r = 2 + \nu$
Sloan's scheme	$\begin{matrix} q & r & s \\ & = & \\ s & r & q \end{matrix}$	$q = (1-\nu)(2-\nu)$ $r = 2(4-\nu^2)$ $s = (1+\nu)(2+\nu)$

Table 2.1: Some unitary schemes for the linear advection equation.

Because these schemes are unitary, the coefficients a_i and b_i are necessarily related. Consequently, the schemes are defined solely in terms of the coefficients a_i. These schemes are given in Table 2.1 in the sparse notation of equation (2.4). What we have called Sloan's scheme has been discovered at least twice: once in Strang and Iserles paper [63], and later, in a quite different way, in Jamieson and Sloan's paper [64]. The Angled-Derivative scheme was introduced by Roberts and Weiss [123]; the Box scheme is due to Keller, Preissmann, Wendroff and others [154, page 101].

The scalar linear advection equation admits wave-like solutions of the form

$$u(x,t) = e^{i(\kappa x + \omega t)},$$

provided that ω satisfies the exact *dispersion relation* $\omega = -a\kappa$. Defining the scaled wave numbers $\eta \equiv \omega k$ and $\xi \equiv \kappa h$, this dispersion relation becomes

$$\eta = -\nu\xi, \tag{2.7}$$

and $u(jh, nk) = \exp(i(j\xi + n\eta))$. On the grid, it is only sensible to consider $\eta, \xi \in (-\pi, \pi]$, since other frequencies are 'aliased' back onto this interval. Alternatively, the discrete dispersion relations can be viewed as being periodic in ξ and η. For this reason, the edges $\xi = \pm\pi$ and $\eta = \pm\pi$ can be identified, and real dispersion relations may be viewed as closed paths on the unit torus $T^2 = \{(z_1, z_2) : |z_1| = |z_2| = 1\}$. Note that exact advection may be regarded as a scheme with the amplification factor $\lambda(z, \nu) = z^{-\nu}$.

Unitary schemes also have wave-like solutions of the form

$$u_i^n = e^{i(j\xi + n\eta)},$$

Scheme	Dispersion relation
Exact	$\eta = -\nu\xi$
Box scheme	$\tan(\eta/2) = -\nu \tan(\xi/2)$
Crank–Nicolson scheme	$\tan(\eta/2) = -\frac{\nu}{2} \sin\xi$
Angled-Derivative scheme	$\tan(\eta/2) = -\frac{\nu}{2} \frac{\sin\xi}{[1+\nu\sin^2(\xi/2)]}$
Sloan's scheme	$\tan(\eta/2) = -\frac{\nu}{2} \frac{6\sin\xi}{[(4-\nu^2)+(2+\nu^2)\cos\xi]}$

Table 2.2: Dispersion relations for various schemes for the linear advection equation.

with a *real* (centre-independent) dispersion relation $\eta = \eta_{scheme}(\xi)$. Since unitary schemes have real dispersion relations, we can define the ***phase velocity***, c_p and ***group velocity***, c_g in the usual way:

$$c_p(\xi) = \frac{\eta}{\xi}, \quad c_g(\xi) = \frac{d\eta}{d\xi}.$$

Of course, the group velocity is the asymptotic velocity of energy propagation [83,152], and the phase velocity is the velocity of propagation of wave crests. For each of the four unitary schemes, the corresponding real dispersion relation is given in Table 2.2. These dispersion relations are graphed for CFL numbers of $\nu = 0.5$, 1.5, 4.0 in Figures 2.1–2.5.

These figures allow the schemes to be divided into two classes: those that have a turning point in their dispersion relations, and those that do not. Specifically:

- The Box scheme and Sloan's scheme for $|\nu| > 1$ do not have turning points in their dispersion relations.
- The Crank–Nicolson scheme, the Angled-Derivative scheme and Sloan's scheme for $\nu \in (-1, 1)$ do have turning points.

This is important, since turning points in the dispersion relation imply the existence of frequencies ξ for which the group velocity is zero; and a zero group velocity (at least for the IVP) means that energy cannot propagate. In Chapter 4 it will be shown that this reasoning carries over to IBVPs for more complex problems: $c_g(\xi) = 0$ slows the escape of energy from finite computational domains, with the result that the discrete transient has a poor rate of decay compared with the rate of decay of the analytical transient. For unitary schemes, the existence of modes with zero group velocities can be predicted directly from the stencil, as we shall now show.

Recall that, in complex analysis, a ***path*** is a continuous image of $[0, 1]$. It can be shown [70, page 143] that any closed path from 1 to 1 in $\mathcal{C} \setminus \{0\}$ is homotopic to some unique path $\gamma^m(s) = e^{2m\pi is}$, with $m \in \mathcal{Z}$; in which case m is the ***winding number*** of the closed path, denoted $W(\gamma)$. When $W(\gamma) = 0$, the closed path is homotopic to the constant path $1 = \gamma^0(s)$ (or any other non-zero constant) in $\mathcal{C} \setminus \{0\}$.

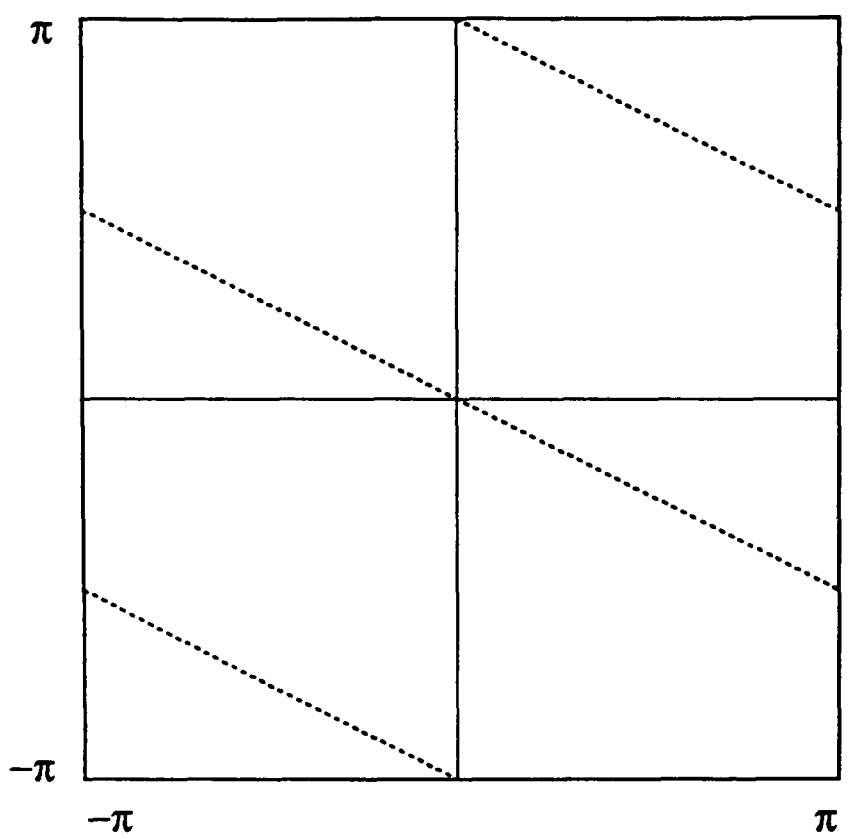


Figure 2.1: The exact dispersion relation for the linear advection equation for $\nu = 0.5$.

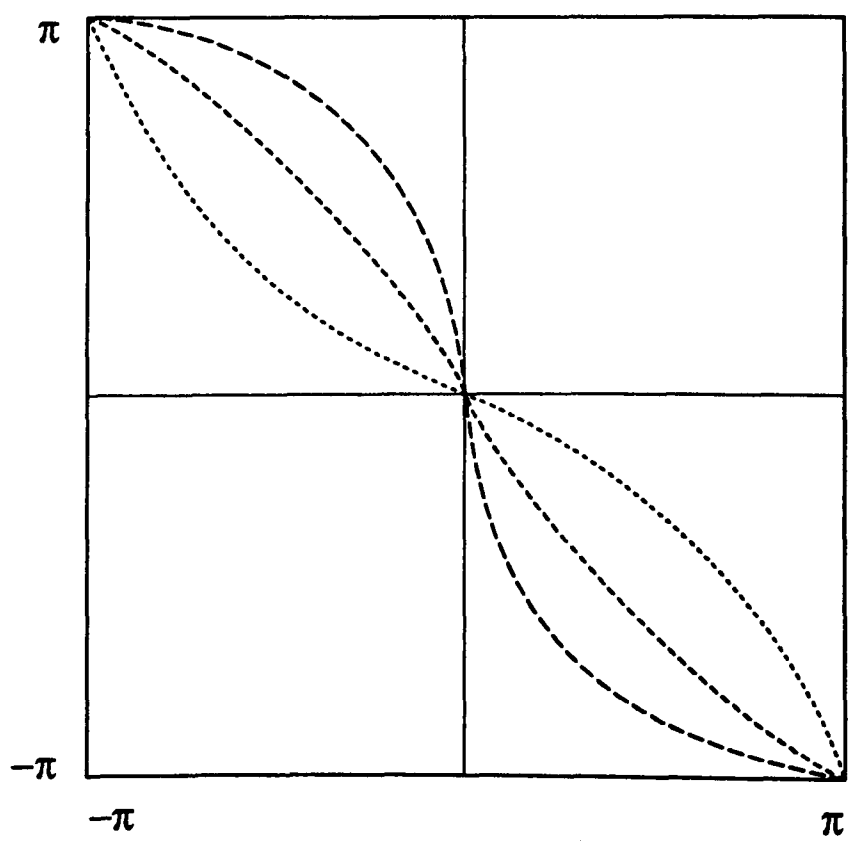


Figure 2.2: The dispersion relation for the Box scheme for $\nu = 0.5$, 1.5, 4.0.

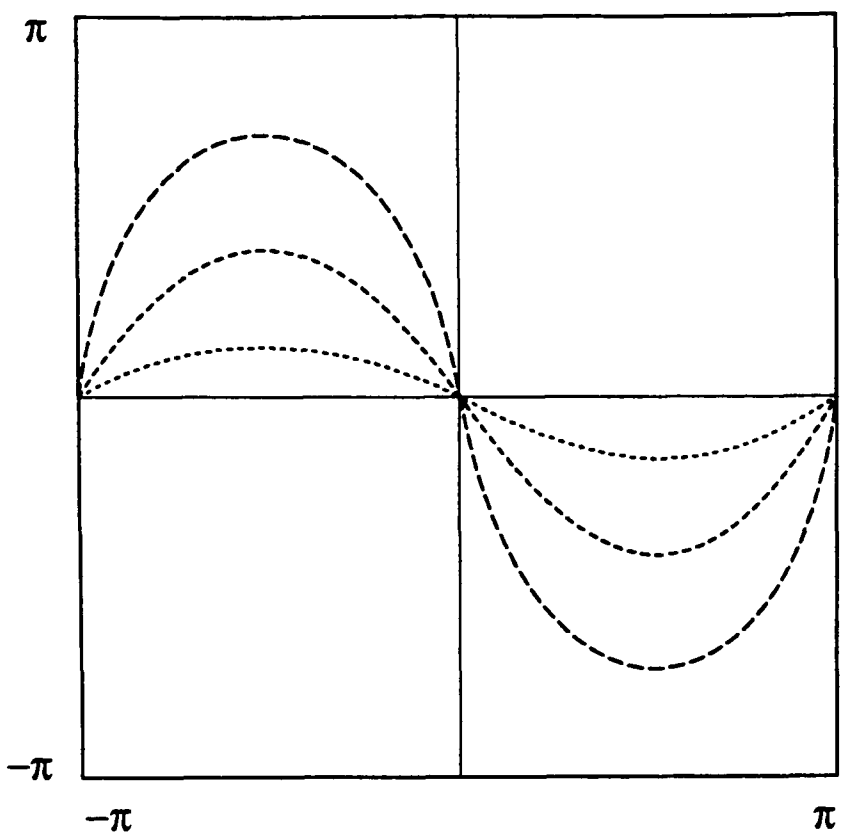


Figure 2.3: The dispersion relation for the Crank–Nicolson scheme for $\nu = 0.5, 1.5, 4.0$.

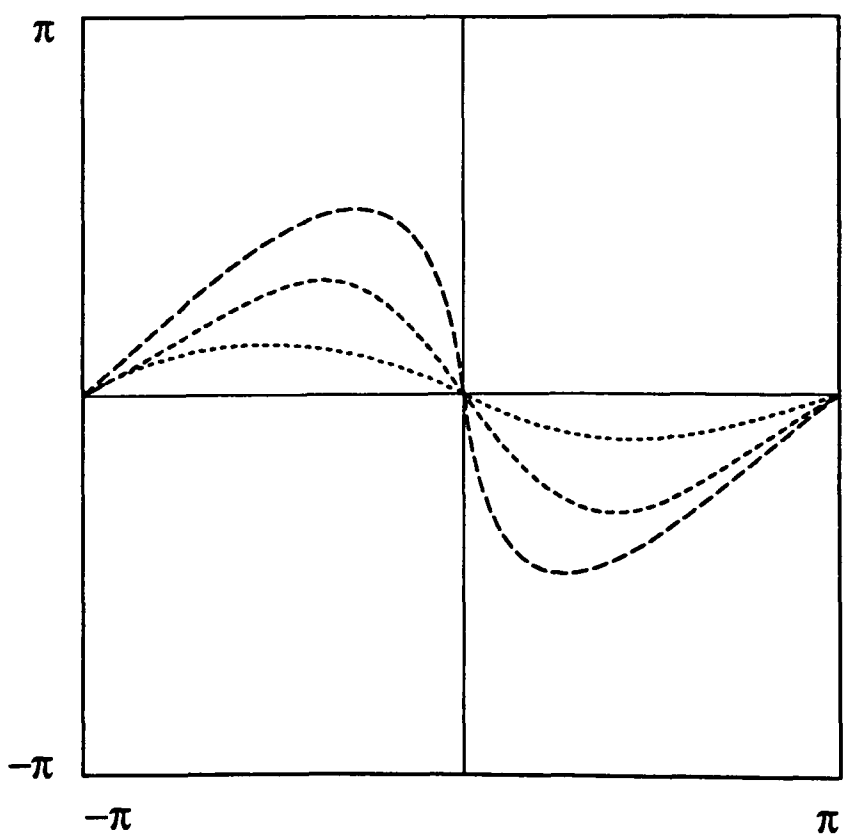


Figure 2.4: The dispersion relation for the Angled-Derivative scheme for $\nu = 0.5, 1.5, 4.0$.

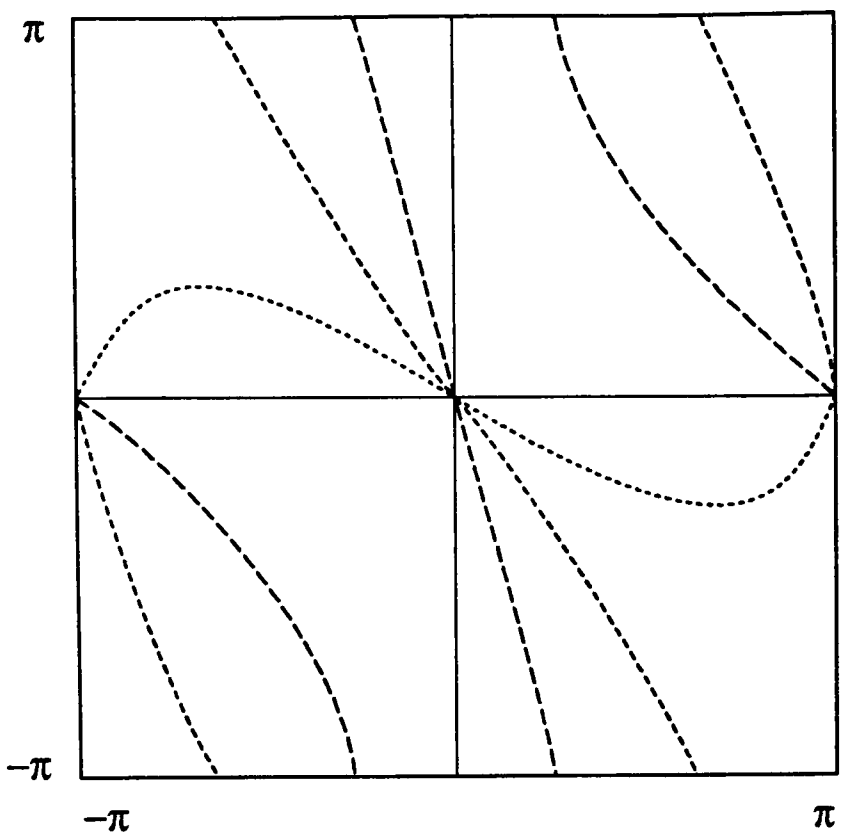


Figure 2.5: The dispersion relation for Sloan's scheme for $\nu = 0.5, 1.5, 4.0$.

Let ϕ be any function that is continuous and non-vanishing on a closed path γ. Then $\phi(\gamma)$ is a closed path in $\mathcal{C} \setminus \{0\}$, so the winding number of ϕ about γ, denoted $W_\gamma(\phi)$, may be defined as the winding number of this path. Thus $W_\gamma(\phi) = W(\phi(\gamma))$. When γ is the unit circle S^1, denote $W(\phi) \equiv W_{S^1}(\phi)$. Note that, for ϕ analytic in some domain containing γ, and non-vanishing on γ,

$$W(\phi) \equiv \frac{1}{2\pi i} \int_\gamma \frac{\phi'(z)}{\phi(z)}\, dz. \tag{2.8}$$

Clearly, when $W(\phi) = 0$, the closed path $\phi(S^1)$ is homotopic to 1 in $\mathcal{C} \setminus \{0\}$. Similarly, if $\phi(z_1, \ldots, z_d)$ is non-vanishing and continuous on the d-torus $T^d \equiv S^1 \times \cdots \times S^1$, for analogous ϕ , the surface $\phi(T^d)$ is said to be homotopic to 1. In this case ϕ is ***non-vanishing and homotopic to*** 1 ***on*** T^d, without further qualification.

Returning to the discussion of group velocity, observe that, for a general two-level scheme,

$$\lambda = \frac{b}{a} \implies W(\lambda) = W(b) - W(a).$$

Since $W(\lambda)$ does not depend on the centring, the centre may be chosen so that $W(a) = 0$ with no loss of generality. (Such a choice is always possible.) A unitary scheme has $b(z) = z^\alpha a(z^{-1})$, where $\alpha = r - L = R - l$ depends on the alignment of the explicit and implicit parts of the stencil, but not on the choice of centre. Consequently,

$$\begin{aligned} W(\lambda) &= \alpha + W(a(z^{-1})) - W(a(z)) \\ &= \alpha - 2W(a) \\ &= \alpha, \end{aligned}$$

which gives the number of times the dispersion relation winds around the unit torus. When $W(\lambda) = 0$, the dispersion relation must have a turning point. Thus, some frequency must have zero group velocity, and we have proved the following result.

Proposition 2.1 *A unitary scheme has* $W(\lambda) = r - L = R - l$. *If* $r = L \iff l = R$, *then there exists* ξ_0 *such that* $c_g(\xi_0) = 0$, *and* ξ_0 *is a turning point.*

This result predicts the admission of zero group velocity modes by the CN, AD and SL ($|\nu| < 1$) schemes. Also, this result predicts that the dispersion relation for the Box scheme will have $W(\lambda) = -1$ for $\nu > 0$, and $W(\lambda) = 1$ for $\nu < 0$. Similarly, the Crank–Nicolson scheme will have $W(\lambda) = -2$ for $\nu > 1$, and $W(\lambda) = 2$ for $\nu < -1$.

For a unitary scheme in d dimensions, we conjecture that if λ is homotopic to 1 on T^d, then there will be some frequency $\xi_0 \in T^d$ at which the scheme's group velocity will vanish, and that this frequency will be a maximum of the dispersion relation.

2.2 Outline of Convergence Theory

This section reviews some aspects of the general theory of numerical methods (or *discretizations*) due to Sanz-Serna [105,106,127,126], and follows the presentation of reference [126] quite closely. For evolutionary problems, Cullen and Morton's paper [25] provides a specific framework.

The Analytic Problem

Assume that the analytic problem is

$$Qu = g, \quad Q : \mathcal{D}(Q) \to \mathcal{R}(Q), \tag{2.9}$$

where $\mathcal{D}(Q) \subseteq X, \mathcal{R}(Q) \subseteq Y$, and X and Y are Banach spaces. Q is some linear (bounded or unbounded) operator with $\ker(Q) = \{0\}$. The initial and boundary data are incorporated into $g \in Y$, whilst $u \in X$ is the sought solution; if u exists, then it is unique, since $\ker(Q) = \{0\}$. In the context of evolutionary problems on a domain Ω, note that u is the solution over Ω and time.

If $g \in \mathcal{R}(Q)$, then the (unique) solution of problem (2.9) is a *genuine solution* of the problem. If, additionally,

(i) $\mathcal{R}(Q)$ is dense in Y, and

(ii) $Q^{-1} : \mathcal{R}(Q) \to X$ is bounded,

then the problem is said to be *well-posed*; in this case Q^{-1} can be uniquely extended from $\mathcal{R}(Q)$ to Y, giving $\widehat{Q^{-1}} \in L(Y, X)$ (the Banach space of bounded linear operators $Y \to X$ with the natural norm). If $g \in Y - \mathcal{R}(Q)$, then $u = \widehat{Q^{-1}} g$ is said to be a *generalized solution* of problem (2.9). Observe that the well-posedness of a problem hinges on the choice of X and Y. It requires that the solution depend continuously on the data.

The Discrete Problem

Next consider some family of discretizations of the analytic problem,

$$Q_h u_h = g_h, \quad Q_h : X_h \to Y_h. \tag{2.10}$$

Here $h > 0$ is a parameter which, in some sense, describes the 'fineness' of the discretization. For example, h may be a parameter upon which the space and time steps of a finite difference

scheme for an evolutionary PDE depend. Successive computations, on increasingly fine meshes, are associated with the limit $h \to 0$. For each fixed $h > 0$, the spaces X_h and Y_h are assumed to be Banach.

The discrete counterpart of well-posedness is termed 'stability'. Our definition differs from that given in reference [126, page 73], but is sufficient for our purposes. A family of discretizations $\{Q_h\}_{h>0}$ is said to be (X_h, Y_h)-*stable* iff there exists $h_0 > 0$, $C > 0$ such that

(i) Q_h is invertible in $L(X_h, Y_h)$ for each $h : 0 < h < h_0$, and

(ii) $\sup_{0<h<h_0} \|Q_h^{-1}\|_{L(Y_h, X_h)} < \infty$.

This sort of ***uniform invertibility*** requirement reappears in a variety of contexts throughout the remainder of this thesis. Clearly, the aim of the definition is to ensure that the continuous dependence of the solution on the data is retained as the computational mesh is refined. However, this definition is sometimes unnecessarily strict. 'Weak' instabilities, whilst reducing the efficiency of a scheme, do not necessarily make it unusable for ***sufficiently smooth*** data g. A family of discretizations is *q-weakly stable* in $L(X_h, Y_h)$ iff there exists $h_0 > 0$ such that

(i) Q_h is invertible in $L(X_h, Y_h)$ for each $h : 0 < h < h_0$, and

(ii) $\|Q_h^{-1}\|_{L(Y_h, X_h)} = O(h^{-q})$.

Prolongation and Restriction Operators

In order to speak of the convergence of the numerical solution to the exact solution, elements $u_h \in X_h$ must be compared with elements $u \in X$. For finite element methods, it is often the case that $X_h \subseteq X$, so this comparison presents no problems. However, for finite difference methods this is rarely the case. There are two ways out of this dilemma. First, ***prolongation operators*** $p_h : X_h \to X$ may be defined, and then $p_h u_h$ may be compared with u in X. Secondly, ***restriction operators*** $r_h : X \to X_h$ may be defined, and u_h may be compared with $r_h u$ in X_h. The former approach is rarely used in finite difference studies, since there is rarely a natural way to interpolate the numbers produced by the method. (Reference [25] has an interesting discussion of this point.) For the time being, the restriction operator approach will be adopted.

Some conditions must be imposed on the restrictions. A map $r_h : X \to X_h$ is a ***restriction operator*** iff

(i) r_h is linear for each $h > 0$, and

(ii) for each $x \in X$ we have $\|x\|_X = \lim_{h \to 0} \|r_h x\|_{X_h}$.

Convergence

With this matter settled, the ***global error*** of a discretization is defined as

$$e_h \equiv r_h u - u_h, \tag{2.11}$$

where u is the exact (possibly generalized) solution of problem (2.9). The discretization is said to be *convergent* iff $\|e_h\|_{X_h} \to 0$ as $h \to 0$; it is said to be *convergent of order p* if $\|e_h\| = O(h^p)$ and $p > 0$.

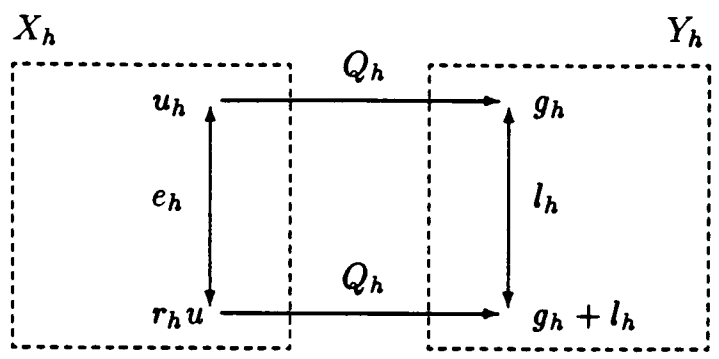


Figure 2.6: The relationship between the discretization error e_h and truncation error l_h. Vertical lines show the distance between elements, and horizontal arrows show the action of Q_h.

To obtain *a priori* estimates of the order of convergence, it is usual to introduce a second collection of restriction operators, $s_h : Y \rightarrow Y_h$. The *truncation error* of a discretization is defined as

$$l_h \equiv Q_h r_h u - s_h g, \tag{2.12}$$

where u is the (possibly generalized) solution of the exact problem. The discretization is said to be *consistent* iff $\|l_h\|_{Y_h} \rightarrow 0$ as $h \rightarrow 0$; it is said to be *consistent of order p* if $\|l_h\| = O(h^p)$ and $p > 0$. Notice that, whilst the stability of a discretization is independent of the function g, its consistency does depend on this function.

Assuming that $g_h = s_h g$, and using the linearity of Q_h, clearly

$$\begin{aligned} Q_h e_h &= Q_h r_h u - Q_h u_h \\ &= Q_h r_h u - s_h g \\ &= l_h. \end{aligned}$$

Consequently, for a stable scheme,

$$\|e_h\| \leq \|Q_h^{-1}\| \, \|l_h\| \leq C \|l_h\|. \tag{2.13}$$

Thus, a discretization that is stable and consistent of order p is convergent of order p for any $g \in Y$. Figure 2.2 illustrates the relationship between discretization and truncation error. The celebrated Lax equivalence Theorem (LET) states — under suitable hypotheses — that consistent discretizations of well-posed exact problems are convergent iff they are stable. The hypotheses will be important in what is to follow, and so are now stated for reference.

(i) There exists $R > 0$ such that

$$\forall h > 0,\ g_h \in Y_h : \|g_h\| \leq 1, \quad \text{there exists } g \in Y \text{such that } \ g_h = s_h g,\ \|g\| \leq R; \tag{2.14}$$

(ii) Q_h is invertible in $L(X_h, Y_h)$ for each $h > 0$; and

(iii) $Q_h^{-1} s_h \in L(Y, X_h)$ for each $h > 0$.

Notice that the first hypothesis links the possible choices of Y_h to the choice of Y. The following form of the LET is proved in reference [126, pages 90,92]; see also references [106,127,105].

Theorem 2.1 *Suppose that the exact problem is well-posed, and that the discretization is stable, and consistent for $g \in Y_0$, a dense subset of Y. Then the discretization is convergent. Conversely, suppose that the exact problem is well-posed and converges for all $g \in Y$. Further, suppose that conditions (i)-(iii) above hold. Then the discretization is stable.*

Note that consistency need only be checked for forcing in a dense subset of Y. This often allows the use of Taylor expansions to check the order of consistency.

If the discretization is merely q-weakly stable ($q < p$), then, from the estimate (2.13), the discretization is convergent of order $p-q$ for *sufficiently smooth* data [135, page 82], say $g \in Y_{smooth} \subseteq Y$. However, for general data $g \in Y$, a weakly stable discretization cannot be expected to be convergent.

L-Convergence

The LET shows that stability is necessary for convergence of a discretization for *all* $g \in Y$. Consequently, showing that a particular method is not stable means that some $g \in Y$ must be found for which the method is not convergent. It is often much easier to show that a scheme is not 'L-convergent' than it is to carry out such a search.

A discretization is said to be *L-convergent* for $g \in Y$ [126, page 78] if, for any $\{\delta_h\}_{h>0} \subseteq Y_h$, with $\lim_{h\to 0} \|\delta_h\| = 0$,

$$u_h : \; Q_h u_h = s_h g + \delta_h,$$

converges to $r_h u = r_h Q^{-1} g$. It can be shown [126, page 79] that a discretization is L-convergent for $g \in Y$ iff it is convergent for $g \in Y$ and stable.

Consequently, if $g \in Y$ and $\{\delta_h\}_{h>0} \subseteq Y_h$ can be found such that $\lim_{h\to 0} \|u_h - r_h u\| \neq 0$, then the discretization cannot be stable. One has considerable latitude in choosing $\{\delta_h\}_{h>0}$, and this is sometimes the easiest way of demonstrating that instability exists.

Approximation Error

Where prolongation operators are available, the prolongation of the discrete solution can be compared with the exact solution. Thus, the error $E_h \in X$ may be defined as

$$\begin{aligned} E_h &\equiv u - p_h u_h \\ &= [u - p_h r_h u] + [p_h r_h u - p_h u_h] \\ &= [I - p_h r_h] u + p_h e_h. \end{aligned}$$

Note that $p_h r_h : X \to X$ is just an interpolation operator. Thus the error E_h has a pure approximation component $[I - p_h r_h]u$, and a component $p_h e_h$ that is the prolongation of the global error.

As was emphasized in Chapter 1, the use of implicit methods at large CFL numbers is usually advocated only where prior knowledge of the exact solution indicates that approximation errors will be small compared with $p_h e_h$. For finite element methods, such arguments are natural, and might be made rigorous. However, for finite difference schemes, there are rarely natural candidates for the (space/time) prolongation operators [25,105,106,127]. Thus

it is simply not possible to discuss the idea of matching the disposition of the grid points to the smoothness of the exact solution. So long as space/time prolongation operators are unavailable, our remarks concerning the suitability of implicit finite difference schemes for super-smooth problems must remain no more than heuristic guidelines.

2.2.1 Lax–Richtmyer Stability for Two-Level Schemes

This section examines how the Lax–Richtmyer stability theory can be deduced from the general stability framework. We shall be considering methods for evolutionary problems that are of the form

$$u_h^{n+1} = G_h u_h^n + k f_h^n, \quad n = 0 \ldots N_h - 1, \tag{2.15}$$

$$u_h^0 = \phi_h, \tag{2.16}$$

over the time interval $[0,T]$. It is assumed that $G_h \in L(B_h)$, where B_h is a Banach space. Recall that N_h is defined by $T = N_h k(h)$. G_h is in some sense an approximation to the exact evolution operator $G(k(h))$. Note that implicit methods can also be written in the above form, provided some care is taken: this point will be discussed later.

The method can be written in the standard form

$$\frac{1}{k(h)} \begin{bmatrix} kI_h & 0 & 0 & \cdots & 0 & 0 \\ -G_h & I_h & 0 & \cdots & 0 & 0 \\ 0 & -G_h & I_h & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -G_h & I_h \end{bmatrix} \begin{bmatrix} u_h^0 \\ u_h^1 \\ u_h^2 \\ \vdots \\ u_h^{N_h} \end{bmatrix} = \begin{bmatrix} \phi_h \\ f_h^0 \\ f_h^1 \\ \vdots \\ f_h^{N_h-1} \end{bmatrix} \tag{2.17}$$

or

$$Q_h u_h = g_h. \tag{2.18}$$

Suppose that $Q_h : X_h \to Y_h$ is a member of $L(X_h, Y_h)$. Of course, stability means that the collection $\{Q_h : h > 0\}$ is uniformly invertible in $L(X_h, Y_h)$. Formally, the inverse of Q_h is given by

$$Q_h^{-1} = \begin{bmatrix} I_h & 0 & 0 & 0 & \ldots & 0 \\ G_h & kI_h & 0 & 0 & \ldots & 0 \\ G_h^2 & kG_h & kI_h & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ G_h^{N_h} & kG_h^{N_h-1} & kG_h^{N_h-2} & kG_h^{N_h-3} & \ldots & kI_h \end{bmatrix}. \tag{2.19}$$

If the spaces (X_h, Y_h) are chosen carefully, Lax–Richtmyer stability may be deduced. Specifically, on a domain Ω, we assume X_h and Y_h have elements in

$$\{w_h : w_h^m \in B_h, \ m = 0 \ldots N_h\},$$

and the norms

$$\|u_h\|_{X_h} = \max_{m=0,\ldots,N_h} \|u_h^m\|_{B_h}, \tag{2.20}$$

$$\|g_h\|_{Y_h} = \|g_h^0\|_{B_h} + k \sum_{m=1}^{N_h} \|g_h^m\|_{B_h}, \tag{2.21}$$

where $k = k(h)$. Using equation (2.19), clearly, Lax–Richtmyer stability for explicit schemes is equivalent to the familiar *power-boundedness* condition

$$\sup_{h>0} \max_{m=0\ldots N_h} \|G_h^m\|_{L(B_h)} < \infty. \tag{2.22}$$

Since $r(G_h) \leq \|G_h\|$, a necessary condition for stability is that

$$r(G_h) \leq 1. \tag{2.23}$$

Of course, for normal G_h, $r(G_h) = \|G_h\|$ and this condition is also sufficient for stability.[1]

For IVPs, the operators G_h are normal, so $r(G_h) = \|G_h\|$; in the power-boundedness condition we can write $\|G_h^m\| = r(G_h)^m$. For the constant coefficient scalar problem, Fourier analysis shows that

$$r(G_h) = \sup_{\xi \in (-\pi,\pi]} |\lambda(e^{i\xi})|,$$

so the power-boundedness condition yields

$$\sup_{|z|=1} |\lambda(z)| \leq 1. \tag{2.24}$$

This is the von Neuman condition.

Finally, note that the definition of Lax–Richtmyer stability is usually restricted to the unforced case, with $g_h = [\phi_h, 0, 0, \ldots, 0]^T$. Where this restriction needs to be made, it will be stated explicitly.

2.2.2 Extended Lax–Richtmyer Stability for Two-Level Schemes

For the problem over the infinite time interval $t \in \Re^+$, choose the spaces

$$X_h \equiv \left\{ (u_h^m) : \max_{m\geq 0} \|u_h^m\|_{B_h} < \infty \right\} \tag{2.25}$$

$$Y_h \equiv \left\{ (g_h^m) : \|g_h^0\|_{B_h} + k \sum_{m=1}^{\infty} \|g_h^m\|_{B_h} < \infty \right\}, \tag{2.26}$$

with the norms for $\{u_h^m\}_{m=0}^{\infty}$ and $\{g_h^m\}_{m=0}^{\infty}$

$$\|u_h\|_{X_h} = \max_{m\geq 0} \|u_h^m\|_{B_h} \tag{2.27}$$

$$\|g_h\|_{Y_h} = \|g_h^0\|_{B_h} + k \sum_{m=1}^{\infty} \|g_h^m\|_{B_h}. \tag{2.28}$$

For brevity, we shall name this form of stability after the time norms as $(\ell^{\infty}_{k(h)}, \ell^1_{k(h)})$-stability, or just $(\ell^{\infty}_k, \ell^1_k)$-stability. We shall also call it *extended L–R stability* (EL–R).

The representation of Q_h given in the previous section is easily modified so that $N_h = \infty$ for all $h > 0$. The formal inverse of Q_h, is given by modifying equation (2.19) reflect the fact that $N_h = \infty$. Evaluating the formal 'norm' of this inverse as before, the inverse is seen to be a member of $L(Y_h, X_h)$ for each $h > 0$, provided

$$\max_{m\geq 0} \|G_h^m\|_{L(B_h)} = C_h < \infty,$$

[1] Interestingly, this conclusion also holds when the G_h are *Hermitian* on a *Banach* (not necessarily a Hilbert!) space: that is,

$$\|e^{itG_h}\| = 1, \quad \forall t \in \Re.$$

For further details, see reference [12, page 31].

for each $h > 0$. Stability in the new spaces is equivalent to

$$\sup_{h>0} \max_{m\geq 0} \|G_h^m\|_{L(B_h)} < \infty, \tag{2.29}$$

which is just a modified power-boundedness condition.

An important limitation of extended L–R stability is now clear. The forcing must lie in ℓ_k^1, and consequently must decay. This supposition precludes the type of forcing that occurs in equilibrium problems.

For constant coefficient IVPs, Fourier analysis can be used to show that stability is equivalent to the VNC of equation (2.24). Thus, for two-level schemes, L–R stability over finite time is equivalent to L–R stability over infinite time.

2.2.3 Stability of Two-Level Implicit Methods

Until now, two-level methods have been written in the explicit form

$$u_h^{n+1} = G_h u_h^n + k f_h^n.$$

However, implicit methods are more naturally written in the form

$$S_h u_h^{n+1} = T_h u_h^n + k g_h^n. \tag{2.30}$$

At first sight, it might be thought that this change is merely notational: by writing $G_h = S_h^{-1} T_h$, the implicit form can be rendered into the explicit form, with a suitably modified forcing function. Thus, it seems that the conditions for the stability of implicit schemes should be the same as those for explicit schemes. However, this reasoning is flawed. When we write software to implement an implicit method, we rarely form G_h explicitly, and never without the introduction of numerical error: we do not simply compute $G_h u_h^n + k f_h^n$ and assign this to u_h^{n+1}. Instead, we compute the term $y_h^n = T_h u_h^n + k g_h^n$, and then solve the linear system $S_h u_h^{n+1} = y_h^n$, using some numerical method.

Solving linear systems of this form accurately requires that the systems be *well conditioned.* That is,

(i) S_h must be invertible for each $h > 0$, and

(ii) the condition number $\mu(S_h) \equiv \|S_h^{-1}\| \, \|S_h\|$ must remain bounded as $h \to 0$ (see references [40] and [62]).

For finite difference schemes, the special form of S_h implies that it has a uniformly (in h) bounded norm, so the condition number requirement simply demands that S_h have a uniformly bounded inverse. This suggests that, for stability, the implicit operator should be uniformly invertible — a point first appreciated by Strang [142,143]. Whilst this conclusion is correct, we would like to prove that it is a direct consequence of the definition of stability.

To this end, the implicit form of method (2.30) over $n \geq 0$[2] is written

$$\frac{1}{k(h)} \begin{bmatrix} kI_h & 0 & 0 & 0 & 0 & \cdots \\ -T_h & S_h & 0 & 0 & 0 & \cdots \\ 0 & -T_h & S_h & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix} \begin{bmatrix} u_h^0 \\ u_h^1 \\ u_h^2 \\ \vdots \end{bmatrix} = \begin{bmatrix} \phi_h \\ g_h^0 \\ g_h^1 \\ \vdots \end{bmatrix}, \tag{2.31}$$

[2]Identical considerations apply over finite time intervals. See reference [126] for further details.

which is in the standard form for discrete methods: $Q_h u_h = g_h$, where $g_h = [\phi_h, f_h^0, f_h^1, \ldots]$ and $u_h = [u_h^0, u_h^1, \ldots]$. Assuming each S_h, $h > 0$ to be invertible, the formal inverse of this block matrix is

$$Q_h^{-1} = \begin{bmatrix} I_h & 0 & 0 & 0 & \ldots \\ G_h & kS_h^{-1} & 0 & 0 & \ldots \\ G_h^2 & kG_hS_h^{-1} & kS_h^{-1} & 0 & \ldots \\ G_h^3 & kG_h^2S_h^{-1} & kG_hS_h^{-1} & kS_h^{-1} & \ldots \\ \vdots & \vdots & \vdots & \vdots & \end{bmatrix}, \tag{2.32}$$

where, naturally, $G_h = S_h^{-1}T_h$. It is important to note that this inverse is different from that given in reference [126, page 102], which I believe to be wrong.

Using the $(\ell_k^\infty, \ell_k^1)$ norms, this representation for Q_h^{-1} shows [126, page 75]

$$\|Q_h^{-1}\| = \max_{m\geq 0} \max\left\{\|G_h^m\| \ , \ \|G_h^m S_h^{-1}\|\right\}. \tag{2.33}$$

Thus, extended L–R stability is equivalent to

(i) $\sup_{h>0} \max_{m\geq 0} \|G_h^m\| < \infty$, and

(ii) $\sup_{h>0} \max_{m\geq 0} \|G_h^m S_h^{-1}\| < \infty$.

These conditions show that power-boundedness and uniform invertibility are *sufficient* for $(\ell_k^\infty, \ell_k^1)$ stability. They also show that power-boundedness is necessary for this form of stability. Furthermore, condition (ii) with $m = 0$ implies the uniform invertibility of $\{S_h\}_{h>0}$. (Our argument differs slightly from that of reference [126, page 102], because of the erroneous inverse of Q_h given in that reference.)

Proposition 2.2 *Extended L–R stability is equivalent to:*

(i) $\sup_{h>0} \max_{m\geq 0} \|G_h^m\| < \infty$ *(power-boundedness of* G_h*), and*

(ii) $\sup_{h>0} \|S_h^{-1}\| < \infty$ *(uniform invertibility of* S_h*).*

For finite-time L–R stability, the power-boundedness need only hold over $m = 0 \ldots N_h$.

2.2.4 The GKS Theory

The GKS paper concerns itself with one-dimensional, constant coefficient, linear, hyperbolic systems of the form

$$u_t = \tilde{A}u_x + \tilde{B}u + F, \tag{2.34}$$

in the quarter plane $x \geq 0$, $t \geq 0$. Here

$$u = [u^I, u^{II}],$$

where $u^I(x,t) \in \Re^\ell$, $u^{II}(x,t) \in \Re^{m-\ell}$, and $F(x,t) \in \Re^m$ for each (x,t). $\tilde{A}$ is an $m \times m$ matrix:

$$\tilde{A} = \begin{bmatrix} \tilde{A}^I & 0 \\ 0 & \tilde{A}^{II} \end{bmatrix}, \quad \tilde{A}^I < 0, \quad \tilde{A}^{II} > 0. \tag{2.35}$$

Here $\tilde{A}^I \in L(\Re^\ell)$ and $\tilde{A}^{II} \in L(\Re^{m-\ell})$ are assumed to have diagonal matrix representations. The boundary conditions are

$$u^I = Su^{II} + g(t) \quad \text{at } x = 0, \tag{2.36}$$

$$u = u^0 \quad \text{at } t = 0, \tag{2.37}$$

where S is a suitable rectangular matrix.

GKS Well-posedness

Of course, equation (2.34), with boundary conditions (2.36, 2.37) can be cast in the standard form

$$Qw = g, \quad Q : X \rightarrow Y. \tag{2.38}$$

Proving well-posedness consists of proving that Q is invertible, and that the extension of the inverse to Y is a member of $L(Y, X)$. The latter, which is the main concern, essentially requires an estimate of the form

$$\|u\|_X \leq C\|(u^0, F, g)\|_Y; \tag{2.39}$$

but first the spaces X and Y must be defined.

The key to (GKS) well-posedness is to employ norms that are weighted (in time) by functions that decay sufficiently rapidly to ensure that *every* problem (2.34, 2.36,2.37) is well posed. Specifically, the following norms are used:

$$\|u\|^2_{X_\alpha} \equiv \|e^{-\alpha t}u(0,.)\|^2_2 + \|e^{-\alpha t}u(.,.)\|^2_2, \tag{2.40}$$

$$\|(u^0, F, g)\|^2_{Y_\alpha} \equiv \|u^0(.)\|^2_2 + \|e^{-\alpha t}F(.,.)\|^2_2 + \|e^{-\alpha t}g(.)\|^2_2. \tag{2.41}$$

The spaces X_0 and Y_0 are defined as

$$\begin{aligned} X_0 &\equiv L_2(\Re^+) \times L_2(\Re^+ \times \Re^2), \\ Y_0 &\equiv L_2(\Re^+) \times L_2(\Re^+ \times \Re^+) \times L_2(\Re^+). \end{aligned}$$

The space X_α is X_0 with norm (2.40); Y_α is Y_0 with norm (2.41). The following result has been well known since the 1950's [74].

Theorem 2.2 *For any $(u^0, F, g) \in Y_0$, there exists $\alpha_0 > 0$ such that (2.34,2.36,2.37) has a unique solution $u \in X_\alpha$ for any $\alpha > \alpha_0$. Furthermore, $\forall\, \alpha > \alpha_0$, $\exists K(\alpha)$:*

$$\|u\|_{X_\alpha} \leq K(\alpha)\|(u^0, F, g)\|_{Y_\alpha}.$$

From the point of view of this thesis, such well-posedness results are simply not helpful, for the following reasons. First, for the purposes of the existence of solutions, the GKS theory assumes that the forcing is square-integrable in time, and so does not admit the non-decaying forcing that occurs in equilibrium problems. Secondly, the time-exponential weights in the norm render it useless for asymptotic studies: from the GKS viewpoint, a pencil standing on its tip is a well posed problem. This circumstance arises because GKS well-posedness is intended to describe systems over *finite time intervals* (as are the attendant numerical stability and convergence results): the quarter plane result implies the well posedness of problems over any bounded domain in $\Re$, and over finite time intervals.

GKS Stability and Convergence

The GKS paper [47] defines stability in terms of the GKS spaces (X_h, Y_h), that are discrete counterparts of X_α and Y_α. The LET then ensures that every consistent, GKS stable method is GKS convergent, since all one-dimensional problems are well-posed for α sufficiently large. Indeed, this seems to be the main conclusion of reference [43].

Unfortunately, this means that error bounds are obtained in the X_h norm [43]. The time-exponential weights, and the use of an ℓ^2 norm in time, mean that the resulting error estimates are useless for discussing the convergence of schemes for equilibrium problems. Futhermore, the results of the GKS paper that allow stability to be checked using extended 'normal mode' analysis all assume the forcing to be square-integrable over $t \geq 0$.

Extensions

Our discussion of (GKS) well-posedness has been restricted to one-dimensional constant coefficient hyperbolic problems. For these problems, well-posedness is a simple matter. In more than one dimension, proving well-posedness requires a substantial amount of work [59]. Multi-dimensional constant coefficient hyperbolic problems were originally analysed by Hersch [56]. Hersch's work was extended by Kreiss [75] to include variable coefficient problems. Kreiss' results have been extended by Rauch [115] to allow inhomogeneous initial data, and by Ralston [114] to allow complex coefficients.

Unfortunately, none of these extensions are of help to us in formulating a theory of well-posedness of equilibrium solutions.

2.3 Uniform Invertibility

Clearly, the study of the uniform invertibility of operators is important to the study of stability. In the following sections, we shall review some uniform invertibility results for the operators that arise from a variety of PDE problems. The connection between uniform invertibility and the Godunov–Ryabenkii theory is also investigated.

2.3.1 Uniform Invertibility for IVPs and PBCPs

Finite difference schemes for one-dimensional scalar problems on regular grids are naturally described in terms of spatial shift operators. We consider the domain $\Omega = [0,1]$ with periodic boundary conditions, and use the ℓ_h^2 norm. Let E_h be the left rotation operator in $L(\ell_h^2)$; its inverse E_h^{-1} is the right rotation operator. Representing these operators as matrices gives a circulant matrix form, which we illustrate for $1/h = J = 6$:

$$E_h \equiv \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad E_h^{-1} \equiv \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}. \tag{2.42}$$

It is easily shown that $\sigma(E_h) = \{\exp(2\pi ij/J) : j = 0, \ldots, J-1\}$, and the spectral mapping theorem [31] gives the spectrum of its inverse.

On the real line, the left shift operation for $w_h \in \ell_h^2(\Re)$ is simply

$$[E_h w_h]_j \equiv [w_h]_{j+1}, \quad j \in Z, \tag{2.43}$$

which is a member of $L(\ell_h^2(\Re))$. Its inverse in $L(\ell_h^2(\Re))$ is the right shift operator. We have $\sigma(E_h) = \{z : |z| = 1\}$, and the spectral mapping theorem can be used to show that $\sigma(E_h^{-1})$ is the same set.

In this notation, the two-level finite difference scheme[3] of equation (2.4) is written

$$A_h u_h^{n+1} = B_h u_h^n + k g_h^n.$$

For IVPs and PBCPs, clearly,

$$A_h = a(E_h, \nu), \quad B_h = b(E_h, \nu).$$

[3] We reserve S_h and T_h for IBVPs. Here $A_h = S_h$ and $T_h = B_h$ are the IBVP or PBCP operators.

The spectral mapping theorem shows that

$$\sigma\left((A_h^{-1})(A_h^{-1})^\star\right) = \left\{|a(z)|^{-2} : z \in \sigma(E_h)\right\},$$

provided all the a_k are real. Consequently (see reference [40]),

$$\|A_h^{-1}\| = \sup_{z\in\sigma(E_h)} \frac{1}{|a(z)|}. \tag{2.44}$$

For the IVP, $E_h \in L(\ell_h^2(\Re))$, and the spectrum of E_h is the unit circle in the complex plane. Consequently, A_h is uniformly invertible iff

$$a(z) \neq 0, \quad \forall\, z : |z| = 1. \tag{2.45}$$

If this condition does not hold, then A_h cannot be invertible in $L(\ell_h^2(\Re))$. Condition (2.45) is required by the VNC.

For the PBCPs, $\sigma(E_h) = \{\exp(2\pi ij/J) : j = 0,\ldots,J-1\}$. Condition (2.45) remains sufficient for uniform invertibility, but its contravention does not necessarily mean that A_h cannot be invertible, as the following example illustrates.

EXAMPLE: PBCP FOR THE BOX SCHEME

The Box scheme for the PBCP with $\nu = 0$ has $a(z,0) = 1+z$. By equation (2.44),

$$\begin{aligned}\|A_h^{-1}\| &= \max_{j=0,\ldots,J-1} \frac{1}{|1+e^{2\pi ij/J}|} \\ &= \begin{cases} \frac{1}{2\sin\pi h} & \text{if } J \text{ is odd} \\ \infty & \text{if } J \text{ is even} \end{cases}.\end{aligned}$$

Thus $\|A_h^{-1}\|$ grows like h^{-1} for odd J, but A_h is singular for even J.

□

2.3.2 Uniform Invertibility for Homogeneous IBVPs

For the domain $\Omega = I = [0,1]$, with homogeneous boundary conditions, E_h is a left shift operator that fills the right of the vector with zeros; E_h^{-1} is a right shift operator that fills the left of the vector with zeros. For example, with $1/h = J = 6$, these operators in $L(\ell_h^2(I))$ have the matrix representations

$$E_h \equiv \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad E_h^{-1} \equiv \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}. \tag{2.46}$$

Notice that this is a definite abuse of notation: E_h and E_h^{-1} are *not* mutual inverses, and this fact must always be borne in mind! As for their spectra, both operators are nilpotent, so $\sigma(E_h) = \sigma(E_h^{-1}) = \{0\}$. However, the spectral mapping theorem can play no part in determining this equality.

Were we to (rather perversely) prescribe homogeneous boundary conditions for the scheme given by equation (2.3), we could once again write $A_h = a(E_h, \nu)$ and $B_h = b(E_h, \nu)$. In these circumstances, we would like to investigate the uniform invertibility of the *Toeplitz matrix* $a(E_h) = a(E_h, \nu)$. This matrix is constant along diagonals, and has the *symbol* $a(z)$.

Strang used Wiener–Hopf techniques to determine the invertibility of Toeplitz operators on the half line from a knowledge of the roots of their symbols, and later extended his theory to determine the uniform invertibility of Toeplitz operators on an interval [142,143]. The crucial points are: no root may lie on the unit circle; for each root inside the unit circle, there must be a corresponding (homogeneous) left-hand boundary condition; and for each root outside the unit circle, there must be a corresponding (homogeneous) right-hand boundary condition. In Chapter 4 we shall see that roots on the unit circle only lead to weak $O(h^{-p})$ growth of the norm of S_h^{-1}, and, consequently, to weak instabilities. However, misplaced boundary conditions lead to catastrophic $O(\rho^J)$, where $\rho > 1$, instabilities.

Strang's result may be rephrased as follows. Suppose that the roots of the polynomial $z^l a(z)$ are, in ascending order of modulus, $z_{-l}, \ldots, z_{-1}, z_1, \ldots, z_r$. Then uniform invertibility in $L(\ell_h^2)$ is equivalent to

(i) $|z_i| < 1$, for $i = -l, \ldots, -1$, and

(ii) $|z_i| > 1$, for $i = 1, \ldots, r$

(and so $|z_i| \neq 1$, for all i). When $z^l a(z)$ has no roots on the unit circle S^1, these conditions are equivalent to $a(z)$ having zero winding number around S^1 [23, page 119]. That is, the path $\{a(z) : |z| = 1\}$ on S^1 can be continuously deformed into the path $\{1\}$ on S^1; thus a is non-vanishing and homotopic to 1 on S^1. Note that the *diagonal dominance condition* [109] is sufficient — but not necessary — for uniform invertibility. Its sufficiency is proved in reference [38, page 35]. The following result summarizes these observations.

Theorem 2.3 *The family of Toeplitz operators $\{A_h\}_{h>0}$ with symbol $a(z)$ that results from the imposition of homogeneous boundary conditions is uniformly invertible in $L(\ell_h^2(I))$ iff $a(z)$ is non-vanishing and homotopic to 1 on S^1.*

Centres

In the previous section it was shown that uniform invertibility for IVPs and PBCPs requires only that the symbol not vanish on the unit circle. Strang's result shows that for IBVPs on $x \in [0, 1]$, the disposition of the roots of the symbol is also crucial. We shall call a_0 the *centre* of the symbol: it is the coefficient that runs down the main diagonal of A_h, and may be changed by multiplying $a(z)$ by a power of z. The limits of the summation in the definition of $a(z)$ are labelled l and r, because we can envisage A_h as having l homogeneous left-hand boundary conditions, and r homogeneous right-hand boundary conditions. The condition that $a(z) \neq 0$ on the unit circle ensures that the boundary conditions are fixed for all frequencies ξ : $z = e^{i\xi}$.

Proper Centres

When the conditions of the theorem are met, we say that the centre a_0 is the (unique) *proper centre* (c.f. reference [63]). Theorem 2.3 may now be phrased as follows: the family $\{A_h\}_{h>0}$ is uniformly invertible iff there are no roots of the symbol on the unit circle, and the chosen centre is the proper centre.

It may seem that the distinction between 'centres' and 'proper centres' is fruitlessly scholastic (c.f. references [65,68,67]). However, Jamieson and Sloan [64] wrongly concluded that Sloan's scheme could not be (GKS) stable for CFL numbers greater than one, precisely because they did not make this distinction. They implicitly chose a centre *in advance*; and this choice was only *proper* for CFL numbers less than one. Had their computer program been able to reposition the boundary conditions (that is, change the scheme's centre) in response to variations of the CFL number, they would have found empirically that Sloan's scheme can work well at all CFL numbers.

EXAMPLE: IBVP FOR THE CN SCHEME

The artificial example of the CN scheme with homogeneous boundary conditions at $x = 0$ and $x = 1$ has an implicit operator A_h with the symbol

$$a(z, \nu) = -z^{-1} + \frac{4}{\nu} + z.$$

This symbol always has one root inside the unit circle and one root outside, irrespective of $\nu > 0$. Thus for all positive ν, the (implicitly) chosen centre, namely the middle stencil point, is the proper centre. Thus $\{A_h\}_{h>0}$ is uniformly invertible.

□

Wiener-Hopf Operators

To clarify the two-dimensional theory presented later, we make some remarks concerning the index of Toeplitz operators on the half line $\Re^+$ (Wiener-Hopf operators). Let $A \in L(X, Y)$, and define

$$\alpha(A) \equiv \dim \ker A, \quad \beta(A) \equiv \operatorname{codim} \mathcal{R}(A).$$

If both $\alpha(A)$ and $\beta(A)$ are finite, A is said to be *Fredholm* with *index*

$$\operatorname{ind}(A) \equiv \alpha(A) - \beta(A).$$

Of course, Fredholm operators satisfy the Fredholm alternative. The following related results are well known [72, page 59].

Theorem 2.4 *Let $A \in L(X, Y)$. Then the following results hold.*

(i) If A is invertible on Y, then A is Fredholm with zero index and $\alpha(A) = \beta(A) = 0$.

(ii) A is Fredholm iff it is almost invertible. *That is, there exists bounded linear operators S_X, S_Y such that*

$$S_X A = 1_X + K_X, \quad A S_Y = 1_Y + K_Y,$$

where $K_X \in \mathcal{K}(X)$ and $K_Y \in \mathcal{K}(Y)$ are compact.

Toeplitz operators on $\Re^+$ are Fredholm iff their symbols do not vanish on the unit circle [28]. When such an operator is Fredholm, its index is minus the winding number of the symbol [28]. By result (i), a zero winding number is necessary for invertibility. In fact, it is also sufficient in one dimension, but not in two. Reference [28, page 1032] gives an example of a scalar Toeplitz operator on the quarter plane which is Fredholm with zero index, but is not invertible.

2.3.3 Uniform Invertibility for General Homogeneous IBVPs

For hyperbolic problems, it is extremely rare to have *homogeneous* boundary conditions at all boundaries, and so we will now extend our results to admit more general homogeneous boundary conditions. With general homogeneous boundary conditions, we shall write our method as

$$S_h u_h^{n+1} = T_h u_h^n + k g_h^n, \tag{2.47}$$

where S_h and T_h are perturbed homogeneous boundary condition operators that arise from the linear boundary conditions at $x = 0, 1$. We shall write

$$S_h = A_h + P_h^1, \quad T_h = B_h + P_h^0. \tag{2.48}$$

The crucial assumption we shall make about the P_h^i $(i = 0, 1)$ is that they have the form

$$P_h^i = \begin{bmatrix} L^i & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & R^i \end{bmatrix},$$

where L^i and R^i are rectangular matrices of fixed, h-independent size.

To present the generalized boundary condition variation of Strang's theorem, some notation must be introduced. Let S_h^r be the operator obtained from S_h by moving the right-hand boundary condition of S_h to $x = \infty$, whilst retaining the boundary condition at $x = 0$. Also, let S_h^l be the operator that results from moving the left-hand boundary condition to $x = -\infty$, whilst retaining the boundary condition at $x = 1$. In this notation we can now state a theorem [101,103] due to Osher.

Theorem 2.5 *The collection $\{S_h\}_{h>0}$, as given by equation (2.48), is uniformly invertible in $L(\ell_h^2([0,1]))$ iff*

(i) a is non-vanishing and homotopic to 1 on S^1,

(ii) $S_h^r u_h^r = 0$ has only trivial solutions $u_h^r \in \ell_h^2([0,\infty))$, and

(iii) $S_h^l u_h^l = 0$ has only trivial solutions $u_h^l \in \ell_h^2((-\infty,1])$.

This result is stated for constant coefficient, one-dimensional systems in [101, Theorem XXV]. It is proved for scalar, variable coefficient problems on an interval in [103, page 19].

EXAMPLE: PLACEMENT OF BOUNDARY CONDITIONS

Consider the Box, CN, AD and SL schemes for the solution of the linear advection equation. When using schemes whose stencil width is greater than one, which includes all the above schemes except the Box scheme, one must provide additional *artificial* or *numerical* boundary conditions to close the system of linear equations, in addition to the genuine boundary data prescribed at $x = 0$. For the CN, AD and SL schemes, it is clear that *one* extra numerical boundary condition is needed. However, it is not clear whether this numerical boundary condition should be placed at $x = 0$ or $x = 1$. Thus, the proper centres of the schemes must be determined. These are found by defining $P(z,\nu) \equiv z^l a(z,\nu)$, and studying the roots of $P(z,\nu) = 0$. When $\nu = 0$, which can occur only if $a = 0$, the (exact) linear advection equation has a characteristic boundary, so no *genuine* boundary condition is required. We

shall not investigate this case. Neither shall we investigate the behaviour of methods for the integer CFL numbers $\nu = 0, \pm 1, \pm 2$, when the stencils of some of our schemes collapse.

To find the proper centre of the Box scheme, note that

$$P(z, \nu) = (1 - \nu) + (1 + \nu)z,$$

so, for all $\nu > 0$, there is only one root inside $C = \{z : |z| = 1\}$: only the genuine boundary condition at $x = 0$ is required. For negative ν, only the genuine boundary condition at $x = 1$ is needed. Thus the Box scheme may be regarded as being *conditionally stable, since a prescribed centre cannot be proper for all CFL numbers.* For the CN scheme,

$$P(z, \nu) = -1 + \frac{4}{\nu}z + z^2,$$

which, for all $\nu \neq 0$, has one root inside S^1, and one root outside S^1. There must always be one boundary condition at each end of the interval, so the numerical boundary condition must always be at the end of the interval opposite the genuine boundary condition. Unlike the Box scheme, the CN scheme has a fixed proper centre, and may possibly be unconditionally stable. The AD scheme has

$$P(z, \nu) = -\nu + (2 + \nu)z,$$

and the situation is the same as that of the CN scheme. More interestingly, the SL scheme has

$$P(z, \nu) = q(\nu) + r(\nu)z + s(\nu)z^2,$$

where

$$q(\nu) = (1 - \nu)(2 - \nu), \quad r(\nu) = 2(4 - \nu^2), \quad s(\nu) = (1 + \nu)(2 + \nu).$$

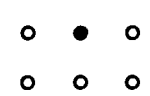
It is easily shown that:

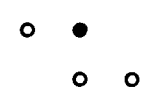
- for $\nu \in (-\infty, -1)$, there are two roots outside S^1, and two right-hand boundary conditions are needed;
- for $\nu \in (-1, 1)$, there is one root inside S^1 and one root outside, and so a left and a right-hand boundary condition are needed;
- for $\nu \in (1, \infty)$, there are two roots inside S^1, and so two left-hand boundary conditions are needed.

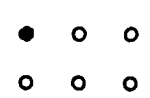
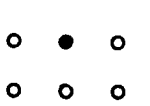
These results completely determine the disposition of boundary conditions for Sloan's scheme.

These proper centring results are summarized in the following diagrams.

Box :  $\nu < 0$  $\nu > 0$

CN :  $\nu \in \Re \setminus \{0\}$

AD :  $\nu \in \Re \setminus \{-2, 0\}$

SL :  $\nu < -1$  $|\nu| < 1$ $\nu > 1$

Note that, if we are to use such diagrams to find the placement of boundary conditions, the stencil points at the old time step must be included. That this is the case is most clear for the AD scheme.

Osher's theorem shows that all the above schemes need one genuine boundary condition at $x = 0$ or $x = 1$, for $\nu > 0$ and $\nu < 0$, respectively. The placement of boundary conditions may be summarized as follows.

- The Box scheme does not require a numerical boundary condition.
- The CN and AD schemes require one boundary condition opposite the genuine boundary condition for all $\nu \neq 0$, and $\nu \neq 0, -2$, respectively.
- The SL scheme for $\nu \in (-1, 1)$ requires a numerical boundary condition opposite the genuine boundary condition; for $|\nu| > 1$, $(\nu \neq \pm 2)$, the scheme requires a numerical boundary condition at the same end as the genuine boundary condition.

□

2.3.4 Uniform Invertibility in Two Dimensions

Implicit finite difference schemes for scalar two-dimensional problems lead naturally to large linear systems that must be solved at each time step. For problems with homogeneous boundary conditions, these systems represent two-dimensional Toeplitz operators. In practice, operator splitting and ADI methods are often used to replace the large two-dimensional system by a series of smaller one-dimensional systems; see, for example, the scheme due to Beam and Warming [5]. [4] For this reason it may seem that there is little practical need for an investigation of the uniform invertibility of two-dimensional Toeplitz operators.

However, information concerning the invertibility of two-dimensional Toeplitz operators will be needed in Chapter 4 — even for an analysis of explicit methods — although not for the stability reasons that we have discussed so far.

On the quarter plane, a two-dimensional Toeplitz equation in w takes the form

$$\sum_{(j,m)\in S} t_{j,m} w_{i+j,n+m} = f_{i,n}, \quad i, n \geq 0, \tag{2.49}$$

where S is a (possibly infinite, but, for finite difference schemes, usually finite) collection of ordered pairs of integers. The boundary conditions are assumed to be homogeneous, and we write this equation as $Tw = f$. The symbol of the operator T is defined as

$$t(z_1, z_2) \equiv \sum_{(j,m)\in S} t_{j,m} z_1^j z_2^m. \tag{2.50}$$

Alternatively, the $\{t_{j,m}\}$ may be viewed as the Fourier coefficients of a given function t. Ideally, we would like conditions that ensure the invertibility of T in $L(\ell^2(\Omega))$, where Ω is the quarter plane. (Since we are dealing with the quarter plane, the parameter h is just a scaling factor, and will be omitted from now on.)

The following result, due to Simonenko (see, for example, reference [30]), allows general, non-polynomial symbols. We use the notation T^2 for the unit torus.

[4] As mentioned in the introduction, this may lead to problems, particularly when the computational grid is not aligned with the boundaries of the domain [139,159,150].

Theorem 2.6 *Suppose that $t \in C(T^2)$ has Fourier coefficients $\{t_{j,m}\}$. Then the operator T defined by equation (2.49) with homogeneous boundary conditions lies in $L(\ell^2(\Omega))$; it is Fredholm iff t is non-vanishing and homotopic to 1 on T^2. If these conditions hold, then T has zero index.*

Note that when t does not vanish on T^2, $t(T^2)$ is homotopic to 1 in $\mathcal{C} \setminus \{0\}$ if and only if

$$W(t(1,.)) = W(t(.,1)) = 0,$$

where the winding numbers are about S^1 [29, page 208].

As we have already mentioned, in two dimensions the Fredholm property and zero index of T do not guarantee its invertibility. However, these two conditions are necessary for invertibility, and are sufficient to define a unique *proper centre* for two-dimensional operators. Thus a necessary condition for invertibility of a quarter-plane Toeplitz operator is that it be properly centred; properly centred operators are Fredholm of zero index, but are not necessarily invertible. Of course, the conditions of the theorem are easily generalized to d dimensions: it is conjectured that the theorem holds generally [28].

Some specialized results are available concerning invertibility: see references [144,92,30, 104] and [28,102,29]. The author does not know of any results concerning the uniform invertibility of Toeplitz operators on finite domains in the plane, whether or not the boundary conditions are homogeneous. However, the author conjectures that the conditions of the above Theorem will be necessary for uniform invertibility.

2.3.5 The Godunov–Ryabenkii Theory

A streamlined account of the Godunov–Ryabenkii theory can be given using the identity

$$(\lambda - G_h)^{-1} = \lambda^{-1} \sum_{m \geq 0} (\lambda^{-1} G_h)^m, \quad \forall |\lambda| > 1,$$

because power-boundedness implies that

$$\|(\lambda - G_h)^{-1}\| \leq C \sum_{m \geq 0} |\lambda|^{-m-1} = \frac{C}{|\lambda| - 1}, \quad \forall |\lambda| > 1.$$

Thus, a necessary condition for power-boundedness is that the collection $\{\lambda - G_h : h > 0\}$ be uniformly invertible for all $|\lambda| > 1$.

Godunov and Ryabenkii call the set of λ for which $\lambda - G_h$ is not uniformly invertible the *family spectrum* of the family $\{G_h : h > 0\}$, denoted $\sigma_f(\{G_h : h > 0\})$. We also write $\rho_f(\{G_h : h > 0\}) \equiv \mathcal{C} \setminus \sigma_f(\{G_h : h > 0\})$. Clearly, a necessary condition for power-boundedness is that the family spectrum be contained in the closed unit disc; this is the usual statement of the Godunov–Ryabenkii theorem [38,135].

In one dimension, the family spectrum of an explicit scheme is easily computed. Write G_h^l for the left quarter-plane operator that is obtained by moving the left hand boundary to $x = -\infty$, G_h^r for the corresponding right quarter-plane operator, and G_h^0 for the corresponding IVP. Osher's Theorem applied to the Toeplitz operator $\lambda - G_h$ shows that

$$\rho_f(\{G_h\}) = \rho_f(\{G_h^l\}) \cap \rho_f(\{G_h^r\}) \cap \rho_f(\{G_h^0\}).$$

Thus, $\sigma_f(\{G_h\})$ is the union of the family spectra of the corresponding left and right quarter plane operators, and the corresponding IVP operator.

This approach is easily modified to include implicit schemes, with uniformly invertible implicit operators S_h. Osher's Theorem is applied to the family $\{\lambda S_h - T_h\}_{h>0}$.

Recall that a necessary condition for the (L–R or EL–R) stability of an implicit scheme is that the family of implicit operators be uniformly invertible, which, it is conjectured, implies that the symbol of the family be non-vanishing and homotopic to 1 on T^d. However, if the family $\{S_h\}_{h>0}$ is uniformly invertible, then the Godunov–Ryabenkii condition implies that $\lambda S_h - T_h$ must be uniformly invertible for all $|\lambda| > 1$. Again, the symbol $\Phi_\lambda \equiv \lambda a - b$ must be non-vanishing and homotopic to 1 on T^d for any $|\lambda| > 1$. (Note that schemes that satisfy the VNC cannot vanish on T^d.) However, the following result shows that this requirement is automatically satisfied. Consequently, only boundary modes must be investigated. The proof of this assertion for $d = 1$ may be found in reference [156], and is easily generalized to $d > 1$.

Proposition 2.3 *Suppose that:*

(i) $a(z_1, \ldots, z_d)$ *is non-vanishing and homotopic to* 1 *on* T^d, *and*

(ii) for each $|\lambda| > 1$, Φ_λ *does not vanish on* T^d.

Then for each $|\lambda| > 1$, $\Phi_\lambda(T^d)$ *is (non-vanishing and) homotopic to* 1 *on* T^d.

Finally, note that for one-dimensional scalar problems on *bounded domains*, the Kreiss matrix theorem [135,99] shows that a scheme that satisfies the resolvent condition can, at worst, have a 2-weak instability in the EL–R sense. This is a much more refined estimate of how 'near' the GR condition is to sufficiency than the result given by Godunov and Ryabenkii [38, page 439].

2.4 Conclusions

From the results surveyed in this chapter, it is clear that the L–R, EL–R and GKS stability theories are of little help in determining the convergence of methods for the computation of equilibrium solutions. Convergence in a time-uniform norm $\|.\|_{X_h}$ is required, and the space of forcing functions Y_h must admit non-decaying forcing.

However, Sanz-Serna [126, Theorem 3.1] shows that well-posedness is necessary for the stability of a consistent scheme: we cannot alter the spaces X_h and Y_h without also altering the spaces X and Y. Furthermore, we are interested in the time decay of the numerical solution, so it makes sense to first study the rate of decay in time of the exact problem. For these reasons, the next chapter is a study of the decay and well-posedness (in suitable norms) of the exact evolutionary problem.

Chapter 3

Solutions of Forced Evolution Equations

3.1 Introduction

This chapter discusses the existence of solutions of evolutionary equations, and gives rigorous definition to the term 'equilibrium solution'. This definition should isolate certain characteristics. If we encounter a forced system — for example, a tidally forced estuary — we say it is in equilibrium in the following circumstances:

(i) The forcing has been going on for all times in the past. The set of possible states the system might have assumed in the past is bounded.

(ii) The present state is independent of the distant history of the system.

(iii) The system's behaviour will remain bounded in the future, provided that the forcing remains bounded.

(iv) If the system is perturbed, it will eventually return to equilibrium.

The equilibrium solution should be a solution of the evolution equation alone, not of any particular IVP for the evolution equation. By property (iv), it is clear that IVPs may be used to approximate a specific equilibrium solution; but property (ii) implies that the definition of an equilibrium solution should be independent of any initial data.

Note that equilibrium solutions share some features of (stationary) boundary value problems. However, equilibrium solutions are, in general, non-stationary. These considerations justify the view that equilibrium solutions of hyperbolic PDEs represent solutions whose behaviour lies between that of truly steady BVPs and truly wave-like IBVPs.

Once a mathematical definition of equilibrium solutions has been settled, we would like to identify a class of evolution equations for which equilibrium solutions exist uniquely. The existence of solutions of evolutionary equations is discussed using semigroup theory in Section 3.2. For linear equations, this allows an explicit representation of equilibrium solutions.

In the linear case, an equilibrium solution $u_{eq}(t; f)$ depends solely on the boundary forcing $f(t)$. For example, the tides in a small estuary depend only on the ocean forcing at the mouth of the estuary. This forcing data is necessarily of limited accuracy, and were $u_{eq}(t; f)$ to be sensitive to changes in f, we could have little confidence that $u_{eq}(t; f)$ would

accurately represent the physical system. Thus, we require the map $f \mapsto u(.;f)$ to be continuous. This curious form of well-posedness is discussed in Section 3.6, and will be termed *BVP stability.*

In practice, very few equilibrium solutions can be computed directly from the forcing. More often, equilibrium solutions can only be found via evolutionary initialization. This exploits the asymptotic stability property implied by condition (ii):

$$\lim_{t\longrightarrow\infty} \|u(t;t^0,u^0,f) - u_{eq}(t;f)\| = 0.$$

Here $u(t;t^0,u^0,f)$ is the solution of the IBVP that results from imposing the arbitrary initial condition $u(t^0) = u^0$ on the evolutionary equation. The *IVP stability* of solutions is discussed in Section 3.3.

3.2 Semigroups and Processes

In this section we recall some important facts about semigroups and processes, and establish notation that will be used frequently in the chapters to come. The material in this section is from references [130,7,31,50,60,2,155,39].

Let B be a Banach space and $\Re^+ \equiv [0,\infty)$. A family of (possibly non-linear) mappings $G(t) : B \rightarrow B$, where $t \in \Re^+$, is a C^0-*semigroup* iff

(i) $G(0) = I$,

(ii) $G(t+s) = G(t)G(s)$ for all $t, s \in \Re^+$, and

(iii) $G(t)x$ is strongly continuous in $t \in \Re^+$ and $x \in B$.

When, additionally, $G(t)$ is a bounded linear operator on B for each $t \in \Re^+$ — that is, $G(t) \in L(B)$ for each $t \in \Re^+$ — G is said to be a *linear* C^0-*semigroup*. Note that only strong continuity is required in G at $t = 0$. In fact, uniform continuity at zero implies that G is an analytic semigroup.

3.2.1 Properties of Linear C^0-Semigroups

For any linear C^0-semigroup, it can be shown that there exist constants $M \geq 0$ and $\beta \in \Re$ such that

$$\|G(t)\| \leq Me^{\beta t}, \quad \forall\, t \in \Re^+. \tag{3.1}$$

In fact [73, page 28] [2, page 334], β may be taken as

$$\beta_G \equiv \lim_{t\longrightarrow\infty} \frac{1}{t} \log \|G(t)\| < \infty. \tag{3.2}$$

β_G is said to be the *type* of the semigroup. When $\beta_G = 0$, the semigroup is said to be *equibounded*. Setting $t = nt_0$, for any $t_0 > 0$,

$$\beta_G = \frac{1}{t_0} \log \left(\lim_{n\longrightarrow\infty} \|G(t_0)^n\|^{1/n} \right) = \frac{1}{t_0} \log \left(r(G(t_0)) \right).$$

Thus $r(G(t_0)) = e^{\beta_G t_0}$. G has negative type iff $r(G(t_0)) < 1$ for all $t_0 > 0$, and G is equibounded iff $r(G(t_0)) \leq 1$ for all $t_0 \geq 0$.

For each $\eta > 0$, let $A_\eta \equiv \frac{1}{\eta}[G(\eta) - I]$, an element of $L(B)$. Define

$$A_0 x \equiv \lim_{\eta \to 0^+} A_\eta x$$

for those $x \in B$ for which the limit exists, denoting these $\mathcal{D}(A_0)$. Now assume that A_0 is *closable*: that is,

$$[\{x_n\} \subseteq \mathcal{D}(A_0),\ x_n \to 0,\ A_0 x_n \to y] \implies y = 0.$$

Then it can be shown [129, page 6], [77, page 537] that A_0 has a *smallest closed extension* $\tilde{A}_0$, in the sense that all other closed extensions of A_0 are extensions of $\tilde{A}_0$. When A_0 is closable in this sense, $\tilde{A}_0$ is said to be the *infinitesimal generator* of the linear C^0-semigroup G.

These observations raise the question of when a *given* closed operator A is the infinitesimal generator of *some* linear C^0 semigroup. The answer is provided by the Hille–Yosida–Phillips Theorem [60, page 360] [73, page 51].

Theorem 3.1 *The following conditions are necessary and sufficient for a closed linear operator $A : \mathcal{D}(A) \subseteq B \to B$ to be the infinitesimal generator of a linear C^0-semigroup:*

(i) $\mathcal{D}(A)$ is dense in B, and

(ii) there exists $M > 0$, $\beta \in \Re$ such that $\forall\ \lambda > \beta, n = 1, 2, \ldots$, $(\lambda - A)^{-n} \in L(B)$, and

$$\|(\lambda - A)^{-n}\| \leq M(\lambda - \beta)^{-n}. \tag{3.3}$$

When these conditions hold, the semigroup G satisfies $\|G(t)\| \leq Me^{\beta t}$.

This result suggests the following definition. A closed linear operator $A : \mathcal{D}(A) \subseteq B \to B$, where B is Banach, is said to be of class $\mathcal{G}(M, \beta; B)$ iff A satisfies conditions (i) and (ii) of Theorem 3.1. When $A \in \mathcal{G}(M, \beta; B)$, A is the infinitesimal generator of a linear C^0-semigroup G, which satisfies $\|G(t)\| \leq Me^{\beta t}$. However, the type of the semigroup β_G may be strictly less than β. In particular, $A \in \mathcal{G}(M, \beta; B)$ with $\beta < 0$ is sufficient, but not necessary, for G to be of negative type.

3.2.2 Autonomous Linear Evolutionary Problems

Consider the evolutionary problem

$$\dot{u} = Au + f(t), \quad u(t^0) = u^0. \tag{3.4}$$

Here $A : \mathcal{D}(A) \subseteq B \to B$ is assumed to be linear and independent of t. Any boundary conditions are enforced implicitly by the definition of the domain $\mathcal{D}(A)$. $f : \Re^+ \to B$ is some forcing function.

A strongly differentiable function u, with strongly continuous derivative, that satisfies (3.4) and $u(t) \in \mathcal{D}(A)$ for all $t \in \Re^+$, is said to be a *strict solution* of (3.4). The following result from references [7, page 167],[39, page 84] shows that strict solutions of problem (3.4) are given by the *variation of parameters* formula

$$u(t) = G(t)u^0 + \int_0^t G(t - s)f(s)\, ds. \tag{3.5}$$

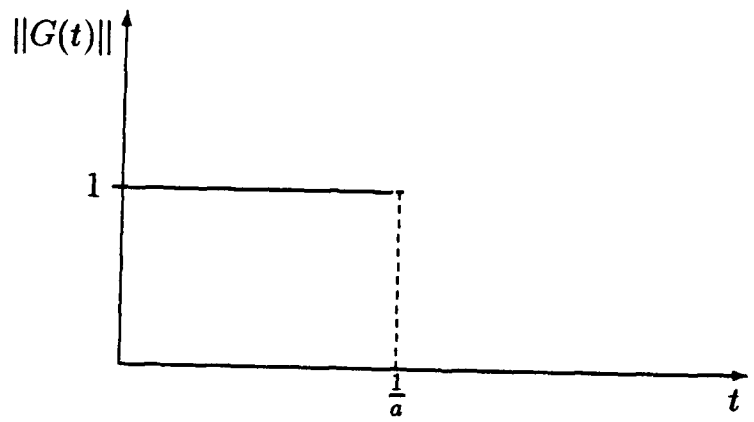


Figure 3.1: The decay of $\|G(t)\|$ for the one-dimensional linear advection equation on $x \in [0,1]$.

Because $\mathcal{D}(A)$ contains C_0^∞ (the space of infinitely differentiable functions with compact support in I) which is dense in L^2 [7, page 141], $\mathcal{D}(A)$ is dense in B. Furthermore, a modification of the argument of reference [7, pages 141, 84–87] shows that $A \in \mathcal{G}(1,0;B)$. The Hille–Yosida–Phillips Theorem shows that A is the infinitesimal generator of an equibounded linear C^0-semigroup, which, for $\phi \in L^2$, is given by

$$[G(t)\phi](x) = \begin{cases} 0 & x \in [0, at] \\ \phi(x - at) & x \in (at, 1] \end{cases}, \quad at \le 1, \quad G(t) = 0, \ at > 1. \tag{3.7}$$

It is clear that $\|G(t)\phi\|_2$ is continuous, as it must be, since G is a C^0 semigroup. However, $\|G(t)\|_{L(B)}$ is not continuous (see Figure 3.1), so the semigroup is not uniformly continuous.

EXAMPLE: TELEGRAPHER'S EQUATIONS

Recall from Chapter 1 that the Telegrapher's equations are

$$\begin{bmatrix} i \\ v \end{bmatrix}_t = \begin{bmatrix} 0 & -1/L \\ -1/C & 0 \end{bmatrix} \begin{bmatrix} i \\ v \end{bmatrix}_x + \begin{bmatrix} -R/L & 0 \\ 0 & -G/C \end{bmatrix} \begin{bmatrix} i \\ v \end{bmatrix},$$

for $x \in I \equiv [0,1]$. The boundary conditions are

$$\begin{aligned} v(0,t) + R_0 i(0,t) &= E(t), && \text{at } x = 0, \\ R_1 v(1,t) + i(1,t) &= C_1 \tfrac{d}{dt} v(1,t), && \text{at } x = 1, \end{aligned}$$

where $E(t)$ is typically (but not necessarily) T-periodic, and the other coefficients are constants.

The boundary conditions at $x = 0$ can be rendered homogeneous by setting $\hat{i} = i$ and $\hat{v} = v - E$. This yields

$$\begin{bmatrix} \hat{i} \\ \hat{v} \end{bmatrix}_t = \begin{bmatrix} 0 & -1/L \\ -1/C & 0 \end{bmatrix} \begin{bmatrix} \hat{i} \\ \hat{v} \end{bmatrix}_x + \begin{bmatrix} -R/L & 0 \\ 0 & -G/C \end{bmatrix} \begin{bmatrix} \hat{i} \\ \hat{v} \end{bmatrix} + \begin{bmatrix} 0 \\ -GE/C - \dot{E} \end{bmatrix},$$

Theorem 3.2 *Suppose that*

(i) $A \in \mathcal{G}(M, \beta; B)$ *is the infinitesimal generator of* G,

(ii) $u^0 \in \mathcal{D}(A)$, *and*

(iii) f *is strongly continuous for each* $t \in \Re^+$.

Then each strict solution of problem (3.4) satisfies the equality (3.5). If, furthermore,

(iv) f *is strongly differentiable with strongly continuous derivative for each* $t \in \Re^+$, *or*

(v) f *is strongly continuous and takes values in* $\mathcal{D}(A)$,

then the function defined by (3.5) is the unique, strict solution of problem (3.4).

Note that conditions (i)–(iii) alone do not guarantee that u, defined by equation (3.5), is a strict solution of problem (3.4). G may be viewed as the solution operator of the unforced ($f(t) = 0$) problem corresponding to (3.4). Also, since $\mathcal{D}(A)$ is assumed to be dense in B, the variation of parameters formula can be used to define a unique solution of (3.4) for *any* $u^0 \in B$. When $u(t)$ is the solution that results from $u^0 \in B - \mathcal{D}(A)$, it is known as the *weakened solution* of (3.4). Questions of well posedness, and the existence of generalized solutions (in the sense of Chapter 2), are deferred until suitable time norms for $u(t)$ have been investigated: see Section 3.4.

Equation (3.5) can be used to define a *weakened solution operator* $S(r, t; f)$ as

$$u(t) = S(t - t^0, t^0)u^0 \equiv G(t - t^0)u^0 + \int_{t^0}^{t} G(t - s)f(s)\, ds, \tag{3.6}$$

which maps $u(t + r) = S(r, t)u(t)$. Observe that $r \in \Re^+$ is the time *increment*. It can be shown that the following equalities hold:

(i) $S(0, s) = I$ for all $s \in \Re^+$,

(ii) $S(t, r + s)S(r, s) = S(t + r, s)$ for all $r, t \in \Re^+$ and $s \in \Re$.

These are the defining characteristics of a *process* (see Section 3.2.4). Processes will be important when more general problems are considered.

When the forcing f is T-periodic, then the solution operator S is T-periodic in the sense $S(r, s) = S(r, s + T)$ for any $r \in \Re^+$ and $s \in \Re$. The *map at a period* is defined to be $S_T \equiv S(T, 0)$.

Example: One-Dimensional Linear Advection on a Semi-Infinite Strip

The one-dimensional scalar linear advection equation can be written in the form

$$\dot{u} = Au + f(t), \quad u(0) = u^0.$$

Here $A = -a\partial_x$ is a scaled generalized derivative, and $a > 0$ is real. Homogeneous boundary conditions are imposed at $x = 0$, and the problem is posed on the domain $I \equiv [0, 1]$. Thus $\mathcal{D}(A) \subseteq B$ is chosen to be

$$\mathcal{D}(A) = \{w \in H^1 : w(0) = 0\}, \quad B = L^2.$$

Of course, $H^1 \equiv H^1(I)$, and $L^2 \equiv L^2(I)$.

with the boundary conditions

$$\begin{aligned} \hat{v}(0,t) + R_0\hat{i}(0,t) = &\quad 0, &&\text{at } x = 0, \\ R_1\hat{v}(1,t) + \hat{i}(1,t) = &\quad C_1\left(\tfrac{d}{dt}\hat{v}(1,t) + \dot{E}\right) - R_1E, &&\text{at } x = 1. \end{aligned}$$

The matrix associated with the spatial derivative can be diagonalized. One obtains

$$u_t = \tilde{A}u_x + \tilde{B}u + f_0, \tag{3.8}$$

where $u = [U, V]$, U and V are the transformed variables, and f_0 is a forcing term. The matrices are given by

$$\tilde{A} = \begin{bmatrix} -c & 0 \\ 0 & c \end{bmatrix}, \quad \tilde{B} = \begin{bmatrix} -a & b \\ b & -a \end{bmatrix},$$

where $c \equiv 1/\sqrt{LC}$ is the characteristic speed, and

$$a \equiv (RC + LG)/2LC, \quad b \equiv (RC - LG)/2LC.$$

The boundary conditions are now

$$U(0,t) + q_0V(0,t) = 0, \quad \text{at } x = 0, \tag{3.9}$$

$$r_1U(1,t) + q_1V(1,t) = \tfrac{d}{dt}\left[U(1,t) + V(1,t)\right] + h_1(t), \quad \text{at } x = 1, \tag{3.10}$$

for some q_1, r_1 and function h_1. The important quantity q_0 is

$$q_0 \equiv \frac{z_0 - R_0}{z_0 + R_0}, \quad z_0 \equiv \sqrt{\frac{L}{C}}. \tag{3.11}$$

Following references [50,100], we can write equation (3.8) in the standard form

$$\dot{w} = Aw + f(t)$$

as follows. For each $w \equiv [u, d] = [U, V, d] \in \mathcal{D}(A)$, set

$$Aw \equiv \left[\tilde{A}u_x + \tilde{B}u,\ r_1U(1) + q_1V(1)\right].$$

To define the domain of the operator A, first let

$$B \equiv L^2 \times L^2 \times \Re,$$

with the norm $\|(\phi, \psi, b)\|_B^2 \equiv \|\phi\|_2^2 + \|\psi\|_2^2 + b^2$. The boundary conditions are enforced by choosing $\mathcal{D}(A) \subseteq B$ as

$$\mathcal{D}(A) \equiv \left\{[u, d] \in B : u \in (H^1)^2, U(0) + q_0V(0) = 0, U(1) + V(1) = d\right\},$$

where $H^1 = H^1(I)$. Finally, define

$$f(t) \equiv \begin{bmatrix} f_0(t) \\ -h_1(t) \end{bmatrix}$$

to complete the specification of the Telegrapher's equations in standard form.

It can be shown that $A \in \mathcal{G}(M, 0; B)$ [100], and is thus the infinitesimal generator of a linear C^0-semigroup. For smooth forcing, the solution of the forced IBVP is given by the variation of parameters formula.

□

EXAMPLE: KLEIN–GORDON EQUATIONS

The Klein–Gordon equations were introduced in Chapter 1, and the procedure for casting them into standard form is exactly the same as for the Telegrapher's equations. The crucial difference is that, after diagonalization, the equations take the form

$$u_t = \tilde{A}u_x + \tilde{B}u + f_0,$$

but with

$$\tilde{A} = \begin{bmatrix} -c & 0 \\ 0 & c \end{bmatrix}, \quad \tilde{B} = \begin{bmatrix} 0 & \beta \\ -\beta & 0 \end{bmatrix}. \tag{3.12}$$

We shall assume that the boundary conditions can be cast in the same form as those for the Telegrapher's equations: that is, in the form given by equations (3.9) and (3.10).

□

3.2.3 Non-Autonomous Linear Problems

Next consider the time-dependent problem

$$\dot{u} = A(t)u + f(t), \quad u(t^0) = u^0. \tag{3.13}$$

Under suitable hypotheses [73, page 195, Theorem 3.1], the unforced problem ($f = 0$) is solved by the *evolution operator*

$$u(t) = G(t, t^0)u^0. \tag{3.14}$$

The evolution operator satisfies the equalities in $L(B)$:

(i) $G(t,t) = I$, and

(ii) $G(t,r)G(r,s) = G(t,s)$,

which are essentially the same as the corresponding equalities for the solution operator (equation (3.6)), except that the arguments of G are (final time, initial time), whilst those of S are (time increment, initial time).

The solution of the forced problem (3.13) is given by the variation of parameters formula

$$u(t) = G(t, t^0)u^0 + \int_{t^0}^{t} G(t,s)f(s)\,ds, \tag{3.15}$$

which defines the solution operator S via $u(t) = S(t - t^0, t^0; f)u^0$. Again, S satisfies the conditions for a process, mentioned earlier.

EXAMPLE: ONE-DIMENSIONAL VARIABLE COEFFICIENT HPDES

The following shows how solutions of a variable coefficient HPDE can be obtained. The presentation relies heavily on reference [100]. Consider the unforced hyperbolic PDE

$$u_t = \tilde{A}u_x + \tilde{B}u, \tag{3.16}$$

posed on the domain $I = [0,1]$, $t \in \Re^+$. Here $u = [u^I, u^{II}]$, where $u^I(x,t) \in \Re^\ell$ and $u^{II}(x,t) \in \Re^{m-\ell}$. This represents the usual split into components with left and right going phase velocities. The boundary conditions are

$$u^I = S_0 u^{II}, \quad \text{at } x = 0, \tag{3.17}$$

$$\frac{d}{dt}\left[u^{II} - T_1 u^I\right] = S_1^1 u^I + S_1^2 u^{II}, \quad \text{at } x = 1. \tag{3.18}$$

Of course, $\tilde{A}$ is an $m \times m$ matrix of the form

$$\tilde{A} = \begin{bmatrix} \tilde{A}^I & 0 \\ 0 & \tilde{A}^{II} \end{bmatrix}, \quad \tilde{A}^I < 0, \quad \tilde{A}^{II} > 0,$$

where the matrices $\tilde{A}^I \in L(\Re^\ell)$ and $\tilde{A}^{II} \in L(\Re^{m-\ell})$ are assumed to be diagonal. Thus write

$$\tilde{A}^I = \operatorname{diag}\left[\lambda_1(x,t), \ldots, \lambda_\ell(x,t)\right], \quad \lambda_i < 0,$$
$$\tilde{A}^{II} = \operatorname{diag}\left[\lambda_{\ell+1}(x,t), \ldots, \lambda_m(x,t)\right], \quad \lambda_i > 0.$$

Following reference [100], assume that:

(i) $\tilde{A}(x,t)$ is C^1 for $(x,t) \in I \times \Re^+$,

(ii) $\tilde{B}(x,t)$ is C^0 for $(x,t) \in I \times \Re^+$, and C^1 for suitable x *or* t,

(iii) $S_0(t)$, $S_1^1(t)$, $S_1^2(t)$ and $T_1(t)$ are members of $C^1(\Re^+)$,

(iv) if, somewhere in $(x,t) \in I \times \Re^+$, either

$$\lambda_i(x,t) = \lambda_j(x,t), \quad 1 \le i,j \le \ell, \text{ or}$$
$$\lambda_i(x,t) = \lambda_j(x,t), \quad \ell+1 \le i,j \le m,$$

for $i \neq j$, then $\tilde{B}_{ij}(x,t) = 0$ for every $(x,t) \in I \times \Re^+$.

Associated with this full problem is the following *reduced problem*:

$$u_t = \tilde{A}u_x + \tilde{B}_0 u, \tag{3.19}$$

$$u^I = S_0 u^{II}, \quad \text{at } x = 0, \tag{3.20}$$

$$u^{II} = T_1 u^I, \quad \text{at } x = 1, \tag{3.21}$$

where $\tilde{B}_0 \equiv \operatorname{diag}\left[\tilde{B}_{11}, \ldots, \tilde{B}_{mm}\right]$.

The standard form of the unforced, full evolution problem is $\dot{u} = A_1(t)u$, whilst the standard form of the unforced, reduced evolution problem is $\dot{u} = A_2(t)u$. To define the operators $A_1(t)$ and $A_2(t)$ properly, let

$$B = \left(L^2\right)^m \times \Re^{m-\ell},$$

and let

$$A_1(t)[u,d] = \left[\tilde{A}u_x + \tilde{B}u,\ S_1^1 u^I(1) + S_1^2(t)u^{II}(1)\right],$$

with domain

$$\mathcal{D}(A_1(t)) = \left\{[u,d] \in B : u \in \left(H^1\right)^m,\ u^I(0) = S_0 u^{II}(0),\ u^{II}(1) - T_1 u^I(1) = d\right\}.$$

The definitions of A_2 and its domain are completely analogous.

The standard theory of Krein [73] can then be used to show that $A_1(t)$ generates $G_1(t,s)$ on B, and that $A_2(t)$ generates $G_2(t,s)$ on B_0, where $B_0 = \{[u,d] \in B : d = 0\}$.

□

3.2.4 More General Problems: Processes

Let B be a Banach space, and let $w : \Re^+ \times \Re \times B \to B$ be a map. Define

$$S(r,s) : B \to B,\ S(r,s)x \equiv w(r,s,x), \quad \forall\, r \in \Re^+,\ s \in \Re.$$

According to references [50, page 41] and [155, page 86], w is a *process* iff

(i) for each $s \in \Re$ and $x \in B$, $w(.,s,x) : \Re^+ \to B$ is continuous at all $r > 0$, and right continuous at $r = 0$,

(ii) for each $r \in \Re^+$, $w(r,.,.) : \Re \times B \to B$ is continuous,

(iii) $S(0,s) = I$ for all $s \in \Re$, and

(iv) $S(t, r+s)S(r,s) = S(t+r,s)$ for all $t, r \in \Re^+$ and $s \in \Re$.

The idea is that $w(r,s,x)$ represents the state of a system at time $r+s$, given that it was in the state x at time s; r is the time *increment*. Processes represent systems that satisfy Hadamard's principle of scientific determinism [60, page 618].

The solution of a general evolutionary IVP with the initial data $u(t^0) = u^0$, denoted $u(t;t^0,u^0)$, *may* define a process via $w(t-t^0,t^0,u^0) \equiv u(t;t^0,u^0)$ for $t - t^0 \geq 0$. When w is a process, $u(t;t^0,u^0) = S(t-t^0,t^0)u^0$ and $S(t,s)$ generalizes the notion of a solution operator.

Dynamical Systems

When a process $w(r,s,x)$ is independent of $s \in \Re$, the process is called a ***dynamical system*** [155, page 13]. Dynamical systems arise from autonomous problems $\dot{u} = f(u)$, where f may be non-linear and involve spatial derivatives, but does not depend on time.

Periodic Processes

A process is said to be *T-periodic* if $S(r,s+T) = S(r,s)$ for all $r \in \Re^+$ and $s \in \Re$. The *map at a period* S_T of a T-periodic process is defined as $S_T \equiv S(T,0)$. Note that for a T-periodic process,

$$S(nT,0) = S((n-1)T,T)S(T,0) = S((n-1)T,0)S(T,0) = \cdots = S(T,0)^n.$$

Thus we have

$$\begin{aligned} S(nT+\hat{t},0) = &\ S(\hat{t},nT)S(nT,0), \\ = &\ S(\hat{t},0)S(nT,0), \quad \text{by periodicity}, \\ = &\ S(\hat{t},0)S(T,0)^n, \quad \text{by above}, \end{aligned}$$

Thus

$$S(nT+\hat{t},0) = S(\hat{t},0)S_T^n, \tag{3.22}$$

a property that will be important when asymptotic stability is examined.

3.3 IVP Stability of Solutions

The requirement that solutions be asymptotically independent of the initial data is tantamount to requiring solutions to be asymptotically stable. Whilst various types of stability can be defined for solutions, it is convenient to recognize that solutions are merely special invariant sets and so, instead, to define stability in terms of invariant sets.

Let G be a (not necessarily linear) C^0-semigroup on a Banach space B. Recall that $S \subseteq B$ is *invariant* if $G(t)S = S$ for all $t \in \Re^+$. An invariant set S is said to be [50]:

Stable if, for any neighbourhood $V \supseteq S$, there exists a neighbourhood $U \supseteq S$ such that

$$G(t)U \subseteq V, \quad \forall\, t \in \Re^+;$$

Asymptotically Stable (or a.s.) if there exists a neighbourhood $U \supseteq S$ such that[1]

$$\forall x \in U \quad \text{dist}(G(t)x, S) \to 0, \quad t \to \infty;$$

Uniformly a.s. (or u.a.s.) if there exists a neighbourhood $U \supseteq S$ such that

$$\text{dist}(G(t)U, S) \to 0, \quad t \to \infty.$$

When U can be taken to be any bounded subset of B, an a.s. (or u.a.s.) invariant set is said to be *globally a.s.* (respectively *globally u.a.s.*).

Since these definitions are given in terms of a one parameter semigroup, there is an implicit assumption that the underlying evolutionary problem is autonomous. Thus no distinction need be made between *uniform* and *Lyapunov* stability [20,71]. Both are equivalent to stability in the sense defined above.

For autonomous linear ODEs, a.s. and u.a.s. are equivalent. However, this is not generally the case: an example is given in the next section. It is helpful to identify a class of C^0-semigroups for which a.s. and u.a.s. are equivalent. According to Hale [50,49], a C^0-semigroup is *asymptotically smooth* if, for any compact $S \subseteq B$, for which $G(t)S \subseteq S$, there exists compact K such that $\text{dist}(G(t)S, K) \to 0$ as $t \to \infty$. The following result is proved in reference [50, page 38].

Theorem 3.3 *Suppose that G is an asymptotically smooth C^0-semigroup. Then a compact invariant set K is a.s. iff it is u.a.s.*

3.3.1 Autonomous Linear Problems

When G is a *linear* C^0-semigroup, the previous stability definitions can be simplified. In fact, an invariant set is a.s. iff it is globally a.s., and u.a.s. iff it is globally u.a.s. Furthermore, the following results hold:

(i) *Any* invariant set S is stable iff G is equibounded (i.e. of zero type).

(ii) *Any* invariant set S is a.s. iff $\|G(t)x\| \to 0$ as $t \to \infty$, for each $x \in S$.

(iii) *Any* invariant set S is u.a.s. iff G is of negative type.

[1] $\text{dist}(R, S) \equiv \sup_{x \in R} \inf_{y \in S} \|x - y\|_B$.

Since each type of stability holds for all invariant sets S, the *semigroup* may be said to be stable, a.s. or u.a.s. Note that a semigroup may be a.s. without being u.a.s., as the following example demonstrates.

EXAMPLE: A.S. AND U.A.S. ARE NOT EQUIVALENT

Consider the linear advection equation on $\Re$, with the solution measured over $x \in \Re^- = (-\infty, 0]$, and choose $B = L^2(\Re^-)$. G is the right shift operator

$$[G(t)\phi](x) = \phi(x - at), \quad x \in \Re^-,\ t \in \Re^+.$$

Clearly, $\|G(t)\|_{L(B)} = 1$. However, for $\phi \in L^2(\Re^-)$,

$$\begin{aligned} \|G(t)\phi\|_B &= \left\{ \int_{\Re^-} (\phi(x - at))^2 \, dx \right\}^{1/2} \\ &= \left\{ \int_{(-\infty, -at]} (\phi(y))^2 \, dy \right\}^{1/2} \\ &\to 0, \text{ as } t \to \infty, \end{aligned}$$

when $a > 0$. Thus the semigroup is a.s., but not u.a.s.

□

Finite dimensional, linear ODEs have the following helpful properties. First, a.s. and u.a.s. are equivalent. We have seen that this property can be ensured for infinite-dimensional problems by insisting that the semigroup be asymptotically smooth. Secondly, a.s./u.a.s. holds iff $\mathrm{Re}\,\sigma(A) = \mathrm{Re}\,\mathrm{P}\sigma(A) < 0$.

This latter result holds because $\sigma(A) = \mathrm{P}\sigma(A)$ in finite dimensions, and for *any* linear C^0-semigroup G with infinitesimal generator A [7, page 273],

$$e^{t\mathrm{P}\sigma(A)} \subseteq \mathrm{P}\sigma(G(t)) \subseteq e^{t\mathrm{P}\sigma(A)} \cup \{0\}, \quad \forall\, t > 0. \tag{3.23}$$

Thus, in finite dimensions, u.a.s. is equivalent to $\exists t_0 > 0 :\ r(G(t_0)) < 1$, which in turn is equivalent to $\mathrm{Re}\,\mathrm{P}\sigma(A) < 0$: an elementary result from the stability theory of linear ODEs [71].

However, in infinite dimensions, the spectrum of a semigroup need not consist solely of eigenvalues, and there are only partial results corresponding to (3.23) for the continuous and residual spectra [7, page 274]. In fact (see reference [100]), there exists a group $\{G(t) : t \in \Re\}$ whose generator A has $\sigma(A) = \emptyset$[2], but $\mathrm{C}\sigma(G(t)) = \{\lambda : |\lambda| = \exp(\pi t)\}$, so $r(G(t)) \geq e^{\pi t}$. Futhermore, $\|G(t)\|$ grows exponentially. In such a case u.a.s. cannot be decided from the (empty!) spectrum of A. We would like to identify a class of infinite-dimensional problems whose stability properties can be determined from the spectra of their generators. The Hille–Yosida–Phillips Theorem cannot help with this identification, since it depends on the behaviour of the norm of the resolvent of the generator.

Essential Spectrum

Let $R \in L(B)$. Recall that $\lambda \in \sigma(R)$ is isolated iff there is a non-zero projection P, commuting with R, such that $(\lambda - R)P$ is quasi-nilpotent and $\alpha P + (\lambda - R)$ is invertible for some $\alpha \neq 0$. We say:

[2]Since A is unbounded in this case, its spectrum may be empty.

(i) λ is a *normal point* of R iff $\lambda \in \rho(R)$ or λ is an isolated spectral point of R, with $\dim \mathcal{R}(P) < \infty$.

(ii) The *essential spectrum* of R, denoted $\sigma_e(R)$, consists of all complex numbers that are not normal points of R.

(iii) The *essential spectral radius* of R, denoted $r_e(R)$, is defined as $r_e(R) \equiv \sup\{|\lambda| : \lambda \in \sigma_e(R)\}$.

Clearly, $\sigma_e(R) \subseteq \sigma(R)$ and $r_e(R) \leq r(R)$. These definitions are the ones given in reference [100]; other equivalent definitions may be found in references [50,12,129]. In particular,

$$\sigma_e(R) = \bigcap_{K \in \mathcal{K}(B)} \sigma(R+K),$$

where $\mathcal{K}(B)$ is the collection of compact operators in $L(B)$. This shows that the essential spectrum of R is invariant under compact perturbations.

Essential Spectrum of Semigroups

The following result identifies a class of semigroups that have stability properties very like those of finite-dimensional ODEs. It is similar in spirit to the remarks made in reference [118].

Proposition 3.1 *Suppose that G is a linear C^0-semigroup with infinitesimal generator A, and that $r_e(G(t_0)) < 1$ for some $t_0 > 0$. Then*

$$a.s. \iff u.a.s. \iff \mathrm{Re}\,\mathrm{P}\sigma(A) \leq \omega < 0.$$

Proof
It is shown in reference [50] that $r_e(G(t_0)) < 1$ implies that G is asymptotically smooth, so a.s. and u.a.s. are equivalent by Theorem 3.3. Furthermore, u.a.s. is equivalent to the existence of t_0 such that $r(G(t_0)) < 1$.

Now, if $\lambda \in \sigma(R)$ has $|\lambda| > r_e(R)$, then λ is a pole (of finite order) of $(\lambda - R)^{-1}$ [50, page 14]. Thus $\lambda \in \mathrm{P}\sigma(R)$ [7, page 258], and the relation (3.23) applies. In particular, the assumption $r_e(G(t_0)) < 1$ implies that $r(G(t_0)) < 1$ iff $\mathrm{Re}\,\mathrm{P}\sigma(A) < 0$.

□

Naturally, this result is only helpful when the essential spectral radius of the semigroup can be computed from the original formulation of the problem. Such computations are the concern of the following examples.

Example: Essential Spectral Radius of One-Dimensional Autonomous Linear HPDEs

Recall the one-dimensional linear HPDE of Section 3.2.3. Assume that the matrices $\tilde{A}$ and $\tilde{B}$ and the boundary conditions are autonomous. The following result is a paraphrase of parts of Theorem B of reference [100].

Theorem 3.4 *Assume that the matrices $\tilde{A}$, $\tilde{B}$, T_1, S_0, S_1^1 and S_1^2 are time-independent. Where $\tilde{A}$ and $\tilde{B}$ depend on x, make the smoothness assumptions of page 59. Then:*

(i) The semigroups $G_1(t)$ and $G_2(t)$, that are generated by A_1 and A_2, have the same essential spectral radius, and

$$r_e(G_1(t)) = e^{\alpha_0 t},$$

where $\alpha_0 \equiv \sup\{\mathrm{Re}\, z : z \in \sigma(A_2)\}$.

(ii) The spectrum of A_2 is given by

$$\sigma(A_2) = \{z \in \mathcal{C} : \ h_2(z) = 0\},$$

where

$$h_2(z) \equiv \det\left[\mathrm{diag}(y_{\ell+1}, \ldots, y_m) - T_1 \mathrm{diag}(y_1, \ldots, y_\ell) S_0\right]$$

and

$$y_i(z) \equiv \exp\left\{\int_I \frac{z - \tilde{B}_{ii}(s)}{\lambda_i(s)}\, ds\right\},$$

for each $i = 1, \ldots, m$.

This result is applied to the Telegrapher's and Klein–Gordon equations in the following examples.

□

EXAMPLE: TELEGRAPHER'S EQUATIONS

Recall the definition of the Telegrapher's equations IBVP from Section 3.2.2. For these equations, $\tilde{B}_{ii} = -a$, $i = 1, 2$ (see equation (3.8)), $\lambda_1 = -c$ and $\lambda_2 = c$, so

$$y_1(z) = e^{-(z+a)/c}, \quad y_2(z) = e^{(z+a)/c}.$$

Since $T_1 = -1$ and $S_0 = -q_0$,

$$h_2(z) = e^{(z+a)/c} - q_0 e^{-(z+a)/c}.$$

Thus[3]

$$h_2(z) = 0 \iff z = -a + \frac{c}{2}\log|q_0| + \frac{cn\pi i}{2}, \quad n \in \mathcal{Z}.$$

Then Theorem 3.4 shows that

$$r_e(G(t)) = e^{\alpha_0 t},$$

where

$$\alpha_0 = -a + \frac{c}{2}\log|q_0|.$$

(Again, typographical errors have been corrected.) Thus the essential spectral radius depends both on the spatial differential operator (via a and c) and on the boundary conditions via q_0. Recalling from equation (3.11) that q_0 is real, with modulus less than or equal to 1, clearly $\alpha_0 < 0$. Thus $r_e(G(t_0)) < 1$ for some $t_0 > 0$, and a.s. and u.a.s. are equivalent whenever the cable is not lossless ($R_0 \neq 0$). For lossless cables ($R_0 = 0$), a.s. and u.a.s. may be distinct [50, page 4].

[3] We have corrected a typographical error in references [50, page 152] and [100, page 344].

Let $A = \tilde{A}\partial_x + \tilde{B}$. If it can be shown that $\mathrm{Re}\,\mathrm{P}\sigma(A) < 0$, then Proposition 3.1 shows $r(G(t_0)) < 1$ for some $t_0 > 0$; that is u.a.s. To do this, observe that if $Au = \lambda u$, then u must satisfy $u_x + Qu = 0$, where

$$Q \equiv \frac{1}{c}\begin{bmatrix} \mu & -b \\ b & -\mu \end{bmatrix},$$

where $\mu \equiv (\lambda + a)$, together with the boundary conditions ($\lambda \neq 1$)

$$\begin{aligned} r_1 U(1) + q_1 V(1) &= 0, \\ U(0) + q_0 V(0) &= 0. \end{aligned}$$

Solutions are of the form $we^{\beta x}$. For non-trivial solutions,

$$\beta = \pm\beta_0, \quad \beta_0 \equiv \frac{\sqrt{\mu^2 - b^2}}{|c|},$$

Eigenvectors corresponding to $\pm\beta_0$ are $w_\pm = [b, \mu \pm \beta_0 c]^T$. Thus the solution is of the form

$$u = C_+ w_+ e^{+\beta_0 x} + C_- w_- e^{-\beta_0 x},$$

where $C_\pm$ are determined by the boundary conditions. For non-trivial solutions

$$\beta_0 : \; e^{2\beta_0} = \frac{f(\beta_0)}{f(-\beta_0)}, \quad f(x) \equiv [b + q_0(\mu + cx)]\,[br_1 + q_1(\mu - cx)]\,. \tag{3.24}$$

Thus we have $\mathrm{Re}\,\mathrm{P}\sigma(A) < 0$ iff equation (3.24) has no solution $\beta_0 = \left[(\lambda + a)^2 - b^2\right]^{1/2}/|c|$ corresponding to $\lambda : \; \mathrm{Re}\lambda \geq 0$.

□

EXAMPLE: KLEIN–GORDON EQUATIONS

The calculation of the essential spectral radius for the Klein–Gordon equations is very similar to the corresponding calculation for the Telegrapher's equation, except that

$$\alpha_0 = \frac{c}{2}\log|q_0|.$$

Here q_0 arises from the boundary conditions given in Section 3.2.2, whilst c arises from the spatial differential operator. Note that $|q_0| < 1$ for a.s. and u.a.s. to be equivalent. The fact that this equivalence does not hold when $|q_0| = 1$ may well be related to the fact that the corresponding IVP admits waves of zero group velocity (see page 16).

□

EXAMPLE: ONE-DIMENSIONAL LINEAR ADVECTION ON A SEMI-INFINITE STRIP

The semigroup generated by this problem is given by equation (3.7). Note that for any $t_0 > 0$, $\mathrm{P}\sigma(G(t_0)) = \emptyset$, so that the set of normal points of $G(t_0)$ is $\rho(G(t_0))$. This latter set is $\mathcal{C} - \{0\}$, thus $\sigma_e(G(t_0)) \subseteq \{0\}$ and $r_e(G(t_0)) = 0$. Thus we have u.a.s., a result that is obvious from the fact that $r(G(t_0)) = 0$ for $t_0 > 1/a$.

□

3.3.2 Non-Autonomous Linear Periodic Problems

Consider a non-autonomous linear problem $\dot{u} = A(t)u$, where A is T-periodic and generates the evolution operator $G(t,s)$. Then $S(t-s,s) \equiv G(t,s)$ is a T-periodic process. Recall that a T-periodic process, with the map at the period $S_T = S(T,0)$, satisfies equation (3.22). Thus if $t = nT + \hat{t}$, where $\hat{t} \in [0,T)$,

$$\|G(t,0)\| \leq C\|G_T^n\|,$$

where $C \equiv \sup_{r\in[0,T]} \|G(r,0)\|$, and $G_T \equiv G(T,0)$. Thus, there is a stability theory like that of autonomous problems (c.f. [118, page 502]).

Proposition 3.2 *Suppose that G is a T-periodic evolution operator. Suppose that $r_e(G_T) < 1$. Then*

(i) any invariant set is a.s. iff it is u.a.s.,

(ii) any invariant set is u.a.s. iff $\text{Re } P\sigma(G_T) \leq \omega < 0$.

This result can, in principle, be used to analyse the stability of solutions of the (unforced) periodically time-dependent hyperbolic problem discussed in Section 3.2.3. However, the computation of the spectrum of G is a current research problem [88,87,151]. A theory analogous to the Floquet theory [20] of ODEs is required.

3.4 Generalized Solutions and Well-Posedness

Suppose that $A : \mathcal{D}(A) \subseteq B \to B$ is the infinitesimal generator of the linear C^0-semigroup G. Define the solution map

$$E : \mathcal{D}(E) \subseteq Y \to \mathcal{R}(E), \quad E(u^0, f)(t) \equiv u(t; u^0, f) = G(t)u^0 + \int_0^t G(t-s)f(s)\,ds.$$

$\mathcal{D}(E) = \mathcal{D}(A) \times C^1(\Re^+, B)$, where $C^1(\Re^+, B)$ is the space of strongly continuously differentiable B-valued functions, equipped with the essential supremum norm on $\Re^+$.

Let Y be the completion of $\mathcal{D}(E)$ in the space $B \times L^\infty(\Re^+, B)$ (with the product norm $\|(x,f)\| \equiv \|x\|_B + \|f\|_{L^\infty(\Re^+,B)}$). Let X be the completion of $C^1(\Re^+, B)$ in $L^\infty(\Re^+, B)$. Since $L^\infty(\Re^+, B)$ is Banach, so too are X and Y.

When $\beta_G < 0$, $E : \mathcal{D}(E) \to X$ and is bounded because

$$\|u(t)\|_B \leq e^{\beta_G t}\|u^0\|_B - \frac{M}{\beta_G}\|f\|_{L^\infty(\Re^+,B)}.$$

By construction, $\mathcal{D}(E)$ is dense in Y. E is linear in (u^0, f). E clearly has a unique continuous extension to Y. That extension can be used to define *generalized solutions* in the sense of Section 2.2. Furthermore, the variaton of parameters formula continues to hold for generalized solutions.

Proposition 3.3 *Suppose that $A \in \mathcal{G}(M, \beta; B)$ is the infinitesimal generator of a linear C^0-semigroup G of type $\beta_G < 0$. Then E has a unique continuous extension $\hat{E} \in L(Y, X)$, so the original problem is well-posed. $\hat{E}$ is given by the variation of parameters formula.*

Proposition 3.4 *Suppose that $A \in \mathcal{G}(M,\beta;B)$ is the infinitesimal generator of a linear C^0-semigroup G, with $r_e(G(t_0)) < 1$ for some $t_0 > 0$. Then well-posedness is equivalent to u.a.s.*

Proof
Since u.a.s. implies $\beta_G < 0$, clearly u.a.s. implies well-posedness. Conversely, suppose the problem is well-posed. Let $f(t) = e^{i\eta t}\phi$, where $\phi \in B$, $\eta \in \Re$, and let $u^0 = 0$. Then $(u^0, f) \in Y$, and $E(u^0, f)(t) = e^{i\eta t}\psi$, where $\psi : \ [i\eta - A]\psi = \phi$. Thus, well posedness implies that $i\eta \in \rho(A)$ for any $\eta \in \Re$. By equation (3.23), no eigenvalue of $G(t_0)$ can have modulus 1. Well-posedness and consideration of the data $(u^0, 0) \in Y$ shows that no eigenvalue of $G(t_0)$ can have modulus greater than 1.

These facts, the assumption that $r_e(G(t_0)) < 1$, and the fact that spectral points $\lambda : |\lambda| > r_e(G(t_0))$ are isolated, imply that $r(G(t_0)) < 1$ (c.f. Proposition 3.1). Thus, by Proposition 3.1, the semigroup is u.a.s.

□

It will prove useful to isolate forcing functions that are defined for all time. Let $Y(\Re)$ be the completion of $\mathcal{D}(A) \times C^1(\Re, B)$ in the space $B \times L^\infty(\Re, B)$. Define

$$Y_f \equiv \{f : \ (0, f) \in Y(\Re)\}. \tag{3.25}$$

3.5 Equilibrium Solutions

The purpose of this section is to give rigorous definition to the term 'equilibrium solution'. Suppose that $u(t; t^0, u^0) \in B$ $(t \geq t^0)$ is the generalized solution of an IVP for an evolution equation, with data $u(t^0) = u^0 \in B$. For any bounded $C \subseteq B$, and each $t \in \Re$, let

$$w(t; C) \equiv \lim_{\substack{t^0 \to -\infty \\ u^0 \in C}} u(t; t^0, u^0). \tag{3.26}$$

When

(i) this limit exists and is finite for each bounded $C \subseteq B$ and for each $t \in \Re$,

(ii) $w(t; C) = w(t)$ is independent of C,

(iii) $\|w(t)\|_B \leq M$ for all $t \in \Re$, and

(iv) for any $t^0 \in \Re$ and $u^0 \in B$, $\lim_{t \to \infty} \|u(t; t^0, u^0) - w(t)\|_B = 0$,

then $w(t)$ is said to be *an equilibrium solution* of the evolution equation, and is denoted $u_{eq}(t)$. These conditions should be compared with the heuristic conditions given on page 52. Next we identify classes of evolution equations for which equilibrium solutions exist uniquely.

3.5.1 Equilibrium Solutions of Autonomous Linear Evolution Equations

Consider the autonomous linear evolution equation

$$\dot{u} = Au + f(t), \tag{3.27}$$

where $A \in \mathcal{G}(M, \beta)$, and $f \in Y_f$ is defined for all time (see equation 3.25). Assume G has negative type. By Proposition 3.3, the unique generalized solution of the IVP, with initial data $u(t^0) = u^0 \in B$, is given by

$$u(t; t^0, u^0, f) = G(t - t^0)u^0 + \int_{t^0}^{t} G(t - s)f(s)\, ds.$$

Then, for bounded $C \subseteq B$,

$$w(t, C) \equiv \lim_{\substack{t^0 \to -\infty \\ u^0 \in C}} u(t; t^0, u^0, f) = 0 + \int_{-\infty}^{t} G(t - s)f(s)\, ds,$$

is independent of C, since G has negative type. This observation suggests the following result.

Proposition 3.5 *Suppose that*

(i) $A \in \mathcal{G}(M, \beta; B)$ *is the infinitesimal generator of* G,

(ii) G *has negative type* $\beta_G < 0$, *and*

(iii) $f \in Y_f$.

Then equation (3.27) has the unique (bounded, u.a.s.) equilibrium solution

$$u_{eq}(t; f) = \int_0^\infty G(r)f(t - r)\, dr. \tag{3.28}$$

Proof

First note that $w(t) = w(t, C)$ is independent of the bounded set $C \subseteq B$, since $\|G(t)\| \leq Me^{\beta_G t}$, and $\beta_G < 0$. Furthermore, $w(t) = \int_0^\infty G(r)f(t - r)\, dr$ satisfies

$$\|w(t)\|_B \leq \|f\|_{Y_f} \int_0^\infty Me^{\beta_G r}\, dr = C\|f\|_{Y_f}, \quad (C \equiv -M/\beta_G > 0),$$

so w is bounded. Finally, w is u.a.s., since $\beta_G < 0$. Thus, w is the equilibrium solution of equation (3.27).

□

Note that the equilibrium solution is a 'moving average' of the forcing, with weighting (in the hyperbolic case) that depends on wave propagation.

Consider an IVP for the forced evolution equation (3.27), with the initial condition $u(t^0) = u^0 \in B$. The *transient* for the IVP is defined as

$$u_{trans}(t; t^0, u^0, f) \equiv u(t; t^0, u^0, f) - u_{eq}(t; f) = G(t)\left[u^0 - u_{eq}(t^0; f)\right]. \tag{3.29}$$

The u.a.s. of the equilibrium solution ensures that the transient will decay with time. Note that the special initial data $u^0 = u_{eq}(t^0)$ eliminates the transient.

EXAMPLE: FORCED ONE-DIMENSIONAL LINEAR ADVECTION ON A SEMI-INFINITE STRIP

Consider the evolution equation $w_t + aw_x = 0$, $w(0,t) = \phi(t)$, posed on $x \in I$. Let $\phi \in C^p(\Re)$ for $p \geq 1$, and $a > 0$. The solution is $w(x,t) = \phi(t - x/a) = w_{eq}(x,t)$.

Setting $u \equiv w - \phi$ (to render the boundary conditions homogeneous), this can be cast in the form $\dot{u} = Au + f(t)$, with $A \equiv -a\partial_x$, $\mathcal{D}(A) = \{u \in H^1 : u(0) = 0\}$, $B = L^2$, and $f : \Re \to B$ defined by $[f(t)](x) = -\phi'(t)$. This is the problem considered on page 55, so the semigroup is given by equation (3.7).

We compute the equilibrium solution. Since f is constant in space, $f(x,t) = \hat{f}(t)$ (say) for all $x \in I$. Thus, for $ar < 1$,

$$[G(r)f(t-r)](x) = \begin{cases} 0 & x \in [0, ar] \\ \hat{f}(t-r) & x \in (ar, 1] \end{cases}.$$

Consequently, writing ϕ' for the generalized derivative of ϕ,

$$\begin{aligned} [u_{eq}(t)](x) &= \left[\int_0^\infty G(r)f(t-r)\,dr\right](x), \\ &= \int_{r\in[0,x/a]} \hat{f}(t-r)\,dr, \\ &= \int_{s\in[t-x/a,t]} -\phi'(s)\,ds, \quad s \equiv t - r, \\ &= \phi(t - x/a) - \phi(t). \end{aligned}$$

Returning to the original variables, $w_{eq}(x,t) = u_{eq}(x,t) + \phi(t) = \phi(t - x/a)$, as anticipated. Furthermore, it has been shown (page 65) that u_{eq} is u.a.s. Observe that the equilibrium solution need be no more smooth in time than the forcing function.

□

EXAMPLE: FORCED ODE

For the forced ODE $\dot{u} = Au + f(t)$, the semigroup is $G(t) = e^{At}$. Obviously, when $f \in C^p(\Re)$ and $A < 0$, $u_{eq} \in C^{p+1}(\Re)$. Unlike the HPDE example considered previously, the equilibrium solution is smoother than the forcing. Note that when f is analytic, then, formally, $u_{eq}(t) = -\frac{1}{A}\sum_{m=0}^{\infty} A^{-m} f^{(m)}(t)$.

□

3.5.2 Steady and Periodic Equilibrium Solutions

When the forcing function of the autonomous linear evolutionary problem $\dot{u} = Au + f(t)$ is steady or T-periodic, the corresponding equilibrium solution will also be steady or T-periodic.

Proposition 3.6 *Suppose that the conditions of Proposition 3.5 are satisfied, and that $f \in Y_f$ is T-periodic. Then the equilibrium solution (3.28) is also T-periodic and u.a.s.*

Proof
The proof is a straightforward adaptation of that in reference [20, page 68] for ODEs. □

The hypotheses of Proposition 3.6 cannot be substantially weakened. For, consider the linear advection equation with (spatially) periodic boundary conditions

$$u_t + au_x = \dot{g}, \quad u(x,0) = u^0(x) + g(0).$$

Here $g : \Re \to \Re$ is any T-periodic differentiable real valued function, whilst u^0 is assumed to satisfy $u^0(x+X) = u^0(x)$. The exact solution of this problem is $u(x,t) = u^0(x-at) + g(t)$, which is time-periodic iff aT/X is rational. The cause of this difficulty is that $\{G(t) : t \in \Re\} \subseteq L(L^2_{per}([0,X]))$ is the group of rotations on $[0,X]$, which has zero type.

EXAMPLE: EQUILIBRIUM SOLUTIONS OF ODES

Consider the ODE $\dot{u} = Au + f(t)$, where $A < 0$ is a real number, and $f(t) = \sin(t)$. This equation has the 2π-periodic equilibrium solution $u_{eq}(t) = -\gamma\,[A\sin t + \cos t]$. The corresponding IVP has the solution $u = u_{eq} + u_{trans}$, where the transient is given by $u_{trans}(t) = [u(0) + \gamma]\, e^{At}$, $\gamma \equiv 1/(1+A^2)$. The solution of the IVP is periodic iff $u(0) = -\gamma$. This result could have been anticipated from Proposition 3.6, since $G(t) = \exp(At)$ has $r(G(t^0)) = \exp(At^0) < 1$.

□

3.5.3 Equilibrium Solutions of Non-Autonomous Linear Equations

Consider the forced non-autonomous, linear evolution equation

$$\dot{u} = A(t)u + f(t). \tag{3.30}$$

When $A(t)$ generates $G(t,s)$, the corresponding IVP, with data $u(t^0) = u^0$, has the solution

$$u(t;t^0,u^0,f) = G(t,t^0)u^0 + \int_{t^0}^{t} G(t,s)f(s)\,ds.$$

For $C \subseteq B$ bounded, define

$$w(t,C) \equiv \lim_{\substack{t^0 \to -\infty \\ u^0 \in C}} u(t;t^0,u^0,f) = \int_{-\infty}^{t} G(t,s)f(s)\,ds. \tag{3.31}$$

Suppose that the following conditions hold.

(i) For each $t \in \Re$, $\|G(t,t^0)\| \to 0$ as $t^0 \to -\infty$ sufficiently quickly for the integral to converge for bounded f, and for $w(t,C)$ to be bounded uniformly on $t \in \Re$. This novel condition ensures that $w(t,C)$ is independent of C.

(ii) For each $t^0 \in \Re$, $\|G(t,t^0)\| \to 0$ as $t \to +\infty$. This ensures that $w(t)$ is u.a.s.

When these conditions hold, $w(t, C)$ is independent of C, $w(t)$ is bounded in B, and $w(t)$ is u.a.s. Hence, the equilibrium solution is

$$u_{eq}(t; f) = \int_{-\infty}^{t} G(t, s) f(s)\, ds. \tag{3.32}$$

In general, it is difficult to ensure these conditions. When the problem is hyperbolic, with periodic spatial boundary conditions, Rauch and Taylor's paper [118] gives conditions that ensure (ii). Note that where $A(t)$ is periodic, the results of Section 3.3.2 can be used. Note that ensuring condition (ii) is tantamount to ensuring u.a.s. Even in the unforced periodic case, the difficulties referred to in Section 3.3.2 will occur.

3.5.4 Equilibrium Solutions of General FDEs: Attractors

Consider the general functional differential equation (FDE)

$$\dot{u} = f(t, u), \quad u(t^0) = u^0, \tag{3.33}$$

where $f : \Re \times B \to B$ is any map. As described in Section 3.2.4, the solution is written $u(t; t^0, u^0)$. The solution may define a process $w(t - t^0; t^0, u^0) = u(t; t^0, u^0)$. The process w defines $S(r, s) : B \to B$ for each $r \geq 0$ and $s \in \Re$ via

$$S(r, s)x \equiv w(r; s, x) \equiv u(r + s; s, x).$$

We make the following definitions. These definitions are adapted from those given in Hale's monograph [50].

(i) A *complete orbit* through $(t^0, u^0) \in \Re \times B$, denoted $\gamma(t^0, u^0)$ when it exists, is defined as

$$\gamma(t^0, u^0) \equiv \{(t, \phi(t)) : t \in \Re\},$$

where $\phi(s) : B \to B$ satisfies $\phi(t^0) = u^0$, and for each $r \geq 0$ and $s \in \Re$, $S(r, s)\phi(s) = \phi(s + r)$.

(ii) A set $M \subseteq \Re \times B$ is *invariant* if for each $(t^0, u^0) \in M$, then $\gamma(t^0, u^0)$ exists and $\gamma(t^0, u^0) \subseteq M$.

(iii) A set $M \subseteq \Re \times B$ is an *attractor* if M is invariant, and for each bounded $C \subseteq B$

$$\lim_{r \to \infty} \text{dist}(S(r, s)C, M(r + s)) = 0,$$

where $M(s) \equiv \{x \in B : (s, x) \in M\}$.

Attractors and Equilibrium Solutions

Recall the definition of an equilibrium solution given in Section 3.5. It is clear that ϕ is an equilibrium solution of $\dot{u} = f(t, u)$ iff, for any bounded set $C \subseteq B$,

(i) $M \equiv \{(t, \phi(t)) : t \in \Re\}$ is invariant,

(ii) $\text{dist}(S(r, s)C, \phi(r + s)) \to 0$ as $s \to -\infty$,

(iii) $\text{dist}(S(r, s)C, \phi(r + s)) \to 0$ as $r \to +\infty$, and

(iv) ϕ is bounded.

The second condition corresponds to requirements (i) and (ii) of the definition of an equilibrium solution (page 67); the third condition corresponds to requirement (iv), whilst the last condition corresponds to requirement (iii). Thus ϕ is an equilibrium solution iff M is bounded (that is, $M(t) \subseteq D$ for all t, where D is bounded) and both an attractor and what might be called an 'anti-attractor'.

This discussion ignores questions of existence and uniqueness. Even dissipative autonomous nonlinear (finite-dimensional) ODEs can have multiple steady equilibria. One example is given by a spherical pendulum with an iron bob, suspended over three magnets. Because of friction in the pivot, the bob will eventually settle into one of three possible steady equilibria, depending on the initial position and velocity of the bob. In this case, there is no unique 'equilibrium solution'.

A general account of equilibrium solutions for non-linear problems could probably be given for the dissipative systems considered by Hale [50]. (See also references [51,50,48, 49].) Hale identifies a large class of problems that have a compact global attractor that is connected. This theory may allow the identification of a class of FDEs that have unique equilibrium solutions.

In particular, there is a substantial body of theory for periodic processes. Recall that for periodic processes, equation (3.22) holds. This may allow the development of a theory quite similar to that of autonomous problems.

3.6 BVP Stability of Equilibrium Solutions

For autonomous linear problems, the equilibrium solution depends on the forcing alone, $u_{eq}(.) = u_{eq}(.; f)$. If the equilibrium solution is to be a good model of a physical system, the map $f \mapsto u_{eq}(.; f)$ should be continuous. This section investigates this *BVP stability* of the equilibrium solution.

BVP stability is equivalent to requiring that the equilibrium solution operator

$$F : Y_f \to L^\infty(\Re, B), \quad (Ff)(t) \equiv u_{eq}(t; f),$$

be bounded. This is not a necessary consequence of the well-posedness discussed in Section 3.4. However, for linear C^0-semigroups G, of type $\beta_G < 0$, the variation of parameters formula gives the estimate

$$\|Ff\|_{L^\infty(\Re,B)} \leq -\frac{M}{\beta_G}\|f\|_{Y_f}, \tag{3.34}$$

which shows BVP stability.

3.6.1 BVP Stability of Periodic Equilibrium Solutions

For the important case of periodic equilibrium, the approach taken so far makes excessively strict smoothness assumptions. This section offers an alternative approach. A time-uniform norm is not used, but Chapter 5 will show that the difficulties this poses can be avoided.

Let P_T be the space of functions $\phi : \Re \to H$ such that $\phi(t+T) = \phi(t)$ almost everywhere on $\Re$, and such that $\|\phi(.)\|_H \in L^2([0,T])$. H is assumed to be a seperable Hilbert space. P_T is also a separable Hilbert space [151], with the inner product

$$(\phi, \psi)_{P_T} \equiv \int_0^T (\phi(s), \psi(s))_H \, ds.$$

If $\{x_j\}_j$ is a basis for H then $\{x_j e^{im\omega t}\}_{j,m}$ ($\omega = 2\pi/T$) is a basis for P_T.

If $\phi \in P_T$ then $\phi = \sum_{m\in\mathcal{Z}} \hat{\phi}^m e^{im\omega t}$ for some $\hat{\phi} \in \hat{P}_T$; the space of H-valued, bi-infinite, square-summable sequences with the norm

$$\|\hat{\phi}\|^2_{\hat{P}_T} = \sum_{m\in\mathcal{Z}} \|\hat{\phi}^m\|^2_H.$$

Also, $\|\phi\|_{P_T} = \|\hat{\phi}\|_{\hat{P}_T}$.

Let $F \in L(P_T)$, and define $\hat{F} : \hat{P}_T \to \hat{P}_T$ via $\hat{F}\hat{\phi} \equiv \widehat{F\phi}$. It is easily seen that $\hat{F} \in L(\hat{P}_T)$, and that $\|\hat{F}\| = \|F\|$. Furthermore,

$$\|F\| = \|\hat{F}\|^2. \tag{3.35}$$

For the evolution equation $\dot{u} = Au + f$, with $f \in P_T$, the solution u has the Fourier coefficients $\hat{u}$ given by $(im\omega - A)\,\hat{u}^m = \hat{f}^m$, where $m \in \mathcal{Z}$. Thus, the equilibrium solution operator $F : P_T \to P_T$ has the Fourier transform

$$\hat{F} = \operatorname{diag}\left(\ldots, \hat{F}^{-1}, \hat{F}^0, \hat{F}^1, \ldots\right), \quad \hat{F}^m \equiv (im\omega - A)^{-1}.$$

By equation (3.35),

$$\|F\| = \max_{m\in\mathcal{Z}} \|(im\omega - A)^{-1}\|_{L(B)}.$$

Consequently, periodic BVP-stability is equivalent to the existence of a constant C such that

$$\max_{m\in\mathcal{Z}} \|(im\omega - A)^{-1}\|_{L(B)} \le C. \tag{3.36}$$

Certainly $\beta_G < 0$ is sufficient for periodic BVP stability.

3.7 Conclusions

In Section 3.5, we defined an equilibrium solution to be a global, generalized solution of an evolutionary problem, that is bounded for all time and is u.a.s. This definition coincides with the heuristic definition given at the start of this chapter, and is believed to be new.

For autonomous linear problems, we have obtained a complete account of the (unique) existence of equilibrium solutions (Section 3.5.1), provided that the linear C^0-semigroup satisfies $r(G(t_0)) < 1$ for some $t_0 > 0$. This assumption is vital. It ensures the existence of generalized solutions, which in the case of hyperbolic PDEs may not be smooth in time or space. It also ensures the uniform asymptotic stability of solutions, and that bounded forcing will always result in bounded solutions.

Furthermore, when the semigroup is 'well behaved' — in the sense that $r_e(G(t_0)) < 1$ for some $t_0 > 0$ — we have seen that well-posedness, a.s. and u.a.s. are all equivalent (Propositions 3.1 and 3.4). All can be decided from an investigation of the spectrum of the generator (Proposition 3.1).

For non-autonomous linear evolution equations, we have presented conditions on the resulting two parameter semigroup that ensure the existence of equilibrium solutions. However, ensuring these conditions is more difficult than in the autonomous case. This is not surprising. Even periodic non-autonomous hyperbolic PDEs are notoriously difficult to analyse [88,87]. Forced hyperbolic problems generally present considerable difficulties [91,100,113,116,117,118].

Our definition of an equilibrium solution is applicable to non-linear evolutionary problems. However, it is difficult to identify classes of non-linear evolution equations for which equilibrium solutions exist uniquely.

The requirement that the semigroup decay makes well-posedness very easy to verify. The well-posedness theory of Kreiss [75] had the very different aim of proving well-posedness over finite time intervals $t \in [0, T]$. Since $e^{\beta_G T}$ is just a constant, Kreiss could prove well-posedness for problems with non-negative type. Unfortunately, this elegant theory does not help the investigation of equilibrium problems.

Chapter 4

Solutions of Forced Difference Equations

4.1 Introduction

Our aim is to find effective methods for the approximation of equilibrium solutions. This chapter is devoted to defining 'numerical equilibrium solutions', and investigating their properties. Questions of convergence are not discussed explicitly until the next chapter.

To define the discrete counterparts of equilibrium solutions, a standard form for the numerical method is needed. The standard form we adopt is suggested by the following considerations. Suppose that A generates the semigroup G. Taking a fixed timestep $k > 0$,

$$\frac{du}{dt} \approx \frac{u(t+k) - u(t)}{k}, \quad A \approx \frac{G(k) - I}{k}.$$

These expressions suggest the time-discretization

$$u((n+1)k) \approx G(k)u(nk) + kf(nk).$$

However, PDEs must also be discretized in the spatial directions. Consequently, a parameter $h > 0$ is introduced to govern the mesh size. In several spatial dimensions, it is assumed that the grid size, h_i, in the x_i direction, is a function of the parameter h. That is, $h_i = h_i(h)$. It is also assumed that the (fixed) timestep is a function of h, so $k = k(h)$. Writing $G_h \approx G(k(h))$ (in some sense) for the fully discretized evolution operator, and u_h^n for an approximation to the exact solution at time $t = nk$, yields the fully discretized evolutionary method

$$u_h^{n+1} = G_h u_h^n + k f_h^n, \quad n \geq 0.$$

The specification of the discrete evolutionary problem is completed by prescribing initial data $u_h^0 \approx u(0)$. This standard form is used throughout for discrete autonomous linear evolutionary problems. It allows an obvious definition of a numerical equilibrium solution:

$$u_{eq,h}^n \equiv k \sum_{m=-\infty}^{n-1} G_h^{n-m-1} f_h^m.$$

There are two crucial differences between exact and discrete semigroups. First, for all practical computations, the discrete semigroup $G_h^n : B_h \to B_h$ is an operator on finite-dimensional spaces B_h. Since $\dim(B_h) < \infty$, the spectrum of G_h consists solely of eigenvalues: that is, $\sigma(G_h) = \mathrm{P}\sigma(G_h)$. This observation avoids many of the problems presented

by the underlying infinite-dimensional PDE in the last chapter. Secondly, the type of the discrete semigroup, β_{G_h}, may depend on h. This is the cause of most of the problems encountered in the present chapter, for reasons that are now discussed.

In the previous chapter, the definition of an equilibrium solution required the discrete semigroup to have negative type. That is,

$$\beta_G < 0 \iff \exists t_0 > 0 : \; r(G(t_0)) < 1.$$

In particular, this condition ensured the uniform asymptotic stability, well-posedness and BVP stability of the equilibrium solution. However, in the discrete case, there is the possibility that $\beta_{G_h} < 0$, but $\lim_{h \to 0} \beta_{G_h} = 0$. This possibility would have two consequences. First, it would imply that the rate of decay of the numerical transient could deteriorate as the mesh is refined. Reducing the size of the numerical transient to acceptable levels would require increasing amounts of computational effort as the mesh is refined. In fact, we shall show in the next chapter that the *rate* of increase of computational effort is tantamount to an increase in the dimensionality of the problem. Secondly, numerical equilibrium solutions could become increasingly dependent on 'old' forcing data (as the mesh is refined) in a way that is physically unrealistic. This phenomenon could prevent the convergence of the numerical equilibrium solution to the exact equilibrium solution.

These difficulties may be avoided by requiring that one of the following three conditions hold.

(i) $\beta_{G_h} \leq \beta < 0$ for all h sufficiently small. In this case we say that the discrete semigroup has *uniformly negative type.*

(ii) $\exists t_0 > 0 : \; r(G_h(t_0)) \leq q < 1$. We term this the *uniform spectral radius condition.*

(iii) $\exists \beta < 0 : \; r(G_h) \leq e^{\beta k(h)}$ for all h sufficiently small. We term this the *modified von Neumann condition.*

It will be shown that these conditions are, in fact, equivalent. The last condition has a particularly straight-forward interpretation. If $\{G_h(t) : \; t \in \Re^+\}$ is to have an exponential rate of decay, then $\hat{G}_h(t) \equiv e^{-\beta t} G_h(t)$ should be uniformly bounded in t and h, for some $\beta < 0$. Applying the (necessary) von Neumann condition to $\{\hat{G}_h(t) : \; t \in \Re^+\}$ gives the modified von Neumann condition.

As in the continuous case, the discrete equilibrium solution is a function of the forcing alone. Thus, a discrete equilibrium solution operator may be defined, that maps the forcing f_h onto the equilibrium solution $u_{eq,h}$; but the notion of BVP stability must be modified to accommodate the parameter h. For discrete problems, 'BVP stability' will require that the map $f_h \mapsto u_{eq,h}$ be uniformly continuous in h.

However, the use of evolutionary initialization procedures suggests that we should also require that the solution depend continuously on the initial data — with the forcing fixed. Consequently, various new definitions of IVP stability are given, and their merits are discussed. Finally, IBVP stability is discussed.

The modified von Neumann condition is necessary for all these types of stability, but is, unfortunately, difficult to verify. The latter half of this chapter concerns necessary conditions for the modified von Neumann condition. These necessary conditions are phrased in terms of the condition

$$\|(e^{i\eta} - G_h)^{-1}\| \leq Ck(h)^{-1},$$

for each $\eta \in \Re$ and each $h > 0$. We shall call this the *near-uniform invertibility* condition. This condition is suggested by analogy with the requirement that Re $P\sigma(A) < 0$ for exponential decay of 'well behaved' semigroups. This requirement implies the invertibility of

$$-i\omega + A = \lim_{k \longrightarrow 0} \frac{e^{-i\omega k} G(k) - I}{k},$$

for each $\omega \in \Re$, which suggests (with some sleight of hand) that

$$A_h(\eta) \equiv \frac{e^{-i\eta} G_h - I}{k}$$

should be *uniformly* invertible for each real η. This requirement is the near-uniform invertibility condition. Unfortunately, this suggestive line of reasoning is wrong. The near-uniform invertibility condition does turn out to be necessary for the modified von Neumann condition, but its necessity must be proved in an entirely different way.

4.2 Discrete Semigroup Theory

Consider the autonomous linear method with initial data ϕ_h,

$$u_h^{n+1} = G_h u_h^n + k f_h^n, \quad n = 0, 1, \ldots, \tag{4.1}$$

$$u_h^0 = \phi_h, \tag{4.2}$$

where h is a mesh parameter: it might typically be the maximum width, diameter or volume of a mesh element. u_h^n is intended to be an approximation to the solution $u(nk(h))$ of an autonomous linear evolution equation. It is assumed that u_h^n is a member of a Banach space B_h for each n. The space B_h enforces the genuine boundary conditions: numerical boundary conditions are enforced through G_h. This form also admits implicit one-step methods, a point we shall return to later. (Some multi-step schemes can also be cast in this form: see, for example, reference [135].)

In general, it is assumed that $k(h) = Ch^p$ for some $p > 0$. For hyperbolic problems, $p = 1$. It will often be convenient to write $G_h(t)$ for G_h^n when $t = nk$. Note that $G_h(nk) = G_h^n$ and $G_h \equiv G_h(k)$, so the superscript can be interpreted both as the number of the timestep, and as a power (repeated multiplication). In some sense, G_h is intended to be an approximation to $G(k(h))$, where G is the semigroup associated with the exact evolution equation.

For each $h > 0$, assume $G_h \in L(B_h)$, where B_h is the solution space at each time step. For practical computations $\dim(B_h) < \infty$, but it is occasionally useful to consider infinite-dimensional B_h. Typically, $B_h = \ell_h^p(\Omega)$, where Ω is some bounded domain in $\Re^d$. The *type* of the discrete semigroup is

$$\beta_h \equiv \beta_{G_h} \equiv \lim_{t \longrightarrow \infty} \frac{1}{t} \log \|G_h(t)\| = \frac{1}{k(h)} \log r(G_h). \tag{4.3}$$

Observe that, for some $M_h > 0$,

$$\|G_h(t)\| \leq M_h e^{\beta_h t}. \tag{4.4}$$

The solution of equations (4.1,4.2), is given by the discrete variation of parameters formula,

$$u_h^n = G_h^n u_h^0 + k \sum_{m=0}^{n-1} G_h^{n-m-1} f_h^m. \tag{4.5}$$

The discrete solution operator $S_h(n,l)$ maps $u_h^l \mapsto u_h^{n+l}$, via

$$u_h^{n+l} = S_h(n,l)u_h^l.$$

S_h satisfies the pair of conditions

(i) $S_h(0,l) = I_h$, and

(ii) $S_h(n,k+l)S_h(k,l) = S_h(n+k,l)$.

For non-autonomous discrete problems, the approximate evolution operator for the unforced problem is $G_h(n,m)$. A solution operator $S_h(n,m)$, which is analogous to the exact solution operator $S(t,s)$ of Chapter 3, may be defined.

4.2.1 Discrete Spaces

This section introduces some generalizations of the ℓ_h^p spaces. For $1 \leq p < \infty$, let

$$\ell_h^p(\Omega, D_h) = \left\{ \{u_j\}_{j \in S_h} : u_j \in D_h, \sum_{j \in S_h} h\|u_j\|_{D_h}^p < \infty \right\},$$

where S_h is the set of multi-indices corresponding to grid points in the domain Ω, and D_h is Banach. The norm is

$$\|u\|_{\ell_h^p(\Omega,D_h)} \equiv \left\{ \sum_{j \in S_h} h\|u_j\|_{D_h}^p \right\}^{1/p}$$

When $p = \infty$, $\ell_h^\infty(\Omega, D_h)$ is the collection of bounded grid functions, with norm

$$\|u\|_{\ell_h^\infty(\Omega,D_h)} \equiv \sup_{j \in S_h} \|u_j\|_{D_h}.$$

When $D_h = \mathcal{C}$, the complex plane, abbreviate $\ell_h^p(\Omega) \equiv \ell_h^p(\Omega, \mathcal{C})$.

4.2.2 A Battery of Methods for the Linear Advection Equation

Most of the examples presented in this chapter will be finite difference schemes for the scalar linear advection equation

$$u_t + au_x = f(x,t), \quad a > 0, \tag{4.6}$$

$$u(x,0) = u^0(x), \tag{4.7}$$

$$u(0,t) = 0, \tag{4.8}$$

posed on $(x,t) \in I \times \Re^+$. Note that homogeneous boundary conditions have been stipulated at $x = 0$. This is not a restriction, since inhomogeneous boundary forcing terms may be incorporated into f.

The way in which f is approximated may affect the accuracy of finite difference schemes. For the Box scheme, centred at $((i+1/2)h, (n+1/2)k)$, take $f((i+1/2)h, (n+1/2)k)$. For the Sloan, Crank–Nicolson and Angled-Derivative schemes, centred at $(ih, (n+1)k)$, take $f(ih, (n+1/2)k)$.

Where appropriate, the following first order accurate numerical boundary conditions are used at $x = 1$:

$$\text{Horizontal Extrapolation (HE)}: \quad u_J^{n+1} = u_{J-1}^{n+1}, \tag{4.9}$$

$$\text{Diagonal Extrapolation (DE)}: \quad u_J^{n+1} = u_{J-1}^{n}. \tag{4.10}$$

The (second order) Box scheme will also be used at $x = 1$ as a boundary condition. When numerical boundary conditions are needed at $x = 0$, analogous formulae are used. Furthermore, for forced problems, the forcing is represented by $f((J-1/2)h,(n+1)k)$ for horizontal extrapolation, and by $f((J-1/2)h,(n+1/2)k)$ for diagonal extrapolation.

For $\nu \in (0,1)$, once proper centring has been taken into account, the four schemes and two numerical boundary conditions result in seven possible combinations of schemes and boundary conditions: that is, seven *methods*. However, for CFL numbers greater than one, the two methods that result from using Sloan's scheme with HE and DE at $x = 0$ coincide, and both will be denoted SLHEDE. Thus:

- For $\nu \in (0,1)$, there are seven methods, denoted BOX, CNHE, CNDE, ADHE, ADDE, SLHE and SLDE.
- For $\nu \in (1,\infty)$, there are six methods, denoted BOX, CNHE, CNDE, ADHE, ADDE and SLHEDE.

However, we shall also be considering the Crank–Nicolson scheme teamed with Box scheme boundary conditions. This method is denoted CNBOX.

For illustration, consider the methods that arise from the AD and CN schemes, and from Sloan's scheme for $\nu \in (0,1)$. These methods take the form

$$S_h u_h^{n+1} = T_h u_h^n + k f_h^n,$$

where $u_h^n = [u_1^n, \ldots, u_J^n] \in \Re^J$ $(J \equiv 1/h)$. All these methods require both a left-hand and a right-hand boundary condition. For $1/h = J = 6$, the matrices S_h and T_h are

$$S_h = \begin{bmatrix} b & c & 0 & 0 & 0 & 0 \\ a & b & c & 0 & 0 & 0 \\ 0 & a & b & c & 0 & 0 \\ 0 & 0 & a & b & c & 0 \\ 0 & 0 & 0 & a & b & c \\ 0 & 0 & 0 & 0 & q & r \end{bmatrix}, \quad T_h = \begin{bmatrix} b & a & 0 & 0 & 0 & 0 \\ c & b & a & 0 & 0 & 0 \\ 0 & c & b & a & 0 & 0 \\ 0 & 0 & c & b & a & 0 \\ 0 & 0 & 0 & c & b & a \\ 0 & 0 & 0 & 0 & s & t \end{bmatrix}, \tag{4.11}$$

where q, r, s and t depend on whether horizontal or diagonal extrapolation is used at the right-hand boundary. The parameters $a(\nu)$, $b(\nu)$ and $c(\nu)$ were defined for each scheme in Chapter 2, but note that $c(\nu) \equiv 0$ for AD. Observe that

$$S_h = A_h + P_{1h}, \quad T_h = B_h + P_{2h},$$

where P_{1h} and P_{2h} are perturbations of

$$A_h = aE_h^{-1} + bI_h + cE_h, \quad B_h = cE_h^{-1} + bI_h + aE_h, \tag{4.12}$$

and E_h and E_h^{-1} are, respectively, the left and right shift operators with homogeneous boundary conditions on $[0,1]$.

For $\nu > 1$, the SL scheme needs a numerical boundary condition at $x = 0$. For both HE and DE, this results in two homogeneous left-hand boundary conditions, thus $S_h = A_h$ and $T_h = B_h$, where

$$A_h = aE_h^{-2} + bE_h^{-1} + cI_h, \quad B_h = cE_h^{-2} + bE_h^{-1} + aI_h,$$

and E_h^{-1} is the homogeneous boundary condition right shift operator. This is the method denoted SLHEDE.

The Box scheme does not need a numerical boundary condition, so the method[1] is $A_h u_h^{n+1} = B_h u_h^n + k f_h^n$, with

$$A_h = aE_h^{-1} + bI_h, \quad B_h = bE_h^{-1} + aI_h.$$

For this method, the approximate evolution operator $G_h = A_h^{-1} B_h$ may be computed exactly. The simplest approach uses a Laurent expansion for

$$\lambda(z, \nu) = \frac{bz^{-1} + a}{az^{-1} + b}$$

and the nilpotence of E_h^{-1}, to obtain $(1/h = J = 6)$

$$G_h = -\frac{\gamma}{\delta} \begin{bmatrix} \delta & 0 & 0 & 0 & 0 & 0 \\ 1 & \delta & 0 & 0 & 0 & 0 \\ \gamma & 1 & \delta & 0 & 0 & 0 \\ \gamma^2 & \gamma & 1 & \delta & 0 & 0 \\ \gamma^3 & \gamma^2 & \gamma & 1 & \delta & 0 \\ \gamma^4 & \gamma^3 & \gamma^2 & \gamma & 1 & \delta \end{bmatrix}, \tag{4.13}$$

where

$$\gamma \equiv -a/b \quad \text{and} \quad \delta \equiv \frac{ab}{b^2 - a^2}. \tag{4.14}$$

Observe that all these methods are GKS stable [47].

4.3 IVP Stability of Solutions

In this section the notions of IVP stability that were discussed in the previous chapter are modified to accommodate numerical evolution operators, with parameter h. The counterpart of IVP stability yields a trivial extension of the standard Lax–Richtmyer numerical stability condition, to accommodate infinite time intervals. However, the counterparts of various types of asymptotic stability yield new and interesting conditions.

4.3.1 Numerical Stability Definitions for Autonomous Linear Problems

Recall that most of the discussion in Chapter 3 was limited to problems with the property that a.s. and u.a.s. are equivalent. In particular, recall that a.s. is equivalent to u.a.s. when $\dim(B) < \infty$, and that both hold iff the semigroup G has negative type.

For each *fixed* $h > 0$, the stability definitions of Chapter 3 apply to the (discrete) semigroup $G_h(.)$. For practical computations, necessarily, $\dim(B_h) < \infty$, so a.s. and u.a.s.

[1] Since there are no numerical boundary conditions involved, no distinction is made here between the terms 'scheme' and 'method' .

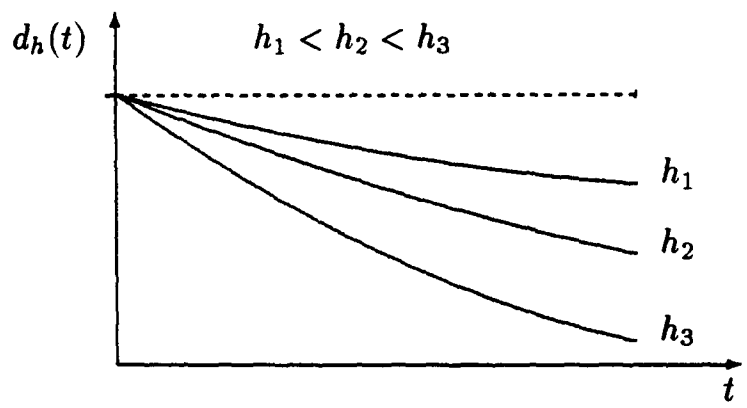


Figure 4.1: A method that is merely n.a.s. may have a rate of decay that slows as the mesh is refined.

are equivalent; both hold iff G_h has negative type (that is $\beta_h < 0$ for each h). Consequently, there is no *practical* need to distinguish asymptotic stability that is ***uniform*** with respect to the initial data. However, there is a need to distinguish asymptotic stability that is ***uniform*** with respect to the mesh parameter h; as mentioned earlier, it is possible for a method to have negative type $\beta_h < 0$, but for $\beta_h \rightarrow 0$ as the mesh is refined. That is, the *rate* of decay may deteriorate as the mesh is refined, as depicted in Figure 4.1. From now on, 'uniform' will denote uniformity with respect to the mesh parameter.

For the numerical method, define

$$d_h(t) \equiv \|G_h(t)\|_{L(B_h)} \tag{4.15}$$

at the times $t = nk$. At intermediate times, we use the linear interpolant, so d_h is continuous for $t \in \Re^+$. We shall use the following definitions of stability for discrete methods.

Numerically Stable (or n.s.) iff $\exists\, h_0 > 0$:

(i) $\sup_{t \in \Re^+} d_h(t) = M_h < \infty$, for each $0 < h < h_0$, and

(ii) $\sup_{0<h<h_0} M_h < \infty$.

Numerically Asymptotically Stable (or n.a.s.) iff

(i) the method is numerically stable, and

(ii) for *each* $0 < h < h_0$, $\lim_{t \rightarrow \infty} d_h(t) = 0$.

Uniformly Numerically Asymptotically Stable (or u.n.a.s.) iff

(i) the method is numerically asymptotically stable, and

(ii) for all $\epsilon > 0$, $\exists\, T(\epsilon)$: $\quad t > T(\epsilon), h > 0 \implies d_h(t) < \epsilon$.

4.3.2 Ensuring n.s.

The definition of numerical stability requires $\{d_h\}_{h>0}$ to be a bounded subset of $C(\Re^+)$. It is equivalent to

$$\sup_{0<h<h_0} \max_{n\geq 0} \|G_h^n\| < \infty, \tag{4.16}$$

which is just Lax–Richtmyer stability extended to the infinite time interval $t \in [0, \infty)$. We shall call condition (4.16) the *power-boundedness condition*.

4.3.3 Ensuring n.a.s.

Proposition 4.1 *A method is n.a.s. iff*

(i) the power-boundedness condition (4.16) holds, and

(ii) $r(G_h) < 1$ *for* each $0 < h < h_0$.

Proof

The result holds because $\|G_h(t)\| \to 0$ as $t \to \infty$ (h fixed) iff G_h has negative type. This is equivalent to $r(G_h) < 1$. Finally, n.s. is equivalent to power-boundedness.

□

Gustafsson [46] has found conditions within the GKS framework which ensure that only time-decaying solutions exist, and Foreman [33] has observed that Gustafsson's conditions imply the spectral radius condition $r(G_h) < 1$. The spectral radius condition is a basic assumption of Foreman's paper. This would suggest that n.a.s. could be suitable as a definition of stability for IVPs. However, we reject this definition. In the case of IVPs, it does not ensure a uniform (in h) rate of decay for the discrete solution. An unfortunate choice of (what is in practice arbitrary) initial data, with a fine mesh, could result in the integration time to reduce the transient to an acceptable level being orders of magnitude larger than the correponding time for the analytic solution.

In fact, the spectral radius condition alone can only ensure that $d_h(t)$ will decay for each fixed h; we are only assured that there exists $\beta_h < 0$, such that (c.f. equation (4.4))

$$d_h(t) \leq M_h e^{\beta_h t}, \quad \forall t \in \Re^+.$$

Note that the spectral radius condition *alone* cannot ensure power-boundedness. This point has been understood since Godunov and Ryabenkii's work in the late 1950s (see reference [38]), and is illustrated in Figure 4.2.

When a method is both power bounded and satisfies the spectral radius condition (so it is n.a.s.), then we are assured that (c.f. equation (4.4))

$$d_h(t) \leq M e^{\beta_h t}, \quad \forall t \in \Re^+, \tag{4.17}$$

where M is independent of h and $\beta_h < 0$. Clearly, problems may still occur if $\beta_h \to 0$ as $h \to 0$. This motivates the idea of u.n.a.s.

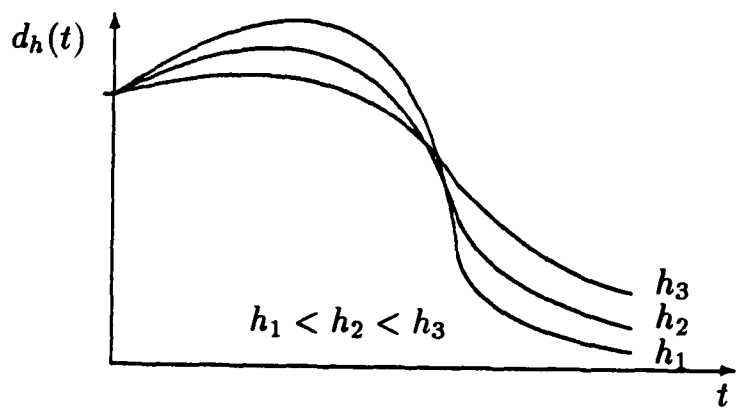


Figure 4.2: Alone, the spectral radius condition cannot ensure even n.s. After reference [38, page 438].

4.3.4 Ensuring u.n.a.s.

Proposition 4.2 *A method is u.n.a.s. iff*

(i) the power-boundedness condition (4.16) holds, and

(ii) there exists $\beta < 0$, such that $r(G_h) \leq e^{\beta k(h)}$ for h sufficiently small.

Proof

Conditions (i) and (ii) imply the existence of $M > 0$, such that $d_h(t) \leq M e^{\beta_h t}$, where β_h is the type of the discrete semigroup: see equation (4.4). The proof will be complete if we can show that condition (ii) is equivalent to the existence of $\beta < 0$ such that $\beta_h \leq \beta$, for all h sufficiently small.

But condition (ii) implies the existence of such a β, since

$$\beta_h \equiv \frac{1}{k} \log(r(G_h)) \leq \beta.$$

Conversely, the existence of such a β implies condition (ii).

□

Notice that the *modified von Neumann condition* (condition (ii) of the Proposition)

$$\exists \beta < 0, h_0 > 0 : \; r(G_h) \leq e^{\beta k(h)}, \; \forall h \in (0, h_0), \tag{4.18}$$

is more stringent than the n.a.s. requirement of $r(G_h) < 1$, since it dictates how the spectral radius may vary for small h. It is equivalent to requiring that G_h have *uniformly negative type*: that is, there exists $\beta < 0$ such that $\beta_h < \beta$ for all h sufficiently small. It is also

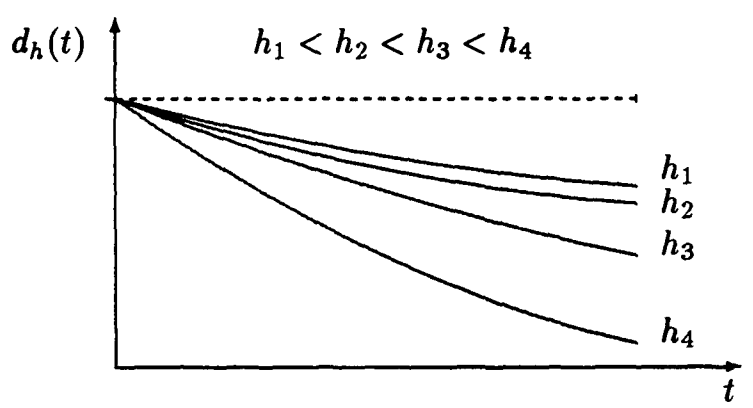


Figure 4.3: A method that is u.n.a.s. has a decay envelope.

equivalent to the condition

$$\exists h_0, t_0 > 0 : \ r(G_h(t_0)) \leq q < 1, \ \forall h \in (0, h_0).$$

Observe that a n.s. method which satisfies the condition

$$\exists C, h_0 > 0 : \quad r(G_h) \leq 1 - Ck(h), \ \forall h \in (0, h_0), \tag{4.19}$$

will be u.n.a.s. It is interesting to compare this condition with the von Neumann condition for finite time IVP stability, which is [120,135]

$$\exists C, h_0 > 0 : \quad r(G_h) \leq 1 + Ck(h), \ \forall h \in (0, h_0).$$

Also, observe that any method which satisfies

$$r(G_h) \geq 1 - Ck^p, \quad C > 0, p > 1, \tag{4.20}$$

cannot be u.n.a.s. This point will prove crucial later in this chapter.

The previous proof shows that u.n.a.s. schemes have a *decay envelope*. In fact, there exists $M > 0$, $\beta < 0$ such that

$$d_h(t) \leq Me^{\beta t}, \quad \forall t \in \Re^+, h > 0. \tag{4.21}$$

This is illustrated in Figure 4.3. Of course, so far, there is no reason to suppose that the worst rate of decay is in any way related to the rate of decay of the analytic solution that the method is intended to model. The following lemma will be helpful later.

Proposition 4.3 *Suppose that a method is u.n.a.s. Then, for all h sufficiently small,*

$$k \sum_{r=0}^{\infty} \|G_h^r\| \leq C.$$

Proof
By inequality (4.21),

$$k\sum_{r=0}^{\infty}\|G_h^r\| \le Mk\sum_{r=0}^{\infty}e^{r\beta k} = \frac{Mk}{1-e^{\beta k}}.$$

For $k(h)$ sufficiently small, $Mk/(1-e^{\beta k}) \le -2M/\beta \equiv C > 0$.

□

In general, verifying the modified von Neumann condition is difficult. The examples that follow illustrate special cases where the condition can be verified directly.

EXAMPLE: IBVP FOR THE BOX SCHEME

Consider the Box scheme for the linear advection equation on $\Omega = [0,1]$, with a homogeneous boundary condition at $x = 0$, that was considered in Section 4.2.2. This method has an approximate evolution operator that is a rational function of E_h^{-1} alone. Since $\sigma(E_h^{-1}) = \{0\}$, the spectral mapping theorem gives

$$\sigma(G_h) = \sigma\left((aE_h^{-1} + bI_h)^{-1}(bE_h^{-1} + aI_h)\right) = \left\{\frac{a}{b}\right\};$$

hence,

$$r(G_h) = \left|\frac{a}{b}\right| = \left|\frac{1-\nu}{1+\nu}\right| < 1, \quad \text{for any } \nu > 0.$$

By Proposition 4.2, the method is u.n.a.s. whenever it is power bounded.

□

EXAMPLE: SLHEDE METHOD

In Section 4.2.2 it was shown that Sloan's scheme must be used with two left-hand boundary conditions for $\nu > 1$, and no right-hand boundary conditions. For $\nu > 1$, as in the last example, the approximate evolution operator is a rational function of E_h^{-1} alone, and similar reasoning yields

$$\sigma(G_h) = \sigma\left((aE_h^{-2} + bE_h^{-1} + cI_h)^{-1}(cE_h^{-2} + bE_h^{-1} + aI_h)\right) = \left\{\frac{a}{c}\right\};$$

hence,

$$r(G_h) = \left|\frac{a}{c}\right| = \left|\frac{(1-\nu)(2-\nu)}{(1+\nu)(2+\nu)}\right| < 1, \quad \text{for any } \nu > 1.$$

By Proposition 4.2, the method is u.n.a.s., provided we have power-boundedness. When $\nu \in (0,1)$, $r(G_h) \le q < 1$, but the boundary conditions are placed wrongly. Power-boundedness does not hold, and the method is not u.n.a.s.

□

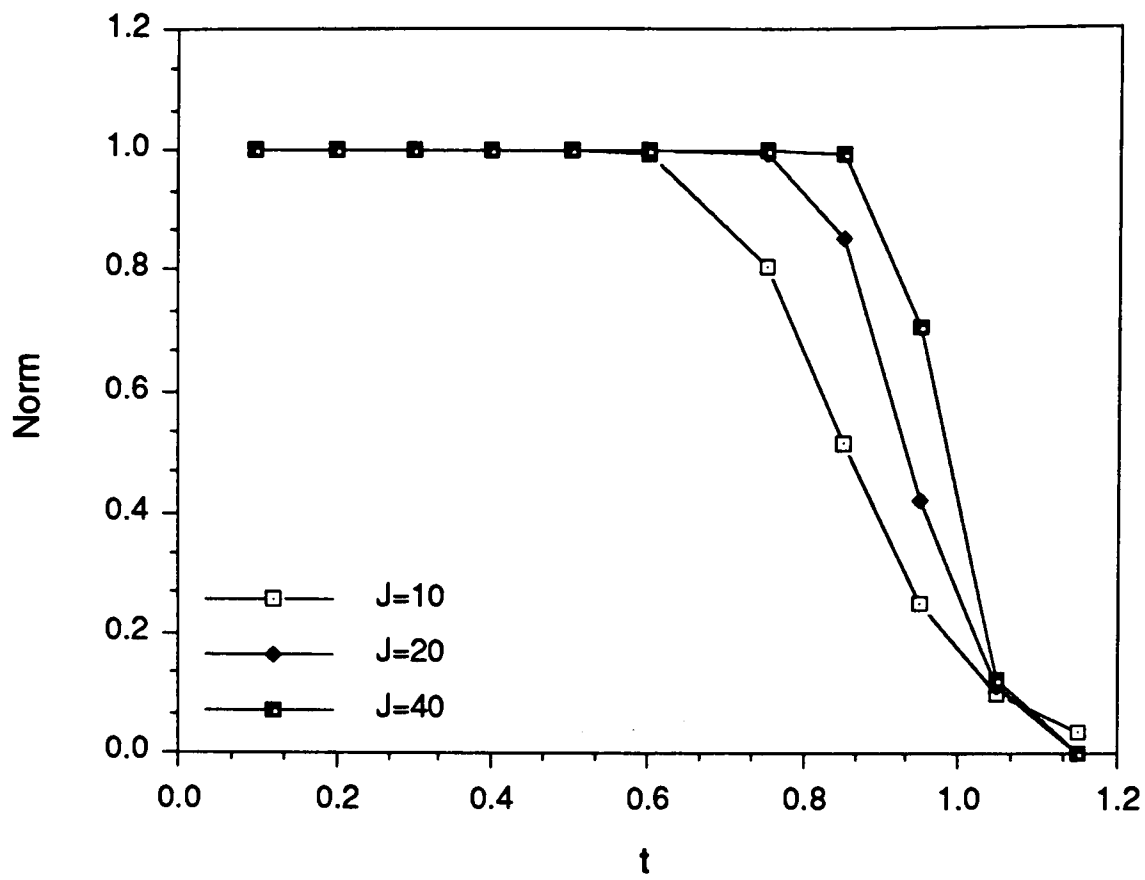


Figure 4.4: $d_h(t)$ for the Box method for $\nu = 0.5$.

EXAMPLE: COMPUTATIONS WITH METHODS FOR THE LINEAR ADVECTION EQUATION IBVP

This section presents some numerical results concerning the u.n.a.s. of the methods that were presented in Section 4.2.2.

Graphs were made of $d_h(t)$ for each of the methods in the battery, with the CFL numbers $\nu = 0.5, 1.5, 4.0$, and mesh sizes $h = 1/10, 1/20, 1/40$. The variation in the rate of decay of $d_h(t)$, as h was reduced, allowed an empirical decision of the u.n.a.s. of a method to be made. These empirical decisions are summarized in Table 4.1. A letter 'y' indicates that the decay rates did not deteriorate as the mesh was refined: that is, the method was (experimentally) u.n.a.s. A letter 'n' indicates that the method was not experimentally u.n.a.s. The symbol '–' indicates that the norm grew too rapidly, in t, to be graphed. Methods with this property are certainly not u.n.a.s.

Representative graphs of $d_h(t)$ are given in Figures 4.4–4.9 for the Box and CNBOX methods. These graphs should be compared with the exact result ($a = 1$)

$$f(t) = \begin{cases} 1 & t \in [0,1) \\ 0 & t \in [1,\infty) \end{cases},$$

irrespective of ν and h.

The norms given in these figures are true $L(\ell_h^2(I))$ norms. They were calculated by computing the (symmetric) matrix $(G_h^n)^*(G_h^n)$ at each n, and then finding its maximum (positive, real) eigenvalue $\lambda_{max}(h,n)$, using an appropriate NAG library routine. The required norm $d_h(nk(h))$ is then $\sqrt{\lambda_{max}}$ [40]. These computations were performed in double precision VAX Fortran.

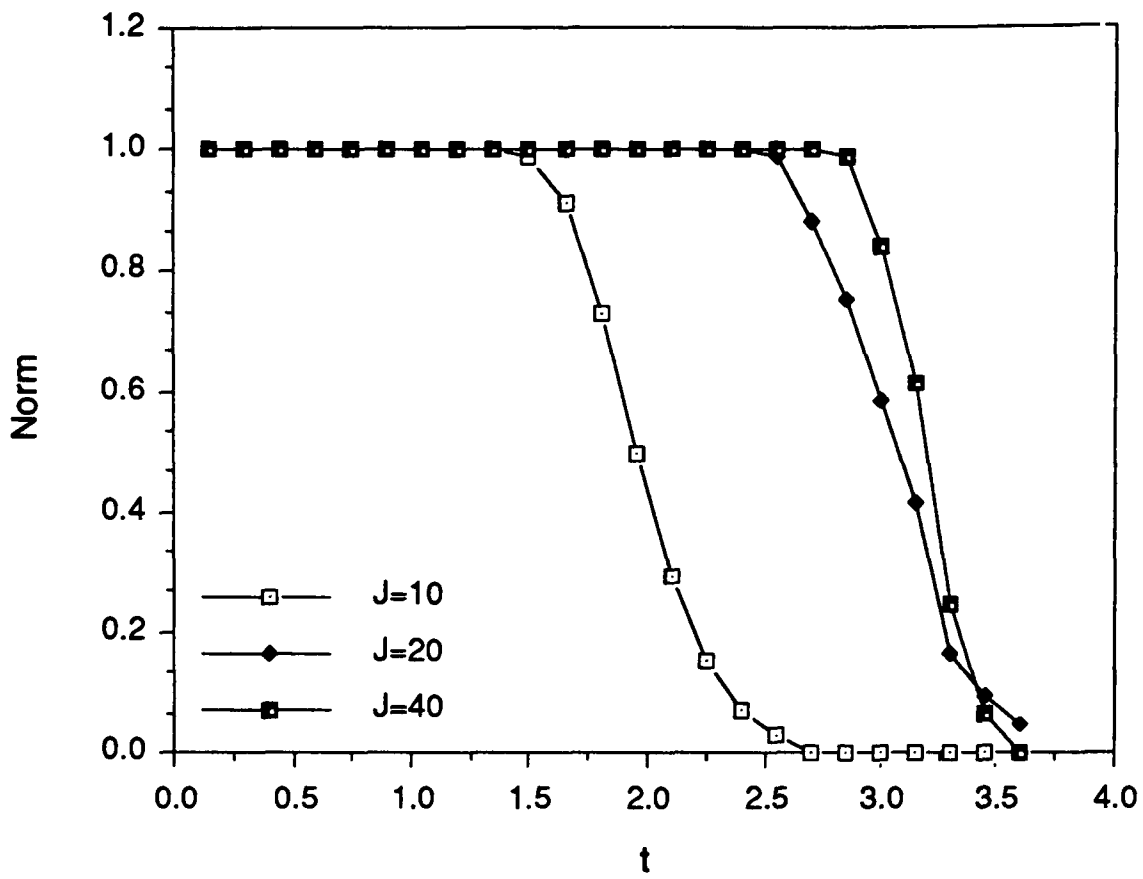


Figure 4.5: $d_h(t)$ for the Box method for $\nu = 1.5$.

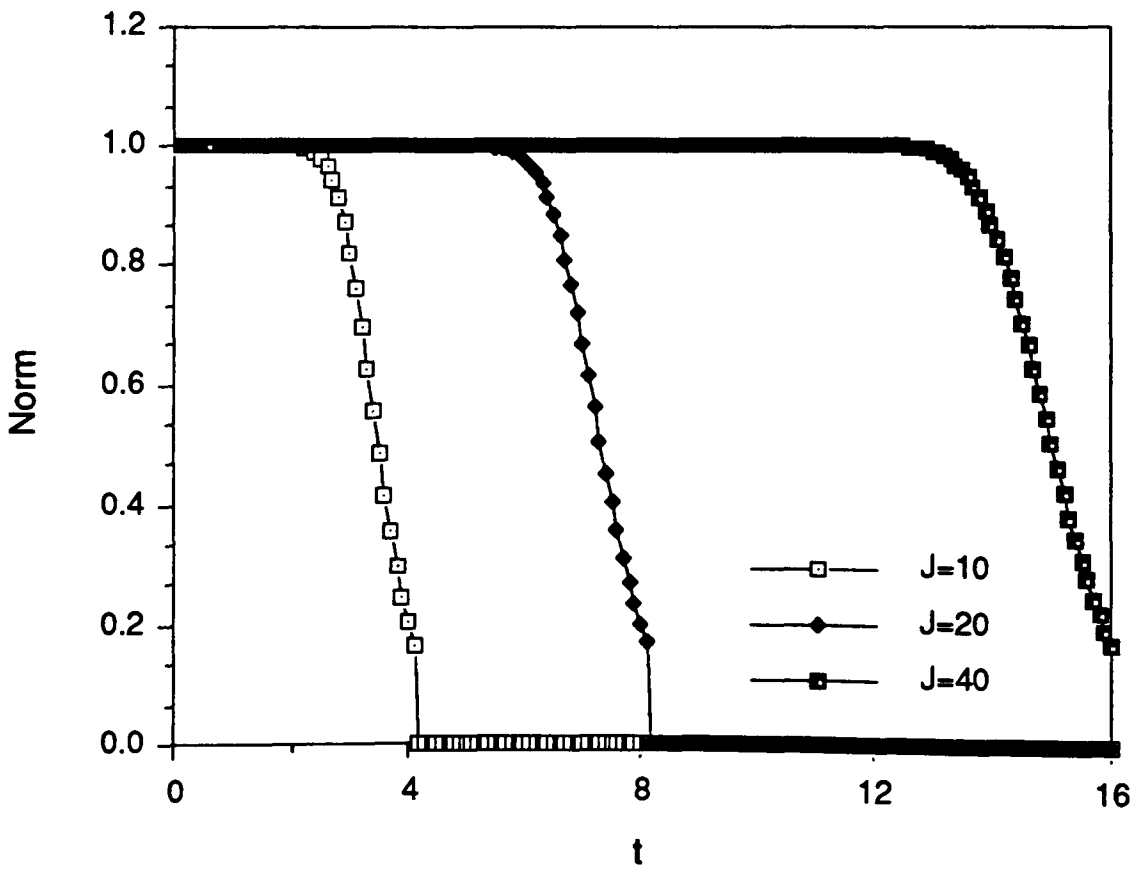


Figure 4.6: $d_h(t)$ for the Box method for $\nu = 4.0$.

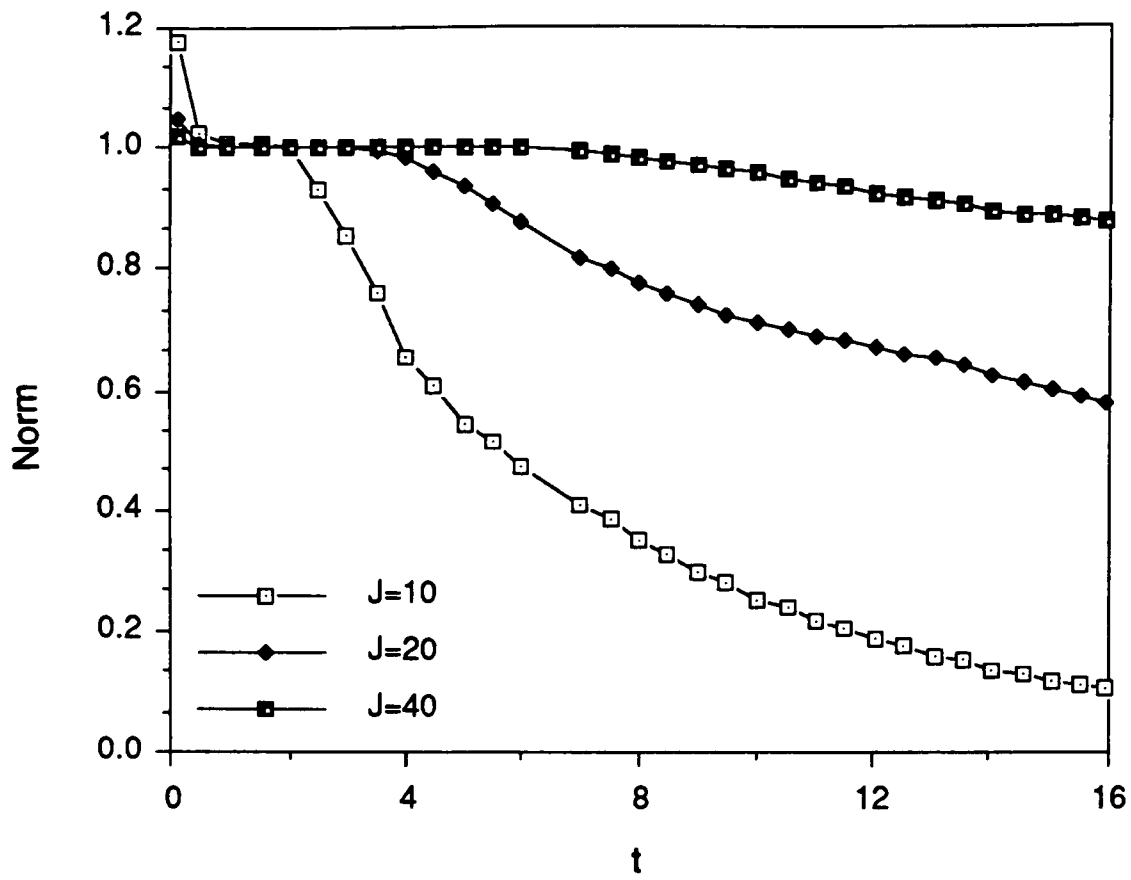


Figure 4.7: $d_h(t)$ for the CNBOX method for $\nu = 0.5$.

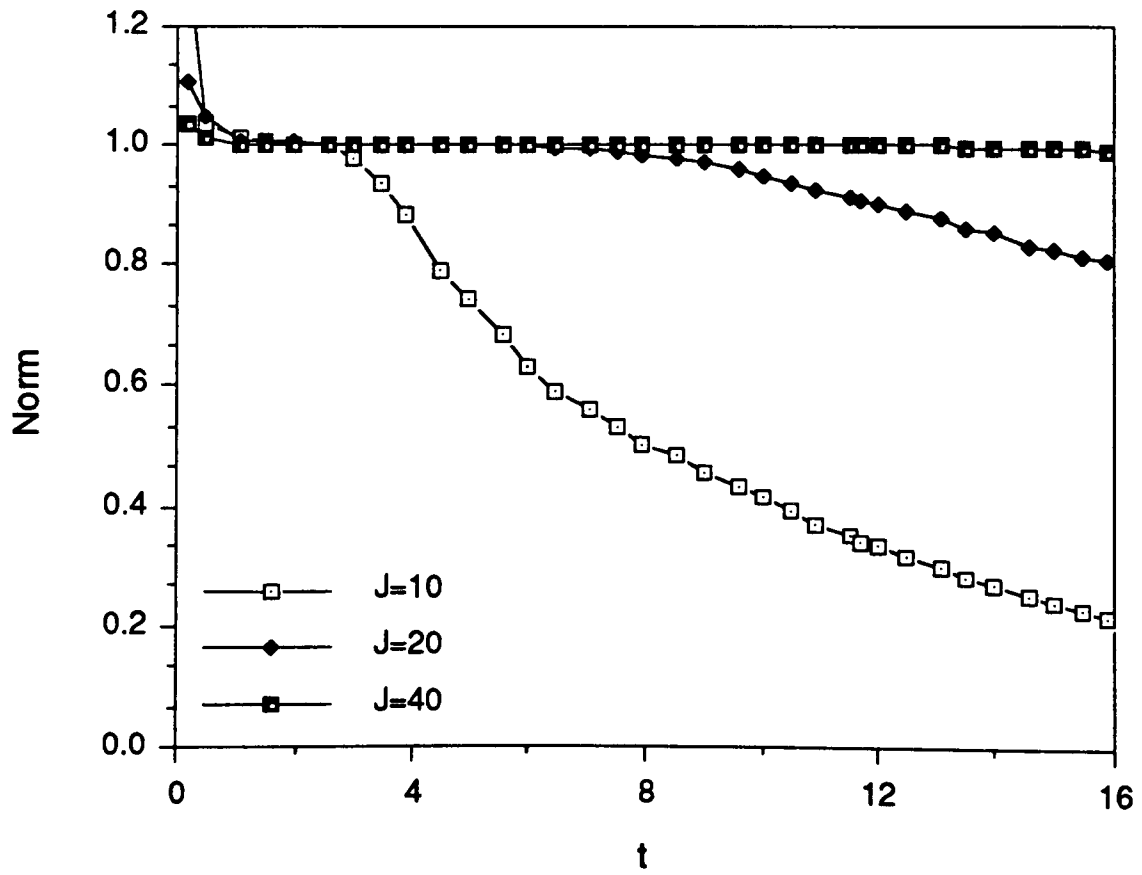


Figure 4.8: $d_h(t)$ for the CNBOX method for $\nu = 1.5$.

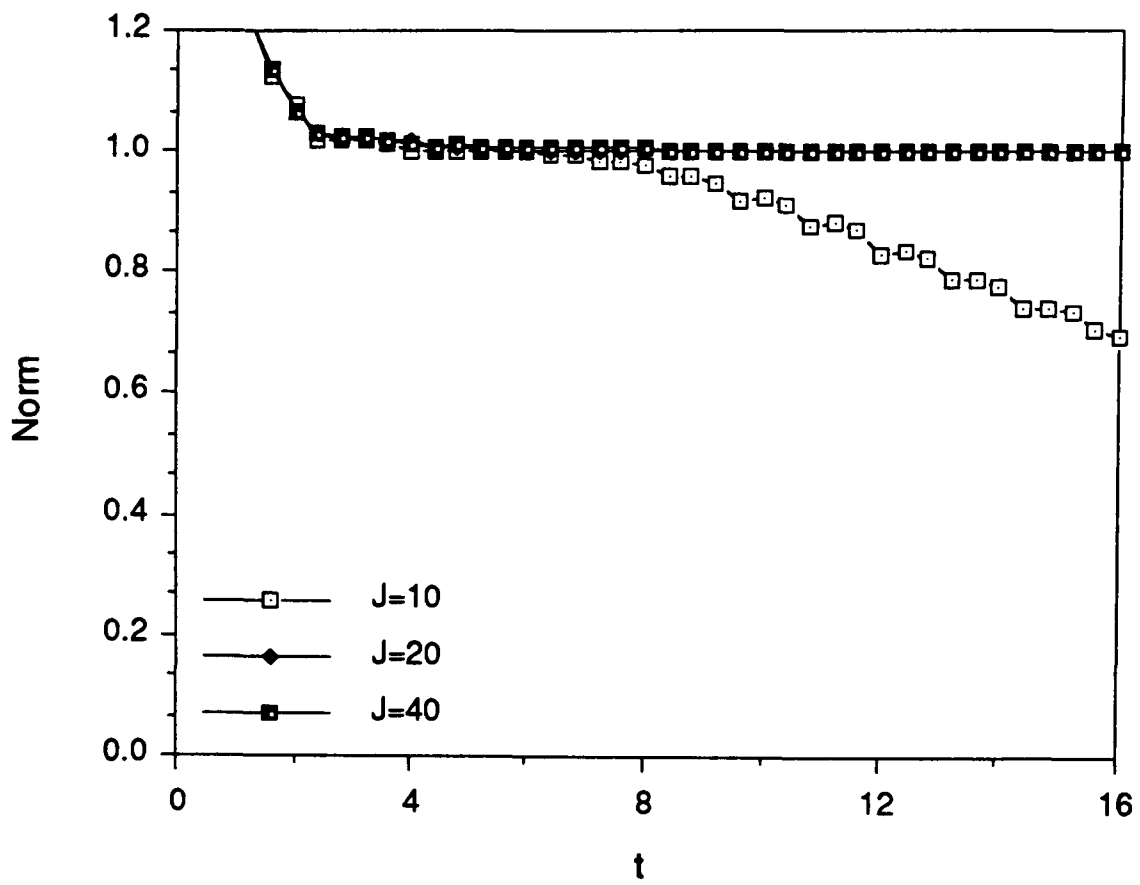


Figure 4.9: $d_h(t)$ for the CNBOX method for $\nu = 4.0$.

Recalling that $k(h) \sim h$, the exponent p in the (assumed) form of the spectral radius

$$r(G_h) = 1 - Ch^p$$

was estimated using

$$p = p(h) = \log\left[\frac{1 - r(G_h)}{1 - r(G_{h/2})}\right] / \log 2.$$

The results obtained are summarized in Tables 4.2–4.4. A value of p for $J = 20$ means that p_h was estimated with $h = 1/20$, and so forth.

Only the Box and SLHEDE methods ($\nu > 1$) had $p \leq 1$. These were also the only methods that were found empirically to be u.n.a.s. This result agrees with Proposition 4.2, and the remarks accompanying equation (4.20). Theoretical results to explain these findings will be presented later in this chapter.

□

4.3.5 P-Stability

Definitions of 'practical stability', usually called simply 'P-stability', vary considerably. Trefethen [149, page 213] says a method is P-stable when, for each fixed $h > 0$, $u_h(t)$ "admits no exponentially growing solutions as $t \to \infty$". Beam et. al. [6, page 209], Gustaffson [46] and Foreman [33] agree that a method is P-stable when it is GKS stable, and satisfies $r(G_h) \leq 1$ for each $h > 0$ [46, page 270]. This definition implies Trefethen's, but the converse does not seem to be true.

Richtmyer and Morton [120] have proposed a P-stability definition for IVPs that has been refined by Burgess [19, page 7]. Suppose that G is the semigroup generated by the

CFL	BOX	CNHE	CNDE	CNBOX	ADHE	ADDE	SLHE	SLDE	SLHEDE
$\nu = 0.5$	y	n	n	n	n	n	n	n	–
$\nu = 1.5$	y	n	n	n	n	n	–	–	y
$\nu = 4.0$	y	n	n	n	n	n	–	–	y

Table 4.1: Numerical determination of the u.n.a.s. of various methods: 'y' denotes u.n.a.s., 'n' denotes not u.n.a.s., and '–' indicates inappropriately placed boundary conditions.

h	BOX	CNHE	CNDE	CNBOX	ADHE	ADDE	SLHE	SLDE	SLHEDE
$J = 10$	0.00	2.89	2.80	3.05	2.88	3.06	2.88	2.83	–
$J = 20$	0.00	2.95	2.90	3.04	2.94	3.04	2.94	2.93	–
$J = 40$	0.00	2.97	2.95	3.02	2.97	3.03	2.97	2.97	–

Table 4.2: Numerical determination of the exponent p in $r(G_h) = 1 - Ch^p$ for $\nu = 0.5$. A dash indicates that one of the spectral radii exceeded 1, so p could not be estimated.

h	BOX	CNHE	CNDE	CNBOX	ADHE	ADDE	SLHE	SLDE	SLHEDE
$J = 10$	0.00	2.92	2.77	3.08	2.86	2.97	–	–	0.00
$J = 20$	0.00	2.95	2.89	3.05	2.94	2.97	–	–	0.00
$J = 40$	0.00	2.97	2.94	3.02	2.97	3.01	–	–	0.00

Table 4.3: Numerical determination of the exponent p in $r(G_h) = 1 - Ch^p$ for $\nu = 1.5$. A dash indicates that one of the spectral radii exceeded 1, so p could not be estimated.

h	BOX	CNHE	CNDE	CNBOX	ADHE	ADDE	SLHE	SLDE	SLHEDE
$J = 10$	0.00	2.96	2.99	3.13	2.80	2.87	–	–	0.00
$J = 20$	0.00	2.97	3.02	3.06	2.92	2.94	–	–	0.00
$J = 40$	0.00	2.98	3.01	3.03	2.96	2.99	–	–	0.00

Table 4.4: Numerical determination of the exponent p in $r(G_h) = 1 - Ch^p$ for $\nu = 4.0$. A dash indicates that one of the spectral radii exceeded 1, so p could not be estimated.

(autonomous) analytic problem. Then Burgess defines a method to be P-stable iff

$$\|u(t)\| \leq Ce^{\alpha t}\|u^0\|, \ \forall u^0 \in B \implies \|G_h(t)u_h^0\| \leq Ce^{\alpha t}\|u_h^0\|, \ \forall u_h^0 \in B_h. \tag{4.22}$$

This condition is equivalent to requiring $\beta_{G_h} \leq \beta_G$, uniformly in h. When the underlying problem has negative type, this definition implies u.n.a.s.

When $\beta_G = 0$, which can always be arranged by transforming the PDE and making corresponding adjustments to the method, this definition requires $\beta_{G_h} \leq 0$, for each $h > 0$. Thus, this definition is essentially equivalent to Trefethen's.

The definition given by Beam et. al. seems like an ungainly attempt to make the GKS theory apply over infinite periods of time — a circumstance that the theory was not originally intended to cover. Consequently, the Richtmyer and Morton/Burgess/Trefethen definition seems best. Unfortunately, this definition connects the discrete decay rate inextricably with the decay rate of the underlying analytic problem. The author believes that any such connection should be a *consequence* of "P-stability" and consistency, not a *requirement* of the "P-stability" definition. For this reason, the author believes that P-stability is best defined as u.n.a.s.

4.4 Discrete Equilibrium Solutions

The *discrete equilibrium solution* of the discrete evolution equation (4.1) is defined by

$$u_{eq,h}^n \equiv k \sum_{m=-\infty}^{n-1} G_h^{n-m-1} f_h^m = k \sum_{r=0}^{\infty} G_h^r f_h^{n-1-r}, \quad n \in \mathcal{Z}, \tag{4.23}$$

whenever the sum converges. Note that a necessary and sufficient condition for the convergence of the sum, for all $f_h \in \ell_k^\infty(\Re, B_h)$, is the non-modified von Neumann condition, $r(G_h) < 1$ for each $h > 0$. When this condition holds, $u_{eq,h} \in \ell_k^\infty(\Re, B_h)$.

The *discrete transient* of the discrete IVP given by equations (4.1,4.2) is defined as

$$u_{trans,h}^n \equiv u_h^n - u_{eq,h}^n = G_h^n \left[u_h^0 - u_{eq,h}^0\right], \quad n \geq 0. \tag{4.24}$$

If $r(G_h) < 1$ for each fixed $h > 0$, then the transient decays as $n \to \infty$.

4.5 BVP Stability of Equilibrium Solutions

It is clear that for any bounded forcing $f_h \in Y_h$, we can define an equilibrium solution $u_{eq,h} \in X_h$. Thus an equilibrium solution operator

$$F_h : Y_h \to X_h,\ F_h f_h \equiv u_{eq,h} \tag{4.25}$$

may be defined. The method is said to be *BVP stable* when $\{F_h\}_{h>0}$ is uniformly bounded in $L(Y_h, X_h)$.

4.5.1 BVP Stability of Periodic and Steady Equilibrium Solutions

Consider an autonomous linear problem with T-periodic forcing. Thus, $f_h^{n+N_h} = f_h^n$, for each $n \in \mathcal{Z}$, where $T = N_h k(h)$. Define $P_{T,k}$ as

$$P_{T,k} \equiv \left\{ \{u_h^n\}_{n\in\mathcal{Z}} : u_h^n \in B_h,\ u_h^{n+N_h} = u_h^n,\ \forall n \in \mathcal{Z} \right\},$$

with the 2-norm over one period. As in Section 3.6.1, it is assumed that B_h is a seperable Hilbert space.

This section proves that the equilibrium solution operator satisfies $F_h \in L(Y_h, X_h)$, and that $\|F_h\|_{L(Y_h,X_h)} \le C$ uniformly in h, when

$$X_h = Y_h = P_{T,k}, \tag{4.26}$$

provided that the method satisfies the modified von Neumann condition. When the norm estimates are available, the method is said to be *BVP stable* in the norm $P_{T,k}$, or simply $P_{T,k}$-stable.

Define the discrete map at a period in $L(B_h)$ by

$$S_{T,h} \equiv S_h(N_h, 0). \tag{4.27}$$

The same reasoning as in the previous chapter shows that the sequence $\{u_h^n\}$ is N_h-periodic iff u_h^0 is a fixed point of the map at a period $S_{T,h}$. Similar reasoning yields the following result. (Foreman's paper [33] contains comments that are similar in spirit to this result.) Since h does not play a part in these results, h subscripts are suppressed wherever possible.

Proposition 4.4 *Suppose that* $1 \in \rho(G_h^{N_h})$, *and* $X_h = Y_h = P_{T,k}$.

(i) For $f_h \in Y_h$, *there is unique initial data* $u_{T,h} \in B_h$ *such that* $u_h \in X_h$.

(ii) This initial data is given by

$$u_h^0 = \Gamma_{T,h} f_h, \tag{4.28}$$

where $\Gamma_{T,h} : Y_h \to B_h$ *is defined by*

$$\Gamma_{T,h} f_h \equiv k \left[I - G_h^{N_h} \right]^{-1} \sum_{m=0}^{N_h-1} G_h^{N_h-m-1} f_h^m. \tag{4.29}$$

(iii) The solution $u_h \in X_h$ *that arises from choosing* $u^0 = \Gamma_{T,h} f_h$ *is the discrete equilibrium solution.*

Proof
Note that, if $1 \in \rho(G^N)$, then u^0, as defined by equation (4.28), is well defined in B, and is a fixed point of the map at the period. This shows that $u \in P_T$. Furthermore, u is the equilibrium solution, because

$$\begin{aligned} u^0_{eq} &= k\sum_{r=0}^{\infty} G^r f^{-r-1} \\ &= k\sum_{q=0}^{\infty}\sum_{s=0}^{N-1} G^{(qN+s)} f^{-s-1} \\ &= k\sum_{q=0}^{\infty} \left(G^N\right)^q \sum_{m=0}^{N-1} G^{N-m-1} f^m, \quad m = N-s-1 \\ &= k\left[I - G^N\right]^{-1} \sum_{m=0}^{N-1} G^{N-m-1} f^m \\ &= \Gamma_T f. \end{aligned}$$

□

Note that $\Gamma_{T,h}$ maps the forcing f_h onto the equilibrium initial data, which in turn defines the equilibrium solution. For initial data other than $\Gamma_{T,h} f_h$, the solution can be decomposed into periodic (equilibrium) and transient components.

Fourier Forcing
The following result concerns the important special case of Fourier forcing — which naturally includes steady forcing.

Proposition 4.5 *Suppose that* $1 \in \rho(G_h^{N_h}$ *and that* $\eta \in \Re$ *is such that* $e^{i\eta} \in \rho(G_h)$. *Let* $\psi_h \in B_h$, *and* $f_h^n = e^{in\eta}\psi_h$. *Then problem (4.1) has a periodic solution (with period* $T = 2\pi/\eta$*) iff*

$$u_h^0 = \Gamma_{T,h} f_h = k\left(e^{i\eta} - G_h\right)^{-1} \psi_h.$$

For this initial data, the solution is $u_h^n = e^{in\eta} u_h^0$.

Proof
Since $I - G^N$ is invertible the unique initial data that results in a T-periodic solution is

$$\begin{aligned} \Gamma_T f &\equiv k\left[I - G^N\right]^{-1} \sum_{m=0}^{N-1} G^{N-m-1} e^{im\eta}\psi \\ &= k\left[I - G^N\right]^{-1}\left[I - G^N\right]\left[e^{i\eta} - G\right]^{-1}\psi \\ &= k\left[e^{i\eta} - G\right]^{-1}\psi. \end{aligned}$$

Furthermore, by equation (4.5),

$$\begin{aligned} u^n &= G^n u^0 + k\sum_{m=0}^{n-1} G^{n-m-1} f^m \\ &= kG^n\left[e^{i\eta} - G\right]^{-1}\psi + k\sum_{m=0}^{n-1} G^{n-m-1} e^{im\eta}\psi \end{aligned}$$

$$\begin{aligned} &= e^{in\eta} k \left[e^{i\eta} - G\right]^{-1} \psi \\ &= e^{in\eta} u^0. \end{aligned}$$

□

General Periodic Forcing

A general T-periodically forced problem can be analysed as follows. Suppose that the forcing $f_h \in Y_h$. The (temporal) Fourier components of f_h are

$$\hat{f}_h = \left[\hat{f}_h^0, \ldots, \hat{f}_h^{N_h-1}\right] \in \hat{Y}_h,$$

where $\hat{Y}_h = \ell^2_k([0,T], B_h)$. These Fourier components are given by

$$\hat{f}_h^m = \sum_{n=0}^{N_h-1} f_h^n e^{-imn\eta_h}, \quad \eta_h \equiv \frac{2\pi}{N_h}. \tag{4.30}$$

Similarly, when $u_h \in X_h$, it has Fourier components $\hat{u}_h \in \hat{X}_h \equiv \hat{Y}_h$. Of course, the discrete Fourier inversion result is

$$f_h^n = \frac{1}{N_h} \sum_{m=0}^{N_h-1} \hat{f}_h^m e^{imn\eta_h}, \quad n \in \mathcal{Z},$$

and the discrete Parseval relation

$$\|f_h\|_{Y_h} = \sqrt{T} \|\hat{f}_h\|_{\hat{Y}_h}.$$

It is clear from Proposition 4.5 that when $f_h \in Y_h$, the T-periodic equilibrium solution is given by

$$u_h^n = k \sum_{m=0}^{N_h-1} \left(e^{im\eta_h} - G_h\right)^{-1} \hat{f}_h^m e^{imn\eta_h}.$$

By inspection,

$$\hat{u}_h^m = \hat{F}_h^m \hat{f}_h^m \equiv k \left(e^{im\eta_h} - G_h\right)^{-1} \hat{f}_h^m, \quad m = 0, \ldots, N_h - 1.$$

Note that $\hat{F}_h^m \in L(B_h)$ for each m. Define $\hat{F}_h : \hat{Y}_h \to \hat{X}_h$ via

$$\hat{F}_h = \operatorname{diag}\left[\hat{F}_h^0, \ldots, \hat{F}_h^{N_h-1}\right].$$

Bearing in mind that $X_h = Y_h = P_{T,k}$, Parseval's relation shows that $\|\hat{F}_h\|_{L(\hat{Y}_h, \hat{X}_h)} = \|F_h\|_{L(Y_h, X_h)}$, so

$$\|F_h\|_{L(Y_h, X_h)} = \max_{m=0,\ldots,N_h-1} \|\hat{F}_h^m\|.$$

Since $P_{T,k}$-stability means that $\{F_h\}_{h>0}$ must be uniformly bounded in $L(Y_h, X_h)$, with $X_h = Y_h = P_{T,k}$, we have obtained the following result (c.f. Section 3.6.1).

Proposition 4.6 *For each h let B_h be a seperable Hilbert space. Define*

$$S_h \equiv \{e^{2m\pi i/N_h} : \ m = 0, \ldots, N_h - 1\}.$$

Then a method is $P_{T,k}$-stable iff $S_h \subseteq \rho(G_h)$ for all $h > 0$, and

$$\sup_{h>0} \sup_{\lambda \in S_h} k(h) \|(\lambda - G_h)^{-1}\| \leq C < \infty. \tag{4.31}$$

In particular, power-boundedness and the modified VNC are sufficient for $P_{T,k}$-stability.

Proof
See proposition 4.10.

□

Next we consider some examples for which the spectrum of G_h is easily found.

EXAMPLE: IVPS AND PBCPS FOR UNITARY SCHEMES

Consider a unitary scheme with symbol $\lambda(z, v)$. If this scheme is used for an IVP, then, since the left and right shift operators are mutual inverses in $L(\ell_h^2(\Re))$, and $\sigma(E_h) = \{z : |z| = 1\}$ for any $h > 0$, the spectral mapping theorem shows that $\sigma(G_h) = \{\lambda(z, \nu) : |z| = 1\}$. Since the scheme is unitary,

$$\sigma(G_h) = \{z : |z| = 1\};$$

so, by Proposition 4.6, the T-periodically forced problem is not $P_{T,k}$-stable.

For PBCPs, the calculation of the spectral radius of G_h is identical, except that now $\sigma(E_h) = \{z : z = \exp(2\pi ij/J),\ j = 0, \ldots, J-1\}$, so

$$\sigma(G_h) = \left\{\lambda(z, \nu) : z = e^{2\pi ij/J},\ j = 0, \ldots, J-1\right\}.$$

Thus, again, $r(G_h) = 1$. If we suppose the unitary scheme to be consistent, then $\lambda(1, \nu) = 1$. Since $1 \in \sigma(E_h)$, this implies that $1 \in \sigma(G_h)$; so, by Proposition 4.6, the periodically forced counterpart of this problem is not $P_{T,k}$-stable.

□

4.5.2 BVP Stability

A u.n.a.s. method is BVP stable, even for non-periodic forcing.

Proposition 4.7 *Let $X_h = Y_h = \ell_k^\infty(\Re, B_h)$. Then u.n.a.s. implies BVP stability in the norm X_h.*

Proof
We must show that the equilibrium solution operators $F_h : Y_h \to X_h$, defined by equation (4.25), are uniformly bounded in $L(Y_h, X_h)$. Now,

$$\begin{aligned}
\|F_h f_h\|_{X_h} &= \max_{n \in \mathcal{Z}} \|u_{eq,h}^n\|_{B_h} \\
&= \max_{n \in \mathcal{Z}} k \| \sum_{r=0}^{\infty} G_h^r f_h^{n-r-1} \| \\
&\leq \left(k \sum_{r=0}^{\infty} \|G_h^r\| \right) \|f_h\|_{Y_h}.
\end{aligned}$$

By Proposition 4.3, the quantity in brackets is bounded in h, which proves BVP stability.

□

4.6 Uniform Stability

In this section we investigate the stability of solutions of discrete IBVPs

$$u_h^{n+1} = G_h u_h^n + k f_h^n, \quad n \geq 0, \quad u_h^0 = \phi_h. \tag{4.32}$$

This differs from the previous IVP and BVP stability investigations, since the solution is now required to depend continuously on both the initial data and the forcing.

Write the method in the standard form $Q_h u_h = g_h$, where

$$g_h = \left[\phi_h, f_h^0, f_h^1, \ldots\right], \quad u_h = \left[u_h^0, u_h^1, u_h^2, \ldots\right],$$

and suppose that $Q_h : X_h \to Y_h$ is linear. (X_h, Y_h)-stability means that the collection $\{Q_h : h > 0\}$ is uniformly invertible in $L(X_h, Y_h)$. This terminology should distinguish IBVP stability from the IVP stability concepts discussed earlier.

Extended Lax–Richtmyer stability has already been discussed in Chapter 2. Recall that it employs the norms

$$\|u_h\|_{X_h} = \max_{m \geq 0} \|u_h^m\|_{B_h}, \text{ and} \tag{4.33}$$

$$\|g_h\|_{Y_h} = \|g_h^0\|_{B_h} + k \sum_{m=1}^{\infty} \|g_h^m\|_{B_h}, \tag{4.34}$$

where, naturally, $k = k(h)$. This form of stability is unsuitable for equilibrium computations. In particular, the space Y_h does not admit non-decaying forcing functions.

We propose enlarging Y_h by setting it equal to X_h: thus both X_h and Y_h have the B_h norm in space and the ℓ_k^∞ norm in time. This choice is the discrete analogue of the norm employed for the well-posedness study of the previous chapter, and results in a new stability condition. For brevity, (X_h, Y_h)-stability, in the new norms, will be called ***uniform stability***. (X_h, Y_h)-stable methods will be called ***uniformly stable***. Observe that uniform stability is the most stringent of all possible $(\ell_k^\infty, \ell_k^p)$ stabilities $(1 \leq p \leq \infty)$.

There are two parts to proving uniform stability. First, the existence of Q_h^{-1} in $L(X_h)$ for *each fixed* $h > 0$ must be shown. Secondly, it must be shown that $\|Q_h^{-1}\| \leq C$ uniformly in h. The following Proposition addresses the first part.

Proposition 4.8 *Let* $h > 0$ *be fixed, and suppose that* $\dim(B_h) < \infty$. *Then* Q_h *is invertible in* $L(X_h)$ *iff* $r(G_h) < 1$.

Proof

Let $h > 0$ be arbitrary. Since h is fixed, it will be omitted in the notation that follows. The formal inverse of Q is

$$\left(Q^{-1}g\right)^n = G^n g^0 + k \sum_{r=1}^{n} G^{n-r} g^r.$$

Suppose that Q is invertible in $L(X)$, but that $r(G) = 1$. (The case of $r(G) > 1$ can be dealt with in a similar way.) Since $\dim(B) < \infty$, there exists $\eta \in \Re$ and $\phi \in B$ such that $G\phi = e^{i\eta}\phi$. Choose $g^0 = \phi$, and $g^m = e^{im\eta}\phi/k$ for $m \geq 1$. Clearly $g \in X$, and

$$\left(Q^{-1}g\right)^n = (n+1)e^{in\eta}\phi.$$

Thus $Q^{-1}g \notin X$, which contradicts the asumption that $Q^{-1} \in L(X)$.

Conversely, suppose that $r(G) < 1$. Then, for $g \in X$,

$$\| \left(Q^{-1}g\right)^n \| \leq \|G^n\|\|g\| + k\|g\| \sum_{r=0}^{\infty} \|G^r\| \equiv C\|g\|_X,$$

since the series converges absolutely [62]. Thus $Q^{-1}g \in X$. Furthermore, since C is independent of g, $Q^{-1} \in L(X)$.

□

The assumption that $\dim(B_h) < \infty$ is not a real restriction, as the following argument shows. Our concern is with the invertibility of $e^{i\eta} - G_h$. Simonenko's result (Theorem 2.6) shows that $e^{i\eta} - G_h$ cannot be invertible on the half line when $e^{i\eta} - \lambda(z, \nu)$ vanishes for $|z| = 1$. Thus, no unitary scheme can have $e^{i\eta} - G_h$ invertible on the half line. Consequently, no unitary scheme can be stable on an unbounded domain Ω. Thus it is reasonable to assume that Ω is bounded, and, in practice, this implies $\dim(B_h) < \infty$.

Next the variation of $\|Q_h^{-1}\|$ with h is considered.

Proposition 4.9 *Suppose that* $\dim(B_h) < \infty$ *for each* $h > 0$. *Then the following conditions are necessary for uniform stability:*

(i) $r(G_h) < 1$ *for each* $h > 0$,

(ii) there exists $C > 0$, *independent of* $\eta \in \Re$, $h > 0$, *such that*

$$\|(e^{i\eta} - G_h)^{-1}\|_{L(B_h)} \leq Ck(h)^{-1}, \quad \forall h > 0, \tag{4.35}$$

(iii) power-boundedness.

Proof

The necessity of condition (i) was proved in the previous Proposition. We now show the necessity of condition (ii). Let $\eta \in \Re$ be arbitrary. For any non-zero $\psi_h \in B_h$, choose

$$f_h^m = e^{im\eta}\psi_h, \quad u_h^0 = 0.$$

The solution is easily shown to be

$$u_h^n = \left(e^{in\eta} - G_h^n\right)\phi_h, \quad n \geq 0,$$

where

$$\phi_h \equiv k\left(e^{i\eta} - G_h\right)^{-1}\psi_h.$$

Recalling that $g_h = [u_h^0, f_h^0, f_h^1, \ldots]$, clearly

$$\begin{aligned} \|g_h\|_{X_h} &= \max\left\{\|u_h^0\|, \|f_h^0\|, \|f_h^1\|, \ldots\right\} = \|\psi_h\|, \\ \|u_h\|_{X_h} &\geq \|e^{in\eta}\phi_h - G_h^n\phi_h\| \\ &\geq \left| \|\phi_h\| - \|G_h^n\phi_h\| \right|, \end{aligned}$$

for any n. Since $r(G_h) < 1$, we can choose n large enough that $\|G_h^n\phi_h\| \leq \frac{1}{2}\|\phi_h\|$. Thus, for n sufficiently large,

$$\left| \|\phi_h\| - \|G_h^n\phi_h\| \right| \geq \frac{1}{2}\|\phi_h\| \implies \|u_h\|_{X_h} \geq \frac{1}{2}\|\phi_h\|.$$

Hence,

$$\|Q_h^{-1}\|_{L(X_h)} \geq \frac{\|u_h\|_{X_h}}{\|g_h\|_{X_h}} \geq \frac{1}{2}\frac{\|\phi_h\|}{\|\psi_h\|}.$$

Taking the supremum over $\psi_h \in B_h$ gives

$$\|Q_h^{-1}\|_{L(X_h)} \geq \frac{k}{2}\|(e^{i\eta} - G_h)^{-1}\|_{L(B_h)}. \tag{4.36}$$

Thus uniform stability implies that there exists C independent of h, η such that

$$k\|(e^{i\eta} - G_h)^{-1}\|_{L(B_h)} \leq C,$$

which proves the necessity of condition (ii).

To prove the necessity of condition (iii) we proceed as before, but choose $u_h^0 = \psi_h$, where $\psi_h \in B_h$ and $f_h^m = 0$. Then $u_h^n = G_h^n u_h^0$, so that

$$\begin{aligned} \|g_h\|_{X_h} &= \max\{\|u_h^0\|, \|f_h^0\|, \|f_h^1\|, \ldots\} = \|\psi_h\|, \\ \|u_h\|_{X_h} &\geq \|G_h^n \psi_h\|, \end{aligned}$$

for any fixed n. Thus

$$\|Q_h^{-1}\|_{L(X_h)} \geq \frac{\|u_h\|_{X_h}}{\|g_h\|_{X_h}} \geq \frac{\|G_h^n \psi_h\|}{\|\psi_h\|}.$$

Taking the supremum over $\psi_h \in B_h$ shows that $\|Q_h^{-1}\| \geq \|G_h^n\|$, for any n. Thus uniform stability implies power-boundedness.

□

We shall call condition (4.35) the *near-uniform invertibility condition.* It should be compared with the weaker Godunov–Ryanbenkii condition, which requires that $(\lambda - G_h)$ be uniformly invertible for each $|\lambda| > 1$. It should also be compared with the sufficient condition given by Godunov and Ryabenkii for the 'stability' of iterative methods for steady problems [38, page 458]. This condition requires that $(\lambda - G_h)$ be uniformly invertible for each $|\lambda| \geq 1$, and does not admit most methods for HPDEs. It should also be compared with the Kreiss resolvent condition [135, page 398][120, page 75][99].

The near-uniform invertibility condition has a clear interpretation for equilibrium solutions of forced evolutionary problems. With Fourier forcing $f_h^n = e^{in\eta}\psi_h$, the equilibrium solution has the asymptotic representation $u_h^n \sim k(e^{i\eta} - G_h)^{-1}\psi_h e^{in\eta}$. If the equilibrium solution is to depend continuously on the forcing, then $k(e^{i\eta} - G_h)^{-1}\psi_h$ must depend continuously on ψ_h, with an h-uniform bound. Thus $\|(e^{i\eta} - G_h)^{-1}\| = O(k^{-1})$ as $h \to 0$.

General forcing results in a superposition over η of such representations. If the forcing is time-smooth, then as $k \to 0$, the η-spectrum will bunch about zero. Thus, for weakly unstable methods, as the mesh is refined, there is a diminishing amount of energy at the critical frequency, which may be amplified by an increasing amount $\|k(e^{i\eta} - G_h)^{-1}\phi_h\|$ — provided that ϕ_h achieves the operator norm.

For smooth forcing, energy at the critical frequency will diminish so quickly that only a loss of accuracy appears. However, unsmooth forcing may actually result in non-convergence. This is discussed in the next chapter.

Proposition 4.10 *For a power bounded method, the near-uniform invertibility condition is equivalent to the modified von Neumann condition.*

Proof
It is well known that $\|(\lambda - T)^{-1}\| \geq 1/\text{dist}(\lambda, \sigma(T))$ for $\lambda \in \rho(T)$. Suppose that the near-uniform invertibility condition holds. Then $e^{i\eta} \in \rho(G_h)$ for any $h > 0$ and $\eta \in \Re$, and

$$\frac{1}{\text{dist}(e^{i\eta}, \sigma(G_h))} \leq \|(e^{i\eta} - G_h)^{-1}\|_{L(B_h)} \leq Ck^{-1}.$$

Taking the supremum over $\eta \in \Re$, and recalling that C is independent of η, gives

$$\frac{1}{1 - r(G_h)} \leq Ck^{-1} \implies r(G_h) \leq 1 - C'k(h),$$

where $C' = 1/C$. This implies the modified von Neumann condition.

Conversely, suppose that the modified von Neumann condition holds. Then $e^{i\eta} \in \rho(G_h)$. Also, by power-boundedness, there exists $M > 0$ and $\beta < 0$ such that $\|G_h^m\| \leq Me^{\beta mk}$. Thus, for any $\eta \in \Re$,

$$\begin{aligned} k\|(e^{i\eta} - G_h)^{-1}\| &= k\|e^{-i\eta} \sum_{m=0}^{\infty} e^{-im\eta} G_h^m\| \\ &\leq Mk \sum_{m=0}^{\infty} e^{\beta mk} \\ &= \frac{Mk}{1 - e^{\beta k}} \leq C, \end{aligned}$$

for k sufficiently small. Since C is independent of h and η, the near-uniform invertibility condition holds.

□

The following result is a discrete analogue of Proposition 3.4 of Chapter 3, which showed the equivalence of u.a.s. and well-posedness for 'well-behaved' semigroups.

Proposition 4.11 *u.n.a.s. is equivalent to uniform stability.*

Proof
Suppose that a method is uniformly stable. Then, by Proposition 4.9, it is power bounded, and satisfies the near-uniform invertibility condition. Hence, by Proposition 4.10, it is power bounded and satisfies the modified von Neumann condition. By Proposition 4.2, the method is u.n.a.s.

Conversely, assume the method is u.n.a.s. Then, by the same argument as Proposition 4.3,

$$\begin{aligned} \|u_h^n\|_{B_h} &\leq Me^{\beta nk}\|u_h^0\|_{B_h} + Mk \max_{m \geq 0} \|f_h^m\|_{B_h} \sum_{m=0}^{\infty} e^{\beta nk} \\ &\leq Me^{\beta nk}\|u_h^0\|_{B_h} + C \max_{m \geq 0} \|f_h^m\|_{B_h}. \end{aligned}$$

Thus,

$$\|u_h\|_{X_h} \leq (C + M) \max\{\max_{n \geq 0} \|f_h^n\|, \|u_h^0\|\} = (C + M)\|g_h\|_{Y_h},$$

where $g_h = [u_h^0, f_h^0, f_h^1, \ldots]$. This implies uniform stability.

□

4.7 Checking the Near-Uniform Invertibility Condition

Proposition 4.11 shows that the most useful form of IVP stability, u.n.a.s., is equivalent to uniform stability. Proposition 4.2 shows that both are, in turn, equivalent to

(i) power-boundedness, and

(ii) the modified von Neumann condition.

Furthermore, by Proposition 4.7, the second condition is *sufficient* for the BVP stability of equilibrium solutions of forced problems.

Unfortunately, properties (i) and (ii) are hard to verify. Even checking the modified von Neumann condition empirically for the scalar linear advection equation requires substantial computations of the type presented in Section 4.3.4. The power-boundedness condition has been studied by many authors [101,99,135,120], and results are available for its verification.

However, Proposition 4.10 shows that, for power bounded schemes, the modified von Neumann condition is equivalent to the near-uniform invertibility condition. Thus, for the remainder of this chapter, our programme is to *assume* that power-boundedness can be checked in some fashion, and concentrate on finding techniques for ensuring the the near-uniform invertibility condition.

In particular, simple *necessary* conditions for the near-uniform invertibility condition are sought. Since our aim is to isolate *schemes* which are potentially efficient for the computation of equilibrium solutions, these conditions should be verifiable from the scheme's stencil alone, and should be independent of the boundary conditions used in the final method. Of course, a 'good' scheme may result in a poor *method* if inappropriate boundary conditions are used. However, we shall not investigate this problem here.

Before proceeding, observe that the near-uniform invertibility condition is rather inconvenient for implicit methods, as it requires analysis of $G_h = S_h^{-1}T_h$. However, the near-uniform invertibility condition implies

$$\|(S_h e^{i\eta} - T_h)^{-1}\| \leq \|S_h^{-1}\|\|(e^{i\eta} - G_h)^{-1}\| \leq \|S_h^{-1}\|Ck^{-1}.$$

Thus, when the implicit operator is uniformly invertible, a necessary condition for the near-uniform invertibility condition is

$$\|(S_h e^{i\eta} - T_h)^{-1}\| \leq Ck^{-1}. \tag{4.37}$$

This condition is also sufficient where $\|S_h\| \leq C$, and has the advantage that $S_h e^{i\eta} - T_h$ is Toeplitz.

4.7.1 Near-Uniform Invertibility of One-Dimensional Toeplitz Matrices

Consider the family of one-dimensional Toeplitz operators $\{A_h\}_{h>0}$, with (fixed order) symbol

$$A(z) \equiv \sum_{k=-l}^{r} a_k z^k. \tag{4.38}$$

The coefficients a_k, $k = -l, \ldots, r$ are supposed constant. Then

$$A_h \equiv A(E_h) = \sum_{k=-l}^{r} a_k E_h^k, \quad h > 0, \tag{4.39}$$

where E_h^k are shift operators.

The uniform invertibility of such operators has already been discussed in some detail in Chapter 2, in Theorems 2.3 and 2.5. However, our aim is to verify the near-uniform invertibility condition. This suggests an investigation of the growth of $\|A_h^{-1}\|$ when the conditions of the uniform invertibility Theorems are contravened.

Proposition 4.12 *Suppose that the roots of the polynomial $z^l A(z) = 0$ are so disposed that:*

(i) m_l roots lie inside the unit circle,

(ii) m_e roots lie on the unit circle, with maximum algebraic multiplicity q, and

(iii) m_r roots lie outside the unit circle,

where $m_l + m_e + m_r = l + r$. Then:

(iv) if $m_l > l$ or $m_r > r$, there exists $\rho > 1$: $\|A_h^{-1}\| \geq C\rho^J$,

(v) if $m_l \leq l$ and $m_r \leq r$, then $\|A_h^{-1}\| \geq Ch^{-q}$,

where all norms are $L(\ell_h^2(I))$ norms and $J = 1/h$.

Proof

See reference [38, page 43].

□

We see that we need at least one left-hand boundary condition for each root inside the unit circle, and at least one right-hand boundary condition for each root outside the unit circle; otherwise $\|A_h^{-1}\|$ grows at a catastrophic rate. The boundary conditions associated with roots on the unit circle may be assigned to *either* end of the interval; but the slowest possible growth of $\|A_h^{-1}\|$ is as h^{-q}. This result may be paraphrased as there being m_e centres that are 'nearly' proper. When the family of matrices $\{A_h\}_{h>0}$ satisfy the bound $\|A_h^{-1}\| \leq Ch^{-q}$, the family is said to be *nearly uniformly invertible of order q*. When $q = 1$, we simply say that the family is *nearly uniformly invertible.*

For hyperbolic problems, $k(h) = Ch$. Thus, a method satisfies the near-uniform invertibility condition when the family of Toeplitz operators $\{S_h e^{i\eta} - T_h\}_{h>0}$ is nearly uniformly invertible (of order 1). For parabolic problems, we would be interested in the near-uniform invertibility of order 2 of the family $\{S_h e^{i\eta} - T_h\}_{h>0}$.

When $m_e = 0$ and $q = 0$, Strang's Theorem (Theorem 2.3) shows that this result is the best possible. Note that the result cannot be extended to $\Omega = \Re^+$ or $\Re$; Theorem 2.6 shows that when $m_e \neq 0$, for any fixed h, A_h cannot be invertible in $L(\ell_h^2(\Omega))$.

For hyperbolic problems, it is extremely rare to have *homogeneous* boundary conditions at both ends of the computational interval. We want to investigate the growth of $\|S_h^{-1}\|$, where

$$S_h = A_h + P_h. \tag{4.40}$$

Here $P_h \in L(\ell_h^2(I))$ is a perturbation of the (homogeneous boundary condition) Toeplitz matrix A_h. The crucial assumption we shall make about P_h is that it has the form

$$P_h = \begin{bmatrix} L & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & R \end{bmatrix},$$

where L and R are rectangular matrices of fixed, h-independent size.

When the pole condition does not hold, we have the following analogue of Proposition 4.12, which proves that inhomogeneous boundary conditions cannot 'cure' any growth in the inverse of the underlying Toeplitz operator.

Proposition 4.13 *Let the roots of $z^l A(z)$ be as in Proposition 4.12. If*

(i) If $\text{rank}(L) < m_l$ *or* $\text{rank}(R) < m_r$ *then* $\|S_h^{-1}\| \geq \rho^J$ *for some* $\rho > 1$.

(ii) If $\text{rank}(L) \geq m_l$ *and* $\text{rank}(R) \geq m_r$ *then* $\|S_h^{-1}\| \geq Ch^{-q}$ *for some* $C > 0$.

Proof
See reference [38, page 43].

□

Notice that, even when $q > 1$, the boundary conditions can be chosen to ensure that $S_h u_h = 0$ has only the zero solution. That is, when $f = 0$, the boundary conditions can be chosen to control the growing modes. However, for general $f \neq 0$, it is to be expected that $\|S_h^{-1} f_h\|$ will grow as h^{-q}, no matter what boundary conditions are applied.

4.7.2 The Scalar One-Dimensional Linear Advection Equation

Suppose that S_h and T_h are the sorts of perturbations of Toeplitz operators discussed in Chapter 2, namely,

$$S_h = A_h + P_{1h}, \quad T_h = B_h + P_{2h},$$

where A_h and B_h are homogeneous Toeplitz operators with symbols $A(z)$ and $B(z)$, respectively. Then, for each η, the operator $e^{i\eta} S_h - T_h$ has the symbol

$$\pi(z, \eta) \equiv A(z)e^{i\eta} - B(z) \equiv \sum_{k=-n_l}^{n_r} \pi_k(\eta) z^k,$$

where $n_l \equiv \max\{l, L\}$ and $n_r \equiv \max\{r, R\}$.

Proposition 4.13 gives necessary conditions for the near-uniform invertibility condition (4.37). In particular, all roots of $\pi(z, \eta) = 0$ with $|z| = 1$ must be simple; otherwise Proposition 4.13 shows that inequality (4.37) cannot hold when $k(h) = Ch$. Writing $\lambda(z) = B(z)/A(z)$, this requires all the members of

$$\{z : \lambda(z) = e^{i\eta}, \text{ and } |z| = 1\}$$

to be simple roots.

This does not exhaust the information that can be gained from Proposition 4.13. Let the roots of $\pi(z, \eta) = 0$ be $z_i(\eta)$ for $i = 1, \ldots, n_l + n_r$. Each root $|z_i(\eta)| < 1$ must have a corresponding left-hand boundary condition. Each root $|z_i(\eta)| > 1$ must have a corresponding

right-hand boundary condition. Simple roots with $|z_i(\eta)| = 1$ may correspond with either a left- or a right-hand boundary condition. Higher order roots on the unit circle necessarily lead to instabilities.

However, the number of left-hand boundary conditions is fixed at n_l, and the number of right-hand boundary conditions is fixed at n_r, independent of η. Thus the number of roots with $|z_i(\eta)| \leq 1$ must be n_l for all η; the number of roots with $|z_i(\eta)| \geq 1$ must be n_r for all η. In particular, no root can 'defect' out of the closed unit circle, unless it is countered by a root 'defecting' into the closed unit circle.

Proposition 4.14 *The following conditions on the roots of $\pi(z,\eta)$ are necessary for the modified von Neumann condition:*

(i) the number of roots with $|z_i(\eta)| \leq 1$ is n_l for all η,

(ii) the number of roots with $|z_i(\eta)| \geq 1$ is n_r for all η,

(iii) roots with $|z_i(\eta)| = 1$, for some η, are simple.

It is interesting to note that this result is analogous to the Dahlquist–Henrici root condition [79,68,66,67] for schemes for IVPs, *but with the roles of x and t interchanged.*

Proposition 4.15 *Suppose that either condition (i) or condition (ii) of Proposition 4.14 does not hold. Then the method is uniformly unstable. Suppose that conditions (i) and (ii) of Proposition 4.14 hold, but that the maximum multiplicity of a root on the unit circle is $q > 1$, for some η. Then the method is $(q-1)$-weakly uniformly unstable.*

Proof
These results follow from equation (4.36) and Proposition 4.13. □

EXAMPLE: CRANK–NICOLSON SCHEME

The CN scheme has the amplification factor (see Table 2.1)

$$\lambda(z,\nu) = \frac{z^{-1} + 4/\nu - z}{-z^{-1} + 4/\nu + z}.$$

Assuming, as usual, that $\nu > 0$, $\pi(z,\eta) = 0$ iff

$$\pi(z,\eta) = z^2 + 2i\delta z - 1 = 0, \quad \delta \equiv \frac{2}{\nu}\tan(\eta/2). \tag{4.41}$$

It is easily seen that:

- for $|\delta| < 1$, this polynomial has two distinct roots with $|z| = 1$,
- for $|\delta| = 1$, this polynomial has one double root with $|z| = 1$, and
- for $|\delta| > 1$, this polynomial has one root with $|z| < 1$, and one with $|z| > 1$.

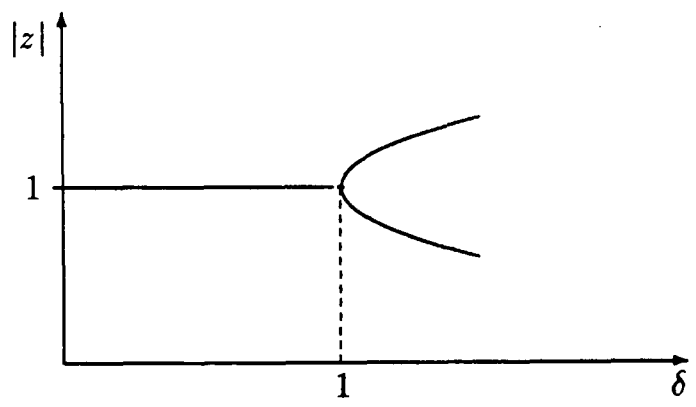


Figure 4.10: The modulus of the roots of $\pi(z,\eta) = 0$. There is a double root at $\delta = 1$.

This situation is illustrated in Figure 4.10.

Observe that when $\delta = 1$ — that is, when $\eta = \eta_{crit}$, or $\eta = \eta_{crit} - \pi$ where

$$\eta_{crit} \equiv 2\arctan(\nu/2),$$

— the polynomial $\pi(z,\eta) = (z-i)^2 = 0$ has a double root on the unit circle. Thus no method based on the CN scheme can be u.n.a.s. or uniformly stable.

In fact, the methods that result from using the CN scheme with HE and DE boundary conditions have

$$\|(e^{i\eta} - G_h)^{-1}\| = \begin{cases} C_1 h^{-1} & \text{if } |\delta| < 1 \\ C_2 h^{-2} & \text{if } |\delta| = 1 \\ C_3 h^0 & \text{if } |\delta| > 1 \end{cases} .$$

The inequalities of Proposition 4.13 hold exactly in these cases. It is not known whether this is the case in general.

Vichnevetsky [152] observes that $\delta = 1$ marks a transition in the behaviour of the solutions of the BVP with forcing $u_0^n = e^{in\eta}$. For $\delta < 1$ the solutions are wave-like, $u_j^n \sim e^{i(j\xi + n\eta)}$. For $\delta > 1$ the solutions are *evanescent*, $u_j^n \sim z^j e^{in\eta}$, where $|z| < 1$. This change of behaviour is illustrated in reference [152] by computations. However, Vichnevetsky's paper does not address the behaviour of solutions with $\delta = 1$.

□

EXAMPLE: BOX SCHEME

The amplification factor of the Box scheme is given in Table 2.1. The equation $\lambda(z) = e^{i\eta}$ is equivalent to $\pi(z,\eta) = (r - qe^{i\eta}) + (q - re^{i\eta})z = 0$. Only simple roots are possible on the

unit circle. Thus, provided the scheme is properly centred, the root condition necessary for the near-uniform invertibility condition holds.

□

EXAMPLE: SLOAN'S SCHEME

The amplification factor is given in Table 2.1. This leads to

$$\pi(z,\eta) = z^2 + \frac{2ir\sin(\eta/2)}{z_0}z - \frac{\overline{z_0}}{z_0},$$

where $z_0 \equiv se^{i\eta/2} - qe^{-i\eta/2}$. Setting $z_0 \equiv \alpha e^{i\theta}$ gives

$$\pi(z,\eta) = z^2 + \frac{2ir\sin(\eta/2)}{\alpha}ze^{-i\theta} - e^{-2i\theta}.$$

Setting $w \equiv ze^{i\theta}$ shows that

$$e^{2i\theta}\pi(z,\eta) = w^2 + 2i\delta w - 1,$$

where $\delta \equiv r\sin(\eta/2)/\alpha(\eta)$. As in the example of the Crank–Nicolson scheme, π has a double root on the unit circle iff $\delta = \pm 1$, which is equivalent to $r^2\sin^2(\eta/2) = \alpha^2$. Using $\alpha^2 = |z_0|^2 = q^2 + s^2 - 2qs\cos(\eta)$, it follows that $\delta = \pm 1$, for some η, iff

$$\frac{2(q^2+s^2)-r^2}{4qs-r^2} = \cos(\eta), \text{ which has a solution iff } \left|\frac{2(q^2+s^2)-r^2}{4qs-r^2}\right| \leq 1. \qquad (4.42)$$

Recalling the definitions of q, r and s from Table 2.1, we have

$$2(q^2+s^2)-r^2 = 12[7\nu^2-4], \quad 4qs-r^2 = 12[\nu^2-4].$$

Thus, $\delta = \pm 1$, for some η, iff

$$\left|\frac{7\nu^2-4}{\nu^2-4}\right| \leq 1.$$

This inequality holds (for positive ν) only when $\nu \in (0,1)$. We conclude that, when $\nu \in (0,1)$, no method based on Sloan's scheme (including the SLDE, SLHE and SLHEDE methods) can satisfy the near-uniform invertibility condition. Hence, for $\nu \in (0,1)$, no method based on Sloan's scheme can be u.n.a.s. or uniformly stable. In particular, this result agrees with the computation for the SLHEDE method given on page 85.

□

EXAMPLE: ANGLED-DERIVATIVE SCHEME

The Angled-Derivative scheme has

$$\pi(z,\eta) = z^2 - \frac{2rie^{i\eta/2}\sin(\eta/2)}{q}z - e^{i\eta}.$$

Setting $w = ze^{-i\eta/2}$ shows that

$$e^{-i\eta}\pi(z,\eta) = w^2 + 2i\delta w - 1,$$

where $\delta \equiv r\sin(\eta/2)/q$. Thus there is a double root on the unit circle when $\delta = \pm 1$: that is, when $|q/r| \leq 1 \Longleftrightarrow |\nu/(2+\nu)| \leq 1$. However, this inequality holds for all $\nu \geq 0$, so no method based on the Angled-Derivative scheme can be u.n.a.s. or uniformly stable.

□

EXAMPLE: DISSIPATIVE SCHEMES

Consider a dissipative scheme with amplification factor $\lambda(z,\nu)$. Thus $|\lambda(z,\nu)| < 1$ for all $z \neq 1$ on the unit circle. Then

$$\pi(z,\eta) = \lambda(z) - e^{i\eta}.$$

The only root of π on the unit circle is $\pi(1,0) = 0$, which must be simple as consistency (of order p) demands that $\lambda(z,\nu) = z^{\nu} + O(|z-1|^{p+1})$. Thus the necessary condition for the near-uniform invertibility condition is satisfied.

□

EXAMPLE: LAX–WENDROFF METHOD

Consider the (dissipative) Lax–Wendroff method for the linear advection equation. This has the symbol [135, page 174]

$$\lambda(z,\nu) = \frac{\nu}{2}(\nu+1)z^{-1} + (1-\nu^2) + \frac{\nu}{2}(\nu-1)z.$$

Thus

$$\pi(z,\eta) = z^2 - 2b(\eta)z + c(\eta),$$

where

$$b(\eta) = \frac{1-\nu^2-e^{i\eta}}{\nu(1-\nu)}, \quad c(\eta) = \frac{\nu+1}{\nu-1}.$$

For all η, the equation $\pi(z,\eta) = 0$ has one root inside C and one root outside. Thus one left-hand boundary condition and one right-hand boundary condition are needed.

□

4.7.3 Group Velocity

The results of the previous section show that, of the schemes in our battery for the linear advection equation, the only ones that can possibly be uniformly stable/u.n.a.s. are Sloan's scheme ($\nu > 1$) and the Box scheme. An examination of the corresponding dispersion relations (see Chapter 2) reveals that these two schemes are the only ones in the battery whose group velocities are non-zero at all frequencies. This result holds generally.

Proposition 4.16 *A necessary condition for the uniform stability/u.n.a.s. of a unitary scheme is the group velocity condition*

$$c_g(\xi) \neq 0, \quad \forall \xi \in (-\pi, \pi].$$

In particular, no method for the scalar linear advection equation, that is based on two-level unitary scheme, and has $r = L \Longleftrightarrow l = R$, can be uniformly stable/u.n.a.s.

Proof
A unitary scheme must have even order accuracy. Thus, if $c_g(\xi_0) = 0$, then ξ_0 must be a local maximum or a minimum of the dispersion relation. However, this implies that $\pi(z, \eta) = \lambda(z) - e^{i\eta}$ has a double root on the unit circle. By Proposition 4.14, the scheme cannot be u.n.a.s. or uniformly stable. Proposition 2.1 proves the final assertion.

□

This result might have been expected on physical grounds. Where Fourier analysis can be used, the group velocity is the velocity of energy propagation. Consequently, $c_g = 0$ implies that energy cannot escape the computational domain, and a rapid decay of initial data cannot be hoped for. It is possible that this result holds for general schemes, applied to hyperbolic systems in several dimensions. If so, it would be analogous to Rauch and Taylor's results [117] for analytic hyperbolic problems.

Armed with Proposition 4.16, an examination of the dispersion relations in Figures 2.1–2.5 allows the prediction of all the uniform instabilities discussed on Page 86.

Proposition 4.16 may well be linked with the following result due to Thomée [120, page 100]. Let 'stability' be in the sense of the Lax–Richtmyer definition, but with the space norm changed to the $\ell_h^p(\Re)$ norm. Consider a scheme for an IVP for the unforced scalar linear advection equation, with an even order of accuracy. Then the result states that the scheme is unstable for $p \neq 2$. Trefethen [149, page 205] observes that the instabilities are weak.

Next, we shall contrast the group velocity result with Trefethen's group velocity interpretation of the GKS theory.

Recall that a method for the scalar linear advection equation (equation (4.6)) requires a genuine boundary condition at $x = 0$, and may require a numerical boundary condition at $x = 1$. The Godunov–Ryabenkii and GKS theories aim to determine the stability of a method for this problem (in appropriate norms) by considering, in turn, the stability of the two corresponding quarter-plane problems (posed on $x \geq 0$ and $x \leq 1$), and the corresponding IVP. The *left quarter-plane problem*, which involves the numerical boundary condition, is of paramount interest.

Trefethen [149, page 208] (see also [148,147]) has shown that a necessary and sufficient condition for GKS stability of a method for the left quarter-plane problem is that the *method* admit no (non-trivial) solution (or *mode*) of the form $u_j^n = Ce^{i(j\xi + n\eta)}$, where

(i) $\mathrm{Im}(\eta) \leq 0$ and $\mathrm{Im}(\xi) < 0$, *or*

(ii) $\xi, \eta \in \Re$ and $c_g \leq 0$.

In case (i), modes with $\mathrm{Im}(\eta) = 0$ and $\mathrm{Im}(\xi) < 0$ are precisely Vichnevetsky's *evanescent modes* (see page 104). Modes with $\mathrm{Im}(\eta) < 0$ and $\mathrm{Im}(\xi) < 0$ are members of $\ell_h^2((-\infty, 1])$, but grow exponentially in time. They are precisely the modes considered by Godunov and Ryabenkii in the original formulation of their theory [38]. Thus, evanescent and exponentially growing modes prevent GKS stability. The modes of case (ii) radiate energy from the numerical boundary into the computational domain. This leads to a build-up of energy that prevents GKS stability.

There are corresponding conditions for the right quarter-plane problem. However, sensible discrete counterparts of the genuine boundary condition at $x = 0$ usually result in

the trivial satisfaction of these conditions.[2] Finally, the IVP is required to satisfy the von Neumann condition.

From this discussion, we see that a unitary scheme with vanishing group velocity may be GKS stable, provided that the boundary conditions exclude any *modes* of zero group velocity. By contrast, uniform stability finds any unitary scheme with vanishing group velocity unstable, regardless of the boundary conditions. This difference is reasonable. The GKS norms measure solutions in norms that contain (time) exponentially decaying weights. Consequently, the polynomial build-up of energy that results from energy trapped in the computational domain need not result in GKS instability. However, this build-up of energy would result in uniform instability.

4.7.4 Near-Uniform Invertibility of Two-Dimensional Toeplitz Operators

The near-uniform invertibility condition $\|(e^{i\eta} - G_h)^{-1}\| = O(k^{-1})$ is still equivalent to the modified von Neumann condition, in two or more dimensions, provided that $G_h = S_h^{-1}T_h$ is power bounded. In particular, methods for the two-dimensional linear advection equation

$$u_t + au_x + bu_y = f(x,y,t), \quad (x,y) \in \Omega, \tag{4.43}$$

result in the family of two-dimensional Toeplitz operators $\{e^{i\eta}S_h - T_h\}_{h>0}$, with possible boundary perturbations. Thus, an investigation of the near-uniform invertibility of such operators would aid understanding of the uniform stability of the corresponding methods.

This section presents necessary conditions for the near-uniform invertibility of families of two-dimensional Toeplitz operators. Our approach exploits the fact that the two-dimensional Toeplitz operators have one-dimensional Toeplitz operators 'embedded' within them. This allows the results of Section 4.7.1 to be used in two-dimensional contexts. The results that follow, whilst in some ways unsatisfactory, are sufficient for a discussion of Morton and Paisley's [97] 'spurious modes'. Further results concerning the *invertibility* of two-dimensional Toeplitz operators can be found in [144,92,30,104,102,29], and [28]. However, the author could find no mention in the literature of (what we have termed) 'near-uniform invertibility'.

Consider a two-dimensional Toeplitz operator, with the fixed order symbol

$$A(z_x, z_y) = \sum_{(j,k)\in S} a_{j,k} z_x^j z_y^k.$$

Here S is some finite collection of ordered pairs of integers, or *stencil points*, and the $a_{j,k}$ are constants. The corresponding family of Toeplitz operators $\{A_h\}_{h>0}$ is given by

$$A_h = \sum_{(j,k)\in S} a_{j,k} E_x^j E_y^k,$$

where E_x is the left shift operator in the x direction, and E_x^{-1} (which need not be the inverse of E_x) is the right shift operator in the x direction. E_y and E_y^{-1} are defined analogously.

[2] Recall from page 104 that the Crank–Nicolson scheme admits evanescent solutions for the right quarter-plane problem, with the boundary condition $u_0^n = e^{in\eta}$ at $x = 0$. However, whilst the counterpart of condition (i) implies that this method is GKS unstable, this conclusion is of little interest, since the GKS theory does not even claim to apply to the problem: Fourier boundary forcing is not square integrable in time. In fact, the only 'Fourier mode' boundary condition admitted by the GKS theory, for this problem, is $u_0^n = 0$.

If the domain Ω is the quarter-plane, so that E_x and E_y are shift operators on the half space, then the symbols of most methods for the two-dimensional linear advection equation will vanish on the unit torus. From Theorem 2.6 of Chapter 2, A_h cannot be even invertible in $L(\ell_h^2(\Omega))^3$, for any h. Hence the modified von Neumann condition cannot hold. For this reason, Ω is assumed to be bounded in this section. In fact, we assume $\Omega = I \times I$.

Let γ be any closed path on the unit torus, parameterized by ξ,

$$\gamma = \{(e^{i\xi_x(\xi)}, e^{i\xi_y(\xi)}) : \xi \in [-\pi, \pi]\}. \tag{4.44}$$

Define $f(\xi) \equiv A(e^{i\xi_x(\xi)}, e^{i\xi_y(\xi)})$. Then $f : [-\pi, \pi] \to \mathcal{C}$ may be extended to a 2π-periodic, continuous function, with Fourier coefficients $\{\hat{f}_j\}_{j\in\mathcal{Z}}$. When only a finite number of $\hat{f}_j \neq 0$, these coefficients define a one-dimensional Toeplitz operator, with the symbol

$$\hat{\hat{f}}(z) \equiv \sum_k \hat{f}_k z^k,$$

that agrees with f on the unit circle. Proposition 4.13 gives necessary conditions for the family of one-dimensional Toeplitz operators to be nearly uniformly invertible. These conditions are necessary for the near-uniform invertibility of the two-dimensional family $\{A_h\}_{h>0}$. In particular, the equation

$$\hat{\hat{f}}(z) = 0$$

may only have simple roots on the unit circle. When $z_0 = e^{i\xi_0}$ is such a root, this implies that the derivative of $f : [-\pi, \pi] \to \mathcal{C}$ must satisfy

$$f'(\xi_0) \neq 0, \text{ whenever } f(\xi_0) = 0. \tag{4.45}$$

This condition may have a simple interpretation in terms of the graph of $q(\xi_x, \xi_y) \equiv |A(e^{i\xi_x}, e^{i\xi_y})|$ (c.f. [97,96,98,90]).

This result can be used to investigate the uniform stability of finite difference methods for the solution of equation (4.43). In particular, it can be used to investigate unitary schemes in two dimensions. Such schemes have received considerable attention in hydraulics circles in the last few years. Stelling [139] and other authors [158,159,150] have observed that the Beam and Warming scheme [5], which is a generalization of the Crank–Nicolson scheme to non-linear, multi-dimensional problems, loses accuracy when it is used in conjunction with the ADI factorization, and when the computational boundaries are not aligned with the (Cartesian) mesh. These instabilities have been attributed to the ADI factorization.

A unitary scheme in two dimensions has a real dispersion relation

$$\phi(\eta, \xi_x, \xi_y) = 0, \quad (\eta, \xi_x, \xi_y) \in [-\pi, \pi]^3.$$

This dispersion relation must be periodic in ξ_x, in ξ_y and in η. It assigns a real number η to each point on the unit torus.

Now consider a smooth closed path γ on the unit torus of the form given by equation (4.44). By flattening out the torus into a square, and replicating the square across the plane, we obtain what we call a *a closed path in the periodic plane*, denoted $\tilde{\gamma}$. Equation (4.45) may be interpreted, and improved, as follows.

[3]This observation does not conflict with the convergence results obtained by Süli in [145]. In that work, the domain was assumed to be a rectangle.

Proposition 4.17 *Consider a unitary scheme for the two-dimensional linear advection equation with (real) dispersion relation $\eta(\xi_x, \xi_y)$. Then a necessary condition for the uniform stability of a method based on the scheme is that, for any smooth closed path $\tilde{\gamma}$ in the periodic plane, $\eta|_{\tilde{\gamma}}$ have no maximum. Thus, a necessary condition is that $c_g(\xi_x, \xi_y) \neq 0$ for all (ξ_x, ξ_y).*

We conjecture that the following result holds. Under the stated hypotheses, the group velocity vector coincides with the energy propagation vector [83].

Conjecture 4.1 *Consider a (possibly dissipative) scheme for the d-dimensional linear advection equation, with (possibly complex) dispersion relation $\eta(\xi_1, \ldots, \xi_d)$. Suppose that there exists $(\xi_1, \ldots, \xi_d) \in [-\pi, \pi]^d$, such that $\eta(\xi_1, \ldots, \xi_d)$ is real and $c_g(\xi_1, \ldots, \xi_d) = 0$. Then any method based on the scheme is uniformly unstable.*

In particular, the Beam and Warming scheme has zero group velocities at some frequencies. This observation may explain the loss of accuracy with this scheme, observed by Stelling, Weare and other authors.

4.7.5 Steady Problems and Spurious Modes

The results of the previous section apply to the steady problem

$$au_x + bu_y = f(x, y), \quad (x, y) \in I^2, \tag{4.46}$$

with suitable boundary conditions. On regular meshes, finite difference methods for this problem lead naturally to linear systems, which involve (perturbed) two-dimensional Toeplitz matrices. These linear systems usually have to be solved using some time-stepping, or some iterative, procedure. However, the stability of methods for equation (4.46) may be considered without reference to the method used for the solution of the linear system.

Suppose that $\{\tilde{Q}_h\}_{h>0}$ is the family of (perturbed) two-dimensional Toeplitz operators that result from the finite difference method. Then the method may be written ($h \equiv h_x$)

$$\frac{1}{h}\tilde{Q}_h u_h = s_h f, \quad \tilde{Q}_h : \ell_h^2(I^2) \to \ell_h^2(I^2),$$

and $s_h : Y \to Y_h$ is a restriction operator: for the cell-centred scheme, $(s_h f)_{i,j}$ would be $f(ih_x, jh_y)$; for the cell-vertex scheme, s_h would represent the average of f over the four neighbouring grid points. In both cases, $s_h \ell_h^2 \subseteq \ell_h^2$, so we have the following observation.

Irrespective of the iterative method, or time-stepping method, actually used to solve the system, stability demands that

$$\|\tilde{Q}_h^{-1}\|_{L(\ell_h^2(I^2))} = O(h^{-1}).$$

This is simply the near-uniform invertibility condition for families of two-dimensional Toeplitz operators, that has just been investigated. Our claims are as follows.

- Morton and Paisley's 'spurious modes' [97] correspond to violations of equation (4.45).
- As a consequence, spurious modes actually represent weak instabilities. Thus, whilst spurious modes can be controlled by the boundary conditions when $f(x, y) = 0$, as claimed in reference [98, page 7], for general $f \neq 0$, a reduction in the order of convergence must be expected. This assertion is demonstrated by simple computations in Chapter 6.

- The necessary conditions for uniform stability presented above represent a natural generalization of the idea of spurious modes to time-dependent problems.
- Condition (4.45) clarifies why some roots on the unit torus represent 'spurious' modes, whilst other roots on the unit torus represent modes that should be considered as 'true' modes (c.f. reference [97]).

EXAMPLE: CELL-VERTEX AND CELL-CENTRED SCHEMES

Cell-vertex and cell-centred schemes for equation (4.46) have been considered in references [90,97,98,96]. Consider the simplest case, where the spatial steps in the x and y directions are fixed at h_x and h_y respectively. In this case, the cell-vertex scheme reduces to the Box scheme

$$p(u_{i,j+1} - u_{i+1,j}) + q(u_{i+1,j+1} - u_{i,j}) = 2h_x \tilde{f}_{i+1/2,j+1/2},$$

where $p \equiv -a + b\rho$, $q \equiv a + b\rho$ and $\rho \equiv h_x/h_y$. We define

$$\tilde{f}_{i+1/2j+1/2} \equiv \frac{1}{4}\left[f_{i,j} + f_{i,j+1} + f_{i+1,j} + f_{i+1,j+1}\right],$$

where $f_{i,j} \equiv f(ih_x, jh_y)$

The cell-centred scheme reduces to the Leapfrog method

$$b\rho(u_{i,j+1} - u_{i,j-1}) + a(u_{i+1,j} - u_{i-1,j}) = 2h_x f_{i,j}.$$

To demonstrate the claims made above, choose $\gamma = \{(e^{i\xi}, e^{i\eta_0}) : \xi \in [-\pi, \pi]\}$, where $\eta_0 \in \Re$ is a constant to be chosen. (In fact, we choose η_0 so that $b\rho/a \sin(\eta_0) = \pm 1$.) Then the cell-centred scheme has

$$\hat{f}(z) = z^2 - 2i\delta z - 1,$$

where $\delta = b\rho/a \sin\eta_0$. This is the same polynomial studied in the context of the Crank–Nicolson scheme.

Let $\nu \equiv b\rho/a$. When $|\nu| \geq 1$, there always exists η_0 such that $\delta = \pm 1$; in which case, $f(z)$ has a double root on the unit circle which violates equation (4.45). Consequently, the family $\{h^{-1}\tilde{Q}_h\}_{h>0}$ cannot be uniformly invertible. Thus, when $\nu \geq 1$, irrespective of the boundary conditions and time-stepping used in the final method, the cell-centred scheme is unstable in the $\ell_h^2(I^2)$ norm.

The previous choice of γ corresponds to one of the two 'corrugated iron' spurious modes reported in [97]. The other arises from the choice $\gamma = \{(e^{i\eta_0}, e^{i\xi}) : \xi \in [-\pi, \pi]\}$. This choice leads to the same polynomial as before, but with $\delta = \nu^{-1} \sin\eta_0$. Thus, the cell-centred scheme is unstable, in the $\ell_h^2(I^2)$ norm, for any $\nu \leq 1$. In sum, the cell-centred scheme is *always* $\ell_h^2(I^2)$-unstable. Numerical results illustrating this assertion are presented in Section 6.3. Finally, note that the 'diagonal' spurious mode, reported in [97], arises from a suitable choice of γ, but does not add to our conclusion.

Note that instabilities analogous to these cannot occur for the Leapfrog method, since they can only manifest themselves for $|\nu| > 1$.

The cell-vertex scheme cannot suffer from this sort of instability, since its symbol is linear in z_x and z_y. The 'chequerboard' mode, given by Morton and Paisley, need not imply $\ell_h^2(I^2)$-instability. In fact, the calculations given in Chapter 6 support the view that

the cell-vertex scheme is $\ell_h^2(I^2)$-stable. For this reason, the author does not feel that the chequerboard mode should be considered spurious.

□

4.8 Multi-Step Schemes

Consider a multi-step finite difference scheme for hyperbolic IVPs, which we will write in the form

$$\sum_{(j,m)\in S} a_j^m(\nu) u_{i+j}^{n+m} = k g_i^n,$$

where S is a finite collection of stencil points. Define

$$\Phi(\lambda, z, \nu) \equiv \sum_{(j,m)\in S} a_j^m(\nu) \lambda^n z^j.$$

For the discrete IVP, Fourier analysis can be used to reduce the problem to a form that is effectively a time-marching scheme for an ODE, which depends on a (spatial frequency) parameter. Consequently, the Dahlquist–Henrici root condition for the stability of ODEs can be applied (see, for example, reference [79]), yielding the following result (see references [67,65]).

Theorem 4.1 *A multi-step scheme for a one-dimensional hyperbolic IVP is n.s. iff, when* $|z| = 1$, *all the roots*

$$\lambda : \ \Phi(\lambda, z, \nu) = 0$$

lie in the closed unit disc $D = \{w : |w| \leq 1\}$, *and those* λ *on the boundary* $\partial D = \{w : |w| = 1\}$ *are simple.*

The conditions of this result are necessary for uniform stability. We conjecture that the following symmetrical requirements are also necessary for uniform stability.

Conjecture 4.2 *A necessary condition for the uniform stability of a multi-step scheme, for a one-dimensional hyperbolic IBVP, is that, when* $|\lambda| = 1$, *all the roots*

$$z : \ \Phi(\lambda, z, \nu) = 0$$

on the boundary $\partial D = \{w : |w| = 1\}$ *are simple.*

Notice that the roots that do not lie on ∂D may reside inside or outside of ∂D. Unlike the situation of Theorem 4.1, numerical boundary conditions can be prescribed at $x = 1$ to control the roots with $|z| > 1$.

These two results could be made more symmetrical by allowing *time-implicit schemes*: that is, schemes that require 'boundary conditions' at $t = T_{final}$ as well as at $t = 0$, in just the same way as we allow numerical boundary conditions at $x = 1$, in addition to the genuine boundary condition at $x = 0$. In fact, the cell-vertex scheme for $au_x + bu_y = f(x, y)$ may be viewed as a time-implicit version of the Leapfrog scheme, by regarding y as a temporal dimension. The main difference is that the stability of the cell-centred scheme would typically be studied in $X_h = \ell_h^2(I \times I)$, whilst the stability of the time-implicit Leapfrog scheme would typically be studied in $X_h = \ell_k^\infty([0, T_{final}], \ell_h^2(I))$.

One drawback of such schemes is that causality is lost for evolutionary problems. Furthermore, time-implicit schemes effectively raise the dimensionality of the problem by one. This is a serious drawback. For these schemes to be viable, a two-dimensional analogue of the Thomas algorithm is needed. Malyšev [92] offers a step in the right direction.

4.9 Conclusions

New Definitions

This chapter has defined and investigated various forms of IVP stability for discrete problems, and concluded that u.n.a.s. is the most valuable. In fact, we have argued that P-stability should be *defined* as u.n.a.s.

We have also given a new IBVP stability definition — uniform stability — that uses uniform norms for both the space of forcing functions and the solution space. Unlike the GKS and Lax–Richtmyer stability theories, uniform stability is applicable to methods for forced problems that have forcing that does not decay in time. This property makes uniform stability well suited, in particular, for the analysis of methods for the computation of equilibrium solutions.

We have also shown (Proposition 4.11) that uniform stability is equivalent to u.n.a.s. Both forms of stability have been shown (Proposition 4.2) to be equivalent to:

(i) power-boundedness, and

(ii) the modified von Neumann condition (which is equivalent to uniformly negative type).

In principle, these conditions allow the uniform stability/u.n.a.s. of a given method to be decided.

We have also defined the discrete counterparts of equilibrium solutions, and shown (Proposition 4.7) that the modified von Neumann condition is sufficient for their BVP stability. (The different types of stability considered in this chapter are summarized in Table 4.5.)

These considerations highlight the importance of the modified von Neumann condition. Unfortunately, this condition is difficult to verify. However, we have shown (Proposition 4.10) that the modified von Neumann condition is equivalent to the near-uniform invertibility condition whenever power-boundedness holds. Furthermore, we have developed Toeplitz operator techniques that make the near-uniform invertibility condition comparatively easy to check.

Necessary Conditions for Uniform Stability

These Toeplitz operator techniques lead to simple conditions, involving only the coefficients of a scheme, that are necessary for the near-uniform invertibility condition to hold in one dimension. These conditions have allowed us to show that only a few schemes in our battery for the scalar linear advection equation have the potential to result in uniformly stable methods. In particular, no method based on the Angled-Derivative, Sloan ($\nu < 1$) or Crank-Nicolson schemes can be u.n.a.s. or uniformly stable, regardless of the boundary conditions employed in the final method. Of the schemes in our battery, only the Box scheme and Sloan's scheme ($\nu > 1$) have the potential for u.n.a.s. and uniform stability. We have shown that this conclusion also follows from a general result: no unitary scheme, that admits waves

Stability Type	*Definition*		*Conditions*
n.s.	$d_h(t) \leq M$	$\Longleftrightarrow$	power-boundedness
n.a.s.	$d_h(t) \leq Me^{\beta_h t}$	$\Longleftrightarrow$	(i) power-boundedness, and (ii) $r(G_h) < 1$
u.n.a.s.	$d_h(t) \leq Me^{\beta t}$	$\Longleftrightarrow$	(i) power-boundedness, and (ii) modified VNC/NUI
BVP	Section 4.5	$\Longleftarrow$	u.n.a.s.
uniform	Section 4.6	$\Longleftrightarrow$	u.n.a.s.

Table 4.5: Summary of the types of stability considered throughout this chapter. Here, $d_h(t) = \|G_h(t)\|$, and 'NUI' denotes the near-uniform invertibility condition.

of zero group velocity for the IVP, can result in a uniformly stable method for an IBVP (Proposition 4.17).

Uniform Invertibility

Our results concerning the invertibility of Toeplitz operators also make clear how quickly the norms of the inverses of these operators grow when the conditions of the Strang and Osher's Theorem are violated. A misplaced boundary condition leads to catastrophic exponential growth. A root on the unit circle of multiplicity q leads to $O(h^{-q})$ growth, which translates into a weak instability.

Spurious Modes

We have also presented necessary conditions for the near-uniform invertibility condition in two dimensions. These conditions have allowed us to show, in the context of steady computations, that Morton and Paisley's spurious modes correspond to instabilities in the $\ell_h^2(I^2)$ norm.

Extensions

Finally, we observe that most of the theory in this chapter also applies to parabolic problems. For scalar parabolic problems, $k(h) = Ch^2$, so the near-uniform invertibility condition, necessary for uniform stability, requires that the family of Toeplitz operators $\{e^{i\eta}S_h - T_h\}_{h>0}$ be nearly uniformly invertible of order 2. Thus, if a scheme for a scalar parabolic problem has the amplification factor λ, a necessary condition for the uniform stability of any method based on this scheme is that, for any $\eta \in \Re$, the roots of $\lambda(z) - e^{i\eta}$ that lie on the unit circle have maximum multiplicity two.

The next chapter concerns the convergence of discrete equilibrium solutions to exact equilibrium solutions in time-uniform norms. Our task is, effectively, to show that there is a Lax equivalence Theorem for uniform stability. When this is done, it will be clear that the schemes that we have identified as being not uniformly stable, such as the Crank–Nicolson scheme, cannot be convergent for arbitrary forcing, no matter what boundary conditions they are teamed with.

Chapter 5

Convergence

5.1 Introduction

There are two aspects to the convergence of a numerical method that uses evolutionary initialization. First, arbitrary initial data must be integrated over a period of time sufficiently large to ensure that the discrete transient is small. Secondly, the discrete equilibrium solution must be a good approximation to the analytic equilibrium solution.

If the same method is to be used for both the initialization and final computational phases, care must be taken in choosing the method. There is no point in using a method that efficiently damps transients, if it is inefficient at the final integration. Similarly, there is no point in using a highly efficient method for the final integration, if that method is not efficient at eliminating transients.

In this chapter, the balance between transient damping and equilibrium convergence is investigated. It is shown that convergence in the uniform norm gives an adequate account of the total efficiency of a method.

It is also shown that there is a Lax equivalence Theorem in the uniform norms. For consistent methods, uniform convergence occurs for all data, if and only if the method is uniformly stable. Unfortunately, this theory requires the boundary forcing to be continuous. For periodic forcing, a result with less restrictive hypotheses is obtained.

5.2 Convergence Orders

5.2.1 Equilibrium Order

Consider a method for the computation of an equilibrium solution. Assume that an evolutionary initialization method is used. Thus, arbitrary initial data is integrated through time $T_h = N_h k(h)$, which depends on some stopping criterion. The data $u_h^{N_h}$ is used to initiate a second computation, through time T. The result is a discrete solution u_h defined for times in $[T_h, T_h + T]$. This is to be compared with the exact result $u_{eq}(t)$, for $t \in [T_h, T_h + T]$. Consequently, we define the *equilibrium error* to be

$$M_h \equiv \|u_h - r_h u_{eq}\|_h,$$

where the norm is the uniform X_h norm[1] restricted to $t \in [T_h, T_h + T]$. When $M_h = O(h^q)$, the *equilibrium order* is said to be q.

[1] That is, the time-uniform norm discussed in Chapter 4.

5.2.2 Uniform Convergence

Suppose that the method is convergent, of order p, in the uniform norm. Then

$$\begin{aligned} M_h = \|u_h - r_h u_{eq}\|_h &\leq \|u_h - r_h u\|_h + \|r_h u_{trans}\|_h \\ &\leq Ch^p + Me^{\beta_G T_h}, \end{aligned}$$

where $\beta_G < 0$ is the type of the semigroup. To allow for the worst imaginable choice of initial data, the optimal stopping procedure should be chosen so that $T_h = \frac{p}{\beta_G} \log h$. A larger choice of T_h would result in the analytical transient being dominated by the discretization error. A smaller choice of T_h would imply the opposite. With the optimal stopping procedure, the equilibrium order is p.

5.2.3 BVP Convergence and u.n.a.s.

Suppose that the method is BVP-convergent, of order p, and u.n.a.s. Then

$$\begin{aligned} M_h = \|u_h - r_h u_{eq}\|_h &= \|u_{eq,h} - r_h u_{eq}\|_h + \|u_{trans,h}\|_h \\ &\leq Ch^p + Me^{\beta_h T_h}. \end{aligned}$$

Since the method is u.n.a.s., there exists $\beta < 0$ such that $\beta_h < \beta$, uniformly in h. For smooth forcing, this approach does not improve on the conclusions of Section 5.2.2, since uniform stability implies both BVP stability and u.n.a.s.

However, for periodic forcing, this approach does improve on the previous conclusions. The proof of well-posedness for periodically forced problems in Section 3.6.1 did not require the forcing to be continuous. Nor did the stability results of Section 4.5.1. Consequently, we shall also prove a Lax equivalence Theorem for BVP-convergence and stability.

5.2.4 Global Equilibrium Order

We take the view that one is primarily interested in the rate of improvement of the equilibrium error as additional computer time is spent on a calculation. That is, if W_h is the amount of computer work (or 'complexity') required for a calculation with mesh parameter h, the main interest is in the connection between M_h and W_h. When $M_h = CW_h^{-g}$, we say the method has *global equilibrium order* g. We assume that a 'good' scheme has a large global equilibrium order. This is an over-simplification. In calculations involving HPDEs, the ability to, say, resolve shocks well may be deemed more important than the global equilibrium order.

Godunov and Ryabenkii [38, page 315] say that a method is *efficient* when the "number of arithmetic operations in the step from u^p to u^{p+1} is proportional to the number of unknowns". In one dimension, explicit methods and implicit methods employing the Thomas algorithm are efficient. In two or more dimensions, any method that uses operator splitting is also efficient.

Consider an evolutionary computation carried out over the time interval $[0, T_h']$ ($T_h' \equiv T_h + T$) with an efficient method. Assume that the domain $\Omega \subseteq \Re^d$ is bounded, and that the mesh is Cartesian. Then

$$W_h = C \frac{T_h'}{k(h)} h^{-d}. \tag{5.1}$$

When $k(h) \sim h^\beta$, $W_h = CT'_h h^{-(d+\beta)}$. (For hyperbolic problems, $\beta = 1$.) Consequently, a method with equilibrium order q has

$$C(T'_h)^{-g} h^{g(d+\beta)} = M_h \leq C' h^q \quad \Longrightarrow \quad g \leq \frac{q + C_h}{d + \beta - \log_h T'_h},$$

where $C_h \to 0$ as $h \to 0$. The stopping procedure should be chosen to maximize q, whilst simultaneously minimizing $-\log_h T'_h$.

With the assumptions of Section 5.2.2, the optimal choice for T_h is $T_h = \frac{p}{\beta_G} \log h$. Then, for small h,

$$g \approx \frac{p}{d + \beta - p/\beta_G}.$$

Since $\beta_G < 0$, g is monotonic in p, and a 'good' scheme is one with a large uniform convergence order p.

5.3 A Lax Equivalence Theorem for Uniform Stability

The considerations above show that a reasonable description of the convergence of methods for the computation of equilibrium solutions follows from convergence in the uniform norm. Stability in the uniform norm was studied in Chapter 4, whilst well-posedness was studied in Chapter 3. All that remains is to prove a Lax equivalence Theorem for uniform convergence. This is tantamount to verifying the assumptions of Theorem 2.1, given on page 37. Consequently, we shall first define appropriate restriction operators, then prove that the assumptions of the equivalence theorem hold for uniform stability. Finally, we shall make some remarks concerning consistency.

Restriction Operators

For uniform well-posedness, X is the closure of $C^1(\Re^+, B)$ in $L^\infty(\Re^+, B)$, and Y is the closure of $\mathcal{D}(A) \times C^1(\Re^+, B)$ in $B \times L^\infty(\Re^+, B)$ (Section 3.4). For uniform stability, $X_h = \ell_k^\infty(\Re^+, B_h)$, and it is easily seen that Y_h is isomorphic with $B_h \times \ell_k^\infty(\Re^+, B_h)$ (Section 4.6). For each $h > 0$, two restriction operators are needed: a restriction $r_h : X \to X_h$, and a restriction $s_h : Y \to Y_h$. The first restriction is defined in two steps.

(i) First, $u \in X$ is restricted in time: $u \mapsto \tilde{u}_h$, where $\tilde{u}_h = [\tilde{u}^0, \tilde{u}^1, \ldots]$. $\tilde{u}^m \equiv \delta_{mk} u$, where, for each $t \in \Re^+$, $\delta_t : X \to B$ is the continuous extension of the point evaluation operator.

(ii) Next, $\tilde{u}_h$ is restricted in space: $\tilde{u}_h \mapsto r_h u$, where $r_h u = [x_h \tilde{u}^0, x_h \tilde{u}^1, \ldots]$, and $x_h : B \to B_h$ is a spatial restriction.

The second restriction $s_h : Y \to Y_h$ is defined analogously, but, for some methods, may not be identical to r_h.

The Lax Equivalence Theorem

Next we verify that our choice of restriction operators satisfies the conditions of Theorem 2.1, given on page 37. Condition (i) (equation (2.14)) is the main concern. Suppose that $g_h \in Y_h$ is such that $\|g_h\|_{Y_h} \leq 1$. Then, for each $n \geq 0$, $\|g_h^n\|_{B_h} \leq 1$. Suppose that x_h satisfies condition (2.14). Then, for each $n \geq 0$, there exists $g^n \in B$ such that $s_h g^n = g_h^n$, and

$\|g^n\|_B \leq R$, where R is independent of n. Let $g \in Y$ be the linear interpolant, in time, of the $\{g^n\}$. Then, by construction, $s_h g = g_h$, and $\|g\|_Y = \sup_{m\geq 0} \|g^m\|_B \leq R$. Thus, condition (i) is verified.

Turning to condition (ii) of page 37, Proposition 4.8 shows that each Q_h is invertible in $L(X_h)$, provided that each G_h : $r(G_h) < 1$ and each B_h : $\dim(B_h) < \infty$. Finally, condition (iii) requires $Q_h^{-1} s_h$: $Y \to Y_h \to X_h$ to be a member of $L(Y, X_h)$ for each $h > 0$. This is easily verified, under the assumptions made so far.

Proposition 5.1 *Suppose that* $\dim(B_h) < \infty$ *and* $r(G_h) < 1$ *for each* $h > 0$. *Suppose that* $\{x_h : B \to B_h\}_{h>0}$ *is a family of restrictions that satisfy condition (2.14). Then the assumptions of the Lax equivalence Theorem are met, for stability and convergence in the uniform norm.*

Consistency

Consistency requires the truncation error $l_h = Q_h r_h u - s_h g$ to satisfy $\lim_{h\to 0} \|l_h\|_{Y_h} = 0$, for each $g \in Y_0$, a dense subset of Y. With r_h as described above,

$$l_h^0 = x_h \tilde{u}^0 - x_h \tilde{\phi}, \quad l_h^{m+1} = \frac{1}{k}\left(x_h \tilde{u}^{m+1} - G_h x_h \tilde{u}^m\right) - x_h \tilde{f}^m, \ m \geq 0.$$

For $g \in Y_0 = \mathcal{D}(A) \times C^1(\Re^+, B)$, then $u \in C^1(\Re^+, B)$, so that $\tilde{u}^m = u(mk) \equiv u^m$, and

$$u^m = G(k)^m \phi + \int_0^{mk} G(mk - s) f(s)\, ds.$$

It is easily seen that

$$k l_h^{m+1} = [x_h G(k) - G_h x_h]\, u^m + x_h \left[\int_0^k G(r) f((m+1)k - r)\, dr - k f^m\right],$$

where G is the exact semigroup. Thus, after considering the analogous expression for l_h^0, consistency is equivalent to

$$\lim_{h\to 0} \max_{m\geq 0} k(h)^{-1} \|\left(x_h G(k) - G_h x_h\right) u^m\|_{B_h} = 0.$$

This condition is the same as the corresponding condition for the unforced problem (c.f. reference [24]).

5.4 A Lax Equivalence Theorem for Periodic Problems

Restriction Operators

For the periodically forced problem the restrictions must be defined differently, since $X = Y = P_T$, whilst $X_h = Y_h = P_{T,k(h)}$. The restriction $r_h : X \to X_h$ is defined in two steps.

(i) First $u \in X$ is restricted in time, via $u \mapsto \tilde{u}_h$, where[2] $\tilde{u}_h = [\tilde{u}^0, \tilde{u}^1, \ldots, \tilde{u}^{N_h - 1}]$, $(N_h \equiv T/k(h))$, and

$$\tilde{u}^m \equiv \frac{1}{k} \int_{(m-1/2)k}^{(m+1/2)k} u(s)\, ds.$$

[2]Note that only a representative cycle is described.

(ii) Next, $\tilde{u}_h$ is restricted in space: $\tilde{u}_h \mapsto r_h u$, where $r_h u = [x_h \tilde{u}^0, x_h \tilde{u}^1, \ldots, x_h \tilde{u}^{N_h - 1}]$, and $x_h : B \to B_h$ is a spatial restriction.

The Lax Equivalence Theorem

Proposition 5.2 *Suppose that* $\dim(B_h) < \infty$ *and* $r(G_h) < 1$ *for each* $h > 0$. *Further, suppose that* $\{x_h : B \to B_h\}_{h>0}$ *is a family of restrictions that satisfy condition (2.14). Then the assumptions of the Lax equivalence Theorem (page 36) are met, for stability and convergence in the time periodic norms.*

5.5 Analysis in Sobolev Spaces

Where the forcing is smooth it is more appropriate to use the Sobolev norm, rather than the L^2 norm, in space. This section sketches the modifications to the uniform convergence theory that are necessary to accommodate this change. For definiteness, an IBVP for the scalar linear advection equation is considered. The problem is posed on $x \in I$, with homogeneous boundary conditions at $x = 0$.

First we must return to the invertibility of Toeplitz operators. Let $\{A_h\}_{h>0} \subseteq L(B_h)$ be a collection of Toeplitz operators. Let $x_h : B \to B_h$ be spatial restrictions. Choose $B = H_0^p(I)$, the Sobolev space of functions that vanish at $x = 0$, and let B_h be the discrete counterpart of B. We consider the invertibility of $\{A_h\} \subseteq L(B_h)$.

Where the symbol of $\{A_h\}_{h>0}$ has a root of order q on the unit circle, $\|A_h^{-1}\|_{L(\ell_h^2)}$ grows as h^{-q}. However, $\|A_h^{-1} w_h\|_2$ will only grow at this rate when w_h is highly oscillatory — say, with 'frequency' ξ_0. However, for $p \neq 0$, B_h will not contain oscillatory grid functions. Indeed, were the boundary conditions to be periodic, it would be expected that the amount of energy at the crucial frequency ξ_0 would fall off as h^p [93]. Consequently, we conjecture that $\|A_h^{-1}\|_{L(B_h)}$ would grow as h^{p-q}. When $p = 0$, $B_h = \ell_h^2(I)$, and Proposition 4.13 is recovered.

The same reasoning used in Chapter 4 suggests that unitary ZGV schemes will have instabilities when the uniform norm is used in time. These instabilities will, at best, be weak of order $p - q + 1$.

For example, the Crank–Nicolson scheme has $q = 2$, so the instability may vanish when the solution lies in $H_0^p(I)$ with $p \geq 1$. Thus, the instability may not be observable when the exact solution is continuous. For $p \in [0, 1)$, the instability is expected to be $(1 - p)$-weak. The Box scheme is not expected to suffer uniform instabilities.

5.6 Conclusions

In Chapter 3, evolutionary problems were considered in the standard form $Qu = g$. Where equilibrium solutions are of interest, well-posedness should be defined by the uniform norms for both u and g. Chapter 4 considered uniform stability, and found the resolvent condition to be necessary. In the present chapter, we have shown that a Lax equivalence Theorem holds for these choices of norms.

This implies that any scheme which violates the resolvent condition will not converge for some data $g \in Y$: that is, for some combination of initial data and boundary forcing. However, the instability associated with the resolvent condition is usually weak ($\|Q_h^{-1}\| = Ch^{-\alpha}$),

unless boundary conditions are actually misplaced. Consequently, when g is sufficiently smooth, convergence may still occur.

The GKS and Godunov–Ryabenkii stability theories, and the uniform theory presented here, all agree that exponentially growing solutions must not be admitted by a method, if that method is to be 'stable' according to the definition of the particular theory. The three theories disagree over the stability of methods that admit wave-like solutions. The author has conjectured (see Conjecture 4.1) that any scheme that admits wave-like solutions (at a possibly isolated frequency), and that has zero group velocity at that frequency, must result in a uniformly unstable method. By contrast, the GKS theory is happy to admit such schemes, provided that the boundary conditions exclude from the final method all solutions with zero group velocity. The Godunov–Ryabenkii theory does not address this issue.

A consequence of these observations is that considerable care is needed to observe the lack of convergence associated with the uniform instability of (say) a GKS-stable method. The boundary forcing must be rough enough that a substantial amount of energy is present at the critical (temporal) frequency η; the spatial forcing must be rough enough that the maximum growth $\|(e^{i\eta} - G_h)^{-1}\|$ does actually occur. This observation explains why uniformly unstable methods are often used successfully in steady computations.

Chapter 6

Empirical Results

6.1 Introduction

For linear problems, the observation of uniform instabilities is not an easy matter. As for the Leapfrog method for the linear advection IVP and $\nu = 1$ [135, page 210], the instability is weak (i.e. the method is weakly stable).

There are at least three ways to observe weak instability of a discretization $Q_h u_h = s_h g$.

(i) Seek $g \in Y$ for which u_h does not converge to $r_h u$. For weak instabilities, this often involves finding $g \in Y$ such that u is a generalized solution of $Qu = g$.

(ii) Seek $g \in Y$ for which u_h converges to $r_h u$, but at a reduced order (c.f. page 37).

(iii) Use L-convergence theory (c.f. page 37).

We pause to consider how each of these approaches might be used to illustrate the (well-understood) weak L–R instability of the Leapfrog scheme when $\nu = 1$.

Instability of the Leapfrog Scheme

Consider the Leapfrog scheme for the linear advection IVP, with initial data $u^0 = s_h \phi$. This scheme has second order accuracy ($q = 2$). Choose the initial data $\phi \in H^p_{per}$: the 1-periodic Sobolev space on $\Re$. $\phi \in H^p_{per}$ iff the Fourier coefficients of this data $\hat{\phi}$ satisfy [93, page 42]

$$\sum_{m \in \mathcal{Z}} (1 + |m|)^{2p} |\hat{\phi}|^2 < \infty.$$

Thus $\hat{\phi}_m \sim |m|^{-(p+1/2+\epsilon)}$, for $\epsilon > 0$, guarantees $\phi \in H^p_{per}$.

A modification of Ansorge's approach [126, page 90] shows that the Leapfrog scheme is L–R convergent over $t \in [0, T]$ with order $pq/(p + 1/2)$, whenever $|\nu| < 1$. However, when $\nu = 1$, the Leapfrog scheme has the convergence order $p(q - 1)/(p + 1/2)$.

Consequently, verifying the weak instability by method (i) above requires computations with $\phi \in H^0_{per}$; smoother data will *always* converge. Verifying the weak instability by method (ii) requires measurement of the order of convergence. For the Leapfrog scheme ($q = 2$) with step function data ($p > 1/2$), we anticipate 1/2-order convergence for $|\nu| < 1$, and 1/4-order convergence for $\nu = 1$.

The weak instability can be verified by method (iii) by considering the sawtooth initial data

$$u^0_h = u^1_h = [\ldots, -1, 0, +1, 0, -1, 0, +1, \ldots].$$

The norm of discrete solution u_h^n grows linearly in n. However, it is not at all obvious that sawtooth initial data $g_h = [u_h^0, u_h^1, 0, 0, \ldots, 0]$ can result from the restriction of any $g \in Y$ for all $h > 0$. Therefore, approach (i) is *not* being used.

The use of such data must stem from an implicit programme of testing L-convergence (see page 37). This requires that the perturbed problem

$$Q_h u_h = s_h g + \delta_h$$

converge to $r_h u$, for any $\{\delta_h\}_{h>0} \subseteq Y_h$, with $\lim_{h\to 0} \|\delta_h\| = 0$. For the Leapfrog example, g and g_h are chosen to be zero, and δ_h is chosen to be a sawtooth of amplitude, say, $h^{1/2}$. Over time $T = N_h k(h)$, this results in $\|u_h\|_{I\times[0,T]} \sim h^{1/2} N_h \sim h^{-1/2}$, instead of zero. The scheme cannot converge to the exact solution (zero). Thus the Leapfrog scheme is not L-convergent, in the Lax–Richtmyer norms, when $\nu = 1$.

The Approach of this Chapter

Each of the above methods for verifying an instability requires different types of computation. Approaches (i)–(ii) require measurements of the order of convergence of methods for computing generalized solutions. Approach (iii) is similar, but requires that the forcing be deliberately perturbed.

In this chapter, we adopt approaches (i) and (iii). Only periodically forced problems are considered. Given this restriction, the convergence theory of Chapter 5 predicts that a consistent, uniformly stable method will be convergent for any (periodic) square integrable forcing $f \in P_T$. Furthermore, such methods should be L-convergent for perturbations $\{\delta_h\} \subseteq P_{T,k}$ such that $\lim_{h\to 0} \|\delta_h\| = 0$.

6.2 One-Dimensional Problems

6.2.1 Linear Advection Equation

This section uses L-convergence theory to demonstrate uniform instabilities for the linear advection equation $u_t + a u_x = f(x,t)$, on $(x,t) \in I$ with the (trivially time-periodic) forcing $f(x,t) = 0$. We regard this equation as being of the form $Qu = g$, where $Q : P_T \to P_T$. Chapter 3 shows that the problem is well-posed.

The numerical methods are of the form $Q_h u_h = s_h g = 0$, where $Q_h : P_{T,k} \to P_{T,k}$ (see Section 4.5.1). L-convergence (which is equivalent to $P_{T,k}$-stability and $P_{T,k}$-convergence) demands $u_h = Q_h^{-1}\delta_h$ to converge to the restriction of the exact solution $r_h u = 0$, whenever

$$\lim_{h\to 0} \|\delta_h\|_{P_{T,k}} = 0. \tag{6.1}$$

The perturbations $\delta_h \in P_{T,k}$ were chosen as follows. Let $\psi_h \in \ell_h^2(I)$ be

$$(\psi_h)_j \equiv \begin{cases} h^{-\alpha}, & jh = 1/2 \\ 0, & \text{otherwise} \end{cases}.$$

Clearly, $\|\psi_h\|_2 = h^{1/2-\alpha}$. Now define

$$\delta_h^m \equiv \sin(m\eta)\psi_h.$$

	$J = 20$		$J = 40$		$J = 80$
CNBOX	1.15(−1)		3.56(−1)		4.50(−1)
EOC		–		–	
Box	5.6550(−1)		3.13058(−1)		1.3370(−1)
EOC		0.85		1.22	

Table 6.1: Equilibrium errors E_h. Case (i).

With $T = 1$ and $N_h k = T$, η may be chosen so that $\delta_h \in P_{T,k}$. Then $\|\delta_h\|_{P_{T,k}} \sim h^{1/2-\alpha}$, and for $\alpha \in [0, 1/2)$ equation (6.1) holds. This choice of δ_h ensures a wide spread of spatial frequencies, but allows control of temporal frequencies.

The Box and Crank–Nicolson schemes were used with θ time-stepping. Specifically, the θ-Box scheme is

$$\frac{1}{2}\left[\frac{u_j^{n+1}-u_j^n}{k}+\frac{u_{j-1}^{n+1}-u_{j-1}^n}{k}\right]+a\theta\left[\frac{u_j^{n+1}-u_{j-1}^{n+1}}{h}\right]+a(1-\theta)\left[\frac{u_j^n-u_{j-1}^n}{h}\right]=$$
$$\frac{\theta}{2}\left[f_j^{n+1}+f_{j-1}^{n+1}\right]+\frac{1-\theta}{2}\left[f_j^n+f_{j-1}^n\right].$$

The θ-Crank–Nicolson scheme,

$$\frac{u_j^{n+1}-u_j^n}{k}+a\theta\left[\frac{u_{j+1}^{n+1}-u_{j-1}^{n+1}}{2h}\right]+a(1-\theta)\left[\frac{u_{j+1}^n-u_{j-1}^n}{2h}\right]=\theta f_j^{n+1}+(1-\theta)f_j^n,$$

was used with θ-Box boundary conditions at $x = 1$, resulting in the θ-CNBOX method. Both the θ-Box and θ-CNBOX methods may be viewed as having the same numerical boundary conditions at $x = 1$. Since $f(x,t) = 0$, $f_j^n = \delta_j^n$.

Both methods were used to integrate zero initial data through a large number p_h of cycles; thus $T_h = p_h$. The number of cycles was chosen so that the norm of the error over the final cycles was constant to the number of figures recorded. This equilibrium error, denoted E_h, was computed with the CFL number $\nu = 0.5$, and mesh sizes $J_h = N_h = 20, 40, 80$. Estimated (equilibrium) orders of convergence (EOCs) were computed.

The following parameters were used:

(i) $\theta = 0.5$, $\eta = \eta_{crit}$, $\alpha = 0.25$,

(ii) $\theta = 0.5$, $\eta = 0$, $\alpha = 0.25$,

(iii) $\theta = 0.6$, $\eta = \eta_{crit}$, $\alpha = 0.25$,

(iv) $\theta = 0.6$, $\eta = 0$, $\alpha = 0.25$.

Note that $\eta_{crit} = 2\arctan(\nu/2)$ is the critical frequency that excites the Crank–Nicolson instability (see Section 4.7.2). Results are presented in Tables 6.1–6.4.

	$J = 20$		$J = 40$		$J = 80$
CNBOX	2.11(−1)		1.25(−1)		7.48(−2)
EOC		0.76		0.74	
Box	1.5139(−1)		8.9461(−2)		5.3033(−2)
EOC		0.76		0.75	

Table 6.2: Equilibrium errors E_h. Case (ii).

	$J = 20$		$J = 40$		$J = 80$
CNBOX	9.4627(−1)		1.0592(−1)		4.2111(−2)
EOC		3.16		1.33	
Box	5.2175(−2)		2.7829(−2)		1.1741(−2)
EOC		0.91		1.25	

Table 6.3: Equilibrium errors E_h. Case (iii).

	$J = 20$		$J = 40$		$J = 80$
CNBOX	2.1147(−1)		1.2574(−1)		7.4767(−2)
EOC		0.75		0.75	
Box	1.5139(−1)		8.9468(−2)		5.3033(−2)
EOC		0.76		0.75	

Table 6.4: Equilibrium errors E_h. Case (iv).

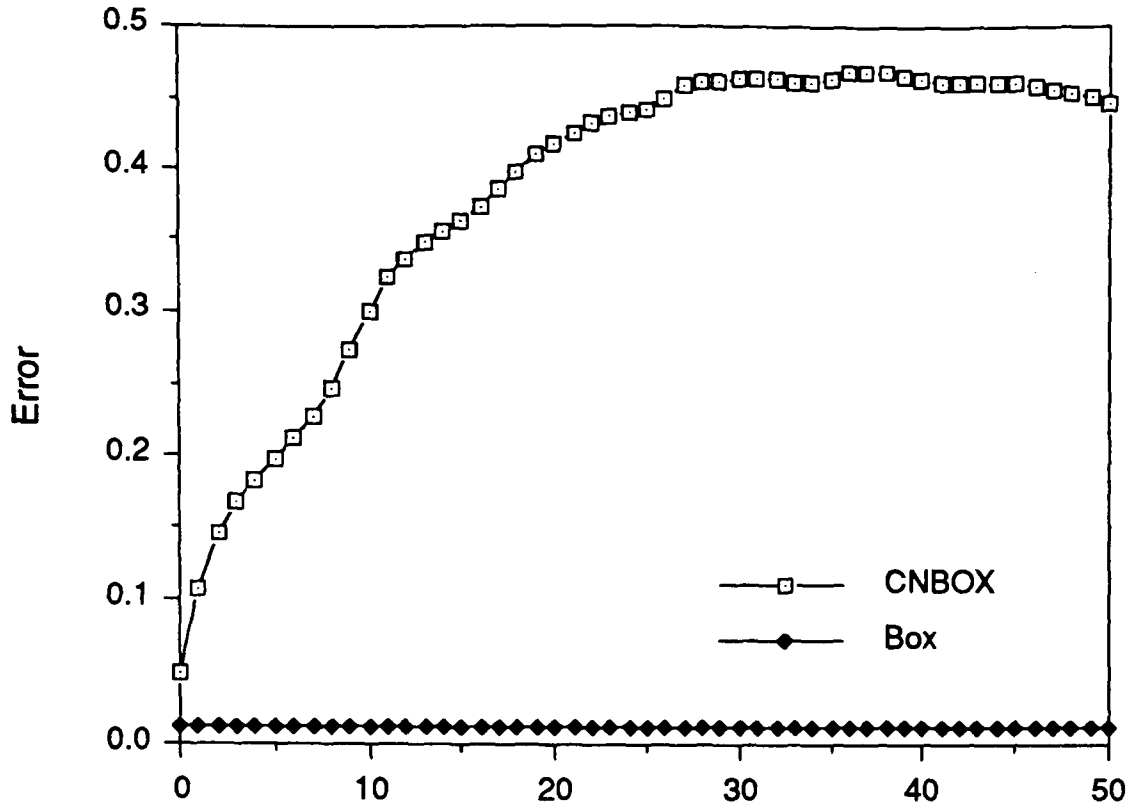


Figure 6.1: The error E over the pth cycle of forcing for the Box and CNBOX methods in Case (i).

In cases (i) and (ii) it was difficult to assess the equilibrium errors of the CNBOX method, since they still fluctuated after 50 cycles (see Figure 6.1). Consequently, these errors are only given to 3 figures in Tables 6.1–6.2. In contrast, the Box equilibrium errors settled down (to 4 figures) after just 1 or 2 cycles. In cases (iii) and (iv), all errors settled quickly.

Table 6.1 shows that the CNBOX method was not L-convergent, whilst the Box scheme was. This observation is in agreement with the predictions of Section 4.7.2. Table 6.2 shows that when the forcing is steady, the CNBOX instability vanishes. As we anticipated in Chapter 4, the instability is only excited for forcing that is non-constant in both space and time. Thus the unform instability need not present problems for steady computations.

Tables 6.3–6.4 show that the instability disappears when dissipation is added by increasing θ. Again, this observation is in agreement with the predictions of Section 4.7.2; dissipative schemes, when Fourier transformed in time, cannot result in polynomials $\pi(z, \eta)$ with double roots on the unit circle (see Proposition 4.14).

CFL numbers other than $\nu = 0.5$ were tested, but the results did not differ significantly from those presented above.

The comparison of the CNBOX and Box methods is appealing, since both methods may be viewed as having the same numerical boundary condition at $x = 1$. However, the possibility exists that the instability of the CNBOX method is the result of an unfortunate interaction between the interior scheme and the numerical boundary condition. For this reason, further numerical experiments were conducted with the Crank–Nicolson scheme, with horizontal and diagonal extrapolation boundary conditions at $x = 1$. The uniform instability remained. This observation is in agreement with the prediction of Chapter 4 that the uniform instability of the Crank–Nicolson scheme cannot be 'cured' by boundary conditions.

6.2.2 St Venant Equations

One-dimensional linear systems are of particular interest. It is well known that schemes that are Lax–Richtmyer stable, may be weakly (LR) unstable when applied to systems. This analogous possibility must be excluded for uniform stability.

The linearized, one-dimensional St Venant equations for a prismatic channel of reference depth $a = h_0$ are (c.f. equations (1.12)–(1.13))

$$h_t + a u_x = 0, \tag{6.2}$$

$$u_t + g h_x + r u = 0, \tag{6.3}$$

where r is a friction coefficient, g is gravitational acceleration, u is the velocity, and h is the depth.

We shall consider the model of a one-dimensional linear tidally forced estuary, or harbour, presented in reference [154, pages 95–99]. This problem is very similar to that considered by Bode and Sobey [11], and analytical solutions are available [11]. The estuary is modelled by equations (6.2,6.3) over $x \in [0, L]$. The sinusoidal forcing at $x = 0$ and no-flow condition at $x = L$ are imposed by

$$h(0,t) = \hat{a} \sin(2\pi\omega t), \quad u(L,t) = 0, \tag{6.4}$$

where $\omega = 2\pi/T$, T is the tidal period, and $\hat{a}$ is the tidal amplitude. The initial data is impulsive:

$$h(x,0) = 0, \; u(x,0) = 0, \quad x \in [0, L]. \tag{6.5}$$

Other choices of initial data are considered in reference [11].

The Exact Solution

The exact equilibrium solution is given by

$$h(x,t) = f(x)\cos\omega t + g(x)\sin\omega t, \quad u(x,t) = -\frac{\omega}{a}\left[\tilde{f}(x)\sin\omega t - \tilde{g}(x)\cos\omega t\right],$$

where

$$f(x) \equiv \frac{\hat{a}}{D}\left[c_1 c_4(x) - c_3 c_2(x)\right], \quad g(x) \equiv \frac{\hat{a}}{D}\left[c_1 c_2(x) + c_3 c_4(x)\right],$$

$$\tilde{f}(x) \equiv \frac{\hat{a}}{D}\left[c_1 s_4(x) - c_3 s_2(x)\right], \quad \tilde{g}(x) \equiv \frac{\hat{a}}{D}\left[c_1 s_2(x) + c_3 s_4(x)\right],$$

$$c_1 \equiv \cosh pL \cos qL, \quad c_3 \equiv \sinh pL \sin qL,$$

$$D \equiv c_1^2 + c_3^2,$$

$$c_2(x) \equiv \cosh p(L-x)\cos q(L-x), \quad c_4(x) \equiv \sinh p(L-x)\sin q(L-x),$$

$$s_2(x) \equiv \int_x^L c_2(s)\,ds, \quad s_4(x) \equiv \int_x^L c_4(s)\,ds$$

and

$$p \equiv \frac{1}{\sqrt{2}}\left[(k^4 + 4l^4)^{1/2} + k^2\right]^{1/2}, \quad q \equiv \frac{1}{\sqrt{2}}\left[(k^4 + 4l^4)^{1/2} - k^2\right]^{1/2},$$

$$k \equiv \frac{\omega}{\sqrt{ag}}, \quad l \equiv \sqrt{\frac{r\omega}{2ag}}.$$

This result is inflicted on the reader, because the result in reference [11, page 1394] contains a minor error.

The Box Scheme

When the Box scheme is applied to the linearized equations, a specially structured, penta-diagonal implicit matrix must be inverted at each time step [154, page 102]. Setting

$$\alpha \equiv \frac{kr}{2}, \quad \beta \equiv a\nu, \quad \gamma \equiv g\nu, \quad \nu \equiv \frac{k}{h},$$

for $J = 2$ the method is of the form $S_h w_h^{n+1} = T_h w_h^n + f_h^n$, where

$$S_h = \begin{bmatrix} -\beta & 1 & \beta & & & \\ 1+\alpha & \gamma & 1+\alpha & & & \\ & 1 & -\beta & 1 & \beta & \\ & -\gamma & 1+\alpha & \gamma & 1+\alpha & \\ & & & 1 & -\beta & 1 \\ & & & -\gamma & 1+\alpha & \gamma \end{bmatrix},$$

$$T_h = \begin{bmatrix} \beta & 1 & -\beta & & & \\ 1-\alpha & -\gamma & 1-\alpha & & & \\ & 1 & \beta & 1 & -\beta & \\ & \gamma & 1-\alpha & -\gamma & 1-\alpha & \\ & & & 1 & \beta & 1 \\ & & & \gamma & 1-\alpha & -\gamma \end{bmatrix},$$

and

$$w_h^n = [u_0^n, h_1^n, u_1^n, h_2^n, u_2^n, h_3^n]^T, \quad f_h^n = [h_0^n - h_0^{n+1}, 0, 0, 0, 0, 0]^T.$$

(Similarly, each Newton iteration of the full non-linear equations involves the solution of a specially structured penta-diagonal linear system.) Such linear systems are easily solved using a modification of the Thomas algorithm.

The Crank–Nicolson Scheme

If the Crank–Nicolson method is applied to the linearized St Venant equations, with suitable staggering of the differences in the momentum and conservation equations [154, page 100], then a tri-diagonal system must be solved at each time step. For example, setting

$$\alpha \equiv \frac{kr}{2}, \quad \beta \equiv \frac{a\nu}{4}, \quad \gamma \equiv \frac{g\nu}{4},$$

where $\nu = k/h$, then for $J = 6$ the method is of the form $S_h w_h^{n+1} = T_h w_h^n + f_h^n$, where

$$S_h = \begin{bmatrix} 1+\alpha & \gamma & & & & \\ -\beta & 1 & \beta & & & \\ & -\gamma & 1+\alpha & \gamma & & \\ & & -\beta & 1 & \beta & \\ & & & -\gamma & 1+\alpha & \gamma \\ & & & & -\beta & 1 \end{bmatrix},$$

$$T_h = \begin{bmatrix} 1-\alpha & -\gamma & & & & \\ \beta & 1 & -\beta & & & \\ & \gamma & 1-\alpha & -\gamma & & \\ & & \beta & 1 & -\beta & \\ & & & \gamma & 1-\alpha & -\gamma \\ & & & & \beta & 1 \end{bmatrix},$$

	$J = 10$		$J = 20$		$J = 40$		$J = 80$
BOX	3.0546(−3)		1.7726(−3)		9.5470(−4)		4.9498(−4)
EOC		0.79		0.89		0.95	
CN	1.2594(−3)		8.6029(−4)		5.0076(−4)		2.6851(−4)
EOC		0.55		0.55		0.89	

Table 6.5: Equilibrium errors and Estimated Orders of Convergence for the forced St Venant equations.

and

$$w_h^n = [u_1^n, h_2^n, u_3^n, h_4^n, u_5^n, h_6^n]^T, \quad f_h^n = [\gamma(h_0^{n+1} + h_0^n), 0, 0, 0, 0, 0]^T.$$

Results

The St Venant equations were integrated using the Box and Crank–Nicolson schemes to periodic equilibrium. The following values were used (c.f. [154, page 97]): $\hat{a} = 0.1m$, $a = 10m$, $L = 10km$, $T = 3260$s, $r = 0.005$. The grids were chosen so that $J = N = 10, 20, 40$, where N is the number of time steps per cycle. The integration was carried out until the error was steady to the number of decimal places given. The numerical solutions were compared with the exact solutions, and EOCs computed. The results appear in Table 6.5. Both schemes seem to have equilibrium orders of roughly 1, although the results for the Crank–Nicolson scheme are not clear.

6.3 Linear Advection Equation and Spurious Modes

Steady equilibrium solutions were discussed in Section 4.7.5. It was claimed that schemes that admit spurious modes cannot result in stable methods, regardless of numerical boundary conditions: boundary conditions may control spurious modes that arise from boundary forcing, but they cannot control spurious modes that arise from interior forcing. This section presents illustrations of this claim.

6.3.1 ℓ_h^2-stability

We consider the two-dimensional scalar linear advection equation

$$au_x + bu_y = f(x, y), \quad (x, y) \in \Omega \equiv I^2,$$

with homogeneous boundary conditions at $x = 0$ and $y = 0$. The problem is well-posed in $L^2(\Omega)$ when $a > 0$ and $b > 0$.

Two methods were used: the cell-vertex method; and the method arising from the cell-centred scheme teamed with Box (i.e. cell-vertex) numerical boundary conditions at $x = 1$ and $y = 1$. Both methods may be viewed as having the same numerical boundary conditions.

	$J = 10$		$J = 20$		$J = 40$
Cell-centred	7.3316(0)		2.1450(1)		6.7319(1)
p		1.54		1.65	
Cell-vertex	4.9310(−1)		4.4817(−1)		4.4414(−1)
p		−0.14		−0.02	

Table 6.6: Approximate values of $\|Q_h^{-1}\|$ for the cell-centred and cell-vertex methods.

The Cell-Centred System

To minimize the possibility of the time-stepping algorithm influencing the result, the linear systems were solved directly. For the cell-vertex method, this presents no difficulty. However, for the cell-centred method, a block tri-diagonal system must be solved. For $J = 4$, this system ($\tilde{Q}_h u_h = h g_h$) takes the form

$$\tilde{Q}_h \equiv \begin{bmatrix} B & C & 0 & 0 \\ A & B & C & 0 \\ 0 & A & B & C \\ 0 & 0 & Q & R \end{bmatrix}, \quad Q \equiv \begin{bmatrix} -r & 0 & 0 & 0 \\ -q & -r & 0 & 0 \\ 0 & -q & -r & 0 \\ 0 & 0 & -q & -r \end{bmatrix}, \quad R \equiv \begin{bmatrix} q & 0 & 0 & 0 \\ r & q & 0 & 0 \\ 0 & r & q & 0 \\ 0 & 0 & r & q \end{bmatrix},$$

and

$$A \equiv \begin{bmatrix} -a & 0 & 0 & 0 \\ 0 & -a & 0 & 0 \\ 0 & 0 & -a & 0 \\ 0 & 0 & -q & -r \end{bmatrix}, \quad B \equiv \begin{bmatrix} 0 & b\rho & 0 & 0 \\ -b\rho & 0 & b\rho & 0 \\ 0 & -b\rho & 0 & b\rho \\ 0 & 0 & r & q \end{bmatrix}, \quad C \equiv \begin{bmatrix} a & 0 & 0 & 0 \\ 0 & a & 0 & 0 \\ 0 & 0 & a & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where $q \equiv a + b\rho$, $r \equiv a - b\rho$, $\rho = h_x/h_y$, and $h_x = h_y = h = 1/J$ are the constant grid spacings. The system was solved using a block variation of the Thomas algorithm, with iterative improvement. Care is required in the solution, as we anticipate that $\tilde{Q}_h$ may be almost singular. Of course, for practical computations, the system would never be solved directly: some type of time-stepping would always be used [97,90].

Results

The Box and cell-centred methods result in systems $Q_h u_h = g_h$. To estimate $\|Q_h^{-1}\|_{L(\ell_h^2(I))}$, a sequence of 40 random vectors $g_h(i)$, $i = 1, \ldots, 40$ were generated. We approximate

$$\|Q_h^{-1}\| \approx \max_{i=1,\ldots,40} \frac{\|u_h(i)\|_2}{\|g_h(i)\|_2}, \tag{6.6}$$

for $h = h_x = h_y = 1/J$ and $J = 10$, 20, 40. For the calculation, we chose $a = 1$ and $b = 0.5$. The results appear in Table 6.6, together with estimates for the exponent p such that $\|Q_h^{-1}\| \sim h^{-p}$.

6.3.2 Generalized Solutions

The problem to be considered is scalar:

$$au_x + bu_y = f(x,y), \quad (x,y) \in \Omega,$$

where $\Omega = I^2$, and

$$u(x,0) = u(0,y) = 0, \quad x,y \in I.$$

The solution u was chosen in advance, and the corresponding interior forcing f was calculated.

Two solutions will be considered: the smooth solution

$$u(x,y) = \sin \pi x \sin \pi y, \quad f(x,y) = \pi \left[a \cos \pi x \sin \pi y + b \sin \pi x \cos \pi y\right], \tag{6.7}$$

and the rough 'cube' solution

$$u(x,y) = \begin{cases} 1, & (x,y) \in [\frac{1}{4}, \frac{3}{4}]^2 \\ 0 & \text{otherwise} \end{cases} \tag{6.8}$$

Restrictions

The forcing space Y and solution space X were both taken as $L^2(\Omega)$, whilst $X_h = Y_h = \ell_h^2(\Omega)$. A restriction $r_h : X \to X_h$ is needed to define the error

$$E_h \equiv \|u_h - r_h u\|_{\ell_h^2(\Omega)}.$$

Also, a restriction $s_h : Y \to Y_h$ is needed to define $f_h \equiv s_h f$.

For the cell-centred method, we take both to be defined by cell averages. Thus for $u \in X = Y$

$$(s_h u)_{i,j} = (r_h u)_{i,j} \equiv \frac{1}{h^2} \int_{\Omega_{i,j}} u(x,y)\, dx\, dy,$$

where $\Omega_{i,j}$ is that part of the cell (of side length h and centred on (ih, jh)) that lies in Ω. It is easily verified that r_h and s_h satisfy the requirements of a restriction operator.

If ϕ is the characteristic function of the interval $[\frac{1}{4}, \frac{3}{4}]$, then $\phi'(s) = \delta_{\frac{1}{4}}(s) - \delta_{\frac{3}{4}}(s)$. With the above choice of restrictions, it is easily seen that for the cube solution,

$$(f_h)_{i,j} = (s_h f)_{i,j} = \begin{cases} a/h, & x = \frac{1}{4}, & y \in (\frac{1}{4}, \frac{3}{4}) \\ -a/h, & x = \frac{3}{4}, & y \in (\frac{1}{4}, \frac{3}{4}) \\ b/h, & x \in (\frac{1}{4}, \frac{3}{4}), & y = \frac{1}{4} \\ -b/h, & x \in (\frac{1}{4}, \frac{3}{4}), & y = \frac{3}{4} \\ (a+b)/2h, & x = \frac{1}{4}, & y = \frac{1}{4} \\ (-a+b)/2h, & x = \frac{3}{4}, & y = \frac{1}{4} \\ (a-b)/2h, & x = \frac{1}{4}, & y = \frac{3}{4} \\ (-a-b)/2h, & x = \frac{3}{4}, & y = \frac{3}{4} \end{cases}$$

Here, $x = \frac{1}{4}$ is a shorthand for $ih = \frac{1}{4}$, and so forth.

For the cell-vertex method, forcing is required at points $((i-1/2)h_x, (j-1/2)h_y)$. This forcing was represented by the average over the neighbours

$$(f_h)_{i-1/2,j-1/2} \equiv \frac{1}{4}\left[(f_h)_{i,j} + (f_h)_{i-1,j} + (f_h)_{i,j-1} + (f_h)_{i-1,j-1}\right].$$

With this choice of restrictions and forcing, the cell-vertex scheme is exact whenever J is divisible by 4.

	$J = 10$		$J = 20$		$J = 40$
Cell-centred	6.3668(−2)		2.0486(−2)		8.3295(−3)
EOC		1.64		1.29	
Cell-vertex	9.0958(−3)		5.1099(−3)		2.7280(−3)
EOC		0.83		0.91	

Table 6.7: Errors E_h with smooth forcing (6.7) and $\rho = 0.5$, for the steady linear advection equation.

	$J = 10$		$J = 20$		$J = 40$
Cell-centred	5.000(−1)		1.242(1)		3.398(1)
EOC		—		—	
Cell-vertex	5.000(−1)		0.000(0)		0.000(0)
EOC		—		—	

Table 6.8: Errors E_h with rough forcing (6.8) and $\rho = 0.5$, for the steady linear advection equation.

Results

For the computations, $a = 1$ and $b = \rho$ were chosen. Thus ρ is the CFL number. The cell-averages $f_h = s_h f$ were calculated analytically. Computations were performed using the two methods for $J = 10$, 20, 40, with $\rho = 0.5$,. The errors E_h for the smooth forcing (6.7) appear in Table 6.7, whilst the errors for the rough forcing (6.8) appear in Table 6.8.

For the smoothly forced problem, both methods converged, and the cell-vertex method converged with order 1. For the roughly forced problem, the cell-vertex scheme converged rapidly, but the cell-centred method did not seem to converge at all.

Finally, the software was tested with $f(x, y) = ay + bx$, corresponding with $u(x, y) = xy$. Both methods obtained the exact answer (to machine precision), even on crude meshes.

6.4 Conclusions

The predictions of the uniform stability theory developed in Chapter 4 have been verified, for the forced one-dimensional linear advection equation, in that we have shown that the Crank–Nicolson scheme is not L-convergent, whilst the Box scheme is L-convergent, for a periodically forced problem. Furthermore, we have shown that the Crank–Nicolson scheme may be L-convergent when steady forcing or θ-time-stepping are used. These observations are in complete agreement with the predictions of uniform convergence theory.

The stability of the cell-centred and cell-vertex methods for the two-dimensional linear advection equation has also been investigated. We conclude that the cell-centred method is weakly unstable in the $L(\ell_h^2(I))$ sense, but the cell-vertex method is stable. These results agree with the predictions of Section 4.7.5, and correspond to the fact that the cell-centred scheme admits spurious modes, whilst the cell-vertex scheme does not.

When non-linear computations are attempted with a method that is uniformly unstable (for linear problems), energy may be transferred between different frequencies in the spatial and temporal power spectra. Shocks may form, and a non-linear problem with smooth forcing may tax a method as much as a linear problem with rough forcing. Indeed, Morton and Paisley [97, page 16] have noted that spurious modes (which we have shown to be related to uniform instabilities)

> ... can be triggered at shocks, boundaries or even non-uniformities in the mesh and usually have to be damped out.

Under circumstances such as these, uniform instabilities may be exacerbated. It would be interesting to test this possibility for non-linear evolutionary problems. Unfortunately, there was insufficient time to perform such computations for inclusion in the present work.

Chapter 7

Conclusions

This thesis began from the observation that there is an interesting class of analytic solutions of evolutionary problems that, in some sense, lie between steady and dynamic solutions. This observation lead to criteria, given in Section 3.5, that it was felt such 'equilibrium solutions' should meet. For linear dissipative problems it has been shown that equilibrium solutions — meeting these criteria — exist uniquely. For more general problems, the criteria have been given succinct expression in terms of processes (Section 3.5.4), though further work is needed to identify a class of functional differential equations for which equilibrium solutions exist uniquely. Hale's work [50] may well offer a suitable class.

The numerical computation of equilibrium solutions requires that the time interval increase as the mesh is refined, to prevent the numerical errors from being swamped by analytical transients. Quite how the length of time increases as the mesh is refined depends on the stopping procedure used for the evolutionary initialization. To ensure convergence irrespective of the stopping procedure, it was felt that the best course was to find a theory that ensured convergence over $t \in \Re^+$.

This course led to difficulties. Existing convergence theories for discretizations apply only over finite time intervals (GKS theory), or fail when the forcing does not decay (extended Lax–Richtmyer theory). Furthermore, the failure of these theories is real, and can be demonstrated in computations. Consequently, a new convergence theory was required.

Proving the convergence of a discretization over $t \in \Re^+$ requires that the analytic problem be well posed over $t \in \Re^+$. The most direct approach is to use the 'semigroup' estimate, which is equivalent to uniform asymptotic stability. However, even for linear hyperbolic problems, proving u.a.s. with only a knowledge of the generator of the semigroup is not easy, and is the topic of recent research. This research was reviewed in Chapter 3.

Once well-posedness has been proved via the semigroup estimate, one has no choice but to define stability in terms of the corresponding discrete norm [126, page 89]. Consequently, Chapter 4 was written to reflect, as closely as possible, the structure of Chapter 3. The crucial difference is that the discrete semigroup depends on the mesh parameter. This means that the various types of asymptotic stability must be defined so that they hold uniformly with regard to the mesh parameter. Then stability in time-uniform norms is equivalent to numerical uniform asymptotic stability.

Unlike the case of the analytic problem, the generator of the discrete semigroup is usually not known. Even if we disregard the added difficulties introduced by the mesh parameter, the analytic theory is of little help in deciding n.u.a.s. However, the coefficients of the interior scheme are known. Chapter 4 derived necessary conditions, involving the scheme's

coefficients, for u.n.a.s., and hence for stability.

It was then shown that the necessary conditions imply that no scheme that admits waves of zero group velocity can be stable in the new norms. In contrast, a scheme that admits waves of zero group velocity may be GKS stable, provided that the boundary conditions exclude those waves. This highlights an essential difference between GKS stability and uniform stability.

The effects of boundary conditions on the uniform stability of schemes have yet to be investigated. Rudy and Strikwerda present some interesting results in reference [124] that can be interpreted in terms of n.u.a.s. and uniform stability.

It is difficult to assess the practical importance of the uniform convergence theory presented in this thesis. It is certainly true that for linear problems the instability only manifests itself seriously when the solution is discontinuous (c.f. Section 5.5), or under local mesh refinement. However, for non-linear problems shocks can form, and discontinuities are a real possibility. Further investigation along these lines would be interesting.

Rather unexpectedly, it has been found that uniform stability also sheds new light on the steady state problem of spurious modes. Indeed, it was shown in Section 4.7.5 that spurious modes are steady uniform instabilities. Section 4.7.5 goes on to present criteria that help resolve why some modes should be considered 'spurious' and others 'genuine'. Section 5.5 lays an analytical foundation for understanding why spurious modes can be troublesome.

The results presented here are also novel, the author believes, in emphasizing the importance of forcing in stability — an aspect which seems to have been neglected in the past. If this thesis has a wider moral, it is that the stability of $Q_h u_h = g_h$ means *more* than $\ker(Q_h) = \{0\}$!

The neglect of forcing in stability studies has led to the situation that the weak instability of the Leapfrog scheme at $\nu = 1$ for IVPs for $u_t + a u_x = 0$ is common knowledge; but the exactly analogous instability for the Crank–Nicolson scheme, applied to the corresponding BVP, is not recognized in the literature. This neglect cannot be ascribed to the slightness of the instability. Indeed, the two instabilities are identical in their severity. Once Fourier-transformed in space, the Leapfrog instability corresponds to a violation of the Dahlquist–Henrici root condition for ODE methods. Similarly, once Fourier-transformed in time, the Crank–Nicolson method corresponds to an ODE method for a two-point BVP, whose symbol has a double root on the unit circle. In both cases, the instability is 1-weak.

Similarly, in the context of steady equilibrium solutions, spurious modes for the problem $a u_x + b u_y = 0$ present no difficulties. However, we have shown that they do present difficulties when we consider the problem $a u_x + b u_y = f$.

Uniform stability may explain a number of other phenomena reported in the literature. In particular, [158,139,159,150] have reported that the ADI-factorized scheme due to Beam and Warming [5] shows poor accuracy when used on domains that are not aligned with the computational grid. This scheme admits waves of zero group velocity. It is possible that the poor accuracy is attributable to uniform instabilities, rather than to the ADI factorization [158].

Uniform instability may also explain certain effects previously attributed to 'non-linear' instability. For example, the following quotation from reference [154, page 105] concerns versions of the Box and Crank–Nicolson schemes with θ time-stepping.

> The factor θ turns out to have a very great effect in the wave damping; apparently you should use a value equal or close to 1/2. In practice, 0.52 or 0.55 is often

used. The reason is that there is no damping at all for $\theta = 1/2$. Any destabilizing effect due to, e.g. nonlinear terms may then lead to instability.

The results presented in this thesis support this conclusion for the CN scheme, but using only linear reasoning. However, our linear analysis did not predict any difficulties with the Box scheme when $\theta = 1/2$. Indeed, none were observed.

Bibliography

[1] R.K. Agarwal. Computational fluid dynamics on parallel processors. In *Proceedings of 5th Distributed Memory Computing Conference, Charleston, South Carolina*, 1990. In press.

[2] J. Aubin. *Applied Functional Analysis.* Pure and Applied Mathematics Series. Wiley–Interscience, New York, 1979.

[3] G.R. Baldock and T. Bridgeman. *The Mathematical Theory of Wave Motion.* Ellis Horwood Series in Mathematics and its Applications. Ellis Horwood (Wiley), Chichester, 1981.

[4] J.W. Barker. Interactions of fast and slow waves in problems with two time scales. *SIAM J. Math. Anal.*, 15(3):500, 1984.

[5] R.M. Beam and R.F. Warming. An implicit finite-difference algorithm for hyperbolic systems in conservation-law form. *J. Comp. Phys.*, 22:87, 1976.

[6] R.M. Beam, R.F. Warming, and H.C. Yee. Stability analysis of numerical boundary conditions and implicit difference approximations for hyperbolic equations. *J. Comp. Phys.*, 48:200, 1982.

[7] A. Belleni-Morante. *Applied Semigroups and Evolution Equations.* Clarendon Press, Oxford, 1979.

[8] P.L. Bhatnagar. *Nonlinear Waves in One-Dimensional Dispersive Systems.* Clarendon Press, Oxford, 1979.

[9] R.L. Bisplinghoff, H. Ashley, and R.L. Halfman. *Aeroelasticity.* Addison–Wesley Series in Mechanics. Addison–Wesley, Reading, Mass., 1955.

[10] D.R. Bland. *Wave Theory and Applications.* Oxford Applied Mathematics and Computing Science Series. Clarendon Press, Oxford, 1988.

[11] L. Bode and R.J. Sobey. Initial transients in long wave computations. *J. Hydraulic Eng.*, 110(10):1371, 1984.

[12] F.F. Bonsall and J. Duncan. Numerical ranges. In R.G. Bartle, editor, *Studies in Functional Analysis.* Mathematical Association of America, 1980. MAA Studies in Mathematics, 21.

[13] R.K. Brayton and W.L. Miranker. A stability theory for nonlinear mixed initial boundary value problems. *Arch. Rath. Mech. Anal.*, 17:358, 1966.

[14] L. Brillouin. *Wave Propagation and Group Velocity.* Academic Press, London, 1960.

[15] G. Browning, A. Kasahara, and H.-O. Kreiss. Initialization of the primitive equations by the bounded derivative method. *J. Atmos. Sci.*, 37:1425, 1980.

[16] G. Browning and H.-O. Kreiss. Initialization of the shallow water equations with open boundaries by the bounded derivative method. *Tellus*, 34:334, 1982.

[17] G. Browning and H.-O. Kreiss. Problems with different times scales for nonlinear partial differential equations. *SIAM J. Appl. Math.*, 42(4):704, 1982.

[18] N. Burgess. *Stable Boundary Conditions for the Shallow Water Equations.* PhD thesis, University of Oxford, 1987.

[19] N.A. Burgess. The stability of a simple IBVP. Technical Report 86/2, Oxford University Computing Laboratory, Numerical Analysis Group, 1986.

[20] T.A. Burton. *Stability and Periodic Solutions of Ordinary and Functional Differential Equations*, volume 178 of *Mathematics in Science and Engineering.* Academic Press, London, 1985.

[21] V. Casulli. Semi-implicit finite difference methods for the two-dimensional shallow water equations. *J. Comp. Phys.*, 86:56–74, 1990.

[22] V. Casulli and D. Greenspan. Pressure method for the numerical solution of transient, compressible fluid flows. *Int. J. Numer. Meth. Fluids*, 4:1001, 1984.

[23] E.T. Copson. *An Introduction to the Theory of Functions of a Complex Variable.* Oxford University Press, Oxford, 1935.

[24] C.W. Cryer. *Numerical Functional Analysis.* Monographs on Numerical Analysis. Clarendon Press, Oxford, 1982.

[25] M.J. Cullen and K.W. Morton. Analysis of evolutionary error in finite element and other methods. *J. Comp. Phys.*, 34(2):245, 1980.

[26] A.S. Davydov. *Quantum Mechanics*, volume 1 of *International Series of Monographs in Natural Philosophy.* Pergamon Press, Oxford, 1965.

[27] K. Dekker and J.G. Verwer. *Stability of Runge–Kutta Methods for Stiff Nonlinear Differential Equations*, volume 2 of *CWI Monographs.* North–Holland, Amsterdam, 1984.

[28] R.G. Douglas. On the invertibility of a class of Toeplitz operators on the quarter-plane. *Indiana Univ. Math. J.*, 21(11):1031, 1972.

[29] R.G. Douglas and R. Howe. On the C^*-algebra of Toeplitz operators on the quarter plane. *Trans. Amer. Math. Soc.*, 158(1):203, 1971.

[30] R.V. Duduchava. Solution of a convolution equation on the quarter-plane. *Math. Notes*, 27:207, 1980.

[31] N. Dunford and J.T. Schwartz. *Linear Operators, Part I*, volume VII of *Pure and Applied Mathematics.* Interscience, New York, 1958.

[32] J.I. Erdos, E. Alzner, and W. McNally. Numerical solution of periodic transonic flow through a fan stage. *AIAA J.*, 15(11):1559, 1977.

[33] M.G. Foreman. An accuracy analysis of boundary conditions for the forced shallow water equations. *J. Comp. Phys.*, 64:334, 1986.

[34] I.E. Garrick. Nonsteady wing characteristics. In A.F. Donovan and H.R. Lawrence, editors, *Aerodynamic components of Aircraft at High Speeds. Volume 7*, chapter F. Oxford University Press, Oxford, 1957.

[35] M. Giles. UNSFLO: A numerical method for unsteady inviscid flow in turbomachinery. Technical Report 195, MIT Gas Turbine Laboratory, October 1988.

[36] M.B. Giles. Developments in the calculation of unsteady turbomachinery flow. In K.W. Morton and M.J. Baines, editors, *Numerical Methods in Fluid Dynamics III.* Clarendon Press, Oxford, 1988.

[37] A.E. Gill. *Atmosphere-Ocean Dynamics*, volume 30 of *International Geopysics Series.* Academic Press, Orlando, 1982.

[38] S.K. Godunov and V.S. Ryabenkii. *Difference Schemes*, volume 19 of *Studies in Mathematics and its Applications.* North–Holland, Amsterdam, 1987.

[39] J.A. Goldstein. *Semigroups of Linear Operators and Applications.* Oxford Mathematical Monographs. Clarendon Press, Oxford, 1985.

[40] G.H. Golub and C.F. Van Loan. *Matrix Computations.* North Oxford Academic, Oxford, 1986.

[41] D.A. Greenberg. Modeling tidal power. *Scientific American*, page 106, November 1987.

[42] J. Guerra and B. Gustafsson. A semi-implicit method for hyperbolic problems with different time-scales. *SIAM J. Numer. Anal.*, 23(4):734, 1986.

[43] B. Gustafsson. The convergence rate for difference approximations to mixed initial boundary value problems. *Math. Comp.*, 29(130):396, 1975.

[44] B. Gustafsson. Asymptotic expansions for hyperbolic problems with different time scales. *SIAM J. Numer. Anal.*, 17(5):623, 1980.

[45] B. Gustafsson. Numerical solution of hyperbolic systems with different time scales using asymptotic expansions. *J. Comp. Phys.*, 36:209, 1980.

[46] B. Gustafsson. The choice of numerical boundary conditions for hyperbolic systems. *J. Comp. Phys.*, 48:270, 1982.

[47] B. Gustafsson, H.-O. Kreiss, and A. Sundström. Stability theory of difference approximations for mixed initial boundary value problems. II. *Math. Comp.*, 26(119):649, 1972.

[48] J.K. Hale. *Theory of Functional Differential Equations*, volume 3 of *Applied Mathematical Sciences.* Springer–Verlag, New York, 1977.

[49] J.K. Hale. Asymptotic behaviour and dynamics in infinite dimensions. In J.K. Hale and P. Martínez-Amores, editors, *Nonlinear Differential Equations.* Pitman, London, 1985. Research Notes in Mathematics. 132.

[50] J.K. Hale. *Asymptotic Behavior of Dissipative Systems*, volume 25 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, Rhode Island, 1988.

[51] J.K. Hale, L.T. Magalhães, and W.M. Oliva. *An Introduction to Infinite Dimensional Dynamical Systems — Geometric Theory*, volume 47 of *Applied Mathematical Sciences*. Springer–Verlag, New York, 1984.

[52] G. Haltiner. *Numerical Weather Prediction*. Wiley, New York, 1971.

[53] F.H. Harlow and A.A. Amsden. Numerical calculation of almost incompressible flow. *J. Comp. Phys.*, 3:80, 1968.

[54] F.H. Harlow and A.A. Amsden. A numerical fluid dynamics calculation method for all flow speeds. *J. Comp. Phys.*, 8:197, 1971.

[55] T.H. Havelock. *The Propagation of Disturbances in Dispersive Media*. Stechert–Hafner Service Agency, New York, 1964.

[56] R. Hersch. Mixed problems in several variables. *J. Math. Mech.*, 12:317, 1963.

[57] J.G. Heywood and R. Rannacher. Finite element approximations of the Navier–Stokes problem. Part II: stability of solutions and error estimates uniform in time. *SIAM J. Numer. Anal.*, 23:750, 1966.

[58] R.L. Higdon. Boundary conditions for suppressing rapidly moving components in hyperbolic systems. *SIAM J. Numer. Anal.*, 21(3):413, 1984.

[59] R.L. Higdon. Initial-boundary value problems for linear hyperbolic systems. *SIAM Rev.*, 28(2):177, 1986.

[60] E. Hille and R.S. Phillips. *Functional Analysis and Semigroups*, volume 31 of *Colloquium Publications*. American Mathematical Society, Providence, Rhode Island, 1957.

[61] J.R. Holton. *An Introduction to Dynamic Meteorology*, volume 23 of *International Geophysics Series*. Academic Press, London, 1979.

[62] E. Isaacson and H.B. Keller. *Analysis of Numerical Methods*. Wiley, New York, 1966.

[63] A. Iserles and G. Strang. The optimal accuracy of difference schemes. *Trans. Amer. Math. Soc.*, 277(2):779, 1983.

[64] S.M. Jamieson and D.M. Sloan. A Petrov–Galerkin method and boundary approximations for hyperbolic initial boundary-value problems. Technical report, University of Strathclyde, Department of Mathematics, 1985.

[65] R. Jeltsch, P. Kiani, and K. Raczek. Counterexamples to a stability barrier. *Numer. Math.*, 52:301, 1988.

[66] R. Jeltsch and O. Nevanlinna. Analysis of time discretizations. In J.K. Hale and P. Martínez-Amores, editors, *Nonlinear Differential Equations*. Pitman, 1985. Research Notes in Mathematics, 132.

[67] R. Jeltsch and K. Raczek. Counter examples to an order barrier for stable multistep discretizations of linear hyperbolic equations. Technical Report 39, Templergraben, Federal Republic of Germany, October 1986.

[68] R. Jeltsch and J.H. Smit. Accuracy barriers of two time level difference schemes for hyperbolic equations. *SIAM J. Numer. Anal.*, 24(1):1, 1987.

[69] D.C. Jespersen. A time-accurate multi-grid algorithm. In *Proceedings of the AIAA 7th Computational Fluid Dynamics Conference, Cincinnati, OH*, 1985. AIAA Paper 85–1493–CP.

[70] G.A. Jones and D. Singerman. *Complex Functions. An algebraic and geometric viewpoint.* Cambridge University Press, Cambridge, 1987.

[71] D. Jordan and P. Smith. *Nonlinear Ordinary Differential Equations.* Clarendon Press, Oxford, 1977.

[72] A.A. Kirillov and A.D. Gvishiani. *Theorems and Problems in Functional Analysis.* Springer–Verlag, 1982.

[73] S.G. Krein. *Linear Differential Equations in Banach Space*, volume 29 of *Translations of Mathematical Monographs.* American Mathematical Society, Providence, Rhode Island, 1972.

[74] H.-O. Kreiss. Stability theory for difference approximations of mixed initial boundary value problems. *Math. Comp.*, 22:703, 1968.

[75] H.-O. Kreiss. Initial boundary value problems for hyperbolic systems. *Comm. Pure Appl. Math.*, XXIII:277, 1970.

[76] H.-O. Kreiss. Problems with different time scales for partial differential equations. *Comm. Pure Appl. Math.*, XXXIII:399, 1980.

[77] E. Kreyszig. *Introductory Functional Analysis with Applications.* Wiley, New York, 1978.

[78] M. Kwizac and A. Robert. A semi-implicit scheme for grid-point atmospheric models of the primitive equations. *Mon. Weather Rev.*, 99:32, 1971.

[79] J.D. Lambert. *Computational Methods in Ordinary Differential Equations.* Introductory Mathematics for Scientists and Engineers. Wiley, New York, 1973.

[80] S. Larsson. The long-time behavior of finite-element approximations of solutions to semilinear parabolic problems. *SIAM J. Numer. Anal.*, 26(2):348, 1989.

[81] A. Lerat and J. Sidès. Efficient solution of the steady Euler equations with a centered implicit method. In K.W. Morton and M.J. Baines, editors, *Numerical Methods in Fluid Dynamics III.* Clarendon Press, Oxford, 1988.

[82] R.J. LeVeque and J. Oliger. Numerical methods based on additive splittings for partial differential equations. *Math. Comp.*, 40(162):469, 1983.

[83] J. Lighthill. *Waves in Fluids.* Cambridge University Press, Cambridge, 1978.

[84] M.J. Lighthill and G.B. Whitham. On kinematic waves. I. Flood movement in long rivers. *Proc. Roy. Soc. A*, 229:281, 1955.

[85] S.G. Lipson and H. Lipson. *Optical Physics.* Cambridge University Press, Cambridge, 1955.

[86] R. Löhner. *Finite Element Methods for Hyperbolic Partial Differential Equations.* PhD thesis, University College of Swansea, 1984.

[87] O. Lopes. On the structure of the spectrum of a linear time periodic wave equation. *J. d'Analyse Math.*, 47:55, 1986.

[88] O. Lopes. Perturbation of the spectrum of evolution operators. *J. Operator Theory*, 19:365, 1988.

[89] J.C. López-Marcos and J.M. Sanz-Serna. Stability and convergence in numerical analysis II: The definition of stability in non-linear problems. preprint.

[90] J. Mackenzie. The cell vertex method for viscous transport problems. Technical Report 89/4, Oxford University Computing Laboratory, Numerical Analysis Group, 1989.

[91] A. Majda and S. Osher. Initial-boundary value problems for hyperbolic equations with uniformly characteristic boundary. *Comm. Pure Appl. Math.*, 28:607, 1975.

[92] V.A. Malyšev. On the solution of discrete Wiener–Hopf equations in a quarter-plane. *Soviet Math. Dokl.*, 10(4):1032, 1969.

[93] J.T. Marti. *Introduction to Sobolev Spaces and Finite Element Solution of Elliptic Boundary Value Problems.* Computational Mathematics and Applications. Academic Press, London, 1986.

[94] K.W. Morton. Generalised Galerkin methods for first-order hyperbolic problems. *J. Comp. Phys.*, 36(2):249, 1980.

[95] K.W. Morton. Generalised Galerkin methods for hyperbolic problems. *Comput. Meth. Appl. Mech. Eng.*, 52:847, 1985.

[96] K.W. Morton and M.F. Paisley. On the cell-centre and cell-vertex approaches to the steady Euler equations and the use of shock fitting. In F.G. Zhuang and Y.L. Zhu, editors, *Tenth International Conference on Numerical Methods in Fluid Dynamics*, volume 264 of *Lecture Notes in Physics*, Heidelberg, 1986. Springer–Verlag.

[97] K.W. Morton and M.F. Paisley. A finite volume scheme with shock fitting for the steady Euler equations. *J. Comp. Phys.*, 80(1):168, 1989.

[98] K.W. Morton and M.A. Rudgyard. Finite volume methods with explicit shock representation. Technical Report 89/11, Oxford University Computing Laboratory, Numerical Analysis Group, 1989.

[99] K.W. Morton and S. Schechter. On the stability of finite difference matrices. *J. SIAM Numer. Anal., Ser. B*, 2(1):119, 1965.

[100] A.F. Neves, H. de Souza Ribeiro, and O. Lopes. On the spectrum of evolution operators generated by hyperbolic systems. *J. Funct. Anal.*, 67:320, 1986.

[101] S. Osher. Systems of difference equations with general homogeneous boundary conditions. *Trans. Amer. Math. Soc.*, 137:177, 1969.

[102] S. Osher. On certain Toeplitz operators in two variables. *Pacific J. Math.*, 34(1):123, 1970.

[103] S. Osher. Stability of parabolic difference approximations to certain mixed initial boundary value problems. *Math. Comp.*, 26(117):13, 1972.

[104] S. Osher. Discrete potential theory and Toeplitz operators on the quarter-plane. I. *Indiana Univ. Math. J.*, 24(9):887, 1975.

[105] C. Palencia and J.M. Sanz-Serna. Equivalence theorems for incomplete spaces: An appraisal. *IMA J. Numer. Anal.*, 4:109, 1984.

[106] C. Palencia and J.M. Sanz-Serna. An extension of the Lax–Richtmyer theory. *Numer. Math.*, 44:279, 1984.

[107] H.-O. Peitgen and P.H. Richter. *The Beauty of Fractals.* Springer–Verlag, Heidelberg, 1986.

[108] D.H. Peregrine, editor. *Floods Due to High Winds and Tides.* Institute of Mathematics and its Applications Conference Series. Academic Press, London, 1981.

[109] R. Peyret and T.D. Taylor. *Computational Methods for Fluid Flow.* Springer Series in Computational Physics. Springer–Verlag, New York, 1983.

[110] T.H. Pulliam. Implicit methods in CFD. In K.W. Morton and M.J. Baines, editors, *Numerical Methods in Fluid Dynamics III.* Clarendon Press, Oxford, 1988.

[111] T.H. Pulliam and J.L. Steger. Implicit finite-difference simulations of three-dimensional compressible flow. *AIAA J.*, 18(2):159, 1980.

[112] Zheng Q.C. and Zhang X. *Scientia Sinica*, 25:866, 1982.

[113] J. Ralston. Local decay of solutions of conservative first order hyperbolic systems in odd dimensional space. *Trans A.M.S.*, 194:27, 1974.

[114] J.V. Ralston. Note on a paper of Kreiss. *Comm. Pure Appl. Math.*, XXIV:759, 1971.

[115] J. Rauch. L_2 is a continuable initial condition for Kriess' mixed problems. *Comm. Pure Appl. Math.*, XXV:265, 1972.

[116] J. Rauch. Asymptotic behavior of solutions to hyperbolic partial differential equations. *Comm. Pure Appl. Math.*, 31:431, 1978.

[117] J. Rauch and M. Taylor. Exponential decay of solutions to hyperbolic equations in bounded domains. *Indiana University Mathematics Journal*, 24(1):79, 1974.

[118] J. Rauch and M. Taylor. Decay of solutions to nondissipative hyperbolic systems on compact manifolds. *Comm. Pure Appl. Math.*, XXVIII:501, 1975.

[119] L.F. Richardson. *Weather Prediction by Numerical Process.* Cambridge University Press, London, 1922.

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

[121] A. Rizzi and B. Engquist. Selected topics in the theory and practice of computational fluid dynamics. *J. Comp. Phys.*, 72(1):1, 1987.

[122] A.J. Robert. The integration of a spectral model of the atmosphere by the implicit method. In *Proceedings WMO/IUGG Symposium on Numerical Weather Prediction.* Meteorological Society of Japan, Tokyo, 1969.

[123] K.V. Roberts and N.O. Weiss. Convective difference schemes. *Math. Comp.*, 20(94), 1966.

[124] D. Rudy and J. Strikwerda. Boundary conditions for subsonic compressible Navier–Stokes calculations. *Computers and Fluids*, 9:327, 1981.

[125] D.J. Salmond. Calculation of aerodynamic forces and pressures on harmonically oscillating aerofoils from the Euler equations. Technical Report TR85051, Royal Aircraft Establishment, 1985.

[126] J.M. Sanz-Serna. Stability and convergence in numerical analysis I: Linear problems, a simple, comprehensive account. In J.K. Hale and P. Martínez-Amores, editors, *Nonlinear Differential Equations.* Pitman, 1985. Research Notes in Mathematics, 132.

[127] J.M. Sanz-Serna and C. Palencia. A general equivalence theorem in the theory of discretization methods. *Math. Comp.*, 45(171):143, 1985.

[128] J.M. Sanz-Serna and A.M. Stuart. A note on uniform in time error estimates for approximations to reaction-diffusion equations. Technical Report 1990/7, Universidad de Valladolid, Applied Mathematics and Computation Reports, October 1990.

[129] M. Schechter. *Spectra of Partial Differential Operators*, volume 14 of *North–Holland Series in Applied Mathematics and Mechanics.* North–Holland, Amsterdam, 1971.

[130] R.E. Showalter. *Hilbert Space Methods for Partial Differential Equations.* Pitman Advanced Publishing Program. Pitman, San Fransisco, 1978.

[131] J. Sidès. Implicit time dependent methods for the solution of the Euler equations. Technical report, von Karman Institute for Fluid Dynamics, 1984. Lecture Series 1984–4.

[132] M. Slemrod. The flip-flop circuit as a neutral equation. In K. Schmitt, editor, *Delay and Functional Differential Equations and Their Applications*, London, 1972. Academic Press.

[133] I.J. Sobey. On flow through furrowed channels. Part 1. Calculated flow patterns. *J. Fluid Mech.*, 96(1):1, 1980.

[134] I.J. Sobey. The occurrence of separation in oscillatory flow. *J. Fluid Mech.*, 134:247, 1983.

[135] G.A. Sod. *Numerical Methods in Fluid Dynamics: Initial and Initial Boundary-Value Problems.* Cambridge University Press, Cambridge, 1985.

[136] A. Sommerfeld. *Optics*, volume IV of *Lectures on Theoretical Physics.* Academic Press, New York, 1964.

[137] J.L. Steger. Implicit finite-difference simulation of flow about arbitrary two-dimensional geometries. *AIAA J.*, 16(7):679, 1978.

[138] J.L. Steger and R.F. Warming. Flux vector splitting of the equations of the inviscid gas dynamic equations with applications to finite difference methods. Technical Report TM–78605, NASA, 1979.

[139] G. Stelling, A. Wiersma, and J.B. Willemse. Practical aspects of accurate tidal computations. *J. Hydraul. Eng. ASCE*, 112:802, 1986.

[140] K.D. Stephanoff, I.J. Sobey, and B.J. Bellhouse. On flow through furrowed channels. Part 2. Observed flow patterns. *J. Fluid Mech.*, 96(1):27, 1980.

[141] P. Stow. CFD applications to the aero-thermodynamics of turbomachinery. In K.W. Morton and M.J. Baines, editors, *Numerical Methods in Fluid Dynamics III.* Clarendon Press, Oxford, 1988.

[142] G. Strang. Wiener–Hopf difference equations. *J. Math. Mech.*, 13(1):85, 1964.

[143] G. Strang. Implicit difference methods for initial-boundary value problems. *J. Math. Anal. Appl.*, 16:188, 1966.

[144] G. Strang. Toeplitz operators in a quarter-plane. *Bull. Amer. Math. Soc.*, 76:1303, 1970.

[145] E. Süli. Finite volume methods on distorted meshes: Stability, accuracy, adaptivity. Technical Report 89/6, Oxford University Computing Laboratory, Numerical Analysis Group, 1989.

[146] R. Timman, A.I. van de Vooren, and J.H. Greidanus. Aerodynamic coefficients of an oscillating airfoil in two-dimensional subsonic flow. *J. Aeronaut. Sci.*, 18(12):797, 1951.

[147] L.N. Trefethen. Group velocity in finite difference schemes. *SIAM Rev.*, 24(2):113, 1982.

[148] L.N. Trefethen. Group Velocity Interpretation of the Stability Theory of Gustafsson, Kreiss and Sundström. *J. Comp. Phys.*, 49(2):199, 1983.

[149] L.N. Trefethen. Dispersion, dissipation and stability. In D.F. Griffiths and G.A. Watson, editors, *Numerical Analysis.* Pitman, London, 1986. Research Notes in Mathematics Series, 140.

[150] T.L. van Stijn, P. Wilders, G.A. Fokkema, and G.S. Stelling. Properties of two computational methods for the shallow water flow problems. In K.W. Morton and M.J. Baines, editors, *Numerical Methods in Fluid Dynamics III.* Clarendon Press, Oxford, 1988.

[151] O. Vejvoda. *Partial differential equations: time-periodic solutions.* Martinus Nijhoff Publishers, The Hague, 1982.

[152] R. Vichnevetsky. Wave propagation analysis of difference schemes for hyperbolic equations: A review. *Int. J. Numer. Meth. Fluids*, 7:409, 1987.

[153] H. Viviand. Technical Report 1984–69, ONERA, 1984.

[154] C.B. Vreugdenhil. *Computational Hydraulics: An Introduction.* Springer–Verlag, Berlin, 1989.

[155] J.A. Walker. *Dynamical Systems and Evolution Equations*, volume 20 of *Mathematical concepts and methods in Science and Engineering.* Plenum Press, New York, 1980.

[156] S.J. Wardrop. Stability of implicit finite difference schemes for the solution of IBVPs. Technical Report 87/1, Oxford University Computing Laboratory, Numerical Analysis Group, 1987.

[157] W.M. Washington and C.L. Parkinson. *Introduction to Three-Dimensional Climate Modeling.* University Science Books (Oxford University Press), Oxford, 1986.

[158] T.J. Weare. Errors arising from irregular boundaries in ADI solutions of the shallow-water equations. *Int. J. Numer. Meth. Eng.*, 14:921, 1979.

[159] P. Wilders, Th.L. van Stijn, G.S. Stelling, and G.A. Fokkema. A fully implicit splitting method for accurate tidal computations. *Int. J. Numer. Meth. Eng.*, 26:2707, 1988.

[160] D.M. Young. *Iterative Solutions of Large Linear Systems.* Computer Science and Applied Mathematics. Academic Press, New York, 1971.

... and so the bird of heaven, with archangelic shrieks, and his imperial beak thrust upwards, and his whole captive form folded in the flag of Ahab, went down with his ship, which, like Satan, would not sink to hell till she had dragged a living part of heaven along with her, and helmetted herself with it.

Moby Dick